EXAMTOPICS

-Expert Verified, Online, **Free**.



CERTIFICATION TEST

-CertificationTest.net - Cheap & Quality Resources With Best Support

Question #1 *Topic 1*

You have a table in an Azure Synapse Analytics dedicated SQL pool. The table was created by using the following Transact-SQL statement.
You need to alter the table to meet the following requirements:
✑ Ensure that users can identify the current manager of employees.
✑ Support creating an employee reporting hierarchy for your entire company.
✑ Provide fast lookup of the managers' attributes such as name and job title.
Which column should you add to the table?

```
CREATE TABLE [dbo].[DimEmployee](
  [EmployeeKey] [int] IDENTITY(1,1) NOT NULL,
  [EmployeeID] [int] NOT NULL,
  [FirstName] [varchar](100) NOT NULL,
  [LastName] [varchar](100) NOT NULL,
  [JobTitle] [varchar](100) NULL,
  [LastHireDate] [date] NULL,
  [StreetAddress] [varchar](500) NOT NULL,
  [City] [varchar](200) NOT NULL,
  [StateProvince] [varchar](50) NOT NULL,
  [Portalcode] [varchar](10) NOT NULL
  )
```

A. [ManagerEmployeeID] [smallint] NULL

B. [ManagerEmployeeKey] [smallint] NULL

C. [ManagerEmployeeKey] [int] NULL

D. [ManagerName] [varchar](200) NULL

Suggested Answer: *C*
We need an extra column to identify the Manager. Use the data type as the EmployeeKey column, an int column. Reference:
https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular

Community vote distribution
C (99%)

jskibick Highly Voted 1 year, 11 months ago
Selected Answer: C
Answer C. Smallint eliminates A and B. But I would name the field [ManagerEmployeeID] [int] NULL since it should reference EmployeeID, not EmployeeKey since this one is IDENTITY.
upvoted 53 times

jskibick Highly Voted 1 year, 11 months ago
Answer C. Smallint eliminates A and B. But I would name the field [ManagerEmployeeID] [int] NULL since it should reference EmployeeID, not EmployeeKey since this one is IDENTITY.Why is smallint not possible. It would still support more than enough range to list all the employees in the company and save space + speed up the lookup.Smallint is up to 32,767: https://docs.microsoft.com/en-us/sql/t-sql/data-types/int-bigint-smallint-and-tinyint-transact-sql?view=sql-server-ver16Because the ManagerEmployeeKey should match the Data type of EmployeeKey which is INTIDENTITY keyword suggests that it's a surrogate key and it's a good practice to have a separate field for the surrogate key which is not a business field like emp no or emp name
upvoted 53 times

azure900test 4 years, 1 month ago
Why is smallint not possible. It would still support more than enough range to list all the employees in the company and save space + speed up the lookup.Smallint is up to 32,767: https://docs.microsoft.com/en-us/sql/t-sql/data-types/int-bigint-smallint-and-tinyint-transact-sql?view=sql-server-ver16Because the ManagerEmployeeKey should match the Data type of EmployeeKey which is INT
upvoted 3 times

azure900test 4 years, 1 month ago

Why is smallint not possible. It would still support more than enough range to list all the employees in the company and save space + speed up the lookup.

upvoted 3 times

azure900test 4 years, 1 month ago

Smallint is up to 32,767: https://docs.microsoft.com/en-us/sql/t-sql/data-types/int-bigint-smallint-and-tinyint-transact-sql?view=sql-server-ver16

upvoted 3 times

yogiazaad 3 years, 6 months ago

Because the ManagerEmployeeKey should match the Data type of EmployeeKey which is INT

upvoted 10 times

Chandra2099 3 years, 9 months ago

IDENTITY keyword suggests that it's a surrogate key and it's a good practice to have a separate field for the surrogate key which is not a business field like emp no or emp name

upvoted 5 times

vigilante89 3 years, 8 months ago

The answer is [ManagerEmployeeKey] [int] NULL because its a foreign key in the DimEmployee table connected to the primary key of the DimManager table. This column can be null because an employee can be a manager and there is no one he/she is reporting to.

upvoted 8 times

tavellie Most Recent 1 year, 3 months ago

can be use to join the other table with manager info

upvoted 1 times

Sayani 1 year, 6 months ago

B EmployeeID is businesskey, C EmployeeKey is audit key. for Reporting audit key will not help

upvoted 1 times

Rayenwalid 1 year, 7 months ago

using a smallint for the EmployeeKey could limit the number of employees you can represent in the future. It's generally safer to use int for larger datasets.

upvoted 1 times

krishna1303 1 year, 7 months ago

the manager is also an employee of the company, in employee I,d, they have given int so we have to give int.

upvoted 1 times

EmnCours 1 year, 8 months ago

Selected Answer: C

upvoted 1 times

loveHT1204 1 year, 9 months ago

To ensure we can identify the manager, managerEmployeeKey or ID must have specific characteristics, such as '123456acb...'. The TINYINT type in SQL Server 2022 ranges from -32,768 to 32,767, which is not suitable for both managerEmployeeKey and ID. Therefore, only C is suitable for this case

upvoted 1 times

mavdplas 1 year, 9 months ago

Answer C.

upvoted 1 times

vigilante89 1 year, 10 months ago

Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage. So DimEmployee and DimManager tables should be the dimension tables in the DW. The answer is [ManagerEmployeeKey] [int] NULL because its the foreign key in the DimEmployee table connected to the primary key of the DimManager table. This column can be null because an employee can be a manager and there is no one he/she is reporting to.

upvoted 1 times

gingerbread_person 1 year, 10 months ago

Based on the case study we need a dimension table(For Managers) and a foreign key in the Employee table that links to the dim table. Its implied that the EmployeeKey is the primary key for the employee table for the simple fact that the identity function increments the numbers for each row in the table. So based off this you can infer that you need to add a foreign key to the Employee table which would be called ManagerEmployeeKey.

upvoted 4 times

UdayKaudgaonkar 1 year, 11 months ago

Smallint can be useful, upto limited range beyond it's defined limit it will not and then need to alter the key to INT.

upvoted 1 times

AnujPanwar 2 years ago

I think C

upvoted 1 times

Anwana 2 years, 10 months ago

You need a key to link the Manager to an employee

upvoted 1 times

74gjd_37 2 years, 11 months ago

This is implied from the schema

upvoted 2 times

kkk5566 2 years, 11 months ago

c is correct

upvoted 1 times

cloudrain 3 years, 1 month ago

Manager is also an employee so the column type should match.

upvoted 2 times

Question #2 *Topic 1*

You have an Azure Synapse workspace named MyWorkspace that contains an Apache Spark database named mytestdb.
You run the following command in an Azure Synapse Analytics Spark pool in MyWorkspace.
CREATE TABLE mytestdb.myParquetTable(
EmployeeID int,
EmployeeName string,
EmployeeStartDate date)
USING Parquet -
You then use Spark to insert a row into mytestdb.myParquetTable. The row contains the following data.
One minute later, you execute the following query from a serverless SQL pool in MyWorkspace.
SELECT EmployeeID -
FROM mytestdb.dbo.myParquetTable
WHERE EmployeeName = 'Alice';
What will be returned by the query?

EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

A. 24

B. an error

C. a null value

Suggested Answer: *A*
Once a database has been created by a Spark job, you can create tables in it with Spark that use Parquet as the storage format. Table names will be converted to lower case and need to be queried using the lower case name. These tables will immediately become available for querying by any of the Azure Synapse workspace Spark pools. They can also be used from any of the Spark jobs subject to permissions. Note: For external tables, since they are synchronized to serverless SQL pool asynchronously, there will be a delay until they appear. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/metadata/table

Community vote distribution
A (56%) B (44%)

gerrie1979 Highly Voted 3 years, 9 months ago
I did a test, waited for one minute and tried the query in a serverless sql pool and received 24 as the result, so I don't understand that B has been voted so much because the answer is A) 24 without a doubt
upvoted 63 times

gerrie1979 Highly Voted 3 years, 9 months ago
I did a test, waited for one minute and tried the query in a serverless sql pool and received 24 as the result, so I don't understand that B has been voted so much because the answer is A) 24 without a doubti test too and confirm that the right answer is ADid you tried the same query that is presented here? with "mytestdb.dbo.myParquetTable"??The table and Column names are case insensitive.I tried with all upper case, and it still return record for name Alice. Answer is A
upvoted 63 times

cecbc1f 2 years, 5 months ago
i test too and confirm that the right answer is A
upvoted 4 times

maximilianogarcia6 3 years, 9 months ago
Did you tried the same query that is presented here? with "mytestdb.dbo.myParquetTable"??
upvoted 4 times

maximilianogarcia6 3 years, 9 months ago
Did you tried the same query that is presented here? with "mytestdb.dbo.myParquetTable"??The table and Column names are case insensitive.I tried with all upper case, and it still return record for name Alice. Answer is A
upvoted 4 times

yogiazaad 3 years, 6 months ago

The table and Column names are case insensitive.

upvoted 3 times

Virul 3 years, 6 months ago

I tried with all upper case, and it still return record for name Alice. Answer is A

upvoted 6 times

dmitriypo 3 years, 10 months ago

Answer is B, but not because of the lowercase. The case has nothing to do with the error. If you look attentively, you will notice that we create table mytestdb.myParquetTable, but the select statement contains the reference to table mytestdb.dbo.myParquetTable (!!! - dbo). Here is the error message I got: Error: spark_catalog requires a single-part namespace, but got [mytestdb, dbo].i think B option is the correct tooi think you don't about sql server bro, Dbo means database object so it is not a issue for this the correct answer is ADbo means database owner actually brokindly clarify, which can be the right option? the conversations are confusing. :(any explanations are appreciated. thank you!!I just tried to run the commands, and that error you had is due to the fact that you queried through Spark pool (!!), I did that as a test and got the exact same error. To query the data using Spark Pool, you don't use the ".dbo" reference, this only works if you're using a Synapse Serverless Pool. So the correct answer is A!is the answer A or B?24 . Because you are using SQL serverless pool to query the data. As the last line.

upvoted 56 times

dmitriypo 3 years, 10 months ago

Answer is B, but not because of the lowercase. The case has nothing to do with the error. If you look attentively, you will notice that we create table mytestdb.myParquetTable, but the select statement contains the reference to table mytestdb.dbo.myParquetTable (!!! - dbo). Here is the error message I got: Error: spark_catalog requires a single-part namespace, but got [mytestdb, dbo].

upvoted 56 times

devnginx 2 years, 9 months ago

i think B option is the correct too

upvoted 1 times

Shaik_Shahul 2 years, 10 months ago

i think you don't about sql server bro, Dbo means database object so it is not a issue for this the correct answer is ADbo means database owner actually bro

upvoted 6 times

Shaik_Shahul 2 years, 10 months ago

i think you don't about sql server bro, Dbo means database object so it is not a issue for this the correct answer is A

upvoted 6 times

__Tom 2 years, 4 months ago

Dbo means database owner actually bro

upvoted 3 times

SenMia 2 years, 8 months ago

kindly clarify, which can be the right option? the conversations are confusing. :(any explanations are appreciated. thank you!!

upvoted 2 times

tlb_20 3 years, 1 month ago

I just tried to run the commands, and that error you had is due to the fact that you queried through Spark pool (!!), I did that as a test and got the exact same error. To query the data using Spark Pool, you don't use the ".dbo" reference, this only works if you're using a Synapse Serverless Pool. So the correct answer is A!is the answer A or B?24 . Because you are using SQL serverless pool to query the data. As the last line.

upvoted 6 times

tlb_20 3 years, 1 month ago

I just tried to run the commands, and that error you had is due to the fact that you queried through Spark pool (!!), I did that as a test and got the exact same error. To query the data using Spark Pool, you don't use the ".dbo" reference, this only works if you're using a Synapse Serverless Pool. So the correct answer is A!

upvoted 6 times

bblord 3 years, 1 month ago

is the answer A or B?

upvoted 2 times

bblord 3 years, 1 month ago

is the answer A or B?24 . Because you are using SQL serverless pool to query the data. As the last line.

upvoted 2 times

darkraijin 2 years, 11 months ago

24 . Because you are using SQL serverless pool to query the data. As the last line.

upvoted 5 times

sam9162 Most Recent 1 year, 1 month ago

In Azure Synapse, Parquet tables created in Spark are accessible from SQL pools—as long as the schema and permissions align. The one-minute delay between insert and query is sufficient for metadata synchronization between Spark and SQL pool. The query syntax is valid and the table exists with the correct data. So, the SQL pool successfully finds EmployeeName = 'Alice' and returns her EmployeeID.

upvoted 1 times

tavellie 1 year, 3 months ago

should be A

upvoted 1 times

technoguy 1 year, 4 months ago

A is correct dbo is default schema for a table.

upvoted 1 times

AMJB 1 year, 5 months ago

The clue is in the name apache cluster database. The database is mytestdb . Basically you can create a table with . . it will default to the dbo schema

upvoted 1 times

Lethahavm 1 year, 5 months ago

Since there's no configuration in the given scenario to link the Spark-managed Parquet table to the serverless SQL pool, an error will occur when attempting to execute the SQL query.

upvoted 1 times

Ciske92 1 year, 6 months ago

I think A is the right answer. Firstly, dbo is added automatically (see examples in documentation). Furthermore, even though the created table in the spark pool will be saved with the name in lower case (as is stated in the documentation), SQL serverless pool in Synapse is case insensitive per default wehn it comes to table names and column names.

upvoted 1 times

Rayenwalid 1 year, 7 months ago

I believe A is correct because the selection query occurs after the insertion operation.

upvoted 1 times

Romanx 1 year, 7 months ago

dbo should not be there.

upvoted 1 times

Asheesh1909 1 year, 7 months ago

option A is correct tested it , there is no issue with the case , since the table and db names are saved in the lowercase , what ever case we use in the query azure / sql engine convert that into lower case before performing the query . 2. issue with .dbo the table is created in the spark pool , the table is store in the hive metadata . so without using the .dbo (schema layer) the serverless sql pool , cannot access the table and gives error . since we are using mytestdb.dbo.myparquetTable --> the serverless poll serches for the table in both its schema layer and hive schema layer giving the output 24 . Also , once a table is created in spark pool , it can be accessed in serverless sql Pool without any issues; there is no need for an external table to be created .

upvoted 1 times

abhi_11kr1 1 year, 8 months ago

The query will fail because the serverless SQL pool cannot directly access Spark-managed tables without additional configuration. Use an external table or expose the data via a view to make it accessible.you should at least test it out before posting wrong answers

upvoted 1 times

abhi_11kr1 1 year, 8 months ago

Selected Answer: B

The query will fail because the serverless SQL pool cannot directly access Spark-managed tables without additional configuration. Use an external table or expose the data via a view to make it accessible.

upvoted 1 times

hypersam 1 year, 7 months ago

you should at least test it out before posting wrong answers

upvoted 2 times

19a3424 1 year, 8 months ago

The example here shows https://learn.microsoft.com/en-us/azure/synapse-analytics/metadata/table#create-an-external-table-in-spark-and-query-from-serverless-sql-pool

upvoted 1 times

EmnCours 1 year, 8 months ago

Selected Answer: A

upvoted 1 times

Anithec0der 1 year, 9 months ago

Option A: dbo is the default schema where the object gets created in synapse if not specified explicitly.

upvoted 1 times

BrilliantBeast 1 year, 9 months ago

Does the order of insertion doesn't matter here? As the column order is different in the table than the one used in insertion.

upvoted 1 times

examdemo 1 year, 10 months ago

24 is correct answer Link : https://learn.microsoft.com/en-us/azure/synapse-analytics/metadata/table#create-a-managed-table-in-spark-and-query-from-serverless-sql-pool

upvoted 1 times

DRAG DROP -

You have a table named SalesFact in an enterprise data warehouse in Azure Synapse Analytics. SalesFact contains sales data from the past 36 months and has the following characteristics:

- ✑ Is partitioned by month
- ✑ Contains one billion rows
- ✑ Has clustered columnstore index

At the beginning of each month, you need to remove data from SalesFact that is older than 36 months as quickly as possible.

Which three actions should you perform in sequence in a stored procedure? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

Actions	**Answer Area**
Switch the partition containing the stale data from SalesFact to SalesFact_Work.	
Truncate the partition containing the stale data.	
Drop the SalesFact_Work table.	
Create an empty table named SalesFact_Work that has the same schema as SalesFact.	
Execute a `DELETE` statement where the value in the Date column is more than 36 months ago.	
Copy the data to a new table by using CREATE TABLE AS SELECT (CTAS).	

Suggested Answer:

Actions	Answer Area
Switch the partition containing the stale data from SalesFact to SalesFact_Work.	Create an empty table named SalesFact_Work that has the same schema as SalesFact.
Truncate the partition containing the stale data.	Switch the partition containing the stale data from SalesFact to SalesFact_Work.
Drop the SalesFact_Work table.	Drop the SalesFact_Work table.
Create an empty table named SalesFact_Work that has the same schema as SalesFact.	
Execute a `DELETE` statement where the value in the Date column is more than 36 months ago.	
Copy the data to a new table by using CREATE TABLE AS SELECT (CTAS).	

Actions	Answer Area
Switch the partition containing the stale data from SalesFact to SalesFact_Work.	Create an empty table named SalesFact_Work that has the same schema as SalesFact.
Truncate the partition containing the stale data.	Switch the partition containing the stale data from SalesFact to SalesFact_Work.
Drop the SalesFact_Work table.	Drop the SalesFact_Work table.
Create an empty table named SalesFact_Work that has the same schema as SalesFact.	
Execute a `DELETE` statement where the value in the Date column is more than 36 months ago.	
Copy the data to a new table by using CREATE TABLE AS SELECT (CTAS).	

Step 1: Create an empty table named SalesFact_work that has the same schema as SalesFact. Step 2: Switch the partition containing the stale data from SalesFact to SalesFact_Work. SQL Data Warehouse supports partition splitting, merging, and switching. To switch partitions between two tables, you must ensure that the partitions align on their respective boundaries and that the table definitions match. Loading data into partitions with partition switching is a convenient way stage new data in a table that is not visible to users the switch in the new data. Step 3: Drop the SalesFact_Work table. Reference: https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-partition

hsetin Highly Voted 4 years, 11 months ago

Given answer D A C is correct.Yes. Once the partition is switched with an empty partition it is equivalent to truncating the partition from the original table

upvoted 49 times

hsetin Highly Voted 4 years, 11 months ago

Given answer D A C is correct.

upvoted 49 times

svik 4 years, 11 months ago

Yes. Once the partition is switched with an empty partition it is equivalent to truncating the partition from the original table

upvoted 4 times

vigilante89 1 year, 10 months ago

The answer should be D --> A --> C. Step 1: Create an empty table SalesFact_Work with same schema as SalesFact. Step 2: Switch the partition (to be removed) from SalesFact to SalesFact_Work. The syntax is: ALTER TABLE SWITCH PARTITION to Step 3: Delete the SalesFact_Work table.

upvoted 13 times

technoguy Most Recent 1 year, 4 months ago

DAC is simpler abd best approah here

upvoted 1 times

EmnCours 1 year, 8 months ago

FAC is correct https://learn.microsoft.com/fr-fr/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition

upvoted 1 times

vrodriguesp 1 year, 10 months ago

Given answer is correct: Step 1: Create an empty table SalesFact_Work with same schema as SalesFact (that it will contains reocords older than 3 years) Step 2: Switch the partition from SalesFact to SalesFact_Work. SO we're only doing metadata operations Step 3: Delete the SalesFact_Work table containing stale data and we're not losing any time or blocking target table

upvoted 4 times

the_frix 1 year, 10 months ago

"Partition switching can be used to quickly remove or replace a section of a table. For example, a sales fact table might contain just data for the past 36 months. At the end of every month, the oldest month of sales data is deleted from the table. This data could be deleted by using a delete statement to delete the data for the oldest month. However, deleting a large amount of data row-by-row with a delete statement can take too much time, as well as create the risk of large transactions that take a long time to rollback if something goes wrong. A more optimal approach is to drop the oldest partition of data. Where deleting the individual rows could take hours, deleting an entire partition could take seconds." https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition#benefits-to-loads

upvoted 5 times

nani191 1 year, 11 months ago

FOR 3RD STEP: why should we drop the table instead of truncating? so that we can use it for subsequent months right?

upvoted 1 times

PavanPusarapu 2 years, 4 months ago

Is given answer correct (D, A, C) ?

upvoted 1 times

valenrami5 2 years, 4 months ago

Why not use 'ALTER TABLE' to avoid the need for creating a new table next time and simply switch the partition instead?

upvoted 1 times

kkk5566 2 years, 11 months ago

correct

upvoted 1 times

rocky48 3 years, 1 month ago

Given answer - Options D A C are correct.

upvoted 2 times

UzairMir 3 years, 3 months ago

Hi Can someone tell me why we cannot simply execute a delete statement? ThanksDELETE is intensive for a database to run. The solution should be faster as possible.https://learn.microsoft.com/en-us/training/modules/analyze-optimize-data-warehouse-storage-azure-synapse-analytics/10-understand-rules-for-minimally-logged-operations

upvoted 3 times

UzairMir 3 years, 3 months ago

Hi Can someone tell me why we cannot simply execute a delete statement? Thanks

upvoted 3 times

Xarvastia 3 years, 3 months ago

DELETE is intensive for a database to run. The solution should be faster as possible.

upvoted 3 times

Aaron1234 3 years, 3 months ago

https://learn.microsoft.com/en-us/training/modules/analyze-optimize-data-warehouse-storage-azure-synapse-analytics/10-understand-rules-for-minimally-logged-operations

upvoted 2 times

Dusica 3 years, 7 months ago

D A C IS CORRECT

upvoted 2 times

Jawidkaderi 3 years, 10 months ago

very interesting questions: Every partition has a name, which indicated by the mmYYYY perhaps. So, if we know the name of the partition, we can drop that partition directly: DROP PARTITION SCHEME partition_scheme_name [;] However, if there is an index on the table DOPR Partition will not

work. So, the it is correct. DAC.

upvoted 1 times

pmc08 3 years, 11 months ago

Answer is F - A - C https://docs.microsoft.com/es-es/archive/blogs/apsblog/azure-sql-dw-performance-ctaspartition-switching-vs-updatedelete

upvoted 2 times

pmc08 3 years, 11 months ago

Answer is F - A - C https://docs.microsoft.com/es-es/archive/blogs/apsblog/azure-sql-dw-performance-ctaspartition-switching-vs-updatedeleteD is incorrect because we also need to copy the data onto the new tableF seems wrong as it says CTAS to copy the data

upvoted 2 times

pmc08 3 years, 11 months ago

D is incorrect because we also need to copy the data onto the new table

upvoted 1 times

supriyako 3 years, 10 months ago

F seems wrong as it says CTAS to copy the data

upvoted 1 times

Deeksha1234 4 years ago

correct ans.

upvoted 2 times

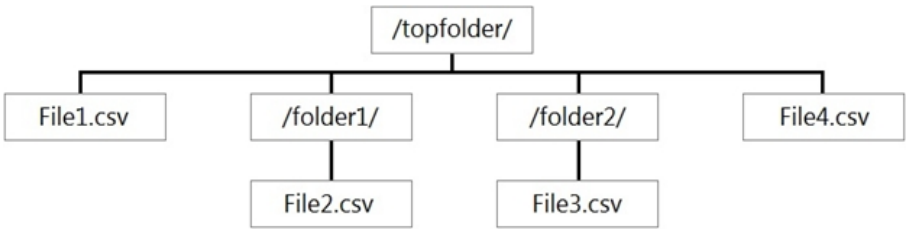
mkthoma3 4 years, 2 months ago

D,A,C Azure Synapse does not support truncating partitions. Currently, that is feature is only tied to MS SQL Server.

upvoted 1 times

Question #4 *Topic 1*

You have files and folders in Azure Data Lake Storage Gen2 for an Azure Synapse workspace as shown in the following exhibit.
You create an external table named ExtTable that has LOCATION='/topfolder/'.
When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, which files are returned?



A. File2.csv and File3.csv only

B. File1.csv and File4.csv only

C. File1.csv, File2.csv, File3.csv, and File4.csv

D. File1.csv only

Suggested Answer: *C*
To run a T-SQL query over a set of files within a folder or set of folders while treating them as a single entity or rowset, provide a path to a folder or a pattern (using wildcards) over a set of files or folders. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-data-storage#query-multiple-files-or-folders

Community vote distribution
B (73%) C (25%) 1

Chillem1900 Highly Voted 1 year, 11 months ago
I believe the answer should be B. In case of a serverless pool a wildcard should be added to the location. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#arguments-create-external-tableI tested and proove you right, the answer is B. Remind the question is referring to serverless SQL and not dedicated SQL pool. "Unlike Hadoop external tables, native external tables don't return subfolders unless you specify /** at the end of path. In this example, if LOCATION='/webdata/', a serverless SQL pool query, will return rows from mydata.txt. It won't return mydata2.txt and mydata3.txt because they're located in a subfolder. Hadoop tables will return all files within any subfolder."Serverless SQL pool can recursively traverse folders if you specify /** at the end of path. The following query will read all files from all folders and subfolders located in the csv/taxi folder. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-folders-multiple-csv-files
upvoted 124 times

Chillem1900 Highly Voted 1 year, 11 months ago
I believe the answer should be B. In case of a serverless pool a wildcard should be added to the location. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#arguments-create-external-table
upvoted 124 times

captainpike 4 years, 10 months ago
I tested and proove you right, the answer is B. Remind the question is referring to serverless SQL and not dedicated SQL pool. "Unlike Hadoop external tables, native external tables don't return subfolders unless you specify /** at the end of path. In this example, if LOCATION='/webdata/', a serverless SQL pool query, will return rows from mydata.txt. It won't return mydata2.txt and mydata3.txt because they're located in a subfolder. Hadoop tables will return all files within any subfolder."
upvoted 41 times

anthony854 2 years, 3 months ago
Serverless SQL pool can recursively traverse folders if you specify /** at the end of path. The following query will read all files from all folders and subfolders located in the csv/taxi folder. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-folders-multiple-csv-files
upvoted 5 times

alain2 1 year, 11 months ago
"Serverless SQL pool can recursively traverse folders only if you specify /** at the end of path." https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-folders-multiple-csv-files
upvoted 31 times

alain2 1 year, 11 months ago

"Serverless SQL pool can recursively traverse folders only if you specify /** at the end of path." https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-folders-multiple-csv-filesThe answer is B however. I could not make "/**" to work. somebody?When you are quoting from Microsoft documentation, do not ADD in words to the sentence. 'Only' is not used.

upvoted 31 times

captainpike 4 years, 10 months ago

The answer is B however. I could not make "/**" to work. somebody?

upvoted 4 times

Preben 5 years, 2 months ago

When you are quoting from Microsoft documentation, do not ADD in words to the sentence. 'Only' is not used.

upvoted 22 times

tavellie Most Recent 1 year, 3 months ago

Serverless SQL pool can recursively traverse folders only if you specify the name of folder

upvoted 1 times

technoguy 1 year, 4 months ago

Selected Answer: D

because we have not given wildcard expression

upvoted 1 times

technoguy 1 year, 4 months ago

because we have not given wildcard expressionsorry it should be B

upvoted 1 times

technoguy 1 year, 4 months ago

sorry it should be B

upvoted 1 times

ayn1 1 year, 5 months ago

you do not need to use LOCATION='/topfolder/*' to read files in subfolders when using external tables in serverless SQL pools. You use wildcards like '/topfolder/*' when querying directly using OPENROWSET

upvoted 1 times

ngabonzic 1 year, 5 months ago

Correct answer is B

upvoted 1 times

AMJB 1 year, 5 months ago

If you specify LOCATION to be a folder, a PolyBase query that selects from the external table will retrieve files from the folder and all of its subfolders. Just like Hadoop, PolyBase doesn't return hidden folders. It also doesn't return files for which the file name begins with an underline (_) or a period (.).

upvoted 1 times

Pey1nkh 1 year, 6 months ago

When you query ExtTable with LOCATION='/topfolder/', only the files directly inside /topfolder/ will be returned.

upvoted 1 times

krishna1303 1 year, 7 months ago

the answer should be B

upvoted 1 times

EmnCours 1 year, 8 months ago

The answer is A, the answer C corresponds to LOCATION='/topfolder/**'

upvoted 1 times

seranvijay 1 year, 9 months ago

https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-folders-multiple-csv-files#traverse-folders-recursively

upvoted 1 times

medsimus 1 year, 10 months ago

Correct answer , I test it and it works : CREATE EXTERNAL TABLE ext2 ([User] varchar(40), Feature varchar(40), Event varchar(40) , Time varchar(40)) WITH (LOCATION = '/topfolder', DATA_SOURCE = ds2, FILE_FORMAT = csvFormat) GO SELECT COUNT(*) as cnt FROM ext2 GOThis is for serverless SQL. What you showed is for dedicated SQL, behaviour is different as explained here: "Unlike Hadoop external tables, native external tables don't return subfolders unless you specify /** at the end of path. In this example, if LOCATION='/webdata/', a serverless SQL pool query, will return rows from mydata.txt. It won't return mydata2.txt and mydata3.txt because they're located in a subfolder. Hadoop tables will return all files within any subfolder."

upvoted 2 times

medsimus 1 year, 10 months ago

Correct answer , I test it and it works :

CREATE EXTERNAL TABLE ext2 ([User] varchar(40), Feature varchar(40), Event varchar(40) , Time varchar(40)) WITH (LOCATION = '/topfolder', DATA_SOURCE = ds2, FILE_FORMAT = csvFormat) GO SELECT COUNT(*) as cnt FROM ext2 GO

upvoted 2 times

captainpike 4 years, 10 months ago

This is for serverless SQL. What you showed is for dedicated SQL, behaviour is different as explained here: "Unlike Hadoop external tables, native external tables don't return subfolders unless you specify /** at the end of path. In this example, if LOCATION='/webdata/', a serverless SQL pool query, will return rows from mydata.txt. It won't return mydata2.txt and mydata3.txt because they're located in a subfolder. Hadoop tables will return all files within any subfolder."

upvoted 2 times

poundmanluffy 1 year, 10 months ago

Option is definitely "B" Below is the documentation given on MS Docs: Recursive data for external tables Unlike Hadoop external tables, native external tables don't return subfolders unless you specify /** at the end of path. In this example, if LOCATION='/webdata/', a serverless SQL pool query, will return rows from mydata.txt. It won't return mydata2.txt and mydata3.txt because they're located in a subfolder. Hadoop tables will return all files within any sub-folder.

upvoted 2 times

Raza12 1 year, 10 months ago

(Seems from this document that Answer is "C")In SQL Server, the CREATE EXTERNAL TABLE statement creates the path and folder if it doesn't already exist. You can then use INSERT INTO to export data from a local SQL Server table to the external data source. For more information, see PolyBase Queries. If you specify LOCATION to be a folder, a PolyBase query that selects from the external table will retrieve files from the folder and all of its subfolders. Just like Hadoop, PolyBase doesn't return hidden folders. It also doesn't return files for which the file name begins with an underline (_) or a period (.). In this example, if LOCATION='/webdata/', a PolyBase query will return rows from mydata.txt and mydata2.txt. It won't return mydata3.txt because it's a file in a hidden folder. And it won't return _hidden.txt because it's a hidden file. Recursive data for external tables https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=sql-server-ver15&tabs=dedicated

upvoted 1 times

rohitbinnani 1 year, 10 months ago

"native external tables don't return subfolders unless you specify /** at the end of path. In this example, if LOCATION='/webdata/', a serverless SQL pool query, will return rows from mydata.txt. It won't return mydata2.txt and mydata3.txt because they're located in a subfolder. Hadoop tables will return all files within any sub-folder." https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#arguments-create-external-table

upvoted 2 times

lesardinha 2 years, 1 month ago

I believe that the right answer is B, because from the root directory, there will be listed two files and two folders

upvoted 1 times

Question #5 *Topic 1*

HOTSPOT -

You are planning the deployment of Azure Data Lake Storage Gen2.

You have the following two reports that will access the data lake:

- Report1: Reads three columns from a file that contains 50 columns.
- Report2: Queries a single record based on a timestamp.

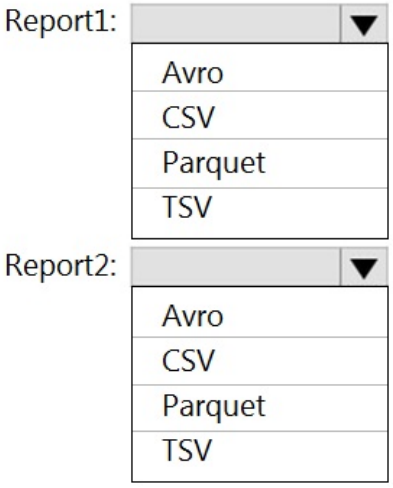
You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.

What should you recommend for each report? To answer, select the appropriate options in the answer area.

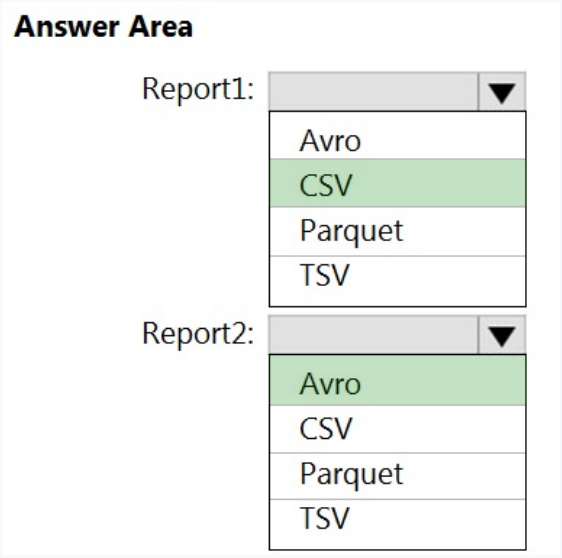
NOTE: Each correct selection is worth one point.

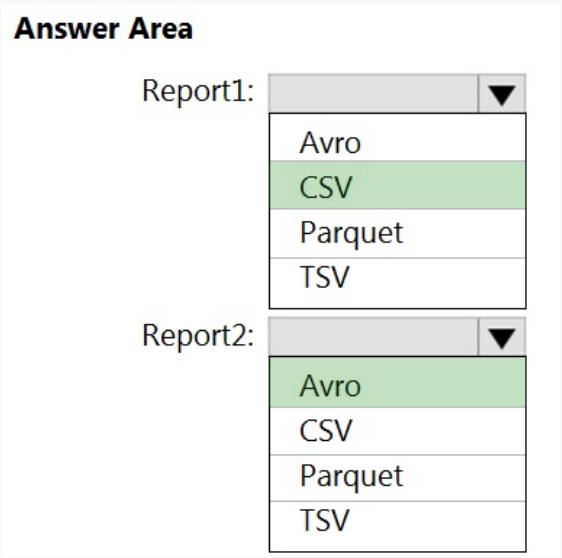
Hot Area:

Answer Area



Suggested Answer:





Report1: CSV - CSV: The destination writes records as delimited data.

Report2: AVRO - AVRO supports timestamps. Not Parquet, TSV: Not options for Azure Data Lake Storage Gen2. Reference: https://streamsets.com/documentation/datacollector/latest/help/datacollector/UserGuide/Destinations/ADLS-G2-D.html

alain2 Highly Voted 5 years, 3 months ago

1: Parquet - column-oriented binary file format 2: AVRO - Row based format, and has logical type timestamp https://youtu.be/UrWthx8T3UY

upvoted 224 times

alain2 Highly Voted 5 years, 3 months ago

1: Parquet - column-oriented binary file format 2: AVRO - Row based format, and has logical type timestamp https://youtu.be/UrWthx8T3UYthis is correct.Thanks for the video share, this really helps. Cheers.the web is full of old information. timestamp support has been added to parquetOk, but in 1st case we need only 3 of 50 columns. Parquet i columnar format. In 2nd Avro because ideal for read full rowThis is Correct since we only have to read three columns its best to use column based storage AVRO for report2 is fine because as per the AVRO format defintin it already got timestmp

upvoted 224 times

azurestudent1498 4 years, 4 months ago

this is correct.

upvoted 3 times

XiltroX 3 years, 9 months ago

Thanks for the video share, this really helps. Cheers.

upvoted 2 times

terajuana 5 years, 2 months ago

the web is full of old information. timestamp support has been added to parquetOk, but in 1st case we need only 3 of 50 columns. Parquet i columnar format. In 2nd Avro because ideal for read full row

upvoted 9 times

terajuana 5 years, 2 months ago

the web is full of old information. timestamp support has been added to parquet

upvoted 9 times

vlad888 5 years, 1 month ago

Ok, but in 1st case we need only 3 of 50 columns. Parquet i columnar format. In 2nd Avro because ideal for read full row

upvoted 25 times

technoguy Most Recent 1 year, 4 months ago

This is Correct since we only have to read three columns its best to use column based storage AVRO for report2 is fine because as per the AVRO format defintin it already got timestmp

upvoted 2 times

Himlo24 5 years, 3 months ago

Shouldn't the answer for Report 1 be Parquet? Because Parquet format is Columnar and should be best for reading a few columns only.

upvoted 33 times

ngabonzic 1 year, 5 months ago

1.Parquet 2.AVRO

upvoted 1 times

krishna1303 1 year, 7 months ago

report 1: Parquet (column-based file format optimized for reading specific columns) report 2:Avro(Row-based file format)

upvoted 1 times

ff5037f 1 year, 9 months ago

the answer should be Parquet for report 1

upvoted 1 times

marcin1212 1 year, 10 months ago

The goal is: The solution must minimize read times.

I made small test on Databrick plus DataLake. The same file saved as Parquet and Avro 9 mln of records. Parquet ~150 MB Avro ~700MB

Reading Parquet is always 10 times faster that Avro. I checked: - for all data or small range of data with condition - all or only one column

So I will select option: - Parquet - Parquet

upvoted 3 times

marcin1212 1 year, 10 months ago

The goal is: The solution must minimize read times. I made small test on Databrick plus DataLake. The same file saved as Parquet and Avro 9 mln of records. Parquet ~150 MB Avro ~700MB Reading Parquet is always 10 times faster that Avro. I checked: - for all data or small range of data with condition - all or only one column So I will select option: - Parquet - Parquethow can be faster read is same as number of reads?

upvoted 3 times

dev2dev 4 years, 7 months ago

how can be faster read is same as number of reads?

upvoted 1 times

ragz_87 1 year, 10 months ago

1. Parquet 2. Avro https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices "Consider using the Avro file format in cases where your I/O patterns are more write heavy, or the query patterns favor retrieving multiple rows of records in their entirety. Consider Parquet and ORC file formats when the I/O patterns are more read heavy or when the query patterns are focused on a subset of columns in the records."Thank you.

upvoted 9 times

ragz_87 1 year, 10 months ago

1. Parquet 2. Avro https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices

"Consider using the Avro file format in cases where your I/O patterns are more write heavy, or the query patterns favor retrieving multiple rows of records in their entirety.

Consider Parquet and ORC file formats when the I/O patterns are more read heavy or when the query patterns are focused on a subset of columns in the records."

upvoted 9 times

SebK 4 years, 5 months ago

Thank you.

upvoted 1 times

akk_1289 1 year, 10 months ago

To minimize read times for the two reports, it is recommended to store the data in the data lake in the parquet format. Parquet is a columnar storage format that is optimized for querying large datasets. It stores data in a compact and efficient manner, allowing for fast querying and filtering of data. In this case, Report1 needs to read only three columns from a file that contains 50 columns. Since parquet stores data in a columnar format, the query can skip reading the unnecessary columns and only read the required ones, which can greatly improve the read performance. Report2 needs to query a single record based on a timestamp. Parquet also supports efficient filtering and querying based on specific values, such as timestamps, making it a good choice for this report as well. Other formats, such as avro, csv, and tsv, may not provide the same level of performance for these types of queries. Therefore, it is recommended to use parquet to store the data in the data lake.

upvoted 9 times

lisa710 1 year, 10 months ago

Report 1: parquet Columnar Storage: Parquet stores data in columns, allowing efficient reading of only the required columns (3 out of 50). Report 2: Avro. Fast Single-Record Access: Optimized row-based formats excel at quickly accessing individual records based on a specific condition, such as a timestamp. I don't understand why incorrect answers are being provided, causing confusion.

upvoted 1 times

ypan 1 year, 10 months ago

Recommendations: Report1: Use Parquet Reason: Parquet is a columnar storage format that is optimized for reading specific columns. Since Report1 needs to read only three columns out of 50, Parquet allows reading just those columns efficiently without scanning the entire file. Report2: Use Avro Reason: Avro is a row-based storage format, which is efficient for retrieving entire rows based on a specific condition, such as a timestamp. It allows quick access to individual records, making it suitable for Report2's requirement.

upvoted 3 times

ItsPayakan 1 year, 11 months ago

Why cant it be Parquet for both reports? Parquet supports timestamp too

upvoted 1 times

207680a 2 years ago

1- Parquet 2- Avro Parquet for Column selection(stored in columnar format) and Avro for row selection (stored in row format)

upvoted 1 times

Vaibhav251999 2 years, 3 months ago

1. Parquet 2. Avro

upvoted 2 times

Alongi 2 years, 5 months ago

Parquet Avro

upvoted 1 times

mykel71 2 years, 6 months ago

1. Parquet - Column oriented data store 2. AVRO - supports timestamps

upvoted 1 times

sdg2844 2 years, 7 months ago

Agree: 1: Parquet - ideal for columnar forma 2: AVRO: Row-based with logical timestamp

upvoted 2 times

matiandal 2 years, 10 months ago

--> TLDR <-- AVRO PARQUET ORC Anal. Queries v v Write Ops (ETL ops) v Nested Data v ACID Properties v Sch.Flexibility v

upvoted 1 times

Question #6 *Topic 1*

You are designing the folder structure for an Azure Data Lake Storage Gen2 container.
Users will query data by using a variety of services including Azure Databricks and Azure Synapse Analytics serverless SQL pools. The data will be secured by subject area. Most queries will include data from the current year or current month.
Which folder structure should you recommend to support fast queries and simplified folder security?

A. /{SubjectArea}/{DataSource}/{DD}/{MM}/{YYYY}/{FileData}_{YYYY}_{MM}_{DD}.csv

B. /{DD}/{MM}/{YYYY}/{SubjectArea}/{DataSource}/{FileData}_{YYYY}_{MM}_{DD}.csv

C. /{YYYY}/{MM}/{DD}/{SubjectArea}/{DataSource}/{FileData}_{YYYY}_{MM}_{DD}.csv

D. /{SubjectArea}/{DataSource}/{YYYY}/{MM}/{DD}/{FileData}_{YYYY}_{MM}_{DD}.csv

Suggested Answer: *D*
There's an important reason to put the date at the end of the directory structure. If you want to lock down certain regions or subject matters to users/groups, then you can easily do so with the POSIX permissions. Otherwise, if there was a need to restrict a certain security group to viewing just the UK data or certain planes, with the date structure in front a separate permission would be required for numerous directories under every hour directory. Additionally, having the date structure in front would exponentially increase the number of directories as time went on. Note: In IoT workloads, there can be a great deal of data being landed in the data store that spans across numerous products, devices, organizations, and customers. It's important to pre-plan the directory layout for organization, security, and efficient processing of the data for down-stream consumers. A general template to consider might be the following layout: {Region}/{SubjectMatter(s)}/{yyyy}/{mm}/{dd}/{hh}/

Community vote distribution
D (96%) (4

sagga Highly Voted 5 years, 3 months ago
D is correct https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices#batch-jobs-structure
upvoted 63 times

technoguy Most Recent 1 year, 4 months ago
since we are querying b MM or year and data is secured by subject area. its better that we design the folder structure that does minimal scan of folder. here in this case we should first create year , month nd then service area
upvoted 1 times

ngabonzic 1 year, 5 months ago
D is correct
upvoted 1 times

EmnCours 1 year, 8 months ago
D is correct
upvoted 1 times

207680a 2 years ago
D - Secured by subject area so top level folder security, Year first to avoid too many subfolder/partitions by each date
upvoted 1 times

sdg2844 2 years, 7 months ago
Definitely D.
upvoted 1 times

dakku987 2 years, 8 months ago
bcz i said so
upvoted 2 times

kkk5566 2 years, 11 months ago
is correct
upvoted 1 times

akhil5432 3 years ago
D is correct
upvoted 1 times

tbluhm 3 years, 3 months ago

D Is the best way, filter subject area first the size queries will reduced just one and after by date.

upvoted 1 times

steveo123 3 years, 3 months ago

Should be D.

upvoted 2 times

pedr0oliveira 3 years, 4 months ago

D is correct

upvoted 1 times

DindaS 3 years, 6 months ago

yes its D :)

upvoted 1 times

vigilante89 3 years, 8 months ago

Serverless SQL Pools offers a straight-forward method of querying data including CSV, JSON, and Parquet format stored in Azure Storage. So, setting up the csv files within azure storage in hive-formated folder hierarchy i.e. /{yyyy}/{mm}/{dd}/ actually helps in sql querying the data much faster since only the partitioned segment of the data is queried.

upvoted 3 times

Fernando_Caemerer 3 years, 8 months ago

D is correct

upvoted 1 times

Deeksha1234 4 years ago

correct

upvoted 1 times

examtopicscap 4 years ago

D is correct

upvoted 1 times

HOTSPOT -

You need to output files from Azure Data Factory.

Which file format should you use for each type of output? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

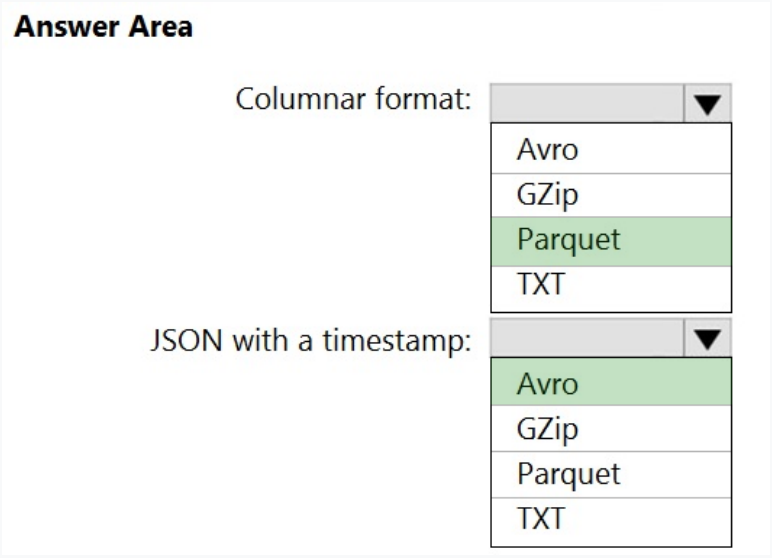
Answer Area

Columnar format: [▼]
- Avro
- GZip
- Parquet
- TXT

JSON with a timestamp: [▼]
- Avro
- GZip
- Parquet
- TXT

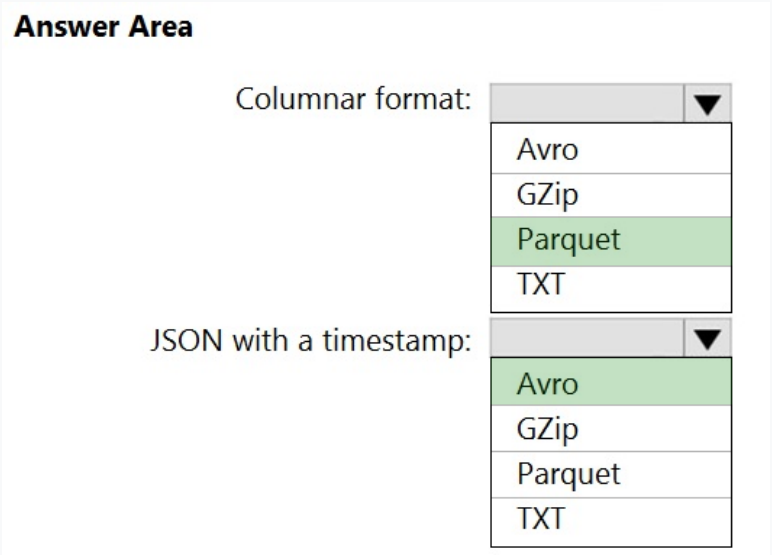
Suggested Answer:

Answer Area



.

Answer Area



Box 1: Parquet - Parquet stores data in columns, while Avro stores data in a row-based format. By their very nature, column-oriented data stores are optimized for read-heavy analytical workloads, while row-based databases are best for write-heavy transactional workloads.

Box 2: Avro - An Avro schema is created using JSON format. AVRO supports timestamps. Note: Azure Data Factory supports the following file formats (not GZip or TXT).

Avro format -

➪ Binary format ➪ Delimited text format ➪ Excel format ➪ JSON format ➪ ORC format ➪ Parquet format ➪ XML format Reference: https://www.datanami.com/2018/05/16/big-data-file-formats-demystified

Mahesh_mm Highly Voted 3 years, 7 months ago

Parquet and AVRO is correct option

upvoted 44 times

krishna1303 Most Recent 1 year, 7 months ago

Parquet and AVRO

upvoted 2 times

h4y 1 year, 8 months ago

Parquet and Avro are correct

upvoted 1 times

hassexat 1 year, 11 months ago

The provided answer is correct: Parquet & AVRO

upvoted 2 times

kkk5566 1 year, 11 months ago

1.parquet 2avro

upvoted 1 times

rocky48 2 years, 3 months ago

1: PARQUET Because Parquet is a columnar file format. 2: AVRO Because Avro is a row-based file format (as JSON) which is connected to logical timestamp.

upvoted 4 times

bubby248 2 years, 6 months ago

Parquet,Avro

upvoted 1 times

vigilante89 2 years, 8 months ago

1: PARQUET Because Parquet is a columnar file format. 2: AVRO Because Avro is a row-based file format (as JSON) which is connected to logical timestamp.

upvoted 2 times

sppdw 2 years, 11 months ago

Parquet and AVRO is correct.

upvoted 1 times

Deeksha1234 3 years ago

correct

upvoted 1 times

Raza12 3 years, 1 month ago

Parquet and AVRO is correct option

upvoted 3 times

Dothy 3 years, 3 months ago

agree with the answer

upvoted 3 times

RalphLiang 3 years, 5 months ago

Parquet and AVRO is correct option

upvoted 2 times

PallaviPatel 3 years, 6 months ago

correct

upvoted 1 times

Skyrocket 3 years, 7 months ago

Parquet and AVRO is right.

upvoted 2 times

edba 3 years, 7 months ago

GZIP file format is one of supported Binary format by ADF. https://docs.microsoft.com/en-us/azure/data-factory/connector-file-system?tabs=data-factory#file-system-as-sink

upvoted 1 times

bad_atitude 3 years, 8 months ago

agree with the answer

upvoted 2 times

Question #8 *Topic 1*

HOTSPOT -

You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools.

Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company.

You need to move the files to a different folder and transform the data to meet the following requirements:

✑ Provide the fastest possible query times.

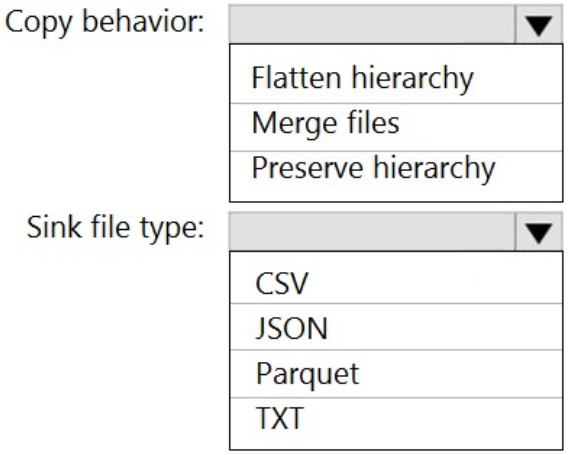
✑ Automatically infer the schema from the underlying files.

How should you configure the Data Factory copy activity? To answer, select the appropriate options in the answer area.

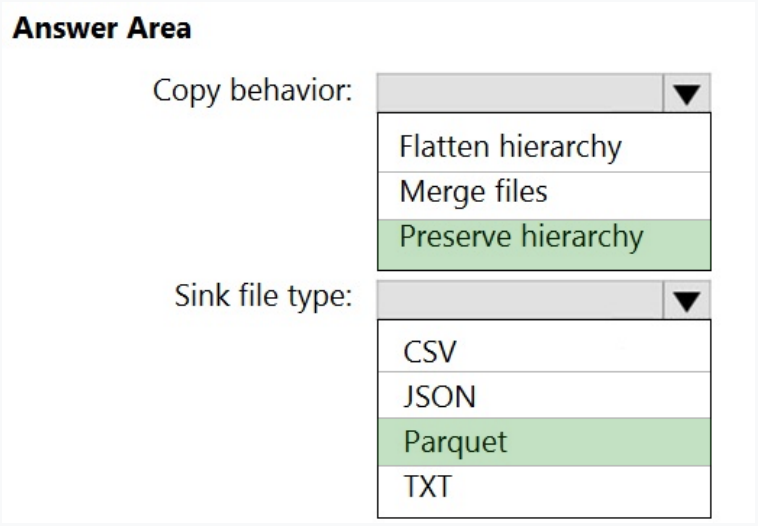
NOTE: Each correct selection is worth one point.

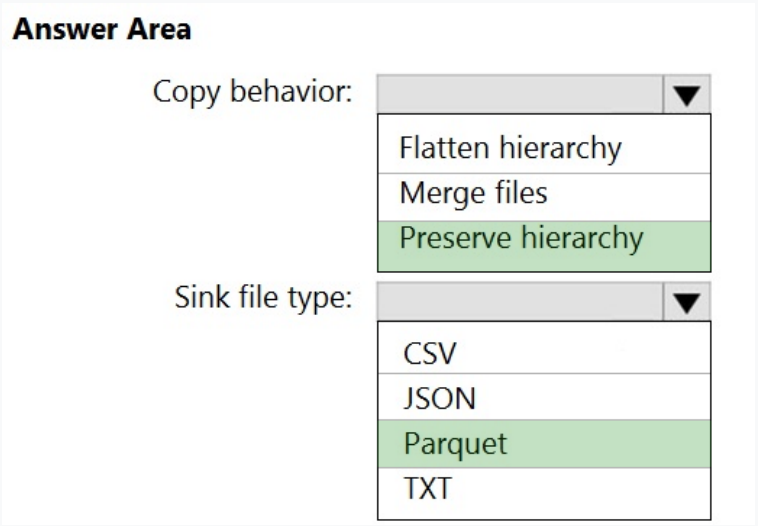
Hot Area:

Answer Area



Suggested Answer:





Box 1: Preserver hierarchy - Compared to the flat namespace on Blob storage, the hierarchical namespace greatly improves the performance of directory management operations, which improves overall job performance.

Box 2: Parquet - Azure Data Factory parquet format is supported for Azure Data Lake Storage Gen2. Parquet supports the schema property.
Reference: https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction https://docs.microsoft.com/en-us/azure/data-factory/format-parquet

alain2 Highly Voted 1 year, 11 months ago

1. Merge Files 2. Parquet https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-performance-tuning-guidance

upvoted 171 times

alain2 Highly Voted 1 year, 11 months ago

1. Merge Files 2. Parquet https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-performance-tuning-guidancejust want to add a bit more reference regarding copyBehavior in ADF plus info mentioned in Best Practice doc, so it shall be MergeFile first. https://docs.microsoft.com/en-us/azure/data-factory/connector-file-system?tabs=data-factory#file-system-as-sinkThe smaller the files, the negative the performance so Merge and Parquet seems to be the right answer.Larger files lead to better performance and reduced costs. Typically, analytics engines such as HDInsight have a per-file overhead that involves tasks such as listing, checking access, and performing various metadata operations. If you store your data as many small files, this can negatively affect performance. In general, organize your data into larger sized files for better performance (256 MB to 100 GB in size). S

upvoted 171 times

edba 4 years, 7 months ago

just want to add a bit more reference regarding copyBehavior in ADF plus info mentioned in Best Practice doc, so it shall be MergeFile first. https://docs.microsoft.com/en-us/azure/data-factory/connector-file-system?tabs=data-factory#file-system-as-sink

upvoted 12 times

Ameenymous 5 years, 2 months ago

The smaller the files, the negative the performance so Merge and Parquet seems to be the right answer.

upvoted 25 times

kilowd 4 years, 10 months ago

Larger files lead to better performance and reduced costs. Typically, analytics engines such as HDInsight have a per-file overhead that involves tasks such as listing, checking access, and performing various metadata operations. If you store your data as many small files, this can negatively affect performance. In general, organize your data into larger sized files for better performance (256 MB to 100 GB in size). S

upvoted 10 times

ThiruthuvaRajan 5 years, 2 months ago

It should be 1)Merge Files - Question clearly says "initially ingested as 10 small json files". There is no hint on hierarchy or partition information. so clearly we need to merge these files for better performance 2) Parquet -> Always gives better performance for columnar based data

upvoted 12 times

Sathya_sree Most Recent 1 year, 4 months ago

Copy Behavior → Merge Files Since there are 10 small JSON files, merging them into a single, optimized file will improve query performance. Azure Synapse Analytics serverless SQL pools perform better with fewer, larger files than many small files. Flatten hierarchy is not suitable since we are not restructuring directories. Preserve hierarchy is not needed because we want better query performance, not maintaining the folder structure. Sink File Type → Parquet Parquet is an efficient columnar format, designed for fast analytical queries. It supports schema inference, which is required for querying in Synapse serverless SQL pools. CSV and JSON are not optimal because they do not support columnar storage and would result in slower queries. TXT is not structured and would not allow schema inference. Thus, the best selections are "Merge files" for Copy behavior and "Parquet" for Sink file type to achieve the fastest query times

upvoted 1 times

Pey1nkh 1 year, 6 months ago

Azure Synapse serverless SQL pools perform best with fewer, larger files in Parquet format - so 1-Merge Files 2- Parquet

upvoted 1 times

KowshikKvt 1 year, 6 months ago

Q says "You need to move the files to a different folder". so preserve for sure

upvoted 1 times

krishna1303 1 year, 7 months ago

1. Copy Behavior: merge files 2. Sink file type: Paquet

upvoted 1 times

Karthikyn 1 year, 9 months ago

1. Preserve Hierarchy 2. Parquet Since Merge Files behavior combines multiple files into a single output file. While this can be useful for reducing the number of files, it may introduce additional performance overhead due to the merging process. Since it ask to provide the fastest possible query time.

upvoted 1 times

7082935 2 years ago

This question is somewhat confusing. The detail states "Each file contains the same data attributes and data from a subsidiary of your company". There is a possibility that each file is for a different subsidiary and would be stored in a different hierarchical folder, so the requirement to "preserve hierarchy" could be the right answer here.

upvoted 2 times

Bakachi55 2 years, 5 months ago

Dear Community, I would like to express my heartfelt gratitude for the thoughtful mock questions that have been shared. Your generosity in providing these valuable resources has been immensely helpful. As we engage in discussions and learn together, I am reminded of the strength and camaraderie that exists within our community. To everyone who has contributed, whether by creating questions, participating in discussions, or simply offering encouragement, thank you. Your collective efforts make this community a vibrant and supportive place for learning and growth. Let us continue to share knowledge, support one another, and celebrate our shared passion for learning

upvoted 5 times

waghsvw53 2 years, 5 months ago

https://learn.microsoft.com/en-us/azure/data-factory/connector-file-system?tabs=data-factory

upvoted 1 times

Azure_2023 2 years, 6 months ago

1. Flatten hierarchy - Performance: Fastest; Schema inference: Straightforward; 2. Parquet

upvoted 2 times

lisa710 2 years, 8 months ago

Merge Files: This option combines the 10 JSON files into a single Parquet file, reducing overhead and improving query performance significantly. Parquet: This columnar format is optimized for fast queries, especially when dealing with large datasets and selective column reads. It also supports

compression and schema inference.

upvoted 4 times

Chemmangat 2 years, 11 months ago

My answer : Merge Since there is no mention of preserving the hierarchy, and the need is to make the process more efficient, merge is the way to go.

upvoted 2 times

hassexat 2 years, 11 months ago

MERGE FILES since you need to make transformation and data have the same attributes PARQUET because is the most effcient file format

upvoted 2 times

kkk5566 2 years, 11 months ago

- Merge - Parquet

upvoted 2 times

ladistar 3 years ago

ChatGPT confirms it's 1. Merge, 2. Parquet

upvoted 5 times

rocky48 3 years, 3 months ago

1. Merge Files 2. Parquet https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-performance-tuning-guidance

upvoted 5 times

Question #9 *Topic 1*

HOTSPOT -

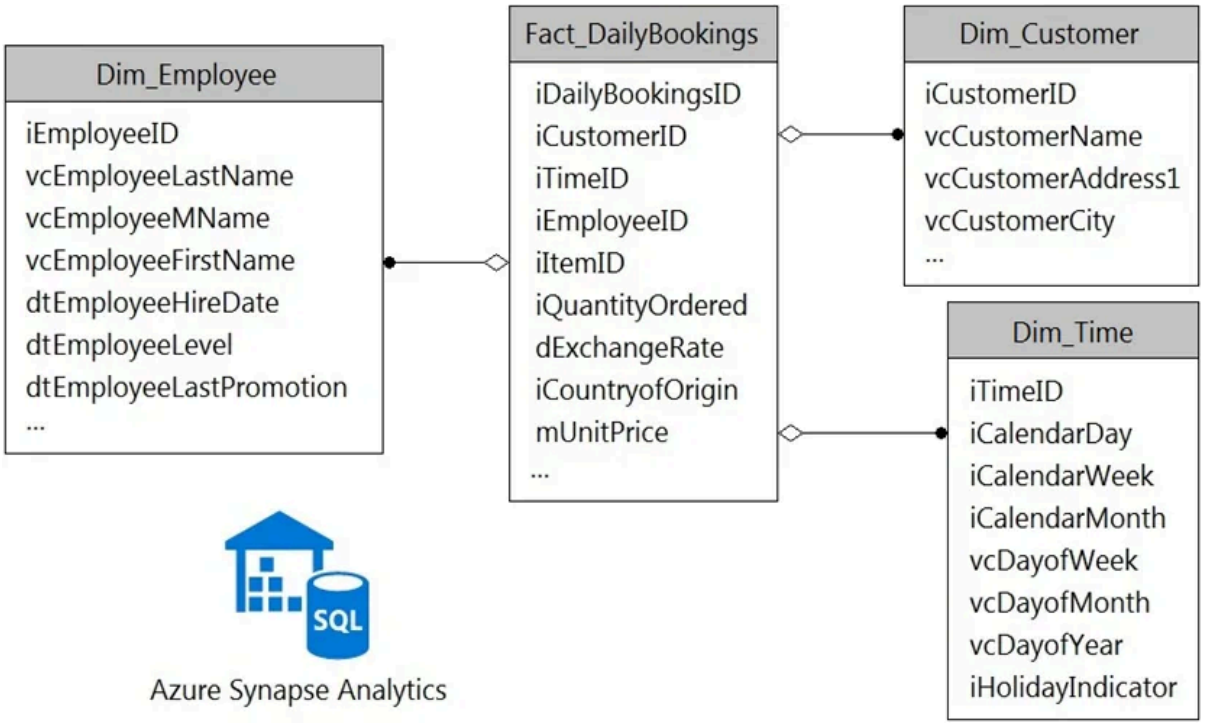
You have a data model that you plan to implement in a data warehouse in Azure Synapse Analytics as shown in the following exhibit.

All the dimension tables will be less than 2 GB after compression, and the fact table will be approximately 6 TB. The dimension tables will be relatively static with very few data inserts and updates.

Which type of table should you use for each table? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

Dim_Customer: ▼
- Hash distributed
- Round-robin
- Replicated

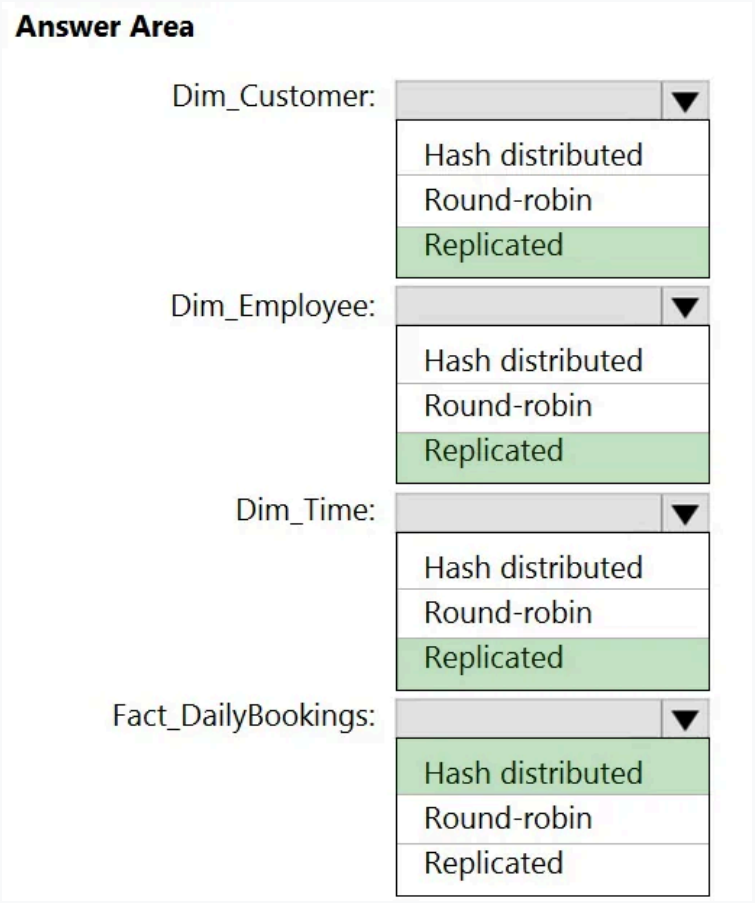
Dim_Employee: ▼
- Hash distributed
- Round-robin
- Replicated

Dim_Time: ▼
- Hash distributed
- Round-robin
- Replicated

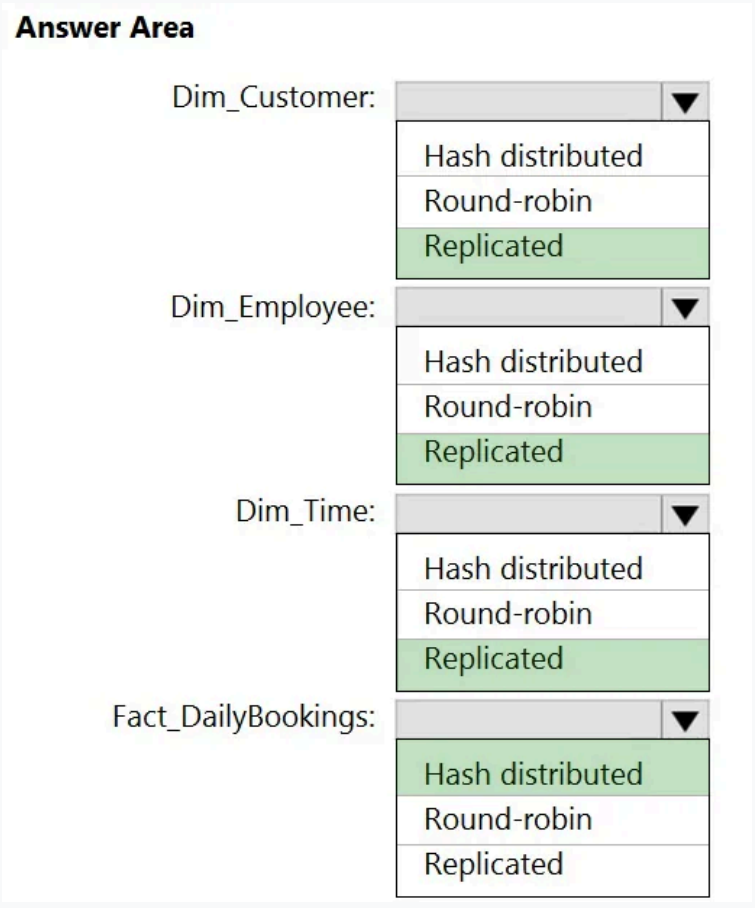
Fact_DailyBookings: ▼
- Hash distributed
- Round-robin
- Replicated

Suggested Answer:

Answer Area



Answer Area



Box 1: Replicated - Replicated tables are ideal for small star-schema dimension tables, because the fact table is often distributed on a column that is not compatible with the connected dimension tables. If this case applies to your schema, consider changing small dimension tables currently implemented as round-robin to replicated.

Box 2: Replicated -

Box 3: Replicated -

Box 4: Hash-distributed - For Fact tables use hash-distribution with clustered columnstore index. Performance improves when two hash tables are joined on the same distribution column. Reference: https://azure.microsoft.com/en-us/updates/reduce-data-movement-and-make-your-queries-more-efficient-with-the-general-availability-of-replicated-tables/ https://azure.microsoft.com/en-us/blog/replicated-tables-now-generally-available-in-azure-sql-data-warehouse/

ian_viana Highly Voted 4 years, 11 months ago

The answer is correct. The Dims are under 2gb so no point in use hash.

Common distribution methods for tables:

The table category often determines which option to choose for distributing the table. Table category Recommended distribution option Fact -Use hash-distribution with clustered columnstore index. Performance improves when two hash tables are joined on the same distribution column. Dimension - Use replicated for smaller tables. If tables are too large to store on each Compute node, use hash-distributed. Staging - Use round-robin for the staging table. The load with CTAS is fast. Once the data is in the staging table, use INSERT...SELECT to move the data to production tables. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overview#common-distribution-methods-for-tables

upvoted 240 times

ian_viana Highly Voted 4 years, 11 months ago

The answer is correct. The Dims are under 2gb so no point in use hash. Common distribution methods for tables: The table category often determines which option to choose for distributing the table. Table category Recommended distribution option Fact -Use hash-distribution with clustered columnstore index. Performance improves when two hash tables are joined on the same distribution column. Dimension - Use replicated for smaller tables. If tables are too large to store on each Compute node, use hash-distributed. Staging - Use round-robin for the staging table. The load with CTAS is fast. Once the data is in the staging table, use INSERT...SELECT to move the data to production tables. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overview#common-distribution-methods-for-tablesThanks, but where in the question does it indicate about Fact table has clustered columnstore index.?Normally for big tables we use clustered columnstore index for optimal performance and compression. Since the table mentioned here is in TBs we can safely assume using this index is the best choice https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-indexhttps://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-overviewCan you please explain for Dimension table it is mentioned that "If tables are too large" use Hash distribution. Here Too large means how much?I am waiting for your reply□□□□exceeding 10 gigabytes (GB) are often considered largeThis is a wonderful explanation. Worth giving a like.

upvoted 240 times

GameLift 4 years, 10 months ago

Thanks, but where in the question does it indicate about Fact table has clustered columnstore index.?

upvoted 3 times

GameLift 4 years, 10 months ago

Thanks, but where in the question does it indicate about Fact table has clustered columnstore index.?Normally for big tables we use clustered columnstore index for optimal performance and compression. Since the table mentioned here is in TBs we can safely assume using this index is the best choice https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-indexhttps://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-overview

upvoted 3 times

berserksap 4 years, 9 months ago

Normally for big tables we use clustered columnstore index for optimal performance and compression. Since the table mentioned here is in TBs we can safely assume using this index is the best choice https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-indexhttps://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-overview

upvoted 3 times

berserksap 4 years, 9 months ago

Normally for big tables we use clustered columnstore index for optimal performance and compression. Since the table mentioned here is in TBs we can safely assume using this index is the best choice https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-index

upvoted 3 times

berserksap 4 years, 9 months ago

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-overview

upvoted 1 times

Tara123 2 years, 8 months ago

Can you please explain for Dimension table it is mentioned that "If tables are too large" use Hash distribution. Here Too large means how much?I am waiting for your reply□□□□

upvoted 1 times

Tara123 2 years, 8 months ago

Can you please explain for Dimension table it is mentioned that "If tables are too large" use Hash distribution. Here Too large means how much?I am waiting for your reply□□□□exceeding 10 gigabytes (GB) are often considered large

upvoted 1 times

lisa710 2 years, 8 months ago

exceeding 10 gigabytes (GB) are often considered large

upvoted 5 times

virendrapsingh 4 years, 2 months ago

This is a wonderful explanation. Worth giving a like.

upvoted 7 times

ohana 4 years, 9 months ago

Took the exam today, this question came out. Ans: All the Dim tables --> Replicated Fact Tables --> Hash Distributed

upvoted 49 times

technoguy Most Recent 1 year, 4 months ago

Replicated Replicated Replicated Hash Dist

upvoted 1 times

krishna1303 1 year, 7 months ago

Dim_customers,Dim_Employees,Dim_time are replicated and Fact_DailyBooking table is hash distributed

upvoted 1 times

EmnCours 1 year, 8 months ago

The answer is correct.

upvoted 1 times

RookieCloudEngineer 2 years ago

this explains better https://www.youtube.com/watch?v=seiJ2xpW0h8

upvoted 1 times

Bakachi55 2 years, 5 months ago

Dear Community, I would like to express my heartfelt gratitude for the thoughtful mock questions that have been shared. Your generosity in providing these valuable resources has been immensely helpful. As we engage in discussions and learn together, I am reminded of the strength and camaraderie that exists within our community. To everyone who has contributed, whether by creating questions, participating in discussions, or simply offering encouragement, thank you. Your collective efforts make this community a vibrant and supportive place for learning and growth. Let us continue to share knowledge, support one another, and celebrate our shared passion for learning

upvoted 4 times

Alongi 2 years, 5 months ago

All Dim Tables, if less than 2 GB, will be REPLICATED. For Fact Table, if major than 10 GB, will be HASH DIST.

upvoted 3 times

dakku987 2 years, 8 months ago

REplicated,REplicated,REplicated and hash

upvoted 1 times

hassexat 2 years, 11 months ago

Replicated / Replicated / Replicated / Hash

upvoted 1 times

kkk5566 2 years, 11 months ago

Ans: All the Dim tables --> Replicated Fact Tables --> Hash Distributed is correct

upvoted 1 times

DataEngDP 3 years, 1 month ago

https://learn.microsoft.com/en-us/training/modules/design-multidimensional-schema-to-optimize-analytical-workloads/3-create-tables

upvoted 2 times

DataEngDP 3 years, 1 month ago

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices#batch-jobs-structure Dimension tables: Replicated distribution since data movement is absent here. Fact Table: Hash distribution to improve performance of moving data.

upvoted 1 times

vigilante89 3 years, 8 months ago

Dim_*: Replicated Since dimension tables are less likely to get frequent updates and are usually smaller in size, replicating them across all partitions makes logical sense. Also, Tables less than 2gb size should be replicated. Fact_*: Hash Distributed Since Fact tables are huge and have frequent insert/delete/updates going on, hash distribution is the perfect distribution candidate. Also, Tables greater than 2gb size should be hash distributed.

upvoted 3 times

Deeksha1234 4 years ago

correct

upvoted 1 times

objecto 4 years, 2 months ago

Just a better link that explains the decisions. Also watch the video, it's cool. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/massively-parallel-processing-mpp-architecture

upvoted 2 times

Dothy 4 years, 3 months ago

The answer is correct.

upvoted 1 times

HOTSPOT -

You have an Azure Data Lake Storage Gen2 container.

Data is ingested into the container, and then transformed by a data integration application. The data is NOT modified after that. Users can read files in the container but cannot modify the files.

You need to design a data archiving solution that meets the following requirements:

- ✑ New data is accessed frequently and must be available as quickly as possible.
- ✑ Data that is older than five years is accessed infrequently but must be available within one second when requested.
- ✑ Data that is older than seven years is NOT accessed. After seven years, the data must be persisted at the lowest cost possible.
- ✑ Costs must be minimized while maintaining the required availability.

How should you manage the data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point

Hot Area:

Answer Area

Five-year-old data:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Seven-year-old data:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Suggested Answer:

Answer Area

Five-year-old data:

- Delete the blob.
- Move to archive storage.
- **Move to cool storage.**
- Move to hot storage.

Seven-year-old data:

- Delete the blob.
- **Move to archive storage.**
- Move to cool storage.
- Move to hot storage.

	Premium performance	Hot tier	Cool tier	Archive tier
Availability	99.9%	99.9%	99%	Offline
Availability (RA-GRS reads)	N/A	99.99%	99.9%	Offline
Usage charges	Higher storage costs, lower access, and transaction cost	Higher storage costs, lower access, and transaction costs	Lower storage costs, higher access, and transaction costs	Lowest storage costs, highest access, and transaction costs
Minimum storage duration	N/A	N/A	30 days[1]	180 days
Latency (Time to first byte)	Single-digit milliseconds	milliseconds	milliseconds	hours[2]

Answer Area

Five-year-old data:

- Delete the blob.
- Move to archive storage.
- **Move to cool storage.**
- Move to hot storage.

Seven-year-old data:

- Delete the blob.
- **Move to archive storage.**
- Move to cool storage.
- Move to hot storage.

Box 1: Move to cool storage -

Box 2: Move to archive storage - Archive - Optimized for storing data that is rarely accessed and stored for at least 180 days with flexible latency requirements, on the order of hours. The following table shows a comparison of premium performance block blob storage, and the hot, cool, and archive access tiers.

Reference: https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blob-storage-tiers

yobllip Highly Voted 1 year, 11 months ago

Answer should be 1 - Cool 2 - Archive

Comparison table shown access time for cool tier ttfb is milliseconds

https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blob-storage-tiers#comparing-block-blob-storage-options

upvoted 94 times

yobllip Highly Voted 1 year, 11 months ago

Answer should be 1 - Cool 2 - Archive Comparison table shown access time for cool tier ttfb is milliseconds https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blob-storage-tiers#comparing-block-blob-storage-optionsRight. #1 is Cool because it's clearly mentioned in the documentation that "Older data sets that are not used frequently, but are expected to be available for immediate access" https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview#comparing-block-blob-storage-options

upvoted 94 times

r00s 4 years, 2 months ago

Right. #1 is Cool because it's clearly mentioned in the documentation that "Older data sets that are not used frequently, but are expected to be available for immediate access" https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview#comparing-block-blob-storage-options

upvoted 5 times

Justbu 4 years, 11 months ago

Tricky question, it says data that is OLDER THAN (> 5 years), must be available within one second when requested But the first question asks for Five-year-old data, which is =5, so it can also be hot storage Similarly for the seven-year-old. Not sure, please confirm?"Costs must be minimized while maintaining the required availability." So cold and archive are the correct answers.If data is frequently being used that is true for till 5 year old data => Hot Older than 5 year upto 10 year rarely used but still needed sometimes on moment notice => Cool Data older than 10 years out of the blue moon used=> Archive

upvoted 13 times

Justbu 4 years, 11 months ago

Tricky question, it says data that is OLDER THAN (> 5 years), must be available within one second when requested

But the first question asks for Five-year-old data, which is =5, so it can also be hot storage

Similarly for the seven-year-old.

Not sure, please confirm?

upvoted 13 times

joshfry 4 years ago

"Costs must be minimized while maintaining the required availability." So cold and archive are the correct answers.

upvoted 4 times

joshfry 4 years ago

"Costs must be minimized while maintaining the required availability." So cold and archive are the correct answers.If data is frequently being used that is true for till 5 year old data => Hot Older than 5 year upto 10 year rarely used but still needed sometimes on moment notice => Cool Data older than 10 years out of the blue moon used=> Archive

upvoted 4 times

Sr18 2 years, 3 months ago

If data is frequently being used that is true for till 5 year old data => Hot Older than 5 year upto 10 year rarely used but still needed sometimes on moment notice => Cool Data older than 10 years out of the blue moon used=> Archive

upvoted 1 times

krishna1303 Most Recent 1 year, 7 months ago

1-cool 2-archive

upvoted 1 times

EmnCours 1 year, 8 months ago

The answer is correct.

upvoted 1 times

Alongi 2 years, 5 months ago

Cool and Archive, correct answer

upvoted 1 times

lisa710 2 years, 8 months ago

Cool Tier: Cost-effective storage for less frequently accessed data, accessible within seconds when needed. Archive Tier: Lowest-cost storage for rarely accessed data, ideal for long-term preservation without immediate access requirements.

upvoted 2 times

hassexat 2 years, 11 months ago

1. Cool --> You will access infrequently but data must be avaliable in few time if you want to access them 2. Archive --> You will never acces them but you need to configure a data archiving solution, so you must retain them always and not delete the blob

upvoted 1 times

kkk5566 2 years, 11 months ago

1.cool and 2.archive

upvoted 1 times

akhil5432 3 years ago

cool and archive

upvoted 1 times

Andrew_Chen 3 years ago

Question: Why can't we just delete the blob to save more cost?

upvoted 2 times

janaki 3 years, 3 months ago

Answer: Cool & Archive

upvoted 2 times

vigilante89 3 years, 8 months ago

5-year old data: Cool Storage Cool Storage: can be retrieved and accessed within seconds/minutes. 7-year old data: Archive Storage Archive Storage: takes hours to retrieve and access the data. Hot Storage: can be retrieved and accessed within a few milliseconds.

upvoted 2 times

PugazhManohar 4 years ago

1.Cool, 2.Archive

upvoted 1 times

Dothy 4 years, 3 months ago

ans is correct

upvoted 1 times

PallaviPatel 4 years, 6 months ago

ans is correct

upvoted 1 times

ANath 4 years, 7 months ago

1. Cool Storage 2. Archive Storage

upvoted 1 times

Mahesh_mm 4 years, 7 months ago

Answer is correct

upvoted 1 times

Question #11 *Topic 1*

DRAG DROP -

You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

Values

- CLUSTERED INDEX
- COLLATE
- DISTRIBUTION
- PARTITION
- PARTITION FUNCTION
- PARTITION SCHEME

Answer Area

```
CREATE TABLE table1
(
 ID INTEGER,
 col1 VARCHAR(10),
 col2 VARCHAR(10)
) WITH
(
[                    ] = HASH(ID),
[                    ] (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
);
```

Suggested Answer:

Values

- CLUSTERED INDEX
- COLLATE
- DISTRIBUTION
- PARTITION
- PARTITION FUNCTION
- PARTITION SCHEME

Answer Area

```
CREATE TABLE table1
(
 ID INTEGER,
 col1 VARCHAR(10),
 col2 VARCHAR(10)
) WITH
(
DISTRIBUTION         = HASH(ID),
PARTITION            (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
);
```

Values

- CLUSTERED INDEX
- COLLATE
- DISTRIBUTION
- PARTITION
- PARTITION FUNCTION
- PARTITION SCHEME

Answer Area

```
CREATE TABLE table1
(
 ID INTEGER,
 col1 VARCHAR(10),
 col2 VARCHAR(10)
) WITH
(
DISTRIBUTION         = HASH(ID),
PARTITION            (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
);
```

Box 1: DISTRIBUTION - Table distribution options include DISTRIBUTION = HASH (distribution_column_name), assigns each row to one distribution by hashing the value stored in distribution_column_name.

Box 2: PARTITION - Table partition options. Syntax: PARTITION (partition_column_name RANGE [LEFT | RIGHT] FOR VALUES ([boundary_value [,...n]])) Reference: https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse ?

Sasha_in_San_Francisco Highly Voted 1 year, 11 months ago

Correct answer by how to remember? Distribution option before the Partition option because… 'D' comes before 'P' or because the system needs to know the algorithm (hash, round-robin, replicate) before it can start to Partition or segment the data. (seem reasonable?)

upvoted 123 times

Sunnyb Highly Voted 5 years, 2 months ago

Answer is correct

upvoted 67 times

krishna1303 Most Recent 1 year, 7 months ago

distribution and partition

upvoted 1 times

dgerok 2 years, 4 months ago

Distribution Partition

upvoted 2 times

hassexat 2 years, 11 months ago

Distribution & Partition

upvoted 1 times

kkk5566 2 years, 11 months ago

correct

upvoted 1 times

TestingCRM 3 years, 2 months ago

The answer is correct. Here you can find an example: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition#syntax-differences-from-sql-server

upvoted 3 times

vigilante89 3 years, 8 months ago

Answer is: DISTRIBUTION = HASH (id) PARTITION (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000)) ---------- The table option syntax for creating a partitioned table within Dedicated SQL pool: ::= { CLUSTERED COLUMNSTORE INDEX -- default for Azure Synapse Analytics } { DISTRIBUTION = HASH (distribution_column_name) } | PARTITION (partition_column_name RANGE [LEFT | RIGHT] -- default is LEFT FOR VALUES ([boundary_value [,...n]]))

upvoted 5 times

Deeksha1234 4 years ago

correct

upvoted 2 times

gursimran_s 4 years, 2 months ago

Go with a logical explanation guys..what is this D before P..if u take it like that then C comes before D as well.. Try to grasp the logics.. answer is correct.SAVAGE XD

upvoted 3 times

gursimran_s 4 years, 2 months ago

Go with a logical explanation guys..what is this D before P..if u take it like that then C comes before D as well.. Try to grasp the logics.. answer is correct.

upvoted 3 times

topgggggggg 2 years, 11 months ago

SAVAGE XD

upvoted 1 times

Dothy 4 years, 3 months ago

Answer is correct

upvoted 1 times

Egocentric 4 years, 4 months ago

provided answer is correct

upvoted 1 times

PallaviPatel 4 years, 6 months ago

correct

upvoted 1 times

Jaws1990 4 years, 7 months ago

Wouldn't VALUES(1,1000000, 200000) create a partition for records with ID = 2000000but only . it is range left for values, right

upvoted 1 times

Jaws1990 4 years, 7 months ago

Wouldn't VALUES(1,1000000, 200000) create a partition for records with ID <= 1 which would mean 1 row?

upvoted 1 times

ploer 4 years, 6 months ago

Having three boundaries will lead to four partitions: 1. Partition for values = 2000000

upvoted 3 times

ploer 4 years, 6 months ago

Having three boundaries will lead to four partitions: 1. Partition for values = 2000000but only . it is range left for values, right

upvoted 3 times

nastyaaa 4 years, 6 months ago

but only . it is range left for values, right

upvoted 2 times

Mahesh_mm 4 years, 7 months ago

Answer is correct

upvoted 1 times

hugoborda 4 years, 10 months ago

Answer is correct

upvoted 1 times

hsetin 4 years, 11 months ago

Indeed! Answer is correct

upvoted 1 times

Question #12 *Topic 1*

You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements:
➪ Can return an employee record from a given point in time.
➪ Maintains the latest employee information.
➪ Minimizes query complexity.
How should you model the employee data?

A. as a temporal table

B. as a SQL graph table

C. as a degenerate dimension table

D. as a Type 2 slowly changing dimension (SCD) table

Suggested Answer: *D*
A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members. Reference: https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types

Community vote distribution
D (100%)

rashjan Highly Voted 4 years, 8 months ago
D is correct (voting comment that people dont have to open discussion always, please upvote to help others)
upvoted 196 times

bc5468521 Highly Voted 5 years, 2 months ago
Answer D; Temporal table is better than SCD2, but it is not supported in Synpase yetHere's the documentation for how to implement temporal tables in Synapse from 2019. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-temporaryTemporal tables and Temporary tables are two very distinct concepts. Your link has absolutely nothing to do with this question.https://docs.microsoft.com/en-us/azure/azure-sql/temporal-tables Answer : A Temporal TablesSorry Answer is D: SCD 2 , as according to microsoft docs ,"Temporal tables keep data closely related to time context so that stored facts can be interpreted as valid only within the specific period." , as in the question it is mentioned "from a given point in time", so D seems to be the correct.though this not something relative to this question. temproal tables looks alike to delta table.Temporarytables is completely offtopic ,the catch here is ...at a point in time hence SCD is the way to gotemPORAL table purpose is to query point in time data which can perform beyond what SCD2 can do, but its not supported in synapse
upvoted 82 times

bc5468521 5 years, 2 months ago
Answer D; Temporal table is better than SCD2, but it is not supported in Synpase yet
upvoted 82 times

Preben 5 years, 2 months ago
Here's the documentation for how to implement temporal tables in Synapse from 2019.

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-temporary
upvoted 1 times

Preben 5 years, 2 months ago
Here's the documentation for how to implement temporal tables in Synapse from 2019. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-temporaryTemporal tables and Temporary tables are two very distinct concepts. Your link has absolutely nothing to do with this question.https://docs.microsoft.com/en-us/azure/azure-sql/temporal-tables Answer : A Temporal TablesSorry Answer is D: SCD 2 , as according to microsoft docs ,"Temporal tables keep data closely related to time context so that stored facts can be interpreted as valid only within the specific period." , as in the question it is mentioned "from a given point in time", so D seems to be the correct.
upvoted 1 times

mbravo 5 years, 2 months ago

Temporal tables and Temporary tables are two very distinct concepts. Your link has absolutely nothing to do with this question.https://docs.microsoft.com/en-us/azure/azure-sql/temporal-tables Answer : A Temporal TablesSorry Answer is D: SCD 2 , as according to microsoft docs ,"Temporal tables keep data closely related to time context so that stored facts can be interpreted as valid only within the specific period." , as in the question it is mentioned "from a given point in time", so D seems to be the correct.

upvoted 20 times

mbravo 5 years, 2 months ago

Temporal tables and Temporary tables are two very distinct concepts. Your link has absolutely nothing to do with this question.

upvoted 20 times

Vaishnav 5 years, 1 month ago

https://docs.microsoft.com/en-us/azure/azure-sql/temporal-tables Answer : A Temporal TablesSorry Answer is D: SCD 2 , as according to microsoft docs ,"Temporal tables keep data closely related to time context so that stored facts can be interpreted as valid only within the specific period." , as in the question it is mentioned "from a given point in time", so D seems to be the correct.

upvoted 2 times

Vaishnav 5 years, 1 month ago

https://docs.microsoft.com/en-us/azure/azure-sql/temporal-tables Answer : A Temporal Tables

upvoted 2 times

Vaishnav 5 years, 1 month ago

Sorry Answer is D: SCD 2 , as according to microsoft docs ,"Temporal tables keep data closely related to time context so that stored facts can be interpreted as valid only within the specific period." , as in the question it is mentioned "from a given point in time", so D seems to be the correct.

upvoted 4 times

sparkchu 4 years, 5 months ago

though this not something relative to this question. temproal tables looks alike to delta table.

upvoted 1 times

GodfreyMbizo 3 years, 6 months ago

Temporarytables is completely offtopic ,the catch here is ...at a point in time hence SCD is the way to go

upvoted 2 times

SolutionA 3 years ago

temPORAL table purpose is to query point in time data which can perform beyond what SCD2 can do, but its not supported in synapse

upvoted 1 times

EmnCours Most Recent 1 year, 8 months ago

Correct Answer: D

upvoted 1 times

mavdplas 1 year, 8 months ago

Answer: D

upvoted 1 times

GBA30 1 year, 9 months ago

Answer is D

upvoted 1 times

dakku987 2 years, 8 months ago

scd 2 maintain the iasactive current time end time like this in table

upvoted 2 times

Bail 2 years, 8 months ago

It's D

upvoted 2 times

hassexat 2 years, 11 months ago

D Because is the way as you can get historical data for the employee

upvoted 1 times

kkk5566 2 years, 11 months ago

correct

upvoted 2 times

Deeksha1234 3 years, 2 months ago

D is correct

upvoted 1 times

steveo123 3 years, 3 months ago

D is correct.

upvoted 2 times

haidebelognime 3 years, 6 months ago

it is D 100%, stop messing around with answers like "temporal table"

upvoted 5 times

bubby248 3 years, 6 months ago

SCD type 2 which has latest boolean flag to retreive

upvoted 1 times

vigilante89 3 years, 8 months ago

Type 1 SCD - This concept overwrites the existing value within the dimension table with the new value without retaining the old data. Type 2 SCD - This concept maintains the versioning of the table. It creates and adds a new row with the new value and maintains the existing row containing the old value which usually is required for historic and reporting purposes. Type 3 SCD - This concept creates a new column with the new value in the existing record but also retains the original column containing the old value which usually is required for historic and reporting purposes.

upvoted 5 times

Fernando_Caemerer 3 years, 8 months ago

Answer D

upvoted 1 times

OldSchool 3 years, 9 months ago

SCD 2 - D

upvoted 1 times

Deeksha1234 4 years ago

correct D

upvoted 2 times

Question #13 *Topic 1*

You have an enterprise-wide Azure Data Lake Storage Gen2 account. The data lake is accessible only through an Azure virtual network named VNET1.
You are building a SQL pool in Azure Synapse that will use data from the data lake.
Your company has a sales team. All the members of the sales team are in an Azure Active Directory group named Sales. POSIX controls are used to assign the
Sales group access to the files in the data lake.
You plan to load data to the SQL pool every hour.
You need to ensure that the SQL pool can load the sales data from the data lake.
Which three actions should you perform? Each correct answer presents part of the solution.
NOTE: Each area selection is worth one point.

A. Add the managed identity to the Sales group.

B. Use the managed identity as the credentials for the data load process.

C. Create a shared access signature (SAS).

D. Add your Azure Active Directory (Azure AD) account to the Sales group.

E. Use the shared access signature (SAS) as the credentials for the data load process.

F. Create a managed identity.

Suggested Answer: *A, B, F*
The managed identity grants permissions to the dedicated SQL pools in the workspace. Note: Managed identity for Azure resources is a feature of Azure Active Directory. The feature provides Azure services with an automatically managed identity in

Azure AD - Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-identity

Community vote distribution
ABF (100%)

IDKol Highly Voted 1 year, 11 months ago
Correct Answer should be F. Create a managed identity. A. Add the managed identity to the Sales group. B. Use the managed identity as the credentials for the data load process.
upvoted 56 times

Backy Highly Voted 4 years, 7 months ago
Is answer A properly worded? "Add the managed identity to the Sales group" should be "Add the Sales group to managed identity"
upvoted 10 times

IMadnan Most Recent 1 year, 6 months ago
F. Create a managed identity. A. Add the managed identity to the Sales group. B. Use the managed identity as the credentials for the data load process.
upvoted 1 times

EmnCours 1 year, 8 months ago
Correct Answer: ABF
upvoted 1 times

acora192 1 year, 9 months ago
ABF, shared access is for external
upvoted 1 times

EvanG 1 year, 10 months ago
Managed Identity for synapse pool is already created and available by default, so I am not sure about F.
upvoted 2 times

KarlGardnerDataEngineering 2 years, 3 months ago
So Since the synapse dedicated sql pool is only accessible through an Azure virtual network wouldn't it make more since for the following steps: 1. Create a managed identity for synapse analytics workspace 2. Give role based access to managed identity 3. Use the managed identity as the

credentials for the data load process. This would not be giving the managed identity to the sales group and follows the least permissions principle

upvoted 2 times

ELJORDAN23 2 years, 7 months ago

Correct!

upvoted 1 times

kkk5566 2 years, 11 months ago

FAB is correct

upvoted 1 times

vigilante89 3 years, 8 months ago

The sequence is: FAB create managed identity --> add managed identity to the sales group --> use managed identity as credentials for data load process.

upvoted 7 times

XiltroX 3 years, 9 months ago

First create a managed ID, Add the managed ID, use the managed ID

upvoted 1 times

coolin 4 years ago

ABF is the correct ans

upvoted 1 times

Deeksha1234 4 years ago

ABF is correct

upvoted 3 times

Dothy 4 years, 3 months ago

correct answer is ABF

upvoted 1 times

Egocentric 4 years, 4 months ago

ABF is correct

upvoted 1 times

praticewizards 4 years, 4 months ago

FAB - create, add to group, use to load data

upvoted 2 times

lukeonline 4 years, 7 months ago

FAB should be correct

upvoted 5 times

HOTSPOT -

You have an Azure Synapse Analytics dedicated SQL pool that contains the users shown in the following table.

User1 executes a query on the database, and the query returns the results shown in the following exhibit.

User1 is the only user who has access to the unmasked data.

Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.

NOTE: Each correct selection is worth one point.

Hot Area:

Name	Role
User1	Server admin
User2	db_datereader

```
SELECT c.name,
    tbl.name as table_name,
    typ.name as datatype,
    c.is_masked,
    c.masking_function
FROM sys.masked_columns AS c
INNER JOIN sys.tables AS tbl ON c.[object_id] = tbl.[object_id]
INNER JOIN sys.types typ ON c.user_type_id = typ.user_type_id
WHERE is_masked = 1;
```

Results Messages

	name	table_name	datatype	is_masked	masking_function
1	BirthDate	DimCustomer	date	1	default()
2	Gender	DimCustomer	nvarchar	1	default()
3	EmailAddress	DimCustomer	nvarchar	1	email()
4	YearlyIncome	DimCustomer	money	1	default()

Answer Area

When User2 queries the YearlyIncome column, the values returned will be **[answer choice]**.

- a random number
- the values stored in the database
- XXXX
- 0

When User1 queries the BirthDate column, the values returned will be **[answer choice]**.

- a random date
- the values stored in the database
- XXXX
- 1900-01-01

Suggested Answer:

Answer Area

When User2 queries the YearlyIncome column, the values returned will be **[answer choice]**.

- a random number
- the values stored in the database
- XXXX
- **0**

When User1 queries the BirthDate column, the values returned will be **[answer choice]**.

- a random date
- **the values stored in the database**
- XXXX
- 1900-01-01

Answer Area

When User2 queries the YearlyIncome column, the values returned will be **[answer choice]**.

- a random number
- the values stored in the database
- XXXX
- **0**

When User1 queries the BirthDate column, the values returned will be **[answer choice]**.

- a random date
- **the values stored in the database**
- XXXX
- 1900-01-01

Box 1: 0 - The YearlyIncome column is of the money data type. The Default masking function: Full masking according to the data types of the designated fields ➪ Use a zero value for numeric data types (bigint, bit, decimal, int, money, numeric, smallint, smallmoney, tinyint, float, real).
Box 2: the values stored in the database Users with administrator privileges are always excluded from masking, and see the original data without any mask. Reference: https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

hsetin Highly Voted 4 years, 5 months ago
user 1 is admin, so he will see the value stored in dbms. 1. 0 2. Value in database2 is wrongConfirmed. Everything explained here https://learn.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview?view=azuresql-mi
upvoted 113 times

hsetin Highly Voted 4 years, 5 months ago
user 1 is admin, so he will see the value stored in dbms. 1. 0 2. Value in database
upvoted 113 times

azurearmy 4 years, 3 months ago
2 is wrong

upvoted 3 times

examtopicsofyannick 2 years, 6 months ago

Confirmed. Everything explained here https://learn.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview?view=azuresql-mi

upvoted 4 times

rjile 4 years, 5 months ago

• Use a zero value for numeric data types (bigint, bit, decimal, int, money, numeric, smallint, smallmoney, tinyint, float, real). • Use 01-01-1900 for date/time data types (date, datetime2, datetime, datetimeoffset, smalldatetime, time).

upvoted 37 times

gplusplus Most Recent 1 year, 10 months ago

I think the answer is correct as admins dont necessarily have access to ddmed data. 0 / values in db

upvoted 1 times

Joanna0 2 years, 1 month ago

* Use a zero value for numeric data types (bigint, bit, decimal, int, money, numeric, smallint, smallmoney, tinyint, float, real).

upvoted 1 times

hassexat 2 years, 5 months ago

User 1 --> Data stored in the database because this user is server_admin and can access to unmasked data User 2 --> 0 because this user is db_reader and can't access to unmasked data and masked_function is set in default() You have to pay attention to masked_function when a user can't access to unmasked data to know what the user will get in the queries

upvoted 2 times

kkk5566 2 years, 5 months ago

1. 0 2. Value stored in database

upvoted 2 times

nicky87654 3 years, 1 month ago

Based on the information provided in the scenario: When User2 queries the YearlyIncome column, the values returned will be [XXXX]. This is because User2 has the role of db datareader, which means that they do not have access to the unmasked data, and the data will be masked (replaced with 'XXXX') when they query it. When User1 queries the BirthDate column, the values returned will be [the values stored in the database]. This is because User1 has the role of Server admin, which means that they have access to the unmasked data, and the data will be shown as it is stored in the database when they query it.Default masking rule is: For numeric data types use a zero value (bigint, bit, decimal, int, money, numeric, smallint, smallmoney, tinyint, float, real). https://learn.microsoft.com/en-us/sql/relational-databases/security/dynamic-data-masking?view=sql-server-ver16

upvoted 5 times

nicky87654 3 years, 1 month ago

Based on the information provided in the scenario:

When User2 queries the YearlyIncome column, the values returned will be [XXXX]. This is because User2 has the role of db datareader, which means that they do not have access to the unmasked data, and the data will be masked (replaced with 'XXXX') when they query it.

When User1 queries the BirthDate column, the values returned will be [the values stored in the database]. This is because User1 has the role of Server admin, which means that they have access to the unmasked data, and the data will be shown as it is stored in the database when they query it.

upvoted 5 times

fiuel 3 years, 1 month ago

Default masking rule is: For numeric data types use a zero value (bigint, bit, decimal, int, money, numeric, smallint, smallmoney, tinyint, float, real). https://learn.microsoft.com/en-us/sql/relational-databases/security/dynamic-data-masking?view=sql-server-ver16

upvoted 11 times

vigilante89 3 years, 2 months ago

First case: output will be 0 User2 queries YearlyIncome column which (is_masked = 1) i.e. its confidential thus has very limited access. Since user2 is a simple db reader, the value wouldn't be viewed by the user. Second case: value stored in the database User1 is the admin or superuser with full access to the entire data within the database sys. So he will be able to view the brithdate column (is_masked=1) in the database 'sys'.

upvoted 4 times

OldSchool 3 years, 3 months ago

User2 is a reader so he will see 0 querying YearlyIncome with default() mask; User1 is admin and only he will see all stored values

upvoted 1 times

Deeksha1234 3 years, 6 months ago

correct

upvoted 5 times

azure900test 3 years, 7 months ago

User 1: The value User 2: XXXX see https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview?view=azuresql AND https://www.sqlshack.com/dynamic-data-masking-in-sql-server

upvoted 2 times

objecto 3 years, 8 months ago

According to https://docs.microsoft.com/en-us/sql/relational-databases/security/dynamic-data-masking?view=sql-server-ver16 "For date and time data types use 01.01.1900 00:00:00.0000000 (date, datetime2, datetime, datetimeoffset, smalldatetime, time)." Data masking as defalut to a date dateime smalldate should be 1900-01-01. Strangely there is no such option. Any ideas anyone?Damn, User 1 only reads YearlyIncome (not date), so yes 0 is the correct answer

upvoted 2 times

objecto 3 years, 8 months ago

According to https://docs.microsoft.com/en-us/sql/relational-databases/security/dynamic-data-masking?view=sql-server-ver16

"For date and time data types use 01.01.1900 00:00:00.0000000 (date, datetime2, datetime, datetimeoffset, smalldatetime, time)."

Data masking as defalut to a date dateime smalldate should be 1900-01-01. Strangely there is no such option. Any ideas anyone?

upvoted 2 times

objecto 3 years, 8 months ago

Damn, User 1 only reads YearlyIncome (not date), so yes 0 is the correct answer

upvoted 2 times

Dothy 3 years, 9 months ago

1. 0 2. Value in database

upvoted 1 times

manan16 3 years, 10 months ago

How user2 can access data as it is masked?

upvoted 1 times

manan16 3 years, 10 months ago

Can Someone explain first option as in doc it says 0

upvoted 1 times

Mahesh_mm 4 years, 1 month ago

1. 0 (Default values for money data type for masked function will written when queried by user2) 2. Value in database (As it is queried by user1 who is admin)

upvoted 9 times

Milan1988 4 years, 3 months ago

CORRECT

upvoted 3 times

Question #15 *Topic 1*

You have an enterprise data warehouse in Azure Synapse Analytics.
Using PolyBase, you create an external table named [Ext].[Items] to query Parquet files stored in Azure Data Lake Storage Gen2 without importing the data to the data warehouse.
The external table has three columns.
You discover that the Parquet files have a fourth column named ItemID.
Which command should you run to add the ItemID column to the external table?
A.
B.
C.
D.

```
ALTER EXTERNAL TABLE [Ext].[Items]
  ADD [ItemID] int;

DROP EXTERNAL FILE FORMAT parquetfile1;
CREATE EXTERNAL FILE FORMAT parquetfile1
WITH (
     FORMAT_TYPE = PARQUET,
     DATA_COMPRESSION = 'org.apache.hadoop.io.compress.SnappyCodec'
);

DROP EXTERNAL TABLE [Ext].[Items]
CREATE EXTERNAL TABLE [Ext].[Items]
([ItemID] [int] NULL,
 [ItemName] nvarchar(50) NULL,
 [ItemType] nvarchar(20) NULL,
 [ItemDescription] nvarchar(250))
WITH
(
     LOCATION= '/Items/',
          DATA_SOURCE = AzureDataLakeStore,
          FILE_FORMAT = PARQUET,
          REJECT_TYPE = VALUE,
          REJECT_VALUE = 0
);

ALTER TABLE [Ext].[Items]
ADD [ItemID] int;
```

Suggested Answer:
Incorrect Answers: A, D: Only these Data Definition Language (DDL) statements are allowed on external tables: ✑ CREATE TABLE and DROP TABLE ✑ CREATE STATISTICS and DROP STATISTICS ✑ CREATE VIEW and DROP VIEW Reference: https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql

Chien_Nguyen_Van Highly Voted 4 years, 5 months ago
C is correct https://www.examtopics.com/discussions/microsoft/view/19469-exam-dp-200-topic-1-question-27-discussion/
upvoted 58 times

baiy Highly Voted 2 years, 12 months ago
C is correct, since "altering the schema or format of an external SQL table is not supported". https://learn.microsoft.com/en-us/azure/data-explorer/kusto/management/external-sql-tables
upvoted 16 times

ELJORDAN23 Most Recent 2 years, 1 month ago
Got this question on my exam on january 17, C is correct
upvoted 6 times

hassexat 2 years, 5 months ago
C is correct since you need to drop first and create again the external table

upvoted 2 times

vigilante89 3 years, 2 months ago

Answer is C. Drop the external table and recreate it. Because the column which needs to be added is the ItemID which seems like a primary key. So we have to drop the table and recreate it. Had it been any other column, we could have used ALTER syntax to add a column like shown below: ALTER EXTERNAL TABLE name action [, ...] where action is one of: ADD [COLUMN] column_name type DROP [COLUMN] column ALTER [COLUMN] column TYPE type [USING expression] OWNER TO new_owner

upvoted 5 times

OldSchool 3 years, 3 months ago

C is Correct

upvoted 1 times

Selma97 3 years, 3 months ago

I still can't understand why it's not D.ALTER statement is not supported on external table, you need to DROP it and CREATE it again

upvoted 1 times

Selma97 3 years, 3 months ago

I still can't understand why it's not D.

upvoted 1 times

anto69 3 years, 2 months ago

ALTER statement is not supported on external table, you need to DROP it and CREATE it again

upvoted 10 times

Deeksha1234 3 years, 6 months ago

correct

upvoted 2 times

dsp17 3 years, 7 months ago

C is correct. Even if you are confuse with other options. The clue here is keyword Location while creating external table, LOCATION = 'folder_or_filepath' : Specifies the folder or the file path and file name for the actual data.

upvoted 2 times

Ozren 3 years, 11 months ago

Good thing the details are shown here: "The external table has three columns." And the solution yet reveals the column details. This doesn't make any sense to me. If C is the correct answer (only one that seems acceptable), then the question itself is flawed.The external table has 3 columns, but the files it references has 4 columns, so the external table has to be altered

upvoted 2 times

Ozren 3 years, 11 months ago

Good thing the details are shown here: "The external table has three columns." And the solution yet reveals the column details. This doesn't make any sense to me. If C is the correct answer (only one that seems acceptable), then the question itself is flawed.

upvoted 2 times

dduque10 3 years, 2 months ago

The external table has 3 columns, but the files it references has 4 columns, so the external table has to be altered

upvoted 2 times

PallaviPatel 4 years ago

c is correct.

upvoted 1 times

hugoborda 4 years, 4 months ago

Answer is correct

upvoted 2 times

HOTSPOT -

You have two Azure Storage accounts named Storage1 and Storage2. Each account holds one container and has the hierarchical namespace enabled. The system has files that contain data stored in the Apache Parquet format.

You need to copy folders and files from Storage1 to Storage2 by using a Data Factory copy activity. The solution must meet the following requirements:

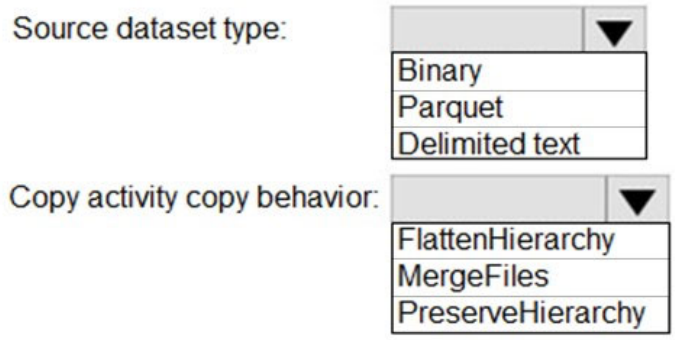
- ✑ No transformations must be performed.
- ✑ The original folder structure must be retained.
- ✑ Minimize time required to perform the copy activity.

How should you configure the copy activity? To answer, select the appropriate options in the answer area.

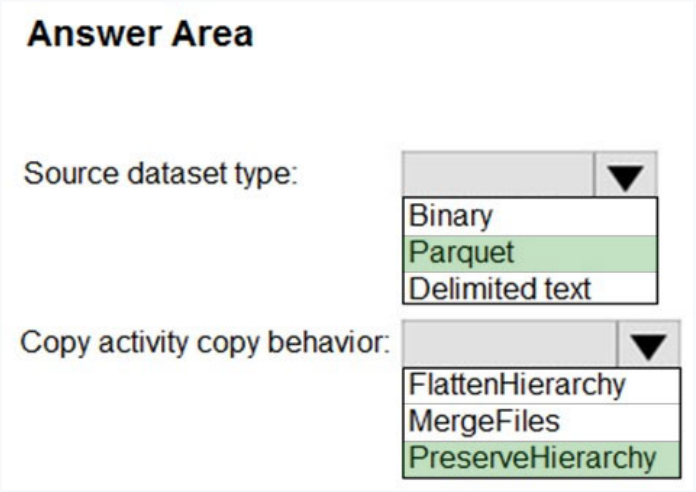
NOTE: Each correct selection is worth one point.

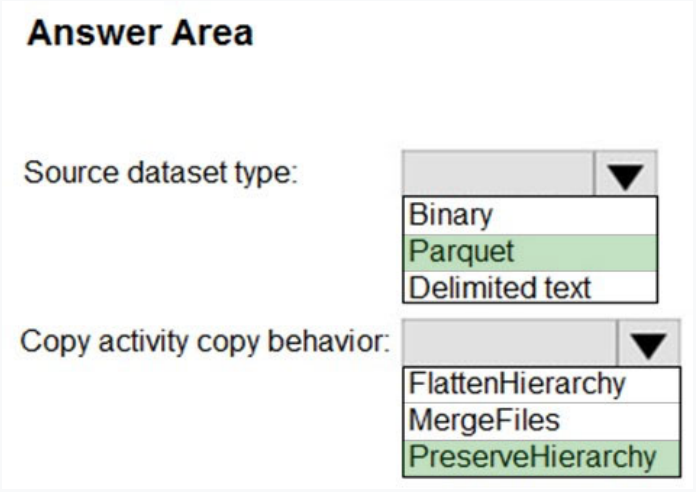
Hot Area:

Answer Area



Suggested Answer:





Box 1: Parquet - For Parquet datasets, the type property of the copy activity source must be set to ParquetSource.

Box 2: PreserveHierarchy - PreserveHierarchy (default): Preserves the file hierarchy in the target folder. The relative path of the source file to the source folder is identical to the relative path of the target file to the target folder. Incorrect Answers: ➪ FlattenHierarchy: All files from the source folder are in the first level of the target folder. The target files have autogenerated names. ➪ MergeFiles: Merges all files from the source folder to one file. If the file name is specified, the merged file name is the specified name. Otherwise, it's an autogenerated file name. Reference: https://docs.microsoft.com/en-us/azure/data-factory/format-parquet https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage

EddyRoboto Highly Voted 1 year, 11 months ago

This could be binary as source and sink, since there are no transformations on files. I tend to believe that would be binary the correct anwer.

upvoted 85 times

EddyRoboto Highly Voted 1 year, 11 months ago

This could be binary as source and sink, since there are no transformations on files. I tend to believe that would be binary the correct anwer.Agree. I've checked it. With binary source and sink datasets it works.no it must be parquet because The type property of the dataset must be set to Binary. and it's parquet hear so answer are correctI agree. If it's just copying then binary is fine and would probably be fasteryes, and a d Binary is fast as it is direct copy, Parquet reads and writs at destination which can be slower.

upvoted 85 times

iooj 4 years, 6 months ago

Agree. I've checked it. With binary source and sink datasets it works.

upvoted 5 times

jed_elhak 4 years, 10 months ago

no it must be parquet because The type property of the dataset must be set to Binary. and it's parquet hear so answer are correct

upvoted 2 times

michalS 4 years, 11 months ago

I agree. If it's just copying then binary is fine and would probably be faster
upvoted 6 times

ayn1 Most Recent 1 year, 4 months ago
yes, and a d Binary is fast as it is direct copy, Parquet reads and writs at destination which can be slower.
upvoted 1 times

AbhiGola 4 years, 11 months ago
Answer seems correct as data is store is parquet already and requirement is to do no transformation so answer is rightAs question has mentioned, Minimize time required to perform the copy activity. And binary is faster than Parquet. Hence, Binary is answerNo: req1 "no transformation", req2 "Minimize time required to perform the copy activity". Both must be met hence it's Parquet cause it's the second fastest choice and it requires no transformations.when doing a binary copy, you're not doing any transformation!
upvoted 64 times

AbhiGola 4 years, 11 months ago
Answer seems correct as data is store is parquet already and requirement is to do no transformation so answer is right
upvoted 64 times

NintyFour 4 years, 3 months ago
As question has mentioned, Minimize time required to perform the copy activity. And binary is faster than Parquet. Hence, Binary is answer
upvoted 6 times

NintyFour 4 years, 3 months ago
As question has mentioned, Minimize time required to perform the copy activity. And binary is faster than Parquet. Hence, Binary is answerNo: req1 "no transformation", req2 "Minimize time required to perform the copy activity". Both must be met hence it's Parquet cause it's the second fastest choice and it requires no transformations.when doing a binary copy, you're not doing any transformation!
upvoted 6 times

anto69 3 years, 8 months ago
No: req1 "no transformation", req2 "Minimize time required to perform the copy activity". Both must be met hence it's Parquet cause it's the second fastest choice and it requires no transformations.
upvoted 6 times

anto69 3 years, 8 months ago
No: req1 "no transformation", req2 "Minimize time required to perform the copy activity". Both must be met hence it's Parquet cause it's the second fastest choice and it requires no transformations.when doing a binary copy, you're not doing any transformation!
upvoted 6 times

mhi 3 years, 3 months ago
when doing a binary copy, you're not doing any transformation!
upvoted 4 times

JustAnotherDBA 1 year, 10 months ago
The answer is correct. 3 reasons.

The file format is Parquet. Parquet has the 2nd fastest load time. No data transformations should happen,

If we are going to quote articles, please read the WHOLE article before posting. Check out the formats that the binary can handle.

"When using Binary dataset in copy activity, you can only copy from Binary dataset to Binary dataset."
upvoted 12 times

JustAnotherDBA 1 year, 10 months ago
The answer is correct. 3 reasons. The file format is Parquet. Parquet has the 2nd fastest load time. No data transformations should happen, If we are going to quote articles, please read the WHOLE article before posting. Check out the formats that the binary can handle. "When using Binary dataset in copy activity, you can only copy from Binary dataset to Binary dataset."https://learn.microsoft.com/en-us/azure/data-factory/format-binaryBinary to binary copies the files as they are, retaining the same content, hence retaining the format and it;s faster than parquet, because it doesn;t require load at all just copy.
upvoted 12 times

JustAnotherDBA 3 years, 7 months ago
https://learn.microsoft.com/en-us/azure/data-factory/format-binary
upvoted 3 times

mtc9 3 years, 2 months ago
Binary to binary copies the files as they are, retaining the same content, hence retaining the format and it;s faster than parquet, because it doesn;t require load at all just copy.

upvoted 2 times

Lestrang 1 year, 10 months ago

According to ChatGPT While "binary" dataset type would be the fastest in terms of copying the data from one Azure storage account to another, it would not be the correct option in this scenario because it does not retain the original format of the files. If the files contain data stored in the Apache Parquet format, specifying the source dataset type as "binary" would cause Data Factory to treat the files as generic binary files, and it would copy the data as is, without recognizing the original format of the files. This would result in losing the original format of the files, and possibly losing the structure of the data, it could also make it more difficult to read the data. Also, When you copy files using binary dataset type, Data Factory will not be able to detect the changes in files and it copies the entire data each time, this can be inefficient in terms of time and storage. it really gives shitty azure answers in general, but ill go for parquet for this one.ChatGPT is plainly wrong, binary type retais the original parquet format, ebcause it means to copy the files as they are and it;s faster than parquet dataset, because it's doesn't require parsing the files. Binary is correct.

upvoted 12 times

Lestrang 1 year, 10 months ago

According to ChatGPT

While "binary" dataset type would be the fastest in terms of copying the data from one Azure storage account to another, it would not be the correct option in this scenario because it does not retain the original format of the files.

If the files contain data stored in the Apache Parquet format, specifying the source dataset type as "binary" would cause Data Factory to treat the files as generic binary files, and it would copy the data as is, without recognizing the original format of the files. This would result in losing the original format of the files, and possibly losing the structure of the data, it could also make it more difficult to read the data.

Also, When you copy files using binary dataset type, Data Factory will not be able to detect the changes in files and it copies the entire data each time, this can be inefficient in terms of time and storage.

it really gives shitty azure answers in general, but ill go for parquet for this one.

upvoted 12 times

mtc9 3 years, 2 months ago

ChatGPT is plainly wrong, binary type retais the original parquet format, ebcause it means to copy the files as they are and it;s faster than parquet dataset, because it's doesn't require parsing the files. Binary is correct.

upvoted 1 times

klayytech 1 year, 10 months ago

The answer is still Source dataset type: Parquet Copy activity copy behavior: Preserve Hierarchy. Even though Binary can be used as the source dataset type, it is not the best option in this scenario. The original folder structure is important, and using Parquet as the source dataset type will ensure that it is preserved. Source dataset type: Parquet Copy activity copy behavior: Preserve Hierarchy This will ensure that the files are copied in their original format, and that the original folder structure is preserved in the destination container. This is the best option for this scenario, as it meets all of the requirements.

upvoted 1 times

auwia 1 year, 10 months ago

Massimo Manganiello 13:36 (49 minuti fa) a me

When it comes to efficiency, copying data from a Parquet file to another Parquet file is generally more efficient than copying to a binary format. This is because Parquet is a columnar storage format specifically designed for efficient data compression and query performance. It leverages advanced compression techniques and data encoding to minimize storage size and optimize query execution.

Copying data from a Parquet file to a binary format may require additional steps and conversions. Binary formats, such as plain text or custom binary formats, may not have the same level of built-in compression and optimization as Parquet. Therefore, the copy process may involve additional serialization and deserialization steps, resulting in increased processing overhead and potentially larger storage requirements.

In summary, when the source and destination formats are both Parquet, copying between Parquet files is generally more efficient in terms of storage utilization and query performance.

In my opinion, the provided answer are corrects!

upvoted 3 times

auwia 1 year, 10 months ago

Massimo Manganiello 13:36 (49 minuti fa) a me When it comes to efficiency, copying data from a Parquet file to another Parquet file is generally more efficient than copying to a binary format. This is because Parquet is a columnar storage format specifically designed for efficient data compression and query performance. It leverages advanced compression techniques and data encoding to minimize storage size and optimize query execution. Copying data from a Parquet file to a binary format may require additional steps and conversions. Binary formats, such as plain text or custom binary formats, may not have the same level of built-in compression and optimization as Parquet. Therefore, the copy process may involve additional serialization and deserialization steps, resulting in increased processing overhead and potentially larger storage requirements. In summary, when the source and destination formats are both Parquet, copying between Parquet files is generally more efficient in terms of storage utilization and query performance. In my opinion, the provided answer are corrects!

upvoted 3 times

Fusejonny1 2 years, 6 months ago

Source dataset type should be set to binary. The reason for this is that you're not performing any transformations on the data, you're simply copying it from one location to another while retaining the original folder structure. The binary dataset in Azure Data Factory is used for copying files as-is without parsing the file data.

upvoted 2 times

kkk5566 2 years, 11 months ago

Binary & PerserveHierarchy

upvoted 3 times

tonyfig 2 years, 12 months ago

Binary & PerserveHierarchy The Parquet option is used when you want to copy data stored in the Apache Parquet format and perform transformations on the data during the copy activity. However, in this scenario, the requirement is to perform no transformations and minimize the time required to perform the copy activity. The Binary option is better suited for this scenario as it copies the data as-is, without performing any transformations, and minimizes the time required to perform the copy activity.

upvoted 4 times

rocky48 3 years, 1 month ago

Answer seems correct as data is store is parquet already and requirement is to do no transformation so answer is right. Source dataset type: Parquet Copy activity copy behavior: Preserve Hierarchy

upvoted 4 times

trantrongw 3 years, 4 months ago

Agree. I've checked it.

upvoted 1 times

Rrk07 3 years, 8 months ago

Answer is correct .

upvoted 1 times

temacc 3 years, 9 months ago

Binary - copy files as is in fastest way. PreserveHierarchy - for saving folder structure.

upvoted 2 times

OldSchool 3 years, 9 months ago

Answer is correct. No transformation and preserve hierarchy

upvoted 1 times

RBKasemodel 3 years, 9 months ago

I believe the answer should be Binary, since it is stated that no transformations must be done. "You can use Binary dataset in Copy activity, GetMetadata activity, or Delete activity. When using Binary dataset, the service does not parse file content but treat it as-is." https://learn.microsoft.com/en-us/azure/data-factory/format-binary I couldn't found any information saying that parquet won't be parsed if the source and sink are parquets files. So I -think- it will parse, and we can understand that it is a transformation.When using Binary dataset in copy activity, you can only copy from Binary dataset to Binary dataset. Why do you pretend to preserver the hierarchy if the same hierarchy "When using Binary dataset in copy activity, you can only copy from Binary dataset to Binary dataset. "

upvoted 2 times

RBKasemodel 3 years, 9 months ago

I believe the answer should be Binary, since it is stated that no transformations must be done.

"You can use Binary dataset in Copy activity, GetMetadata activity, or Delete activity. When using Binary dataset, the service does not parse file content but treat it as-is." https://learn.microsoft.com/en-us/azure/data-factory/format-binary

I couldn't found any information saying that parquet won't be parsed if the source and sink are parquets files. So I -think- it will parse, and we can understand that it is a transformation.

upvoted 2 times

alphilla 2 years, 6 months ago

When using Binary dataset in copy activity, you can only copy from Binary dataset to Binary dataset. Why do you pretend to preserver the hierarchy if the same hierarchy "When using Binary dataset in copy activity, you can only copy from Binary dataset to Binary dataset. "

upvoted 1 times

allagowf 3 years, 10 months ago

Answer seems correct, advice don't overthink, the source is parquet and it's one of the options so it is parquet.

upvoted 7 times

Deeksha1234 4 years ago

given ans is correct

upvoted 2 times

Question #17 *Topic 1*

You have an Azure Data Lake Storage Gen2 container that contains 100 TB of data.
You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region.
The solution must minimize costs.
Which type of data redundancy should you use?

A. geo-redundant storage (GRS)

B. read-access geo-redundant storage (RA-GRS)

C. zone-redundant storage (ZRS)

D. locally-redundant storage (LRS)

Suggested Answer: *B*
Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages. However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region. When you enable read access to the secondary region, your data is available to be read at all times, including in a situation where the primary region becomes unavailable. Incorrect Answers: A: While Geo-redundant storage (GRS) is cheaper than Read-Access Geo-Redundant Storage (RA-GRS), GRS does NOT initiate automatic failover. C, D: Locally redundant storage (LRS) and Zone-redundant storage (ZRS) provides redundancy within a single region. Reference: https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy

Community vote distribution
A (61%) B (39%)

meetj Highly Voted 1 year, 11 months ago
B is right Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages. However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region. When you enable read access to the secondary region, your data is available to be read at all times, including in a situation where the primary region becomes unavailable.
upvoted 127 times

meetj Highly Voted 1 year, 11 months ago
B is right Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages. However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region. When you enable read access to the secondary region, your data is available to be read at all times, including in a situation where the primary region becomes unavailable.A looks correct answer. RA-GRS is always avialable because its auto failover. Since this is not asked in the question but more importantly the question is about reducing cost which GRS.It should be A because of two reasons: 1. Minimize cost 2. When primary is unavailable. Hence No need for RA_GRSExactly. This is the point. It clearly states ' in case of an outage' RA-GRS --> secondary region can be read also not in a case of outageits not A, dude, if you dont understand the difference between GRS and RA-GRS then u need az 101. With GRS, the 2nd region is NEVER available for access until Microsoft fails over the first failed region. Otherwise, you can NEVER access the 2nd regions data. Hence RA-GRS.No need to be rude. The question specifies that the data in the second region needs to be available IF an outage occurs. So GRS is more than enough. It's not because you think otherwise that you're right.The question clearly says "is available for read workloads in a secondary region". This is only available when choosing RA-GRS.* With GRS, when a disaster happens in the primary region, the user has to initiate a failover so that the secondary region becomes the primary region**. At no point you are reading from your secondary region with GRS. Hence i believe the answers should be B. *https://learn.microsoft.com/en-us/azure/storage/common/storage-redundancy#geo-redundant-storage **https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidanceYou misunderstanding the question : GRS also
upvoted 127 times

dev2dev 4 years, 7 months ago
A looks correct answer. RA-GRS is always avialable because its auto failover. Since this is not asked in the question but more importantly the question is about reducing cost which GRS.It should be A because of two reasons: 1. Minimize cost 2. When primary is unavailable. Hence No need for RA_GRSExactly. This is the point. It clearly states ' in case of an outage' RA-GRS --> secondary region can be read also not in a case of outageits not A, dude, if you dont understand the difference between GRS and RA-GRS then u need az 101. With GRS, the 2nd region is NEVER available for access until Microsoft fails over the first failed region. Otherwise, you can NEVER access the 2nd regions data. Hence RA-GRS.No need to be rude. The question specifies that the data in the second region needs to be available IF an outage occurs. So GRS is more than enough. It's not because you think otherwise that you're right.The question clearly says "is available for read workloads in a secondary region". This is only available when

choosing RA-GRS.* With GRS, when a disaster happens in the primary region, the user has to initiate a failover so that the secondary region becomes the primary region**. At no point you are reading from your secondary region with GRS. Hence i believe the answers should be B. *https://learn.microsoft.com/en-us/azure/storage/common/storage-redundancy#geo-redundant-storage **https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidanceYou misunderstanding the question : GRS also give the possibilities to read. it's not specified that we need to read from the second region when the first is available + You have to reduce the cost : GRS is cheaper than RA-GRS because GRS will be available only if the first region failover (in the subject we can read IF AN OUTAGE OCCURES) : https://azure.microsoft.com/en-us/pricing/details/storage/blobs/

upvoted 36 times

dev2dev 4 years, 7 months ago

A looks correct answer. RA-GRS is always avialable because its auto failover. Since this is not asked in the question but more importantly the question is about reducing cost which GRS.

upvoted 36 times

BK10 4 years, 6 months ago

It should be A because of two reasons: 1. Minimize cost 2. When primary is unavailable. Hence No need for RA_GRSExactly. This is the point. It clearly states ' in case of an outage' RA-GRS --> secondary region can be read also not in a case of outageits not A, dude, if you dont understand the difference between GRS and RA-GRS then u need az 101. With GRS, the 2nd region is NEVER available for access until Microsoft fails over the first failed region. Otherwise, you can NEVER access the 2nd regions data. Hence RA-GRS.No need to be rude. The question specifies that the data in the second region needs to be available IF an outage occurs. So GRS is more than enough. It's not because you think otherwise that you're right.

upvoted 32 times

BK10 4 years, 6 months ago

It should be A because of two reasons: 1. Minimize cost 2. When primary is unavailable. Hence No need for RA_GRS

upvoted 32 times

Billybob0604 3 years, 8 months ago

Exactly. This is the point. It clearly states ' in case of an outage' RA-GRS --> secondary region can be read also not in a case of outage

upvoted 7 times

AnonymousJhb 3 years, 3 months ago

its not A, dude, if you dont understand the difference between GRS and RA-GRS then u need az 101. With GRS, the 2nd region is NEVER available for access until Microsoft fails over the first failed region. Otherwise, you can NEVER access the 2nd regions data. Hence RA-GRS.No need to be rude. The question specifies that the data in the second region needs to be available IF an outage occurs. So GRS is more than enough. It's not because you think otherwise that you're right.

upvoted 5 times

AnonymousJhb 3 years, 3 months ago

its not A, dude, if you dont understand the difference between GRS and RA-GRS then u need az 101. With GRS, the 2nd region is NEVER available for access until Microsoft fails over the first failed region. Otherwise, you can NEVER access the 2nd regions data. Hence RA-GRS.

upvoted 5 times

semauni 3 years ago

No need to be rude. The question specifies that the data in the second region needs to be available IF an outage occurs. So GRS is more than enough. It's not because you think otherwise that you're right.

upvoted 10 times

kenmexam 3 years, 9 months ago

The question clearly says "is available for read workloads in a secondary region". This is only available when choosing RA-GRS.* With GRS, when a disaster happens in the primary region, the user has to initiate a failover so that the secondary region becomes the primary region**. At no point you are reading from your secondary region with GRS. Hence i believe the answers should be B. *https://learn.microsoft.com/en-us/azure/storage/common/storage-redundancy#geo-redundant-storage **https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance

upvoted 16 times

kenmexam 3 years, 9 months ago

The question clearly says "is available for read workloads in a secondary region". This is only available when choosing RA-GRS.* With GRS, when a disaster happens in the primary region, the user has to initiate a failover so that the secondary region becomes the primary region**. At no point you are reading from your secondary region with GRS. Hence i believe the answers should be B. *https://learn.microsoft.com/en-us/azure/storage/common/storage-redundancy#geo-redundant-storage **https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidanceYou misunderstanding the question : GRS also give the possibilities to read. it's not specified that we need to read from the second region when the first is available + You have to reduce the cost : GRS is cheaper than RA-GRS because GRS will be available only if the first region failover (in the subject we can read IF AN OUTAGE OCCURES) : https://azure.microsoft.com/en-us/pricing/details/storage/blobs/

upvoted 16 times

dylan_t 3 years, 2 months ago

You misunderstanding the question : GRS also give the possibilities to read. it's not specified that we need to read from the second region when the first is available + You have to reduce the cost : GRS is cheaper than RA-GRS because GRS will be available only if the first region failover (in the subject we can read IF AN OUTAGE OCCURES) : https://azure.microsoft.com/en-us/pricing/details/storage/blobs/

upvoted 3 times

Sasha_in_San_Francisco 1 year, 11 months ago

In my opinion, I believe the and answer is A, and this is why.

In the question they state "...available for read workloads in a secondary region IF AN OUTAGE OCCURES in the primary...". Well, answer B (RA-GRS) states in Microsoft documentation that RA-GRS is for when "...your data is available to be read AT ALL TIMES, including in a situation where the primary region becomes unavailable."

To me, the nature of the question is what is the cheapest solution which allows for failover to read workload, when there is an outage. Answer (A).

Common sense would be 'A' too because that is probably the most often real-life use case.

upvoted 79 times

Sasha_in_San_Francisco 1 year, 11 months ago

In my opinion, I believe the and answer is A, and this is why. In the question they state "...available for read workloads in a secondary region IF AN OUTAGE OCCURES in the primary...". Well, answer B (RA-GRS) states in Microsoft documentation that RA-GRS is for when "...your data is available to be read AT ALL TIMES, including in a situation where the primary region becomes unavailable." To me, the nature of the question is what is the cheapest solution which allows for failover to read workload, when there is an outage. Answer (A). Common sense would be 'A' too because that is probably the most often real-life use case.It's not about common sense rather about technology. With GRS, data remains available even if an entire data center becomes unavailable or if there is a widespread regional failure. There would be a down time when a region becomes unavailable. Alternately, you could implement read-access geo-redundant storage (RA-GRS), which provides read-access to the data in alternate locations.

upvoted 79 times

SabaJamal2010AtGmail 4 years, 7 months ago

It's not about common sense rather about technology. With GRS, data remains available even if an entire data center becomes unavailable or if there is a widespread regional failure. There would be a down time when a region becomes unavailable. Alternately, you could implement read-access geo-redundant storage (RA-GRS), which provides read-access to the data in alternate locations.

upvoted 4 times

technoguy Most Recent 1 year, 4 months ago

A is correct since cost is concern

upvoted 1 times

Jolyboy 1 year, 4 months ago

For the keyword "is available for read workloads in a secondary region" B is the right choice

upvoted 1 times

ayn1 1 year, 4 months ago

'Data must be available for read workloads in a secondary region', even is GRS is cheaper, no data read in secondary region unless manually initiated i.e. not useful for Read workload from secondary region.

upvoted 1 times

IMadnan 1 year, 6 months ago

read-access geo-zone-redundant storage (RA-GZRS)

upvoted 1 times

jxs221 1 year, 6 months ago

Read-Access Geo-Redundant Storage (RA-GRS): Data is replicated to a secondary region, and read access is enabled in the secondary region. This means that if the primary region fails, the data can still be read from the secondary region

upvoted 1 times

JustImperius 1 year, 7 months ago

I was Team B up until the last sentence "the solution must minimize costs". To minimize costs and to ensure the data is available for read IF an outage occurs points to A. They never said anything about it being available at the same time. B goes against the minimizing of cost so A. Not the best questions in my opinion...leaves too much room for interpretation.

upvoted 1 times

Manash_chottu 1 year, 7 months ago

Understating the question very clearly that asking for an outage not failover, so I would say my answer to "B". Here is the reference to follow: Link: https://learn.microsoft.com/en-us/azure/storage/common/storage-redundancy#read-

upvoted 1 times

Manash_chottu 1 year, 7 months ago

Question is clearly mentioned that "read workloads in a secondary region if an outage occurs in the primary region + optimize cost" - where GRS solution is the best option compared to others.

upvoted 1 times

s1852 1 year, 8 months ago

B is right Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages. However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region. When you enable read access to the secondary region, your data is available to be read at all times, including in a situation where the primary region becomes unavailable.

upvoted 1 times

DP2032024 1 year, 8 months ago

Question has asked about "read workloads in a secondary region" - Hence answer would be B

upvoted 1 times

moize 1 year, 8 months ago

Pour assurer la disponibilité des données dans une région secondaire en cas de panne dans la région principale tout en minimisant les coûts, vous devriez utiliser le stockage géo-redondant à accès en lecture (RA-GRS) (Option B). Voici pourquoi : RA-GRS réplique vos données de manière asynchrone vers une région secondaire géographiquement éloignée, offrant ainsi une protection contre les pannes régionales1. En cas de panne dans la région principale, RA-GRS permet un accès en lecture aux données répliquées dans la région secondaire, ce qui assure la continuité des charges de travail de lecture1. Cette option équilibre bien la disponibilité et les coûts, car elle offre une haute disponibilité pour les lectures sans les coûts plus élevés associés à d'autres options de redondance.

upvoted 1 times

shinypriti23 1 year, 8 months ago

RA-GRS offers the benefits of geo-redundant storage with additional read access to the secondary region. This means your data will be available for reading in a secondary region if the primary region goes down, ensuring high availability at a relatively low cost.

upvoted 1 times

EmnCours 1 year, 8 months ago

Correct Answer: A

upvoted 1 times

Lestrang 1 year, 10 months ago

While it is true that the customer/Microsoft has to initiate the failover, this is not elaborated in any sense in the question. What is the point of GRS if you cannot read from it after a failover? It provides the service needed, at the lowest cost. This would be different if there were keywords like "available immediately without downtime" or "automatically" but there are none, so well, if a region fails, you fail over, and read from secondary region. Bottom line: A. GRS

upvoted 3 times

semauni 1 year, 10 months ago

In this scenario, the data in the secondary region only needs to be available IF the data isn't available in the primary region. Both GRS and RA-GRS accomplish that. The difference between GRS and RA-GRS is that the data in RA-GRS is always readable, even if the primary region is up, which also makes it more expensive. That is not necessary in this case, so GRS is the answer. Source: https://learn.microsoft.com/en-us/azure/storage/common/storage-redundancy

upvoted 3 times

Question #18 *Topic 1*

You plan to implement an Azure Data Lake Gen 2 storage account.
You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs.
Which type of replication should you use for the storage account?

A. geo-redundant storage (GRS)

B. geo-zone-redundant storage (GZRS)

C. locally-redundant storage (LRS)

D. zone-redundant storage (ZRS)

Suggested Answer: *D*
Zone-redundant storage (ZRS) copies your data synchronously across three Azure availability zones in the primary region. Incorrect Answers: C: Locally redundant storage (LRS) copies your data synchronously three times within a single physical location in the primary region. LRS is the least expensive replication option, but is not recommended for applications requiring high availability or durability Reference: https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy

Community vote distribution



MadEgg Highly Voted 4 years, 7 months ago

Selected Answer: D

First, about the Question: What fails? -> The (complete) DataCenter, not the region and not components inside a DataCenter.

So, what helps us in this situation? LRS: "..copies your data synchronously three times within a single physical location in the primary region." Important is here the SINGLE PHYSICAL LOCATION (meaning inside the same Data Center. So in our scenario all copies wouldn't work anymore.) -> C is wrong. ZRS: "...copies your data synchronously across three Azure availability zones in the primary region" (meaning, in different Data Centers. In our scenario this would meet the requirements) -> D is right GRS/GZRS: are like LRS/ZRS but with the Data Centers in different azure regions. This works too but is more expensive than ZRS. So ZRS is the right answer.

https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy

upvoted 112 times

MadEgg Highly Voted 4 years, 7 months ago

First, about the Question: What fails? -> The (complete) DataCenter, not the region and not components inside a DataCenter. So, what helps us in this situation? LRS: "..copies your data synchronously three times within a single physical location in the primary region." Important is here the SINGLE PHYSICAL LOCATION (meaning inside the same Data Center. So in our scenario all copies wouldn't work anymore.) -> C is wrong. ZRS: "...copies your data synchronously across three Azure availability zones in the primary region" (meaning, in different Data Centers. In our scenario this would meet the requirements) -> D is right GRS/GZRS: are like LRS/ZRS but with the Data Centers in different azure regions. This works too but is more expensive than ZRS. So ZRS is the right answer. https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancyYes, well said, that's the correct answer.Well explained!

upvoted 112 times

Ozren 4 years, 5 months ago

Yes, well said, that's the correct answer.

upvoted 2 times

Narasimhap 4 years, 6 months ago

Well explained!

upvoted 2 times

JohnMasipa 4 years, 12 months ago

This can't be correct. Should be D.

upvoted 80 times

JohnMasipa 4 years, 12 months ago

This can't be correct. Should be D.Why, LRS is cheaper?It is cheaper but LRS helps to replicate data in the same data center while ZRS replicates data synchronously across three storage clusters in one region. So if one data center fails you should go for ZRS.Also, note that the question talks about

failure in "a data center". As long as other data centers are running fine(as in ZRS which will have many), ZRS would be the least expensive option.

upvoted 80 times

JayBird 4 years, 11 months ago

Why, LRS is cheaper?It is cheaper but LRS helps to replicate data in the same data center while ZRS replicates data synchronously across three storage clusters in one region. So if one data center fails you should go for ZRS.Also, note that the question talks about failure in "a data center". As long as other data centers are running fine(as in ZRS which will have many), ZRS would be the least expensive option.

upvoted 1 times

JayBird 4 years, 11 months ago

Why, LRS is cheaper?

upvoted 1 times

Vitality 4 years, 11 months ago

It is cheaper but LRS helps to replicate data in the same data center while ZRS replicates data synchronously across three storage clusters in one region. So if one data center fails you should go for ZRS.Also, note that the question talks about failure in "a data center". As long as other data centers are running fine(as in ZRS which will have many), ZRS would be the least expensive option.

upvoted 13 times

Vitality 4 years, 11 months ago

It is cheaper but LRS helps to replicate data in the same data center while ZRS replicates data synchronously across three storage clusters in one region. So if one data center fails you should go for ZRS.

upvoted 13 times

azurearmy 4 years, 9 months ago

Also, note that the question talks about failure in "a data center". As long as other data centers are running fine(as in ZRS which will have many), ZRS would be the least expensive option.

upvoted 6 times

Maor23 Most Recent 1 year, 6 months ago

GZRS combines the benefits of both geo-redundant storage and zone-redundant storage. It provides redundancy across both regions and availability zones, ensuring data availability even if a data center fails.

upvoted 1 times

EmnCours 1 year, 8 months ago

Correct Answer: D

upvoted 2 times

GameLift 1 year, 10 months ago

D Redundancy in the primary region Data in an Azure Storage account is always replicated three times in the primary region. Azure Storage offers two options for how your data is replicated in the primary region: Locally redundant storage (LRS) copies your data synchronously three times within a single physical location in the primary region. LRS is the least expensive replication option, but is not recommended for applications requiring high availability or durability. Zone-redundant storage (ZRS) copies your data synchronously across three Azure availability zones in the primary region. For applications requiring high availability, Microsoft recommends using ZRS in the primary region, and also replicating to a secondary region.

upvoted 3 times

Daarora 1 year, 10 months ago

LRS is the lowest-cost redundancy option and offers the least durability compared to other options. LRS protects your data against server rack and drive failures. However, if a disaster such as fire or flooding occurs within the data center, all replicas of a storage account using LRS may be lost or unrecoverable. To mitigate this risk, Microsoft recommends using zone-redundant storage (ZRS), geo-redundant storage (GRS), or geo-zone-redundant storage (GZRS). Source: https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy#locally-redundant-storage

upvoted 5 times

hardenlv 1 year, 11 months ago

The Correct Answer is B - based on microsoft documentation ZRS provides excellent performance, low latency, and resiliency for your data if it becomes temporarily unavailable. However, ZRS by itself might not protect your data against a regional disaster where multiple zones are permanently affected. For protection against regional disasters, we recommend using geo-zone-redundant storage (GZRS), which uses ZRS in the primary region and also geo-replicates your data to a secondary region. https://learn.microsoft.com/en-us/azure/storage/files/files-redundancy#zone-redundant-storage

upvoted 1 times

BlessedChild 2 years, 5 months ago

RA-GRS is an option on top of GRS. So, the regions defined for GRS are the same for RA-GRS

upvoted 1 times

lisa710 2 years, 8 months ago

zone-redundant storage (ZRS)

upvoted 1 times

SillyChili 2 years, 9 months ago

I don't understand why the answer is ZRS. ZRS redundancy is across availability zones, not other region. The question mentioned that "if data center fails in the primary Azure region". Isn't it ZRS will not be available when primary Azure region fails? Correct answer should then be GRS then, which redundancy is across region, and is lower cost compare to GZRS.

upvoted 2 times

74gjd_37 2 years, 11 months ago

According to Microsoft "Locally redundant storage (LRS) replicates your storage account three times within a single data center in the primary region." Therefore, if a the center fails, all three copies will be unavailable. However, according to the condition, the data lake should remain available if a data center fails. Azure Data Lake Gen 2 storage is based on blob storage. There is no separate data redundancy options for the Azure Data Lake Gen 2 comparing to that of blob storage within a storage account. Thereofer, option C is incorrect.

upvoted 1 times

kkk5566 2 years, 11 months ago

is correct

upvoted 1 times

Mani_V 3 years, 1 month ago

LRS/ZRS doesnt come to picture if anything needs to available in other regions. so GRS is rite one.

upvoted 2 times

deutscher 3 years, 4 months ago

I understood it this way, 1. LRS : single 3-storey building in Frankfurt > Each floor has a data center > if the data center fails then everything is lost 2. LRS: Single 3-storey building in Frankfurt and Berlin, if data in Frankfurt center is lost, then we still have in Berlin Hence it's even cheaper because they are in the same Geolocation

upvoted 2 times

anoj_cha 3 years, 6 months ago

Has the question recently changed? Most of the conversation below is talking about zone failures (Availability Zone) whereas the question is talking about a "primary Azure region" (region). In case of a region failure, would a GRS not be required as the ZRS will protect one if one of the availability zones go down (and not the entire region)?sorry ignore above... i reread the question "one of the data centres in the primary region".

upvoted 1 times

anoj_cha 3 years, 6 months ago

Has the question recently changed? Most of the conversation below is talking about zone failures (Availability Zone) whereas the question is talking about a "primary Azure region" (region). In case of a region failure, would a GRS not be required as the ZRS will protect one if one of the availability zones go down (and not the entire region)?

upvoted 1 times

anoj_cha 3 years, 6 months ago

sorry ignore above... i reread the question "one of the data centres in the primary region".

upvoted 1 times

nicky87654 3 years, 7 months ago

that the data lake remains available if a data center fails in the primary Azure region, while minimizing costs, you should use geo-redundant storage (GRS) for the storage account. GRS stores 3 copies of the data across 2 regions, so that if a data center fails in the primary region, the data can still be accessed from the secondary region and you only pay for the primary region's storage cost.

upvoted 1 times

greenlever 3 years, 10 months ago

Microsoft recommends using ZRS in the primary region for Azure Data Lake Storage Gen2 workloads.

upvoted 1 times

HOTSPOT -

You have a SQL pool in Azure Synapse.

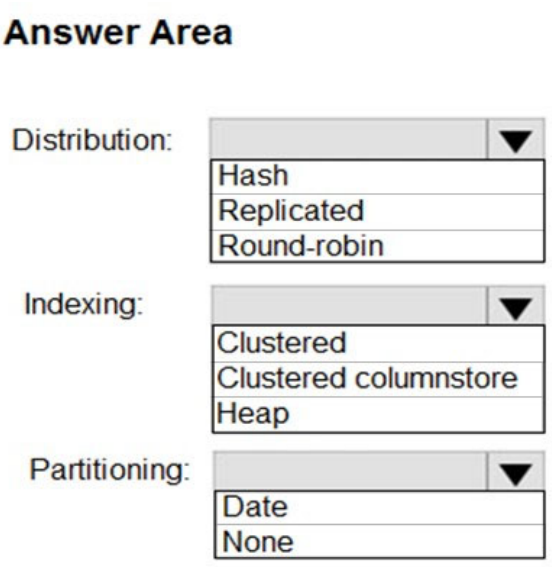
You plan to load data from Azure Blob storage to a staging table. Approximately 1 million rows of data will be loaded daily. The table will be truncated before each daily load.

You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.

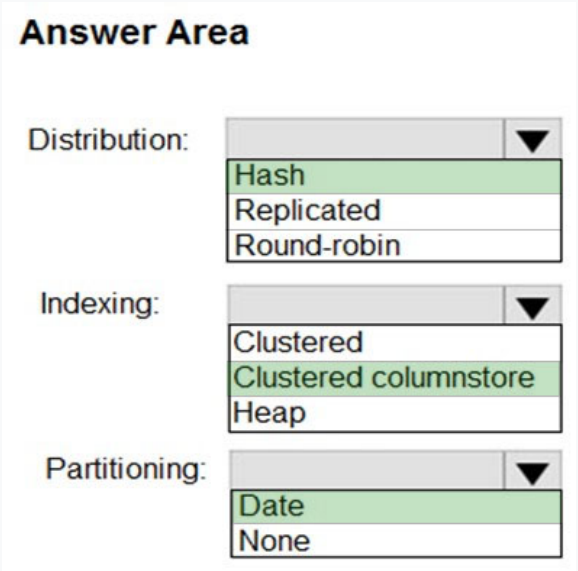
How should you configure the table? To answer, select the appropriate options in the answer area.

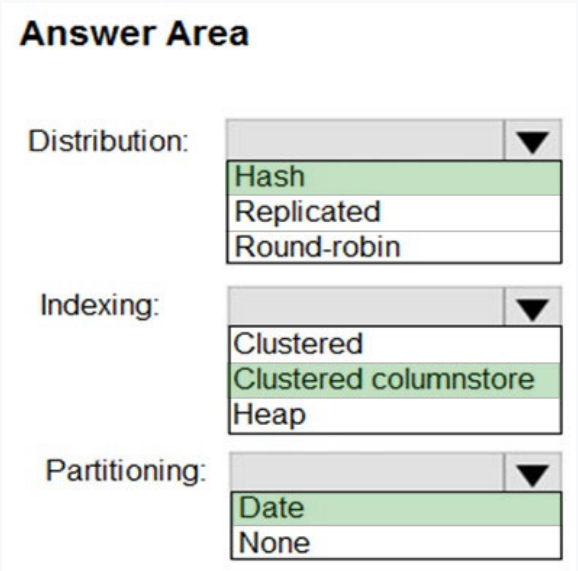
NOTE: Each correct selection is worth one point.

Hot Area:



Suggested Answer:





Box 1: Hash - Hash-distributed tables improve query performance on large fact tables. They can have very large numbers of rows and still achieve high performance. Incorrect Answers: Round-robin tables are useful for improving loading speed.

Box 2: Clustered columnstore - When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed.

Box 3: Date - Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column. Partition switching can be used to quickly remove or replace a section of a table. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

A1000 Highly Voted 4 years, 12 months ago

Round-Robin Heap NoneRound- Robin Heap None. No brainer for this question.Agree - Round Robin, Heap and None are the correct options. The solution doesn't make any sense for a staging table to truncate every day without any JOINs.Don't know why it says Hash/Columnstored/Date.Not sure if the table is already to larg for heap? "Do not use a heap when there are no nonclustered indexes and the table is large, unless you intend to return the entire table content without any specified order. In a heap, all rows of the heap must be read to find any row." https://docs.microsoft.com/en-us/sql/relational-databases/indexes/heaps-tables-without-clustered-indexes?view=sql-server-ver16I agree too

upvoted 428 times

A1000 Highly Voted 4 years, 12 months ago

Round-Robin Heap None

upvoted 428 times

Narasimhap 4 years, 6 months ago

Round- Robin Heap None. No brainer for this question.

upvoted 28 times

Homer23 2 years, 5 months ago

Agree - Round Robin, Heap and None are the correct options. The solution doesn't make any sense for a staging table to truncate every day without any JOINs.Don't know why it says Hash/Columnstored/Date.

upvoted 2 times

azure900test 4 years, 1 month ago

Not sure if the table is already to larg for heap? "Do not use a heap when there are no nonclustered indexes and the table is large, unless you intend to return the entire table content without any specified order. In a heap, all rows of the heap must be read to find any row."
https://docs.microsoft.com/en-us/sql/relational-databases/indexes/heaps-tables-without-clustered-indexes?view=sql-server-ver16

upvoted 1 times

anto69 4 years, 7 months ago

I agree too

upvoted 7 times

laszek 4 years, 11 months ago

Round-robin - this is the simplest distribution model, not great for querying but fast to process Heap - no brainer when creating staging tables No partitions - this is a staging table, why add effort to partition, when truncated daily?Can you explain me why should we use heap?The term heap basically refers to a table without a clustered index. Adding a clustered index to a temp table makes absolutely no sense and is a waste of compute resources for a table that would be entirely truncated daily. no clustered index = heap.DrTaz is right, in addition, when you populate an indexed table, you are also writing to the index, so this adds an additional overhead in the write processHad doubts regarding why there is no need for a partition. While what you suggested is true won't it be better if there is a date partition to truncate the table ?There is no filter on a truncate statement so no benefit in having a partition

upvoted 68 times

laszek 4 years, 11 months ago

Round-robin - this is the simplest distribution model, not great for querying but fast to process Heap - no brainer when creating staging tables No partitions - this is a staging table, why add effort to partition, when truncated daily?

upvoted 68 times

Vardhan_Brahmanapally 4 years, 9 months ago

Can you explain me why should we use heap?The term heap basically refers to a table without a clustered index. Adding a clustered index to a temp table makes absolutely no sense and is a waste of compute resources for a table that would be entirely truncated daily. no clustered index = heap.DrTaz is right, in addition, when you populate an indexed table, you are also writing to the index, so this adds an additional overhead in the write process

upvoted 1 times

Vardhan_Brahmanapally 4 years, 9 months ago

Can you explain me why should we use heap?

upvoted 1 times

DrTaz 4 years, 7 months ago

The term heap basically refers to a table without a clustered index. Adding a clustered index to a temp table makes absolutely no sense and is a waste of compute resources for a table that would be entirely truncated daily.

no clustered index = heap.

upvoted 13 times

DrTaz 4 years, 7 months ago

The term heap basically refers to a table without a clustered index. Adding a clustered index to a temp table makes absolutely no sense and is a waste of compute resources for a table that would be entirely truncated daily. no clustered index = heap.DrTaz is right, in addition, when you populate an indexed table, you are also writing to the index, so this adds an additional overhead in the write process

upvoted 13 times

SQLDev0000 4 years, 5 months ago

DrTaz is right, in addition, when you populate an indexed table, you are also writing to the index, so this adds an additional overhead in the write process

upvoted 3 times

berserksap 4 years, 9 months ago

Had doubts regarding why there is no need for a partition. While what you suggested is true won't it be better if there is a date partition to truncate the table ?

upvoted 2 times

berserksap 4 years, 9 months ago

Had doubts regarding why there is no need for a partition. While what you suggested is true won't it be better if there is a date partition to truncate the table ?There is no filter on a truncate statement so no benefit in having a partition

upvoted 2 times

andy_g 4 years, 6 months ago

There is no filter on a truncate statement so no benefit in having a partition

upvoted 4 times

HaliBrickclay Most Recent 1 year, 10 months ago

as per Microsoft document Load to a staging table To achieve the fastest loading speed for moving data into a data warehouse table, load data into a staging table. Define the staging table as a heap and use round-robin for the distribution option. Consider that loading is usually a two-step process in which you first load to a staging table and then insert the data into a production data warehouse table. If the production table uses a hash distribution, the total time to load and insert might be faster if you define the staging table with the hash distribution. Loading to the staging table takes longer, but the second step of inserting the rows to the production table does not incur data movement across the distributions.It doesn't mention the prd table. Only the staging. So, round Robin/Heap is the answer, correct? tricky questions. :)

upvoted 5 times

HaliBrickclay 1 year, 10 months ago

as per Microsoft document

Load to a staging table To achieve the fastest loading speed for moving data into a data warehouse table, load data into a staging table. Define the staging table as a heap and use round-robin for the distribution option.

Consider that loading is usually a two-step process in which you first load to a staging table and then insert the data into a production data warehouse table. If the production table uses a hash distribution, the total time to load and insert might be faster if you define the staging table with the hash distribution. Loading to the staging table takes longer, but the second step of inserting the rows to the production table does not incur data movement across the distributions.

upvoted 5 times

VeroDon 4 years, 7 months ago

It doesn't mention the prd table. Only the staging. So, round Robin/Heap is the answer, correct? tricky questions. :)

upvoted 2 times

Sasha_in_San_Francisco 1 year, 10 months ago

Answer: Round-Robin (1), Heap (2), None (3). Within this doc: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-overview #1. Search for “Use round-robin for the staging table.” #2. Search for: “A heap table can be especially useful for loading data, such as a staging table,…” Within this doc: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition?context=/azure/synapse-analytics/context/context #3. Partitioning by date is useful when stage destination has data because you can hide the inserting data’s new partition (to keep users from hitting it), complete the load and then unhide the new partition. However, in this question it states, “the table will be truncated before each daily load”, so, it appears it’s a true Staging table and there are no users with access, no existing data, and I see no reason to have a Date partition. To me, such a partition would do nothing but slow the load.

upvoted 17 times

jhargett1 1 year, 10 months ago

Since it's the staging table, the main focus should be minimizing the load time, as noted in the question. Heap table does not have any index and it's the fastest option for loading large amounts of data. Using a round-robin distribution will help to evenly distribute the data across all the distributions, further reducing the load time. As the data is truncated before each load, partitioning is not necessary, so it is best to choose None. Round-robin Heap None

upvoted 1 times

akk_1289 1 year, 10 months ago

Distribution: Round Robin Indexing: Clustered Columnstore Partitioning: Date The recommended configuration for a staging table that will be loaded daily with approximately 1 million rows of data and truncated before each load is to use a round robin distribution, a clustered columnstore index, and date-based partitioning. Round robin distribution will evenly distribute the data across nodes, reducing the load time. Clustered columnstore index provides efficient compression and supports fast bulk load operations. Date-based partitioning will allow for easy archiving and maintenance of the table.

upvoted 2 times

74gjd_37 1 year, 10 months ago

Round-Robing, Heap, None. The question is to configure the staging table. According to the conditions, "The solution must minimize how long it takes to load the data to the staging table." Therefore, loading time is the most essential condition here. According to Microsoft documentation at https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-overview Distribution: A round-robin table distributes table rows evenly across all distributions. The rows are distributed randomly. Loading data into a round-robin table is fast. But, queries can require more data movement than the other distribution methods. Indexing: A heap table can be especially useful for loading transient data, such as a staging table, which is transformed into a final table. Partitioning: None. Since the table is truncated before each daily load, we can not benefit of partitioning to drop date ranges.

upvoted 6 times

MBRSDG 1 year, 10 months ago

1 --> REPLICATED Since table has always 1mln rows (it is a small staging table, loaded in truncate-insert everyday!) Also Round-Robin is wrong, since 1mln/60nodes = almost 17k rows per compute node, which is a unrecommended situation. You haven't data enough to choose round-robin. 2 --> HEAP Why do you have to choose clustered columnstore index? This is not a big table. 3 --> None Since you haven't data enough to create a efficient partition. It's more convenient to have only one partition in this case.

upvoted 3 times

iceberge 2 years ago

Copilot Sent by Copilot: To optimize the loading process for your staging table in Azure Synapse, here are the recommended configurations: Distribution: Use Round-robin distribution. This method evenly distributes the data across all distributions, which helps in achieving balanced data loading and minimizes data skew. Indexing: Use Heap. Since the table will be truncated before each load, using a heap (a table without a clustered index) will speed up the data loading process as it avoids the overhead of maintaining indexes during the load. Partitioning: Do not partition the staging table. Partitioning can add overhead to the data loading process. Since the table is truncated daily, partitioning is not necessary and can be avoided to keep the load process efficient. These configurations will help minimize the time it takes to load data into the staging table.

upvoted 2 times

RakshithaReddy 2 years, 2 months ago

RR Heap Date

upvoted 1 times

Dusica 2 years, 3 months ago

Round robin Heap None

upvoted 2 times

Alongi 2 years, 4 months ago

RR Heap None

upvoted 1 times

__Tom 2 years, 9 months ago

Round-Robin Heap None

upvoted 5 times

kkk5566 2 years, 12 months ago

Round-Robin Heap None

upvoted 1 times

janaki 3 years, 2 months ago

Never ever use date partioning with hash distribution. The correct answer is: Round robin, Heap and None

upvoted 2 times

rocky48 3 years, 3 months ago

Round-robin - this is the simplest distribution model, not great for querying but fast to process Heap - no brainer when creating staging tables No partitions - this is a staging table

upvoted 1 times

Question #20 *Topic 1*

You are designing a fact table named FactPurchase in an Azure Synapse Analytics dedicated SQL pool. The table contains purchases from suppliers for a retail store. FactPurchase will contain the following columns.

FactPurchase will have 1 million rows of data added daily and will contain three years of data.

Transact-SQL queries similar to the following query will be executed daily.

```
SELECT -
SupplierKey, StockItemKey, IsOrderFinalized, COUNT(*)
FROM FactPurchase -
WHERE DateKey >= 20210101 -
AND DateKey <= 20210131 -
GROUP By SupplierKey, StockItemKey, IsOrderFinalized
```

Which table distribution will minimize query times?

Name	Data type	Nullable
PurchaseKey	Bigint	No
DateKey	Int	No
SupplierKey	Int	No
StockItemKey	Int	No
PurchaseOrderID	Int	Yes
OrderedQuantity	Int	No
OrderedOuters	Int	No
ReceivedOuters	Int	No
Package	Nvarchar(50)	No
IsOrderFinalized	Bit	No
LineageKey	Int	No

A. replicated

B. hash-distributed on PurchaseKey

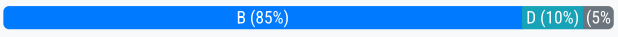
C. round-robin

D. hash-distributed on IsOrderFinalized

Suggested Answer: *B*

Hash-distributed tables improve query performance on large fact tables. To balance the parallel processing, select a distribution column that: ✑ Has many unique values. The column can have duplicate values. All rows with the same value are assigned to the same distribution. Since there are 60 distributions, some distributions can have > 1 unique values while others may end with zero values. ✑ Does not have NULLs, or has only a few NULLs. ✑ Is not a date column. Incorrect Answers: C: Round-robin tables are useful for improving loading speed. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

Community vote distribution



FredNo Highly Voted 4 years, 8 months ago

CorrectB is correct

upvoted 47 times

FredNo Highly Voted 4 years, 8 months ago

Selected Answer: B

Correct

upvoted 47 times

Deepshikha1228 4 years ago

B is correct

upvoted 2 times

GameLift 4 years, 11 months ago

Is it hash-distributed on PurchaseKey and not on IsOrderFinalized because 'IsOrderFinalized' yields less distributions(rows either contain yes,no values) compared to PurchaseKey?

upvoted 31 times

GameLift 4 years, 11 months ago

Is it hash-distributed on PurchaseKey and not on IsOrderFinalized because 'IsOrderFinalized' yields less distributions(rows either contain yes,no values) compared to PurchaseKey?Yes, your logic is correct!

upvoted 31 times

Podavenna 4 years, 11 months ago

Yes, your logic is correct!

upvoted 9 times

adaod14 Most Recent 1 year, 5 months ago

Correct Answer: B. Hash-Distributed on PurchaseKey Explanation: Azure Synapse Analytics supports three table distribution methods: Replicated, Hash-Distributed, and Round-Robin. The best choice depends on the workload and query patterns. 1 Why Use Hash Distribution? The FactPurchase table will have 1 million rows added daily and contain three years of data → This results in over 1 billion rows, making hash distribution the best option for efficient querying. Queries GROUP BY SupplierKey, StockItemKey, and IsOrderFinalized → A hash-distributed table will ensure these columns are evenly distributed across compute nodes, reducing data movement and improving query performance. 2 Why Hash on PurchaseKey? PurchaseKey is the primary key (unique for each purchase). Ideal for distributing data evenly across compute nodes, avoiding data skew. Improves aggregation queries where GROUP BY is used on multiple dimensions.

upvoted 4 times

RAG11 1 year, 5 months ago

D, Since the query doesn't reference PurchaseKey for filtering or grouping, hash distribution on PurchaseKey doesn't help with optimizing the query and would not be the best choice.

upvoted 1 times

thanglai 1 year, 7 months ago

Hash-Distributed: For large fact tables where joins or aggregations are common

upvoted 2 times

EmnCours 1 year, 8 months ago

Correct Answer: B

upvoted 1 times

Okkier 2 years, 1 month ago

Selected Answer: C

Optimal Distribution Given that the query performs a GROUP BY on SupplierKey, StockItemKey, and IsOrderFinalized, the most balanced approach is to use Round-robin distribution. While it does not ensure that rows with the same key are stored together, it avoids data skew and ensures even distribution, which helps in achieving better performance for aggregate queries.

upvoted 3 times

Okkier 2 years, 1 month ago

Optimal Distribution Given that the query performs a GROUP BY on SupplierKey, StockItemKey, and IsOrderFinalized, the most balanced approach is to use Round-robin distribution. While it does not ensure that rows with the same key are stored together, it avoids data skew and ensures even distribution, which helps in achieving better performance for aggregate queries.I hear where you are coming from. The fact that the PurchaseKey is not used in the query also made me think this for a sec. So even when using the hash the situation becomes like a round-robin distribution in terms of data access in this specific query. But generally hash will be more efficient because of the predictable hash distribution structure, which is better than a completely random approach. Seeing as we are using the purchaskey which i assume is unique, we don't need to worry about skew at all. So I Still think B is the correct choice here.hash distribution on high cardinatlity column can also avoid data skew

upvoted 3 times

JustImperius 1 year, 7 months ago

I hear where you are coming from. The fact that the PurchaseKey is not used in the query also made me think this for a sec. So even when using the hash the situation becomes like a round-robin distribution in terms of data access in this specific query. But generally hash will be more efficient because of the predictable hash distribution structure, which is better than a completely random approach. Seeing as we are using the purchaskey which i assume is unique, we don't need to worry about skew at all. So I Still think B is the correct choice here.

upvoted 1 times

hypersam 1 year, 7 months ago

hash distribution on high cardinatlity column can also avoid data skew

upvoted 1 times

MBRSDG 2 years, 4 months ago

almost exactly what's shown in a example of the official docs --> https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute#choose-a-distribution-column

upvoted 2 times

sdg2844 2 years, 7 months ago

Correct. Column with many unique values. Also, it's USUALLY not a column that is used in whereclauses or groupings or such, which this isn't.

upvoted 1 times

pperf 2 years, 10 months ago

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute#choosing-a-distribution-column

upvoted 1 times

jiajiani 2 years, 11 months ago

why the answer says it cannot be a date column?

upvoted 1 times

74gjd_37 2 years, 11 months ago

Hash-distributed tables improve query performance on large fact tables. The PurchaseKey has many unique values, does not have NULLs and is not a date column.why we cannot use data column?

upvoted 2 times

74gjd_37 2 years, 11 months ago

Selected Answer: B

Hash-distributed tables improve query performance on large fact tables. The PurchaseKey has many unique values, does not have NULLs and is not a date column.

upvoted 2 times

jiajiani 2 years, 11 months ago

why we cannot use data column?

upvoted 3 times

kkk5566 2 years, 11 months ago

B is correct

upvoted 1 times

SolutionA 3 years ago

in this case the sql where condition is on datekey so hash-distributed on PurchaseKey or Round robin distributed table the sql cost will be the same as it will be full table scanon second thought if purchasekey is not unique what is the constraint and how its created , as the question didn't mention more details , i would go with round robin not the has distributed

upvoted 1 times

SolutionA 3 years ago

in this case the sql where condition is on datekey so hash-distributed on PurchaseKey or Round robin distributed table the sql cost will be the same as it will be full table scan

upvoted 1 times

SolutionA 3 years ago

on second thought if purchasekey is not unique what is the constraint and how its created , as the question didn't mention more details , i would go with round robin not the has distributed

upvoted 3 times

mamahani 3 years, 3 months ago

B is correct

upvoted 1 times

henryphchan 3 years, 3 months ago

B. Hash the purchasekey to evenly distribute the data into 60 distributions.

upvoted 1 times

SHENOOOO 3 years, 6 months ago

B is the Correct Answer

upvoted 3 times

HOTSPOT -

From a website analytics system, you receive data extracts about user interactions such as downloads, link clicks, form submissions, and video plays.

The data contains the following columns.

You need to design a star schema to support analytical queries of the data. The star schema will contain four tables including a date dimension.

To which table should you add each column? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Name	Sample value
Date	15 Jan 2021
EventCategory	Videos
EventAction	Play
EventLabel	Contoso Promotional
ChannelGrouping	Social
TotalEvents	150
UniqueEvents	120
SessionWithEvents	99

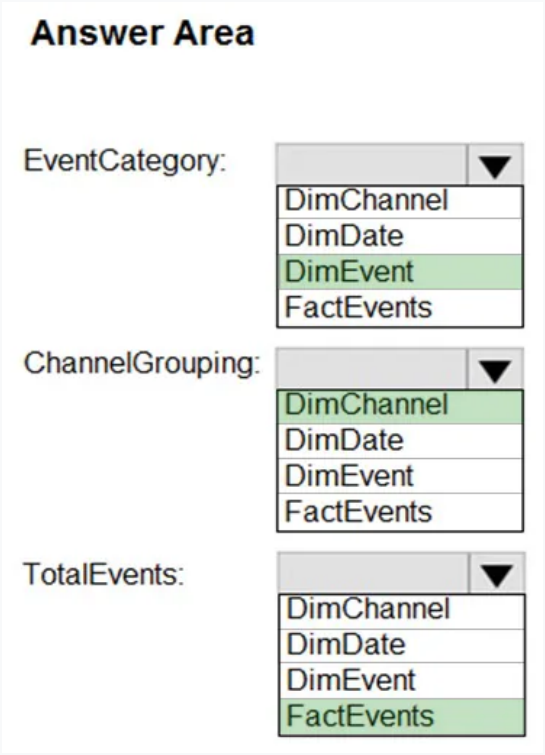
Answer Area

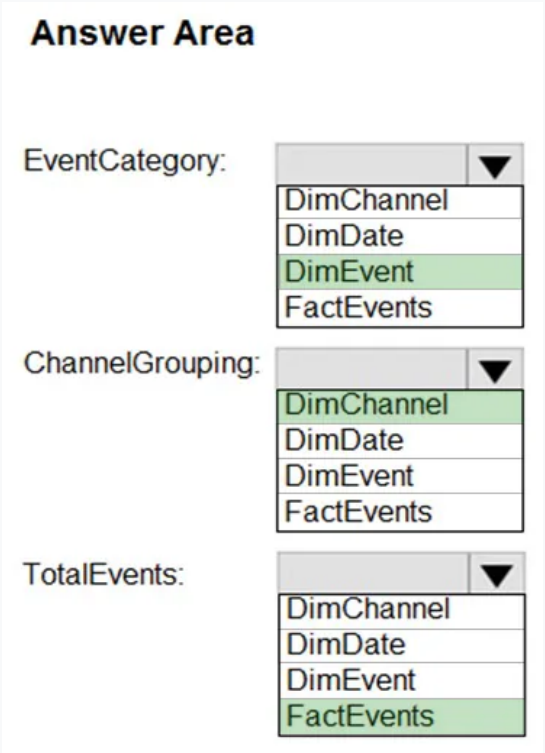
EventCategory: DimChannel / DimDate / DimEvent / FactEvents

ChannelGrouping: DimChannel / DimDate / DimEvent / FactEvents

TotalEvents: DimChannel / DimDate / DimEvent / FactEvents

Suggested Answer:





Box 1: DimEvent -

Box 2: DimChannel -

Box 3: FactEvents - Fact tables store observations or events, and can be sales orders, stock balances, exchange rates, temperatures, etc

Reference: https://docs.microsoft.com/en-us/power-bi/guidance/star-schema

gssd4scoder Highly Voted 4 years, 3 months ago

It seems to be correct

upvoted 73 times

DingDongSingSong Highly Voted 3 years, 11 months ago

What is this question? It is poorly written. I couldn't even understand what's being asked here. It talks about 4 tables, yet the answer shows 3. Then, the columns mentioned in the question don't match the column/attributes shown in the 3 tables noted in the answer.the question is clear try to read it again, there is no need to design DIM_DATEQuestion says,including a dimTime table so we need to design only 3 tables.

upvoted 27 times

DingDongSingSong 3 years, 11 months ago

What is this question? It is poorly written. I couldn't even understand what's being asked here. It talks about 4 tables, yet the answer shows 3. Then, the columns mentioned in the question don't match the column/attributes shown in the 3 tables noted in the answer.

upvoted 27 times

allagowf 3 years, 4 months ago

the question is clear try to read it again, there is no need to design DIM_DATE

upvoted 1 times

sdegcp 3 years, 7 months ago

Question says,including a dimTime table so we need to design only 3 tables.

upvoted 4 times

dgerok Most Recent 1 year, 10 months ago

Correct answer

upvoted 1 times

Alongi 1 year, 11 months ago

Correct answer

upvoted 1 times

hassexat 2 years, 5 months ago

Correct provided answer

upvoted 1 times

kkk5566 2 years, 5 months ago

EventCategory -> dimEvent channelGrouping -> dimChannel TotalEvents -> factEven

upvoted 1 times

mamahani 2 years, 9 months ago

DimEvent / DimChannel / FactEvents

upvoted 1 times

SHENOOOO 3 years ago

Category or Group will go with the table DIM, Total Events will go with the Fact

upvoted 2 times

Deeksha1234 3 years, 6 months ago

correct

upvoted 1 times

Rrk07 3 years, 8 months ago

EventCategory -> dimEvent channelGrouping -> dimChannel TotalEvents -> factEven

upvoted 4 times

Dothy 3 years, 9 months ago

EventCategory -> dimEvent channelGrouping -> dimChannel TotalEvents -> factEven

upvoted 1 times

JJdeWit 3 years, 9 months ago

EventCategory ==> dimEvent channelGrouping ==> dimChannel TotalEvents ==> factEvent Explanation: A bit of knowledge of Google Analytics Universal helps to understand this question. eventCategory, eventAction and eventLabel all contain information about the event/action done on the website, and can be logically be grouped together. ChannelGrouping is about how the user came on the website (through Google, and advertisement, an email link, etc.) and is not related to events at all. It therefore would make sense to put it in a second dim table.

upvoted 4 times

Mahesh_mm 4 years, 1 month ago

Answer is correct

upvoted 4 times

laszek 4 years, 5 months ago

I would add ChannelGrouping to DimEvents table. What would DimChannel table contain? only one column? No sense to mel mean if you think about it, ChannelName (facebook,google,youtube), ChannelType (paid media, free posts, ads), ChannleDelivery (chrome, etc etc). Just thinking out loudIt is supposed to contain 4 tables. Date, Event, Fact so the logical conclusion would be to include the channel dimension. If it were up to me though I'd use the channel as a degenerate dimension and store it in fact table if it's the only information that we have provided.

upvoted 4 times

laszek 4 years, 5 months ago

I would add ChannelGrouping to DimEvents table. What would DimChannel table contain? only one column? No sense to me

upvoted 4 times

Seansmyrke 3 years, 11 months ago

I mean if you think about it, ChannelName (facebook,google,youtube), ChannelType (paid media, free posts, ads), ChannleDelivery (chrome, etc etc). Just thinking out loud

upvoted 1 times

manquak 4 years, 5 months ago

It is supposed to contain 4 tables. Date, Event, Fact so the logical conclusion would be to include the channel dimension. If it were up to me though I'd use the channel as a degenerate dimension and store it in fact table if it's the only information that we have provided.

upvoted 3 times

Question #22 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly.
Solution: You convert the files to compressed delimited text files.
Does this meet the goal?

A. Yes

B. No

Suggested Answer: *A*
All file formats have different performance characteristics. For the fastest load, use compressed delimited text files. Reference: https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

Community vote distribution
A (78%) B (22%)

Fahd92 Highly Voted 4 years, 10 months ago
They said you need to prepare the files to copy, maybe the mean we should make them less than 1MB ? so it will be A else would be B !!!!
upvoted 18 times

Fahd92 Highly Voted 4 years, 10 months ago
They said you need to prepare the files to copy, maybe the mean we should make them less than 1MB ? so it will be A else would be B !!!!The answer should be A. https://azure.microsoft.com/en-gb/blog/increasing-polybase-row-width-limitation-in-azure-sql-data-warehouse/After reading the other questions oh this topic I go with A because the relevant part seems to be the compression.i think when compression is in question we should go for parquet/avro bcz only they give compression
upvoted 18 times

ANath 4 years, 7 months ago
The answer should be A. https://azure.microsoft.com/en-gb/blog/increasing-polybase-row-width-limitation-in-azure-sql-data-warehouse/
upvoted 5 times

Thij 4 years, 10 months ago
After reading the other questions oh this topic I go with A because the relevant part seems to be the compression.i think when compression is in question we should go for parquet/avro bcz only they give compression
upvoted 4 times

Thij 4 years, 10 months ago
After reading the other questions oh this topic I go with A because the relevant part seems to be the compression.
upvoted 4 times

dakku987 2 years, 8 months ago
i think when compression is in question we should go for parquet/avro bcz only they give compression
upvoted 1 times

Amino23 Most Recent 1 year, 4 months ago
NO, what would accelarate the copying is using polybase and hence multiple splitted files.
upvoted 1 times

thespy 1 year, 5 months ago
Parquet for column level compression and better compression
upvoted 1 times

Pey1nkh 1 year, 6 months ago
While compressing delimited text files reduces file size, it does not optimize for fast loading into Azure Synapse Analytics.Use Parquet format !
upvoted 1 times

iam_momo88 1 year, 7 months ago

Yes, converting the files to compressed delimited text files is a good solution to ensure the data copies quickly to Azure Synapse Analytics.

upvoted 1 times

EmnCours 1 year, 8 months ago

Correct Answer: A

upvoted 1 times

akhil5432 1 year, 11 months ago

"a" is correct option

upvoted 1 times

kkk5566 1 year, 11 months ago

Answer is yes

upvoted 1 times

moneytime 1 year, 11 months ago

The answer is A Compression doesn't not only help to reduce the size or space occupied by a file in a storage but also increases the speed of file movement during transfer

upvoted 2 times

SHENOOOO 3 years, 6 months ago

A will do the job

upvoted 3 times

Rrk07 3 years, 8 months ago

Delimited text file is true,

upvoted 1 times

greenlever 3 years, 10 months ago

For the fastest load, use compressed delimited text files https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/data-loading-best-practices

upvoted 4 times

Deeksha1234 4 years ago

yes, answer is A

upvoted 1 times

Janisys 4 years ago

Selected Answer: A PolyBase can't load rows that have more than 1,000,000 bytes of data. When you put data into the text files in Azure Blob storage or Azure Data Lake Store, they must have fewer than 1,000,000 bytes of data. This byte limitation is true regardless of the table schema. All file formats have different performance characteristics. For the fastest load, use compressed delimited text files. Split large compressed files into smaller compressed files.PolyBase can load more than 1 mln rows since SQL Server 2019 https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-versioned-feature-summary?view=sql-server-ver16

upvoted 3 times

Janisys 4 years ago

Selected Answer: A PolyBase can't load rows that have more than 1,000,000 bytes of data. When you put data into the text files in Azure Blob storage or Azure Data Lake Store, they must have fewer than 1,000,000 bytes of data. This byte limitation is true regardless of the table schema. All file formats have different performance characteristics. For the fastest load, use compressed delimited text files. Split large compressed files into smaller compressed files.

upvoted 3 times

dgerok 2 years, 4 months ago

PolyBase can load more than 1 mln rows since SQL Server 2019 https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-versioned-feature-summary?view=sql-server-ver16

upvoted 1 times

Deepshikha1228 4 years ago

A is correct ,with copy command PolyBase COPY Needs CONTROL permission Relaxed permission Has row width limits No row width limit No delimiters within text Supports delimiters in text Fixed line delimiter Supports custom column and row delimiters Complex to set up in code Reduces amount of code

upvoted 2 times

objecto 4 years, 2 months ago

It's just a copy to storage so zipping it will work fine.

upvoted 4 times

Rrk07 4 years, 2 months ago

It says about files compression. which will reduce the file size. so Answer is correct

upvoted 1 times

Question #23 *Topic 1*

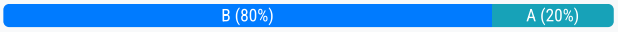
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly.
Solution: You copy the files to a table that has a columnstore index.
Does this meet the goal?

A. Yes

B. No

Suggested Answer: *B*
Instead convert the files to compressed delimited text files. Reference: https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

Community vote distribution



bhanuprasad9331 Highly Voted 4 years, 6 months ago
From the documentation, loads to heap table are faster than indexed tables. So, better to use heap table than columnstore index table in this case.
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-index#heap-tables
upvoted 12 times

Odoxtoom 4 years, 10 months ago
Consider this sets one question: What should you do to improve loading times? What | Yes | No | compressed | O | O | columnstore | O | O | > 1MB | O | O | So now answers should be clearCan You explain this in more details?I really didn't understand this , can you explain?it's a virtualized chart that the guy want to simplified the question set.but it becomes more complicated with this chart hahaI Think what he tried to show was: Set Answer Matrix What should you do to improve loading times? What | Yes | No | compressed | X | O | columnstore | O | X | > 1MB | O | X | So all three variations of this question and x is marking the correct answer.
upvoted 9 times

Odoxtoom 4 years, 10 months ago
Consider this sets one question: What should you do to improve loading times? What | Yes | No | compressed | O | O | columnstore | O | O | > 1MB | O | O |

So now answers should be clear
upvoted 9 times

Julius7000 Highly Voted 4 years, 10 months ago
Can You explain this in more details?
upvoted 11 times

helly13 4 years, 8 months ago
I really didn't understand this , can you explain?it's a virtualized chart that the guy want to simplified the question set.but it becomes more complicated with this chart haha
upvoted 7 times

helly13 4 years, 8 months ago
I really didn't understand this , can you explain?
upvoted 7 times

aurorafang 4 years ago
it's a virtualized chart that the guy want to simplified the question set.
upvoted 2 times

aurorafang 4 years ago
it's a virtualized chart that the guy want to simplified the question set.but it becomes more complicated with this chart haha

upvoted 2 times

ML_Novice 3 years, 11 months ago

but it becomes more complicated with this chart haha

upvoted 6 times

tbhtp 3 years, 11 months ago

I Think what he tried to show was: Set Answer Matrix What should you do to improve loading times? What | Yes | No | compressed | X | O | columnstore | O | X | > 1MB | O | X | So all three variations of this question and x is marking the correct answer.

upvoted 6 times

Pey1nkh Most Recent 1 year, 6 months ago

- A columnstore index stores columns separately, allowing queries to only read relevant columns, making queries much faster! - Columnstore indexes automatically compress data, reducing storage size and I/O. - This is especially useful for large text data like descriptions and numerical values

upvoted 1 times

moize 1 year, 8 months ago

Moi je vote A). Selon Copilot : Oui, copier les fichiers dans une table avec un index columnstore dans Azure Synapse Analytics peut répondre à l'objectif de copier les données rapidement. Les index columnstore sont conçus pour optimiser les performances des requêtes et la compression des données, ce qui est particulièrement utile pour les grandes quantités de données.

upvoted 1 times

EmnCours 1 year, 8 months ago

Correct Answer: B

upvoted 1 times

kkk5566 2 years, 11 months ago

Answer is no ,u use HEAP idx

upvoted 1 times

youngbug 3 years, 6 months ago

For fast loading to a table, using a staging table which is a heap table.

upvoted 1 times

DindaS 3 years, 6 months ago

its always recommended to load the data into a staging where the table should be a heap table and data will be loaded using ROUND_ROBIN mechanism

upvoted 1 times

Deeksha1234 4 years ago

B is right

upvoted 1 times

Janisys 4 years ago

Correct Answer: B To achieve the fastest loading speed for moving data into a data warehouse table, load data into a staging table. Define the staging table as a heap and use round-robin for the distribution option

upvoted 5 times

Deepshikha1228 4 years ago

B is right

upvoted 1 times

Amsterliese 4 years, 4 months ago

Columnstore index would be used for faster reading, but the question is only about faster loading. So for faster loading you want the least possible overhead. So the answer should be no. Am I right?Yes load to a table without indexes for faster load right?

upvoted 4 times

Amsterliese 4 years, 4 months ago

Columnstore index would be used for faster reading, but the question is only about faster loading. So for faster loading you want the least possible overhead. So the answer should be no. Am I right?

upvoted 4 times

Muishkin 4 years, 3 months ago

Yes load to a table without indexes for faster load right?

upvoted 1 times

lionurag 4 years, 5 months ago

B is correct

upvoted 3 times

PallaviPatel 4 years, 6 months ago

B is correct.

upvoted 1 times

DE_Sanjay 4 years, 7 months ago

NO is the answer.

upvoted 1 times

Mahesh_mm 4 years, 7 months ago

B is correct

upvoted 1 times

rashjan 4 years, 8 months ago

Correct Answer: No.

upvoted 2 times

Question #24 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly.
Solution: You modify the files to ensure that each row is more than 1 MB.
Does this meet the goal?

A. Yes

B. No

Suggested Answer: *B*
Instead convert the files to compressed delimited text files. Reference: https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

Community vote distribution
B (100%)

Gilvan Highly Voted 4 years, 11 months ago
No, rows need to have less than 1 MB. A batch size between 100 K to 1M rows is the recommended baseline for determining optimal batch size capacity.
upvoted 13 times

amarG1996 Highly Voted 4 years, 7 months ago
PolyBase can't load rows that have more than 1,000,000 bytes of data. When you put data into the text files in Azure Blob storage or Azure Data Lake Store, they must have fewer than 1,000,000 bytes of data. This byte limitation is true regardless of the table schema.
upvoted 8 times

amarG1996 4 years, 7 months ago
PolyBase can't load rows that have more than 1,000,000 bytes of data. When you put data into the text files in Azure Blob storage or Azure Data Lake Store, they must have fewer than 1,000,000 bytes of data. This byte limitation is true regardless of the table schema.is it stated anywhere that we have to use PolyBase? What about COPY command?https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/data-loading-best-practices#prepare-data-in-azure-storage
upvoted 8 times

kamil_k 4 years, 5 months ago
is it stated anywhere that we have to use PolyBase? What about COPY command?
upvoted 2 times

amarG1996 4 years, 7 months ago
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/data-loading-best-practices#prepare-data-in-azure-storage
upvoted 2 times

moize Most Recent 1 year, 8 months ago
Selon Copilot : Non, modifier les fichiers pour garantir que chaque ligne fait plus de 1 Mo n'est pas nécessairement la meilleure solution pour garantir une copie rapide des données vers Azure Synapse Analytics. En fait, cela pourrait même compliquer le processus de copie et de traitement des données.
upvoted 2 times

EmnCours 1 year, 8 months ago
Correct Answer: B
upvoted 1 times

kkk5566 2 years, 11 months ago
Answer is no
upvoted 1 times

vigilante89 3 years, 8 months ago

B is correct!!!

upvoted 2 times

Deeksha1234 4 years ago

B is correct, agree with explanation by Amar

upvoted 1 times

PallaviPatel 4 years, 6 months ago

B is correct.

upvoted 4 times

Mahesh_mm 4 years, 7 months ago

Answer is No

upvoted 1 times

rashjan 4 years, 8 months ago

Correct Answer: No.

upvoted 2 times

Odoxtoom 4 years, 10 months ago

Consider this sets one question: What should you do to improve loading times? What | Yes | No | compressed | O | O | columnstore | O | O | > 1MB | O | O |

So now answers should be clear

upvoted 1 times

Odoxtoom 4 years, 10 months ago

Consider this sets one question: What should you do to improve loading times? What | Yes | No | compressed | O | O | columnstore | O | O | > 1MB | O | O | So now answers should be clear@Odoxtoom, can you please explain your answer and specify based on this matrix which option is correct.Yes No No

upvoted 1 times

Aslam208 4 years, 8 months ago

@Odoxtoom, can you please explain your answer and specify based on this matrix which option is correct.Yes No No

upvoted 5 times

Aslam208 4 years, 8 months ago

@Odoxtoom, can you please explain your answer and specify based on this matrix which option is correct.

upvoted 5 times

Bishtu 4 years, 7 months ago

Yes No No

upvoted 2 times

Question #25 *Topic 1*

You build a data warehouse in an Azure Synapse Analytics dedicated SQL pool.
Analysts write a complex SELECT query that contains multiple JOIN and CASE statements to transform data for use in inventory reports. The inventory reports will use the data and additional WHERE parameters depending on the report. The reports will be produced once daily.
You need to implement a solution to make the dataset available for the reports. The solution must minimize query times.
What should you implement?

A. an ordered clustered columnstore index

B. a materialized view

C. result set caching

D. a replicated table

Suggested Answer: *B*
Materialized views for dedicated SQL pools in Azure Synapse provide a low maintenance method for complex analytical queries to get fast performance without any query change. Incorrect Answers: C: One daily execution does not make use of result cache caching. Note: When result set caching is enabled, dedicated SQL pool automatically caches query results in the user database for repetitive use. This allows subsequent query executions to get results directly from the persisted cache so recomputation is not needed. Result set caching improves query performance and reduces compute resource usage. In addition, queries using cached results set do not use any concurrency slots and thus do not count against existing concurrency limits. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-materialized-views https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-caching

Community vote distribution
B (97%) (3

ANath Highly Voted 4 years, 6 months ago
B is correct. Materialized view and result set caching These two features in dedicated SQL pool are used for query performance tuning. Result set caching is used for getting high concurrency and fast response from repetitive queries against static data. To use the cached result, the form of the cache requesting query must match with the query that produced the cache. In addition, the cached result must apply to the entire query. Materialized views allow data changes in the base tables. Data in materialized views can be applied to a piece of a query. This support allows the same materialized views to be used by different queries that share some computation for faster performance.
upvoted 28 times

Canary_2021 Highly Voted 4 years, 8 months ago
B is the correct answer. A materialized view is a database object that contains the results of a query. A materialized view is not simply a window on the base table. It is actually a separate object holding data in itself. So query data against a materialized view with different filters should be quick. Difference Between View and Materialized View: https://techdifferences.com/difference-between-view-and-materialized-view.html
upvoted 14 times

20b1837 Most Recent 1 year, 5 months ago
B is incorrect. A materialized view MUST contains one of the following a select that contains an aggregate function and/or a Group by clause. See the below article https://learn.microsoft.com/en-us/sql/t-sql/statements/create-materialized-view-as-select-transact-sql?toc=%2Fazure%2Fsynapse-analytics%2Fsql-data-warehouse%2Ftoc.json&bc=%2Fazure%2Fsynapse-analytics%2Fsql-data-warehouse%2Fbreadcrumb%2Ftoc.json&view=azure-sqldw-latest&preserve-view=true
upvoted 1 times

hypersam 1 year, 7 months ago
also why C is incorrect: "When cached results are used: There is an exact match between the new query and the previous query that generated the result set cache." but in our case we need additional WHERE clause https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-caching#when-cached-results-are-used
upvoted 1 times

EmnCours 1 year, 8 months ago
Correct Answer: B
upvoted 1 times

ff5037f 1 year, 9 months ago

As materalized views stores data. we can write once and can read from it mutiple times

upvoted 1 times

ELJORDAN23 2 years, 7 months ago

Got this question on my exam on january 17, B is correct

upvoted 3 times

Joanna0 2 years, 7 months ago

Materialized Views: Create materialized views that store the results of the complex SELECT queries. Materialized views are precomputed views stored as tables, and they can significantly reduce query times by avoiding the need to recompute the results every time the query is executed.

upvoted 2 times

ll94 2 years, 7 months ago

A. an ordered clustered columnstore index => this will impact the where clause so it is a valid option B. a materialized => can't be used since there is no aggregation C. result set caching => We don't know if the output query respects the limitation (10 gb) so no D. a replicated table => sizes of lookups tables not mentioned so even if it is a possible solution it's not a suggested approach

upvoted 1 times

phydev 2 years, 9 months ago

Was on my exam today (31.10.2023).

upvoted 4 times

kkk5566 2 years, 11 months ago

Materialized view

upvoted 1 times

norbitek 3 years, 7 months ago

There is no information that this query aggregates data. "SELECT list in the materialized view definition needs to meet at least one of these two criteria: The SELECT list contains an aggregate function. GROUP BY is used in the Materialized view definition and all columns in GROUP BY are included in the SELECT list. Up to 32 columns can be used in the GROUP BY clause." I'm not sure that B is correct answer. Unfortunately I cannot see better answerA properly designed materialized view provides : Reduce the execution time for complex queries with JOINs and aggregate functions. The more complex the query, the higher the potential for execution-time saving. The most benefit is gained when a query's computation cost is high and the resulting data set is small. So the right answer is B

upvoted 3 times

norbitek 3 years, 7 months ago

There is no information that this query aggregates data.

"SELECT list in the materialized view definition needs to meet at least one of these two criteria:

The SELECT list contains an aggregate function. GROUP BY is used in the Materialized view definition and all columns in GROUP BY are included in the SELECT list. Up to 32 columns can be used in the GROUP BY clause."

I'm not sure that B is correct answer. Unfortunately I cannot see better answer

upvoted 3 times

ck8.kakade 2 years ago

A properly designed materialized view provides : Reduce the execution time for complex queries with JOINs and aggregate functions. The more complex the query, the higher the potential for execution-time saving. The most benefit is gained when a query's computation cost is high and the resulting data set is small. So the right answer is B

upvoted 1 times

Deepshikha1228 4 years ago

B is correct

upvoted 2 times

SKN0865 4 years, 1 month ago

B is correct acc to: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-materialized-views
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-materialized-views

upvoted 1 times

SandipSingha 4 years, 3 months ago

B materialized view

upvoted 2 times

Egocentric 4 years, 4 months ago

B is correct without a doubt

upvoted 2 times

DingDongSingSong 4 years, 5 months ago

Why isn't the answer "A" when the query may have additional WHERE parameters depending on the report. That mean's the query isn't static and will change depending on the report. A clustered columstore index would provide a bettery query performance in case of a complex query where query isn't static.I was thinking on the same level initially but there are multiple tables informed and apply column store indexes an all tables would not ensure good results materialized view would store the results of complex calculations and it would be faster to just query those results i believe even if extra where clauses are applied

upvoted 1 times

DingDongSingSong 4 years, 5 months ago

Why isn't the answer "A" when the query may have additional WHERE parameters depending on the report. That mean's the query isn't static and will change depending on the report. A clustered columstore index would provide a bettery query performance in case of a complex query where query isn't static.

upvoted 1 times

uzairahm 4 years, 1 month ago

I was thinking on the same level initially but there are multiple tables informed and apply column store indexes an all tables would not ensure good results materialized view would store the results of complex calculations and it would be faster to just query those results i believe even if extra where clauses are applied

upvoted 3 times

Question #26 *Topic 1*

You have an Azure Synapse Analytics workspace named WS1 that contains an Apache Spark pool named Pool1.
You plan to create a database named DB1 in Pool1.
You need to ensure that when tables are created in DB1, the tables are available automatically as external tables to the built-in serverless SQL pool.
Which format should you use for the tables in DB1?

A. CSV

B. ORC

C. JSON

D. Parquet

Suggested Answer: *D*
Serverless SQL pool can automatically synchronize metadata from Apache Spark. A serverless SQL pool database will be created for each database existing in serverless Apache Spark pools. For each Spark external table based on Parquet or CSV and located in Azure Storage, an external table is created in a serverless SQL pool database. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-storage-files-spark-tables

Community vote distribution
D (100%)

KevinSames Highly Voted 4 years, 8 months ago
Both A and D are correct "For each Spark external table based on Parquet or CSV and located in Azure Storage, an external table is created in a serverless SQL pool database. As such, you can shut down your Spark pools and still query Spark external tables from serverless SQL pool."
upvoted 24 times

ZIMARAKI 3 years, 10 months ago
"For each Spark external table based on Parquet or CSV and located in Azure Storage, an external table is created in a serverless SQL pool database. As such, you can shut down your Spark pools and still query Spark external tables from serverless SQL pool." So A and D. Parquet are faster so DBut they never asked about faster, so why it cant be AIn this business, time is money.
upvoted 5 times

ZIMARAKI 3 years, 10 months ago
Selected Answer: D
"For each Spark external table based on Parquet or CSV and located in Azure Storage, an external table is created in a serverless SQL pool database. As such, you can shut down your Spark pools and still query Spark external tables from serverless SQL pool."

So A and D. Parquet are faster so D
upvoted 5 times

Jiviify 3 years, 9 months ago
But they never asked about faster, so why it cant be A
upvoted 3 times

Jiviify 3 years, 9 months ago
But they never asked about faster, so why it cant be AIn this business, time is money.
upvoted 3 times

shakes103 Highly Voted 3 years, 7 months ago
In this business, time is money.
upvoted 19 times

EmnCours Most Recent 1 year, 8 months ago
Correct Answer: D
upvoted 1 times

sergio_eduardo 2 years, 2 months ago
Both A and D, but for priority i decide D
upvoted 1 times

kkk5566 2 years, 11 months ago

Prefer to D

upvoted 2 times

ExamWiner 3 years, 2 months ago

Tipos de archivo adecuados para consultas analíticas - Si tiene cargas de trabajo basadas en Hive o Presto, vaya con ORC. - Si tiene cargas de trabajo basadas en Spark o Drill, vaya con Parquet.

upvoted 3 times

Deeksha1234 4 years ago

both A and D; D - Parquet is faster

upvoted 2 times

jainparag1 4 years, 1 month ago

Option D as the explanation suggests. Parquet is always faster than CSV being columnar data store.

upvoted 3 times

Dicer 4 years, 1 month ago

both A and D are okay, but Paraquet is faster than CSV, so answer is D.

upvoted 2 times

Rrk07 4 years, 2 months ago

Both A & D are correct , as the explanation also suggest same.

upvoted 2 times

RehanRajput 4 years, 2 months ago

Both A and D

upvoted 2 times

RehanRajput 4 years, 2 months ago

Both A and Dhttps://docs.microsoft.com/en-us/azure/synapse-analytics/metadata/database

upvoted 2 times

RehanRajput 4 years, 2 months ago

https://docs.microsoft.com/en-us/azure/synapse-analytics/metadata/database

upvoted 1 times

MatiCiri 4 years, 3 months ago

Looks correct to me

upvoted 2 times

AhmedDaffaie 4 years, 5 months ago

JSON is also supported by Serverless SQL Pool but it is kinda complicated. Why is it not selected?

upvoted 2 times

Ajitk27 4 years, 5 months ago

Looks correct to me

upvoted 2 times

VijayMore 4 years, 6 months ago

Correct

upvoted 1 times

PallaviPatel 4 years, 6 months ago

Both A and D are correct. as CSV and Parquet are correct answers.

upvoted 1 times

Mahesh_mm 4 years, 7 months ago

Parquet and CSV are correct

upvoted 4 times

Question #27 *Topic 1*

You are planning a solution to aggregate streaming data that originates in Apache Kafka and is output to Azure Data Lake Storage Gen2. The developers who will implement the stream processing solution use Java.
Which service should you recommend using to process the streaming data?

A. Azure Event Hubs

B. Azure Data Factory

C. Azure Stream Analytics

D. Azure Databricks

Suggested Answer: *D*

Capability	Azure Stream Analytics	HDInsight with Spark Streaming	Apache Spark in Azure Databricks	HDInsight with Storm
Programmability	Stream analytics query language, JavaScript	C#/F#, Java, Python, Scala	C#/F#, Java, Python, R, Scala	C#, Java

Capability	Azure Stream Analytics	HDInsight with Spark Streaming	Apache Spark in Azure Databricks	HDInsight with Storm
Inputs	Azure Event Hubs, Azure IoT Hub, Azure Blob storage	Event Hubs, IoT Hub, Kafka, HDFS, Storage Blobs, Azure Data Lake Store	Event Hubs, IoT Hub, Kafka, HDFS, Storage Blobs, Azure Data Lake Store	Event Hubs, IoT Hub, Storage Blobs, Azure Data Lake Store
Sinks	Azure Data Lake Store, Azure SQL Database, Storage Blobs, Event	HDFS, Kafka, Storage Blobs, Azure Data Lake Store, Cosmos DB	HDFS, Kafka, Storage Blobs, Azure Data Lake Store, Cosmos DB	Event Hubs, Service Bus, Kafka

Capability	Azure Stream Analytics	HDInsight with Spark Streaming	Apache Spark in Azure Databricks	HDInsight with Storm
Programmability	Stream analytics query language, JavaScript	C#/F#, Java, Python, Scala	C#/F#, Java, Python, R, Scala	C#, Java

The following tables summarize the key differences in capabilities for stream processing technologies in Azure.

General capabilities -

Integration capabilities -

Reference: https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/stream-processing

Community vote distribution



Nifl91 Highly Voted 3 years, 8 months ago

Correct!

upvoted 24 times

davidkhala Most Recent 1 year, 8 months ago

Answer list is very strange. Except for JAR deployment in Databricks Job, there is almost no space for Java as a language. Scala is another syntax need training for regular java dev to adopt

upvoted 1 times

kkk5566 1 year, 11 months ago

DataBricks with Java lang

upvoted 2 times

auwia 2 years, 2 months ago

D is correct.

upvoted 1 times

Mohitsain 2 years, 5 months ago

D is the correct one for sure.

upvoted 2 times

Fernando_Caemerer 2 years, 8 months ago

D is correct

upvoted 2 times

Aaashu 2 years, 11 months ago

correct

upvoted 2 times

Deepshikha1228 3 years ago

D is correct

upvoted 3 times

jainparag1 3 years, 1 month ago

Azure Databrics as the question is clearly asking the support for Java programming.

upvoted 3 times

NewTuanAnh 3 years, 4 months ago

why not C: Azure Stream Analytics?I see, Azure Stream Analytics does not associate with Javaor databrickskafka*Yes Azure stream Analytics for streaming data?

upvoted 1 times

NewTuanAnh 3 years, 4 months ago

why not C: Azure Stream Analytics?

upvoted 1 times

NewTuanAnh 3 years, 4 months ago

I see, Azure Stream Analytics does not associate with Javaor databrickskafka*

upvoted 2 times

NewTuanAnh 3 years, 4 months ago

I see, Azure Stream Analytics does not associate with Java

upvoted 2 times

sdokmak 3 years, 2 months ago

or databrickskafka*

upvoted 1 times

sdokmak 3 years, 2 months ago

or databricks

upvoted 1 times

sdokmak 3 years, 2 months ago

kafka*

upvoted 2 times

Muishkin 3 years, 3 months ago

Yes Azure stream Analytics for streaming data?

upvoted 1 times

PallaviPatel 3 years, 6 months ago

correct.

upvoted 2 times

Mahesh_mm 3 years, 7 months ago

Answer is correct

upvoted 3 times

alexleonvalencia 3 years, 8 months ago

Respuesta correcta Azure DataBricks.

upvoted 4 times

Question #28 *Topic 1*

You plan to implement an Azure Data Lake Storage Gen2 container that will contain CSV files. The size of the files will vary based on the number of events that occur per hour.
File sizes range from 4 KB to 5 GB.
You need to ensure that the files stored in the container are optimized for batch processing.
What should you do?

A. Convert the files to JSON

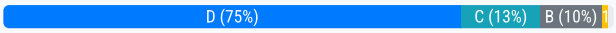
B. Convert the files to Avro

C. Compress the files

D. Merge the files

Suggested Answer: *B*
Avro supports batch and is very relevant for streaming. Note: Avro is framework developed within Apache's Hadoop project. It is a row-based storage format which is widely used as a serialization process. AVRO stores its schema in JSON format making it easy to read and interpret by any program. The data itself is stored in binary format by doing it compact and efficient. Reference:
https://www.adaltas.com/en/2020/07/23/benchmark-study-of-different-file-format/

Community vote distribution



VeroDon Highly Voted 1 year, 11 months ago
You can not merge the files if u don't know how many files exist in ADLS2. In this case, you could easily create a file larger than 100 GB in size and decrease performance. so B is the correct answer. Convert to AVRO
upvoted 51 times

VeroDon Highly Voted 1 year, 11 months ago
You can not merge the files if u don't know how many files exist in ADLS2. In this case, you could easily create a file larger than 100 GB in size and decrease performance. so B is the correct answer. Convert to AVROI can understand why you say not merge, but why avro?Because we need to ensure files stored in the container are optimized for batch processing. converting the files to AVRO would be suitable for optimized for batch processing. So, the answer is "Convert to AVRO"The information about the file size is already given which is between 5KB to 5GB. So option D seems to be correct.Option B: Convert the files to Avro (WRONG FOR ME) While converting the files to Avro is a valid option for optimizing data storage and processing, it may not be the most suitable choice in this specific scenario. Avro is a binary serialization format that is efficient for compact storage and fast data processing. It provides schema evolution support and is widely used in big data processing frameworks like Apache Hadoop and Apache Spark. However, in the given scenario, the files are already in CSV format. Converting them to Avro would require additional processing and potentially introduce complexity. Avro is better suited for scenarios where data is generated or consumed by systems that natively support Avro or for cases where schema evolution is a critical requirement. On the other hand, merging the files (Option D) is a more straightforward and common approach to optimize batch processing. It helps reduce the overhead associated with managing a large number of small files, improves data scanning efficiency, and enhances overall processing performance. Merging files is a recommended practice to achieve better performance and cost efficiency in scenarios where file sizes vary.hi , please clarify if we can conclude this as convert to avro as the right option? thank you very much!
upvoted 51 times

Massy 4 years, 3 months ago
I can understand why you say not merge, but why avro?
upvoted 4 times

Massy 4 years, 3 months ago
I can understand why you say not merge, but why avro?Because we need to ensure files stored in the container are optimized for batch processing. converting the files to AVRO would be suitable for optimized for batch processing. So, the answer is "Convert to AVRO"
upvoted 4 times

anks84 3 years, 11 months ago
Because we need to ensure files stored in the container are optimized for batch processing. converting the files to AVRO would be suitable for optimized for batch processing. So, the answer is "Convert to AVRO"
upvoted 2 times

bhrz 3 years, 11 months ago

The information about the file size is already given which is between 5KB to 5GB. So option D seems to be correct.

upvoted 3 times

auwia 3 years, 2 months ago

Option B: Convert the files to Avro (WRONG FOR ME) While converting the files to Avro is a valid option for optimizing data storage and processing, it may not be the most suitable choice in this specific scenario. Avro is a binary serialization format that is efficient for compact storage and fast data processing. It provides schema evolution support and is widely used in big data processing frameworks like Apache Hadoop and Apache Spark. However, in the given scenario, the files are already in CSV format. Converting them to Avro would require additional processing and potentially introduce complexity. Avro is better suited for scenarios where data is generated or consumed by systems that natively support Avro or for cases where schema evolution is a critical requirement. On the other hand, merging the files (Option D) is a more straightforward and common approach to optimize batch processing. It helps reduce the overhead associated with managing a large number of small files, improves data scanning efficiency, and enhances overall processing performance. Merging files is a recommended practice to achieve better performance and cost efficiency in scenarios where file sizes vary.

upvoted 17 times

SenMia 2 years, 8 months ago

hi , please clarify if we can conclude this as convert to avro as the right option? thank you very much!

upvoted 1 times

Canary_2021 1 year, 11 months ago

If you store your data as many small files, this can negatively affect performance. In general, organize your data into larger sized files for better performance (256 MB to 100 GB in size). https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices#optimize-for-data-ingest

upvoted 27 times

Jolyboy Most Recent 1 year, 4 months ago

I choose B, To optimize the files stored in your Azure Data Lake Storage Gen2 container for batch processing, the best option is to convert the files to Avro (Option B). Here's why: Avro is a binary format that is highly efficient for both storage and processing. It supports schema evolution, which is beneficial for handling changes in data structure over time1. Compression: Avro files can be compressed, reducing storage costs and improving read/write performance. Splittable: Avro files are splittable, which means they can be processed in parallel, enhancing batch processing efficiency. While compressing the files (Option C) can also help reduce storage size and improve transfer speeds, converting to Avro provides additional benefits like schema support and better performance for large-scale data processing.

upvoted 1 times

Januaz 1 year, 5 months ago

Copilot says Nerge the files

upvoted 1 times

IMadnan 1 year, 6 months ago

merging the files (Option D)

upvoted 1 times

samianae 1 year, 6 months ago

so Confusing for me C and D are seem correct for me

upvoted 1 times

RMK2000 1 year, 7 months ago

Larger files lead to better performance and reduced costs. Typically, analytics engines such as HDInsight have a per-file overhead that involves tasks such as listing, checking access, and performing various metadata operations. If you store your data as many small files, this can negatively affect performance. In general, organize your data into larger sized files for better performance (256 MB to 100 GB in size). Some engines and applications might have trouble efficiently processing files that are greater than 100 GB in size. Increasing file size can also reduce transaction costs. Read and write operations are billed in 4 megabyte increments so you're charged for operation whether or not the file contains 4 megabytes or only a few kilobytes. For pricing information, see Azure Data Lake Storage pricing. https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices#file-size

upvoted 1 times

moize 1 year, 8 months ago

Pour optimiser les fichiers CSV stockés dans Azure Data Lake Storage Gen2 pour le traitement par lots, la meilleure option est de fusionner les fichiers (option D). Fusionner les fichiers permet de réduire le nombre de petits fichiers, ce qui améliore les performances de traitement par lots en réduisant la surcharge liée à la gestion de nombreux petits fichiers

upvoted 1 times

EmnCours 1 year, 8 months ago

Correct Answer: D

upvoted 1 times

roopansh.gupta2 1 year, 11 months ago

D. Merge the files Explanation: File Merging: Merging small files into larger ones helps to reduce the overhead associated with processing many small files, which can negatively impact performance. Large files are typically more efficient for batch processing in distributed systems like Azure Data Lake and Azure Databricks because they minimize the number of tasks and the amount of metadata the system needs to manage. Compressing the Files (Option C Incorrect): While compression can reduce storage costs and improve I/O efficiency, it doesn't address the issue of small file sizes, which can still lead to inefficiencies in distributed processing. Converting to Avro (Option B Incorrect) or JSON (Option A Incorrect): Converting to a different file format like Avro or JSON could be beneficial for specific use cases, especially where schema evolution or specific query optimizations are required. However, this does not address the fundamental issue of optimizing file size for batch processing. Therefore, merging the files (Option D) is the most direct and effective way to optimize for batch processing in this scenario.

upvoted 1 times

d39f475 2 years, 2 months ago

Compress the files is my choise

upvoted 1 times

Dusica 2 years, 3 months ago

pay attention to the "number of events that occur per hour." That means it is streaming and you can group small files with streaming engines - it is D

upvoted 1 times

f214eb2 2 years, 4 months ago

Convert to AVRO is Legit

upvoted 1 times

Charley92 2 years, 4 months ago

C. Compress the files Compression is beneficial for batch processing as it can significantly reduce the file size, which leads to faster transfer rates and can improve performance during batch processing tasks. It’s particularly effective for large files, making it easier to handle and process them efficiently

upvoted 1 times

da257c2 2 years, 4 months ago

option d

upvoted 1 times

Khadija10 2 years, 6 months ago

The answer is compress the files - we can't merge because we don't know how many files we will receive, and the performance will be decreased if the size of the merged files is too large. - converting the files to Avro will require additional processing and if we receive too many files it may cause complexity. - compressing the CSV files is the best choice in our scenario, compressing the files is a common practice for optimizing batch processing. It helps reduce storage space, minimize data transfer times, and improve overall performance.

upvoted 3 times

alphilla 2 years, 6 months ago

Crazy how OPTION D (which is wrong) has such a high ammount of votes. I'm going for C.compress files. Guys you don't know how many files you have, you can't design a system where this kind of randomness can make your system to fail.

upvoted 5 times

HOTSPOT -

You store files in an Azure Data Lake Storage Gen2 container. The container has the storage policy shown in the following exhibit.

Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.

NOTE: Each correct selection is worth one point.

Hot Area:

```
{
   "rules": [
     {
       "enabled": true,
       "name": "contosorule",
       "type": "Lifecycle",
       "definition": {
         "actions": {
           "version": {
             "delete": {
               "daysAfterCreationGreaterThan": 60
             }
           },
           "baseBlob": {
             "tierToCool": {
               "daysAfterModificationGreaterThan":
30
             },
             },
           }
         },
         "filters": {
           "blobTypes": [
             "blockBlob"
           ],
           "prefixMatch": [
             "container1/contoso"
           ]
         }
       }
     }
   ]
}
```

Answer Area

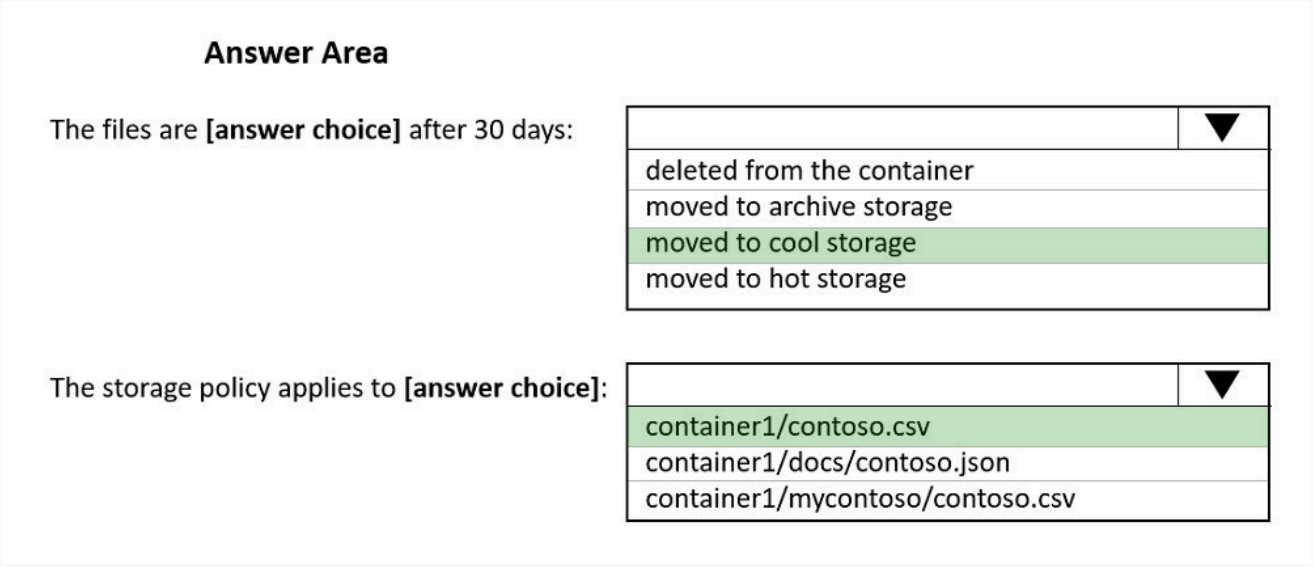
The files are **[answer choice]** after 30 days:

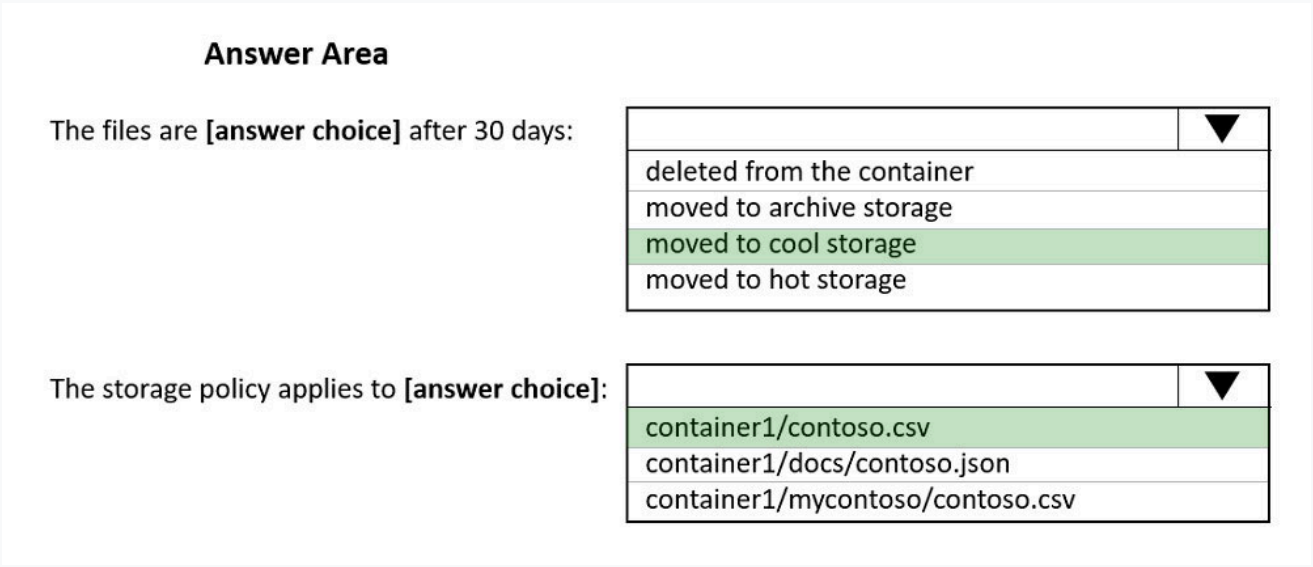
- deleted from the container
- moved to archive storage
- moved to cool storage
- moved to hot storage

The storage policy applies to **[answer choice]**:

- container1/contoso.csv
- container1/docs/contoso.json
- container1/mycontoso/contoso.csv

Suggested Answer:





Box 1: moved to cool storage - The ManagementPolicyBaseBlob.TierToCool property gets or sets the function to tier blobs to cool storage. Support blobs currently at Hot tier.

Box 2: container1/contoso.csv - As defined by prefixMatch. prefixMatch: An array of strings for prefixes to be matched. Each rule can define up to 10 case-senstive prefixes. A prefix string must start with a container name. Reference: https://docs.microsoft.com/en-us/dotnet/api/microsoft.azure.management.storage.fluent.models.managementpolicybaseblob.tiertocool

bad_atitude Highly Voted 3 years, 8 months ago

correctwhy the .csv?It matches anything that starts with "container1/contoso" and the csv in the answer is the only one that matches.

upvoted 42 times

bad_atitude Highly Voted 3 years, 8 months ago

correct

upvoted 42 times

adfgasd 3 years, 8 months ago

why the .csv?It matches anything that starts with "container1/contoso" and the csv in the answer is the only one that matches.

upvoted 3 times

adfgasd 3 years, 8 months ago

why the .csv?

upvoted 3 times

Lewistrick 3 years, 8 months ago

It matches anything that starts with "container1/contoso" and the csv in the answer is the only one that matches.

upvoted 19 times

alexleonvalencia 3 years, 8 months ago

Respuesta Cool Tier & Container1/contoso.csv

upvoted 8 times

kwokeric97 1 year, 8 months ago

The tireToCool values is empty, should this rule being skipped? Should the data keep in hot storage at day 30, and being deleted at day 60?

"tierToCool": { "daysAfterModificationGreaterThan":

https://stackoverflow.com/questions/62368999/skip-rule-creation-if-parameter-is-empty-in-azure-storage-life-cycle-management

upvoted 1 times

kwokeric97 1 year, 8 months ago

The tireToCool values is empty, should this rule being skipped? Should the data keep in hot storage at day 30, and being deleted at day 60? "tierToCool": { "daysAfterModificationGreaterThan": https://stackoverflow.com/questions/62368999/skip-rule-creation-if-parameter-is-empty-in-azure-storage-life-cycle-managementIt's not empty: the value "30" is on the next line

upvoted 1 times

data_guy Most Recent 1 year, 8 months ago

It's not empty: the value "30" is on the next line

upvoted 2 times

kkk5566 1 year, 11 months ago

Move to Cool Tier & Container1/contoso.csv

upvoted 3 times

darth_vader_007 2 years, 3 months ago

Move to cold storage is the right answer, but the statement, has to be "files which are idle for more than 30 days have to be moved to a cold storage", If the files get modified before the 30 day period, say weekly, it remains in the hot storage tier.

upvoted 1 times

tembal 2 years, 8 months ago

ah yes move to Cool storage and Container1/contoso.csv because prefixmatch is Container1/contoso

upvoted 1 times

Deeksha1234 3 years ago

Cool Tier & Container1/contoso.csv

upvoted 2 times

Deepshikha1228 3 years ago

given answer is correct

upvoted 2 times

azure900test 3 years, 1 month ago

I think the responses do not match the question. There is no policy for cold storage here (it only says delete after 60 days) and as far as I know there is no such thing as a default duration for moving things to the cold storage if lifecycle is enabled https://docs.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-policy-configure?tabs=azure-portalsorry overlooked the cold tier policy, please ignore

upvoted 1 times

azure900test 3 years, 1 month ago

I think the responses do not match the question. There is no policy for cold storage here (it only says delete after 60 days) and as far as I know there is no such thing as a default duration for moving things to the cold storage if lifecycle is enabled https://docs.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-policy-configure?tabs=azure-portal

upvoted 1 times

azure900test 3 years, 1 month ago

sorry overlooked the cold tier policy, please ignore

upvoted 3 times

AJ01 3 years, 7 months ago

shouldn't the question be greater than 60 days?

upvoted 2 times

AJ01 3 years, 7 months ago

shouldn't the question be greater than 60 days?The files get deleted after 60 days but after 30 days they are moved to the cool storage.

upvoted 2 times

stunner85_ 3 years, 6 months ago

The files get deleted after 60 days but after 30 days they are moved to the cool storage.

upvoted 4 times

Mahesh_mm 3 years, 7 months ago

correct

upvoted 3 times

Question #30 *Topic 1*

You are designing a financial transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:

✑ TransactionType: 40 million rows per transaction type

✑ CustomerSegment: 4 million per customer segment

✑ TransactionMonth: 65 million rows per month

AccountType: 500 million per account type

You have the following query requirements:

✑ Analysts will most commonly analyze transactions for a given month.

✑ Transactions analysis will typically summarize transactions by transaction type, customer segment, and/or account type

You need to recommend a partition strategy for the table to minimize query times.

On which column should you recommend partitioning the table?

.

A. CustomerSegment

B. AccountType

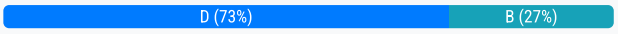
C. TransactionType

D. TransactionMonth

Suggested Answer: *D*

For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases. Example: Any partitioning added to a table is in addition to the distributions created behind the scenes. Using this example, if the sales fact table contained 36 monthly partitions, and given that a dedicated SQL pool has 60 distributions, then the sales fact table should contain 60 million rows per month, or 2.1 billion rows when all months are populated. If a table contains fewer than the recommended minimum number of rows per partition, consider using fewer partitions in order to increase the number of rows per partition.

Community vote distribution



SinSS Highly Voted 1 year, 11 months ago

Considering row count, the answer is Transaction Month or Account Type. Partitioning with Trans. Month, it will end up with hot partition and cannot utilize parallel processing, because all data for the query will be in the same partition. To minimize the processing time, parallel processing should happen. So Transaction Month can not be the answer. My answer is Account Type.

There is a similar question in Microsoft site and the explanation was given with answer as follows. Product ensures parallelism when querying data from a given month within the same region, or multiple regions. Using date and partitioning by month, all sales for a month will be in the same partition, not providing parallelism. All sales for a given region will be in the same partition, not providing parallelism....

upvoted 9 times

SinSS Highly Voted 1 year, 11 months ago

Considering row count, the answer is Transaction Month or Account Type. Partitioning with Trans. Month, it will end up with hot partition and cannot utilize parallel processing, because all data for the query will be in the same partition. To minimize the processing time, parallel processing should happen. So Transaction Month can not be the answer. My answer is Account Type. There is a similar question in Microsoft site and the explanation was given with answer as follows. Product ensures parallelism when querying data from a given month within the same region, or multiple regions. Using date and partitioning by month, all sales for a month will be in the same partition, not providing parallelism. All sales for a given region will be in the same partition, not providing parallelism....I think you are confusing partition with distribution. Each partition will still be distributed over 60 distributions.The Microsoft question is also about partitioning.

upvoted 9 times

Rob77 3 years, 2 months ago

I think you are confusing partition with distribution. Each partition will still be distributed over 60 distributions.

upvoted 3 times

Rob77 3 years, 2 months ago

I think you are confusing partition with distribution. Each partition will still be distributed over 60 distributions.The Microsoft question is also about partitioning.

upvoted 3 times

semauni 3 years ago

The Microsoft question is also about partitioning.

upvoted 1 times

gf2tw 4 years, 8 months ago

Agree with D, should not be confused with Distribution column for Hash-distributed tables.

upvoted 7 times

gf2tw 4 years, 8 months ago

Agree with D, should not be confused with Distribution column for Hash-distributed tables.Partitioning and Distributing are different concepts In most cases data is partitioned on a date column that is closely tied to the order in which the data is loaded into the SQL pool.
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition

upvoted 7 times

emna2022 4 years ago

Partitioning and Distributing are different concepts In most cases data is partitioned on a date column that is closely tied to the order in which the data is loaded into the SQL pool. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition

upvoted 4 times

Jolyboy Most Recent 1 year, 4 months ago

To minimize query times for your financial transactions table in Azure Synapse Analytics, you should recommend partitioning the table on the TransactionMonth column (Option D). Here's why: Query Requirements: Analysts will most commonly analyze transactions for a given month. Partitioning by TransactionMonth allows queries to quickly filter and access the relevant partitions, significantly improving query performance1. Data Distribution: With 65 million rows per month, partitioning by TransactionMonth ensures that each partition contains a manageable amount of data, optimizing both storage and query efficiency1. Partitioning by TransactionMonth aligns well with your query patterns and helps in efficiently managing and querying large datasets.

upvoted 1 times

s1852 1 year, 7 months ago

For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

upvoted 1 times

EmnCours 1 year, 8 months ago

Correct Answer: D

upvoted 1 times

ibardonr 1 year, 9 months ago

D. xD

upvoted 1 times

dgerok 1 year, 11 months ago

Selected Answer: B

While you choose month, there is no parallelization at all, so this is the worst possible answer. The correct answer is AccountType. This ensures the PARALELLIZATION, which is improving the performance. CustomerSegment and TransactionType are below 60 million, which is not enough for 60 nodes with 1 mln for each...

upvoted 4 times

dgerok 1 year, 11 months ago

While you choose month, there is no parallelization at all, so this is the worst possible answer. The correct answer is AccountType. This ensures the PARALELLIZATION, which is improving the performance. CustomerSegment and TransactionType are below 60 million, which is not enough for 60 nodes with 1 mln for each...Parallelization is for distributions I believe, and every partition will automatically be distributed as well. Partitioning is done to reduce the amount of data that is scanned. If I partition the data by months, then each month partition is further distributed into 60 distributions which are split across different nodes, thereby ensuring parallelism

upvoted 4 times

Ahan 2 years, 3 months ago

Parallelization is for distributions I believe, and every partition will automatically be distributed as well. Partitioning is done to reduce the amount of data that is scanned. If I partition the data by months, then each month partition is further distributed into 60 distributions which are split across different nodes, thereby ensuring parallelism

upvoted 5 times

otapi 1 year, 11 months ago

Selected Answer: B

I assume, the listed figures are the number of unique values in each columns. TransactionMonth is a bad partition, as queries would typically fall into one partition. AccountType has the most unique values, those are better for partitioning.

upvoted 1 times

otapi 1 year, 11 months ago

I assume, the listed figures are the number of unique values in each columns. TransactionMonth is a bad partition, as queries would typically fall into one partition. AccountType has the most unique values, those are better for partitioning.Don't worry about the listed figures. They are not specific to the question, but is copied text from an example provided on MS-Docs.

upvoted 1 times

roopansh.gupta2 1 year, 12 months ago

Don't worry about the listed figures. They are not specific to the question, but is copied text from an example provided on MS-Docs.

upvoted 1 times

jppdks 2 years, 4 months ago

The best column to partition the table on would be D. TransactionMonth. Partitioning is a database design technique which is used to improves performance, manageability, simplifies maintenance and reduce the response time of the query. Given that analysts will most commonly analyze transactions for a given month, partitioning on the TransactionMonth column would allow the database to quickly isolate the relevant rows for a given month, thereby minimizing query times. This is particularly beneficial when dealing with large volumes of data, as is the case here. Moreover, partitioning on TransactionMonth would also align with the typical transaction analysis patterns, which summarize transactions by transaction type, customer segment, and/or account type. This would allow for efficient querying and analysis of the data. Therefore, the TransactionMonth column would be the most suitable choice for partitioning the table.

upvoted 2 times

Khadija10 2 years, 6 months ago

Selected Answer: D

the answer is transaction month "In most cases data is partitioned on a date column that is closely tied to the order in which the data is loaded into the SQL pool." https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition?context=%2Fazure%2Fsynapse-analytics%2Fcontext%2Fcontext https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

upvoted 1 times

Khadija10 2 years, 6 months ago

the answer is transaction month "In most cases data is partitioned on a date column that is closely tied to the order in which the data is loaded into the SQL pool." https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition?context=%2Fazure%2Fsynapse-analytics%2Fcontext%2Fcontext https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-poolHello. I have a question. What is difference between partitioning and distribution?

upvoted 1 times

Bakhtiyor 2 years, 5 months ago

Hello. I have a question. What is difference between partitioning and distribution?

upvoted 1 times

DooperMan 2 years, 9 months ago

i felt partitioning then going with something that is not unique, if it is like hashing or even distribution depends on the case to go with more unique.

upvoted 2 times

ellala 2 years, 10 months ago

Partitioning by date will lead to hot partitions since the most common query is by date. Therefore this is not a good idea. Next best option is AccountType. We want to take advantage of parallel processing to improve efficiency

upvoted 2 times

mav2000 2 years, 11 months ago

Answer is D, because the analyst will be querying transactions for month, and then its mentioned that transaction analysis will be done on Transaction_type, customer_segment and account_type, meaning they won't be querying for an individual columns but all 3 at the same time, which means it's pointless to partition between these columns, so transaction month is the answer

upvoted 3 times

hassexat 2 years, 11 months ago

Clustered columnstore index has 60 partitions and each partition needs a minimum of 1 million rows so TransactionMonth is the only that has more than 60 millions of rows. Correct answer: D

upvoted 2 times

kkk5566 2 years, 11 months ago

Partition by month is a good idra

upvoted 1 times

akhil5432 3 years ago

transaction month

upvoted 1 times

Shanks111 3 years, 2 months ago

B, as selecting the date column as a partition will make one partition a hot partition and won't be able to make use of parallel processing, so Account Type should be the correct answer.

upvoted 4 times

HOTSPOT -

You have an Azure Data Lake Storage Gen2 account named account1 that stores logs as shown in the following table.

You do not expect that the logs will be accessed during the retention periods.

You need to recommend a solution for account1 that meets the following requirements:

✑ Automatically deletes the logs at the end of each retention period

✑ Minimizes storage costs

What should you include in the recommendation? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Type	Designated retention period
Application	360 days
Infrastructure	60 days

Answer Area

To minimize storage costs:

- Store the infrastructure logs and the application logs in the Archive access tier
- Store the infrastructure logs and the application logs in the Cool access tier
- Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier

To delete logs automatically:

- Azure Data Factory pipelines
- Azure Blob storage lifecycle management rules
- Immutable Azure Blob storage time-based retention policies

Suggested Answer:

Answer Area

To minimize storage costs:

- Store the infrastructure logs and the application logs in the Archive access tier
- Store the infrastructure logs and the application logs in the Cool access tier
- Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier

To delete logs automatically:

- Azure Data Factory pipelines
- Azure Blob storage lifecycle management rules
- Immutable Azure Blob storage time-based retention policies

Answer Area

To minimize storage costs:

- Store the infrastructure logs and the application logs in the Archive access tier
- Store the infrastructure logs and the application logs in the Cool access tier
- Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier

To delete logs automatically:

- Azure Data Factory pipelines
- Azure Blob storage lifecycle management rules
- Immutable Azure Blob storage time-based retention policies

Box 1: Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier For infrastructure logs: Cool tier - An online tier optimized for storing data that is infrequently accessed or modified. Data in the cool tier should be stored for a minimum of 30 days. The cool tier has lower storage costs and higher access costs compared to the hot tier. For application logs: Archive tier - An offline tier optimized for storing data that is rarely accessed, and that has flexible latency requirements, on the order of hours. Data in the archive tier should be stored for a minimum of 180 days. Box 2: Azure Blob storage lifecycle management rules Blob storage lifecycle management offers a rule-based policy that you can use to transition your data to the desired access tier when your specified conditions are met. You can also use lifecycle management to expire data at the end of its life. Reference: https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview

gf2tw Highly Voted 4 years, 2 months ago

"Data must remain in the Archive tier for at least 180 days or be subject to an early deletion charge. For example, if a blob is moved to the Archive tier and then deleted or moved to the Hot tier after 45 days, you'll be charged an early deletion fee equivalent to 135 (180 minus 45) days of storing that blob in the Archive tier." <- from the sourced link.

This explains why we have to use two different access tiers rather than both as archive.

upvoted 92 times

gf2tw Highly Voted 4 years, 2 months ago

"Data must remain in the Archive tier for at least 180 days or be subject to an early deletion charge. For example, if a blob is moved to the Archive tier and then deleted or moved to the Hot tier after 45 days, you'll be charged an early deletion fee equivalent to 135 (180 minus 45) days of storing that blob in the Archive tier." <- from the sourced link. This explains why we have to use two different access tiers rather than both as archive.Thanks a ton for explaining.Thanks for sharing this info."You do not expect that the logs will be accessed during the retention periods." - including deletes, i suppose. you just let lifecycle management rule do the deletes after the retention period ... Archiving for cost-reduction?

upvoted 92 times

dsp17 3 years, 7 months ago

Thanks a ton for explaining.

upvoted 2 times

Anshuman_B 3 years ago

Thanks for sharing this info.

upvoted 1 times

RoyP654 2 years, 8 months ago

"You do not expect that the logs will be accessed during the retention periods." - including deletes, i suppose. you just let lifecycle management rule do the deletes after the retention period ... Archiving for cost-reduction?

upvoted 4 times

ANath 4 years, 1 month ago

The answers are correct. Data must remain in the Archive tier for at least 180 days or be subject to an early deletion charge. For example, if a blob is moved to the Archive tier and then deleted or moved to the Hot tier after 45 days, you'll be charged an early deletion fee equivalent to 135 (180 minus 45) days of storing that blob in the Archive tier. A blob in the Cool tier in a general-purpose v2 accounts is subject to an early deletion penalty if it is deleted or moved to a different tier before 30 days has elapsed. This charge is prorated. For example, if a blob is moved to the Cool tier and then deleted after 21 days, you'll be charged an early deletion fee equivalent to 9 (30 minus 21) days of storing that blob in the Cool tier.
https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview

upvoted 20 times

d39f475 Most Recent 1 year, 8 months ago

configuring a trigger in a pipeline to delete logs when they reach their expiration date is a valid approach to automatically manage log retention in Azure.

upvoted 1 times

Alongi 1 year, 11 months ago

Correct answers

upvoted 1 times

kkk5566 2 years, 5 months ago

correct

upvoted 1 times

Rashmi007 2 years, 8 months ago

Given answer is correct.

upvoted 3 times

maochi 2 years, 8 months ago

The answers are correct

upvoted 1 times

mamahani 2 years, 9 months ago

Infrastructure: archive access tier (and pay penalty for early deletion) / application: archive access tier ; Azure Blob storage lifecycle management rules

upvoted 2 times

mamahani 2 years, 10 months ago

as per microsoft price lists: https://azure.microsoft.com/en-us/pricing/details/storage/blobs/ to store 50 terabytes in cool storage is around 500 usd/month to store 50 terabytes in archive storage is around 50 usd/month the retention for infrastructure logs is 60 days; files must remain in archive for minimum of 180 otherwise early delete penalty is to be paid; oki doki, 180 days minus 60 days, is 120 days; so around 4 month; 4*50 usd = 200 usd; its still WAY cheaper than cool storage; and the logs will not be accessed during retention (and even if they are , in archive they can be also retrieved, who cares if it takes hours); id go for archive for both

upvoted 2 times

NeerajGarg 2 years, 12 months ago

As per my calculations of storing the data in both the tiers and early deletion penalty, the archive storage (for both) is much cheaper than cool tier and archive tier. Assumption is that the data is not accessed from archive tier. The reference link for calculations : https://azure.microsoft.com/en-us/pricing/details/storage/blobs/

upvoted 1 times

SHENOOOO 3 years ago

Given Answers are correct

upvoted 2 times

zilvakas 3 years, 1 month ago

While I understand the reason for such answer, I still have doubts. If we think about costs of storage then answer will be different. E.g. (west europe region, all costs for data retrevial are removed) Cool storage for 1 month for 1TB costs 11.10 Eur for 1 month (that is 11,10x6=66,6 for 180 days) and archive storage costs 3,41 Eur per month (that is 3,41x6=20,46 Eur for 6 months). Therefore Archive tear is cheeper even you need to keep the data for 180 days instead of 30. I would go for Archive tier in both cases.

upvoted 3 times

zilvakas 3 years, 1 month ago

While I understand the reason for such answer, I still have doubts. If we think about costs of storage then answer will be different. E.g. (west europe region, all costs for data retrevial are removed) Cool storage for 1 month for 1TB costs 11.10 Eur for 1 month (that is 11,10x6=66,6 for 180 days) and

archive storage costs 3,41 Eur per month (that is 3,41x6=20,46 Eur for 6 months). Therefore Archive tear is cheeper even you need to keep the data for 180 days instead of 30. I would go for Archive tier in both cases.Ok, my comment is wrong because you MUST remove data at the end of retention period.Your comment is still valid though. You can delete data in the Archive tier in 30 days in subject to an early deletion penalty. But even if you got an early deletion penalty, the total cost of the Archive tier is still cheaper than that of the Cool tier

upvoted 3 times

zilvakas 3 years, 1 month ago

Ok, my comment is wrong because you MUST remove data at the end of retention period.

upvoted 2 times

patvn 3 years ago

Your comment is still valid though. You can delete data in the Archive tier in 30 days in subject to an early deletion penalty. But even if you got an early deletion penalty, the total cost of the Archive tier is still cheaper than that of the Cool tier

upvoted 1 times

Deeksha1234 3 years, 6 months ago

given answer is correct

upvoted 2 times

georgiakon 3 years, 8 months ago

I think that infrastructure logs should be Cool tier due to the 60 days retention period. At the Archive tier they will be early deletion charge (they should remain at least 180 days in Archive), as mentioned in the answer's provided link.

upvoted 1 times

Backy 3 years, 9 months ago

The question says "You do not expect that the logs will be accessed during the retention periods" - so there is no reason to keep any of them as Cool, so the correct answer should be to put them both in Archive

upvoted 4 times

Backy 3 years, 9 months ago

The question says "You do not expect that the logs will be accessed during the retention periods" - so there is no reason to keep any of them as Cool, so the correct answer should be to put them both in Archiveyeah but because the infrastructure logs are <180 days before deleting, there is a considerable fee to delete if in archive, so not the cheapest option.

upvoted 4 times

sdokmak 3 years, 8 months ago

yeah but because the infrastructure logs are <180 days before deleting, there is a considerable fee to delete if in archive, so not the cheapest option.

upvoted 6 times

Muishkin 3 years, 9 months ago

But the question says 360 days and 60 days for the 2 logs...whereas archive tier could store only upto 180 days .Also the cool tier has lesser storage cost /- hour as compared to archive tier.So should'nt the answer be cool tier for both?

upvoted 1 times

Muishkin 3 years, 9 months ago

But the question says 360 days and 60 days for the 2 logs...whereas archive tier could store only upto 180 days .Also the cool tier has lesser storage cost /- hour as compared to archive tier.So should'nt the answer be cool tier for both?Data in the archive tier should be stored for a minimum of 180 days - not a maximum.

upvoted 1 times

JNunn 3 years, 1 month ago

Data in the archive tier should be stored for a minimum of 180 days - not a maximum.

upvoted 1 times

Mahesh_mm 4 years, 1 month ago

Answers are correct

upvoted 2 times

Question #32 *Topic 1*

You plan to ingest streaming social media data by using Azure Stream Analytics. The data will be stored in files in Azure Data Lake Storage, and then consumed by using Azure Databricks and PolyBase in Azure Synapse Analytics.
You need to recommend a Stream Analytics data output format to ensure that the queries from Databricks and PolyBase against the files encounter the fewest possible errors. The solution must ensure that the files can be queried quickly and that the data type information is retained.
What should you recommend?

A. JSON

B. Parquet

C. CSV

D. Avro

Suggested Answer: *B*
Need Parquet to support both Databricks and PolyBase. Reference: https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-file-format-transact-sql

Community vote distribution
B (100%)

demirsamuel Highly Voted 4 years, 2 months ago
Avro schema definitions are JSON records. Polybase does not support JSON so why supporting Avro then. A CSV does not contain the schema as it is everything marked as string. so only parquet is left to choose.
upvoted 34 times

hrastogi7 Highly Voted 4 years, 8 months ago
Parquet can be quickly retrieved and maintain metadata in itself. Hence Parquet is correct answer.
upvoted 23 times

EmnCours Most Recent 1 year, 8 months ago
Selected Answer: B
upvoted 1 times

kkk5566 2 years, 11 months ago
should be correct
upvoted 1 times

akhil5432 3 years ago
Parquet
upvoted 1 times

Deeksha1234 4 years ago
Parquet is correct
upvoted 3 times

Rrk07 4 years, 2 months ago
Parquet is correct
upvoted 2 times

Muishkin 4 years, 3 months ago
Isnt JSON good for batch processing/streaming?Indeed. However, we also want to query the data using PolyBase. Polybase doesn't support Avro. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/load-data-overview#polybase-external-file-formats
upvoted 1 times

Muishkin 4 years, 3 months ago
Isnt JSON good for batch processing/streaming?
upvoted 1 times

RehanRajput 4 years, 2 months ago
Indeed. However, we also want to query the data using PolyBase. Polybase doesn't support Avro. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/load-data-overview#polybase-external-file-formats
upvoted 6 times

AhmedDaffaie 4 years, 4 months ago

I am confused! Avro has self-describing schema and good for quick loading (patching), why parquet?Apparently, the deciding factor is the fact that PolyBase doesn't support AVRO, but it does support Parquet."Polybase currently supports only delimeted text, rcfile, orc and parquet formats." R: https://msdn.microsoft.com/en-us/library/dn935025.aspx

upvoted 5 times

AhmedDaffaie 4 years, 4 months ago

I am confused! Avro has self-describing schema and good for quick loading (patching), why parquet?

upvoted 5 times

Boompiee 4 years, 3 months ago

Apparently, the deciding factor is the fact that PolyBase doesn't support AVRO, but it does support Parquet."Polybase currently supports only delimeted text, rcfile, orc and parquet formats." R: https://msdn.microsoft.com/en-us/library/dn935025.aspx

upvoted 7 times

Boompiee 4 years, 3 months ago

Apparently, the deciding factor is the fact that PolyBase doesn't support AVRO, but it does support Parquet.

upvoted 7 times

matiandal 2 years, 9 months ago

"Polybase currently supports only delimeted text, rcfile, orc and parquet formats." R: https://msdn.microsoft.com/en-us/library/dn935025.aspx

upvoted 1 times

PallaviPatel 4 years, 6 months ago

correct.

upvoted 1 times

EmmettBrown 4 years, 6 months ago

Parquet is the correct answer

upvoted 1 times

alexleonvalencia 4 years, 8 months ago

Respuesta correcta PARQUET

upvoted 1 times

Question #33 *Topic 1*

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a partitioned fact table named dbo.Sales and a staging table named stg.Sales that has the matching table and partition definitions.
You need to overwrite the content of the first partition in dbo.Sales with the content of the same partition in stg.Sales. The solution must minimize load times.
What should you do?

A. Insert the data from stg.Sales into dbo.Sales.

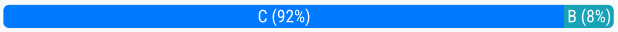
B. Switch the first partition from dbo.Sales to stg.Sales.

C. Switch the first partition from stg.Sales to dbo.Sales.

D. Update dbo.Sales from stg.Sales.

Suggested Answer: *B*
A way to eliminate rollbacks is to use Metadata Only operations like partition switching for data management. For example, rather than execute a DELETE statement to delete all rows in a table where the order_date was in October of 2001, you could partition your data monthly. Then you can switch out the partition with data for an empty partition from another table Note: Syntax: SWITCH [PARTITION source_partition_number_expression] TO [schema_name.] target_table [PARTITION target_partition_number_expression] Switches a block of data in one of the following ways: ➪ Reassigns all data of a table as a partition to an already-existing partitioned table. ➪ Switches a partition from one partitioned table to another. ➪ Reassigns all data in one partition of a partitioned table to an existing non-partitioned table.
Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

Community vote distribution



Aslam208 Highly Voted 4 years, 8 months ago
The correct answer is C
upvoted 73 times

Nifl91 Highly Voted 4 years, 8 months ago
this must be C. since the need is to overwrite dbo.Sales with the content of stg.Sales. SWITCH source TO target
upvoted 36 times

moize Most Recent 1 year, 8 months ago
Pour remplacer le contenu de la première partition dans dbo.Sales par le contenu de la même partition dans stg.Sales tout en minimisant les temps de chargement, vous devez basculer la première partition de stg.Sales vers dbo.Sales (C). Cette méthode est efficace car elle utilise le basculement de partition, qui est une opération rapide et minimise les temps d'arrêt et de chargement par rapport à l'insertion ou à la mise à jour des données.
upvoted 1 times

EmnCours 1 year, 8 months ago
Correct Answer: C
upvoted 1 times

AmitJais 1 year, 8 months ago
The correct answer is C
upvoted 1 times

xac123 1 year, 11 months ago
As switching partition does not involve movement of data, dbo.sales partition should be switched to the partition of stg.Sales so that partition from stg.Sales will be available in dbo.Sales
upvoted 2 times

devereux 1 year, 11 months ago
must be c
upvoted 1 times

SheShanD 1 year, 11 months ago
it's C
upvoted 1 times

roopansh.gupta2 1 year, 12 months ago

(C) because By switching the first partition from stg.Sales to dbo.Sales, you effectively replace the content of the first partition in dbo.Sales with the content of the corresponding partition in stg.Sales. This approach ensures that the partition in dbo.Sales is updated almost instantly, with minimal load time. Incorrect Options: A. Insert the data: Inserting data into dbo.Sales from stg.Sales would involve more overhead, as it requires data movement and processing, which can be time-consuming. B. Switch the first partition from dbo.Sales to stg.Sales: This would move the data out of dbo.Sales but would not achieve the goal of updating dbo.Sales with the data from stg.Sales. D. Update dbo.Sales: Updating the data would involve reading and writing the data, which would take significantly more time compared to partition switching.

upvoted 2 times

Okkier 2 years, 1 month ago

SWITCH [PARTITION source_partition_number_expression] TO [schema_name.] target_table [PARTITION target_partition_number_expression] SOURCE TO DETENTION

upvoted 1 times

binhdortmund 2 years, 1 month ago

actually B is correct. We have to understand the sentence as: Change the content of the first partition of dbo.Sales with the content of the first partition of stage.Sales. Thus, switch the content of first partition of dbo.Sales to/with the first partition of stage.Sales.

upvoted 1 times

ageorgieva 2 years, 2 months ago

Logical is C

upvoted 1 times

Dusica 2 years, 3 months ago

B is correct ALTER TABLE FactInternetSales SWITCH PARTITION 2 TO FactInternetSales_20000101 PARTITION 2; ALTER TABLE FactInternetSales SPLIT RANGE (20010101); https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition?context=%2Fazure%2Fsynapse-analytics%2Fcontext%2Fcontext

upvoted 2 times

dgerok 2 years, 4 months ago

The correct answer is C

upvoted 1 times

6152ee4 2 years, 4 months ago

switch the first partition of sales to stage.sales's partition

upvoted 1 times

6152ee4 2 years, 4 months ago

it is the way on wording....tricky. might be not native speaker or they shall ask differt way...

upvoted 1 times

dgerok 2 years, 4 months ago

Partition switching is a powerful technique that allows you to move data between tables quickly and with minimal logging. It's especially useful when dealing with large datasets. You switch the staging table into the production table, and the production table becomes the staging table. https://dba.stackexchange.com/questions/231918/best-solution-to-switch-staging-table-to-main-table-avoid-access-when-main-tabl

upvoted 1 times

Question #34 *Topic 1*

You are designing a slowly changing dimension (SCD) for supplier data in an Azure Synapse Analytics dedicated SQL pool.
You plan to keep a record of changes to the available fields.
The supplier data contains the following columns.
Which three additional columns should you add to the data to create a Type 2 SCD? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

Name	Description
SupplierSystemID	Unique supplier ID in an enterprise resource planning (ERP) system
SupplierName	Name of the supplier company
SupplierAddress1	Address of the supplier company
SupplierAddress2	Second address of the supplier company
SupplierCity	City of the supplier company
SupplierStateProvince	State or province of the supplier company
SupplierCountry	Country of the supplier company
SupplierPostalCode	Postal code of the supplier company
SupplierDescription	Free-test description of the supplier company
SupplierCategory	Category of goods provided by the supplier company

A. surrogate primary key

B. effective start date

C. business key

D. last modified date

E. effective end date

F. foreign key

Suggested Answer: *B, C, E*
C: The Slowly Changing Dimension transformation requires at least one business key column. BE: Historical attribute changes create new records instead of updating existing ones. The only change that is permitted in an existing record is an update to a column that indicates whether the record is current or expired. This kind of change is equivalent to a Type 2 change. The Slowly Changing Dimension transformation directs these rows to two outputs: Historical Attribute Inserts Output and New Output. Reference: https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/slowly-changing-dimension-transformation

Community vote distribution
ABE (87%) BCE (12%)

ItHYMeRIsh Highly Voted 4 years, 8 months ago

Selected Answer: ABE
The answer is ABE. A type 2 SCD requires a surrogate key to uniquely identify each record when versioning.

See https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types under SCD Type 2 " the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member."

A business key is already part of this table - SupplierSystemID. The column is derived from the source data.

upvoted 130 times

ItHYMeRIsh Highly Voted 4 years, 8 months ago

The answer is ABE. A type 2 SCD requires a surrogate key to uniquely identify each record when versioning. See https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types under SCD Type 2 " the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member." A business key is already part of this table - SupplierSystemID. The column is derived from the source data.CorrectCorrect, the answer is ABE indeed ! Type 2 SCD involves keeping a full history of changes to the data by creating new records for changes rather than overwriting old data. Each version of the record is identified by a surrogate key and is valid for a specific period, defined by the effective start and end dates.

upvoted 130 times

Achu24 3 years, 7 months ago

Correct

upvoted 4 times

Bouhdy 1 year, 11 months ago

Correct, the answer is ABE indeed ! Type 2 SCD involves keeping a full history of changes to the data by creating new records for changes rather than overwriting old data. Each version of the record is identified by a surrogate key and is valid for a specific period, defined by the effective start and end dates.

upvoted 1 times

CHOPIN 4 years, 7 months ago

WHAT ARE YOU GUYS TALKING ABOUT??? You are really misleading other people!!! No issue with the provided answer. Should be BCE!!! Check this out: https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/slowly-changing-dimension-transformation?view=sql-server-ver15 "The Slowly Changing Dimension transformation requires at least one business key column." [Surrogate key] is not mentioned in this Microsoft documentation AT ALL!!!Yes, because SupplierSystemID is unique. But Microsoft questions are terribly misleading here. People think that SupplierSystemID is business key because of Supplier in it. Also, there are some really not good and not sufficient examples on Learn. See https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-typesI don't understand. 1) What in your opinion should then be the business key. Can you explain please. 2) SupplierSysteID ist uniqe in the source system. Is there a definition that the column need to be unique also in the DataWarehouse? If no, there ist the possibility to use it as business key. Am I wrong?No you're not wrong, the unique identifier form the ERP system is the Business Keyhttps://learn.microsoft.com/en-us/training/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types Type 2 SCD: "The table must also define a surrogate key because the business key (in this instance, employee ID) won't be unique."https://www.mssqltips.com/sqlservertip/5602/why-surrogate-keys-are-needed-for-a-sql-server-data-warehouse/https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types Under Type 2 SCD, it indicates that it must have a surrogate keyLook at the following Question #35 and you will understand quickly! ;-)

upvoted 24 times

CHOPIN 4 years, 7 months ago

Selected Answer: BCE

WHAT ARE YOU GUYS TALKING ABOUT??? You are really misleading other people!!! No issue with the provided answer. Should be BCE!!!

Check this out: https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/slowly-changing-dimension-transformation?view=sql-server-ver15 "The Slowly Changing Dimension transformation requires at least one business key column." [Surrogate key] is not mentioned in this Microsoft documentation AT ALL!!!

upvoted 24 times

assU2 4 years, 7 months ago

Yes, because SupplierSystemID is unique. But Microsoft questions are terribly misleading here. People think that SupplierSystemID is business key because of Supplier in it. Also, there are some really not good and not sufficient examples on Learn. See https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types

upvoted 1 times

assU2 4 years, 7 months ago

Yes, because SupplierSystemID is unique. But Microsoft questions are terribly misleading here. People think that SupplierSystemID is business key because of Supplier in it. Also, there are some really not good and not sufficient examples on Learn. See https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-typesI don't understand. 1) What in your opinion should then be the business key. Can you explain please. 2) SupplierSysteID ist uniqe in the source system. Is there a definition that the column need to be unique also in the DataWarehouse? If no, there ist the possibility to use it as business key. Am I wrong?No you're not wrong, the unique identifier form the ERP system is the Business Keyhttps://learn.microsoft.com/en-us/training/modules/populate-slowly-changing-

dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types Type 2 SCD: "The table must also define a surrogate key because the business key (in this instance, employee ID) won't be unique."

upvoted 1 times

Mad_001 4 years, 5 months ago

I don't understand. 1) What in your opinion should then be the business key. Can you explain please. 2) SupplierSysteID ist uniqe in the source system. Is there a definition that the column need to be unique also in the DataWarehouse? If no, there ist the possibility to use it as business key. Am I wrong?No you're not wrong, the unique identifier form the ERP system is the Business Key

upvoted 4 times

Mad_001 4 years, 5 months ago

I don't understand. 1) What in your opinion should then be the business key. Can you explain please. 2) SupplierSysteID ist uniqe in the source system. Is there a definition that the column need to be unique also in the DataWarehouse? If no, there ist the possibility to use it as business key. Am I wrong?

upvoted 4 times

Onobhas01 4 years, 4 months ago

No you're not wrong, the unique identifier form the ERP system is the Business Key

upvoted 3 times

practia 2 years ago

https://learn.microsoft.com/en-us/training/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types Type 2 SCD: "The table must also define a surrogate key because the business key (in this instance, employee ID) won't be unique."

upvoted 1 times

muove 4 years ago

https://www.mssqltips.com/sqlservertip/5602/why-surrogate-keys-are-needed-for-a-sql-server-data-warehouse/

upvoted 1 times

SHENOOOO 3 years, 6 months ago

https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types Under Type 2 SCD, it indicates that it must have a surrogate key

upvoted 3 times

auwia 3 years, 1 month ago

Look at the following Question #35 and you will understand quickly! ;-)

upvoted 1 times

IMadnan Most Recent 1 year, 6 months ago

The answer is ABE. A type 2 SCD requires a surrogate key to uniquely identify each record when versioning.

upvoted 1 times

The_OTC 1 year, 7 months ago

Why not D: Last Modified Date ??????

upvoted 1 times

EmnCours 1 year, 8 months ago

Correct Answer: ABE

upvoted 1 times

jarobdataeng 1 year, 11 months ago

In my humble opinion the SupplierSystemID column is the business key: it remains an IDENTIFY column to represent the surrogate key

upvoted 1 times

iceberge 2 years ago

Copilot Sent by Copilot: To create a Type 2 Slowly Changing Dimension (SCD) in Azure Synapse Analytics, you should add the following three columns: A. Surrogate primary key This uniquely identifies each version of a record, allowing you to track changes over time. B. Effective start date This indicates when a particular version of the record became active. E. Effective end date This indicates when a particular version of the record was superseded by a new version. These columns help maintain historical data by creating new records for each change, rather than updating existing records. This way, you can track the history of changes to the supplier data.

upvoted 1 times

Okkier 2 years, 1 month ago

The surrogate key is a must, so the proposed answer is incorrect in my view. The start date and end date for each record have always been an over-design in my view. This means that you have to maintain both records. The latest record of the set with the same business key is the latest record, and you only need one date, the valid from date, in that row. I have long stopped versioning with valid from and valid to dates. You can easily use a simple lag function to grab a previous valid to date for the previous record, which is then the valid from date for the latest record. So why still use both dates?

upvoted 2 times

sergio_eduardo 2 years, 2 months ago

The answer is ABE, you need subrogate key,

upvoted 1 times

codedev91 2 years, 3 months ago

ABE SCD Type 2 doesn't work with business key it requires a surrogate key to uniquely identify each record when versioning., for example - an employee can have a unique key(i.e. Business key) in the organization, but when you create SCD Type 2, you need to preserve historical data, hence then even a unique key of an Employee can't uniquely different the record but Surrogate key will do it.

upvoted 2 times

Dusica 2 years, 3 months ago

ABE; SupplierSystemId is a business key

upvoted 1 times

olayvo 2 years, 4 months ago

I think ABE, see as below: "A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members." https://learn.microsoft.com/en-us/training/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types

upvoted 1 times

dgerok 2 years, 4 months ago

Surrogate primary key (ensure to include edited records) - A, effective start date - B, effective end date - E

upvoted 2 times

ulrom 2 years, 4 months ago

Should be ABE. According to a Microsoft learning SCD Type 2 requires a surrogate key. "A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member." https://learn.microsoft.com/en-us/training/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types

upvoted 2 times

j888 2 years, 6 months ago

The surrogate key is an artificially generated key, usually an incremental or globally unique identifier. Business key derived from the actual business data and has business meaning. Examples could include a product code, customer ID, employee ID, or any other identifier that has relevance in the business context. I think this is it's business key Effective end and start date

upvoted 1 times

rocky48 2 years, 6 months ago

Selected Answer: ABE The answer is ABE. A type 2 SCD requires a surrogate key to uniquely identify each record when versioning. See https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types under SCD Type 2 " the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member." A business key is already part of this table - SupplierSystemID. The column is derived from the source data.

upvoted 2 times

blazy001 2 years, 8 months ago

I'm working now more then 13years with this stuff, ABE is correct. CHOPIN is wrong. a 2 SCD needs a unique ID, this is the surrogate key, besides, in the table given, there is already a business key , is the first column A business key is NOT unique in an 2 SCD hallo

upvoted 3 times

HOTSPOT -

You have a Microsoft SQL Server database that uses a third normal form schema.

You plan to migrate the data in the database to a star schema in an Azure Synapse Analytics dedicated SQL pool.

You need to design the dimension tables. The solution must optimize read operations.

What should you include in the solution? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Transform data for the dimension tables by:

- Maintaining to a third normal form
- Normalizing to a fourth normal form
- Denormalizing to a second normal form

For the primary key columns in the dimension tables, use:

- New IDENTITY columns
- A new computed column
- The business key column from the source sys

Suggested Answer:

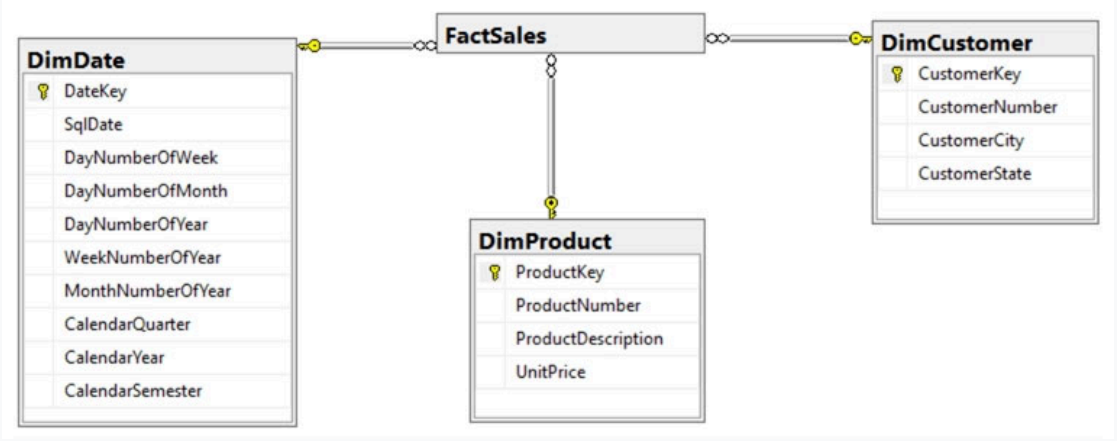
Answer Area

Transform data for the dimension tables by:

- Maintaining to a third normal form
- Normalizing to a fourth normal form
- Denormalizing to a second normal form

For the primary key columns in the dimension tables, use:

- New IDENTITY columns
- A new computed column
- The business key column from the source sys



Answer Area

Transform data for the dimension tables by:

- Maintaining to a third normal form
- Normalizing to a fourth normal form
- Denormalizing to a second normal form

For the primary key columns in the dimension tables, use:

- New IDENTITY columns
- A new computed column
- The business key column from the source sys

Box 1: Denormalize to a second normal form Denormalization is the process of transforming higher normal forms to lower normal forms via storing the join of higher normal form relations as a base relation. Denormalization increases the performance in data retrieval at cost of bringing update anomalies to a database.

Box 2: New identity columns - The collapsing relations strategy can be used in this step to collapse classification entities into component entities to obtain flat dimension tables with single-part keys that connect directly to the fact table. The single-part key is a surrogate key generated to ensure it remains unique over time. Example:

Note: A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance. Reference: https://www.mssqltips.com/sqlservertip/5614/explore-the-role-of-normal-forms-in-dimensional-modeling/ https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identity

PallaviPatel Highly Voted 4 years, 6 months ago

Answer correct.

upvoted 26 times

ajhak Highly Voted 3 years, 3 months ago

Always denormalize when moving from database to date warehouse.

upvoted 8 times

ajhak 3 years, 3 months ago

Always denormalize when moving from database to date warehouse.God no! Only dimensional models, read Inmon and Lindstedt approach.

upvoted 8 times

Rob77 3 years, 2 months ago

God no! Only dimensional models, read Inmon and Lindstedt approach.

upvoted 2 times

roopansh.gupta2 Most Recent 1 year, 12 months ago

Why Denormalizing to 2nd Normal Form (2NF) is CORRECT?: In a star schema, dimension tables are typically denormalized to reduce the number of joins required during query execution. This denormalization improves query performance, particularly in read-heavy workloads typical in analytics environments like Azure Synapse. Denormalizing to 2NF involves reducing the number of tables and increasing the redundancy within the dimension tables, which is beneficial for faster reads. Maintaining 3NF: While 3rd Normal Form (3NF) minimizes redundancy and ensures data integrity, it also requires more complex joins during queries, which can slow down read operations in a star schema designed for analytics.

upvoted 3 times

kkk5566 2 years, 11 months ago

denormalizing and IDENTITY

upvoted 2 times

SHENOOOO 3 years, 6 months ago

Answer Is correct

upvoted 3 times

Deepshikha1228 4 years ago

Answer is correct

upvoted 3 times

JimZhang4123 4 years, 3 months ago

'The solution must optimize read operations.' means denormalization

upvoted 6 times

Mahesh_mm 4 years, 7 months ago

Answers are correct

upvoted 1 times

PallaviPatel 4 years, 8 months ago

answer is correct

upvoted 4 times

moreinva43 4 years, 8 months ago

While denormalizing does require implementing a lower level of normalization, the second normal form ONLY applies when a table has a composite primary key. https://www.geeksforgeeks.org/second-normal-form-2nf/

upvoted 1 times

HOTSPOT -

You plan to develop a dataset named Purchases by using Azure Databricks. Purchases will contain the following columns:

- ProductID
- ItemPrice
- LineTotal
- Quantity
- StoreID
- Minute
- Month
- Hour

Year -

- Day

You need to store the data to support hourly incremental load pipelines that will vary for each Store ID. The solution must minimize storage costs.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

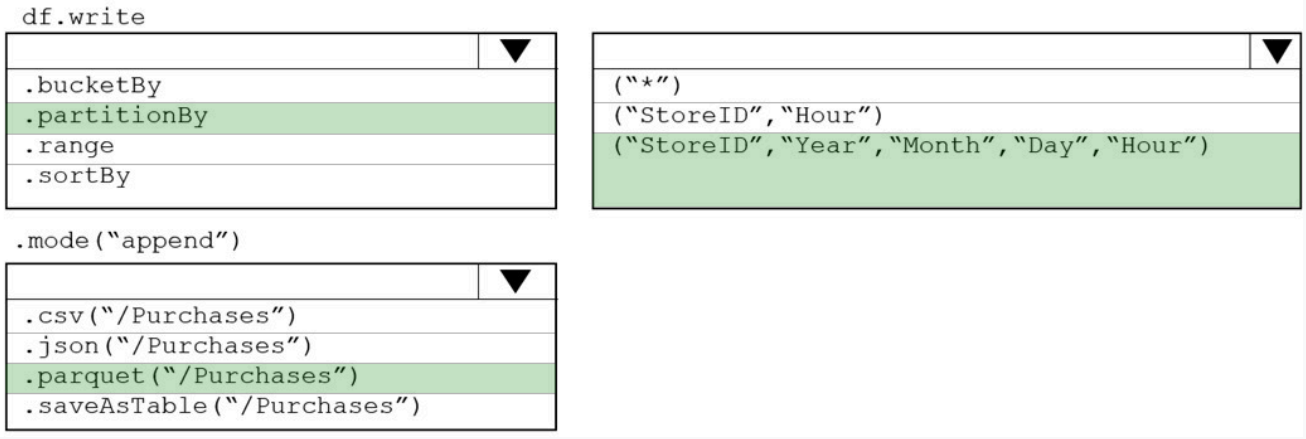
.

Answer Area

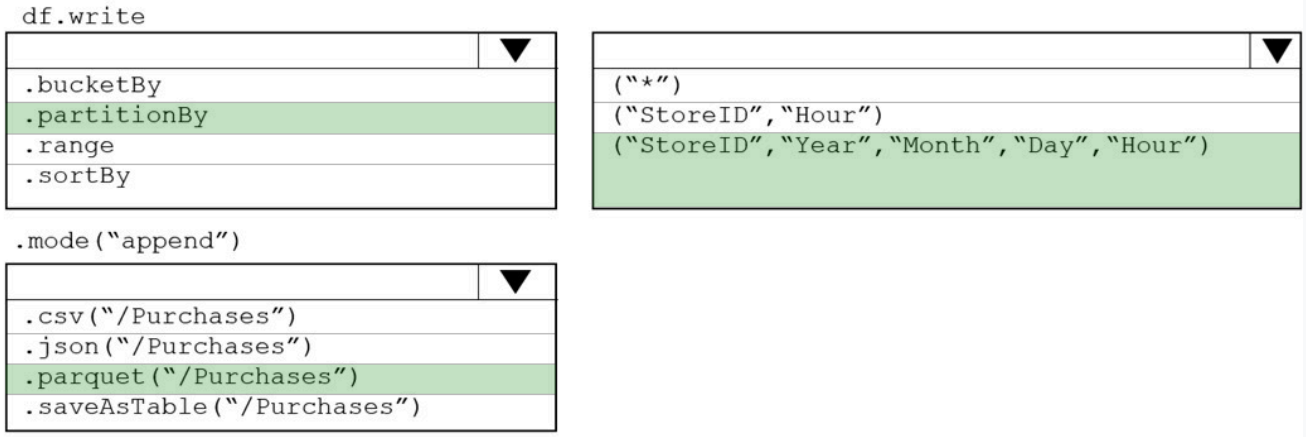
```
df.write
[ ▼ ]                                  [ ▼ ]
.bucketBy                              ("*")
.partitionBy                           ("StoreID","Hour")
.range                                 ("StoreID","Year","Month","Day","Hour")
.sortBy
.mode("append")
[ ▼ ]
.csv("/Purchases")
.json("/Purchases")
.parquet("/Purchases")
.saveAsTable("/Purchases")
```

Suggested Answer:

Answer Area



Answer Area



Box 1: partitionBy - We should overwrite at the partition level. Example: df.write.partitionBy("y","m","d") .mode(SaveMode.Append) .parquet("/data/hive/warehouse/db_name.db/" + tableName) Box 2: ("StoreID", "Year", "Month", "Day", "Hour", "StoreID") Box 3: parquet("/Purchases") Reference: https://intellipaat.com/community/11744/how-to-partition-and-write-dataframe-in-spark-without-deleting-partitions-with-no-new-data

Mahesh_mm 4 years, 1 month ago

Answers are correct

upvoted 24 times

Aslam208 4 years, 2 months ago

correct

upvoted 9 times

dgerok 1 year, 10 months ago

The answer is correct

upvoted 2 times

ELJORDAN23 2 years, 1 month ago

Got this question on my exam on january 17, the answer is correct

upvoted 7 times

hodashiyam 2 years, 5 months ago

Answers are correct

upvoted 1 times

kkk5566 2 years, 5 months ago

correct

upvoted 1 times

Rrk07 3 years, 2 months ago

Why parquet option? Can anyone explain.Because Parquet is always the answer.The solution must minimize storage costs.I guess because of the requirement "reducing storage costs"

upvoted 3 times

Rrk07 3 years, 2 months ago

Why parquet option? Can anyone explain.

upvoted 3 times

phydev Highly Voted 2 years, 3 months ago

Because Parquet is always the answer.

upvoted 32 times

steveo123 2 years, 9 months ago

The solution must minimize storage costs.

upvoted 6 times

DataSaM 2 years, 7 months ago

I guess because of the requirement "reducing storage costs"

upvoted 6 times

gabrielkuka 3 years, 2 months ago

Can somebody explain why are we partitioning by StoreId, Year, Month, Day and Hour instead of just StoreID and Hour?if partitioned by storeid and hour only, the same hours from different days would go to the same partition, that would be innefficient

upvoted 6 times

gabrielkuka 3 years, 2 months ago

Can somebody explain why are we partitioning by StoreId, Year, Month, Day and Hour instead of just StoreID and Hour?

upvoted 6 times

dduque10 Highly Voted 3 years, 2 months ago

if partitioned by storeid and hour only, the same hours from different days would go to the same partition, that would be innefficient

upvoted 41 times

Keerthi24 3 years, 5 months ago

Can someone explain why parquet and not saveAsTable option?Parquet is columnar, so faster to be read by Azure Synapse Analytics via CETAs.saveAsTable default saves as delta table, which is parquet but with additional _delta_log

upvoted 3 times

Keerthi24 3 years, 5 months ago

Can someone explain why parquet and not saveAsTable option?

upvoted 3 times

uira 3 years, 2 months ago

Parquet is columnar, so faster to be read by Azure Synapse Analytics via CETAs.saveAsTable default saves as delta table, which is parquet but with additional _delta_log

upvoted 7 times

uira 3 years, 2 months ago

Parquet is columnar, so faster to be read by Azure Synapse Analytics via CETAs.

upvoted 7 times

hypersam Most Recent 1 year, 7 months ago

saveAsTable default saves as delta table, which is parquet but with additional _delta_log

upvoted 1 times

Deeksha1234 3 years, 6 months ago

given answers are correct

upvoted 3 times

hm358 3 years, 8 months ago

Correct

upvoted 2 times

sparkchu 3 years, 10 months ago

ans should be saveAsTable. format is defined by format() method.

upvoted 4 times

assU2 4 years, 1 month ago

Can anyone explain why it's Partitioning and not Bucketing pls?

upvoted 6 times

assU2 4 years, 1 month ago

Can anyone explain why it's Partitioning and not Bucketing pls?There should be a different folder for each store. Partitioning will create separate folder for each storeId. In bucketing, multiple stores having same hash value can be present in the same file, so multiple storeIds can be present under a single file.Is it a question of correct syntax (numBuckets int the number of buckets to save) or is it smth else?Bucketing feature (part of data skipping index) was removed and microsoft recommends using DeltaLake, which uses the partition syntax. https://docs.microsoft.com/en-us/azure/databricks/spark/latest/spark-sql/dataskipping-index

upvoted 6 times

bhanuprasad9331 4 years ago

There should be a different folder for each store. Partitioning will create separate folder for each storeId. In bucketing, multiple stores having same hash value can be present in the same file, so multiple storeIds can be present under a single file.

upvoted 11 times

assU2 4 years, 1 month ago

Is it a question of correct syntax (numBuckets int the number of buckets to save) or is it smth else?

upvoted 2 times

KashRaynardMorse 3 years, 9 months ago

Bucketing feature (part of data skipping index) was removed and microsoft recommends using DeltaLake, which uses the partition syntax. https://docs.microsoft.com/en-us/azure/databricks/spark/latest/spark-sql/dataskipping-index

upvoted 6 times

Question #37 *Topic 1*

You are designing a partition strategy for a fact table in an Azure Synapse Analytics dedicated SQL pool. The table has the following specifications:

✑ Contain sales data for 20,000 products.

Use hash distribution on a column named ProductID.

✑ Contain 2.4 billion records for the years 2019 and 2020.

Which number of partition ranges provides optimal compression and performance for the clustered columnstore index?

.

A. 40

B. 240

C. 400

D. 2,400

Suggested Answer: *A*

Each partition should have around 1 millions records. Dedication SQL pools already have 60 partitions. We have the formula: Records/(Partitions*60)= 1 million Partitions= Records/(1 million * 60) Partitions= 2.4 x 1,000,000,000/(1,000,000 * 60) = 40 Note: Having too many partitions can reduce the effectiveness of clustered columnstore indexes if each partition has fewer than 1 million rows. Dedicated SQL pools automatically partition your data into 60 databases. So, if you create a table with 100 partitions, the result will be 6000 partitions.

Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

Community vote distribution



Aslam208 Highly Voted 4 years, 8 months ago

correct

upvoted 26 times

AZLearn111 Highly Voted 3 years, 6 months ago

No of automatic Distributions is 60. So each distribution will have 2.4 B / 60 = 40M. For a good performance each partition within a distribution (some time called buckets of data) should have 1M rows per bucket. So 40M / 1M = 40 partitions.Thank you for your comment. I do not agree. Your logic conflates rowgroups with table partitions. You don't define partitions specifically for each distribution. The 60 distributions operate automatically, regardless of how many partitions you define.Table partitions are logical divisions, and Azure Synapse automatically handles the data distribution within each partition across its 60 distributions. If you only define 40 partitions, each partition would hold 2.4 billion ÷ 40 = 60 million rows.2.4 billion ÷ 40 = 60 million rows. This would make partitions very large and reduce the effectiveness of partition elimination, especially for date-based queries. 240 partitions aligns better with the guideline of 1–10 million rows per partition, allowing better granularity and performance for large-scale data.Upon further review of the documentation it is clearly stated...Guess I was wrong. "If you partition your data, each partition will need to have 1 million rows to benefit from a clustered columnstore index. For a table with 100 partitions, it needs to have at least 6 billion rows to benefit from a clustered columns store (60 distributions 100 partitions 1 million rows). If your table doesn't have 6 billion rows, you have two main options. Either reduce the number of partitions or consider using a heap table instead. It also may be worth experimenting to see if better performance can be gained by using a heap table with secondary indexes rather than a columnstore table.?" https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-poolAnother way to think about this: The number of records for the period stated = 2.4 billion Number of underlying ("automatic") distributions: 60 2.4 billion / 60 distribut

upvoted 19 times

AZLearn111 3 years, 6 months ago

No of automatic Distributions is 60. So each distribution will have 2.4 B / 60 = 40M. For a good performance each partition within a distribution (some time called buckets of data) should have 1M rows per bucket. So 40M / 1M = 40 partitions.

upvoted 19 times

JustImperius 1 year, 7 months ago

Thank you for your comment. I do not agree. Your logic conflates rowgroups with table partitions. You don't define partitions specifically for each distribution. The 60 distributions operate automatically, regardless of how many partitions you define.Table partitions are logical divisions, and Azure Synapse automatically handles the data distribution within each partition across its 60 distributions. If you only define 40 partitions, each partition

would hold 2.4 billion ÷ 40 = 60 million rows.2.4 billion ÷ 40 = 60 million rows. This would make partitions very large and reduce the effectiveness of partition elimination, especially for date-based queries. 240 partitions aligns better with the guideline of 1–10 million rows per partition, allowing better granularity and performance for large-scale data.Upon further review of the documentation it is clearly stated...Guess I was wrong. "If you partition your data, each partition will need to have 1 million rows to benefit from a clustered columnstore index. For a table with 100 partitions, it needs to have at least 6 billion rows to benefit from a clustered columns store (60 distributions 100 partitions 1 million rows). If your table doesn't have 6 billion rows, you have two main options. Either reduce the number of partitions or consider using a heap table instead. It also may be worth experimenting to see if better performance can be gained by using a heap table with secondary indexes rather than a columnstore table.?" https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

upvoted 1 times

JustImperius 1 year, 7 months ago

Thank you for your comment. I do not agree. Your logic conflates rowgroups with table partitions. You don't define partitions specifically for each distribution. The 60 distributions operate automatically, regardless of how many partitions you define.Table partitions are logical divisions, and Azure Synapse automatically handles the data distribution within each partition across its 60 distributions. If you only define 40 partitions, each partition would hold 2.4 billion ÷ 40 = 60 million rows.2.4 billion ÷ 40 = 60 million rows. This would make partitions very large and reduce the effectiveness of partition elimination, especially for date-based queries.

240 partitions aligns better with the guideline of 1–10 million rows per partition, allowing better granularity and performance for large-scale data.

upvoted 1 times

JustImperius 1 year, 7 months ago

Upon further review of the documentation it is clearly stated...Guess I was wrong. "If you partition your data, each partition will need to have 1 million rows to benefit from a clustered columnstore index. For a table with 100 partitions, it needs to have at least 6 billion rows to benefit from a clustered columns store (60 distributions 100 partitions 1 million rows). If your table doesn't have 6 billion rows, you have two main options. Either reduce the number of partitions or consider using a heap table instead. It also may be worth experimenting to see if better performance can be gained by using a heap table with secondary indexes rather than a columnstore table.?" https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

upvoted 1 times

zekescookies 3 years, 3 months ago

Another way to think about this: The number of records for the period stated = 2.4 billion Number of underlying ("automatic") distributions: 60 2.4 billion / 60 distributions = 40 million rows 40 million / 40 partitions = 1 million rows As stated, 1 million rows per distribution are optimal for compression and performance. Divide the 40 million rows with the other partitioning options and you have too few rows per distribution -> suboptimal.

upvoted 5 times

samirarian Most Recent 1 year, 6 months ago

1 million in every partition is the best performance

upvoted 1 times

iam_momo88 1 year, 7 months ago

240 partitions balances the data across partitions, with each partition handling approximately 10 million rows: Rows per partition = 2.4 billion rows 240 partitions = 10 million rows per partition . Rows per partition= 240 partitions 2.4 billion rows =10 million rows per partition

upvoted 2 times

Parakkal 1 year, 7 months ago

Isn't the question about finding the total number of partitions, not partitions per distribution. Dividing the total rows (2.4B) by 10M gives 240 partitions, which falls within the optimal range of 1–60M rows per partition. Am I missing something? The most voted answer and the other comments are to find partition within each distribution.

upvoted 2 times

moize 1 year, 8 months ago

Les index columnstore fonctionnent de manière optimale lorsqu'une partition contient environ 1 million de lignes par segment de colonne. Vous avez 2,4 milliards d'enregistrements. Diviser ces données pour que chaque partition contienne au moins 1 million de lignes donne : Nombre optimal de partitions = 2 , 4 milliards 1 million = 2400 Nombre optimal de partitions= 1million 2,4milliards =2400 2400 partitions garantiront que chaque partition est suffisamment compacte pour une compression efficace et des performances optimales.

upvoted 2 times

e56bb91 2 years, 1 month ago

2.4B = 2400M 2400/60 = 40

upvoted 1 times

Danweo 2 years, 1 month ago

A is correct

upvoted 1 times

hassexat 2 years, 11 months ago

2,400,000,000 / 60,000,000 = 40

upvoted 2 times

kkk5566 2 years, 11 months ago

A is correct

upvoted 1 times

akhil5432 3 years ago

OPTION A

upvoted 1 times

vrodriguesp 3 years, 6 months ago

Considering that: Having too many partitions can reduce the effectiveness of clustered columnstore indexes if each partition has fewer than 1 million rows. Dedicated SQL pools automatically partition your data into 60 databases So a table with no partiton (or just one partition) has 60Milion of records I have use this logic, simple proportion: 1 partion : 60M = x = 2.4 B ==> 1 : 60 M = x : 2400 M ==> x = 2400 / 60 ==> x = 40 partitions1 partion : 60M = x : 2.4 B 1 partion : 60 M = x : 2400 M ==> x = 2400 / 60 ==> x = 40 partitions

upvoted 3 times

vrodriguesp 3 years, 6 months ago

Selected Answer: A

Considering that: Having too many partitions can reduce the effectiveness of clustered columnstore indexes if each partition has fewer than 1 million rows. Dedicated SQL pools automatically partition your data into 60 databases So a table with no partiton (or just one partition) has 60Milion of records

I have use this logic, simple proportion:

1 partion : 60M = x = 2.4 B ==> 1 : 60 M = x : 2400 M ==> x = 2400 / 60 ==> x = 40 partitions

upvoted 3 times

vrodriguesp 3 years, 6 months ago

1 partion : 60M = x : 2.4 B 1 partion : 60 M = x : 2400 M ==> x = 2400 / 60 ==> x = 40 partitions

upvoted 3 times

SHENOOOO 3 years, 6 months ago

Correct Answer

upvoted 1 times

NORLI 3 years, 10 months ago

Very simple go with the smallest partition because too many partitions affect peformance

upvoted 4 times

dom271219 3 years, 11 months ago

2,4bn/60=40M

upvoted 13 times

Deeksha1234 4 years ago

correct

upvoted 1 times

hm358 4 years, 2 months ago

Optimal distribution is up to 60 instances

upvoted 1 times

HOTSPOT -

You are creating dimensions for a data warehouse in an Azure Synapse Analytics dedicated SQL pool.

You create a table by using the Transact-SQL statement shown in the following exhibit.

Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.

NOTE: Each correct selection is worth one point.

Hot Area:

```
CREATE TABLE [DBO].[DimProduct](
	[ProductKey] [int] IDENTITY(1,1) NOT NULL,
	[ProductSourceID] [int] NOT NULL,
	[ProductName] [nvarchar](100) NOT NULL,
	[ProductNumber] [nvarchar](25) NOT NULL,
	[Color] [nvarchar](15) NULL,
	[Size] [nvarchar](5) NULL,
	[Weight] [decimal](8, 2) NULL,
	[ProductCategory] [nvarchar](100) NULL,
	[SellStartDate] [date] NOT NULL,
	[SellEndDate] [date] NULL,
	[RowInsertedDateTime] [datetime] NOT NULL,
	[RowUpdatedDateTime] [datetime] NOT NULL,
	[ETLAuditID] [int] NOT NULL
)
```

Answer Area

DimProduct is a **[answer choice]** slowly changing dimension (SCD).

- Type 0
- Type 1
- Type 2

The ProductKey column is **[answer choice]**.

- a surrogate key
- a business key
- an audit column

Suggested Answer:

Answer Area

DimProduct is a **[answer choice]** slowly changing dimension (SCD).

- Type 0
- Type 1
- Type 2

The ProductKey column is **[answer choice]**.

- a surrogate key
- a business key
- an audit column

Answer Area

DimProduct is a **[answer choice]** slowly changing dimension (SCD).

- Type 0
- Type 1
- Type 2

The ProductKey column is **[answer choice]**.

- a surrogate key
- a business key
- an audit column

Box 1: Type 2 - A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members. Incorrect Answers: A Type 1 SCD always reflects the latest values, and when changes in source data are detected, the dimension table data is overwritten.

Box 2: a business key - A business key or natural key is an index which identifies uniqueness of a row based on columns that exist naturally in a table according to business rules. For example business keys are customer code in a customer table, composite of sales order header number and sales order item line number within a sales order details table. Reference: https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types

nkav Highly Voted 5 years, 3 months ago

product key is a surrogate key as it is an identity columnAgree on the surrogate key, exactly. "In data warehousing, IDENTITY functionality is particularly important as it makes easier the creation of surrogate keys." Why ProductKey is certainly not a business key: "The IDENTITY value in Synapse is not guaranteed to be unique if the user explicitly inserts a duplicate value with 'SET IDENTITY_INSERT ON' or reseeds IDENTITY". Business key is an index which identifies uniqueness of a row and here Microsoft says that identity doesn't guarantee uniqueness. References: https://azure.microsoft.com/en-us/blog/identity-now-available-with-azure-sql-data-warehouse/ https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identityType 2 In order to support type 2 changes, we need to add four columns to our table: · Surrogate Key – the original ID will no longer be sufficient to identify the specific record we require, we therefore need to create a new ID that the fact records can join to specifically. · Current Flag – A quick method of returning only the current version of each record · Start Date – The date from which the specific historical version is active · End Date – The date to which the specific historical version record is active With these elements in place, our table will now look like:You know the start and end date of the selling, but not the version of the product (size, color....) If the waterpump

size has changed for a car by rhe producer but the function product id are the same, selling is continuous you can not store it into this structure. But you can store at the timepoints and see validation point in creation date of the record. Therefore it is type 1TOTALLY AGREE HERE !!!! The ProductKey column is defined as an IDENTITY(1,1) column, which means it is an auto-incremented value and serves as a unique identifier for each row. This is characteristic of a surrogate key, which is used in dimensional modeling to uniqu

upvoted 238 times

nkav Highly Voted 5 years, 3 months ago

product key is a surrogate key as it is an identity column

upvoted 238 times

111222333 5 years, 3 months ago

Agree on the surrogate key, exactly. "In data warehousing, IDENTITY functionality is particularly important as it makes easier the creation of surrogate keys." Why ProductKey is certainly not a business key: "The IDENTITY value in Synapse is not guaranteed to be unique if the user explicitly inserts a duplicate value with 'SET IDENTITY_INSERT ON' or reseeds IDENTITY". Business key is an index which identifies uniqueness of a row and here Microsoft says that identity doesn't guarantee uniqueness. References: https://azure.microsoft.com/en-us/blog/identity-now-available-with-azure-sql-data-warehouse/ https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identityType 2 In order to support type 2 changes, we need to add four columns to our table: · Surrogate Key – the original ID will no longer be sufficient to identify the specific record we require, we therefore need to create a new ID that the fact records can join to specifically. · Current Flag – A quick method of returning only the current version of each record · Start Date – The date from which the specific historical version is active · End Date – The date to which the specific historical version record is active With these elements in place, our table will now look like:You know the start and end date of the selling, but not the version of the product (size, color....) If the waterpump size has changed for a car by rhe producer but the function product id are the same, selling is continuous you can not store it into this structure. But you can store at the timepoints and see validation point in creation date of the record. Therefore it is type 1

upvoted 17 times

111222333 5 years, 3 months ago

Agree on the surrogate key, exactly.

"In data warehousing, IDENTITY functionality is particularly important as it makes easier the creation of surrogate keys."

Why ProductKey is certainly not a business key: "The IDENTITY value in Synapse is not guaranteed to be unique if the user explicitly inserts a duplicate value with 'SET IDENTITY_INSERT ON' or reseeds IDENTITY". Business key is an index which identifies uniqueness of a row and here Microsoft says that identity doesn't guarantee uniqueness.

References: https://azure.microsoft.com/en-us/blog/identity-now-available-with-azure-sql-data-warehouse/ https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identity

upvoted 17 times

rikku33 4 years, 10 months ago

Type 2

In order to support type 2 changes, we need to add four columns to our table:

· Surrogate Key – the original ID will no longer be sufficient to identify the specific record we require, we therefore need to create a new ID that the fact records can join to specifically.

· Current Flag – A quick method of returning only the current version of each record

· Start Date – The date from which the specific historical version is active

· End Date – The date to which the specific historical version record is active

With these elements in place, our table will now look like:

upvoted 14 times

rikku33 4 years, 10 months ago

Type 2 In order to support type 2 changes, we need to add four columns to our table: · Surrogate Key – the original ID will no longer be sufficient to identify the specific record we require, we therefore need to create a new ID that the fact records can join to specifically. · Current Flag – A quick method of returning only the current version of each record · Start Date – The date from which the specific historical version is active · End Date – The date to which the specific historical version record is active With these elements in place, our table will now look like:You know the start and end

date of the selling, but not the version of the product (size, color....) If the waterpump size has changed for a car by rhe producer but the function product id are the same, selling is continuous you can not store it into this structure. But you can store at the timepoints and see validation point in creation date of the record. Therefore it is type 1

upvoted 14 times

Gikan 2 years, 6 months ago

You know the start and end date of the selling, but not the version of the product (size, color....) If the waterpump size has changed for a car by rhe producer but the function product id are the same, selling is continuous you can not store it into this structure. But you can store at the timepoints and see validation point in creation date of the record. Therefore it is type 1

upvoted 7 times

Bouhdy 1 year, 11 months ago

TOTALLY AGREE HERE !!!! The ProductKey column is defined as an IDENTITY(1,1) column, which means it is an auto-incremented value and serves as a unique identifier for each row. This is characteristic of a surrogate key, which is used in dimensional modeling to uniquely identify records in a dimension table, especially in a Type 2 SCD.

upvoted 5 times

sagga 5 years, 3 months ago

Type2 because there are start and end columns and ProductKey is a surrogate key. ProductNumber seems a business key.

upvoted 51 times

sagga 5 years, 3 months ago

Type2 because there are start and end columns and ProductKey is a surrogate key. ProductNumber seems a business key.The start and end columns are for when to when the product was being sold, not for metadata purposes. That makes it: Type 1 – No History Update record directly, there is no record of historical values, only current stateExactly how I saw itWhen the product is not being sold anymore, it becomes a historical record. Hence Type 2.With type 2, you normally don't update any column of a row other than row start date and end date.It is type 1 not 2I agree with the first part. From just the table it's impossible to know if the changes in the products are ignored or are updated, if you don't see the ETL. I suppose there is some mistake in the name of the fields start end effective fields.

upvoted 51 times

DrC 5 years, 2 months ago

The start and end columns are for when to when the product was being sold, not for metadata purposes. That makes it: Type 1 – No History Update record directly, there is no record of historical values, only current stateExactly how I saw itWhen the product is not being sold anymore, it becomes a historical record. Hence Type 2.With type 2, you normally don't update any column of a row other than row start date and end date.It is type 1 not 2I agree with the first part. From just the table it's impossible to know if the changes in the products are ignored or are updated, if you don't see the ETL. I suppose there is some mistake in the name of the fields start end effective fields.

upvoted 124 times

DrC 5 years, 2 months ago

The start and end columns are for when to when the product was being sold, not for metadata purposes. That makes it: Type 1 – No History Update record directly, there is no record of historical values, only current state

upvoted 124 times

captainbee 5 years, 2 months ago

Exactly how I saw it

upvoted 2 times

Kyle1 4 years, 11 months ago

When the product is not being sold anymore, it becomes a historical record. Hence Type 2.

upvoted 2 times

Kyle1 4 years, 11 months ago

When the product is not being sold anymore, it becomes a historical record. Hence Type 2.With type 2, you normally don't update any column of a row other than row start date and end date.It is type 1 not 2

upvoted 2 times

Yuri1101 4 years, 8 months ago

With type 2, you normally don't update any column of a row other than row start date and end date.

upvoted 1 times

rockyc05 4 years, 6 months ago

It is type 1 not 2

upvoted 4 times

borinot 3 years, 9 months ago

I agree with the first part. From just the table it's impossible to know if the changes in the products are ignored or are updated, if you don't see the ETL. I suppose there is some mistake in the name of the fields start end effective fields.

upvoted 2 times

AMJB Most Recent 1 year, 5 months ago

I agree with the surrogate key cos of the identiy (1,1) clause however, if you look at the adventureworks dw tables, product key is a business key or primary key on the dimproduct table. It is confusing but i guess the gve away is that the table is a SCD table so yea I'm swayed towards Dimensin table

upvoted 1 times

roopansh.gupta2 1 year, 12 months ago

SURROGATE KEY AND SCD 1: The absence of a Current_Flag or similar indicator and the RowUpdatedDateTime column might suggest that the DimProduct table is not strictly adhering to a Type 2 Slowly Changing Dimension (SCD). Instead, it seems to be implementing a Type 1 Slowly Changing Dimension (SCD). Further evidence for SCD 1: RowUpdatedDateTime: The RowUpdatedDateTime column indicates when the row was last modified, which is consistent with Type 1 SCDs where data is overwritten. No Historical Tracking: There is no mechanism to track historical changes (e.g., Current_Flag, EffectiveDate, EndDate). Without these, the table does not store previous versions of the data, reinforcing that this is a Type 1 SCD.

upvoted 5 times

jonatahs18 2 years ago

The ProductKey is a surrogate key, as you can se the query use IDENTITY(1,1) to identify unique values in the table. If was a Business Key, the query don't need to specific the INDENTITY, usually we use Concat to create new business Key or naturally the database return that.

upvoted 1 times

Okkier 2 years, 1 month ago

Operational transaction systems don't use keys as their ID columns. They generally use product IDs, which refer to a business key. When we hit the dimension design for star schemas, we generally name them product keys. The second column in the table, which is product source ID, I've read as being an ID combined with the source where the record has come from, which makes it the business key. This is generally what we do for multi-sourced dimensions, i.e., different systems that can deliver products into a single dimension table. I wouldn't create a product source ID column; I would create a source column and a product ID column, which then become the combined business key. These questions are overly simplified and not providing enough clear guidance, which is simply confusing and leads to incorrect answers and inaccurate results.

upvoted 1 times

jpgsa11 2 years, 3 months ago

A Product can stop selling gets a sellendDate and the table does not receive a new version of the record. On the other hand a SCD1 can have a SK to simplify joins

upvoted 2 times

Dusica 2 years, 3 months ago

This is a far too easy question to get it wrong by the site admins

upvoted 2 times

olayvo 2 years, 3 months ago

I think: Type 1 and Surrogate keywhy need surrogate key for type 1 ?

upvoted 1 times

olayvo 2 years, 3 months ago

I think: Type 1 and Surrogate key

upvoted 1 times

SheShanD 1 year, 11 months ago

why need surrogate key for type 1 ?

upvoted 1 times

Pankaj_128 2 years, 4 months ago

Correct Answer : Type 2 SCD (start date , end date , surrogate key) Surrogate key as its an IDENTITY column.

upvoted 1 times

Alongi 2 years, 4 months ago

Type 1 because there is a time of row update, and not a column validity flag. Surrogate key, identity column is always a surrogate .

upvoted 2 times

s_unsworth 2 years, 5 months ago

The SellStartDate and SellEndDate wont be updated. They seem to be fixed columns. As you have a RowInsert and a RowUpdate it would be deemed to be a Type 1 because the row would be updated in place rather than a new row created (Type 2). Plus as the ID is an identity column it is a Surrogate Key - it has been created by the system not by the business. Business Keys are Keys that are provided by the business with business logic, such as an order number or a customer number (e.g. CUST-00001).

upvoted 2 times

Jdfer34 2 years, 6 months ago

Type 1 plus surrogate Key

upvoted 4 times

blazy001 2 years, 8 months ago

13 years in this stuff product key is a surrogate , business key comes from the ERP business app and it is a type 1 NOT 2 an insert and update column does not tell you the date from and to , hallo

upvoted 16 times

hassexat 2 years, 11 months ago

Type 2 & Surrogate Key

upvoted 2 times

AvSUN 2 years, 11 months ago

Ans 1 - TYPE 2 Ans 2 - Product key is a surrogate key (identity column) Note: Product number would be the business key if I had to pick one

upvoted 3 times

kkk5566 2 years, 11 months ago

t2 & is a surrogate key.

upvoted 2 times

Question #39 *Topic 1*

You are designing a fact table named FactPurchase in an Azure Synapse Analytics dedicated SQL pool. The table contains purchases from suppliers for a retail store. FactPurchase will contain the following columns.

FactPurchase will have 1 million rows of data added daily and will contain three years of data.

Transact-SQL queries similar to the following query will be executed daily.

```
SELECT -
SupplierKey, StockItemKey, COUNT(*)
FROM FactPurchase -
WHERE DateKey >= 20210101 -
AND DateKey <= 20210131 -
GROUP By SupplierKey, StockItemKey
```

Which table distribution will minimize query times?

Name	Data type	Nullable
PurchaseKey	Bigint	No
DateKey	Int	No
SupplierKey	Int	No
StockItemKey	Int	No
PurchaseOrderID	Int	Yes
OrderedQuantity	Int	No
OrderedOuters	Int	No
ReceivedOuters	Int	No
Package	Nvarchar(50)	No
IsOrderFinalized	Bit	No
LineageKey	Int	No

A. replicated

B. hash-distributed on PurchaseKey

C. round-robin

D. hash-distributed on DateKey

Suggested Answer: *B*

Hash-distributed tables improve query performance on large fact tables, and are the focus of this article. Round-robin tables are useful for improving loading speed. Incorrect: Not D: Do not use a date column. . All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

Community vote distribution



AugustineUba Highly Voted 5 years ago

From the documentation the answer is clear enough. B is the right answer. When choosing a distribution column, select a distribution column that: "Is not a date column. All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work."

upvoted 65 times

AugustineUba Highly Voted 5 years ago

From the documentation the answer is clear enough. B is the right answer. When choosing a distribution column, select a distribution column that: "Is not a date column. All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work."To minimize data movement, select a distribution column that: Is used in JOIN, GROUP BY, DISTINCT, OVER, and HAVING clauses. "PurchaseKey" is not used in the group byA distribution column should have high cardinality to ensure even distribution over nodes.Consider using the round-robin distribution for your table in the following scenarios: When getting started as a simple starting point since it is

the default If there is no obvious joining key If there is no good candidate column for hash distributing the table If the table does not share a common join key with other tables If the join is less significant than other joins in the query

upvoted 65 times

YipingRuan 4 years, 10 months ago

To minimize data movement, select a distribution column that: Is used in JOIN, GROUP BY, DISTINCT, OVER, and HAVING clauses. "PurchaseKey" is not used in the group byA distribution column should have high cardinality to ensure even distribution over nodes.

upvoted 8 times

YipingRuan 4 years, 10 months ago

To minimize data movement, select a distribution column that:

Is used in JOIN, GROUP BY, DISTINCT, OVER, and HAVING clauses.

"PurchaseKey" is not used in the group by

upvoted 8 times

cem_kalender 3 years, 9 months ago

A distribution column should have high cardinality to ensure even distribution over nodes.

upvoted 3 times

YipingRuan 4 years, 10 months ago

Consider using the round-robin distribution for your table in the following scenarios: When getting started as a simple starting point since it is the default If there is no obvious joining key If there is no good candidate column for hash distributing the table If the table does not share a common join key with other tables If the join is less significant than other joins in the query

upvoted 7 times

waterbender19 5 years ago

I think the answer should be D for that specific query. If you look at the datatypes, DateKey is an INT datatype not a DATE datatype.and thet statement that Fact table will be added 1 million rows daily means that each datekey value has an equal amount of rows associated with that value.But the DateKey is used in the WHERE clause.I agree, date key is int, and besides, even if it was a date, when you query a couple days then 1 million rows per distribution is not that much. So what if you are going to use only a couple distributions to do the job? Isn't it still faster than using all distributions to process all of the records to get the required date range?https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute this link says date filed , NOT a date Data type. B is correctn.b. if we look at the example query itself the date range is 31 days so we will use 31 distributions out of 60, and only process ~31 million records

upvoted 21 times

waterbender19 5 years ago

I think the answer should be D for that specific query. If you look at the datatypes, DateKey is an INT datatype not a DATE datatype.

upvoted 21 times

waterbender19 5 years ago

and thet statement that Fact table will be added 1 million rows daily means that each datekey value has an equal amount of rows associated with that value.But the DateKey is used in the WHERE clause.I agree, date key is int, and besides, even if it was a date, when you query a couple days then 1 million rows per distribution is not that much. So what if you are going to use only a couple distributions to do the job? Isn't it still faster than using all distributions to process all of the records to get the required date range?

upvoted 5 times

waterbender19 5 years ago

and thet statement that Fact table will be added 1 million rows daily means that each datekey value has an equal amount of rows associated with that value.

upvoted 5 times

Lucky_me 4 years, 7 months ago

But the DateKey is used in the WHERE clause.I agree, date key is int, and besides, even if it was a date, when you query a couple days then 1 million rows per distribution is not that much. So what if you are going to use only a couple distributions to do the job? Isn't it still faster than using all distributions to process all of the records to get the required date range?

upvoted 4 times

Lucky_me 4 years, 7 months ago

But the DateKey is used in the WHERE clause.

upvoted 4 times

kamil_k 4 years, 5 months ago

I agree, date key is int, and besides, even if it was a date, when you query a couple days then 1 million rows per distribution is not that much. So what if you are going to use only a couple distributions to do the job? Isn't it still faster than using all distributions to process all of the records to get the

required date range?

upvoted 1 times

AnandEMani 4 years, 11 months ago

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute this link says date filed , NOT a date Data type. B is correct

upvoted 7 times

kamil_k 4 years, 5 months ago

n.b. if we look at the example query itself the date range is 31 days so we will use 31 distributions out of 60, and only process ~31 million records

upvoted 2 times

Ammy_b Most Recent 1 year, 5 months ago

The Query is filtering particular date range,. so fast processing hash distribution is applied in date column.

upvoted 1 times

jay0809 1 year, 5 months ago

D, this avoids scanning irrelevant data which helps other query operations which triggers data movement to move less data as it's filtered B does not make sense at all coz everytime this will be a unique value

upvoted 1 times

IMadnan 1 year, 6 months ago

Hash distribution is generally preferred for large fact tables. Among the hash distribution options, hash-distributed on DateKey seems most beneficial because the query has a WHERE clause filtering on DateKey. Distributing on DateKey will help to localize data access for the given query pattern, minimizing data movement and potentially improving query performance. Therefore, Option D: hash-distributed on DateKey is the most likely correct answer.

upvoted 1 times

dippip123 1 year, 6 months ago

In addition to not using a Date column, the azure documentation also says: To minimize data movement, select a distribution column or set of columns that: Is not used in WHERE clauses. When a query's WHERE clause and the table's distribution columns are on the same column, the query could encounter high data skew, leading to processing load falling on only few distributions. This impacts query performance, ideally many distributions share the processing load.

upvoted 1 times

samirarian 1 year, 6 months ago

d is correct

upvoted 1 times

JustImperius 1 year, 7 months ago

Hi Community, it's B. We are not talking about partitions. We are talking about distribution and we know that its a poor choice to choose a low cardinality column like datekey for a distribution.

upvoted 3 times

nockda 1 year, 8 months ago

In the Where clause, DateKey is the one which can affect to query. Partition should be with DateKey

upvoted 1 times

nockda 1 year, 8 months ago

Answer should be D. PurchaseKey is not related with this query.

upvoted 1 times

Daniel627 1 year, 8 months ago

Even if Datekey is int, it is used in "WHERE" clause so it is not suitable as well. 1: "Is used in JOIN, GROUP BY, DISTINCT, OVER, and HAVING clauses. When two large fact tables have frequent joins, query performance improves when you distribute both tables on one of the join columns. When a table is not used in joins, consider distributing the table on a column or column set that is frequently in the GROUP BY clause." 2: "Is not used in WHERE clauses. When a query's WHERE clause and the table's distribution columns are on the same column, the query could encounter high data skew, leading to processing load falling on only few distributions. This impacts query performance, ideally many distributions share the processing load."

upvoted 1 times

f7c717f 1 year, 8 months ago

The answer is "D" since the question is referring to "three years of data" so they are asking for date partitioning

upvoted 1 times

moize 1 year, 8 months ago

D. Distribué par hachage sur DateKey Cette distribution optimise les requêtes qui filtrent par DateKey en minimisant les mouvements de données entre les nœuds.

upvoted 1 times

EmnCours 1 year, 8 months ago

Correct Answer: D

upvoted 1 times

nockda 1 year, 8 months ago

It should be D.

upvoted 1 times

nockda 1 year, 9 months ago

The answer should be D. because this query is searched by datekey

upvoted 1 times

renan_ineu 1 year, 11 months ago

To optimize parallel processing in Azure Synapse Analytics, it's important to select a distribution column or set of columns that: Has many unique values. Does not have many nulls. Is not a date column. Partitioning by DateKey does not meet these criteria. Date columns typically have limited unique values, can contain many nulls, and, by nature, are date fields. Given these constraints, DateKey is not suitable for distribution. For fact tables, which usually exceed 2 GB in size, a hash distribution is recommended. In this scenario, using PurchaseKey as the distribution key is ideal as it aligns with all the specified criteria: it has many unique values, is unlikely to have nulls, and is not a date column. This approach ensures more efficient data distribution and query performance. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

upvoted 1 times

Question #40 *Topic 1*

You are implementing a batch dataset in the Parquet format.
Data files will be produced be using Azure Data Factory and stored in Azure Data Lake Storage Gen2. The files will be consumed by an Azure Synapse Analytics serverless SQL pool.
You need to minimize storage costs for the solution.
What should you do?

A. Use Snappy compression for the files.

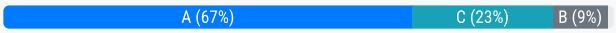
B. Use OPENROWSET to query the Parquet files.

C. Create an external table that contains a subset of columns from the Parquet files.

D. Store all data as string in the Parquet files.

Suggested Answer: *C*
An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

Community vote distribution



m2shines Highly Voted 4 years, 8 months ago
Answer should be A, because this talks about minimizing storage costs, not querying costs
upvoted 75 times

m2shines Highly Voted 4 years, 8 months ago
Answer should be A, because this talks about minimizing storage costs, not querying costsI found this comparison of compression methods, which explained that A should not be the answer. https://www.linkedin.com/pulse/comparison-compression-methods-parquet-file-format-saurav-mohapatra/ "BROTLI : This is a relatively new codec which offers very high compression ratio , but with lower compression and decompression speeds. This codec is useful when storage space is a major constraint. This technique also offers parallel processing that other methods don't."Isn't snappy a default compressionCodec for parquet in azure? https://docs.microsoft.com/en-us/azure/data-factory/format-parquetVery confused at first, after thinking about it and rereading this is what I found: It says we are implementing the batch process in parquet format, so we should think about a situation where we write the file and specify snappy compression as an argument explicitly. The phrasing is very confusing I have to say, but if you argue from a 'query externally' perspective, then B and C would yield the same benefit. Therefore, A makes the most sense and connects best with the question.
upvoted 75 times

Homer23 2 years, 5 months ago
I found this comparison of compression methods, which explained that A should not be the answer. https://www.linkedin.com/pulse/comparison-compression-methods-parquet-file-format-saurav-mohapatra/ "BROTLI : This is a relatively new codec which offers very high compression ratio , but with lower compression and decompression speeds. This codec is useful when storage space is a major constraint. This technique also offers parallel processing that other methods don't."
upvoted 2 times

assU2 4 years, 7 months ago
Isn't snappy a default compressionCodec for parquet in azure? https://docs.microsoft.com/en-us/azure/data-factory/format-parquetVery confused at first, after thinking about it and rereading this is what I found: It says we are implementing the batch process in parquet format, so we should think about a situation where we write the file and specify snappy compression as an argument explicitly. The phrasing is very confusing I have to say, but if you argue from a 'query externally' perspective, then B and C would yield the same benefit. Therefore, A makes the most sense and connects best with the question.
upvoted 24 times

assU2 4 years, 7 months ago
Isn't snappy a default compressionCodec for parquet in azure? https://docs.microsoft.com/en-us/azure/data-factory/format-parquet
upvoted 24 times

jongert 2 years, 7 months ago

Very confused at first, after thinking about it and rereading this is what I found: It says we are implementing the batch process in parquet format, so we should think about a situation where we write the file and specify snappy compression as an argument explicitly. The phrasing is very confusing I have to say, but if you argue from a 'query externally' perspective, then B and C would yield the same benefit. Therefore, A makes the most sense and connects best with the question.

upvoted 2 times

Aslam208 4 years, 8 months ago

C is the correct answer, as an external table with a subset of columns with parquet files would be cost-effective.

upvoted 23 times

Aslam208 4 years, 8 months ago

C is the correct answer, as an external table with a subset of columns with parquet files would be cost-effective.in serverless sql pool you don't create a copy of the data, so how could be cost effective?Don't forget that there is Transaction cost part of storage cost, so taking a subset of columns will lower transaction cost consequently storage cost.This is not correct. 1. External tables are are not saved in the database. (This is why they're external) 2. You're assuming that the SQL Serverless pools have a local storage. They don't -- > https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-serverless-sql-poolwell there is a possibility to create an external table and load only the required columns using openrowset in serverless sql pool to a different container in ADLS. Remember serverless sql pool does support cetas with openrowset but dedicated pool doesn't support loading data using openrowset. So basically the solution could be load the required columns using cetas using openrowset to a differnet container and delete the source data from previous container after loading the filtered data to a different container in ADLScheck this https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas. Answer C is correct

upvoted 23 times

Massy 4 years, 3 months ago

in serverless sql pool you don't create a copy of the data, so how could be cost effective?Don't forget that there is Transaction cost part of storage cost, so taking a subset of columns will lower transaction cost consequently storage cost.

upvoted 2 times

Massy 4 years, 3 months ago

in serverless sql pool you don't create a copy of the data, so how could be cost effective?

upvoted 2 times

Bro111 3 years, 8 months ago

Don't forget that there is Transaction cost part of storage cost, so taking a subset of columns will lower transaction cost consequently storage cost.

upvoted 1 times

RehanRajput 4 years, 2 months ago

This is not correct. 1. External tables are are not saved in the database. (This is why they're external) 2. You're assuming that the SQL Serverless pools have a local storage. They don't -- > https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-serverless-sql-pool

upvoted 5 times

RehanRajput 4 years, 2 months ago

This is not correct. 1. External tables are are not saved in the database. (This is why they're external) 2. You're assuming that the SQL Serverless pools have a local storage. They don't -- > https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-serverless-sql-poolwell there is a possibility to create an external table and load only the required columns using openrowset in serverless sql pool to a different container in ADLS. Remember serverless sql pool does support cetas with openrowset but dedicated pool doesn't support loading data using openrowset. So basically the solution could be load the required columns using cetas using openrowset to a differnet container and delete the source data from previous container after loading the filtered data to a different container in ADLScheck this https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas. Answer C is correct

upvoted 5 times

Aditya0891 4 years, 2 months ago

well there is a possibility to create an external table and load only the required columns using openrowset in serverless sql pool to a different container in ADLS. Remember serverless sql pool does support cetas with openrowset but dedicated pool doesn't support loading data using openrowset. So basically the solution could be load the required columns using cetas using openrowset to a differnet container and delete the source data from previous container after loading the filtered data to a different container in ADLScheck this https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas. Answer C is correct

upvoted 2 times

Aditya0891 4 years, 2 months ago

well there is a possibility to create an external table and load only the required columns using openrowset in serverless sql pool to a different container in ADLS. Remember serverless sql pool does support cetas with openrowset but dedicated pool doesn't support loading data using openrowset. So basically the solution could be load the required columns using cetas using openrowset to a differnet container and delete the source data from previous container after loading the filtered data to a different container in ADLS

upvoted 2 times

Aditya0891 4 years, 2 months ago

check this https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas. Answer C is correct

upvoted 4 times

PrasadMP Most Recent 1 year, 4 months ago

C is correct answer

upvoted 1 times

IMadnan 1 year, 6 months ago

Parquet is a columnar storage format that inherently provides compression benefits. Applying Snappy compression on top of Parquet's internal compression will further reduce the storage footprint of the files in Azure Data Lake Storage Gen2. Snappy is a well-suited compression codec for analytical workloads, offering a good balance between compression ratio and decompression speed, which is important for Azure Synapse Analytics serverless SQL pool to efficiently query the data. Options B, C, and D do not directly address minimizing storage costs for the Parquet files themselves. Option B is about query access, Option C is about query efficiency but not storage, and Option D is counterproductive to storage cost minimization.

upvoted 1 times

moize 1 year, 8 months ago

A. Utilisez la compression Snappy pour les fichiers Cette approche permet de réduire la taille des fichiers Parquet, ce qui minimise les coûts de stockage dans Azure Data Lake Storage Gen2 tout en restant compatible avec Azure Synapse Analytics.

upvoted 1 times

EmnCours 1 year, 8 months ago

https://learn.microsoft.com/en-us/azure/data-factory/supported-file-formats-and-compression-codecs-legacy

upvoted 1 times

a85becd 1 year, 11 months ago

Using Snappy compression (Option A) is specifically designed to reduce the size of Parquet files, thereby directly minimizing storage costs.

upvoted 1 times

Danweo 2 years, 1 month ago

The question is confusing but I believe it is C, because you can use CETAS to store this external table in Gen2 (this is the storage solution), from there you will query it using serverless SQL pool.

upvoted 1 times

Dusica 2 years, 3 months ago

A, B and C they are all acceptable, D is just stupid But pay attention to "You need to minimize storage costs for the solution" that means snappy parquet compresson - A is correct

upvoted 2 times

dgerok 2 years, 4 months ago

Use Snappy compression for the files is the only answer, which is about minimizing cost of storage. While one is using serverless SQL pool, the external tables are available, which are the only metadata...

upvoted 3 times

Elanche 2 years, 4 months ago

Using Snappy compression for the Parquet files helps minimize storage costs while still maintaining good compression efficiency. Snappy is a compression library that offers a good balance between compression ratio and processing speed. By compressing the data using Snappy, you can significantly reduce the amount of storage required for your dataset. Option B, using OPENROWSET to query the Parquet files, doesn't directly impact storage costs. It's a method for querying data but doesn't address storage optimization. Option C, creating an external table with a subset of columns, may help reduce query costs by minimizing the amount of data that needs to be processed during queries. However, it doesn't directly address storage costs. Option D, storing all data as strings in the Parquet files, would likely increase storage costs rather than minimize them. Storing data as strings without appropriate compression would result in larger file sizes compared to using efficient compression algorithms like Snappy.

upvoted 6 times

ankeshpatel2112 2 years, 4 months ago

A. Use Snappy compression for the files.

upvoted 2 times

Zen9nez 2 years, 5 months ago

The answer is C - Parquet has default SNAPPY compression which cannot be overwritten so why would I apply SNAPPY again?

upvoted 3 times

s_unsworth 2 years, 5 months ago

Further information required for this question. There isn't enough information to go off as to what is being asked. The initial question is in regards to storage which would result in using the snappy compression answer. If you are asking about querying the data then this should be clearly defined in

the question. If someone was to create a User Story with regards to this (As a Manager I want to store data in the data lake at the reduced cost) then you wouldn't be providing them with an External table. You would give them information on storage.

upvoted 2 times

Joanna0 2 years, 7 months ago

Snappy compression can reduce the size of Parquet files by up to 70%. This can save you a significant amount of money on storage costs.

upvoted 1 times

kkk5566 2 years, 11 months ago

using compression

upvoted 2 times

kkk5566 2 years, 12 months ago

To minimize storage costs for the solution, you should use Snappy compression for the files. Snappy is a fast and efficient data compression and decompression library that can be used to compress Parquet files. This will help reduce the size of the data files and minimize storage costs in Azure Data Lake Storage Gen2. So, the correct answer is A. Use Snappy compression for the files

upvoted 1 times

Question #41 *Topic 1*

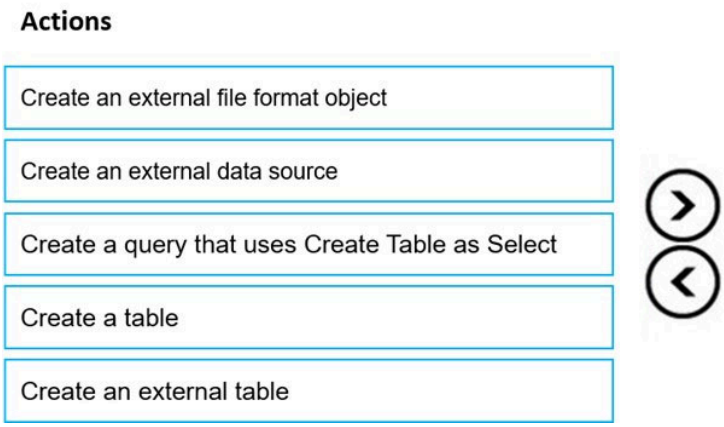
DRAG DROP -

You need to build a solution to ensure that users can query specific files in an Azure Data Lake Storage Gen2 account from an Azure Synapse Analytics serverless SQL pool.

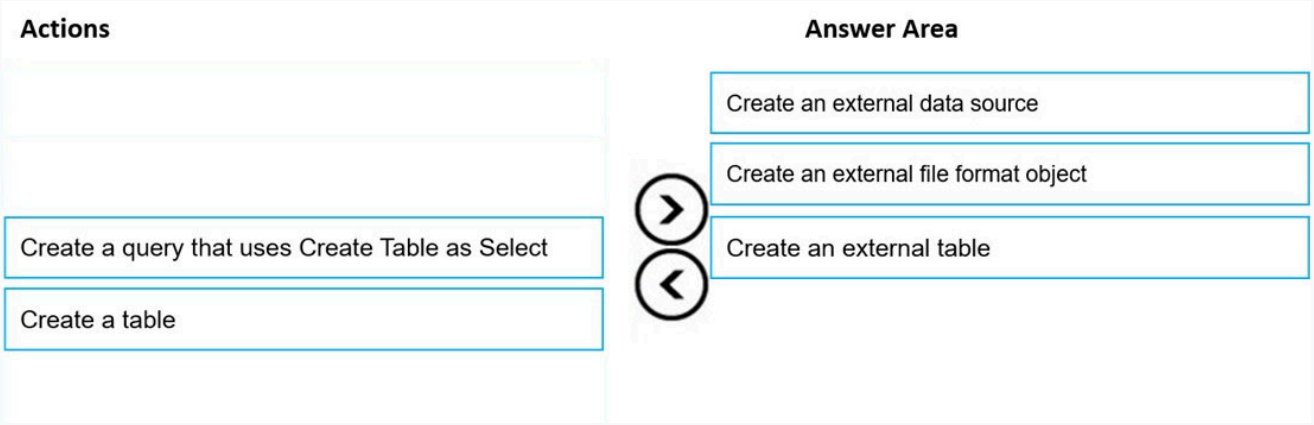
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

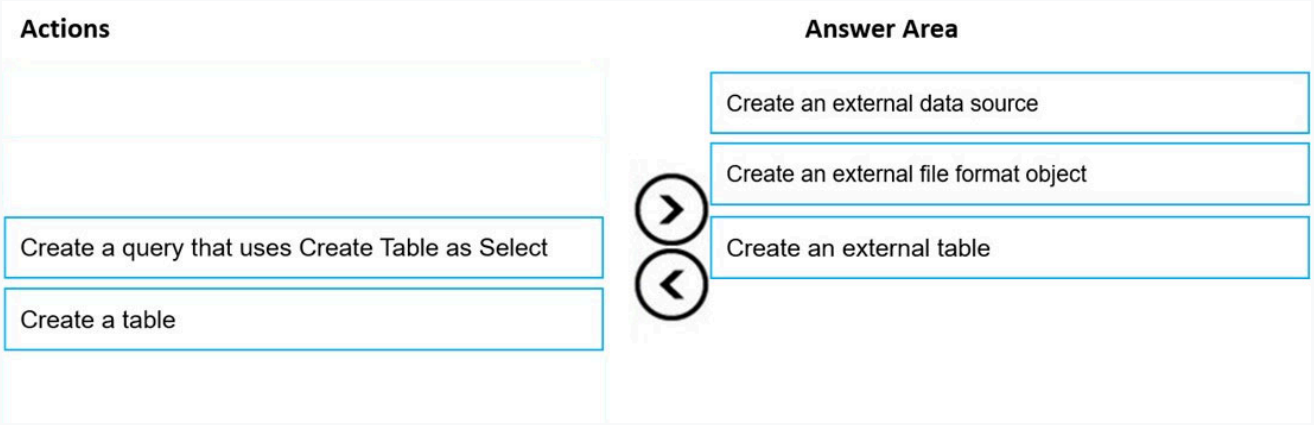
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Select and Place:



Suggested Answer:





Step 1: Create an external data source You can create external tables in Synapse SQL pools via the following steps: 1. CREATE EXTERNAL DATA SOURCE to reference an external Azure storage and specify the credential that should be used to access the storage. 2. CREATE EXTERNAL FILE FORMAT to describe format of CSV or Parquet files. 3. CREATE EXTERNAL TABLE on top of the files placed on the data source with the same file format. Step 2: Create an external file format object Creating an external file format is a prerequisite for creating an external table. Step 3: Create an external table Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

avijitd Highly Voted 4 years, 2 months ago

Looks correct answer

upvoted 41 times

vctrhugo Highly Voted 2 years, 7 months ago

1. CREATE EXTERNAL DATA SOURCE to reference an external Azure storage and specify the credential that should be used to access the storage. 2. CREATE EXTERNAL FILE FORMAT to describe format of CSV or Parquet files. 3. CREATE EXTERNAL TABLE on top of the files placed on the data source with the same file format.

upvoted 21 times

dgerok Most Recent 1 year, 10 months ago

The answer is correct

upvoted 1 times

kkk5566 2 years, 5 months ago

source, format ,external

upvoted 1 times

CrazyHorse 3 years, 5 months ago

If I run CREATE EXTERNAL FILE FORMAT before CREATE EXTERNAL DATA SOURCE, will it change something?

upvoted 6 times

CrazyHorse 3 years, 5 months ago

If I run CREATE EXTERNAL FILE FORMAT before CREATE EXTERNAL DATA SOURCE, will it change something?It doesn't really matter in which order you create a file format or a data source. I have done it in different orders before.Can anyone who knows answer this, would like to know too. Asked chatgpt and it says there's no functional difference. However, it is generally considered a best practice to create the external file format before the external data source."More than one order of answer choices is correct", so it seems to me that CREATE EXTERNAL FILE FORMAT before CREATE EXTERNAL DATA SOURCE should be accepted too.

upvoted 6 times

zekescookies 2 years, 9 months ago

It doesn't really matter in which order you create a file format or a data source. I have done it in different orders before.

upvoted 4 times

almachg 2 years, 11 months ago

Can anyone who knows answer this, would like to know too. Asked chatgpt and it says there's no functional difference. However, it is generally considered a best practice to create the external file format before the external data source."More than one order of answer choices is correct", so it seems to me that CREATE EXTERNAL FILE FORMAT before CREATE EXTERNAL DATA SOURCE should be accepted too.

upvoted 3 times

almachg 2 years, 11 months ago

Can anyone who knows answer this, would like to know too. Asked chatgpt and it says there's no functional difference. However, it is generally considered a best practice to create the external file format before the external data source.

upvoted 3 times

aemilka 2 years, 10 months ago

"More than one order of answer choices is correct", so it seems to me that CREATE EXTERNAL FILE FORMAT before CREATE EXTERNAL DATA SOURCE should be accepted too.

upvoted 5 times

Deeksha1234 3 years, 6 months ago

given answer is correct

upvoted 1 times

Dicer 3 years, 7 months ago

why creating a query that creates a table is not correct?

upvoted 1 times

Dicer 3 years, 7 months ago

why creating a query that creates a table is not correct?CeTAS also exports the query results to Blob storage or Data LAke. Its not a requirement in this question :) https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas

upvoted 1 times

VeroDon 3 years, 6 months ago

CeTAS also exports the query results to Blob storage or Data LAke. Its not a requirement in this question :) https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas

upvoted 3 times

SKN0865 3 years, 7 months ago

Correct: See Microsoft docs: You can create external tables in Synapse SQL pools via the following steps: 1) CREATE EXTERNAL DATA SOURCE to reference an external Azure storage and specify the credential that should be used to access the storage. 2) CREATE EXTERNAL FILE FORMAT to describe format of CSV or Parquet files. 3) CREATE EXTERNAL TABLE on top of the files placed on the data source with the same file format.

upvoted 5 times

Rrk07 3 years, 8 months ago

correct

upvoted 2 times

SandipSingha 3 years, 9 months ago

correct

upvoted 1 times

lotuspetall 3 years, 11 months ago

correct

upvoted 1 times

PallaviPatel 4 years ago

correct

upvoted 2 times

ANath 4 years, 1 month ago

Correct

upvoted 1 times

gf2tw 4 years, 2 months ago

Correct

upvoted 1 times

Question #42 *Topic 1*

You are designing a data mart for the human resources (HR) department at your company. The data mart will contain employee information and employee transactions.
From a source system, you have a flat extract that has the following fields:
✑ EmployeeID
FirstName -
✑ LastName
✑ Recipient
✑ GrossAmount
✑ TransactionID
✑ GovernmentID
✑ NetAmountPaid
✑ TransactionDate
You need to design a star schema data model in an Azure Synapse Analytics dedicated SQL pool for the data mart.
Which two tables should you create? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

▪

A. a dimension table for Transaction

B. a dimension table for EmployeeTransaction

C. a dimension table for Employee

D. a fact table for Employee

E. a fact table for Transaction

Suggested Answer: *C, E*
C: Dimension tables contain attribute data that might change but usually changes infrequently. For example, a customer's name and address are stored in a dimension table and updated only when the customer's profile changes. To minimize the size of a large fact table, the customer's name and address don't need to be in every row of a fact table. Instead, the fact table and the dimension table can share a customer ID. A query can join the two tables to associate a customer's profile and transactions. E: Fact tables contain quantitative data that are commonly generated in a transactional system, and then loaded into the dedicated SQL pool. For example, a retail business generates sales transactions every day, and then loads the data into a dedicated SQL pool fact table for analysis. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overview

Community vote distribution
CE (100%)

avijitd Highly Voted 4 years, 8 months ago
Correct Answer . Emp info as Dimension & trans table as fact
upvoted 24 times

moize Most Recent 1 year, 8 months ago
Table de fait et table de dimension.
upvoted 1 times

EmnCours 1 year, 8 months ago
Correct Answer: CE
upvoted 1 times

mavdplas 1 year, 8 months ago
Currect answqer - CE
upvoted 1 times

ellala 2 years, 10 months ago
Correct. A dimension is like a dictionary of information, therefore will hold the customer data such as name, address, date of birth, whatever. The fact table contains the facts. Usually numbers, usually the data that we are getting from a system from which we will create metrics later on. Therefore for

transactions.

upvoted 4 times

hassexat 2 years, 11 months ago

Correct answers

upvoted 1 times

kkk5566 2 years, 11 months ago

CE is correct

upvoted 1 times

SHENOOOO 3 years, 6 months ago

Correct Answers

upvoted 3 times

kl8585 3 years, 9 months ago

Correct answer

upvoted 3 times

Deeksha1234 4 years ago

correct answer - CE

upvoted 2 times

Remedios79 4 years, 2 months ago

correct

upvoted 2 times

hm358 4 years, 2 months ago

Correct

upvoted 3 times

SandipSingha 4 years, 3 months ago

correct

upvoted 1 times

tg2707 4 years, 3 months ago

why not fact table for employee and dim table for transactionsdo you even know what is fact or dim table? If you know you wouldn't be asking this questionIf we would know we wouldn't be here.

upvoted 1 times

tg2707 4 years, 3 months ago

why not fact table for employee and dim table for transactions

upvoted 1 times

Aditya0891 4 years, 2 months ago

do you even know what is fact or dim table? If you know you wouldn't be asking this question

upvoted 3 times

Aditya0891 4 years, 2 months ago

do you even know what is fact or dim table? If you know you wouldn't be asking this questionIf we would know we wouldn't be here.

upvoted 3 times

gabrysr1997 Highly Voted 4 years, 1 month ago

If we would know we wouldn't be here.

upvoted 13 times

Egocentric 4 years, 4 months ago

CE is correct

upvoted 1 times

NewTuanAnh 4 years, 4 months ago

CE is the correct answer

upvoted 2 times

SebK 4 years, 5 months ago

CE is correct

upvoted 2 times

Question #43 *Topic 1*

You are designing a dimension table for a data warehouse. The table will track the value of the dimension attributes over time and preserve the history of the data by adding new rows as the data changes.
Which type of slowly changing dimension (SCD) should you use?

A. Type 0

B. Type 1

C. Type 2

D. Type 3

Suggested Answer: *C*
A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members. Incorrect Answers: B: A Type 1 SCD always reflects the latest values, and when changes in source data are detected, the dimension table data is overwritten. D: A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD. Reference: https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types

Community vote distribution
C (100%)

gf2tw Highly Voted 4 years, 8 months ago
Correct
upvoted 18 times

virendrapsingh Highly Voted 4 years, 2 months ago
Kind of question that teacher leaves in paper for one free mark.
upvoted 11 times

Stuartsen Most Recent 1 year, 5 months ago
since adding a new row
upvoted 1 times

EmnCours 1 year, 8 months ago
Correct Answer: C
upvoted 1 times

hassexat 2 years, 11 months ago
Preserve history of changes... Type 2 is correct
upvoted 1 times

kkk5566 2 years, 11 months ago
SCD2 is correct
upvoted 1 times

akhil5432 3 years ago
Type 2
upvoted 1 times

GodfreyMbizo 3 years, 6 months ago
Type 2 is correct
upvoted 3 times

LokeshJ 3 years, 7 months ago
Type 2- Mainitains Row for each version on dimension data.
upvoted 3 times

dimbrici 3 years, 8 months ago

Correct

upvoted 3 times

kl8585 3 years, 9 months ago

Correct: type 1 - does not keep history of previuos values type 3 - updates all rows, not required

upvoted 4 times

Deeksha1234 4 years ago

right answer given

upvoted 1 times

NikeJDI 4 years, 1 month ago

C is right answer

upvoted 1 times

Remedios79 4 years, 2 months ago

correct

upvoted 1 times

SandipSingha 4 years, 3 months ago

correct

upvoted 1 times

AZ99979897989797897989797897 4 years, 3 months ago

Correct

upvoted 1 times

DRAG DROP -

You have data stored in thousands of CSV files in Azure Data Lake Storage Gen2. Each file has a header row followed by a properly formatted carriage return (/ r) and line feed (/n).

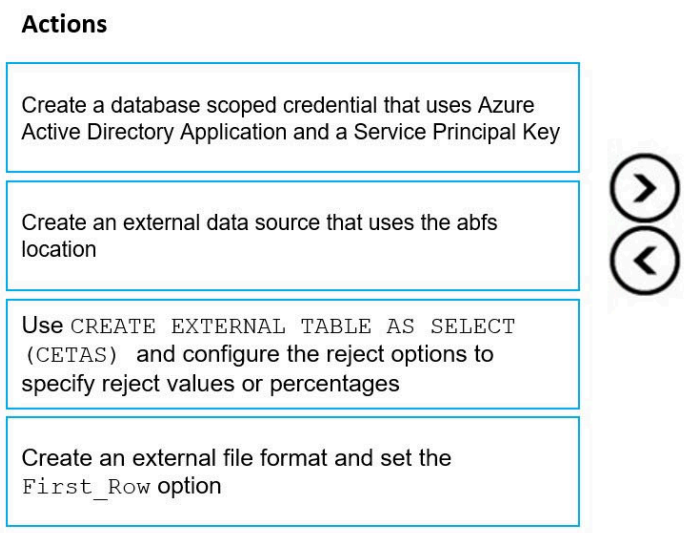
You are implementing a pattern that batch loads the files daily into a dedicated SQL pool in Azure Synapse Analytics by using PolyBase.

You need to skip the header row when you import the files into the data warehouse. Before building the loading pattern, you need to prepare the required database objects in Azure Synapse Analytics.

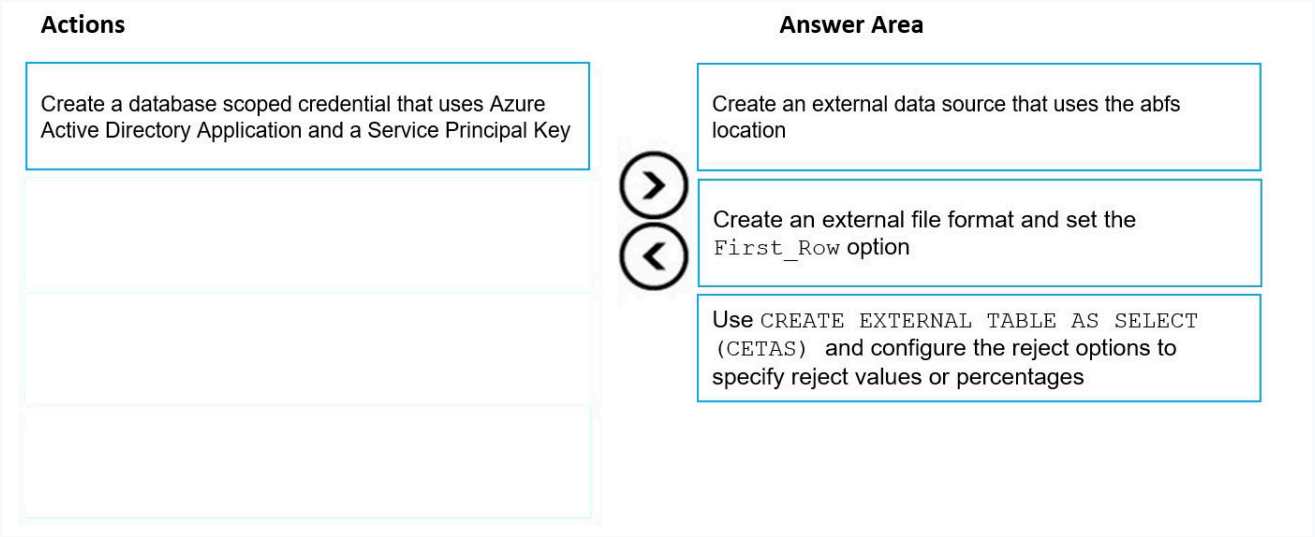
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

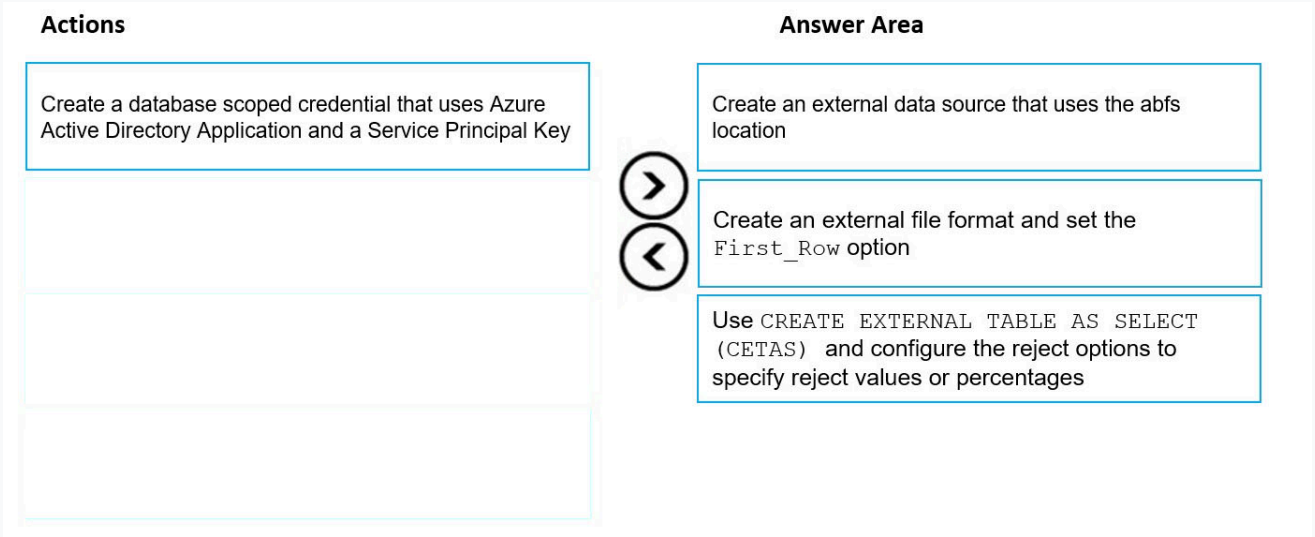
NOTE: Each correct selection is worth one point

Select and Place:



Suggested Answer:





Step 1: Create an external data source that uses the abfs location Create External Data Source to reference Azure Data Lake Store Gen 1 or 2 Step 2: Create an external file format and set the First_Row option. Create External File Format. Step 3: Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages To use PolyBase, you must create external tables to reference your external data. Use reject options. Note: REJECT options don't apply at the time this CREATE EXTERNAL TABLE AS SELECT statement is run. Instead, they're specified here so that the database can use them at a later time when it imports data from the external table. Later, when the CREATE TABLE AS SELECT statement selects data from the external table, the database will use the reject options to determine the number or percentage of rows that can fail to import before it stops the import. Reference: https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-as-select-transact-sql

sunil_smile Highly Voted 3 years, 5 months ago

1) create database scoped credentials 2) create external source 3) create file format 4) create external table (it not supports CTAS)It supports, it is a dedicated SQL pool (means not severless), reading the question: You are implementing a pattern that batch loads the files daily into a dedicated SQL pool in Azure Synapse Analytics by using PolyBase. --> provided answers are correct in my opinion.Please see this. Given answer is correct. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/load-data-overview#3-prepare-the-data-for-loading

upvoted 39 times

sunil_smile Highly Voted 3 years, 5 months ago

1) create database scoped credentials 2) create external source 3) create file format 4) create external table (it not supports CTAS)

upvoted 39 times

auwia 2 years, 7 months ago

It supports, it is a dedicated SQL pool (means not severless), reading the question: You are implementing a pattern that batch loads the files daily into a dedicated SQL pool in Azure Synapse Analytics by using PolyBase. --> provided answers are correct in my opinion.

upvoted 6 times

RMK2000 1 year, 7 months ago

Please see this. Given answer is correct. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/load-data-overview#3-prepare-the-data-for-loading

upvoted 2 times

OldSchool 3 years, 3 months ago

Because it's saying "You have data stored in thousands of CSV files in Azure Data Lake Storage Gen2" and "You are implementing a pattern that batch loads the files daily into a dedicated SQL pool in Azure Synapse Analytics by using PolyBase" assumption is that we already have database credentials, so the answer is: 1) create external source 2) create file format 3) create external tableNo, CETAS is not used for loading Azure Synapse Analytics. It's used to export data from and not to!

upvoted 19 times

OldSchool 3 years, 3 months ago

Because it's saying "You have data stored in thousands of CSV files in Azure Data Lake Storage Gen2" and "You are implementing a pattern that batch loads the files daily into a dedicated SQL pool in Azure Synapse Analytics by using PolyBase" assumption is that we already have database credentials, so the answer is: 1) create external source 2) create file format 3) create external table

upvoted 19 times

Rob77 2 years, 8 months ago

No, CETAS is not used for loading Azure Synapse Analytics. It's used to export data from and not to!

upvoted 5 times

imatheushenrique Most Recent 1 year, 5 months ago

1) create database scoped credentials 2) create external source 3) create file format

upvoted 1 times

dgerok 1 year, 10 months ago

The provided answer is correct. 1) source 2) file format 3) CETAS - this is dedicated SQL Pool. So, it is preferred. While external table is the only option for SERVERLESS SQL Pool... https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-as-select-transact-sql?view=azure-sqldw-latest&tabs=powershell#examples

upvoted 1 times

ankeshpatel2112 1 year, 10 months ago

The correct answer are : 1) create database scoped credentials 2) create external source 3) create file format Why ? : Focus on this below statement in question. Before building the loading pattern, you need to prepare the required database objects in Azure Synapse Analytics. Explanation: You just need to specify about prerequisite database object (Not Loading pattern which is about Creating External Table)

upvoted 9 times

hydmt07 2 years, 2 months ago

I think the answer (A, B, D) is very clearly explained on this page: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables

upvoted 2 times

ellala 2 years, 4 months ago

I believe the only reason why "create database scoped credential" is not a right answer is because it should be a managed identity instead of a service principal. Service principals are used for applications outside of the Azure Environment (such as SQL Server, as some comments here refer to SQL Server documentation). But since we are using Synapse Analytics environment, we use managed identities. Check the link: https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-as-select-transact-sql?view=azure-sqldw-latest&preserve-view=true&tabs=powershell#g-use-create-external-table-as-select-with-a-view-as-the-source And if you want to compare to SQL Server documentation, where indeed they use Service principal keys (which is not the situation we are given in this question): https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects?view=sql-server-ver16#create-external-tables-for-azure-data-lake-store

upvoted 2 times

pperf 2 years, 4 months ago

The provided Answer is correct check yourself, goto F section in the following link https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-as-select-transact-sql?view=azure-sqldw-latest&tabs=powershell#examples

upvoted 2 times

pperf 2 years, 4 months ago

The provided Answer is correct check yourself, goto F section in the following link https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-as-select-transact-sql?view=azure-sqldw-latest&tabs=powershell#examplesF section shows all the options given in this questions :) meaning including CETAS and data credential also

upvoted 2 times

SushilJinder 1 year, 9 months ago

F section shows all the options given in this questions :) meaning including CETAS and data credential also

upvoted 1 times

kkk5566 2 years, 5 months ago

source ,format ,external

upvoted 3 times

eladioyovera 2 years, 6 months ago

The answer is correct, - Create database scoped credential: "This step is required only for Kerberos-secured Hadoop clusters." In this case, the previous step does not apply.

upvoted 1 times

vctrhugo 2 years, 7 months ago

1. Create database scoped credential 2. Create external data source 3. Create external file format https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects?view=sql-server-ver16#create-external-tables-for-azure-blob-storage

upvoted 5 times

mamahani 2 years, 10 months ago

the given answer is correct imo; '"PolyBase loads can be run using CTAS or INSERT INTO. CTAS will minimize transaction logging and is the fastest way to load your data. '" https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool#use-polybase-to-load-and-export-data-quicklysorry, wrong link: https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects?view=sql-server-ver16 '"PolyBase can now use CETAS to create an external table and then export, in parallel, the result of a Transact-SQL SELECT statement to Azure Data Lake Storage Gen2, Azure Storage Account V2, and S3-compatible object storage.'" '"Creates an external table and then exports, in parallel, the results of a Transact-SQL SELECT statement. For Azure Synapse Analytics and Analytics Platform System, Hadoop or Azure Blob storage are supported.'"

upvoted 1 times

mamahani 2 years, 10 months ago

the given answer is correct imo; '"PolyBase loads can be run using CTAS or INSERT INTO. CTAS will minimize transaction logging and is the fastest way to load your data. '" https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool#use-polybase-to-load-and-export-data-quickly

upvoted 1 times

mamahani 2 years, 10 months ago

sorry, wrong link: https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects?view=sql-server-ver16 '"PolyBase can now use CETAS to create an external table and then export, in parallel, the result of a Transact-SQL SELECT statement to Azure Data Lake Storage Gen2, Azure Storage Account V2, and S3-compatible object storage.'" '"Creates an external table and then exports, in parallel, the results of a Transact-SQL SELECT statement. For Azure Synapse Analytics and Analytics Platform System, Hadoop or Azure Blob storage are supported.'"

upvoted 3 times

olegjdll 3 years ago

When using serverless SQL pool, CETAS is used to create an external table and export query results to Azure Storage Blob or Azure Data Lake Storage Gen2 and we need to implement a pattern that batch loads the files daily into a dedicated SQL pool in Azure Synapse Analytics, so: 1) create database scoped credentials 2) create external source 3) create file format

upvoted 6 times

Rakrah 3 years ago

In this question clearly stating that, "Before building the loading pattern, you need to prepare the required database objects" So Database objects list 1) Data Source 2) Data File Format 3) Table (need to set up skip the header row) My Answer is 1) Create external source; 2) Create File Format ; 3) Create External Table

upvoted 1 times

aws123 3 years, 1 month ago

CTAS for external table is to write the result of the query (select) in a destination folder. So the good answer for this question is :) create database scoped credentials 2) create external source 3) create file format

upvoted 3 times

rohanb1986 3 years, 1 month ago

Should be - 1) create database scoped credentials 2) create external source 3) create file format As per https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects?view=sql-server-ver16 - Check under : Create external tables for Azure Blob Storage - CTAS is not an optionyes it is an option PolyBase loads can be run using CTAS or INSERT INTO. CTAS will minimize transaction logging and is the fastest way to load your data. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool#use-polybase-to-load-and-export-data-quicklysorry, wrong link: https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects?view=sql-server-ver16 '"PolyBase can now use CETAS to create an external table and then export, in parallel, the result of a Transact-SQL SELECT statement to Azure Data Lake Storage Gen2, Azure Storage Account V2, and S3-compatible object storage.'" '"Creates an external table and then exports, in parallel,

the results of a Transact-SQL SELECT statement. For Azure Synapse Analytics and Analytics Platform System, Hadoop or Azure Blob storage are supported.'"

upvoted 2 times

rohanb1986 3 years, 1 month ago

Should be - 1) create database scoped credentials 2) create external source 3) create file format

As per https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects?view=sql-server-ver16 - Check under : Create external tables for Azure Blob Storage - CTAS is not an option

upvoted 2 times

mamahani 2 years, 10 months ago

yes it is an option PolyBase loads can be run using CTAS or INSERT INTO. CTAS will minimize transaction logging and is the fastest way to load your data. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool#use-polybase-to-load-and-export-data-quicklysorry, wrong link: https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects?view=sql-server-ver16 '"PolyBase can now use CETAS to create an external table and then export, in parallel, the result of a Transact-SQL SELECT statement to Azure Data Lake Storage Gen2, Azure Storage Account V2, and S3-compatible object storage.'" '"Creates an external table and then exports, in parallel, the results of a Transact-SQL SELECT statement. For Azure Synapse Analytics and Analytics Platform System, Hadoop or Azure Blob storage are supported.'"

upvoted 1 times

mamahani 2 years, 10 months ago

yes it is an option PolyBase loads can be run using CTAS or INSERT INTO. CTAS will minimize transaction logging and is the fastest way to load your data. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool#use-polybase-to-load-and-export-data-quickly

upvoted 1 times

Bro111 3 years, 2 months ago

What is azure active directory application? is it managed identity?

upvoted 2 times

HOTSPOT -

You are building an Azure Synapse Analytics dedicated SQL pool that will contain a fact table for transactions from the first half of the year 2020.

You need to ensure that the table meets the following requirements:

- ✑ Minimizes the processing time to delete data that is older than 10 years
- ✑ Minimizes the I/O for queries that use year-to-date values

How should you complete the Transact-SQL statement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE TABLE [dbo].[FactTransaction]
(
      [TransactionTypeID]    int      NOT NULL
,     [TransactionDateID]    int      NOT NULL
,     [CustomerID]           int      NOT NULL
,     [RecipientID]          int      NOT NULL
,     [Amount]               money    NOT NU::
)
WITH
(
```

[▼]
- CLUSTERED COLUMNSTORE INDEX
- DISTRIBUTION
- PARTITION
- TRUNCATE_TARGET

```
(
```

[▼]
- [TransactionDateID]
- [TransactionDateID], [TransactionTypeID]
- HASH([TransactionTypeID])
- ROUND_ROBIN

```
RANGE RIGHT FOR VALUES
(20200101,20200201,20200301,20200401,20200501,20200601)
```

Suggested Answer:

Answer Area

```
CREATE TABLE [dbo].[FactTransaction]
(
      [TransactionTypeID]    int      NOT NULL
,     [TransactionDateID]    int      NOT NULL
,     [CustomerID]           int      NOT NULL
,     [RecipientID]          int      NOT NULL
,     [Amount]               money    NOT NU::
)
WITH
(
```

- CLUSTERED COLUMNSTORE INDEX
- DISTRIBUTION
- **PARTITION**
- TRUNCATE_TARGET

```
    (
```

- **[TransactionDateID]**
- [TransactionDateID], [TransactionTypeID]
- HASH([TransactionTypeID])
- ROUND_ROBIN

```
RANGE RIGHT FOR VALUES
(20200101,20200201,20200301,20200401,20200501,20200601)
```

Answer Area

```
CREATE TABLE [dbo].[FactTransaction]
(
      [TransactionTypeID]    int      NOT NULL
,     [TransactionDateID]    int      NOT NULL
,     [CustomerID]           int      NOT NULL
,     [RecipientID]          int      NOT NULL
,     [Amount]               money    NOT NU::
)
WITH
(
```

- CLUSTERED COLUMNSTORE INDEX
- DISTRIBUTION
- **PARTITION**
- TRUNCATE_TARGET

```
    (
```

- **[TransactionDateID]**
- [TransactionDateID], [TransactionTypeID]
- HASH([TransactionTypeID])
- ROUND_ROBIN

```
RANGE RIGHT FOR VALUES
(20200101,20200201,20200301,20200401,20200501,20200601)
```

Box 1: PARTITION - RANGE RIGHT FOR VALUES is used with PARTITION. Part 2: [TransactionDateID] Partition on the date column. Example: Creating a RANGE RIGHT partition function on a datetime column The following partition function partitions a table or index into 12 partitions, one for each month of a year's worth of values in a datetime column. CREATE PARTITION FUNCTION [myDateRangePF1] (datetime) AS RANGE RIGHT FOR VALUES ('20030201', '20030301', '20030401', '20030501', '20030601', '20030701', '20030801', '20030901', '20031001', '20031101', '20031201'); Reference: https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql

gf2tw Highly Voted 3 years, 8 months ago

Correct

upvoted 25 times

OldSchool Highly Voted 2 years, 9 months ago

Correct answer, giveaway is "RANGE RIGHT"

upvoted 7 times

M_Anas_007 1 year, 10 months ago

Have a small doubt. We are creating partition on what field? Shouldn't it be the columnstore index.

upvoted 1 times

M_Anas_007 1 year, 10 months ago

Have a small doubt. We are creating partition on what field? Shouldn't it be the columnstore index.on "TransactionDateId" field, which is the second answer

upvoted 1 times

hypersam Most Recent 1 year, 7 months ago

on "TransactionDateId" field, which is the second answer

upvoted 1 times

kkk5566 1 year, 11 months ago

Partition

upvoted 1 times

GodfreyMbizo 2 years, 6 months ago

correct answer given

upvoted 5 times

Deeksha1234 3 years ago

ans is correct

upvoted 3 times

NikeJDI 3 years, 1 month ago

I can see Keyword "Range right for values" pointing to "Partition", then "TransactionDateID" is the column on which partition needs to be done, rather than the TransactionID.

upvoted 4 times

gabdu 3 years, 3 months ago

How are we ensuring "Minimizes the processing time to delete data that is older than 10 years"?while deleting we can use switch partition. It is efficient than delete statement, so partition by date column in good but the question says the TransactionDate as int field which is wrong. It should be date typeit's a date_ID and it's correct it will be linked to a date table via the date_ID to get the date

upvoted 2 times

gabdu 3 years, 3 months ago

How are we ensuring "Minimizes the processing time to delete data that is older than 10 years"?

upvoted 2 times

Aditya0891 3 years, 2 months ago

while deleting we can use switch partition. It is efficient than delete statement, so partition by date column in good but the question says the TransactionDate as int field which is wrong. It should be date type

upvoted 4 times

Aditya0891 3 years, 2 months ago

while deleting we can use switch partition. It is efficient than delete statement, so partition by date column in good but the question says the TransactionDate as int field which is wrong. It should be date typeit's a date_ID and it's correct it will be linked to a date table via the date_ID to get the date

upvoted 4 times

allagowf 2 years, 10 months ago

it's a date_ID and it's correct it will be linked to a date table via the date_ID to get the date

upvoted 3 times

wwdba 3 years, 5 months ago

Correct

upvoted 2 times

PallaviPatel 3 years, 6 months ago

correct

upvoted 2 times

saupats 3 years, 7 months ago

correct

upvoted 2 times

Question #46 *Topic 1*

You are performing exploratory analysis of the bus fare data in an Azure Data Lake Storage Gen2 account by using an Azure Synapse Analytics serverless SQL pool.
You execute the Transact-SQL query shown in the following exhibit.
What do the query results include?

```
SELECT
    payment_type,
    SUM(fare_amount) AS fare_total
FROM OPENROWSET (
        BULK 'csv/busfare/tripdata_2020*.csv',
        DATA_SOURCE = 'BusData',
        FORMAT = 'CSV', PARSER_VERSION = '2.0',
        FIRSTROW = 2
    )
    WITH (
        payment_type INT 10,
        fare_amount FLOAT 11
    ) AS nyc
GROUP BY payment_type
ORDER BY payment_type;
```

A. Only CSV files in the tripdata_2020 subfolder.

B. All files that have file names that beginning with "tripdata_2020".

C. All CSV files that have file names that contain "tripdata_2020".

D. Only CSV that have file names that beginning with "tripdata_2020".

Suggested Answer: *D*
Official Answer: D

Community vote distribution
D (100%)

gf2tw Highly Voted 4 years, 8 months ago
CorrectJust a tricky question to think about: Is the csv/bufare/tripdata_2020_something/something/something.csv file will be loaded? I think YES :) But it is not matched with answer A, because not ONLY the tripdata_2020 folder will be loaded.And therefore the D is not exactly true.
upvoted 19 times

gf2tw Highly Voted 4 years, 8 months ago
Correct
upvoted 19 times

Gikan 2 years, 6 months ago
Just a tricky question to think about: Is the csv/bufare/tripdata_2020_something/something/something.csv file will be loaded? I think YES :) But it is not matched with answer A, because not ONLY the tripdata_2020 folder will be loaded.And therefore the D is not exactly true.
upvoted 1 times

Gikan 2 years, 6 months ago
Just a tricky question to think about: Is the csv/bufare/tripdata_2020_something/something/something.csv file will be loaded? I think YES :) But it is not matched with answer A, because not ONLY the tripdata_2020 folder will be loaded.
upvoted 1 times

Gikan 2 years, 6 months ago
And therefore the D is not exactly true.
upvoted 1 times

akshaynag95 3 years, 6 months ago

D is the correct Answer

upvoted 5 times

EmnCours Most Recent 1 year, 8 months ago

Correct Answer: D

upvoted 1 times

hassexat 2 years, 11 months ago

Correct is D

upvoted 1 times

kkk5566 2 years, 11 months ago

D is correct

upvoted 1 times

henryphchan 3 years, 3 months ago

correct

upvoted 3 times

SHENOOOO 3 years, 6 months ago

D is the correct Answer

upvoted 3 times

panda_azzurro 3 years, 6 months ago

Selected Answer: D

Sorry but I don't understand. File or Directory that start with "tripdata_2020" can selected. /tripdata_2020/a.csv /tripdata_2020_a_b.csv /tripdata_2020/2020/1/1/myfile.csv So question is very not clear. D question is partially correct

upvoted 2 times

panda_azzurro 3 years, 6 months ago

Sorry but I don't understand. File or Directory that start with "tripdata_2020" can selected. /tripdata_2020/a.csv /tripdata_2020_a_b.csv /tripdata_2020/2020/1/1/myfile.csv So question is very not clear. D question is partially correctOnly file, there is no / after 2020 in the exercise.

upvoted 2 times

Anton2020 3 years, 6 months ago

Only file, there is no / after 2020 in the exercise.

upvoted 2 times

Stokstaartje420 3 years, 6 months ago

Answer D is correct. Would be great if it was also correct grammatically.

upvoted 4 times

kl8585 3 years, 9 months ago

Correct

upvoted 1 times

GauravPurandare 4 years ago

CSV that have file names that beginning with "tripdata_2020" .

upvoted 2 times

Deeksha1234 4 years ago

D is correct

upvoted 1 times

objecto 4 years, 2 months ago

Correct

upvoted 2 times

Rrk07 4 years, 2 months ago

D is correct

upvoted 1 times

Egocentric 4 years, 4 months ago

on this one you need to pay attention to wording

upvoted 3 times

jskibick 4 years, 4 months ago

D all good

upvoted 1 times

sarapaisley 4 years, 4 months ago

D is correct

upvoted 1 times

DRAG DROP -

You use PySpark in Azure Databricks to parse the following JSON input.

You need to output the data in the following tabular format.

How should you complete the PySpark code? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the spit bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

```
{
    "persons":[
        {
            "name":"Keith",
            "age":30,
            "dogs":["Fido", "Fluffy"]
        },
        {
            "name":"Donna",
            "age":46,
            "dogs":["Spot"]
        }
    ]
}
```

owner	age	dog
Keith	30	Fido
Keith	30	Fluffy
Donna	46	Spot

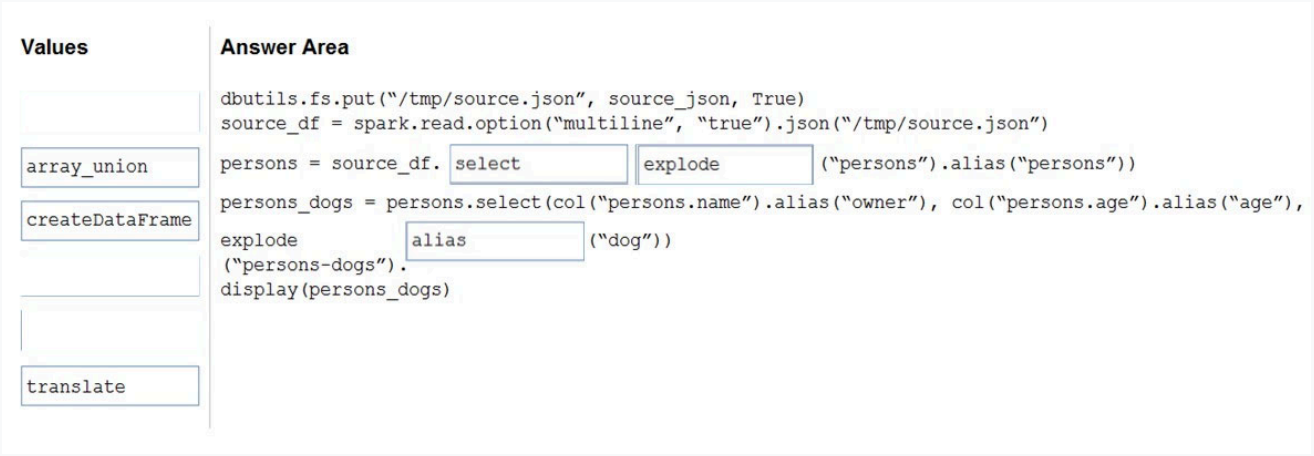
Values

- alias
- array_union
- createDataFrame
- explode
- select
- translate

Answer Area

```
dbutils.fs.put("/tmp/source.json", source_json, True)
source_df = spark.read.option("multiline", "true").json("/tmp/source.json")
persons = source_df. [Value] [Value] ("persons").alias("persons"))
persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"),
explode [Value] ("dog"))
("persons-dogs").
display(persons_dogs)
```

Suggested Answer:





Box 1: select -

Box 2: explode -

Bop 3: alias - pyspark.sql.Column.alias returns this column aliased with a new name or names (in the case of expressions that return more than one column, such as explode). Reference: https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.sql.Column.alias.html https://docs.microsoft.com/en-us/azure/databricks/sql/language-manual/functions/explode

kb8bo Highly Voted 3 years, 2 months ago

The final line with the blank looks incorrect... surely it should be: explode("persons.dogs").alias("dog")) (Assuming this, the answer is correct, otherwise I don't think it makes any sense).

upvoted 24 times

urassi Highly Voted 2 years, 5 months ago

ah "persons".alias("persons") what a fun and useful and nice alias

upvoted 18 times

urassi 2 years, 5 months ago

ah "persons".alias("persons") what a fun and useful and nice aliasactually if you don't use alias after explode("persons"), the column name would be "col" so it's mandatory here.

upvoted 18 times

hypersam Most Recent 1 year, 7 months ago

actually if you don't use alias after explode("persons"), the column name would be "col" so it's mandatory here.

upvoted 2 times

kkk5566 1 year, 11 months ago

syntax is correct

upvoted 2 times

esaade 2 years, 5 months ago

dbutils.fs.put("/tmp/source.json", source_json, True) source_df = spark.read.option("multiline", "true").json("/tmp/source.json") persons = source_df.select(explode("persons").alias("persons")) persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"), explode(col("persons.dog")).alias("dog_name")) persons_dogs.display()

upvoted 12 times

Deeksha1234 3 years ago

Correct

upvoted 4 times

Dicer 3 years ago

Correct, but last .alias("dog") is quite unnecessary because the column name is alredy 'dog'. I guess that is for safety measurement.

upvoted 5 times

Dicer 3 years ago

Correct, but last .alias("dog") is quite unnecessary because the column name is alredy 'dog'. I guess that is for safety measurement.The column name in the json is dogs, not dog

upvoted 5 times

Anton2020 2 years, 5 months ago

The column name in the json is dogs, not dog

upvoted 3 times

galacaw 3 years, 3 months ago

Correct

upvoted 4 times

HOTSPOT -

You are designing an application that will store petabytes of medical imaging data.

When the data is first created, the data will be accessed frequently during the first week. After one month, the data must be accessible within 30 seconds, but files will be accessed infrequently. After one year, the data will be accessed infrequently but must be accessible within five minutes.

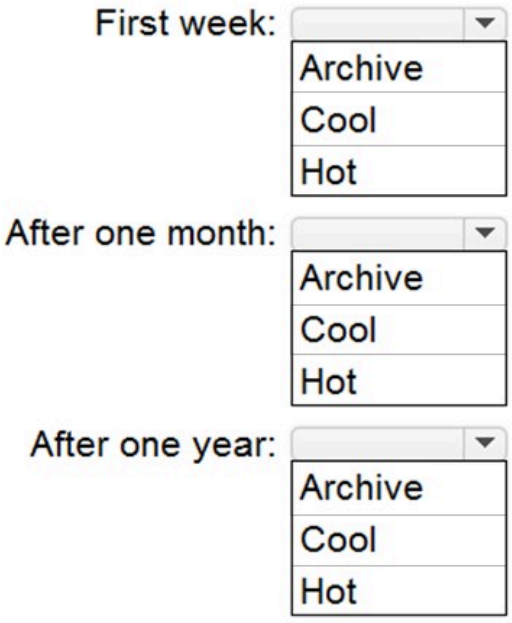
You need to select a storage strategy for the data. The solution must minimize costs.

Which storage tier should you use for each time frame? To answer, select the appropriate options in the answer area.

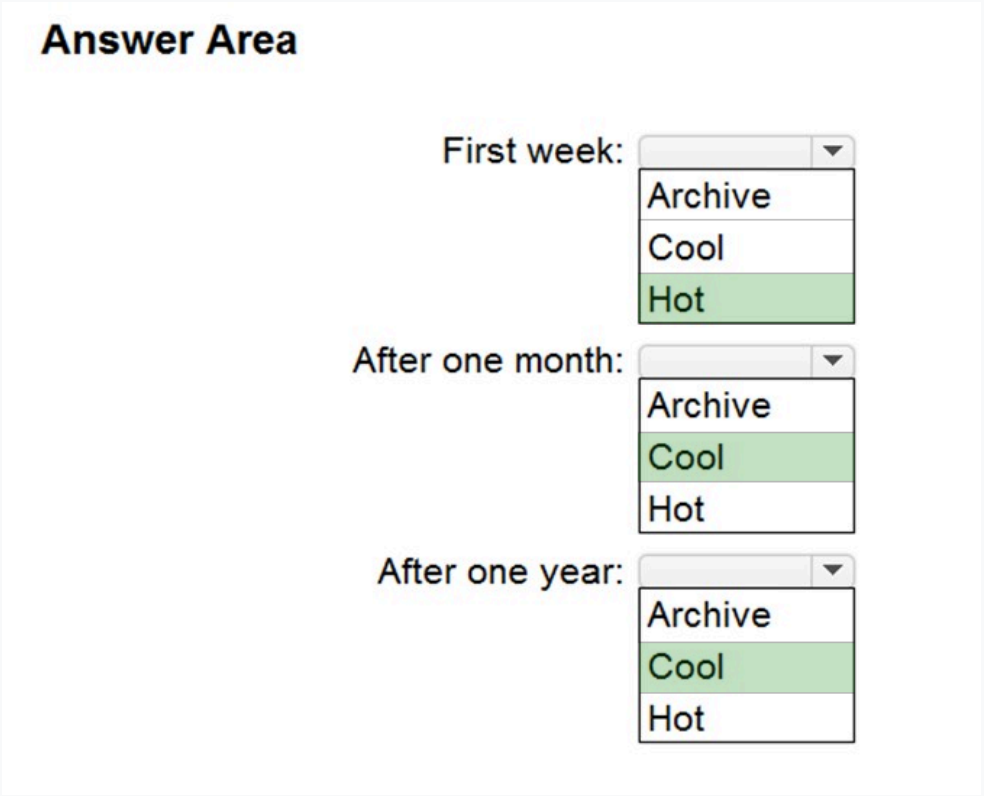
NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

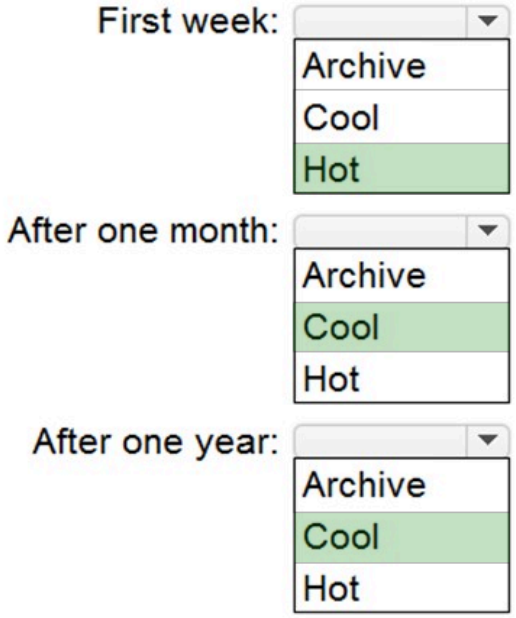


Suggested Answer:



	Premium performance	Hot tier	Cool tier	Archive tier
Availability	99.9%	99.9%	99%	Offline
Availability (RA-GRS reads)	N/A	99.99%	99.9%	Offline
Usage charges	Higher storage costs, lower access, and transaction cost	Higher storage costs, lower access, and transaction costs	Lower storage costs, higher access, and transaction costs	Lowest storage costs, highest access, and transaction costs
Minimum object size	N/A	N/A	N/A	N/A
Minimum storage duration	N/A	N/A	30 days[1]	180 days
Latency (Time to first byte)	Single-digit milliseconds	milliseconds	milliseconds	hours[2]

Answer Area



Box 1: Hot - Hot tier - An online tier optimized for storing data that is accessed or modified frequently. The Hot tier has the highest storage costs, but the lowest access costs.

Box 2: Cool - Cool tier - An online tier optimized for storing data that is infrequently accessed or modified. Data in the Cool tier should be stored for a minimum of 30 days. The Cool tier has lower storage costs and higher access costs compared to the Hot tier.

Box 3: Cool - Not Archive tier - An offline tier optimized for storing data that is rarely accessed, and that has flexible latency requirements, on the order of hours. Data in the Archive tier should be stored for a minimum of 180 days.

Reference: https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview https://www.altaro.com/hyper-v/azure-archive-storage/

Deeksha1234 Highly Voted 3 years ago
right , Hot-cool-cool
upvoted 23 times

Andy91 3 years, 3 months ago
Correct answer! Hot, Cool, Cool
upvoted 10 times

hassexat Most Recent 1 year, 11 months ago
Hot Cool Cool
upvoted 3 times

kkk5566 1 year, 11 months ago
hot,cool,cool
upvoted 3 times

rocky48 2 years, 3 months ago
Hot, Cool, Cool
upvoted 4 times

Fishy_Marcy 3 years, 1 month ago
Isn't the cool storage enough for initial requirements and also required for the other options? So shouldn't the answer be cool in all places? that would be coolReads in Cool Tier are more expensive than in Hot Tier. Since the data will be accessed frequently in the first week, it makes sense to store it in hot tier to minimize costs.
upvoted 2 times

Fishy_Marcy 3 years, 1 month ago

Isn't the cool storage enough for initial requirements and also required for the other options? So shouldn't the answer be cool in all places? that would be cool

upvoted 2 times

lucasramos 3 years, 1 month ago

Reads in Cool Tier are more expensive than in Hot Tier. Since the data will be accessed frequently in the first week, it makes sense to store it in hot tier to minimize costs.

upvoted 6 times

nefarious_smalls 3 years, 3 months ago

Why would it not be be Hot Cool and ArchiveAfter one year, the data will be accessed infrequently but must be accessible within five minutes.hydration on the archive tier is hours, not minutes. hence its hot, cool, cool."After one year, the data will be accessed infrequently but must be accessible within five minutes" The latency for the first bytes, is "hours" for the archive. so because they want to be able to access the data within 5 min, you need to place it in "cool" So the answer is correct.

upvoted 2 times

nefarious_smalls 3 years, 3 months ago

Why would it not be be Hot Cool and Archive

upvoted 2 times

SandipSingha 3 years, 3 months ago

After one year, the data will be accessed infrequently but must be accessible within five minutes.

upvoted 5 times

AnonymousJhb 2 years, 3 months ago

hydration on the archive tier is hours, not minutes. hence its hot, cool, cool.

upvoted 2 times

Guincimund Highly Voted 3 years, 3 months ago

"After one year, the data will be accessed infrequently but must be accessible within five minutes" The latency for the first bytes, is "hours" for the archive. so because they want to be able to access the data within 5 min, you need to place it in "cool" So the answer is correct.

upvoted 23 times

nefarious_smalls 3 years, 3 months ago

I dont know

upvoted 1 times

Question #49 *Topic 1*

You have an Azure Synapse Analytics Apache Spark pool named Pool1.
You plan to load JSON files from an Azure Data Lake Storage Gen2 container into the tables in Pool1. The structure and data types vary by file.
You need to load the files into the tables. The solution must maintain the source data types.
What should you do?

A. Use a Conditional Split transformation in an Azure Synapse data flow.

B. Use a Get Metadata activity in Azure Data Factory.

C. Load the data by using the OPENROWSET Transact-SQL command in an Azure Synapse Analytics serverless SQL pool.

D. Load the data by using PySpark.

Suggested Answer: *C*
Serverless SQL pool can automatically synchronize metadata from Apache Spark. A serverless SQL pool database will be created for each database existing in serverless Apache Spark pools. Serverless SQL pool enables you to query data in your data lake. It offers a T-SQL query surface area that accommodates semi-structured and unstructured data queries. To support a smooth experience for in place querying of data that's located in Azure Storage files, serverless SQL pool uses the OPENROWSET function with additional capabilities. The easiest way to see to the content of your JSON file is to provide the file URL to the OPENROWSET function, specify csv FORMAT. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-data-storage

Community vote distribution
D (90%) C (10%)

galacaw Highly Voted 4 years, 3 months ago
Should be D, it's about Apache Spark pool, not serverless SQL pool.
upvoted 40 times

Joanna0 Highly Voted 2 years, 7 months ago
If your JSON files have a consistent structure and data types, then OPENROWSET is a good option. However, if your JSON files have a varying structure and data types, then PySpark is a better option.
upvoted 6 times

EmnCours Most Recent 1 year, 8 months ago
Correct Answer: D
upvoted 1 times

vaibhavs120 2 years ago
The answer is C because with external table you can load the data with solution must maintain the source data types.
upvoted 1 times

e56bb91 2 years, 1 month ago
ChatGPT 4o Using PySpark in an Apache Spark pool within Azure Synapse Analytics is the most flexible and powerful way to handle JSON files with varying structures and data types. PySpark can infer schema and handle complex data transformations, making it well-suited for loading heterogeneous JSON data into tables while preserving the original data types.
upvoted 1 times

Okkier 2 years, 1 month ago
When loading data into an Apache Spark pool, especially when dealing with inconsistent file structures, PySpark (the Python API for Spark) is generally the better choice over OpenRowset. This is because PySpark offers greater flexibility, better performance, and more robust handling of varied and complex data structures.
upvoted 1 times

kldakdlsa 2 years, 1 month ago
should be D
upvoted 1 times

ellala 2 years, 10 months ago
We have an "Azure Synapse Analytics Apache Spark pool" therefore, we use Spark. There is no information about a serverless SQL Pool
upvoted 2 times

kkk5566 2 years, 11 months ago

Should be D

upvoted 2 times

vctrhugo 3 years, 1 month ago

PySpark provides a powerful and flexible programming interface for processing and loading data in Azure Synapse Analytics Apache Spark pools. With PySpark, you can leverage its JSON reader capabilities to infer the schema and maintain the source data types during the loading process.

upvoted 3 times

vctrhugo 3 years, 2 months ago

To load JSON files from an Azure Data Lake Storage Gen2 container into tables in an Azure Synapse Analytics Apache Spark pool, you can use PySpark. PySpark provides a flexible and powerful framework for working with big data in Apache Spark. Therefore, the correct answer is: D. Load the data by using PySpark. You can use PySpark to read the JSON files from Azure Data Lake Storage Gen2, infer the schema, and load the data into tables in the Spark pool while maintaining the source data types. PySpark provides various functions and methods to handle JSON data and perform transformations as needed before loading it into tables.

upvoted 4 times

janaki 3 years, 2 months ago

Option D: Load the data by using PySpark

upvoted 1 times

henryphchan 3 years, 3 months ago

The question stated that "You have an Azure Synapse Analytics Apache Spark pool named Pool1.", so this question is about Spark pool

upvoted 1 times

Victor_Kings 3 years, 4 months ago

Selected Answer: C

As stated by Microsoft, "Serverless SQL pool can automatically synchronize metadata from Apache Spark. A serverless SQL pool database will be created for each database existing in serverless Apache Spark pools.". So even though the files in Azure Storage were created with Apache Spark, you can still query them using OPENROWSET with a serverless SQL Pool

https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-storage-files-spark-tables

upvoted 4 times

Victor_Kings 3 years, 4 months ago

As stated by Microsoft, "Serverless SQL pool can automatically synchronize metadata from Apache Spark. A serverless SQL pool database will be created for each database existing in serverless Apache Spark pools.". So even though the files in Azure Storage were created with Apache Spark, you can still query them using OPENROWSET with a serverless SQL Pool https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-storage-files-spark-tablesAs the question states that "You need to load the files into the tables" , through serverless sql pool we cannot load data. so the answer should be DWe are dealing with varying JSON. There is nothing about this option by the link you've provided. The correct answer is D...

upvoted 4 times

Tejashu 2 years, 9 months ago

As the question states that "You need to load the files into the tables" , through serverless sql pool we cannot load data. so the answer should be D

upvoted 2 times

dgerok 2 years, 4 months ago

We are dealing with varying JSON. There is nothing about this option by the link you've provided. The correct answer is D...

upvoted 1 times

esaade 3 years, 5 months ago

To load JSON files from an Azure Data Lake Storage Gen2 container into the tables in an Apache Spark pool in Azure Synapse Analytics while maintaining the source data types, you should use PySpark.

upvoted 3 times

haidebelognime 3 years, 5 months ago

PySpark is the Python API for Apache Spark, which is a distributed computing framework that can handle large-scale data processing.

upvoted 2 times

brzhanyu 3 years, 8 months ago

Should be D, it's about Apache Spark pool, not serverless SQL pool.

upvoted 2 times

Question #50 *Topic 1*

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier. Workspace1 contains an all-purpose cluster named cluster1.
You need to reduce the time it takes for cluster1 to start and scale up. The solution must minimize costs.
What should you do first?

A. Configure a global init script for workspace1.

B. Create a cluster policy in workspace1.

C. Upgrade workspace1 to the Premium pricing tier.

D. Create a pool in workspace1.

Suggested Answer: *D*
You can use Databricks Pools to Speed up your Data Pipelines and Scale Clusters Quickly. Databricks Pools, a managed cache of virtual machine instances that enables clusters to start and scale 4 times faster. Reference: https://databricks.com/blog/2019/11/11/databricks-pools-speed-up-data-pipelines.html

Community vote distribution
D (73%) C (27%)

hkay Highly Voted 4 years, 2 months ago
Answer D is correct. Azure Databricks pools reduce cluster start and auto-scaling times by maintaining a set of idle, ready-to-use instances.
upvoted 13 times

kim32 Highly Voted 3 years, 3 months ago
D is accurate answer and this link show this info explicty https://learn.microsoft.com/en-us/azure/databricks/clusters/cluster-config-best-practices
upvoted 9 times

EmnCours Most Recent 1 year, 8 months ago
Correct Answer: D
upvoted 1 times

ff5037f 1 year, 9 months ago
both c and d are correct, but option c is expensive. as question clearly states that solution should be cheaper
upvoted 1 times

Alongi 2 years, 3 months ago
Selected Answer: C
I found this question on my exam yesterday, 30/04/2024, and I put C. I passed the exam with a high score, but I'm not sure if the answer is correct.
upvoted 2 times

Alongi 2 years, 3 months ago
I found this question on my exam yesterday, 30/04/2024, and I put C. I passed the exam with a high score, but I'm not sure if the answer is correct.The answer C seems to be false, given that we have to minimize the costs and the prenium tier has a higher cost
upvoted 2 times

Slimshaddy 2 years, 3 months ago
The answer C seems to be false, given that we have to minimize the costs and the prenium tier has a higher cost
upvoted 1 times

dgerok 2 years, 4 months ago
The solution must minimize costs, so upgrade to Premium is not an option.
upvoted 1 times

Ajjjiii 2 years, 6 months ago
This came in my today's exam . Answer D is correct .
upvoted 4 times

Joanna0 2 years, 7 months ago
The best solution to reduce the time it takes for cluster1 to start and scale up while minimizing costs is to create a pool in workspace1. A pool is a group of preconfigured clusters that share resources and can be quickly provisioned and scaled up or down as needed. This can significantly reduce the time it takes for cluster1 to start and scale up, as the resources are already provisioned and ready to use.

upvoted 4 times

positivitypeople 2 years, 8 months ago

Got this question today on the exam. Selected Upgrade to Premium tier, was incorrect of course

upvoted 3 times

kkk5566 2 years, 11 months ago

D is correct

upvoted 1 times

Abdullah77 3 years ago

C is the answer

upvoted 1 times

auwia 3 years, 2 months ago

To reduce the time it takes for cluster1 to start and scale up in Azure Databricks workspace, the first step you should take is to create a pool in workspace1.

upvoted 1 times

aga444 3 years, 2 months ago

The correct answer in this scenario would be B. Create a cluster policy in workspace1. Creating a cluster policy allows you to define a set of rules and configurations that apply to all clusters within the workspace. By creating a cluster policy, you can optimize the cluster startup and scaling behavior. With a cluster policy, you can specify settings such as the minimum and maximum number of worker nodes, the idle timeout duration, and the auto-termination behavior. By tuning these settings, you can reduce the time it takes for cluster1 to start and scale up, ensuring that resources are efficiently utilized. Configuring a global init script (option A) or creating a pool (option D) may have other benefits, but they are not directly related to reducing the time it takes for cluster1 to start and scale up. Upgrading the workspace to the Premium pricing tier (option C) may offer additional features but is not necessary to address the specific requirement of minimizing cluster startup and scaling time while minimizing costs.

upvoted 2 times

aga444 3 years, 2 months ago

How can it be C), as the question clearly states 'minimize costs'

upvoted 1 times

janaki 3 years, 2 months ago

I think it's B. Create a cluster policy in workspace1 Cluster policies in Azure Databricks allow you to define rules and configurations for cluster creation and termination. By creating a cluster policy, you can optimize the cluster start time and scale-up process based on your specific requirements.

upvoted 1 times

mamahani 3 years, 3 months ago

D is correct answer

upvoted 1 times

rocky48 3 years, 3 months ago

Correct Answer: D

upvoted 1 times

HOTSPOT -

You are building an Azure Stream Analytics job that queries reference data from a product catalog file. The file is updated daily.

The reference data input details for the file are shown in the Input exhibit. (Click the Input tab.)

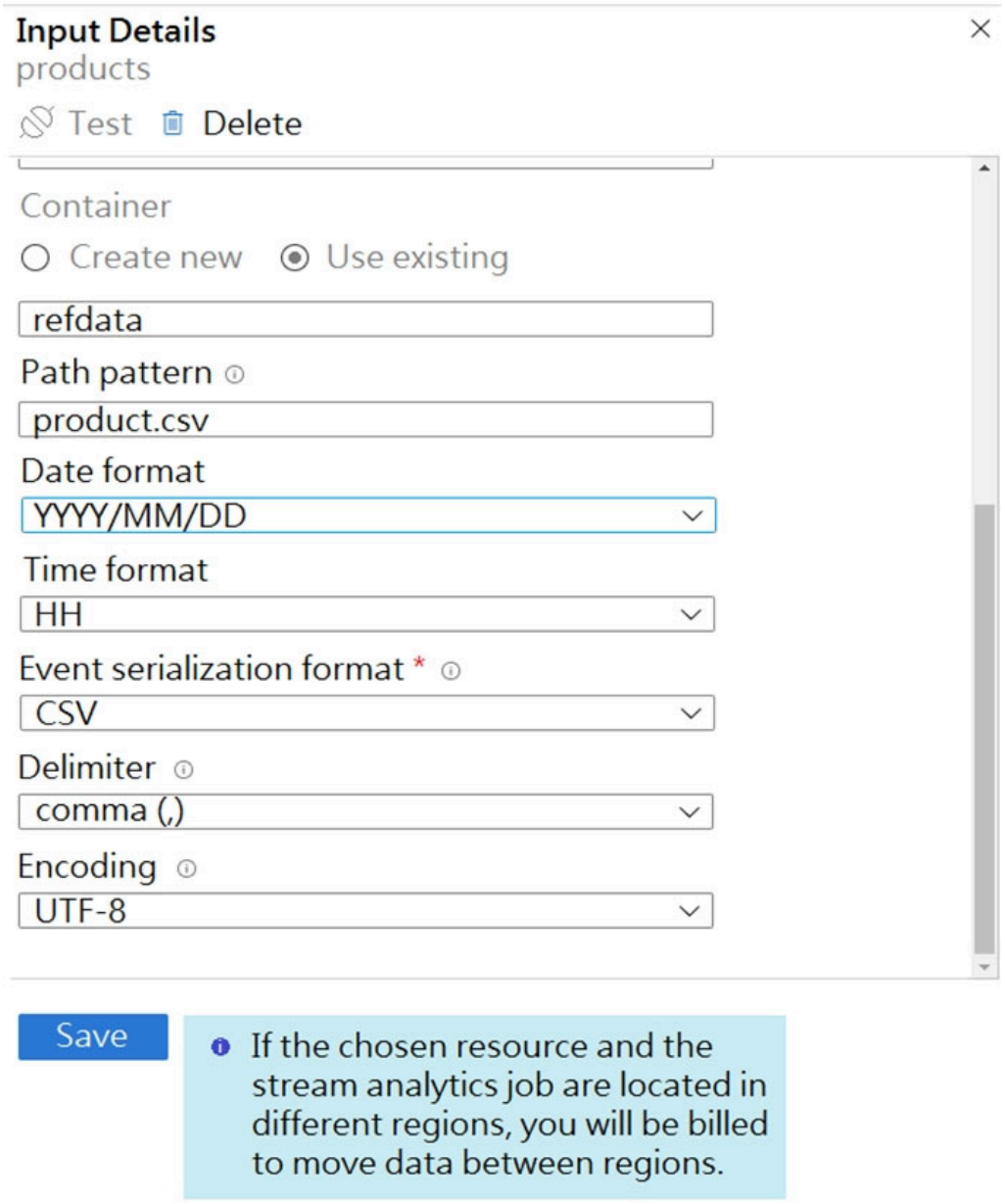
The storage account container view is shown in the Refdata exhibit. (Click the Refdata tab.)

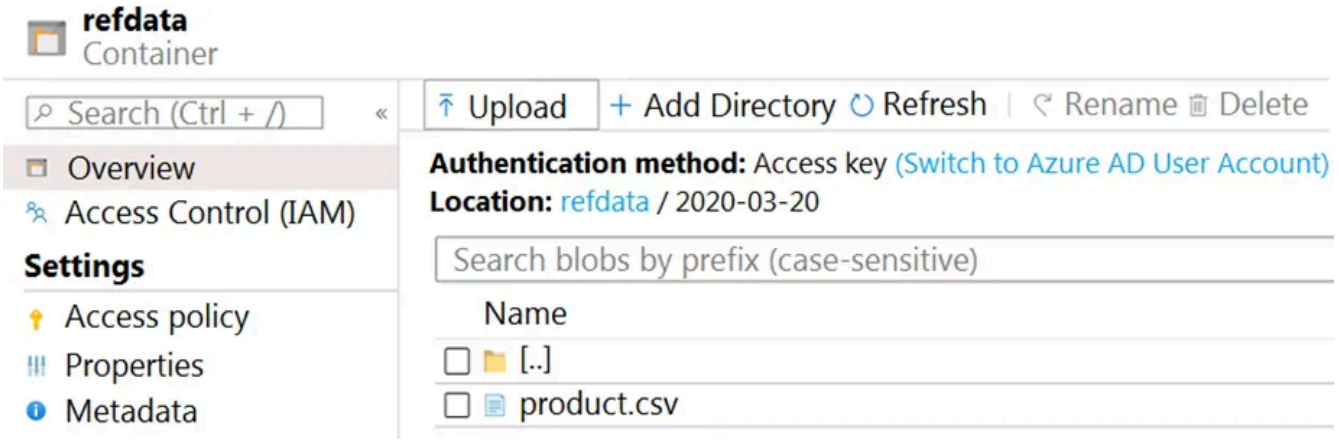
You need to configure the Stream Analytics job to pick up the new reference data.

What should you configure? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:





Answer Area

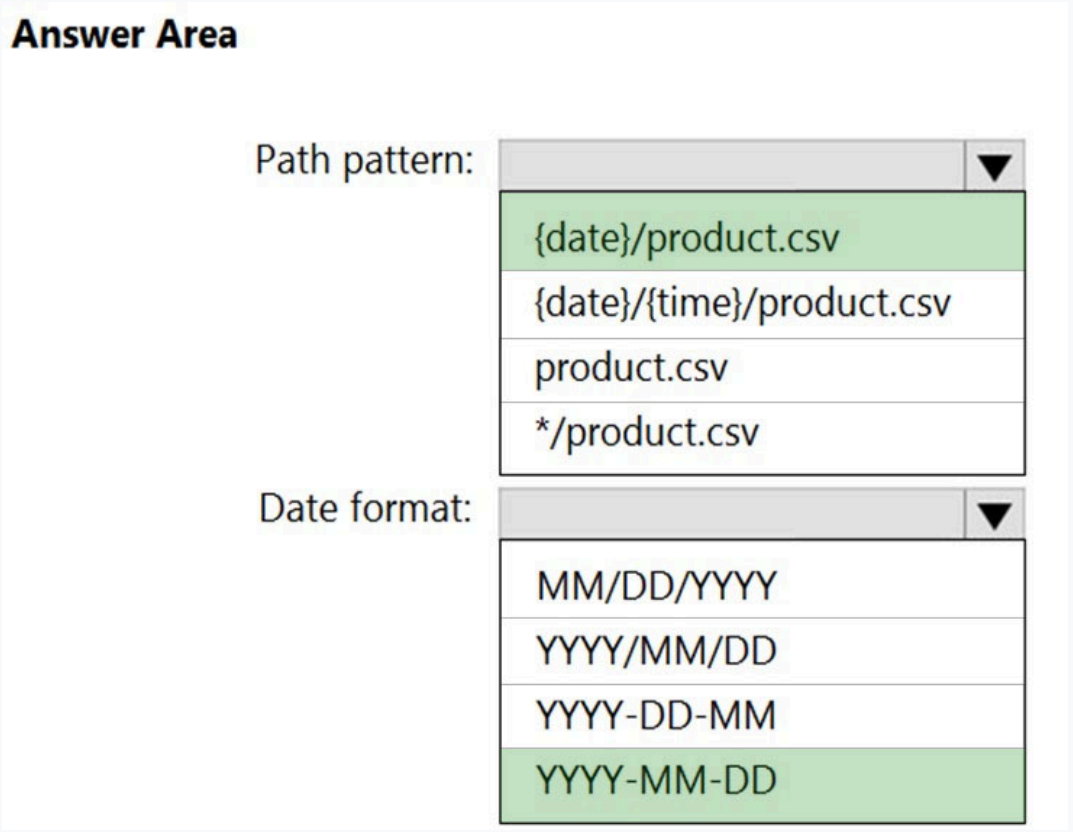
Path pattern:

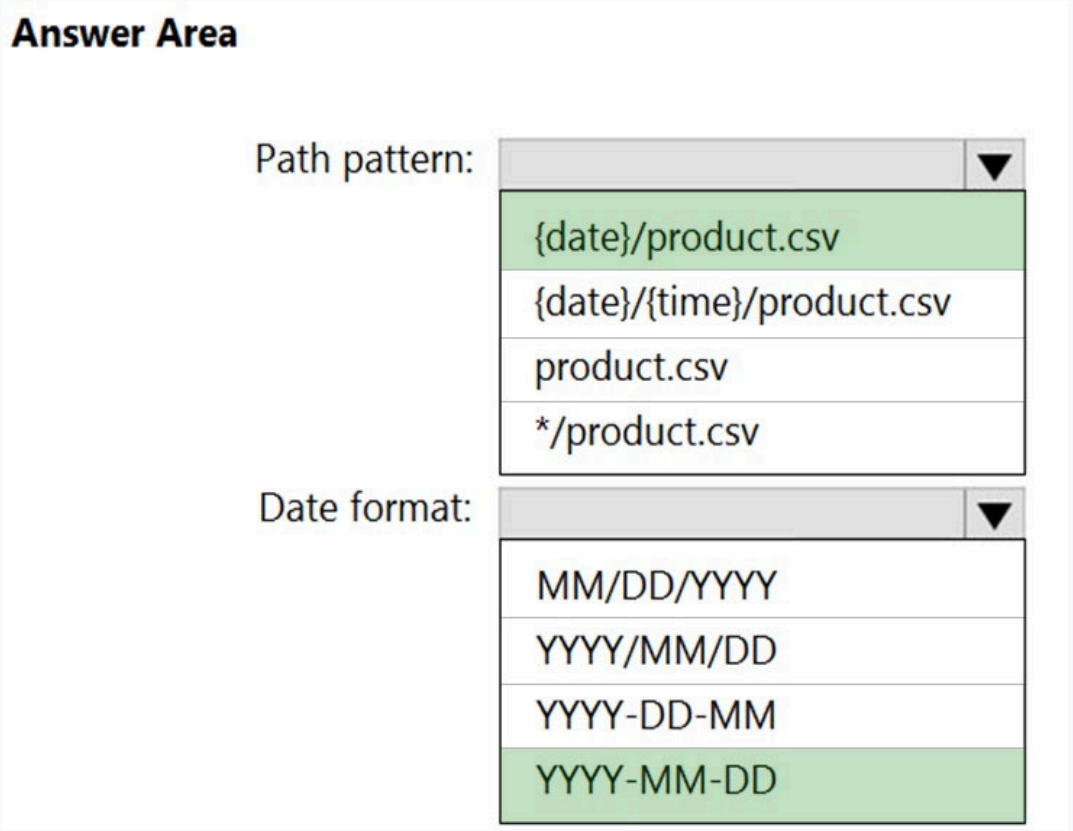
- {date}/product.csv
- {date}/{time}/product.csv
- product.csv
- */product.csv

Date format:

- MM/DD/YYYY
- YYYY/MM/DD
- YYYY-DD-MM
- YYYY-MM-DD

Suggested Answer:





Box 1: {date}/product.csv - In the 2nd exhibit we see: Location: refdata / 2020-03-20 Note: Path Pattern: This is a required property that is used to locate your blobs within the specified container. Within the path, you may choose to specify one or more instances of the following 2 variables: {date}, {time} Example 1: products/{date}/{time}/product-list.csv Example 2: products/{date}/product-list.csv

Example 3: product-list.csv -

Box 2: YYYY-MM-DD - Note: Date Format [optional]: If you have used {date} within the Path Pattern that you specified, then you can select the

date format in which your blobs are organized from the drop-down of supported formats. Example: YYYY/MM/DD, MM/DD/YYYY, etc. Reference: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data

Dicer Highly Voted 3 years ago

correct. Reasons: 1. {dat}/{time}/product.csv More detailed things should be put at the last. 2. YYYY-MM-DD if you choose YYYY/MM/DD, the system will think this is a file path.

upvoted 26 times

Rakrah Highly Voted 2 years, 7 months ago

second box is straight forwarded answer YYYY-MM-DD First Box = {date}/product.csv - Because the requirement is reference data loaded on daily basis, so it may be once in a day not hourly or timely.

upvoted 21 times

kkk5566 1 year, 11 months ago

correct

upvoted 1 times

Rakrah 2 years, 7 months ago

The recommended way to refresh reference data is to: Use {date}/{time} in the path pattern. Box 1 should be {dat}/{time}/product.csv

upvoted 2 times

Deeksha1234 3 years ago

given answer is correct

upvoted 3 times

Franz58 3 years ago

1. {date}/product.csv 2. YYYY-MM-DD

upvoted 2 times

temacc 3 years, 1 month ago

Path pattern: This required property is used to locate your blobs within the specified container. Within the path, you might choose to specify one or more instances of the variables {date} and {time}. Example 1: products/{date}/{time}/product-list.csv Example 2: products/{date}/product-list.csv Example 3: product-list.csv If the blob doesn't exist in the specified path, the Stream Analytics job waits indefinitely for the blob to become available. ##### Date format [optional]: If you used {date} within the path pattern you specified, select the date format in which your blobs are organized from the dropdown list of supported formats. Example: YYYY/MM/DD or MM/DD/YYYY ##### Time format [optional]: If you used {time} within the path pattern you specified, select the time format in which your blobs are organized from the dropdown list of supported formats. Example: HH, HH/mm, or HH-mm https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data

upvoted 2 times

Revave2 3 years, 1 month ago

I think the {date}/product.csv is correct, however it's formatted as YYYY/MM/DD, so why hyphenate it? In the example provided in the link the date was formatted with - instead of /, but in the question it's all /....

upvoted 3 times

Revave2 3 years, 1 month ago

I think the {date}/product.csv is correct, however it's formatted as YYYY/MM/DD, so why hyphenate it? In the example provided in the link the date was formatted with - instead of /, but in the question it's all /.....and to answer my own question, the second exhibit shows the product.csv file in refdata/yyyy-mm-dd . So that'll be the pathYes it through me at first because the first graphic refers to the file contents date format. When they looking for directory naming formatThe prints don't look consistent with each other, the 2nd image the location is: refdata/2020-03-20/product.csv should be "YYYY-DD-MM". But the 1st image shows a different format: YYYY/MM/DD.first image only shows the default, you need to change them to the actual path shown in image 2

upvoted 3 times

Revave2 3 years, 1 month ago

....and to answer my own question, the second exhibit shows the product.csv file in refdata/yyyy-mm-dd . So that'll be the path

upvoted 5 times

Revave2 3 years, 1 month ago

....and to answer my own question, the second exhibit shows the product.csv file in refdata/yyyy-mm-dd . So that'll be the pathYes it through me at first because the first graphic refers to the file contents date format. When they looking for directory naming format

upvoted 5 times

TriumphMC 2 years, 6 months ago

Yes it through me at first because the first graphic refers to the file contents date format. When they looking for directory naming format

upvoted 1 times

JayOR 2 years ago

The prints don't look consistent with each other, the 2nd image the location is: refdata/2020-03-20/product.csv should be "YYYY-DD-MM". But the 1st image shows a different format: YYYY/MM/DD.

upvoted 2 times

JayOR 2 years ago

The prints don't look consistent with each other, the 2nd image the location is: refdata/2020-03-20/product.csv should be "YYYY-DD-MM". But the 1st image shows a different format: YYYY/MM/DD.first image only shows the default, you need to change them to the actual path shown in image 2

upvoted 2 times

hypersam Most Recent 1 year, 7 months ago

first image only shows the default, you need to change them to the actual path shown in image 2

upvoted 1 times

demirsamuel 3 years, 2 months ago

answers are correct

upvoted 3 times

inotbf83 3 years, 3 months ago

I should change box 2 to YYYY/MM/DD (as shows 1st exhibit). A bit confusing with time format in the box 1.

upvoted 9 times

jackttt 3 years, 3 months ago

The file is updated daily, i think `{date}/product.csv` is correct

upvoted 5 times

Lotusss 3 years, 3 months ago

Wrong! Path Pattern: {dat}/{time}/product.csv Dat format: yyyy-mm-dd https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data

upvoted 3 times

Lotusss 3 years, 3 months ago

Wrong! Path Pattern: {dat}/{time}/product.csv Dat format: yyyy-mm-dd https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-dataSee that the file is stored under the date folder, and there is no time folder. Your link does recommend the time part, but the the link also says it's optional, and ultimately you need to answer the question, which states the path without the time.

upvoted 3 times

KashRaynardMorse 3 years, 3 months ago

See that the file is stored under the date folder, and there is no time folder. Your link does recommend the time part, but the the link also says it's optional, and ultimately you need to answer the question, which states the path without the time.

upvoted 8 times

HOTSPOT -

You have the following Azure Stream Analytics query.

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Hot Area:

```
WITH

step1 AS (SELECT *
     FROM input1
     PARTITION BY StateID
     INTO 10),
step2 AS (SELECT *
     FROM input2
     PARTITION BY StateID
     INTO 10)

SELECT *
INTO output
FROM step1
PARTITION BY StateID
UNION
SELECT * INTO output
       FROM step2
       PARTITION BY StateID
```

Answer Area

Statements	Yes	No
The query combines two streams of partitioned data.	○	○
The stream scheme key and count must match the output scheme.	○	○
Providing 60 streaming units will optimize the performance of the query.	○	○

Suggested Answer:

Answer Area

Statements	Yes	No
The query combines two streams of partitioned data.	○	○
The stream scheme key and count must match the output scheme.	○	○
Providing 60 streaming units will optimize the performance of the query.	○	○

Answer Area

Statements	Yes	No
The query combines two streams of partitioned data.	○	○
The stream scheme key and count must match the output scheme.	○	○
Providing 60 streaming units will optimize the performance of the query.	○	○

Box 1: No - Note: You can now use a new extension of Azure Stream Analytics SQL to specify the number of partitions of a stream when reshuffling the data. The outcome is a stream that has the same partition scheme. Please see below for an example: WITH step1 AS (SELECT * FROM [input1] PARTITION BY DeviceID INTO 10), step2 AS (SELECT * FROM [input2] PARTITION BY DeviceID INTO 10) SELECT * INTO [output] FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceID Note: The new extension of Azure Stream Analytics SQL includes a keyword INTO that allows you to specify the number of partitions for a stream when performing reshuffling using a PARTITION BY statement.

Box 2: Yes - When joining two streams of data explicitly repartitioned, these streams must have the same partition key and partition count.

Box 3: Yes - Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job. In general, the best practice is to start with 6 SUs for queries that don't use PARTITION BY. Here there are 10 partitions, so 6x10 = 60 SUs is good. Note: Remember, Streaming Unit (SU) count, which is the unit of scale for Azure Stream Analytics, must be adjusted so the number of physical resources available to the job can fit the partitioned flow. In general, six SUs is a good number to assign to each partition. In case there are insufficient resources assigned to the job, the system will only apply the repartition if it benefits the job. Reference: https://azure.microsoft.com/en-in/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/ https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption

objecto Highly Voted 4 years, 2 months ago

I feel its all YES. Since it does use a UNION and UNION combines. No matter it repartitions the result is the combination of two sources, a UNION of two sources. Am I missing something here?

upvoted 40 times

objecto Highly Voted 4 years, 2 months ago

I feel its all YES. Since it does use a UNION and UNION combines. No matter it repartitions the result is the combination of two sources, a UNION of two sources. Am I missing something here?I believe the answer to the first question heavily relies on creator's understanding "what is query". If it is the last part only (without CTEs) than the answer should be "yes", because you have partitioned data that come from CTEs as input for the main query. But if creator's understaning that query is the whole thing, than probably answer should be "no", because you're receiving non-partitioned data from sensors.

upvoted 40 times

oleg25 2 years, 10 months ago

I believe the answer to the first question heavily relies on creator's understanding "what is query". If it is the last part only (without CTEs) than the answer should be "yes", because you have partitioned data that come from CTEs as input for the main query. But if creator's understaning that query is the whole thing, than probably answer should be "no", because you're receiving non-partitioned data from sensors.

upvoted 1 times

auwia 3 years, 1 month ago

False, True, False. https://learn.microsoft.com/en-us/azure/stream-analytics/repartition The first is False, because this: "The following example query joins two streams of repartitioned data." It's extracted from the link above, and it's pointing to our query! Repartitioned and not partitioned. Second is True, it's explicitly written The output scheme should match the stream scheme key and count so that each substream can be flushed independently. Third is False, "In general, six SUs are needed for each partition." In the example we have 10 positions for step 1 and 10 for step 2, it should be 120 and not 60.

upvoted 26 times

auwia 3 years, 1 month ago

False, True, False. https://learn.microsoft.com/en-us/azure/stream-analytics/repartition The first is False, because this: "The following example query joins two streams of repartitioned data." It's extracted from the link above, and it's pointing to our query! Repartitioned and not partitioned. Second is True, it's explicitly written The output scheme should match the stream scheme key and count so that each substream can be flushed independently. Third is False, "In general, six SUs are needed for each partition." In the example we have 10 positions for step 1 and 10 for step 2, it should be 120 and not 60.they are unioned not joined

upvoted 26 times

Dusica 2 years, 3 months ago

they are unioned not joined

upvoted 2 times

samianae Most Recent 1 year, 6 months ago

yes -yes -no

upvoted 2 times

ff5037f 1 year, 9 months ago

the first question is definetly yes https://learn.microsoft.com/en-us/stream-analytics-query/union-azure-stream-analytics

upvoted 2 times

dgerok 2 years, 4 months ago

1) Y 2) Y The output scheme should match the stream scheme key and count so that each substream can be flushed independently. https://learn.microsoft.com/en-us/azure/stream-analytics/repartition 3) Y

upvoted 2 times

rocky48 2 years, 6 months ago

False, True, False. https://learn.microsoft.com/en-us/azure/stream-analytics/repartition The first is False, because this: "The following example query joins two streams of repartitioned data." It's extracted from the link above, and it's pointing to our query! Repartitioned and not partitioned. Second is True, it's explicitly written The output scheme should match the stream scheme key and count so that each substream can be flushed independently. Third is False, "In general, six SUs are needed for each partition." In the example we have 10 positions for step 1 and 10 for step 2, it should be 120 and not 60.

upvoted 3 times

Ha_Tran 2 years, 6 months ago

The first question should be yes. "Partitioning lets you divide data into subsets based on a partition key". https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization

upvoted 1 times

Anto___ 2 years, 6 months ago

Yes NO NO

upvoted 3 times

Momoanwar 2 years, 8 months ago

Chatgpt say : yes tes no. Based on the provided information and additional documentation on Azure Stream Analytics:

1. **Yes**: The query combines two streams of partitioned data [[”]](https://azure.microsoft.com/fr-fr/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/). 2. **Yes**: The stream scheme key and count should match the output scheme for optimal independent processing of each substream [[”]](https://azure.microsoft.com/fr-fr/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/). 3. **No**: The documentation does not specify that providing 60 streaming units will optimize the query's performance. The appropriate number of streaming units depends on experimentation and resource usage observation [[”]](https://azure.microsoft.com/fr-fr/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/).

upvoted 8 times

Momoanwar 2 years, 8 months ago

Chatgpt say : yes tes no. Based on the provided information and additional documentation on Azure Stream Analytics: 1. **Yes**: The query combines two streams of partitioned data [[”]](https://azure.microsoft.com/fr-fr/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/). 2. **Yes**: The stream scheme key and count should match the output scheme for optimal independent processing of each substream [[”]](https://azure.microsoft.com/fr-fr/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/). 3. **No**: The documentation does not specify that providing 60 streaming units will optimize the query's performance. The appropriate number of streaming units depends on experimentation and resource usage observation [[”]](https://azure.microsoft.com/fr-fr/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/).You gice one reasoning to chatgpt it will change its answer so i dont think it is reliable resource

upvoted 8 times

dakku987 2 years, 7 months ago

You gice one reasoning to chatgpt it will change its answer so i dont think it is reliable resource

upvoted 7 times

ellala 2 years, 10 months ago

The answer you need for first and second questions is in Microsoft Documentation: "The following example query joins two streams of repartitioned data. When joining two streams of repartitioned data, the streams must have the same partition key and count. The outcome is a stream that has the same partition scheme. (...) The output scheme should match the stream scheme key and count so that each substream can be flushed independently. " https://learn.microsoft.com/en-us/azure/stream-analytics/repartition#repartition-input-within-a-single-stream-analytics-job So its not two streams of partitioned data, but two streams of REpartitioned data. And the output stream must have the same partition key and count. For the third question, a bit lower in the same link, we get: In general, six SUs are needed for each partition. Therefore, if we have 10 partitions, 6*10 = 60.I will correct my previous message. After reading through SU calculation documentation, I concluded it should be 120 for SU V1 or 20 for SU V2. Therefore none of them would be 60. Explanation is this sentence here: "All non-partitioned steps together can scale up to one streaming unit (SU V2s) for a Stream Analytics job. In addition, you can add 1 SU V2 for each partition in a partitioned step. So tecnically, it would be 21 SU V2. Therefore, 60 SU is not correct. Should be FALSE TRUE FALSECalculate the max streaming units for a job All non-partitioned steps together can scale up to one streaming unit (SU V2s) for a Stream Analytics job. In addition, you can add 1 SU V2 for each partition in a partitioned step. You can see some examples in the table below. Query Max SUs for the job The query contains one step. The step isn't partitioned. 1 SU V2 The input data stream is partitioned by 16. The query contains one step. The step is partitioned. 16 SU V2 (1 * 16 partitions) The query contains two steps. Neither of the steps is partitioned. 1 SU V2 The input data stream is partitioned by 3. The query contains two st

upvoted 3 times

ellala 2 years, 10 months ago

The answer you need for first and second questions is in Microsoft Documentation: "The following example query joins two streams of repartitioned data. When joining two streams of repartitioned data, the streams must have the same partition key and count. The outcome is a stream that has the same partition scheme. (...) The output scheme should match the stream scheme key and count so that each substream can be flushed independently. " https://learn.microsoft.com/en-us/azure/stream-analytics/repartition#repartition-input-within-a-single-stream-analytics-job

So its not two streams of partitioned data, but two streams of REpartitioned data. And the output stream must have the same partition key and count.

For the third question, a bit lower in the same link, we get: In general, six SUs are needed for each partition. Therefore, if we have 10 partitions, 6*10 = 60.

upvoted 3 times

ellala 2 years, 10 months ago

I will correct my previous message. After reading through SU calculation documentation, I concluded it should be 120 for SU V1 or 20 for SU V2. Therefore none of them would be 60. Explanation is this sentence here: "All non-partitioned steps together can scale up to one streaming unit (SU V2s) for a Stream Analytics job. In addition, you can add 1 SU V2 for each partition in a partitioned step. So tecnically, it would be 21 SU V2. Therefore, 60 SU is not correct. Should be FALSE TRUE FALSECalculate the max streaming units for a job All non-partitioned steps together can scale up to one streaming unit (SU V2s) for a Stream Analytics job. In addition, you can add 1 SU V2 for each partition in a partitioned step. You can see some examples in the table below. Query Max SUs for the job The query contains one step. The step isn't partitioned. 1 SU V2 The input data stream is partitioned by 16. The query contains one step. The step is partitioned. 16 SU V2 (1 * 16 partitions) The query contains two steps. Neither of the steps is partitioned. 1 SU V2 The input data stream is partitioned by 3. The query contains two steps. The input step is partitioned and the second step isn't. The SELECT statement reads from the partitioned input. 4 SU V2s (3 for partitioned steps + 1 for non-partitioned steps Base on the above from MS documentation, why do we need to multiply by 6SUs?

upvoted 3 times

ellala 2 years, 10 months ago

I will correct my previous message. After reading through SU calculation documentation, I concluded it should be 120 for SU V1 or 20 for SU V2. Therefore none of them would be 60. Explanation is this sentence here: "All non-partitioned steps together can scale up to one streaming unit (SU

V2s) for a Stream Analytics job. In addition, you can add 1 SU V2 for each partition in a partitioned step. So tecnically, it would be 21 SU V2. Therefore, 60 SU is not correct.

Should be FALSE TRUE FALSE

upvoted 3 times

ahmadsayeed 2 years, 9 months ago

Calculate the max streaming units for a job All non-partitioned steps together can scale up to one streaming unit (SU V2s) for a Stream Analytics job. In addition, you can add 1 SU V2 for each partition in a partitioned step. You can see some examples in the table below. Query Max SUs for the job The query contains one step. The step isn't partitioned. 1 SU V2 The input data stream is partitioned by 16. The query contains one step. The step is partitioned. 16 SU V2 (1 * 16 partitions) The query contains two steps. Neither of the steps is partitioned. 1 SU V2 The input data stream is partitioned by 3. The query contains two steps. The input step is partitioned and the second step isn't. The SELECT statement reads from the partitioned input. 4 SU V2s (3 for partitioned steps + 1 for non-partitioned steps Base on the above from MS documentation, why do we need to multiply by 6SUs?

upvoted 1 times

Vanq69 2 years, 10 months ago

I think this is an older question since there is SU V2 now. According to this: https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization Each partition would consume 1 Streamiung Unit (SU) V2 and since we have 2 inputs with 10 partitions each it would add up to 20 SU V2, now we have 2 Select statements after the WITH Step which each consume 1 SU V2, so it should add up to 22 SU V2 which would equal 22*6=132 SU V1.

upvoted 1 times

mav2000 2 years, 10 months ago

Based on recommended Streaming units, Step 1: 10 partitions Step 2: 10 partitions (1*10+1*10) = 20 SU's is the optimal, if you have more, it's not ideal because some SU's are inactive and if you have less, it can cause a bottleneck https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization#calculate-the-maximum-streaming-units-of-a-job

upvoted 1 times

Saintu 2 years, 11 months ago

Question 1 is false; the question says the union of streams and not data. The union combines 2 streams which are the same and thus, the output is the same stream.

upvoted 1 times

auwia 3 years, 1 month ago

1. True 2. False: In the context of a UNION operation in Azure Stream Analytics, the stream scheme key and count do not need to match the output schema. The key and count of the output schema are determined based on the input streams being unioned. When performing a UNION operation, the input streams must have compatible schemas, which means that the data types and field names should align. However, the key and count are determined by the input streams themselves and do not need to match the output schema. 3. TrueAgain for point 2-False: https://learn.microsoft.com/en-us/stream-analytics-query/union-azure-stream-analytics The following are basic rules for combining the result sets of two queries by using UNION: - The number and the order of the columns must be the same in all queries. - The data types must be compatible. - Streams must have the same partition and partition count (not scheme key and count! :-))All answer is no.forgot it, no, yes, noFalse, true, false. I've changed mind completely looking and reading accurately this official link where all 3 questions are answered: https://learn.microsoft.com/en-us/azure/stream-analytics/repartition The first is False, because this: "The following example query joins two streams of repartitioned data." It's extracted from the link above, and it's pointing to our query! Repartitioned and not partitioned. Second is True, it's explicitly written The output scheme should match the stream scheme key and count so that each substream can be flushed independently. Third is False, "In general, six SUs are needed for each partition." In the example we have 10 positions for step 1 and 10 for step 2, it should be 120 and not 60.

upvoted 2 times

auwia 3 years, 1 month ago

1. True 2. False: In the context of a UNION operation in Azure Stream Analytics, the stream scheme key and count do not need to match the output schema. The key and count of the output schema are determined based on the input streams being unioned.

When performing a UNION operation, the input streams must have compatible schemas, which means that the data types and field names should align. However, the key and count are determined by the input streams themselves and do not need to match the output schema. 3. True

upvoted 2 times

auwia 3 years, 1 month ago

Again for point 2-False: https://learn.microsoft.com/en-us/stream-analytics-query/union-azure-stream-analytics The following are basic rules for combining the result sets of two queries by using UNION: - The number and the order of the columns must be the same in all queries. - The data types must be compatible. - Streams must have the same partition and partition count (not scheme key and count! :-))All answer is no.forgot it, no, yes, no

upvoted 1 times

auwia 3 years, 1 month ago

Again for point 2-False: https://learn.microsoft.com/en-us/stream-analytics-query/union-azure-stream-analytics The following are basic rules for combining the result sets of two queries by using UNION: - The number and the order of the columns must be the same in all queries. - The data types must be compatible. - Streams must have the same partition and partition count (not scheme key and count! :-))

upvoted 1 times

kkk5566 2 years, 11 months ago

All answer is no.

upvoted 1 times

kkk5566 2 years, 11 months ago

All answer is no.forgot it, no, yes, no

upvoted 1 times

kkk5566 2 years, 11 months ago

forgot it, no, yes, no

upvoted 1 times

auwia 3 years, 1 month ago

False, true, false. I've changed mind completely looking and reading accurately this official link where all 3 questions are answered: https://learn.microsoft.com/en-us/azure/stream-analytics/repartition The first is False, because this: "The following example query joins two streams of repartitioned data." It's extracted from the link above, and it's pointing to our query! Repartitioned and not partitioned. Second is True, it's explicitly written The output scheme should match the stream scheme key and count so that each substream can be flushed independently. Third is False, "In general, six SUs are needed for each partition." In the example we have 10 positions for step 1 and 10 for step 2, it should be 120 and not 60.

upvoted 4 times

UristMcFarmer 3 years, 7 months ago

I believe the answer to the First Question is No because the execution of the last statement will result in an error because the second query has "SELECT INTO output" rather than a straight SELECT. Can anyone confirm that you can UNION two data sets both directed to the same stream with SELECT INTO? It's not functionality shown in any example I've been able to find (in the given answer, linked in other comments on here, and my own research).

upvoted 1 times

rohanb1986 3 years, 7 months ago

Is the first option NO because it mentions partitioned data instead of repartitioned data? Reference : https://learn.microsoft.com/en-us/azure/stream-analytics/repartition The following example query joins two streams of repartitioned data. When joining two streams of repartitioned data, the streams must have the same partition key and count. The outcome is a stream that has the same partition scheme. WITH step1 AS (SELECT * FROM input1 PARTITION BY DeviceID), step2 AS (SELECT * FROM input2 PARTITION BY DeviceID) SELECT * INTO output FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceIDyes, the only difference I see is this: 1)this example: SELECT * INTO output FROM step1 PARTITION BY StateID UNION SELECT * INTO output FROM step2 PARTITION BY StateID #is using another select * on top of step2 2)the doc example: SELECT * INTO output FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceID I think both query joins two streams of repartitioned/partitioned data, so first answer should be yes

upvoted 5 times

rohanb1986 3 years, 7 months ago

Is the first option NO because it mentions partitioned data instead of repartitioned data?

Reference : https://learn.microsoft.com/en-us/azure/stream-analytics/repartition

The following example query joins two streams of repartitioned data. When joining two streams of repartitioned data, the streams must have the same partition key and count. The outcome is a stream that has the same partition scheme.

WITH step1 AS (SELECT * FROM input1 PARTITION BY DeviceID), step2 AS (SELECT * FROM input2 PARTITION BY DeviceID)

SELECT * INTO output FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceID

upvoted 5 times

vrodriguesp 3 years, 6 months ago

yes, the only difference I see is this: 1)this example: SELECT * INTO output FROM step1 PARTITION BY StateID UNION SELECT * INTO output FROM step2 PARTITION BY StateID #is using another select * on top of step2 2)the doc example: SELECT * INTO output FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceID I think both query joins two streams of repartitioned/partitioned data, so first answer should be yes

upvoted 2 times

XiltroX 3 years, 8 months ago

All 3 are YES

upvoted 3 times

HOTSPOT -

You are building a database in an Azure Synapse Analytics serverless SQL pool.

You have data stored in Parquet files in an Azure Data Lake Storege Gen2 container.

Records are structured as shown in the following sample.

{

"id": 123,

"address_housenumber": "19c",

"address_line": "Memory Lane",

"applicant1_name": "Jane",

"applicant2_name": "Dev"

}

The records contain two applicants at most.

You need to build a table that includes only the address fields.

How should you complete the Transact-SQL statement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
[CREATE EXTERNAL TABLE | CREATE TABLE | CREATE VIEW] applications
WITH (
    LOCATION = 'applications/',
    DATA_SOURCE = applications_ds,
    FILE_FORMAT = applications_file_format
)
AS
SELECT id, [address_housenumber] as addresshousenumber, [address_line1] as addressline1
FROM
[CROSS APPLY | OPENJSON | OPENROWSET] (BULK 'https://contoso1.dfs.core.windows.net/applications/year=*/*.parquet',
FORMAT='PARQUET') AS [r]
GO
```

Suggested Answer:

Answer Area

```
[CREATE EXTERNAL TABLE | CREATE TABLE | CREATE VIEW] applications
WITH (
    LOCATION = 'applications/',
    DATA_SOURCE = applications_ds,
    FILE_FORMAT = applications_file_format
)
AS
SELECT id, [address_housenumber] as addresshousenumber, [address_line1] as addressline1
FROM
[CROSS APPLY | OPENJSON | OPENROWSET] (BULK 'https://contoso1.dfs.core.windows.net/applications/year=*/*.parquet',
FORMAT='PARQUET') AS [r]
GO
```

Answer Area

```
[CREATE EXTERNAL TABLE | CREATE TABLE | CREATE VIEW] applications
WITH (
    LOCATION = 'applications/',
    DATA_SOURCE = applications_ds,
    FILE_FORMAT = applications_file_format
)
AS
SELECT id, [address_housenumber] as addresshousenumber, [address_line1] as addressline1
FROM
[CROSS APPLY | OPENJSON | OPENROWSET] (BULK 'https://contoso1.dfs.core.windows.net/applications/year=*/*.parquet',
FORMAT='PARQUET') AS [r]
GO
```

Box 1: CREATE EXTERNAL TABLE - An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool. Syntax: CREATE EXTERNAL TABLE { database_name.schema_name.table_name | schema_name.table_name | table_name } ([,...n]) WITH (LOCATION = 'folder_or_filepath', DATA_SOURCE = external_data_source_name, FILE_FORMAT = external_file_format_name

Box 2. OPENROWSET - When using serverless SQL pool, CETAS is used to create an external table and export query results to Azure Storage Blob or Azure Data Lake Storage Gen2. Example:

AS - SELECT decennialTime, stateName, SUM(population) AS population

FROM - OPENROWSET(BULK 'https://azureopendatastorage.blob.core.windows.net/censusdatacontainer/release/us_population_county/year=*/*.parquet', FORMAT='PARQUET') AS [r] GROUP BY decennialTime, stateName

GO - Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

Deeksha1234 Highly Voted 4 years ago

Correct answer

upvoted 20 times

Piantoni Most Recent 1 year, 9 months ago

1) CREATE EXTERNAL TABLE 2) OPENROWSET

upvoted 1 times

dgerok 2 years, 4 months ago

1) CREATE EXTERNAL TABLE (because this is SERVERLESS sql pool) 2) OPENROWSET

upvoted 3 times

kkk5566 2 years, 11 months ago

correct

upvoted 1 times

cjb0 3 years, 11 months ago

Why is it External Table if it's a Serverless SQL pool?

upvoted 2 times

cjb0 3 years, 11 months ago

Why is it External Table if it's a Serverless SQL pool?"You have data stored in Parquet files in an Azure Data Lake Storege Gen2 container."because serverless SQL pool does not have internal tables

upvoted 2 times

vctrhugo 3 years, 1 month ago

"You have data stored in Parquet files in an Azure Data Lake Storege Gen2 container."

upvoted 2 times

NORLI Highly Voted 3 years, 11 months ago

because serverless SQL pool does not have internal tables

upvoted 15 times

dkamat 4 years, 1 month ago

correct

upvoted 3 times

SandipSingha 4 years, 3 months ago

correct

upvoted 3 times

Feljoud 4 years, 3 months ago

correct

upvoted 4 times

HOTSPOT -

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named Account1.

You plan to access the files in Account1 by using an external table.

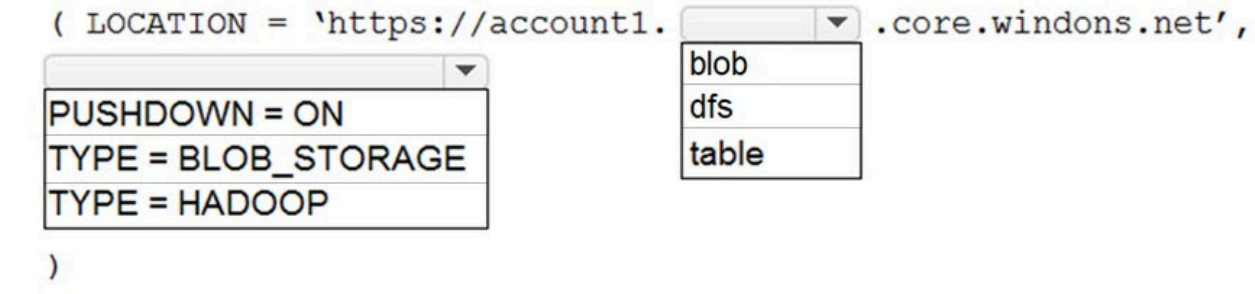
You need to create a data source in Pool1 that you can reference when you create the external table.

How should you complete the Transact-SQL statement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area



Suggested Answer:

Answer Area

```
CREATE EXTERNAL DATA SOURCE source1
WITH
   ( LOCATION = 'https://account1. [ blob / dfs / table ] .core.windons.net',
   [ PUSHDOWN = ON / TYPE = BLOB_STORAGE / TYPE = HADOOP ]
   )
```

Answer Area

```
CREATE EXTERNAL DATA SOURCE source1
WITH
   ( LOCATION = 'https://account1. [ blob / dfs / table ] .core.windons.net',
   [ PUSHDOWN = ON / TYPE = BLOB_STORAGE / TYPE = HADOOP ]
   )
```

Box 1: blob - The following example creates an external data source for Azure Data Lake Gen2 CREATE EXTERNAL DATA SOURCE YellowTaxi WITH (LOCATION = 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/', TYPE = HADOOP)

Box 2: HADOOP - Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

galacaw Highly Voted 4 years, 3 months ago

1. dfs (for Azure Data Lake Storage Gen2)

upvoted 69 times

galacaw Highly Voted 4 years, 3 months ago

1. dfs (for Azure Data Lake Storage Gen2)dfs is not validdfs is valid Data Lake Storage Gen2 abfs[s] @ .dfs.core.windows.net http[s] .dfs.core.windows.net/ /subfolders wasb[s] @ .blob.core.windows.netThis table corroborates that "dfs" should be used for ADLS Gen 2: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#locationCREATE EXTERNAL DATA SOURCE mydatasource WITH (LOCATION = 'abfss://', CREDENTIAL = AzureStorageCredential, TYPE = HADOOP)Correct, https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw-latest&preserve-view=true&tabs=dedicated#location--prefixpath

upvoted 69 times

panda_azzurro 3 years, 6 months ago

dfs is not validdfs is valid Data Lake Storage Gen2 abfs[s] @ .dfs.core.windows.net http[s] .dfs.core.windows.net/ /subfolders wasb[s] @ .blob.core.windows.net

upvoted 1 times

panda_azzurro 3 years, 6 months ago

dfs is not valid

upvoted 1 times

suvec 3 years, 4 months ago

dfs is valid Data Lake Storage Gen2 abfs[s] @ .dfs.core.windows.net http[s] .dfs.core.windows.net/ /subfolders wasb[s] @ .blob.core.windows.net

upvoted 7 times

jds0 3 years, 4 months ago

This table corroborates that "dfs" should be used for ADLS Gen 2: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#location

upvoted 1 times

Vedjha 3 years, 6 months ago

CREATE EXTERNAL DATA SOURCE mydatasource WITH (LOCATION = 'abfss://data@storageaccount.dfs.core.windows.net', CREDENTIAL = AzureStorageCredential, TYPE = HADOOP)

upvoted 9 times

Vedjha 3 years, 6 months ago

CREATE EXTERNAL DATA SOURCE mydatasource WITH (LOCATION = 'abfss://', CREDENTIAL = AzureStorageCredential, TYPE = HADOOP)

upvoted 9 times

Rob77 3 years, 2 months ago

Correct, https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw-latest&preserve-view=true&tabs=dedicated#location--prefixpath

upvoted 5 times

Kure87 3 years, 8 months ago

1. blob. Acoording with this article https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop we only use DFS (abfss endpoint) when your account has secure transfer enabled. On the question the location starts with "https://account1." not "abfss://"That's not correct, you use abfss:// if you have secure transfer enabled. There is nothing wrong with using https:// when you don't have secure transfer enabled. However, for DLSv2 you need to specify .dfs. ... The correct answer is is: dfs hadoop@kure87 DFS is valid Data Lake Storage Gen2 abfs[s] @ .dfs.core.windows.net http[s] .dfs.core.windows.net/ /subfolders wasb[s] @ .blob.core.windows.netI agree with DFS but remember this is possible on blob storage: http[s] .blob.core.windows.net/ /subfolders I created an linked service from my synapse to storage account explicitly using blob.please upvote thisAs it is written in this MSF example on how to create an External data source based on Azure storage account: " TYPE = HADOOP, -- For dedicated SQL pool -- TYPE = BLOB_STORAGE, -- For serverless SQL pool " https://learn.microsoft.com/en-us/training/modules/use-azure-synapse-serverless-sql-pools-for-transforming-data-lake/2-transform-data-using-create-external-table-select-statement

upvoted 39 times

Kure87 3 years, 8 months ago

1. blob. Acoording with this article https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop we only use DFS (abfss endpoint) when your account has secure transfer enabled.

On the question the location starts with "https://account1." not "abfss://"

upvoted 39 times

vadiminski_a 3 years, 4 months ago

That's not correct, you use abfss:// if you have secure transfer enabled. There is nothing wrong with using https:// when you don't have secure transfer enabled. However, for DLSv2 you need to specify .dfs. ... The correct answer is is: dfs hadoop

upvoted 7 times

suvec 3 years, 4 months ago

@kure87 DFS is valid Data Lake Storage Gen2 abfs[s] @ .dfs.core.windows.net http[s] .dfs.core.windows.net/ /subfolders wasb[s] @ .blob.core.windows.net

upvoted 6 times

suvec 3 years, 4 months ago

@kure87 DFS is valid Data Lake Storage Gen2 abfs[s] @ .dfs.core.windows.net http[s] .dfs.core.windows.net/ /subfolders wasb[s] @ .blob.core.windows.netI agree with DFS but remember this is possible on blob storage: http[s] .blob.core.windows.net/ /subfolders I created an linked service from my synapse to storage account explicitly using blob.

upvoted 6 times

JustImperius 1 year, 7 months ago

I agree with DFS but remember this is possible on blob storage: http[s] .blob.core.windows.net/ /subfolders I created an linked service from my synapse to storage account explicitly using blob.

upvoted 1 times

Sebastian1677 3 years, 8 months ago

please upvote this

upvoted 3 times

tlb_20 2 years, 4 months ago

As it is written in this MSF example on how to create an External data source based on Azure storage account: " TYPE = HADOOP, -- For dedicated SQL pool -- TYPE = BLOB_STORAGE, -- For serverless SQL pool " https://learn.microsoft.com/en-us/training/modules/use-azure-synapse-serverless-sql-pools-for-transforming-data-lake/2-transform-data-using-create-external-table-select-statement

upvoted 5 times

samianae Most Recent 1 year, 6 months ago

dfs and hadoop

upvoted 1 times

ff5037f 1 year, 9 months ago

answer is correct. as location prefix is https. if the container location prefix is abfs then it is dfs, if it is wbfs then it is blob . https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw-latest&preserve-view=true&tabs=dedicated#location--prefixpath Data Lake Storage Gen2 http[s] .dfs.core.windows.net/ /subfolders

upvoted 2 times

ypan 2 years, 1 month ago

Dedicated SQL Pool: Use CREATE EXTERNAL DATA SOURCE with TYPE = HADOOP for accessing Azure Data Lake Storage Gen2. Serverless SQL Pool: Use OPENROWSET for direct querying of the external data.

upvoted 1 times

Charley92 2 years, 4 months ago

CREATE EXTERNAL DATA SOURCE MyDataSource WITH (TYPE = HADOOP, LOCATION = 'abfss:// @ .dfs.core.windows.net/', CREDENTIAL =);

upvoted 2 times

dgerok 2 years, 4 months ago

answer is correct see the MS example https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#example-for-create-external-data-source

upvoted 1 times

ankeshpatel2112 2 years, 4 months ago

Correct Answer : DFS Explnation : If your source is ADLS Gen2 then it would be "DFS" and if your source is Azure Blob Storage then "Blob". Please refer below table from Microsoft Documentation. External Data Source Connector location prefix Location path Data Lake Storage* Gen1 adl .azuredatalake.net Data Lake Storage Gen2 abfs[s] @ .dfs.core.windows.net Azure Blob Storage wasbs @ .blob.core.windows.net Azure Blob Storage https .blob.core.windows.net/ /subfolders Data Lake Storage Gen1 http[s] .azuredatalakestore.net/webhdfs/v1 Data Lake Storage Gen2 http[s] .dfs.core.windows.net/ /subfolders Data Lake Storage Gen2 wasb[s] @ .blob.core.windows.net

upvoted 2 times

Alongi 2 years, 4 months ago

Should be DFS for Datalake Gen2

upvoted 1 times

Momoanwar 2 years, 8 months ago

Both `blob` and `dfs` endpoints work when connecting to Azure Data Lake Storage Gen2, but they serve different purposes. The `blob` endpoint is typically used for standard storage operations, while the `dfs` endpoint is optimized for hierarchical file system operations and is preferred for analytics workloads with Azure Synapse Analytics.

upvoted 1 times

Momoanwar 2 years, 8 months ago

Both `blob` and `dfs` endpoints work when connecting to Azure Data Lake Storage Gen2, but they serve different purposes. The `blob` endpoint is typically used for standard storage operations, while the `dfs` endpoint is optimized for hierarchical file system operations and is preferred for analytics workloads with Azure Synapse Analytics.To simply access files in Azure Data Lake Storage Gen2 for reading and analysis, without the need for Data Lake specific features like directory management or fine-grained ACLs, using the `blob` endpoint is sufficient. If your operations are primarily related to accessing files for reading, the `blob` endpoint can be used in the external data source definition within Azure Synapse Analytics.

upvoted 1 times

Momoanwar 2 years, 8 months ago

To simply access files in Azure Data Lake Storage Gen2 for reading and analysis, without the need for Data Lake specific features like directory management or fine-grained ACLs, using the `blob` endpoint is sufficient. If your operations are primarily related to accessing files for reading, the `blob` endpoint can be used in the external data source definition within Azure Synapse Analytics.

upvoted 2 times

fahfouhi94 2 years, 11 months ago

Ans : dfs & hadoop https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw-latest&preserve-view=true&tabs=dedicated

upvoted 1 times

kkk5566 2 years, 11 months ago

CREATE EXTERNAL DATA SOURCE AzureDataLakeStore WITH -- Please note the abfss endpoint when your account has secure transfer enabled (LOCATION = 'abfss://' , CREDENTIAL = ADLS_credential , TYPE = HADOOP) ; CREATE EXTERNAL DATA SOURCE YellowTaxi WITH (LOCATION = 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/', TY HADOOP, blob

upvoted 1 times

kkk5566 2 years, 11 months ago

CREATE EXTERNAL DATA SOURCE AzureDataLakeStore WITH -- Please note the abfss endpoint when your account has secure transfer enabled (LOCATION = 'abfss://data@newyorktaxidataset.dfs.core.windows.net' , CREDENTIAL = ADLS_credential , TYPE = HADOOP) ;

CREATE EXTERNAL DATA SOURCE YellowTaxi WITH (LOCATION = 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/', TY

HADOOP, blob

upvoted 1 times

kdp203 2 years, 11 months ago

dfs should be the correct answer (ADLS Gen2)

upvoted 1 times

auwia 3 years, 1 month ago

Confirmed HADOOP and DFS: External Data Source | Connector | Location path ---------- Data Lake Storage Gen1 | adl | .azuredatalake.net Data Lake Storage Gen2 | abfs[s] | @ .dfs.core.windows.net Azure Blob Storage | wasbs | @ .blob.core.windows.net Azure Blob Storage | https | .blob.core.windows.net/ /subfolders Data Lake Storage Gen1 | http[s] | .azuredatalakestore.net/webhdfs/v1 Data Lake Storage Gen2 | http[s] | .dfs.core.windows.net/ /subfolders Data Lake Storage Gen2 | wasb[s] | @ .blob.core.windows.net

upvoted 3 times

auwia 3 years, 2 months ago

CREATE EXTERNAL DATA SOURCE source1 WITH (LOCATION = 'https://account1.dfs.core.windows.net', TYPE = HADOOP)

upvoted 1 times

aga444 3 years, 2 months ago

CREATE EXTERNAL DATA SOURCE DataSourceName WITH (TYPE = HADOOP, LOCATION = 'adl://Account1.dfs.core.windows.net/', CREDENTIAL = SqlPoolCredential);

upvoted 1 times

janaki 3 years, 2 months ago

CREATE EXTERNAL DATA SOURCE WITH (TYPE = HADOOP, LOCATION = 'adl:// .dfs.core.windows.net', CREDENTIAL =); So answer is dfs and Type = Hadoop

upvoted 1 times

Question #55 *Topic 1*

You have an Azure subscription that contains an Azure Blob Storage account named storage1 and an Azure Synapse Analytics dedicated SQL pool named
Pool1.
You need to store data in storage1. The data will be read by Pool1. The solution must meet the following requirements:
Enable Pool1 to skip columns and rows that are unnecessary in a query.
➪ Automatically create column statistics.
➪ Minimize the size of files.
Which type of file should you use?

▪

A. JSON

B. Parquet

C. Avro

D. CSV

Suggested Answer: *B*
Automatic creation of statistics is turned on for Parquet files. For CSV files, you need to create statistics manually until automatic creation of CSV files statistics is supported. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-statistics

Community vote distribution
B (100%)

namtn6 Highly Voted 4 years, 1 month ago
If the answer has Parquet. Of course, you should choose that answer. :D
upvoted 29 times

ClassMistress Highly Voted 4 years, 3 months ago
Automatic creation of statistics is turned on for Parquet files. For CSV files, you need to create statistics manually until automatic creation of CSV files statistics is supported.Good point, also better cost
upvoted 15 times

ClassMistress 4 years, 3 months ago
Selected Answer: B
Automatic creation of statistics is turned on for Parquet files. For CSV files, you need to create statistics manually until automatic creation of CSV files statistics is supported.
upvoted 15 times

sdokmak 4 years, 2 months ago
Good point, also better cost
upvoted 3 times

GigachadPrime Most Recent 1 year, 8 months ago
If it has Parquet, odds are very good.
upvoted 1 times

EmnCours 1 year, 8 months ago
Selected Answer: B
upvoted 1 times

dgerok 2 years, 4 months ago
PARQUET
upvoted 2 times

phydev 2 years, 9 months ago
Parquet is always the answer ;-)
upvoted 5 times

kkk5566 2 years, 11 months ago

is correct

upvoted 1 times

akhil5432 3 years ago

Parquet

upvoted 1 times

ajhak 3 years, 3 months ago

When in doubt, select Parquet.

upvoted 8 times

greenlever 3 years, 10 months ago

When reading from Parquet files, you can specify only the columns you want to read and skip the rest.

upvoted 5 times

Rahuar 3 years, 11 months ago

CORRECT

upvoted 2 times

monibun 3 years, 11 months ago

Question is bit contradictory: it mentions reading blob storage data in dedicated sql , which could be done by External Tables, however, dedicated sql pool do NOT support automatic stats for external tables (as mentioned on "automatic stats creation for dedicated sql pool" section- https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-statistics

upvoted 3 times

monibun 3 years, 11 months ago

Question is bit contradictory: it mentions reading blob storage data in dedicated sql , which could be done by External Tables, however, dedicated sql pool do NOT support automatic stats for external tables (as mentioned on "automatic stats creation for dedicated sql pool" section- https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-statistics@monibun : great point. I agree with you . I am not sure if anyone wants to advise on your point .

upvoted 3 times

Ranjan6214 2 years, 7 months ago

@monibun : great point. I agree with you . I am not sure if anyone wants to advise on your point .

upvoted 1 times

Deeksha1234 4 years ago

Parquet

upvoted 2 times

shachar_ash 4 years, 3 months ago

Correct

upvoted 2 times

DRAG DROP -

You plan to create a table in an Azure Synapse Analytics dedicated SQL pool.

Data in the table will be retained for five years. Once a year, data that is older than five years will be deleted.

You need to ensure that the data is distributed evenly across partitions. The solution must minimize the amount of time required to delete old data.

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

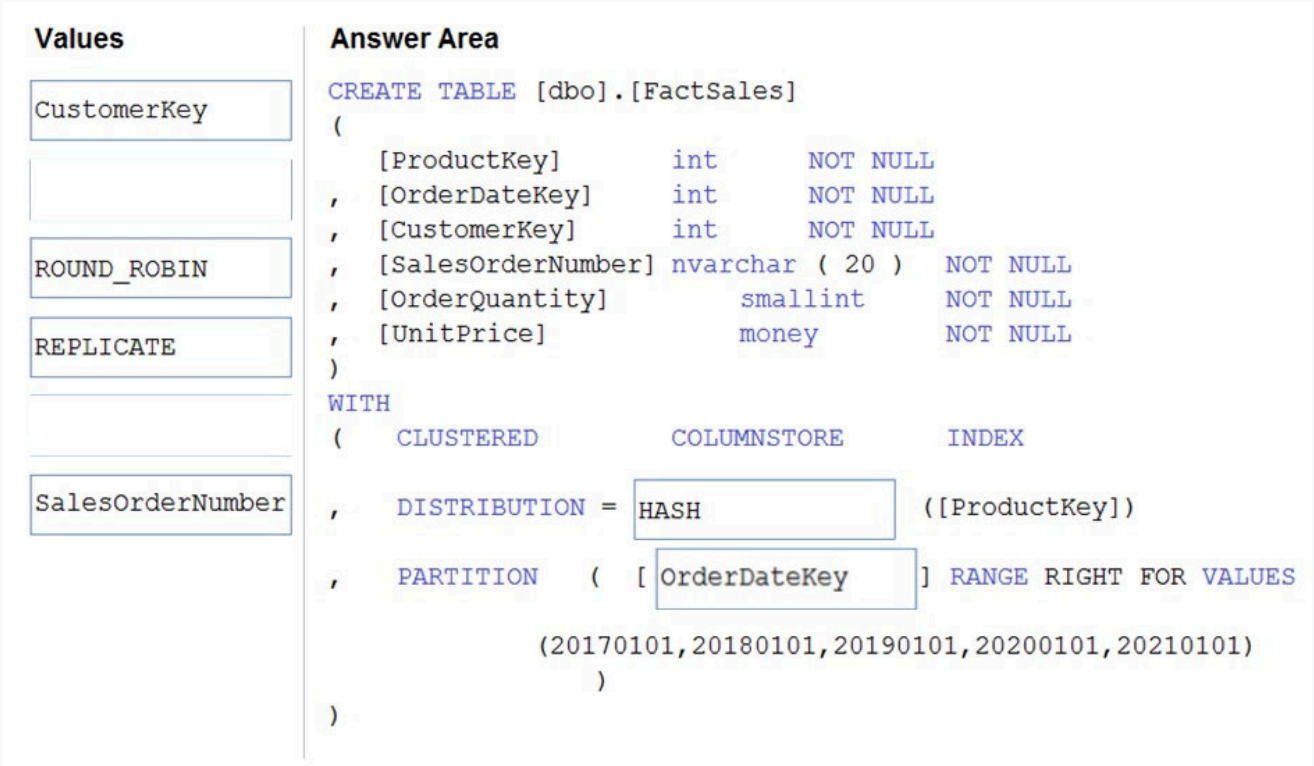
Values

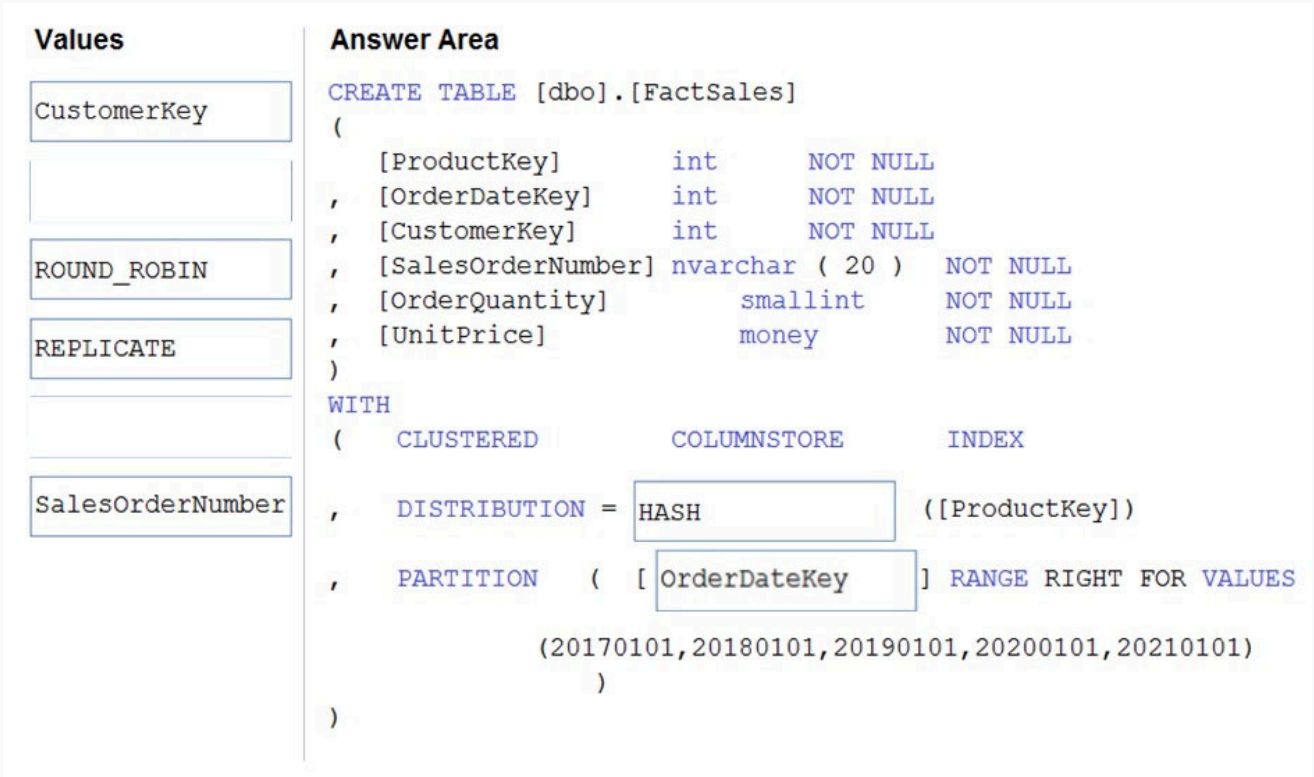
- CustomerKey
- HASH
- ROUND_ROBIN
- REPLICATE
- OrderDateKey
- SalesOrderNumber

Answer Area

```
CREATE TABLE [dbo].[FactSales]
(
   [ProductKey]       int      NOT NULL
,  [OrderDateKey]     int      NOT NULL
,  [CustomerKey]      int      NOT NULL
,  [SalesOrderNumber] nvarchar ( 20 )   NOT NULL
,  [OrderQuantity]        smallint     NOT NULL
,  [UnitPrice]            money        NOT NULL
)
WITH
(   CLUSTERED        COLUMNSTORE      INDEX

,   DISTRIBUTION =  [ Value ]  ([ProductKey])

,   PARTITION   (  [ Value ] ] RANGE RIGHT FOR VALUES

            (20170101,20180101,20190101,20200101,20210101)
               )
)
```

Suggested Answer:





Box 1: HASH -

Box 2: OrderDateKey - In most cases, table partitions are created on a date column. A way to eliminate rollbacks is to use Metadata Only operations like partition switching for data management. For example, rather than execute a DELETE statement to delete all rows in a table where the order_date was in October of 2001, you could partition your data early. Then you can switch out the partition with data for an empty partition from another table. Reference: https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

ClassMistress 3 years, 9 months ago

I think it is Hash because the question refer to a Fact table.

upvoted 23 times

hereiamken 2 years, 12 months ago

1. Hash -> Fact Table 2. DateKey -> for Partition

upvoted 17 times

dgerok Most Recent 1 year, 10 months ago

1) HASH, because this is a fact table 2) OrderDateKey (simply see the partition values - these are dates)

upvoted 2 times

kkk5566 2 years, 5 months ago

the syntax is ok only for HASH & Datekey

upvoted 3 times

kumarsunny 2 years, 5 months ago

Why not 'Product Key' for partition? can anyone explain me please.

upvoted 1 times

kumarsunny 2 years, 5 months ago

Why not 'Product Key' for partition? can anyone explain me please.Because partitioning on the date key will help in deleting older data quickly since the older records' partition can be moved to a different table and the table truncated.

upvoted 1 times

MJamesP 2 years, 4 months ago

Because partitioning on the date key will help in deleting older data quickly since the older records' partition can be moved to a different table and the table truncated.

upvoted 4 times

VittalManikonda 2 years, 7 months ago

if it is round robin, there is no key to specify, so hash

upvoted 2 times

SHENOOOO 3 years ago

The Answer is correct

upvoted 3 times

astone42 3 years ago

The answer is correct.

upvoted 2 times

DindaS 3 years ago

Should be Round Robin as the requirement is to have the data evenly. the second one should be on the date

upvoted 3 times

DindaS 3 years ago

Should be Round Robin as the requirement is to have the data evenly. the second one should be on the dateIt should, but they have given attributes. So only hash supports attributeYou would use Round-Robin for staging table and not Fact table.

upvoted 3 times

Ritik37 2 years, 11 months ago

It should, but they have given attributes. So only hash supports attribute

upvoted 1 times

auwia 2 years, 8 months ago

You would use Round-Robin for staging table and not Fact table.

upvoted 3 times

allagowf 3 years, 3 months ago

::= { DISTRIBUTION = HASH (distribution_column_name) | DISTRIBUTION = ROUND_ROBIN | DISTRIBUTION = REPLICATE } + fact table it's for sure ::: hash

upvoted 4 times

greenlever 3 years, 4 months ago

data is distributed evenly across partitions and data is deleted once a year not frequently. So it should be Round-robin distribution.

upvoted 3 times

Rajashekharc 3 years, 5 months ago

Cannot be Round Robin, the syntax of distribution for round robin don't mention/include Column Name. So it has to be HASH

upvoted 5 times

Deeksha1234 3 years, 6 months ago

Answer is correct

upvoted 2 times

Dicer 3 years, 6 months ago

should be round robin because of distributed evenly.syntax used is for HASH distribution.

upvoted 1 times

Dicer 3 years, 6 months ago

should be round robin because of distributed evenly.

upvoted 1 times

anks84 3 years, 5 months ago

syntax used is for HASH distribution.

upvoted 4 times

Franz58 3 years, 6 months ago

correct

upvoted 1 times

jebias 3 years, 9 months ago

I think the first answer should be Round-Robin as it should be distributed evenly. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distributeyes i think so toothe syntax is ok only for HASHWhile you are right, that Round-Robin guarantees an even distribution, it is only recommended to use on small tables < 2 GB (see your link). Using the Hash of the ProductKey will also allow for an even distribution but in a more efficient manner. Also, the Syntax here would be wrong if you would insert Round-Robin. As in that case it would only say: "DISTRIBUTION = ROUND-ROBIN" (no ProductKey)You are exactly righty@Feljoud : Thanks for the clarification. Even I opted for Roundrobin, considering the keywords = "distributed evenly", but that's incorrect.For small tables is recommended replicated, not round robin

upvoted 2 times

jebias 3 years, 9 months ago

I think the first answer should be Round-Robin as it should be distributed evenly. https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

upvoted 2 times

Muishkin 3 years, 9 months ago

yes i think so too

upvoted 1 times

Massy 3 years, 9 months ago

the syntax is ok only for HASH

upvoted 6 times

Feljoud Highly Voted 3 years, 9 months ago

While you are right, that Round-Robin guarantees an even distribution, it is only recommended to use on small tables < 2 GB (see your link). Using the Hash of the ProductKey will also allow for an even distribution but in a more efficient manner. Also, the Syntax here would be wrong if you would insert Round-Robin. As in that case it would only say: "DISTRIBUTION = ROUND-ROBIN" (no ProductKey)You are exactly righty@Feljoud : Thanks for the clarification. Even I opted for Roundrobin, considering the keywords = "distributed evenly", but that's incorrect.For small tables is recommended replicated, not round robin

upvoted 26 times

Feljoud Highly Voted 3 years, 9 months ago

While you are right, that Round-Robin guarantees an even distribution, it is only recommended to use on small tables < 2 GB (see your link). Using the Hash of the ProductKey will also allow for an even distribution but in a more efficient manner. Also, the Syntax here would be wrong if you would insert Round-Robin. As in that case it would only say: "DISTRIBUTION = ROUND-ROBIN" (no ProductKey)

upvoted 26 times

nefarious_smalls 3 years, 9 months ago

You are exactly righty

upvoted 1 times

sivva 3 years, 7 months ago

@Feljoud : Thanks for the clarification. Even I opted for Roundrobin, considering the keywords = "distributed evenly", but that's incorrect.

upvoted 2 times

dduque10 3 years, 2 months ago

For small tables is recommended replicated, not round robin

upvoted 1 times

HOTSPOT -

You have an Azure Data Lake Storage Gen2 service.

You need to design a data archiving solution that meets the following requirements:

✑ Data that is older than five years is accessed infrequently but must be available within one second when requested.

✑ Data that is older than seven years is NOT accessed.

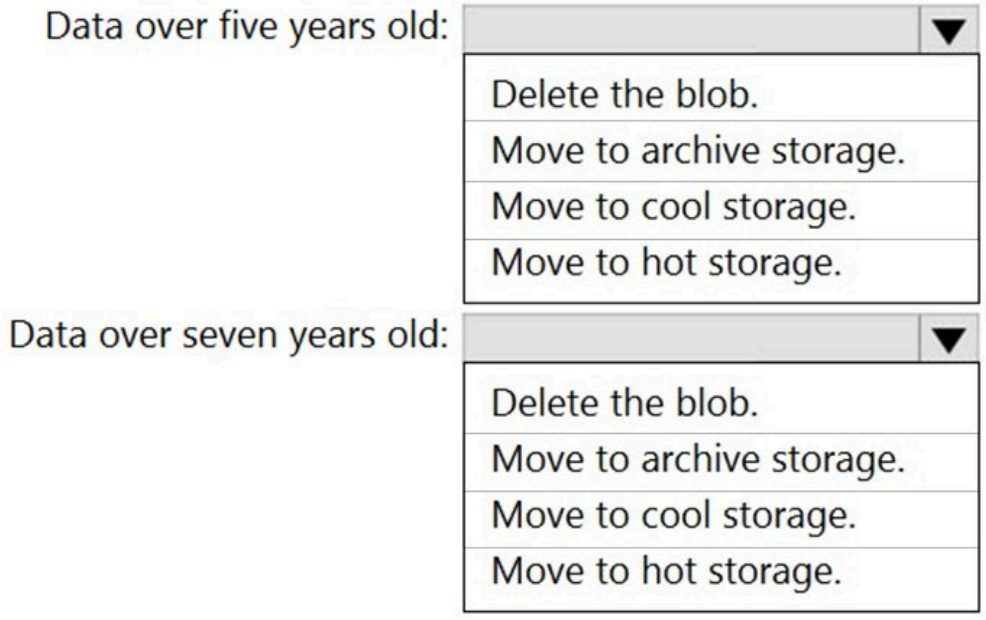
✑ Costs must be minimized while maintaining the required availability.

How should you manage the data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area



Suggested Answer:

Answer Area

Data over five years old:

- Delete the blob.
- Move to archive storage.
- **Move to cool storage.**
- Move to hot storage.

Data over seven years old:

- Delete the blob.
- **Move to archive storage.**
- Move to cool storage.
- Move to hot storage.

	Premium performance	Hot tier	Cool tier	Archive tier
Availability	99.9%	99.9%	99%	Offline
Availability (RA-GRS reads)	N/A	99.99%	99.9%	Offline
Usage charges	Higher storage costs, lower access, and transaction cost	Higher storage costs, lower access, and transaction costs	Lower storage costs, higher access, and transaction costs	Lowest storage costs, highest access, and transaction costs
Minimum storage duration	N/A	N/A	30 days[1]	180 days
Latency (Time to first byte)	Single-digit milliseconds	milliseconds	milliseconds	hours[2]

Answer Area



Box 1: Move to cool storage -

Box 2: Move to archive storage - Archive - Optimized for storing data that is rarely accessed and stored for at least 180 days with flexible latency requirements, on the order of hours. The following table shows a comparison of premium performance block blob storage, and the hot, cool, and archive access tiers.

Reference: https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blob-storage-tiers

sagur Highly Voted 4 years, 4 months ago

If "Data that is older than seven years is NOT accessed" then this data can be deleted to minimize the storage costs, right?

upvoted 40 times

sagur Highly Voted 4 years, 4 months ago

If "Data that is older than seven years is NOT accessed" then this data can be deleted to minimize the storage costs, right?I agree, should be deletedDeleting data older than 7 years is not an option available in the answer list. Becareful of the gotcha; 'Delete the blob' is an option but it would delete all the data, included the ones that are e.g. 5 years old. So you can't choose that answer. So the next best thing to do is to put it into archive.I'm confused by your comment. It clearly does state an option to delete the blob after 7 years.whet he meant to say is all the data be it in hot, cool or achieve resides in the blob. So if you delete the blob it will delete all the data be it 5 years or 7 years or more recent data in hot tier. Delete blob option is just to make it a tricky questionWhy it will delete data in hot tier? I have a blob. Blob between 5 and 7 go to cool, more than 7 delete.God no... A single piece of data could be a BLOB (Binary Large OBJECT).Be careful with wording. Any answer given must be an "Archiving solution" & Delete the blob is not an archiving solution.of course it is.Yes, right!Makes sense to meMost important part about that question is "Costs must be minimized while maintaining the required availability.". So deleting isn't an option, since it won't be available anymore. So ARCHIVE seems correctThe documentation, which is linked into the answer: "Archive tier - An offline tier optimized for storing data that is rarely accessed, and ..."

upvoted 40 times

Massy 4 years, 3 months ago

I agree, should be deletedDeleting data older than 7 years is not an option available in the answer list. Becareful of the gotcha; 'Delete the blob' is an option but it would delete all the data, included the ones that are e.g. 5 years old. So you can't choose that answer. So the next best thing to do is to put it into archive.I'm confused by your comment. It clearly does state an option to delete the blob after 7 years.whet he meant to say is all the data be it in hot, cool or achieve resides in the blob. So if you delete the blob it will delete all the data be it 5 years or 7 years or more recent data in hot tier. Delete blob option is just to make it a tricky questionWhy it will delete data in hot tier? I have a blob. Blob between 5 and 7 go to cool, more than 7 delete.God no... A single piece of data could be a BLOB (Binary Large OBJECT).Be careful with wording. Any answer given must be an "Archiving solution" & Delete the blob is not an archiving solution.of course it is.Yes, right!

upvoted 5 times

Massy 4 years, 3 months ago

I agree, should be deleted

upvoted 5 times

KashRaynardMorse 4 years, 3 months ago

Deleting data older than 7 years is not an option available in the answer list. Becareful of the gotcha; 'Delete the blob' is an option but it would delete all the data, included the ones that are e.g. 5 years old. So you can't choose that answer. So the next best thing to do is to put it into archive.

upvoted 30 times

KashRaynardMorse 4 years, 3 months ago

Deleting data older than 7 years is not an option available in the answer list. Becareful of the gotcha; 'Delete the blob' is an option but it would delete all the data, included the ones that are e.g. 5 years old. So you can't choose that answer. So the next best thing to do is to put it into archive.I'm confused by your comment. It clearly does state an option to delete the blob after 7 years.whet he meant to say is all the data be it in hot, cool or achieve resides in the blob. So if you delete the blob it will delete all the data be it 5 years or 7 years or more recent data in hot tier. Delete blob option is just to make it a tricky questionWhy it will delete data in hot tier? I have a blob. Blob between 5 and 7 go to cool, more than 7 delete.God no... A single piece of data could be a BLOB (Binary Large OBJECT).

upvoted 30 times

Boompiee 4 years, 3 months ago

I'm confused by your comment. It clearly does state an option to delete the blob after 7 years.

upvoted 2 times

Boompiee 4 years, 3 months ago

I'm confused by your comment. It clearly does state an option to delete the blob after 7 years.whet he meant to say is all the data be it in hot, cool or achieve resides in the blob. So if you delete the blob it will delete all the data be it 5 years or 7 years or more recent data in hot tier. Delete blob option is just to make it a tricky questionWhy it will delete data in hot tier? I have a blob. Blob between 5 and 7 go to cool, more than 7 delete.

upvoted 2 times

Aditya0891 4 years, 2 months ago

whet he meant to say is all the data be it in hot, cool or achieve resides in the blob. So if you delete the blob it will delete all the data be it 5 years or 7 years or more recent data in hot tier. Delete blob option is just to make it a tricky questionWhy it will delete data in hot tier? I have a blob. Blob between 5 and 7 go to cool, more than 7 delete.

upvoted 2 times

Aditya0891 4 years, 2 months ago

whet he meant to say is all the data be it in hot, cool or achieve resides in the blob. So if you delete the blob it will delete all the data be it 5 years or 7 years or more recent data in hot tier. Delete blob option is just to make it a tricky question

upvoted 2 times

temacc 4 years, 1 month ago

Why it will delete data in hot tier? I have a blob. Blob between 5 and 7 go to cool, more than 7 delete.

upvoted 1 times

vctrhugo 3 years, 1 month ago

God no... A single piece of data could be a BLOB (Binary Large OBJECT).

upvoted 2 times

shakes103 3 years, 7 months ago

Be careful with wording. Any answer given must be an "Archiving solution" & Delete the blob is not an archiving solution.

upvoted 7 times

shakes103 3 years, 7 months ago

Be careful with wording. Any answer given must be an "Archiving solution" & Delete the blob is not an archiving solution.of course it is.Yes, right!

upvoted 7 times

AnonymousJhb 3 years, 3 months ago

of course it is.

upvoted 4 times

Piantoni Most Recent 1 year, 9 months ago

Yes, right!

upvoted 1 times

noobprogrammer 4 years, 3 months ago

Makes sense to me

upvoted 2 times

vernillen 2 years, 7 months ago

Most important part about that question is "Costs must be minimized while maintaining the required availability.". So deleting isn't an option, since it won't be available anymore. So ARCHIVE seems correct

upvoted 2 times

Gikan 2 years, 6 months ago

The documentation, which is linked into the answer: "Archive tier - An offline tier optimized for storing data that is rarely accessed, and ..."

upvoted 1 times

PeteZaria 3 years, 10 months ago

Answer is correct. RE: Why is 'Delete the blob' not a valid option? Well I agree that 7 years may seem a long time to most who commented here BUT there is ABSOLUTELY NO mention to DELETE here, In several context NOT ACCESSED can easily refer to be drawn OFFLINE in ARCHIVE IOW: "Your data files may be stored in the archive access tier in Azure Blob storage based on different context needs. According to the Azure documentation: While a blob is in archive storage, the blob data is OFFLINE and CANNOT BE ACCESSED that is: read, copied, overwritten, or modified. You also can't take snapshots of a blob in archive storage. However, the blob METADA remains online and available, allowing you to list the blob and its properties. For blobs in archive, the only valid operations are GetBlobProperties, GetBlobMetadata, ListBlobs, SetBlobTier, and DeleteBlob. For more information about Azure Blob storage tiers, see the Azure documentation: https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blob-storage-tiers.

upvoted 9 times

PeteZaria 3 years, 10 months ago

Answer is correct. RE: Why is 'Delete the blob' not a valid option? Well I agree that 7 years may seem a long time to most who commented here BUT there is ABSOLUTELY NO mention to DELETE here, In several context NOT ACCESSED can easily refer to be drawn OFFLINE in ARCHIVE IOW: "Your data files may be stored in the archive access tier in Azure Blob storage based on different context needs. According to the Azure documentation: While a blob is in archive storage, the blob data is OFFLINE and CANNOT BE ACCESSED that is: read, copied, overwritten, or modified. You also can't take snapshots of a blob in archive storage. However, the blob METADA remains online and available, allowing you to list the blob and its properties. For blobs in archive, the only valid operations are GetBlobProperties, GetBlobMetadata, ListBlobs, SetBlobTier, and DeleteBlob. For more information about Azure Blob storage tiers, see the Azure documentation: https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blob-storage-tiers.The question states that data older than 7 years is NOT accessed. However, there is nothing stating that the company no longer needs the data. It should be archived because the company may still have a regulatory or legal requirement to keep this data.

upvoted 9 times

7082935 2 years ago

The question states that data older than 7 years is NOT accessed. However, there is nothing stating that the company no longer needs the data. It should be archived because the company may still have a regulatory or legal requirement to keep this data.

upvoted 1 times

tmz1 1 year, 9 months ago

I think the keyword in this question is "data archiving solution". Which means that option to delete anything is not considered

upvoted 2 times

rlnd2000 2 years, 1 month ago

In my opinion, the retention policy clearly states that any file or blob older than seven years is a violation. The policy specifies: "Data that is older than seven years is NOT accessed." Therefore, any data older than seven years should be deleted to avoid unnecessary costs, as we should not pay for storage we don't need. This approach aligns with the requirement to "minimize costs while maintaining the required availability." For data retention: Data older than five years should be moved to the cool storage tier. Data older than seven years should be deleted.

upvoted 2 times

otapi 2 years, 2 months ago

Why would keep any data in archive if it is NOT accessed? That would be just waste of money and possible violation of give data protection regulations.

upvoted 2 times

dgerok 2 years, 4 months ago

1) Cool, because it takes hours to access Archive data, but cool is cheaper than Hot 2) archive. For those, who thinks, that "Data that is older than seven years is NOT accessed" is equal to "the data can be deleted", you're fired :)

upvoted 4 times

dgerok 2 years, 4 months ago

1) Cool, because it takes hours to access Archive data, but cool is cheaper than Hot 2) archive. For those, who thinks, that "Data that is older than seven years is NOT accessed" is equal to "the data can be deleted", you're fired :)Damn Right!! With my 10 years exp never ever delete the data until its being asked by business owners for cost reduction. It will end up reduction in work force in form of you.

upvoted 4 times

Sr18 2 years, 2 months ago

Damn Right!! With my 10 years exp never ever delete the data until its being asked by business owners for cost reduction. It will end up reduction in work force in form of you.

upvoted 2 times

Little_Soap 2 years, 4 months ago

Data older than 7 years old is NOT accessed. Archived blobs are not accessible and deleting the blobs would be too aggressive so I'll go with archive for the second answer with the question as is.

upvoted 1 times

Alongi 2 years, 4 months ago

Cool Storage and Delete job, because in this way you can seriously minimize costs

upvoted 1 times

dataalex 2 years, 5 months ago

i think you guys are really overthinking it. just f....g delete data that will never be accesed.

upvoted 1 times

Charley92 2 years, 6 months ago

Costs must be minimized while maintaining the required availability. -- > DO NOT DELETE

upvoted 1 times

d046bc0 2 years, 8 months ago

correct. deleting files is not a type of archiving them (chatGPT)

upvoted 1 times

hassexat 2 years, 11 months ago

Cool & Archive No deletion of data here because question require to perform a data archiving solution

upvoted 1 times

kkk5566 2 years, 11 months ago

correct

upvoted 2 times

Saintu 2 years, 12 months ago

The question says: ✑ Provides the highest degree of data resiliency and this would be RA-GZRS with customer failover. Data is replicated to 3 zones in the primary region and async to a secondary region. If primary zone fails, there is read and write to other zones in the region. Microsoft only engages a failover if "original primary region is deemed unrecoverable within a reasonable amount of time due to a major disaster.", which is not the case here. The answer provided is correct

upvoted 1 times

auwia 3 years, 1 month ago

"Costs must be minimized while maintaining the required availability." ==> it means Move to archive storage"

upvoted 1 times

GodfreyMbizo 3 years, 6 months ago

The Given answer is correct,the big tip here is archiving solution,deleting is not archiving solution

upvoted 5 times

sumanthss 3 years, 9 months ago

Archive might be the right option as we need to retrive before we access the data. so we cannot access without retrieval.

upvoted 1 times

HOTSPOT -

You plan to create an Azure Data Lake Storage Gen2 account.

You need to recommend a storage solution that meets the following requirements:

✑ Provides the highest degree of data resiliency

✑ Ensures that content remains available for writes if a primary data center fails

What should you include in the recommendation? To answer, select the appropriate options in the answer area.

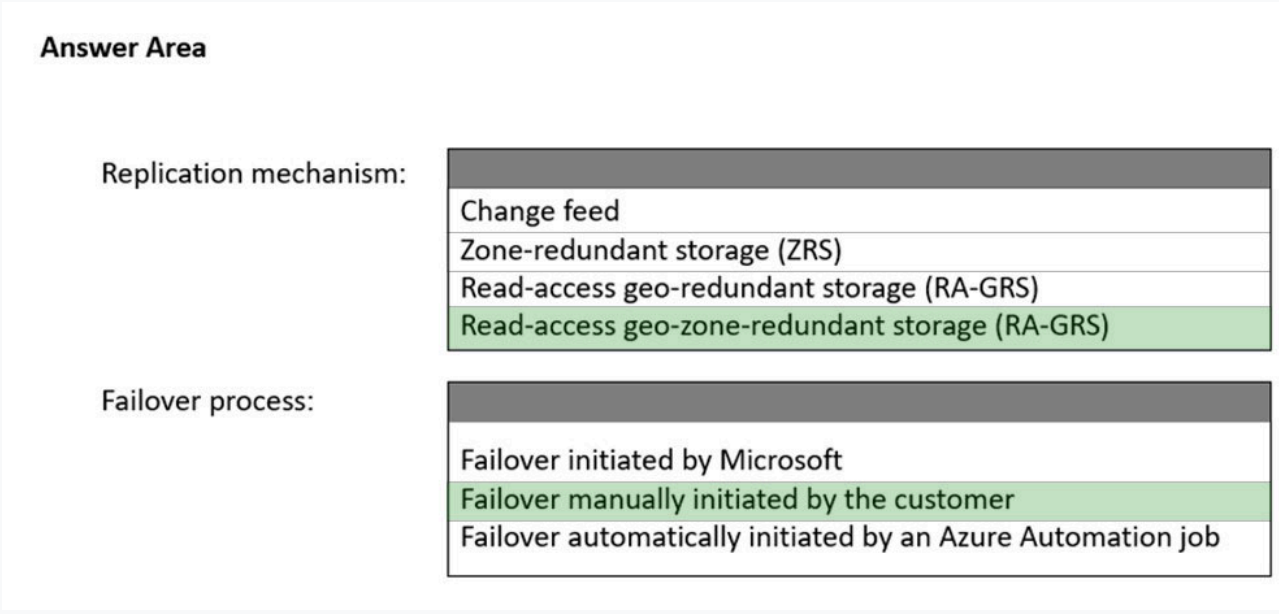
NOTE: Each correct selection is worth one point.

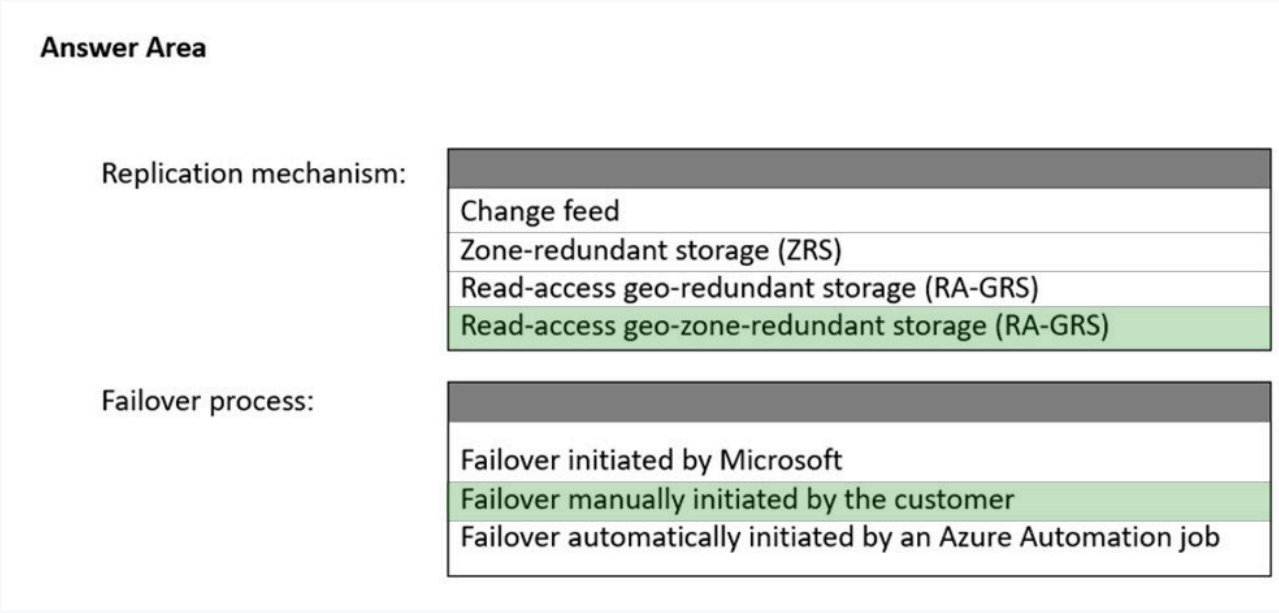
Hot Area:

Answer Area

Replication mechanism:	Change feed Zone-redundant storage (ZRS) Read-access geo-redundant storage (RA-GRS) Read-access geo-zone-redundant storage (RA-GRS)
Failover process:	Failover initiated by Microsoft Failover manually initiated by the customer Failover automatically initiated by an Azure Automation job

Suggested Answer:





Reference: https://docs.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance?toc=/azure/storage/blobs/toc.json
https://docs.microsoft.com/en-us/answers/questions/32583/azure-data-lake-gen2-disaster-recoverystorage-acco.html

aniagnesighile1 Highly Voted 3 years, 9 months ago

I am surprised you all missed this requirement 'Ensures that content remains available for writes if a primary data center fails'. RA-GRS and RAGZRS provide read access only after failover. The correct answer is ZRS as t=stated in the link below "Microsoft recommends using ZRS in the primary region for Azure Data Lake Storage Gen2 workloads." https://learn.microsoft.com/en-us/azure/storage/common/storage-redundancy?toc=%2Fazure%2Fstorage%2Fblobs%2Ftoc.json

upvoted 81 times

aniagnesighile1 Highly Voted 3 years, 9 months ago

I am surprised you all missed this requirement 'Ensures that content remains available for writes if a primary data center fails'. RA-GRS and RAGZRS provide read access only after failover. The correct answer is ZRS as t=stated in the link below "Microsoft recommends using ZRS in the primary region for Azure Data Lake Storage Gen2 workloads." https://learn.microsoft.com/en-us/azure/storage/common/storage-redundancy?toc=%2Fazure%2Fstorage%2Fblobs%2Ftoc.jsonyou're right. https://bluexp.netapp.com/blog/azure-anf-blg-azure-storage-replication-explained-lrs-zrs-grs-ra-grs#h_h3You can still write data to second region if first one fails. RA only allows you to read data in second region even if the first does not fail.What about the highest level of data resiliency that cant be in ZRS .... answer is RA-GZRSHaving "Provides the highest degree of data resiliency" as a requirement automatically sends you to the most redundant storage option.its data center failure not region failure. ZRS is correct

upvoted 81 times

Billybob0604 3 years, 8 months ago

you're right. https://bluexp.netapp.com/blog/azure-anf-blg-azure-storage-replication-explained-lrs-zrs-grs-ra-grs#h_h3

upvoted 1 times

vctrhugo 3 years, 1 month ago

You can still write data to second region if first one fails. RA only allows you to read data in second region even if the first does not fail.

upvoted 12 times

suvec 3 years, 4 months ago

What about the highest level of data resiliency that cant be in ZRS .... answer is RA-GZRS

upvoted 10 times

suvec 3 years, 4 months ago

What about the highest level of data resiliency that cant be in ZRS .... answer is RA-GZRSHaving "Provides the highest degree of data resiliency" as a requirement automatically sends you to the most redundant storage option.

upvoted 10 times

7082935 2 years ago

Having "Provides the highest degree of data resiliency" as a requirement automatically sends you to the most redundant storage option.

upvoted 1 times

SM94 2 years, 8 months ago

its data center failure not region failure. ZRS is correct

upvoted 3 times

chinomoreno 3 years, 11 months ago

Failover initiated by Microsoft. Customer-managed account failover is not yet supported in accounts that have a hierarchical namespace (Azure Data Lake Storage Gen2). To learn more, see Blob storage features available in Azure Data Lake Storage Gen2.RA-GZRS Failover initiated by Microsoft."Until the Microsoft-managed failover has completed, you won't have write access to your storage account." Hence, I guess it should NOT be Microsoft-managed failover. https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance?toc=%2Fazure%2Fstorage%2Fblobs%2Ftoc.json

upvoted 26 times

chinomoreno 3 years, 11 months ago

Failover initiated by Microsoft. Customer-managed account failover is not yet supported in accounts that have a hierarchical namespace (Azure Data Lake Storage Gen2). To learn more, see Blob storage features available in Azure Data Lake Storage Gen2.

upvoted 26 times

Gg2 3 years, 11 months ago

RA-GZRS Failover initiated by Microsoft."Until the Microsoft-managed failover has completed, you won't have write access to your storage account." Hence, I guess it should NOT be Microsoft-managed failover. https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance?toc=%2Fazure%2Fstorage%2Fblobs%2Ftoc.json

upvoted 22 times

Gg2 3 years, 11 months ago

RA-GZRS Failover initiated by Microsoft.

upvoted 22 times

Metaalverf 2 years, 9 months ago

"Until the Microsoft-managed failover has completed, you won't have write access to your storage account." Hence, I guess it should NOT be Microsoft-managed failover. https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance?toc=%2Fazure%2Fstorage%2Fblobs%2Ftoc.json

upvoted 1 times

prem__raj Most Recent 1 year, 6 months ago

Use customer-managed failover options to develop, test, and implement your disaster recovery plans. Do not rely on Microsoft-managed failover, which might only be used in extreme circumstances. A Microsoft-managed failover would be initiated for an entire physical unit, such as a region or a datacenter. It can't be initiated for individual storage accounts, subscriptions, or tenants. If you need the ability to selectively failover your individual storage accounts

upvoted 1 times

vishnumaya 1 year, 9 months ago

RA-GRS and failure initiated by Microsoft

upvoted 2 times

clowin 1 year, 10 months ago

Region wasn't mention in the question, so how did you guys come about it? Answer: ZRS and Failover initiated by Microsoft.

upvoted 1 times

ypan 2 years, 1 month ago

Failover Process: Failover initiated by Microsoft: In this scenario, Microsoft would initiate the failover process, which might not meet the requirement for maintaining availability for writes during a failure. CORRECT: Failover manually initiated by the customer: Allows the customer to control when the failover happens, ensuring that writes can be directed to the secondary region during a primary region failure. Failover automatically initiated by an Azure Automation job: This could provide automation for failover, but it may not give the same level of control as manual initiation by the customer.

upvoted 2 times

Dusica 2 years, 3 months ago

Zone redundant, microsoft initiated

upvoted 1 times

Alongi 2 years, 3 months ago

Should be: - ZRS because the question talk about on a failure on the primary region for Azure Data Lake Storage Gen2 Workloads, so a Zone. - Failover initiated by Microsoft.

upvoted 1 times

Elanche 2 years, 4 months ago

Type Failover Scope Use case Expected data loss HNS supported Customer-managed Storage account The storage service endpoints for the primary region become unavailable, but the secondary region is available. You received an Azure Advisory in which Microsoft advises you to perform a failover operation of storage accounts potentially affected by an outage. Yes Yes (In preview) Microsoft-managed Entire region or scale unit The primary region becomes completely unavailable due to a significant disaster, but the secondary region is available. Yes Yes

upvoted 1 times

j888 2 years, 5 months ago

"Provides the highest degree of data resiliency" and it didn't meantioned about the cost. I think GZRS will be the one to go for and it is microsoft initiated failover

upvoted 4 times

Anto____ 2 years, 6 months ago

RA-GZRS Failover manually initiated by the customer

upvoted 6 times

Souvik_79 2 years, 7 months ago

Why is the ET Azure community so divided for most of the Multiple Correct Questions? This was not the case for GCP ET community.

upvoted 1 times

Joanna0 2 years, 7 months ago

Failover initiated by Microsoft: In general, automatic failover initiated by Microsoft is the best option for most organizations, as it is the most reliable and efficient. However, if you need more control over the failover process, then you can choose to use automatic failover initiated by an Azure Automation job. RA-GZRS Available using RA-GZRS. RA-GZRS replicates data asynchronously to a secondary data center in a different Azure region. This means that there is a longer latency for replication, but it also means that your data is protected from regional outages. Additionally, RA-GZRS supports data compression and encryption, which can further improve storage efficiency.

upvoted 1 times

BitacTeam 2 years, 8 months ago

accroding to microsoft Documentation: https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance it mentioned 1-Customer-managed -- scope--> Storage account-- Case--> The storage service endpoints for the primary region become unavailable, but the secondary region is available. You received an Azure Advisory in which Microsoft advises you to perform a failover operation of storage accounts potentially affected by an outage. 2-Microsoft-managed-- Scope--> Entire region or scale unit -- Case--> The primary region becomes completely unavailable due to a significant disaster, but the secondary region is available. in the question it talks about Storage i.e. Data Center not a region failure. so the anwer is Customer managed

upvoted 1 times

ec255af 2 years, 8 months ago

anser is correct. RA-GZRS (Read Access Geo Zone Redundancy) provides read AND WRITE and has the highest availability (16 9's). https://learn.microsoft.com/en-us/azure/storage/common/storage-redundancy#geo-zone-redundant-storage For failover: iniciated by customer. In Microsofr case they MAY failover only if greater disaster happens. https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance?toc=%2Fazure%2Fstorage%2Fblobs%2Ftoc.json&bc=%2Fazure%2Fstorage%2Fblobs%2Fbreadcrumb%2Ftoc.json#microsoft-managed-failover

upvoted 2 times

Andrew_Chen 2 years, 9 months ago

GZRS Because it mentioned that when a data center failed, not the entire region. When a data centre failed, with ZRS, we can still read and write to the other two datacentres in the same region. So for the primary region, it must be ZRS. We still need a backup for the secondary region because it reqiures the highest degree of data resiliency. Customer Managed Failover, No Doubt!!!

upvoted 2 times

ellala 2 years, 10 months ago

Careful as this question will change pretty soon: "Customer-managed account failover for accounts that have a hierarchical namespace (Azure Data Lake Storage Gen2) is currently in PREVIEW and only supported in specific regions." Therefore it might soon be an option for ADLS2 as well.

upvoted 1 times

Question #59 *Topic 1*

You need to implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool.

You have a table that was created by using the following Transact-SQL statement.

Which two columns should you add to the table? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A.

B.

C.

D.

E.

```
CREATE TABLE [DB0].[DimProduct](
[ProductKey] [int] IDENTITY(1,1) NOT NULL,
[ProductSourceID] [int] NOT NULL,
[ProductNane] [nvarchar](100) NOT NULL,
[Color] [nvarchar] (15) NULL,
[SellStartDate] [date] NOT NULL,
[SellEnd0ate] [date] NULL,
[RowInsertedDateTime] [datetime] NOT NULL,
[RowipdatedDateTine] [datetime] NOT NULL,
[ETLAuditID] [int] NOT NULL
)
```

```
[EffectiveEndDate] [datetime] NULL,
```

```
[CurrentProductCategory] [nvarchar] (100) NOT NULL,
```

```
[ProductCategory] [nvarchar](100) NOT NULL,
```

```
[EffectiveStartDate] [datetime] NOT NULL,
```

```
[OriginalProductCategory] [nvarchar] (100) NOT NULL,
```

Suggested Answer:



CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	donna0@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-20

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	dc3@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-22

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	donna0@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-20

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	dc3@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-22

A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD. This type of tracking may be used for one or two columns in a dimension table. It is not common to use it for many members of the same table. It is often used in combination with Type 1 or Type 2 members.

Reference: https://k21academy.com/microsoft-azure/azure-data-engineer-dp203-q-a-day-2-live-session-review/

OldSchool Highly Voted 3 years, 8 months ago

B & E is correct answer

upvoted 22 times

erhard Highly Voted 3 years, 9 months ago

If BE is correct, then CE is also correct.

upvoted 13 times

erhard 3 years, 9 months ago

If BE is correct, then CE is also correct.Both BE and CE can be correct solutions. However, you should choose the one that is most descriptive to any business user who is consuming this data. So, describing each column as "Current" or "Original" is the better design.I agree with you; however, in the table's design, there's already a column called [RowInsertedDateTime], which implies that it is tracking the insertion of the ProductCategory values. So, it would make no sense for one of the answers to be Option B. Therefore, I believe the correct answers are C and E.

upvoted 13 times

7082935 2 years ago

Both BE and CE can be correct solutions. However, you should choose the one that is most descriptive to any business user who is consuming this data. So, describing each column as "Current" or "Original" is the better design.I agree with you; however, in the table's design, there's already a column called [RowInsertedDateTime], which implies that it is tracking the insertion of the ProductCategory values. So, it would make no sense for one of the answers to be Option B. Therefore, I believe the correct answers are C and E.

upvoted 2 times

7082935 2 years ago

Both BE and CE can be correct solutions. However, you should choose the one that is most descriptive to any business user who is consuming this data. So, describing each column as "Current" or "Original" is the better design.

upvoted 2 times

renan_ineu Most Recent 1 year, 11 months ago

I agree with you; however, in the table's design, there's already a column called [RowInsertedDateTime], which implies that it is tracking the insertion of the ProductCategory values. So, it would make no sense for one of the answers to be Option B. Therefore, I believe the correct answers are C and E.

upvoted 1 times

sergio_eduardo 2 years, 2 months ago

BE is correct but CE is also correct

upvoted 2 times

dgerok 2 years, 4 months ago

B & E is correct answer

upvoted 1 times

hassexat 2 years, 11 months ago

B & E is the correct answer

upvoted 1 times

kkk5566 2 years, 11 months ago

B & E is correct answer for SCD 3

upvoted 2 times

anks84 3 years, 11 months ago

Answer is correct

upvoted 4 times

dom271219 3 years, 11 months ago

Correct but SCD is always type 2. Type 3 is not SCD.

upvoted 1 times

dom271219 3 years, 11 months ago

Correct but SCD is always type 2. Type 3 is not SCD.A SCD can be of type 0, 1, 2, 3 and so on, please read the documentation

upvoted 1 times

gerrie1979 3 years, 9 months ago

A SCD can be of type 0, 1, 2, 3 and so on, please read the documentation

upvoted 12 times

MuhammadK 3 years, 11 months ago

Given answer is correct

upvoted 1 times

DRAG DROP -

You have an Azure subscription.

You plan to build a data warehouse in an Azure Synapse Analytics dedicated SQL pool named pool1 that will contain staging tables and a dimensional model.

Pool1 will contain the following tables.

You need to design the table storage for pool1. The solution must meet the following requirements:

✑ Maximize the performance of data loading operations to Staging.WebSessions.

✑ Minimize query times for reporting queries against the dimensional model.

Which type of table distribution should you use for each table? To answer, drag the appropriate table distribution types to the correct tables. Each table distribution type may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

Name	Number of rows	Update frequency	Description
Common. Date	7,300	New rows inserted yearly	• Contains one row per date for the last 20 years • Contains columns named Year, Month, Quarter, and IsWeekend
Marketing.WebSessions	1,500,500,000	Hourly inserts and updates	Fact table that contains counts of and updates sessions and page views, including foreign key values for date, channel, device, and medium
Staging.WebSessions	300,000	Hourly truncation and inserts	Staging table for web session data, truncation and including descriptive fields for inserts channel, device, and medium

Table distribution types

Hash

Replicated

Round-robin

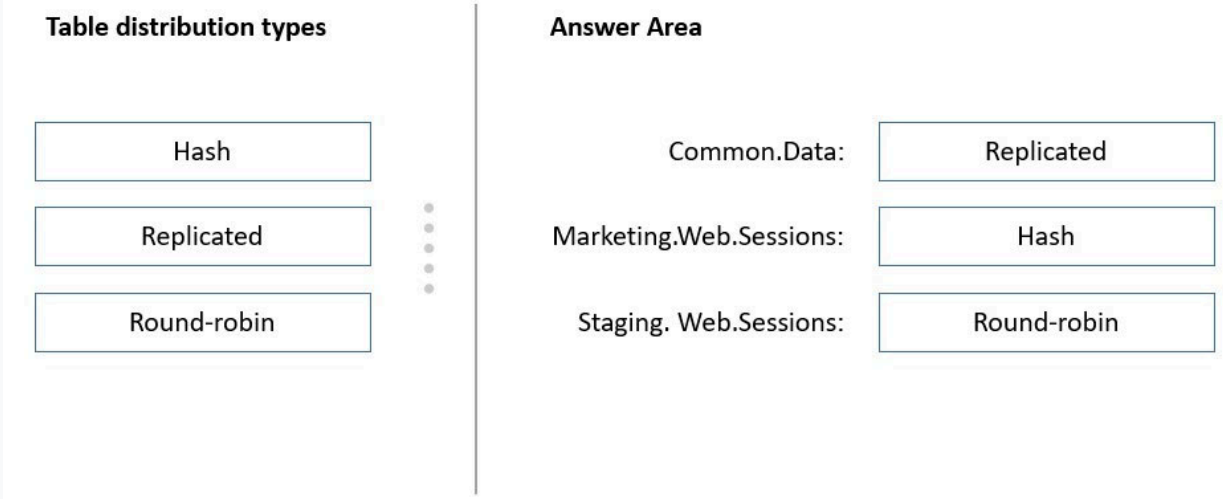
Answer Area

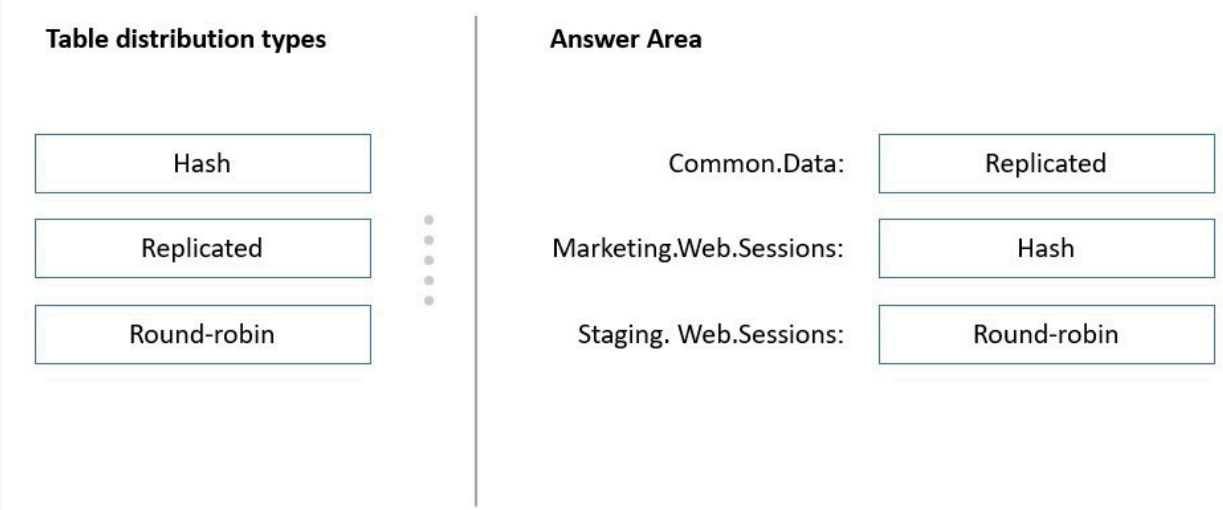
Common.Data: [ ]

Marketing.Web.Sessions: [ ]

Staging. Web.Sessions: [ ]

Suggested Answer:





Box 1: Replicated - The best table storage option for a small table is to replicate it across all the Compute nodes.

Box 2: Hash - Hash-distribution improves query performance on large fact tables.

Box 3: Round-robin - Round-robin distribution is useful for improving loading speed. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

anks84 Highly Voted 3 years, 11 months ago
Replicated (Because its a Dimension table) Hash (Fact table with High volume of data) Round-Robin (Staging table)
upvoted 37 times

Piantoni Most Recent 1 year, 9 months ago
Correct
upvoted 1 times

ff5037f 1 year, 9 months ago
CORRECT
upvoted 1 times

Alongi 2 years, 4 months ago
Correct
upvoted 1 times

kkk5566 2 years, 11 months ago
Replicated (Because its a Dimension table) Hash (Fact table with High volume of data) Round-Robin (Staging table)
upvoted 3 times

spramanik_de 3 years, 1 month ago

Shouldn't the fact table has round-robin as well since we need to prioritize data loading, hash will definitely improve the read query performance but will impact the data load speed?

upvoted 2 times

spramanik_de 3 years, 1 month ago

Shouldn't the fact table has round-robin as well since we need to prioritize data loading, hash will definitely improve the read query performance but will impact the data load speed?The requirements stated are: - Maximize the performance of data loading operations to Staging.WebSessions. - Minimize query times for reporting queries against the dimensional model. So only the staging table needs fast data loading.

upvoted 2 times

wanchihh 3 years ago

The requirements stated are: - Maximize the performance of data loading operations to Staging.WebSessions. - Minimize query times for reporting queries against the dimensional model. So only the staging table needs fast data loading.

upvoted 1 times

SHENOOOO 3 years, 6 months ago

Given answer is correct

upvoted 4 times

DindaS 3 years, 6 months ago

The dimension should be a replicated one. so that it will be available in all noes for a better performance fact table should be a HASH staging table is always Round_ROBIN

upvoted 1 times

nb1000 3 years, 7 months ago

correct

upvoted 1 times

allagowf 3 years, 10 months ago

the giving answer is correct, as the requirements

upvoted 4 times

HOTSPOT -

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a table named FactInternetSales that will be a large fact table in a dimensional model. FactInternetSales will contain 100 million rows and two columns named SalesAmount and OrderQuantity. Queries executed on FactInternetSales will aggregate the values in SalesAmount and OrderQuantity from the last year for a specific product. The solution must minimize the data size and query execution time.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE TABLE [dbo].[FactInternetSales]
( [ProductKey] int NOT NULL
, [OrderDateKey] int NOT NULL
, [CustomerKey] int NOT NULL
, [PromotionKey] int NOT NULL
, [SalesOrderNumber] nvarchar(20) NOT NULL
, [OrderQuantity] smallint NOT NULL
, [UnitPrice] money NOT NULL
, [SalesAmount] money NOT NULL
)
WITH
```

(CLUSTERED COLUMNSTORE INDEX
(CLUSTERED INDEX ([OrderDateKey])
(HEAP
(INDEX on [ProductKey]

```
, DISTRIBUTION =
);
```

Hash([OrderDateKey])
Hash([ProductKey])
REPLICATE
ROUND_ROBIN

Suggested Answer:

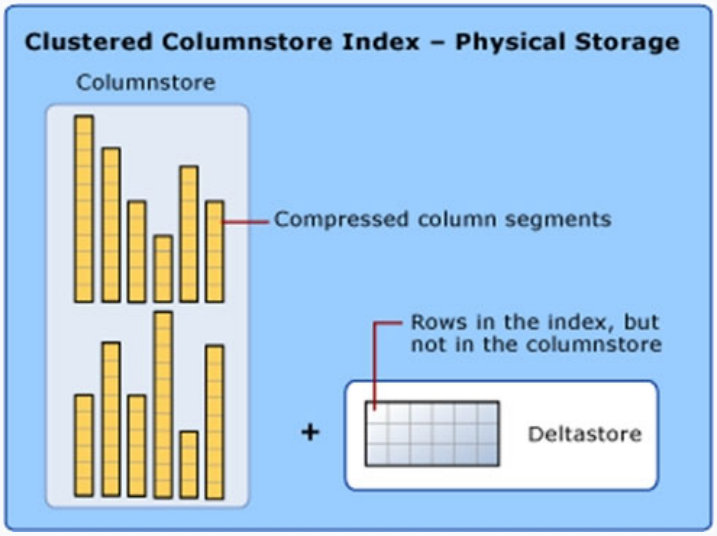
Answer Area

```
CREATE TABLE [dbo].[FactInternetSales]
( [ProductKey] int NOT NULL
, [OrderDateKey] int NOT NULL
, [CustomerKey] int NOT NULL
, [PromotionKey] int NOT NULL
, [SalesOrderNumber] nvarchar(20) NOT NULL
, [OrderQuantity] smallint NOT NULL
, [UnitPrice] money NOT NULL
, [SalesAmount] money NOT NULL
)
WITH
```

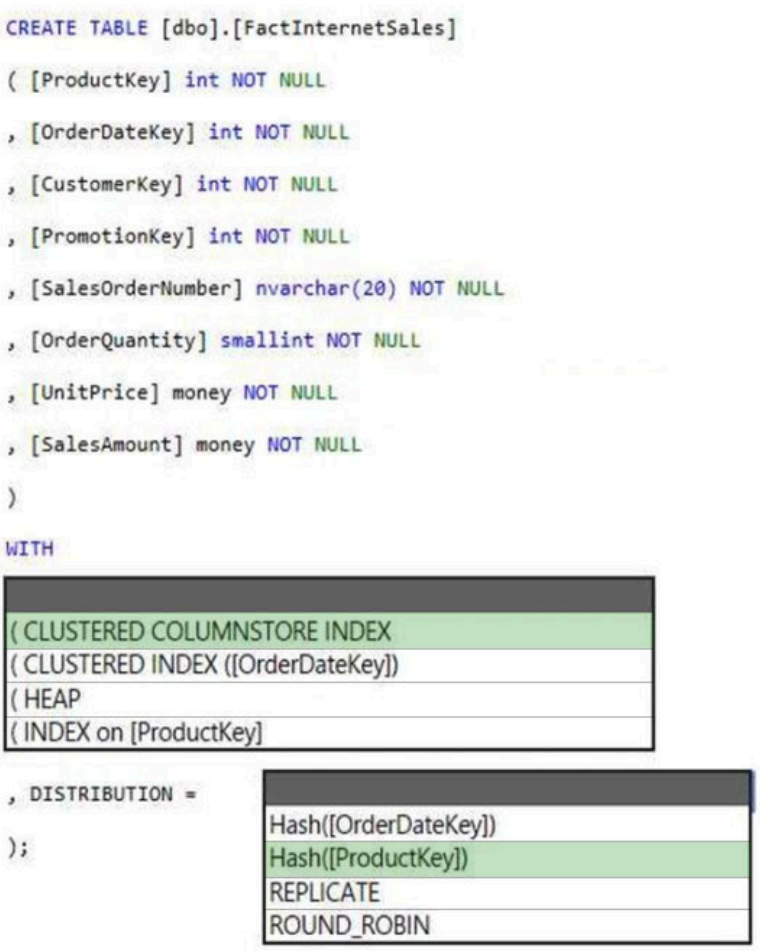
(CLUSTERED COLUMNSTORE INDEX
(CLUSTERED INDEX ([OrderDateKey])
(HEAP
(INDEX on [ProductKey]

```
, DISTRIBUTION =
);
```

Hash([OrderDateKey])
Hash([ProductKey])
REPLICATE
ROUND_ROBIN



Answer Area



Box 1: (CLUSTERED COLUMNSTORE INDEX

CLUSTERED COLUMNSTORE INDEX - Columnstore indexes are the standard for storing and querying large data warehousing fact tables. This index uses column-based data storage and query processing to achieve gains up to 10 times the query performance in your data warehouse over traditional row-oriented storage. You can also achieve gains up to 10 times the data compression over the uncompressed data size. Beginning with SQL Server 2016 (13.x) SP1, columnstore indexes enable operational analytics: the ability to run performant real-time analytics on a transactional workload. Note: Clustered columnstore index A clustered columnstore index is the physical storage for the entire table.

To reduce fragmentation of the column segments and improve performance, the columnstore index might store some data temporarily into a clustered index called a deltastore and a B-tree list of IDs for deleted rows. The deltastore operations are handled behind the scenes. To return the correct query results, the clustered columnstore index combines query results from both the columnstore and the deltastore. Box 2: HASH([ProductKey]) A hash distributed table distributes rows based on the value in the distribution column. A hash distributed table is designed to achieve high performance for queries on large tables. Choose a distribution column with data that distributes evenly Incorrect: * Not HASH([OrderDateKey]). Is not a date column. All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work * A replicated table has a full copy of the table available on every Compute node. Queries run fast on replicated tables since joins on replicated tables don't require data movement. Replication requires extra storage, though, and isn't practical for large tables. * A round-robin table distributes table rows evenly across all distributions. The rows are distributed randomly. Loading data into a round-robin table is fast. Keep in mind that queries can require more data movement than the other distribution methods. Reference: https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-overview https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overview https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

ted0809 Highly Voted 3 years, 9 months ago

you don't hash the date.. never..Usually you don't use the hash date, because if the query is made daily, all record contains the same date, so the same node will do the job. In this example " from the last year for a specific product", so we have 365 separate node from date, but we have only one product!!This sentence helped me soooo much - Hash the date NEVER.. thank youIf you cant read don't bother commentingCheck the question 39,

which is quite similar: SELECT SupplierKey, StockItemKey, COUNT(*) FROM FactPurchase WHERE DateKey >= 20210101 AND DateKey <= 20210131 GROUP By SupplierKey, StockItemKey B. hash-distributed on PurchaseKey D. hash-distributed on DateKey Bear in mind the type of the orderDateKey column, it's int, not date, not datetime... Doesn't it make a difference? In this case I'd choose the ProductKey since it says queries will be done "for specific products", but in the previous case the WHERE clause uses the DateKey field.you don't hash dateKey because same dateKey goes to same distribution, thus only a few out of total 60 distributions are working in a query. Instead, within each distribution we can partition by dateKey, so that all distributions are used in a query and each distribution can filter data efficiently because of dateKey partitionDate key is same as date, it would have distributed by date and that is a no no

upvoted 55 times

ted0809 Highly Voted 3 years, 9 months ago

you don't hash the date.. never..

upvoted 55 times

Gikan 2 years, 6 months ago

Usually you don't use the hash date, because if the query is made daily, all record contains the same date, so the same node will do the job. In this example " from the last year for a specific product", so we have 365 separate node from date, but we have only one product!!

upvoted 3 times

Data_Analytics 2 years, 10 months ago

This sentence helped me soooo much - Hash the date NEVER.. thank youIf you cant read don't bother commenting

upvoted 6 times

Data_Analytics 2 years, 10 months ago

This sentence helped me soooo much - Hash the date NEVER.. thank you

upvoted 6 times

Euanm28 2 years, 10 months ago

If you cant read don't bother commenting

upvoted 3 times

tlb_20 2 years, 4 months ago

Check the question 39, which is quite similar:

SELECT SupplierKey, StockItemKey, COUNT(*) FROM FactPurchase WHERE DateKey >= 20210101 AND DateKey <= 20210131 GROUP By SupplierKey, StockItemKey

B. hash-distributed on PurchaseKey D. hash-distributed on DateKey

Bear in mind the type of the orderDateKey column, it's int, not date, not datetime... Doesn't it make a difference?

In this case I'd choose the ProductKey since it says queries will be done "for specific products", but in the previous case the WHERE clause uses the DateKey field.

upvoted 1 times

tlb_20 2 years, 4 months ago

Check the question 39, which is quite similar: SELECT SupplierKey, StockItemKey, COUNT(*) FROM FactPurchase WHERE DateKey >= 20210101 AND DateKey <= 20210131 GROUP By SupplierKey, StockItemKey B. hash-distributed on PurchaseKey D. hash-distributed on DateKey Bear in mind the type of the orderDateKey column, it's int, not date, not datetime... Doesn't it make a difference? In this case I'd choose the ProductKey since it says queries will be done "for specific products", but in the previous case the WHERE clause uses the DateKey field.you don't hash dateKey because same dateKey goes to same distribution, thus only a few out of total 60 distributions are working in a query. Instead, within each distribution we can partition by dateKey, so that all distributions are used in a query and each distribution can filter data efficiently because of dateKey partitionDate key is same as date, it would have distributed by date and that is a no no

upvoted 1 times

hypersam 1 year, 7 months ago

you don't hash dateKey because same dateKey goes to same distribution, thus only a few out of total 60 distributions are working in a query. Instead, within each distribution we can partition by dateKey, so that all distributions are used in a query and each distribution can filter data efficiently because of dateKey partition

upvoted 1 times

Dusica 2 years, 3 months ago

Date key is same as date, it would have distributed by date and that is a no no

upvoted 1 times

Lestrang 3 years, 6 months ago

By using the product key as the distribution key, the data for a specific product will be stored on the same node, allowing for faster aggregation of the values in SalesAmount and OrderQuantity for that product.

upvoted 11 times

Pey1nkh Most Recent 1 year, 6 months ago

correct answer! why Clustered Columnstore Index ? cause its best for large fact tables. Reduces data storage by up to 10x. Speeds up aggregation queries.

upvoted 1 times

Dusica 2 years, 3 months ago

distriubtion hash(product key) clustered index (date key)

upvoted 1 times

Dusica 2 years, 3 months ago

distriubtion hash(product key) clustered index (date key)Answer to 1. Question and why other options are wrong: Correct Answer: CLUSTERED COLUMNSTORE INDEX Reasoning: A clustered columnstore index (CCI) is the default and most efficient index for large fact tables in Synapse Analytics because it compresses data significantly, reducing storage and improving query performance for analytical workloads. It is optimized for aggregations, such as summing up SalesAmount and OrderQuantity, which are typical in analytical queries. Why Others Are Incorrect: CLUSTERED INDEX ([OrderDateKey]): A clustered index is better suited for transactional systems, not for analytical systems with large fact tables. HEAP: Heap tables lack any indexing, making them inefficient for queries that require aggregation or filtering on large datasets. INDEX on [ProductKey]: A non-clustered index might help in specific queries but doesn't optimize overall performance for large-scale analytics.

upvoted 1 times

17lan 1 year, 8 months ago

Answer to 1. Question and why other options are wrong: Correct Answer: CLUSTERED COLUMNSTORE INDEX Reasoning: A clustered columnstore index (CCI) is the default and most efficient index for large fact tables in Synapse Analytics because it compresses data significantly, reducing storage and improving query performance for analytical workloads. It is optimized for aggregations, such as summing up SalesAmount and OrderQuantity, which are typical in analytical queries. Why Others Are Incorrect: CLUSTERED INDEX ([OrderDateKey]): A clustered index is better suited for transactional systems, not for analytical systems with large fact tables. HEAP: Heap tables lack any indexing, making them inefficient for queries that require aggregation or filtering on large datasets. INDEX on [ProductKey]: A non-clustered index might help in specific queries but doesn't optimize overall performance for large-scale analytics.

upvoted 1 times

dgerok 2 years, 4 months ago

The answer is right

upvoted 1 times

kkk5566 2 years, 11 months ago

Clustered columnstore Hash(ProductKey)

upvoted 6 times

smsme323 3 years, 10 months ago

https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute#choose-a-distribution-column-with-data-that-distributes-evenly To balance the parallel processing, select a distribution column or set of columns that: Has many unique values. The distribution column(s) can have duplicate values. All rows with the same value are assigned to the same distribution. Since there are 60 distributions, some distributions can have > 1 unique values while others may end with zero values. Does not have NULLs, or has only a few NULLs. For an extreme example, if all values in the distribution column(s) are NULL, all the rows are assigned to the same distribution. As a result, query processing is skewed to one distribution, and does not benefit from parallel processing. Is not a date column. All data for the same date lands in the same distribution, or will cluster records by date. If several users are all filtering on the same date (such as today's date), then only 1 of the 60 distributions do all the processing work. Ans: Hash(ProductKey)

upvoted 9 times

Phund 3 years, 11 months ago

must hash on OrderDateKey because that field was not a date and it was used for filter condition "from the last year for a specific product"https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute the example in the sample-test is in this page and they used productkey for hashing, so yeah, the answer is productkey

upvoted 8 times

Phund 3 years, 11 months ago

must hash on OrderDateKey because that field was not a date and it was used for filter condition "from the last year for a specific product"

upvoted 8 times

Lestrang 3 years, 11 months ago

https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute the example in the sample-test is in this page and they used productkey for hashing, so yeah, the answer is productkey

upvoted 14 times

anks84 3 years, 11 months ago

correct

upvoted 4 times

Question #62 *Topic 1*

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. Table1 contains the following:

- ✑ One billion rows
- ✑ A clustered columnstore index
- ✑ A hash-distributed column named Product Key
- ✑ A column named Sales Date that is of the date data type and cannot be null

Thirty million rows will be added to Table1 each month.

You need to partition Table1 based on the Sales Date column. The solution must optimize query performance and data loading.

How often should you create a partition?

A. once per month

B. once per year

C. once per day

D. once per week

Suggested Answer: *B*

Need a minimum 1 million rows per distribution. Each table is 60 distributions. 30 millions rows is added each month. Need 2 months to get a minimum of 1 million rows per distribution in a new partition. Note: When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributions. Any partitioning added to a table is in addition to the distributions created behind the scenes. Using this example, if the sales fact table contained 36 monthly partitions, and given that a dedicated SQL pool has 60 distributions, then the sales fact table should contain 60 million rows per month, or 2.1 billion rows when all months are populated. If a table contains fewer than the recommended minimum number of rows per partition, consider using fewer partitions in order to increase the number of rows per partition. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition

Community vote distribution



vrodriguesp Highly Voted 3 years, 8 months ago

Remembering that we have data splitted in distribution (60 nodes) and considering that we Need a MINMIUM 1 million rows per distribution, we have:

A. once per month = 30 milion / 60 = 500k record per partition B. once per year = 360 milion / 60 = 6 milion record per partition C. once per day = about 1 milion / 60 = 16k record per partition D. once per week =about 7.5 milion / 60 = 125k record per partition

correct should be B

upvoted 99 times

vrodriguesp Highly Voted 3 years, 8 months ago

Remembering that we have data splitted in distribution (60 nodes) and considering that we Need a MINMIUM 1 million rows per distribution, we have: A. once per month = 30 milion / 60 = 500k record per partition B. once per year = 360 milion / 60 = 6 milion record per partition C. once per day = about 1 milion / 60 = 16k record per partition D. once per week =about 7.5 milion / 60 = 125k record per partition correct should be Bactually to be more accurate, I should have written record per distribution. We have 1 milion rows per distribution and 60 milion rows per partition..I think you left out the fact that the table already has 1 billion records. This will change your calculations.You should consider the partition, not the tableagree, the question ask about partition, not the table distributionno, we should consider everything including the table in order to come up with the most efficient solutionI think you mix up the concept of distribution and partition, we always do partitions with date, distribution with Product Key. I think you over think, no need to think about the calculationsorry may be you are rightI have contributor access which is purchased for $47.99 but no downloadable PDF with questions and explanations received yet.Contact their Customer Care for answer with your complain and they will respond to you.Following this logic. Although A is less than 1m requirement, but it is the closest on to 1m. B met the requirement, but is is way too big hence loading to memory is slower than 500k one, already A may result in higher number of partitions, which is larger storage space.

upvoted 99 times

vrodriguesp 3 years, 6 months ago

actually to be more accurate, I should have written record per distribution. We have 1 milion rows per distribution and 60 milion rows per partition..

upvoted 2 times

yogiazaad 3 years, 6 months ago

I think you left out the fact that the table already has 1 billion records. This will change your calculations.

upvoted 9 times

yogiazaad 3 years, 6 months ago

I think you left out the fact that the table already has 1 billion records. This will change your calculations.You should consider the partition, not the tableagree, the question ask about partition, not the table distributionno, we should consider everything including the table in order to come up with the most efficient solution

upvoted 9 times

CCCool77 3 years, 6 months ago

You should consider the partition, not the tableagree, the question ask about partition, not the table distributionno, we should consider everything including the table in order to come up with the most efficient solution

upvoted 5 times

CCCool77 3 years, 6 months ago

You should consider the partition, not the table

upvoted 5 times

hiyoww 3 years ago

agree, the question ask about partition, not the table distribution

upvoted 2 times

d47320d 2 years, 3 months ago

no, we should consider everything including the table in order to come up with the most efficient solution

upvoted 1 times

hiyoww 3 years ago

I think you mix up the concept of distribution and partition, we always do partitions with date, distribution with Product Key. I think you over think, no need to think about the calculation

upvoted 1 times

hiyoww 3 years ago

I think you mix up the concept of distribution and partition, we always do partitions with date, distribution with Product Key. I think you over think, no need to think about the calculationsorry may be you are rightI have contributor access which is purchased for $47.99 but no downloadable PDF with questions and explanations received yet.Contact their Customer Care for answer with your complain and they will respond to you.

upvoted 1 times

hiyoww 3 years ago

sorry may be you are rightI have contributor access which is purchased for $47.99 but no downloadable PDF with questions and explanations received yet.Contact their Customer Care for answer with your complain and they will respond to you.

upvoted 1 times

hiyoww 3 years ago

sorry may be you are right

upvoted 1 times

Gcplearner8888 3 years ago

I have contributor access which is purchased for $47.99 but no downloadable PDF with questions and explanations received yet.

upvoted 3 times

Gcplearner8888 3 years ago

I have contributor access which is purchased for $47.99 but no downloadable PDF with questions and explanations received yet.Contact their Customer Care for answer with your complain and they will respond to you.

upvoted 3 times

Paulkuzzio 2 years, 10 months ago

Contact their Customer Care for answer with your complain and they will respond to you.

upvoted 1 times

gggmaaster 2 years, 9 months ago

Following this logic. Although A is less than 1m requirement, but it is the closest on to 1m. B met the requirement, but is is way too big hence loading to memory is slower than 500k one, already A may result in higher number of partitions, which is larger storage space.

upvoted 3 times

anks84 3 years, 11 months ago

Correct, Considering the high volume of data, for faster queries its recommended to create fewer partitions. " If a table contains fewer than the recommended minimum number of rows per partition(i.e. 60 million rows per month for 60 distributed partitions, consider using fewer partitions in

order to increase the number of rows per partition."

upvoted 11 times

ShdwZephyr Most Recent 1 year, 5 months ago

I would say its option A. - Efficiently load new data (in this case, 30 million new rows each month). - Improve query performance for time-based queries.

upvoted 1 times

prem__raj 1 year, 6 months ago

I think it should be by month. There are 12 months in a year. The table already has 1 billion rows, we have 60 distribution nodes. So 1 billion/60 = 16.6 Million approx per node. We get 30 million data per month, we distribute it to 60 nodes, that is 30 Million/60 = 50,000 rows per node. Now we have 16.6 million + 50,000 rows per node that's 16,650,00 rows. Now if we partition by month that's 16,650,00/12 we'll have close to 138750 which is 1.3 million. And the optimal size per partition is close to 1 million. Therefore A

upvoted 1 times

abhi_11kr1 1 year, 8 months ago

Creating a partition once per month strikes the right balance between performance, manageability, and data loading efficiency

upvoted 1 times

mavdplas 1 year, 9 months ago

I think the loading performance over here is the determining factor. With a partition by month the date can be quickly swapped into the table with a partition swap. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition

upvoted 1 times

renan_ineu 1 year, 11 months ago

The table currently has 1 billion rows. With 60 partitions, this means there are approximately 16,666,666 rows per partition.

Each month, an additional 30 million rows will be added. After the first month, the table will have 1.03 billion rows, resulting in approximately 17,166,666 rows per partition.

I recommend going with Option A: partitioning once per month.

upvoted 1 times

renan_ineu 1 year, 11 months ago

The table currently has 1 billion rows. With 60 partitions, this means there are approximately 16,666,666 rows per partition. Each month, an additional 30 million rows will be added. After the first month, the table will have 1.03 billion rows, resulting in approximately 17,166,666 rows per partition. I recommend going with Option A: partitioning once per month.wrong. New data will be saved in a new partition, not distributed across all previous partitions. so it doesn't matter how much data in previous partitionsData will be partitioned only after it's distributed. The 30 million would be distributed among the 60 nodes and then partitioned in each node inside.

upvoted 1 times

hypersam 1 year, 6 months ago

wrong. New data will be saved in a new partition, not distributed across all previous partitions. so it doesn't matter how much data in previous partitions

upvoted 1 times

hypersam 1 year, 6 months ago

wrong. New data will be saved in a new partition, not distributed across all previous partitions. so it doesn't matter how much data in previous partitionsData will be partitioned only after it's distributed. The 30 million would be distributed among the 60 nodes and then partitioned in each node inside.

upvoted 1 times

prem__raj 1 year, 6 months ago

Data will be partitioned only after it's distributed. The 30 million would be distributed among the 60 nodes and then partitioned in each node inside.

upvoted 1 times

d47320d 2 years, 3 months ago

We have to consider all information given to us about Table1, including existing data, in order to optimize query performance and data loading for the specific table. So, a year after Table1 will contain 1360M i.e. around 22.6M per distribution having specified one partition per year. Instead, if we specify one partition by month, then we will have around 1.8M per distribution and partition. Typical analytical queries involve filter predicates for a specific month. Isn't it more efficient to process 1.8M instead of 22.6M ? (optimize query performance). Loading data to a new empty or smaller partition isn't it faster than loading it to one with pre-existing data or larger partition (optimize data loading)? The above leads us to "A. once per month".

upvoted 1 times

Alongi 2 years, 4 months ago

Selected Answer: B

It's B for that reason: 12month * 30 mln / 60 nodes = 6 mln record per partition

upvoted 1 times

Alongi 2 years, 4 months ago

It's B for that reason: 12month * 30 mln / 60 nodes = 6 mln record per partitionyou are totally ignoring existing table data

upvoted 1 times

d47320d 2 years, 3 months ago

you are totally ignoring existing table data

upvoted 1 times

dgerok 2 years, 4 months ago

B) once per year, because your partition should contain more than 60 mln rows (1 mln for each of 60 nodes)

upvoted 2 times

dgerok 2 years, 4 months ago

B. once per year = 360 milion / 60 = 6 milion record per partition

upvoted 2 times

mav2000 2 years, 5 months ago

A partition is divided into distribution units. A) eleven per month => 30 million / 60 = 500k rows per partition B) eleven per year => 360 million / 60 = 6 million rows per partition C) eleven per day => 1 million / 60 = 15k rows per partition D) eleven per week => 7.5~ million/60 = 125k rows per partition But since I already have 1 billion rows distributed, each distribution node would have 16.7 million rows, fulfilling more than the minimum, since I am interested in having a certain reasonable number of partitions to increase the speed of the queries, I would choose once per month, because once per year would create very few partitions. Therefore the answer is A)you'll have new partitions whenever you update the table's partitions. if you create a new partition every month, this new partition will have only 500k rows per distribution. I agree that the previous partitions have more than enough records, but the problem is with the new partition.

upvoted 4 times

mav2000 2 years, 5 months ago

Selected Answer: A

A partition is divided into distribution units.

A) eleven per month => 30 million / 60 = 500k rows per partition B) eleven per year => 360 million / 60 = 6 million rows per partition C) eleven per day => 1 million / 60 = 15k rows per partition D) eleven per week => 7.5~ million/60 = 125k rows per partition

But since I already have 1 billion rows distributed, each distribution node would have 16.7 million rows, fulfilling more than the minimum, since I am interested in having a certain reasonable number of partitions to increase the speed of the queries, I would choose once per month, because once per year would create very few partitions.

Therefore the answer is A)

upvoted 4 times

SajadAhm 2 years ago

you'll have new partitions whenever you update the table's partitions. if you create a new partition every month, this new partition will have only 500k rows per distribution. I agree that the previous partitions have more than enough records, but the problem is with the new partition.

upvoted 2 times

rocky48 2 years, 6 months ago

Given that you'll be adding 30 million rows per month, consider partitioning once per month based on the Sales Date column. This approach balances data maintenance and query performance while avoiding excessive partitioning. Therefore, the correct answer is A. once per month

upvoted 2 times

rocky48 2 years, 6 months ago

To optimize query performance and data loading for your Azure Synapse Analytics dedicated SQL pool table, you should create a partition based on the Sales Date column. Partitioning Frequency: The decision on how often to create a partition depends on the data volume and your specific use case. Minimum Rows per Partition: For optimal compression and performance of clustered columnstore tables, it's recommended to have a minimum of 1 million rows per distribution and partition. Automatic Distribution: Dedicated SQL pools already divide each table into 60 distributed databases. Recommendation: Given that you'll be adding 30 million rows per month, consider partitioning once per month based on the Sales Date column. This approach balances data maintenance and query performance while avoiding excessive partitioning. Therefore, the correct answer is A. once per month

upvoted 1 times

moneytime 2 years, 6 months ago

I chose D Reason: For optimal performance and compression the followings are true; The minimum record to be load is 60M permonth. This implies that for ; *Per week = 60/4= 15M records *Pe rday = 15/7 =2.1M records *Per year= 12 *60M = 740M and so on for hours ,minutes ,sec. in the right proportions. Considering ,the case at hand. Adding 30M rows monthly will have negative impact on the performance and compression of the table .The reason is that the quantity of records to load is less than the required minimum value of 60M for monthly partition. In other to correct this ,30M minimum rows show be added every 2 weeks(not permonth) or 10M minimum rows per week. So ,the most reasonable advice is to load the table perweek for optimal performance and compression.Your calculation logic is incorrect. A year after Table1 will contain 1360M i.e. around 22.6M per distribution having specified one partition per year. Instead, if we specify one partition by week, then we will have around 0.4M per distribution and partition, which is below than 1M Microsoft recommendation. This makes your choice wrong.

upvoted 1 times

moneytime 2 years, 6 months ago

I chose D Reason: For optimal performance and compression the followings are true; The minimum record to be load is 60M permonth. This implies that for ; *Per week = 60/4= 15M records *Pe rday = 15/7 =2.1M records *Per year= 12 *60M = 740M and so on for hours ,minutes ,sec. in the right proportions. Considering ,the case at hand. Adding 30M rows monthly will have negative impact on the performance and compression of the table .The reason is that the quantity of records to load is less than the required minimum value of 60M for monthly partition.

In other to correct this ,30M minimum rows show be added every 2 weeks(not permonth) or 10M minimum rows per week. So ,the most reasonable advice is to load the table perweek for optimal performance and compression.

upvoted 1 times

d47320d 2 years, 3 months ago

Your calculation logic is incorrect. A year after Table1 will contain 1360M i.e. around 22.6M per distribution having specified one partition per year. Instead, if we specify one partition by week, then we will have around 0.4M per distribution and partition, which is below than 1M Microsoft recommendation. This makes your choice wrong.

upvoted 2 times

Charley92 2 years, 6 months ago

To partition Table1 based on the Sales Date column, you can create a partition for each month. Since thirty million rows will be added to Table1 each month, creating a partition once per month would be the most optimal solution for query performance and data loading 1. This way, the data can be easily queried and loaded, and the partitioning will not be too granular or too coarse 1. Therefore, the answer is A. once per month.

upvoted 1 times

be8a152 2 years, 7 months ago

The table already has 1 billion records (ie 100 Million) + an additional 30 Million per month. So going by the theory that a minimum of 1 million rows should be accommodated by each of the 60 distributions created in the background each distribution will end up with 130 Million / 60 = 2.6 millions. So A. Once per month is the right answer

upvoted 1 times

Question #63 *Topic 1*

You have an Azure Databricks workspace that contains a Delta Lake dimension table named Table1.
Table1 is a Type 2 slowly changing dimension (SCD) table.
You need to apply updates from a source table to Table1.
Which Apache Spark SQL operation should you use?

A. CREATE

B. UPDATE

C. ALTER

D. MERGE

Suggested Answer: *D*
The Delta provides the ability to infer the schema for data input which further reduces the effort required in managing the schema changes. The Slowly Changing Data(SCD) Type 2 records all the changes made to each key in the dimensional table. These operations require updating the existing rows to mark the previous values of the keys as old and then inserting new rows as the latest values. Also, Given a source table with the updates and the target table with dimensional data, SCD Type 2 can be expressed with the merge. Example: // Implementing SCD Type 2 operation using merge function customersTable .as("customers") .merge(stagedUpdates.as("staged_updates"), "customers.customerId = mergeKey") .whenMatched("customers.current = true AND customers.address <> staged_updates.address") .updateExpr(Map("current" -> "false", "endDate" -> "staged_updates.effectiveDate")) .whenNotMatched() .insertExpr(Map("customerid" -> "staged_updates.customerId", "address" -> "staged_updates.address", "current" -> "true", "effectiveDate" -> "staged_updates.effectiveDate", "endDate" -> "null")) .execute() }
Reference: https://www.projectpro.io/recipes/what-is-slowly-changing-data-scd-type-2-operation-delta-table-databricks

Community vote distribution



labriji Highly Voted 3 years, 7 months ago
When applying updates to a Type 2 slowly changing dimension (SCD) table in Azure Databricks, the best option is to use the MERGE operation in Apache Spark SQL. This operation allows you to combine the data from the source table with the data in the destination table, and then update or insert the appropriate r ecords. The MERGE operation provides a powerful and flexible way to handle updates for SCD tables, as it can handle both updates and inserts in a single operation. Additionally, this operation can be performed on Delta Lake tables, which can easily handle the ACID transactions needed for handling SCD updates.
upvoted 13 times

Naman1605 Highly Voted 3 years, 11 months ago
correct D
upvoted 9 times

EmnCours Most Recent 1 year, 8 months ago
Correct Answer: D
upvoted 1 times

dgerok 2 years, 4 months ago
Type 2 SCD means you have a new row to insert for the changes. The only MERGE supports this operation among the possible answers
upvoted 1 times

rocky48 2 years, 6 months ago
MERGE Operation: The MERGE operation in Apache Spark SQL (and specifically in Delta Lake) combines the capabilities of both INSERT and UPDATE. It allows you to upsert data (insert new records or update existing records) based on a specified condition. When you have a slowly changing dimension (SCD) table like Table1, where historical data needs to be maintained. Therefore, the correct answer is D. MERGE
upvoted 3 times

kkk5566 2 years, 11 months ago
To update or upsert records in a delta lake in Databricks use the "Merge" command.
upvoted 4 times

ajhak 3 years, 2 months ago
To update or upsert records in a delta lake in Databricks use the "Merge" command. https://learn.microsoft.com/en-us/azure/databricks/delta/merge
upvoted 4 times

anks84 3 years, 11 months ago

Correct

upvoted 3 times

Question #64 *Topic 1*

You are designing an Azure Data Lake Storage solution that will transform raw JSON files for use in an analytical workload.
You need to recommend a format for the transformed files. The solution must meet the following requirements:
✑ Contain information about the data types of each column in the files.
✑ Support querying a subset of columns in the files.
✑ Support read-heavy analytical workloads.
✑ Minimize the file size.
What should you recommend?

A. JSON

B. CSV

C. Apache Avro

D. Apache Parquet

Suggested Answer: *D*
Parquet, an open-source file format for Hadoop, stores nested data structures in a flat columnar format. Compared to a traditional approach where data is stored in a row-oriented approach, Parquet file format is more efficient in terms of storage and performance. It is especially good for queries that read particular columns from a ɔ€wideɔ€ (with many columns) table since only needed columns are read, and IO is minimized. Incorrect: Not C: The Avro format is the ideal candidate for storing data in a data lake landing zone because: 1. Data from the landing zone is usually read as a whole for further processing by downstream systems (the row-based format is more efficient in this case). 2. Downstream systems can easily retrieve table schemas from Avro files (there is no need to store the schemas separately in an external meta store). 3. Any source schema change is easily handled (schema evolution). Reference: https://www.clairvoyant.ai/blog/big-data-file-formats

Community vote distribution
D (100%)

ajhak Highly Voted 3 years, 3 months ago
Parquet has columnar format, best for reading a few columns. Also when in doubt, just select Parquet. More often than not, that is gonna be the correct answer if you don't know it.
upvoted 13 times

anks84 Highly Voted 3 years, 11 months ago
Correct as the requirement is Column based.
upvoted 7 times

EmnCours Most Recent 1 year, 8 months ago
Correct Answer: D
upvoted 1 times

Piantoni 1 year, 9 months ago
The answer is "D"
upvoted 1 times

kkk5566 2 years, 11 months ago
D is correct
upvoted 2 times

Reloadedvn 3 years, 3 months ago
Parquet format satisfies all reqs
upvoted 2 times

GodfreyMbizo 3 years, 6 months ago
Parquet
upvoted 3 times

youngbug 3 years, 8 months ago
Avro is totally a serialization format. It combines of JSON and Raw binary files. Why is the explanation like this? It's misleading.
upvoted 4 times

Question #65 *Topic 1*

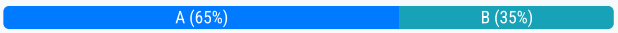
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly.
Solution: You modify the files to ensure that each row is less than 1 MB.
Does this meet the goal?

A. Yes

B. No

Suggested Answer: *A*
Polybase loads rows that are smaller than 1 MB. Note on Polybase Load: PolyBase is a technology that accesses external data stored in Azure Blob storage or Azure Data Lake Store via the T-SQL language. Extract, Load, and Transform (ELT) Extract, Load, and Transform (ELT) is a process by which data is extracted from a source system, loaded into a data warehouse, and then transformed. The basic steps for implementing a PolyBase ELT for dedicated SQL pool are: Extract the source data into text files. Land the data into Azure Blob storage or Azure Data Lake Store. Prepare the data for loading. Load the data into dedicated SQL pool staging tables using PolyBase. Transform the data. Insert the data into production tables. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-service-capacity-limits https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/load-data-overview

Community vote distribution



Tj87 Highly Voted 3 years, 11 months ago
I think we had this question in the previous pages and the correct answer was set as " compress the files"Exactly compress because a lot of row have more than 1MB lengthThe question before was more than that 1 MB but here is less than 1 MB. since, it is less, then answer is Yes.More than 1 solution might be right. The question here is: if row size is reduced to 1MB, will loading go faster? The answer then is yes: whether compression is better or not, is not relevant.
upvoted 34 times

Tj87 Highly Voted 3 years, 11 months ago
I think we had this question in the previous pages and the correct answer was set as " compress the files"
upvoted 34 times

dom271219 3 years, 11 months ago
Exactly compress because a lot of row have more than 1MB length
upvoted 4 times

kim32 3 years, 3 months ago
The question before was more than that 1 MB but here is less than 1 MB. since, it is less, then answer is Yes.
upvoted 5 times

semauni 3 years ago
More than 1 solution might be right. The question here is: if row size is reduced to 1MB, will loading go faster? The answer then is yes: whether compression is better or not, is not relevant.
upvoted 5 times

Phund 3 years, 11 months ago
"ensure that each row is less than 1 MB" and the condition for polybase is <1M, whatever method you used
upvoted 16 times

ecalvo Most Recent 1 year, 5 months ago
correct answer is B, it's the same question 24 topic 1
upvoted 1 times

imatheushenrique 1 year, 5 months ago

Answer: B. No While ensuring that each row is less than 1 MB may help in some cases, it does not fully address the performance challenges of copying data efficiently into Azure Synapse Analytics.

upvoted 1 times

moize 1 year, 7 months ago

Réduire la taille des lignes à moins de 1 Mo ne garantit pas nécessairement une copie plus rapide des données vers Azure Synapse Analytics. Il serait plus efficace d'utiliser des techniques d'optimisation de transfert de données, comme la compression des fichiers ou l'utilisation de formats de fichiers optimisés pour les performances, tels que Parquet ou ORC.

upvoted 1 times

dgerok 2 years, 4 months ago

Selected Answer: B

If you modify the files to ensure that each row is less than 1 MB, you might end up truncating or losing data from those rows. To achieve faster data copying, consider alternative approaches such as: Compression: Compress the files before transferring them to Azure Synapse Analytics. This can reduce the overall size of the data and improve transfer speed. Parallelization: Split the data into smaller chunks and copy them in parallel to take advantage of multiple resources. Optimized Data Types: Ensure that numerical values are stored using appropriate data types (e.g., integers, floats) to minimize storage space. Batch Processing: Process the data in batches rather than row by row to optimize data transfer. The question should be read carefully and attentively. You need to prepare the files to ensure that the data copies QUICKLY. Nobody asks if this improve... I agree with rocky48. You need do more to copy the data QUICKLY. So, the answer is B (NO)

upvoted 4 times

dgerok 2 years, 4 months ago

If you modify the files to ensure that each row is less than 1 MB, you might end up truncating or losing data from those rows. To achieve faster data copying, consider alternative approaches such as: Compression: Compress the files before transferring them to Azure Synapse Analytics. This can reduce the overall size of the data and improve transfer speed. Parallelization: Split the data into smaller chunks and copy them in parallel to take advantage of multiple resources. Optimized Data Types: Ensure that numerical values are stored using appropriate data types (e.g., integers, floats) to minimize storage space. Batch Processing: Process the data in batches rather than row by row to optimize data transfer. The question should be read carefully and attentively. You need to prepare the files to ensure that the data copies QUICKLY. Nobody asks if this improve... I agree with rocky48. You need do more to copy the data QUICKLY. So, the answer is B (NO)I am with this answer. It should never be a good practise modifying the data at loading time.

upvoted 4 times

stornati 2 years, 1 month ago

I am with this answer. It should never be a good practise modifying the data at loading time.

upvoted 2 times

ItsAB 2 years, 6 months ago

Given the large size of the table, I will utilize PolyBase for data transfer. Additionally, considering PolyBase's constraint that it cannot load rows exceeding 1MB in size, I will compress rows to ensure compliance with this requirement, thereby making PolyBase the optimal choice for data transfer. => Option A: yes

upvoted 1 times

stickslinger 2 years, 6 months ago

Shouldn't this be "B: No". Where does the question ask about Polybase? Modifying the rows of data could affect the integrity of the data

upvoted 1 times

rocky48 2 years, 6 months ago

No, this solution does not meet the goal. While modifying the files to ensure that each row is less than 1 MB might help with individual row sizes, it won't necessarily improve the overall data transfer speed. The total size of the files in the storage account is still 100 GB, and copying large volumes of data can be time-consuming regardless of individual row sizes. To optimize data transfer speed, consider other strategies such as parallelizing the data transfer, optimizing network bandwidth, or using appropriate data loading techniques in Azure Synapse Analytics.

upvoted 3 times

Charley92 2 years, 6 months ago

No, this solution does not meet the goal. The files contain rows of text and numerical values, and 75% of the rows contain description data that has an average length of 1.1 MB. If you modify the files to ensure that each row is less than 1 MB, you may end up splitting the description data into multiple rows, which could affect the integrity of the data

upvoted 3 times

ExamDestroyer69 2 years, 7 months ago

Variations Solution: You convert the files to compressed delimited text files. Does this meet the goal? **YES** Solution: You copy the files to a table that has a columnstore index. Does this meet the goal? **NO** Solution: You modify the files to ensure that each row is more than 1 MB. Does this meet the goal? **NO** Solution: You modify the files to ensure that each row is less than 1 MB. Does this meet the goal? **YES**

upvoted 14 times

hcq31818 2 years, 9 months ago

PolyBase enables Azure Synapse Analytics to import and export data from Azure Data Lake Store, and from Azure Blob Storage. And it supports row sizes up to 1MB. https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-guide?view=sql-server-ver16#:~:text=Azure%20integration,and%20from%20Azure%20Blob%20Storage. https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-versioned-feature-summary?view=sql-server-ver16

upvoted 1 times

kkk5566 2 years, 11 months ago

A is correct

upvoted 1 times

auwia 3 years, 1 month ago

Yes, with less of 1Mb file we increase performance.

upvoted 2 times

e5019c6 3 years, 1 month ago

i thought that polybase just query the tables and dont do any process of ETL or ELT.

upvoted 2 times

rocky48 3 years, 3 months ago

"You modify the files to ensure that each row is more than 1 MB" and the answer was "No". This particular question asks if "You modify the files to ensure that each row is less than 1 MB", and the answer given is "Yes".

upvoted 4 times

Question #66 *Topic 1*

You plan to create a dimension table in Azure Synapse Analytics that will be less than 1 GB.
You need to create the table to meet the following requirements:
➪ Provide the fastest query time.
➪ Minimize data movement during queries.
Which type of table should you use?

A. replicated

B. hash distributed

C. heap

D. round-robin

Suggested Answer: *A*
A replicated table has a full copy of the table accessible on each Compute node. Replicating a table removes the need to transfer data among Compute nodes before a join or aggregation. Since the table has multiple copies, replicated tables work best when the table size is less than 2 GB compressed. 2 GB is not a hard limit. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-guidance-for-replicated-tables

Community vote distribution
A (100%)

anks84 Highly Voted 3 years, 11 months ago
Given answer is correct ! Replicated because - Dimension table - Less than 2 GB (less than 1 GB in this case)
upvoted 11 times

allagowf Highly Voted 3 years, 10 months ago
correct answer
upvoted 5 times

samianae Most Recent 1 year, 6 months ago
answer correct
upvoted 1 times

EmnCours 1 year, 8 months ago
Correct Answer: A
upvoted 1 times

17lan 1 year, 8 months ago
correct
upvoted 1 times

Thameur01 1 year, 10 months ago
The question is: Which type of table should you use? Actually, it should be Which type of distribution should you use? Table types = Clustered Columnstore, heap, clustered index Distribution types = hash, round robin and replicated
upvoted 1 times

rocky48 2 years, 6 months ago
Replicated Table: Caches a full copy of the table on each compute node. While it eliminates data movement during joins or aggregations, it requires extra storage and is not suitable for large tables.
upvoted 2 times

hcq31818 2 years, 9 months ago
Replicated
upvoted 2 times

kkk5566 2 years, 11 months ago
A is correct
upvoted 2 times

Deeksha1234 3 years ago
A is correct

upvoted 2 times

akhil5432 3 years ago

replicated

upvoted 1 times

mamahani 3 years, 3 months ago

A is correct answer

upvoted 3 times

GodfreyMbizo 3 years, 6 months ago

answer is correct for Dimension tables and less than 2 GB microsoft recommends replicated tables

upvoted 3 times

vigilante89 3 years, 8 months ago

Since the dim table is under 1 GB size which is quite small, it should be replicated across all the partitions so that data movement is less and efficiency is more.

upvoted 3 times

MEIRONGD 3 years, 8 months ago

correct answer

upvoted 3 times

Question #67 *Topic 1*

You are designing a dimension table in an Azure Synapse Analytics dedicated SQL pool.
You need to create a surrogate key for the table. The solution must provide the fastest query performance.
What should you use for the surrogate key?

A. a GUID column

B. a sequence object

C. an IDENTITY column

Suggested Answer: *C*
Use IDENTITY to create surrogate keys using dedicated SQL pool in AzureSynapse Analytics. Note: A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance. Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identity

Community vote distribution
C (96%) (4

anks84 Highly Voted 3 years, 11 months ago
Given answer is correct
upvoted 10 times

nicky87654 Highly Voted 3 years, 7 months ago
C. an IDENTITY column When designing a dimension table in a data warehouse, it's important to consider the types of queries that will be run against it. IDENTITY columns are generally the best option for surrogate keys in dimension tables because they provide the fastest query performance. IDENTITY columns are auto-incremented and indexed by default, which makes them ideal for use as primary keys. They also require less storage space than GUID columns and are less likely to cause fragmentation in indexes.
upvoted 6 times

EmnCours Most Recent 1 year, 8 months ago
Correct Answer: C
upvoted 1 times

rocky48 2 years, 6 months ago
Remember that the IDENTITY value doesn't guarantee order due to the distributed architecture, but it ensures uniqueness within each distribution.The IDENTITY property generates unique values automatically. It scales out across all distributions in the dedicated SQL pool. It doesn't impact load performance. Each distribution maintains its own set of IDENTITY values. Ideal for creating surrogate keys without manual intervention.
upvoted 2 times

kkk5566 2 years, 11 months ago
C is the answer.
upvoted 1 times

Deeksha1234 3 years ago
C is correct
upvoted 3 times

akhil5432 3 years ago
Identity column
upvoted 3 times

vigilante89 3 years, 8 months ago
Surrogate key is substitute key used if there is no natural or business key within the table. It should be always instantiated by IDENTITY(1,1).
upvoted 2 times

Question #68 *Topic 1*

HOTSPOT

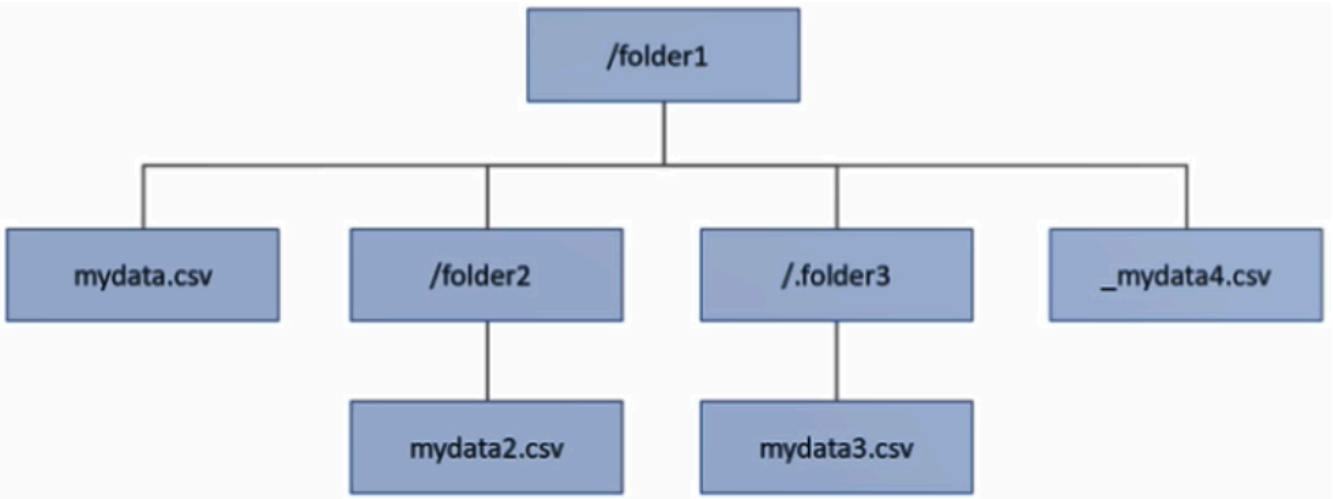
-

You have an Azure Data Lake Storage Gen2 account that contains a container named container1. You have an Azure Synapse Analytics serverless SQL pool that contains a native external table named dbo.Table1. The source data for dbo.Table1 is stored in container1. The folder structure of container1 is shown in the following exhibit.

The external data source is defined by using the following statement.

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.




```
CREATE EXTERNAL DATA SOURCE DataLake
WITH
(    LOCATION         = 'https://mydatalake.dfs.core.windows.net/container1/folder1/**'
     , CREDENTIAL = DataLakeCred
);
```

Answer Area

Statements	Yes	No
When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.	○	○
When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	○	○
When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	○	○

Suggested Answer:

Answer Area

Statements	Yes	No
When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.	◉	○
When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	◉	○
When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	○	◉

Answer Area

Statements	Yes	No
When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.	◉	○
When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	◉	○
When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	○	◉

shoottheduck Highly Voted 3 years, 6 months ago

I have just tested this on Synapse Serverless: ./Folder3 AND _mydata4.csv were ignored. Therefor; Yes, No, No

upvoted 97 times

shoottheduck Highly Voted 3 years, 6 months ago

I have just tested this on Synapse Serverless: ./Folder3 AND _mydata4.csv were ignored. Therefor; Yes, No, Noits not ./Folder3 . it is /.Folder3 still ignored?Folder or file that starts with . or _ Reference documentation: https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1Thanks for the linkIts same for AD credz of team members / brass that get das-boot at work. Recent times seeing a whole lotta " _ " lol. Trying to exit hence here before catching " _ "thanks, and did you try it with a native external table ?I think it should be: 1 - NO: It would be needed to specify the "/**" option to traverse folders: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-folders-multiple-csv-files 2 - NO: since it's a hidden folder: https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=sql-server-ver16&tabs=dedicated Not sure if this documentation is well suited... 3 - NO: It starts with "_": Same link as question 2For sure the corrected answers are: - YES - NO - NO EXPLAIN: Recursive Wildcard (/**): The wildcard at the end (/**) means all files in /folder1 and its subfolders will be included. Hidden Files and Folders: Files or folders starting with . (e.g., .folder3) or _ (e.g., _mydata4.csv) are ignored by Synapse serverless SQL pool. Files Included: 1. /folder1/mydata.csv → Included 2. /folder1/folder2/mydata2.csv → Included 3. /folder1/.folder3/mydata3.csv → Ignored (starts with a dot .). 4. /folder1/_mydata4.csv → Ignored (starts with an underscore _). Final Answer: The query will return mydata.csv and mydata2.csv only.

upvoted 97 times

MuruAzure 3 years, 4 months ago

its not ./Folder3 . it is /.Folder3 still ignored?Folder or file that starts with . or _ Reference documentation: https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1Thanks for the linkIts same for AD credz of team members / brass that get das-boot at work. Recent times seeing a whole lotta " _ " lol. Trying to exit hence here before catching " _ "

upvoted 5 times

MuruAzure 3 years, 4 months ago

its not ./Folder3 . it is /.Folder3 still ignored?

upvoted 5 times

VikkiC 3 years, 2 months ago

Folder or file that starts with . or _ Reference documentation: https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1Thanks for the linkIts same for AD credz of team members / brass that get das-boot at work. Recent times seeing a whole lotta " _ " lol. Trying to exit hence here before catching " _ "

upvoted 26 times

VikkiC 3 years, 2 months ago

Folder or file that starts with . or _ Reference documentation: https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1

upvoted 26 times

AvSUN 2 years, 11 months ago

Thanks for the link

upvoted 1 times

darkraijin 2 years, 10 months ago

Its same for AD credz of team members / brass that get das-boot at work. Recent times seeing a whole lotta " _ " lol. Trying to exit hence here before catching " _ "

upvoted 1 times

RMK2000 Most Recent 1 year, 7 months ago

thanks, and did you try it with a native external table ?

upvoted 1 times

tlb_20 2 years, 4 months ago

I think it should be: 1 - NO: It would be needed to specify the "/**" option to traverse folders: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-folders-multiple-csv-files 2 - NO: since it's a hidden folder: https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=sql-server-ver16&tabs=dedicated Not sure if this documentation is well suited... 3 - NO: It starts with "_": Same link as question 2

upvoted 2 times

HUSSAIN_ALYAFEI 1 year, 7 months ago

For sure the corrected answers are: - YES - NO - NO EXPLAIN: Recursive Wildcard (/**): The wildcard at the end (/**) means all files in /folder1 and its subfolders will be included. Hidden Files and Folders: Files or folders starting with . (e.g., .folder3) or _ (e.g., _mydata4.csv) are ignored by Synapse serverless SQL pool. Files Included: 1. /folder1/mydata.csv → Included 2. /folder1/folder2/mydata2.csv → Included 3.

/folder1/.folder3/mydata3.csv → Ignored (starts with a dot .). 4. /folder1/_mydata4.csv → Ignored (starts with an underscore _). Final Answer: The query will return mydata.csv and mydata2.csv only.

upvoted 9 times

PGiagkoulas 3 years, 7 months ago

1.Yes, 2.Yes: "Unlike Hadoop external tables, native external tables don't return subfolders unless you specify /** at the end of path" which is the case here. 3. No: "Both Hadoop and native external tables will skip the files with the names that begin with an underline (_) or a period (.)", refers to files, not directories, so the last file with the underscore will be exluded. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#arguments-create-external-tableένα like για το nickname ;-)@PGiagkoulas, read this link again : https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1 See this part in the link: native external tables don't return subfolders unless you specify /** at the end of path. In this example, if LOCATION='/webdata/', a serverless SQL pool query, will return rows from mydata.txt. It won't return mydata2.txt and mydata3.txt because they're located in a subfolder. Based on this, the answer is Yes, No and NoBut /** is specified at the end?yes thats why it will return the all folders

upvoted 47 times

PGiagkoulas 3 years, 7 months ago

1.Yes, 2.Yes: "Unlike Hadoop external tables, native external tables don't return subfolders unless you specify /** at the end of path" which is the case here. 3. No: "Both Hadoop and native external tables will skip the files with the names that begin with an underline (_) or a period (.)", refers to files, not directories, so the last file with the underscore will be exluded. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#arguments-create-external-table

upvoted 47 times

matiandal 3 years, 2 months ago

ένα like για το nickname ;-)

upvoted 3 times

Paulkuzzio 3 years, 1 month ago

@PGiagkoulas, read this link again : https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1 See this part in the link: native external tables don't return subfolders unless you specify /** at the end of path. In this example, if LOCATION='/webdata/', a serverless SQL pool query, will return rows from mydata.txt. It won't return mydata2.txt and mydata3.txt because they're located in a subfolder. Based on this, the answer is Yes, No and NoBut /** is specified at the end?yes thats why it will return the all folders

upvoted 7 times

Paulkuzzio 3 years, 1 month ago

@PGiagkoulas, read this link again : https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1

See this part in the link:

native external tables don't return subfolders unless you specify /** at the end of path. In this example, if LOCATION='/webdata/', a serverless SQL pool query, will return rows from mydata.txt. It won't return mydata2.txt and mydata3.txt because they're located in a subfolder.

Based on this, the answer is Yes, No and No

upvoted 7 times

semauni 3 years ago

But /** is specified at the end?

upvoted 4 times

semauni 3 years ago

But /** is specified at the end?yes thats why it will return the all folders

upvoted 4 times

dakku987 2 years, 7 months ago

yes thats why it will return the all folders

upvoted 1 times

8ac3742 1 year, 8 months ago

"Unlike Hadoop and PolyBase, serverless SQL pool doesn't return subfolders unless you specify /** at the end of path. Just like Hadoop and PolyBase, it doesn't return files for which the file name begins with an underline (_) or a period (.)." The answer is Yes, Yes, No

upvoted 1 times

mavdplas 1 year, 9 months ago

Yes, No, No If you specify LOCATION to be a folder, a PolyBase query that selects from the external table will retrieve files from the folder and all of its subfolders. Just like Hadoop, PolyBase doesn't return hidden folders. It also doesn't return files for which the file name begins with an underline (_) or a period (.). https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1

upvoted 1 times

jarobdataeng 1 year, 10 months ago

yes-no-no -> https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql

upvoted 1 times

uncledana 1 year, 10 months ago

Selected Answer is correct: 1) Yes 2) Yes 3) No Link: https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=sql-server-ver16&tabs=dedicated Link

upvoted 1 times

34c89bf 2 years ago

All should be No, Since we cannot provide ** in data source , it should be there in external table.

upvoted 1 times

dgerok 2 years, 4 months ago

you can use a wildcard character (*) in the path to reference multiple files within a directory. However, the wildcard character should not be used consecutively. You will get an error message trying to select data from the external table: Consecutive wildcard characters present in path '/container1/*'. So, I suppose, all the answers should be NO

upvoted 1 times

Nitin9in 2 years, 5 months ago

YES,NO,NO The "." and "-" are hidden. The native external tables and Hadoop external tables won't return anything from hidden folders. https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&preserve-view=true&tabs=dedicated

upvoted 7 times

rocky48 2 years, 6 months ago

Yes, No, No. /.Folder 3 2ill be ignored because it is a hidden folder and _mydata4.csv will also be ignored because it is a hidden file

upvoted 1 times

be8a152 2 years, 6 months ago

Yes, No, No. /.Folder 3 2ill be ignored because it is a hidden folder and _mydata4.csv will also be ignored because it is a hidden file

upvoted 2 times

Joanna0 2 years, 7 months ago

Y - Visible / N - Hidden Folder / N - Hidden File https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-guidance-for-replicated-tables

upvoted 2 times

blazy001 2 years, 8 months ago

yes, no, no >> Both Hadoop and native external tables will skip the files with the names that begin with an underline (_) or a period (.).

upvoted 2 times

Lscranio 2 years, 8 months ago

Yes / No / No https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1

upvoted 2 times

ellala 2 years, 10 months ago

In my opinion YES NO NO from this doc: https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1

upvoted 5 times

Chemmangat 2 years, 11 months ago

Yes, No, No Anything that starts with '_' or '.' will be ignored. And if you are having a doubt that files are in the form of /.{name_of_folder}, Note that these are folders and '/' is used for representing the folders. Pretty basic thing, but hope it helps someone.

upvoted 4 times

hassexat 2 years, 11 months ago

YES / NO / NO

upvoted 1 times

Question #69 *Topic 1*

You have an Azure Synapse Analytics dedicated SQL pool.
You need to create a fact table named Table1 that will store sales data from the last three years. The solution must be optimized for the following query operations:
• Show order counts by week.
• Calculate sales totals by region.
• Calculate sales totals by product.
• Find all the orders from a given month.
Which data should you use to partition Table1?

A. product

B. month

C. week

D. region

Suggested Answer: *B*
Official Answer: B

Community vote distribution
B (100%)

AlviraTony Highly Voted 2 years, 11 months ago
• Show order counts by week. • Calculate sales totals by region. • Calculate sales totals by product. For these, Group By is required while querying, hence cannot be a parition. But fourth one, requires you to use WHERE clause, so month is ideal for a partition
upvoted 44 times

AlviraTony Highly Voted 2 years, 11 months ago
• Show order counts by week. • Calculate sales totals by region. • Calculate sales totals by product. For these, Group By is required while querying, hence cannot be a parition. But fourth one, requires you to use WHERE clause, so month is ideal for a partitionThanks, this was helpful
upvoted 44 times

ellala 2 years, 10 months ago
Thanks, this was helpful
upvoted 7 times

nicky87654 3 years, 7 months ago
When designing a fact table in a data warehouse, it is important to consider the types of queries that will be run against it. In this case, the queries that need to be optimized include: show order counts by week, calculate sales totals by region, calculate sales totals by product, and find all the orders from a given month. Partitioning the table by month would be the best option in this scenario as it would allow for efficient querying of data by month, which is necessary for the query operations described above. For example, it would be easy to find all the orders from a given month by only searching the partition for that specific month.
upvoted 18 times

EmnCours Most Recent 1 year, 8 months ago
Correct Answer: B
upvoted 1 times

renan_ineu 1 year, 11 months ago
I choose B, partitioning by month. This will make finding all orders for a specific month much faster. It also makes it easier to count orders by week. The other queries, like calculating sales totals by region and product, are simpler and don't need special partitioning.
upvoted 1 times

Nitin9in 2 years, 5 months ago
Partition by week. Partition by month sounds correct, but it is not. If partition by week, then get the four weeks data from 4 partitions. If it is a month, retrieve the month partition and get the data for a week. The first one is simpler than the second one.
upvoted 2 times

ELJORDAN23 2 years, 7 months ago
Got this question on my exam on january 17, answer B is correct

upvoted 3 times

kkk5566 2 years, 11 months ago

Should be B month.

upvoted 1 times

Deeksha1234 3 years ago

B is right

upvoted 1 times

YikesYikes2023 3 years, 1 month ago

Can someone please link documentation to where this is stated? I feel like any of these categories would be an effective partition strategy as there is a use case for each. I am confusedSame. Every question refers to a different category, so why is month dominant over the others?

upvoted 4 times

YikesYikes2023 3 years, 1 month ago

Can someone please link documentation to where this is stated? I feel like any of these categories would be an effective partition strategy as there is a use case for each. I am confused

upvoted 4 times

semauni 3 years ago

Same. Every question refers to a different category, so why is month dominant over the others?

upvoted 5 times

Rossana 3 years, 4 months ago

Chat GPT: Based on the given usage patterns and requirements, the recommended folder structure would be option B: \DataSource\SubjectArea\YYYY-WW\FileData_YYYY_MM_DD.parquet This structure allows for easy filtering of data by year and week, which aligns with the identified usage pattern of most queries filtering by the current year or week. It also organizes the data by data source and subject area, which simplifies folder security. By using a flat structure, with the data files directly under the year-week folder, query times can be minimized as the data is organized for efficient partition pruning. Option A is similar but includes an additional level of hierarchy for the year, which is unnecessary given the requirement to filter by year-week. Options C, D, and E do not follow a consistent hierarchy, making it difficult to navigate and locate specific data files.

upvoted 2 times

rohit 3 years, 4 months ago

How can we partition by unique months unless we have the year too?If the Month column is the format YYYYMM, this won't be a problem.

upvoted 2 times

rohit 3 years, 4 months ago

How can we partition by unique months unless we have the year too?

upvoted 2 times

hydmt07 2 years, 8 months ago

If the Month column is the format YYYYMM, this won't be a problem.

upvoted 1 times

groozyn 3 years, 6 months ago

Why B) month and not C) week?because it's doing aggregation (like the others answer A and D), instead partitions are powerful for where clause query

upvoted 3 times

groozyn 3 years, 6 months ago

Why B) month and not C) week?

upvoted 3 times

vrodriguesp 3 years, 6 months ago

because it's doing aggregation (like the others answer A and D), instead partitions are powerful for where clause query

upvoted 4 times

DindaS 3 years, 6 months ago

B should be the correct answer here

upvoted 1 times

ZIMARAKI 3 years, 7 months ago

Correct

upvoted 1 times

MrWood47 3 years, 7 months ago

Answer is correct

upvoted 1 times

Question #70 *Topic 1*

You are designing the folder structure for an Azure Data Lake Storage Gen2 account.
You identify the following usage patterns:
• Users will query data by using Azure Synapse Analytics serverless SQL pools and Azure Synapse Analytics serverless Apache Spark pools.
• Most queries will include a filter on the current year or week.
• Data will be secured by data source.
You need to recommend a folder structure that meets the following requirements:
• Supports the usage patterns
• Simplifies folder security
• Minimizes query times
Which folder structure should you recommend?

A. \DataSource\SubjectArea\YYYY\WW\FileData_YYYY_MM_DD.parquet

B. \DataSource\SubjectArea\YYYY-WW\FileData_YYYY_MM_DD.parquet

C. DataSource\SubjectArea\WW\YYYY\FileData_YYYY_MM_DD.parquet

D. \YYYY\WW\DataSource\SubjectArea\FileData_YYYY_MM_DD.parquet

E. WW\YYYY\SubjectArea\DataSource\FileData_YYYY_MM_DD.parquet

Suggested Answer: *A*
Official Answer: A

Community vote distribution
A (89%) B (11%)

nicky87654 Highly Voted 3 years, 7 months ago
A. \DataSource\SubjectArea\YYYY\WW\FileData_YYYY_MM_DD.parquet The recommended folder structure that best meets the requirements is option A. It separates data by data source, year and week. It allows for easy filtering of data by year or week, which aligns with the usage pattern where most queries include a filter on the current year or week.
upvoted 16 times

EmnCours Most Recent 1 year, 8 months ago
Correct Answer: A
upvoted 1 times

Charley92 2 years, 6 months ago
Based on the usage patterns and requirements, I recommend the following folder structure:
\DataSource\SubjectArea\YYYY\WW\FileData_YYYY_MM_DD.parquet This structure is designed to support the usage patterns by organizing data by data source and subject area. The year and week folders will help optimize queries that filter by current year or week. This structure also simplifies folder security by grouping data by data source and subject area, making it easier to apply security policies. Finally, this structure minimizes query times by organizing data in a way that is optimized for the most common queries.
upvoted 2 times

ExamDestroyer69 2 years, 8 months ago
I believe it would be B. "Minimises query time" and "Most queries will include a filter on the current year OR week." The "OR WEEK" suggests that we may filter by only week and not the year. In this event it would (to my knowledge) take longer to query through every single year to select the week you want as opposed to selecting all the folders containing the WW target value in their name. If someone with query optimisation knowledge could confirm this below it would be appreciated.
upvoted 1 times

kkk5566 2 years, 11 months ago
Selected Answer: B
Minimizes query times && Most queries will include a filter on the current year OR week. It is B.
upvoted 1 times

kkk5566 2 years, 11 months ago
Minimizes query times && Most queries will include a filter on the current year OR week. It is B.forhot it ,Option A can set permissions at the year level or at the week, should be A.

upvoted 1 times

kkk5566 2 years, 11 months ago

forhot it ,Option A can set permissions at the year level or at the week, should be A.

upvoted 2 times

Deeksha1234 3 years ago

my opinion it should be A, as it can clearly filter on year or on week , both the options will be available.

upvoted 2 times

auwia 3 years, 1 month ago

We need to query by week too, so better YYYY-WWW.I got another good point in favour of B from the question, look the requirement: "Most queries will include a filter on the current year OR week." (OR) let's suppose they ask you to give back the fourth week for example, you need to go year by year (folders) instead of have it in 1 page. Definitely for me it's option B.But why is it YYY-WW for you then instead of YYYY\WW?

upvoted 1 times

auwia 3 years, 1 month ago

Selected Answer: B

We need to query by week too, so better YYYY-WWW.

upvoted 1 times

auwia 3 years, 1 month ago

I got another good point in favour of B from the question, look the requirement: "Most queries will include a filter on the current year OR week." (OR) let's suppose they ask you to give back the fourth week for example, you need to go year by year (folders) instead of have it in 1 page. Definitely for me it's option B.But why is it YYY-WW for you then instead of YYYY\WW?

upvoted 1 times

auwia 3 years, 1 month ago

I got another good point in favour of B from the question, look the requirement: "Most queries will include a filter on the current year OR week." (OR) let's suppose they ask you to give back the fourth week for example, you need to go year by year (folders) instead of have it in 1 page. Definitely for me it's option B.

upvoted 1 times

semauni 3 years ago

But why is it YYY-WW for you then instead of YYYY\WW?

upvoted 2 times

VittalManikonda 3 years, 2 months ago

Option B seems right answer as we can directly access the given yyyy-ww .

upvoted 2 times

Rossana 3 years, 4 months ago

chat GPT: Based on the given usage patterns and requirements, the recommended folder structure would be option B: \DataSource\SubjectArea\YYYY-WW\FileData_YYYY_MM_DD.parquet This structure allows for easy filtering of data by year and week, which aligns with the identified usage pattern of most queries filtering by the current year or week. It also organizes the data by data source and subject area, which simplifies folder security. By using a flat structure, with the data files directly under the year-week folder, query times can be minimized as the data is organized for efficient partition pruning. Option A is similar but includes an additional level of hierarchy for the year, which is unnecessary given the requirement to filter by year-week. Options C, D, and E do not follow a consistent hierarchy, making it difficult to navigate and locate specific data files.No, A it will give you additional separate column for week (WW)."Most queries will include a filter on the current year OR week."I would not trust using Chat GPT for studying a certification...

upvoted 4 times

Rossana 3 years, 4 months ago

chat GPT: Based on the given usage patterns and requirements, the recommended folder structure would be option B:

\DataSource\SubjectArea\YYYY-WW\FileData_YYYY_MM_DD.parquet

This structure allows for easy filtering of data by year and week, which aligns with the identified usage pattern of most queries filtering by the current year or week. It also organizes the data by data source and subject area, which simplifies folder security. By using a flat structure, with the data files directly under the year-week folder, query times can be minimized as the data is organized for efficient partition pruning.

Option A is similar but includes an additional level of hierarchy for the year, which is unnecessary given the requirement to filter by year-week. Options C, D, and E do not follow a consistent hierarchy, making it difficult to navigate and locate specific data files.

upvoted 4 times

Rob77 3 years, 2 months ago

No, A it will give you additional separate column for week (WW).

upvoted 1 times

Rob77 3 years, 2 months ago

No, A it will give you additional separate column for week (WW)."Most queries will include a filter on the current year OR week."

upvoted 1 times

auwia 3 years, 1 month ago

"Most queries will include a filter on the current year OR week."

upvoted 1 times

DP203Cert2023 Highly Voted 3 years, 2 months ago

I would not trust using Chat GPT for studying a certification...

upvoted 5 times

akshaynag95 3 years, 6 months ago

A is correct answer

upvoted 1 times

vrodriguesp 3 years, 6 months ago

A is correct

upvoted 1 times

DindaS 3 years, 6 months ago

The answer A is correct

upvoted 1 times

ZIMARAKI 3 years, 7 months ago

Correct

upvoted 1 times

MrWood47 3 years, 7 months ago

Answer is correct. The reason is that this folder structure allows for the data to be organized by data source and subject area, which can help with securing the data by data source. Additionally, it organizes the data by year and week, which can minimize query times for the queries that include a filter on the current year or week. And also the file name format is consistent with the folder structure, which makes it easy to understand where the data comes from.

upvoted 2 times

alexnicolita 3 years, 7 months ago

My choice is A

upvoted 1 times

Question #71 *Topic 1*

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a table named table1.
You load 5 TB of data into table1.
You need to ensure that columnstore compression is maximized for table1.
Which statement should you execute?

A. DBCC INDEXDEFRAG (pool1, table1)

B. DBCC DBREINDEX (table1)

C. ALTER INDEX ALL on table1 REORGANIZE

D. ALTER INDEX ALL on table1 REBUILD

Suggested Answer: *D*
Official Answer: D

Community vote distribution
D (84%) C (16%)

MrWood47 Highly Voted 3 years, 7 months ago
D. ALTER INDEX ALL on table1 REBUILD This statement will rebuild all indexes on table1, which can help to maximize columnstore compression. The other options are not appropriate for this task. DBCC INDEXDEFRAG (pool1, table1) is for defragmenting the indexes and DBCC DBREINDEX (table1) is for recreating the indexes. ALTER INDEX ALL on table1 REORGANIZE is for reorganizing the indexes.
upvoted 30 times

aemilka 3 years, 4 months ago
Reorganizing an index is less resource intensive than rebuilding an index. For that reason it should be your preferred index maintenance method, unless there is a specific reason to use index rebuild. https://learn.microsoft.com/en-us/sql/relational-databases/indexes/reorganize-and-rebuild-indexes?view=sql-server-ver16https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-index // For a table with an ordered clustered columnstore index, ALTER INDEX REORGANIZE does not re-sort the data. To re-sort data, use ALTER INDEX REBUILD // The question state "columnstored" table, so DAs far as I can see, your quoted article does not refer to Azure Synapse Analytics dedicated SQL pool. I think rebuild is the only supported option for dedicated SQL as can be found here: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-indexYes, I agree with you, I haven't noticed that the article does not apply to Synapse Analytics. D seems to be only possible answer.
upvoted 7 times

aemilka 3 years, 4 months ago
Selected Answer: C
Reorganizing an index is less resource intensive than rebuilding an index. For that reason it should be your preferred index maintenance method, unless there is a specific reason to use index rebuild. https://learn.microsoft.com/en-us/sql/relational-databases/indexes/reorganize-and-rebuild-indexes?view=sql-server-ver16
upvoted 7 times

Homer23 2 years, 5 months ago
https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-index // For a table with an ordered clustered columnstore index, ALTER INDEX REORGANIZE does not re-sort the data. To re-sort data, use ALTER INDEX REBUILD // The question state "columnstored" table, so D
upvoted 1 times

[Removed] Highly Voted 3 years, 3 months ago
As far as I can see, your quoted article does not refer to Azure Synapse Analytics dedicated SQL pool. I think rebuild is the only supported option for dedicated SQL as can be found here: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-index
upvoted 8 times

[Removed] 3 years, 3 months ago
As far as I can see, your quoted article does not refer to Azure Synapse Analytics dedicated SQL pool. I think rebuild is the only supported option for dedicated SQL as can be found here: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-indexYes, I agree with you, I haven't noticed that the article does not apply to Synapse Analytics. D seems to be only possible answer.

upvoted 8 times

aemilka 3 years, 3 months ago

Yes, I agree with you, I haven't noticed that the article does not apply to Synapse Analystics. D seems to be only possible answer.

upvoted 6 times

JustImperius Most Recent 1 year, 6 months ago

Definitely D. REORGANIZE is for light maintenance, to compact small rowgroups and clean up delta stores without a full rebuild. It's faster and less resource-intensive. REBUILD when the table has significant fragmentation or many delta stores. It recreates the entire index, providing MAXIMUM compression and performance improvement. Use for large-scale maintenance.

upvoted 1 times

EmnCours 1 year, 8 months ago

Correct Answer: D

upvoted 1 times

Okea 1 year, 9 months ago

Answer is C https://learn.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-query-performance?view=sql-server-ver16 If the number of deleted rows for several rowgroups is large enough to be merged into fewer rowgroups, reorganizing the columnstore increases the quality of the index and query performance improves. If your data deletion process usually empties entire rowgroups, consider using table partitioning

upvoted 1 times

renan_ineu 1 year, 11 months ago

I saw all the documentations here and found another one that makes pretty clear the correct option is option D: "By default, tables are defined as a clustered columnstore index. After a load completes, some of the data rows might not be compressed into the columnstore. There's a variety of reasons why this can happen. To learn more, see manage columnstore indexes. To optimize query performance and columnstore compression after a load, rebuild the table to force the columnstore index to compress all the rows." https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-load-from-azure-data-lake-store#optimize-columnstore-compression

upvoted 2 times

Yoleri 1 year, 12 months ago

The answer is D. Why not C because ALTER INDEX ALL on table1 REORGANIZE is used for defragmenting the index, but it doesn't achieve the same level of compression as a full rebuild.

upvoted 1 times

Charley92 2 years, 6 months ago

To maximize columnstore compression for table1 in Azure Synapse Analytics dedicated SQL pool, you should execute the following statement: ALTER INDEX ALL on table1 REBUILD The REBUILD option is used to rebuild all indexes on the table, which will maximize columnstore compression for table1. This option is recommended when a large amount of data has been added to the table, as in this case where 5 TB of data has been loaded into table1

upvoted 2 times

ukivanlamlpi 2 years, 9 months ago

what is the different between "Alter Index Rebuild" or "DBCC DBREINDEX"

upvoted 2 times

kkk5566 2 years, 11 months ago

D. ALTER INDEX ALL on table1 REBUILD

upvoted 1 times

Deeksha1234 3 years ago

D is correct

upvoted 1 times

vctrhugo 3 years, 1 month ago

Selected Answer: D

ALTER INDEX REORGANIZE is used for rebuilding or reorganizing indexes, but it does not maximize columnstore compression.

upvoted 1 times

vctrhugo 3 years, 1 month ago

ALTER INDEX REORGANIZE is used for rebuilding or reorganizing indexes, but it does not maximize columnstore compression.I agree. A rebuild can compress the data more efficiently within each combination of distribution and partition: It can open such existing columnstore segments and shuffle data within them (and the deltastore) to maximize compression for the resulting compressed columnstore segments. That is not possible when reorganizing. That process only changes compressed columnstore segments by physically deleting logically deleted rows and combining small columnstore segments into larger ones.

upvoted 1 times

Matt2000 3 years ago

I agree. A rebuild can compress the data more efficiently within each combination of distribution and partition: It can open such existing columnstore segments and shuffle data within them (and the deltastore) to maximize compression for the resulting compressed columnstore segments. That is not possible when reorganizing. That process only changes compressed columnstore segments by physically deleting logically deleted rows and combining small columnstore segments into larger ones.

upvoted 1 times

Rajan191083 3 years, 3 months ago

Reorganize is for row store indexes. The question here clearly mentions column store indexes. Correct answer is D

upvoted 2 times

MuhilMahil 3 years, 3 months ago

Selected Answer is C. reorganizing only help in optimizing compression and performance.

upvoted 1 times

Vedjha 3 years, 6 months ago

Why not C? When reorganizing a columnstore index, the Database Engine compresses each closed row group in delta store into columnstore as a compressed row group. Starting with SQL Server 2016 (13.x) and in Azure SQL Database, the REORGANIZE command performs the following additional defragmentation optimizations online: Physically removes rows from a row group when 10% or more of the rows have been logically deleted. For example, if a compressed row group of 1 million rows has 100,000 rows deleted, the Database Engine will remove the deleted rows and recompress the row group with 900,000 rows, reducing storage footprint.

upvoted 1 times

Question #72 *Topic 1*

You have an Azure Synapse Analytics dedicated SQL pool named pool1.

You plan to implement a star schema in pool and create a new table named DimCustomer by using the following code.

You need to ensure that DimCustomer has the necessary columns to support a Type 2 slowly changing dimension (SCD).

Which two columns should you add? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

```
CREATE TABLE dbo.[DimCustomer](
    [CustomerKey] int NOT NULL,
    [CustomerSourceID] [int] NOT NULL,
    [Title] [nvarchar](8) NULL,
    [FirstName] [nvarchar](50) NOT NULL,
    [MiddleName] [nvarchar](50) NULL,
    [LastName] [nvarchar](50) NOT NULL,
    [Suffix] [nvarchar](10) NULL,
    [CompanyName] [nvarchar](128) NULL,
    [SalesPerson] [nvarchar](256) NULL,
    [EmailAddress] [nvarchar](50) NULL,
    [Phone] [nvarchar](25) NULL,
    [InsertedDate] [datetime] NOT NULL,
    [ModifiedDate] [datetime] NOT NULL,
    [HashKey] [varchar](100) NOT NULL,
    [IsCurrentRow] [bit] NOT NULL
)
WITH
(
    DISTRIBUTION = REPLICATE,
    CLUSTERED COLUMNSTORE INDEX
);
GO
```

A. [HistoricalSalesPerson] [nvarchar] (256) NOT NULL

B. [EffectiveEndDate] [datetime] NOT NULL

C. [PreviousModifiedDate] [datetime] NOT NULL

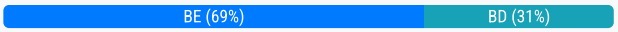
D. [RowID] [bigint] NOT NULL

E. [EffectiveStartDate] [datetime] NOT NULL

Suggested Answer: *B, D*

Official Answer: BD

Community vote distribution



aditya816 Highly Voted 3 years, 4 months ago

Selected Answer: BE

Surrogate is already there as customerkey column

upvoted 18 times

aditya816 Highly Voted 3 years, 4 months ago

Surrogate is already there as customerkey columnthat's the business key, not the surrogate key. If a new record is generated, there would be a duplicate key. SK is necessary to identify the recordNo, the 'CustomerKey' is the Surrogate Key. Moreover, a Business Key also already exists in DimCustomer table by the name 'CustomerSourceID'. So, B&E are the correct options.

upvoted 18 times

laurasscastro 3 years, 3 months ago

that's the business key, not the surrogate key. If a new record is generated, there would be a duplicate key. SK is necessary to identify the recordNo, the 'CustomerKey' is the Surrogate Key. Moreover, a Business Key also already exists in DimCustomer table by the name 'CustomerSourceID'. So, B&E are the correct options.

upvoted 8 times

laurasscastro 3 years, 3 months ago

that's the business key, not the surrogate key. If a new record is generated, there would be a duplicate key. SK is necessary to identify the record

upvoted 8 times

phydev 2 years, 9 months ago

No, the 'CustomerKey' is the Surrogate Key. Moreover, a Business Key also already exists in DimCustomer table by the name 'CustomerSourceID'. So, B&E are the correct options.

upvoted 3 times

[Removed] 3 years, 3 months ago

Selected Answer: BD

I think, there is already a column called InsertedDate, therefore E is not necessary. So we just need another column to track the end date, which is B. And RowID should be a surrogate key in this case.

upvoted 11 times

[Removed] 3 years, 3 months ago

I think, there is already a column called InsertedDate, therefore E is not necessary. So we just need another column to track the end date, which is B. And RowID should be a surrogate key in this case.The date of insertion and the expiration date from when to when is something else. You can insert data now, but either with future validity or with past validity (correcting errors, for example). So options : BE

upvoted 11 times

jiriz 2 years, 10 months ago

The date of insertion and the expiration date from when to when is something else. You can insert data now, but either with future validity or with past validity (correcting errors, for example). So options : BE

upvoted 7 times

JustImperius 1 year, 6 months ago

Selected Answer: BD

Its a bad question. Looking at the table creation query CustomerKey is not clearly a surrogate key due to the lack of IDENTITY or NEWID() etc. You cannot have a SCD 2 table without a surrogate key. So assuming CustomerKey is not a surrogate key (due to the evidence provided in the query) you need that -> D We only have one spot left. I would prefer to have both effective start and end date but if we had to choose the choice is effectiveEndDate because InsertedDate can act as the defacto effectiveStartDate -> B

If you can prove to me the CustomerKey is a surrogate key I will change my mind. The code says no. And before you tell me CustomerSourceId is the business key I would say the CustomerSourceId in the context of a DimCustomer table likely represents an identifier used to link the customer record to its original source system. But its really all conjecture...go with the code and the code says that there ain't no surrogate key present.

upvoted 1 times

JustImperius 1 year, 6 months ago

Its a bad question. Looking at the table creation query CustomerKey is not clearly a surrogate key due to the lack of IDENTITY or NEWID() etc. You cannot have a SCD 2 table without a surrogate key. So assuming CustomerKey is not a surrogate key (due to the evidence provided in the query) you need that -> D We only have one spot left. I would prefer to have both effective start and end date but if we had to choose the choice is effectiveEndDate because InsertedDate can act as the defacto effectiveStartDate -> B If you can prove to me the CustomerKey is a surrogate key I will change my mind. The code says no. And before you tell me CustomerSourceId is the business key I would say the CustomerSourceId in the context of a DimCustomer table likely represents an identifier used to link the customer record to its original source system. But its really all conjecture...go with the code and the code says that there ain't no surrogate key present.You're overcomplicating things. If it needs to be an identity, why is answer D just a BIGINT? Why is your surrogate key a BIGINT; nothing in the question mentions any volumes for you to be concerned of running out of values. Secondly, it is not good practice to "kind-of" implement a model because you have columns that seem similar. Inserted data and EffectiveStartDate serve different purposes. Are your EffectiveStartDate and InsertedDate the same for late arriving rows?

upvoted 1 times

PreQL Most Recent 1 year, 6 months ago

You're overcomplicating things. If it needs to be an identity, why is answer D just a BIGINT? Why is your surrogate key a BIGINT; nothing in the question mentions any volumes for you to be concerned of running out of values. Secondly, it is not good practice to "kind-of" implement a model because you have columns that seem similar. Inserted data and EffectiveStartDate serve different purposes. Are your EffectiveStartDate and InsertedDate the same for late arriving rows?

upvoted 1 times

monakl23 1 year, 8 months ago

Answer: BE. Link - https://learn.microsoft.com/en-us/training/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types

upvoted 1 times

EmnCours 1 year, 8 months ago

Correct Answer: BE

upvoted 1 times

ahana1074 1 year, 11 months ago

wrong answer by Xam topic correct answer with solution is-: [EffectiveStartDate] [datetime] NOT NULL: This column tracks when the current record started being valid, helping to define the validity period of the data. [EffectiveEndDate] [datetime] NOT NULL: This column marks the end of the record's validity period when a new version of the data is inserted. These two columns will allow you to differentiate between active and historical records and handle Type 2 SCD effectively. BE

upvoted 1 times

f2a9aa5 2 years ago

We need a surrogate key. The Inserteddate is NOT NULL, so will act as EffectiveStartDate. However, ModifiedDate is also NOT NULL, so it cannot work as EffectiveEndDate because the current record must have NULL for its EffectiveEndDate. So, we need EffectiveEndDate. Answer: D,E.

upvoted 1 times

DanielCruz 2 years, 1 month ago

the right answer depends on a few topics not clear: - CustumerKey may be a surrogate key? - InsertedDate may be a the "EffectiveStartDate" - modifiedDate may be the "EffectiveEndDate"

upvoted 1 times

sergio_eduardo 2 years, 2 months ago

The subrogatekey is already

upvoted 1 times

MBRSDG 2 years, 4 months ago

B --> end date is missing, and required by SCD2 D --> (not E), RowID is required since in SCD2 you're adding the same CustomerID twice, even with a different end date. So, you need a way to uniquely identify a row in the table, that's going no longer to be the customer identifier in general.

upvoted 5 times

AbhiJain1993 2 years, 4 months ago

It should be BD. If CustomerKey was surrogate key then IDENTITY should have been mentioned in Column definition.

upvoted 2 times

sdg2844 2 years, 7 months ago

There is already a hash key that serves as the surrogate, if I'm not mistaken. Inserted and modified are probably dates from the source data, not from the work being done here, so you need to add the start/end dates.

upvoted 2 times

hassexat 2 years, 11 months ago

B and E

upvoted 1 times

AvSUN 2 years, 11 months ago

B and D we need a unique row identifier

upvoted 2 times

kkk5566 2 years, 11 months ago

B and D ,its a star schema on which has a fact table include a customerID property.

upvoted 1 times

kkk5566 2 years, 11 months ago

B and D ,its a star schema on which has a fact table include a customerID property.after think twice ,B&E

upvoted 1 times

kkk5566 2 years, 11 months ago

after think twice ,B&E

upvoted 1 times

Deeksha1234 3 years ago

B and D makes more sense, since inserted date is there already

upvoted 1 times

HOTSPOT

-

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool.

You plan to deploy a solution that will analyze sales data and include the following:

• A table named Country that will contain 195 rows

• A table named Sales that will contain 100 million rows

• A query to identify total sales by country and customer from the past 30 days

You need to create the tables. The solution must maximize query performance.

How should you complete the script? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

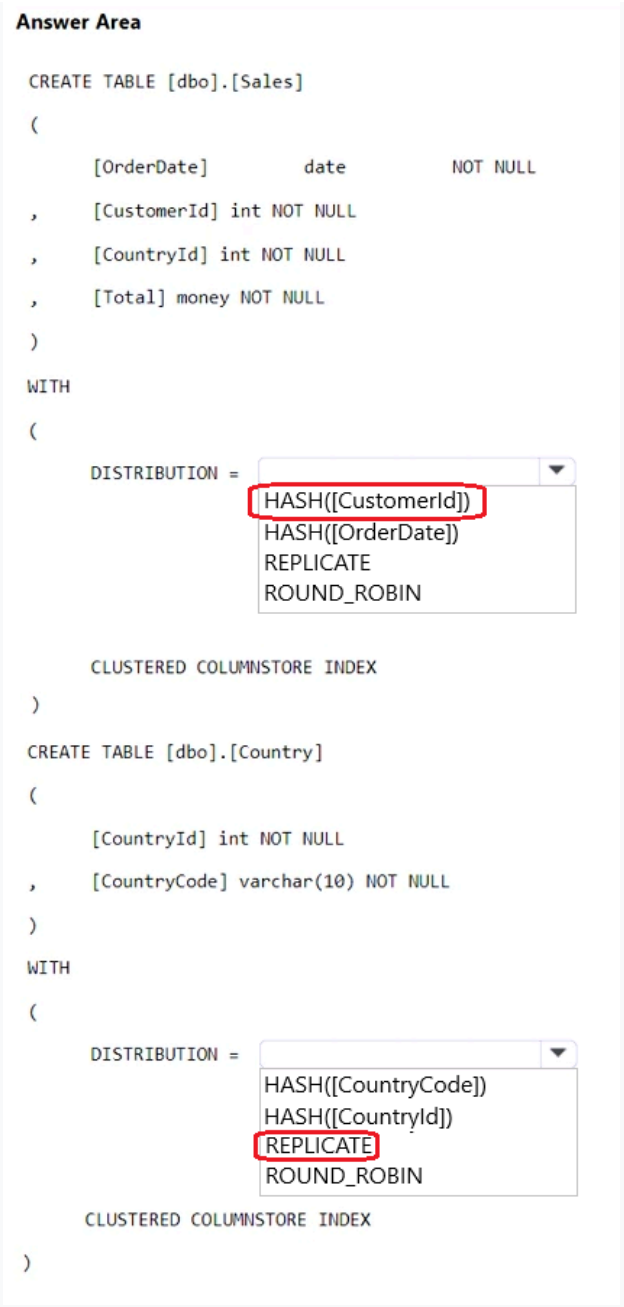
```
CREATE TABLE [dbo].[Sales]
(
      [OrderDate]         date          NOT NULL
,     [CustomerId] int NOT NULL
,     [CountryId] int NOT NULL
,     [Total] money NOT NULL
)
WITH
(
      DISTRIBUTION =
                     HASH([CustomerId])
                     HASH([OrderDate])
                     REPLICATE
                     ROUND_ROBIN

      CLUSTERED COLUMNSTORE INDEX
)
CREATE TABLE [dbo].[Country]
(
      [CountryId] int NOT NULL
,     [CountryCode] varchar(10) NOT NULL
)
WITH
(
      DISTRIBUTION =
                     HASH([CountryCode])
                     HASH([CountryId])
                     REPLICATE
                     ROUND_ROBIN
     CLUSTERED COLUMNSTORE INDEX
)
```

Suggested Answer:

Answer Area

```
CREATE TABLE [dbo].[Sales]
(
      [OrderDate]         date          NOT NULL
,     [CustomerId] int NOT NULL
,     [CountryId] int NOT NULL
,     [Total] money NOT NULL
)
WITH
(
      DISTRIBUTION =
                    HASH([CustomerId])
                    HASH([OrderDate])
                    REPLICATE
                    ROUND_ROBIN

      CLUSTERED COLUMNSTORE INDEX
)
CREATE TABLE [dbo].[Country]
(
      [CountryId] int NOT NULL
,     [CountryCode] varchar(10) NOT NULL
)
WITH
(
      DISTRIBUTION =
                    HASH([CountryCode])
                    HASH([CountryId])
                    REPLICATE
                    ROUND_ROBIN
     CLUSTERED COLUMNSTORE INDEX
)
```

Answer Area



FRANCIS_A_M Highly Voted 3 years, 4 months ago

Correct! 1. Hash(CustomerID) 2. Replicate

upvoted 21 times

FRANCIS_A_M Highly Voted 3 years, 4 months ago

Correct! 1. Hash(CustomerID) 2. ReplicateCould you please explain why 1. Hash([CustomerID]) is correct, and 2. Hash([OrderDate]) is incorrect.Never distribute on Date.Don't hash on date, only partition on dateIt is hash because it is a fact table (you can tell because there is the "total" column being created which is numerical). Rule of thumb, never hash on a date field, so in this case you would hash on 'CustomerID'. You want the hash to have as many unique values as possible.

upvoted 21 times

zafnad 3 years, 3 months ago

Could you please explain why 1. Hash([CustomerID]) is correct, and 2. Hash([OrderDate]) is incorrect.Never distribute on Date.Don't hash on date, only partition on dateIt is hash because it is a fact table (you can tell because there is the "total" column being created which is numerical). Rule of thumb, never hash on a date field, so in this case you would hash on 'CustomerID'. You want the hash to have as many unique values as possible.

upvoted 2 times

zafnad 3 years, 3 months ago

Could you please explain why 1. Hash([CustomerID]) is correct, and 2. Hash([OrderDate]) is incorrect.

upvoted 2 times

vctrhugo 3 years, 1 month ago

Never distribute on Date.

upvoted 5 times

Spam_Account 3 years, 1 month ago

Don't hash on date, only partition on date

upvoted 16 times

ajhak 3 years, 3 months ago

It is hash because it is a fact table (you can tell because there is the "total" column being created which is numerical). Rule of thumb, never hash on a date field, so in this case you would hash on 'CustomerID'. You want the hash to have as many unique values as possible.

upvoted 7 times

AnhHoang Most Recent 2 years, 1 month ago

1. You would want hash distribution to improve query performance. You don't want to hash on Date column since it can cause bottlenecks if many people query on a same date (eg. getdate()), so hash on customerid is the way. 2. You would want to replicate small tables across all distributions, so it can pick up any distribution and still have full data.

upvoted 2 times

kkk5566 2 years, 11 months ago

1. Hash(CustomerID) 2. Replicate

upvoted 2 times

Deeksha1234 3 years ago

given answer is correct

upvoted 1 times

examtopicsofyannick 3 years ago

Correct. Hash on Sales Table(Fact) and Replicate on Country table(Dimension)

upvoted 2 times

nmnm22 3 years, 4 months ago

correct

upvoted 4 times

Question #74 *Topic 1*

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named account1 and an Azure Synapse Analytics workspace named workspace1.
You need to create an external table in a serverless SQL pool in workspace1. The external table will reference CSV files stored in account1. The solution must maximize performance.
How should you configure the external table?

A. Use a native external table and authenticate by using a shared access signature (SAS).

B. Use a native external table and authenticate by using a storage account key.

C. Use an Apache Hadoop external table and authenticate by using a shared access signature (SAS).

D. Use an Apache Hadoop external table and authenticate by using a service principal in Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra.

Suggested Answer: *A*
Official Answer: A

Community vote distribution
A (96%) (4

FRANCIS_A_M Highly Voted 3 years, 4 months ago
Selected Answer: A
Correct! Serverless SQL Pools cannot use Hadoop, Only Native. Access Key Auth is never best practice therefore leaving only A as a viable answer.
https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop
upvoted 34 times

FRANCIS_A_M Highly Voted 3 years, 4 months ago
Correct! Serverless SQL Pools cannot use Hadoop, Only Native. Access Key Auth is never best practice therefore leaving only A as a viable answer.
https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoopthanks a lot for the good explanationIt's not about the best practice - there is no option to use storage keys... https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#securityIndeed, to narrow it down, it's shown in the table from this link, from the row "Storage authentication", the Native column only shows Shared Access Signature(SAS), Microsoft Entra passthrough, Managed identity, Custom application Microsoft Entra identity.
upvoted 34 times

nmnm22 3 years, 4 months ago
thanks a lot for the good explanation
upvoted 5 times

Rob77 3 years, 2 months ago
It's not about the best practice - there is no option to use storage keys... https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#security
upvoted 7 times

Rob77 3 years, 2 months ago
It's not about the best practice - there is no option to use storage keys... https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#securityIndeed, to narrow it down, it's shown in the table from this link, from the row "Storage authentication", the Native column only shows Shared Access Signature(SAS), Microsoft Entra passthrough, Managed identity, Custom application Microsoft Entra identity.
upvoted 7 times

hypersam 1 year, 6 months ago
Indeed, to narrow it down, it's shown in the table from this link, from the row "Storage authentication", the Native column only shows Shared Access Signature(SAS), Microsoft Entra passthrough, Managed identity, Custom application Microsoft Entra identity.
upvoted 1 times

vctrhugo 3 years, 1 month ago
The other options provided (B, C, and D) are not the recommended configurations for maximizing performance in this scenario. Using a storage account key for authentication (option B) poses a security risk and should be avoided. Apache Hadoop external tables (options C and D) do not

provide the same level of performance optimization as native external tables in Azure Synapse Analytics.

upvoted 5 times

samirarian Most Recent 1 year, 6 months ago

Why Native External Table? Native external tables (CREATE EXTERNAL TABLE) in Synapse serverless SQL pools are optimized for querying external data sources such as Azure Data Lake Storage Gen2. They provide better query optimization than Apache Hadoop external tables. Why Storage Account Key for Authentication? Using the storage account key ensures fast and direct access to the storage account without requiring token exchanges or additional configurations. While a shared access signature (SAS) works for temporary scenarios, it isn't as straightforward for long-term performance and scalability.

upvoted 1 times

moize 1 year, 7 months ago

Pour optimiser les performances lors de la création d'une table externe dans un pool SQL sans serveur dans Azure Synapse Analytics, il est recommandé d'utiliser une table externe native et de s'authentifier à l'aide d'une clé de compte de stockage (option B) L'utilisation d'une table externe native permet de tirer parti des optimisations spécifiques à Azure Synapse pour les fichiers CSV. L'authentification avec une clé de compte de stockage est simple à configurer et offre de bonnes performances pour accéder aux données stockées dans Azure Data Lake Storage Gen2.

upvoted 1 times

moize 1 year, 8 months ago

Pour créer une table externe dans un pool SQL sans serveur dans Azure Synapse Analytics et optimiser les performances, vous devez utiliser une table externe native. Voici la configuration recommandée : A. Utilisez une table externe native et authentifiez-vous à l'aide d'une signature d'accès partagé (SAS). Raisons : Tables Externes Natives : Les tables externes natives sont optimisées pour les pools SQL sans serveur dans Azure Synapse Analytics et offrent de meilleures performances par rapport aux tables externes Hadoop1. Authentification avec SAS : Utiliser une signature d'accès partagé (SAS) est une méthode sécurisée et flexible pour accéder aux données stockées dans Azure Data Lake Storage Gen2.

upvoted 1 times

EmnCours 1 year, 8 months ago

Correct Answer: A

upvoted 1 times

Dusica 2 years, 3 months ago

A for sure

upvoted 1 times

moneytime 2 years, 6 months ago

A is correct. Native external tables perform better than hadoop external table. Also ,storage account is more secure with the scheme of SAS
https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables

upvoted 1 times

Charley92 2 years, 6 months ago

To maximize performance when creating an external table in a serverless SQL pool in Azure Synapse Analytics workspace1 that references CSV files stored in account1, you should use a native external table and authenticate by using a shared access signature (SAS). Native external tables are designed to read and export data in various data formats such as CSV and Parquet. They are available in serverless SQL pools and are in public preview in dedicated SQL pools. Using an SAS to authenticate provides a secure way to access the data in account1 without exposing the storage account key.

upvoted 1 times

kkk5566 2 years, 11 months ago

is correct

upvoted 1 times

Deeksha1234 3 years ago

A is correct

upvoted 1 times

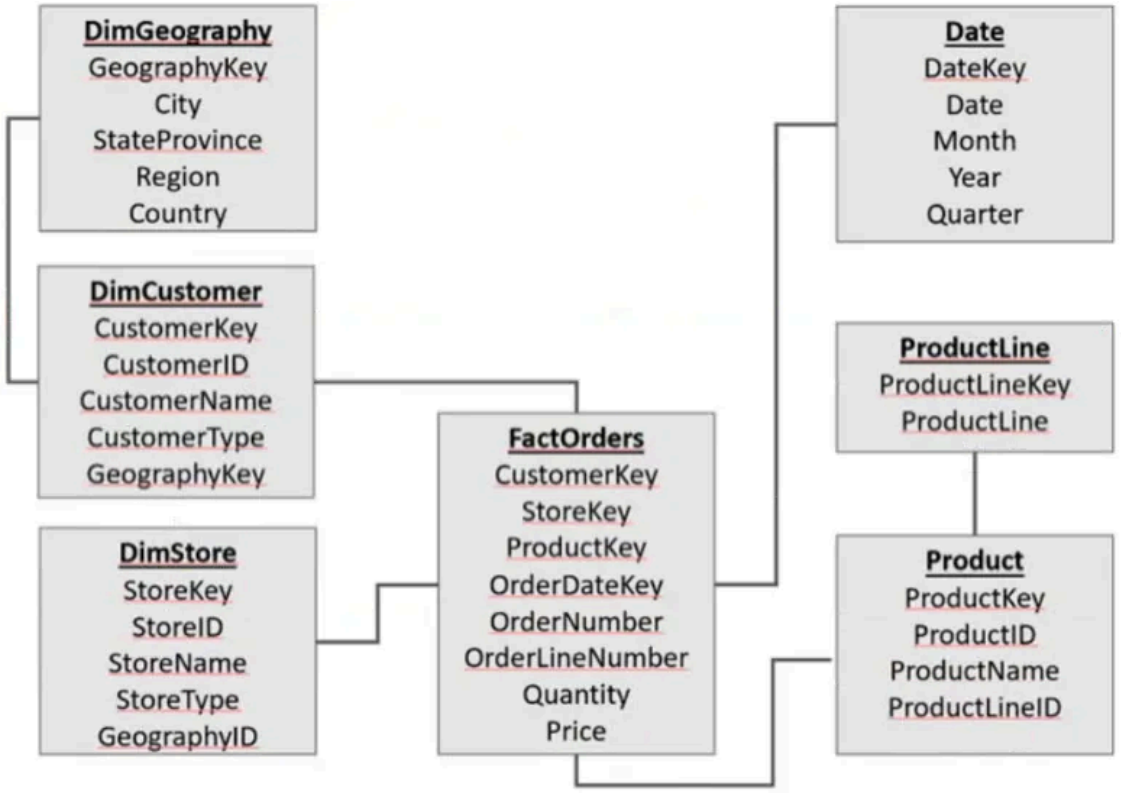
HOTSPOT

-

You have an Azure Synapse Analytics serverless SQL pool that contains a database named db1. The data model for db1 is shown in the following exhibit.

Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the exhibit.

NOTE: Each correct selection is worth one point.



Answer Area

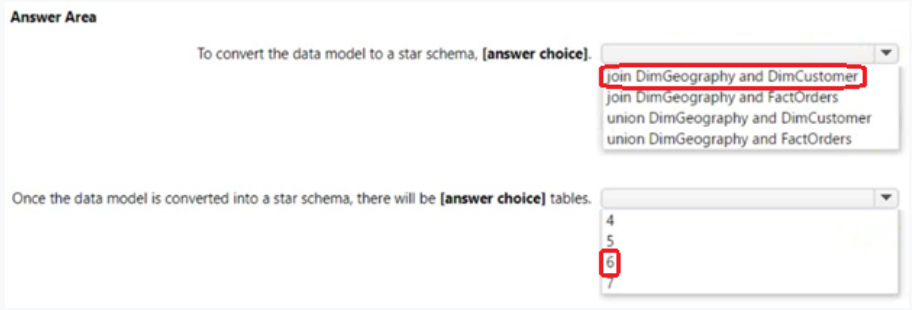
To convert the data model to a star schema, **[answer choice]**.

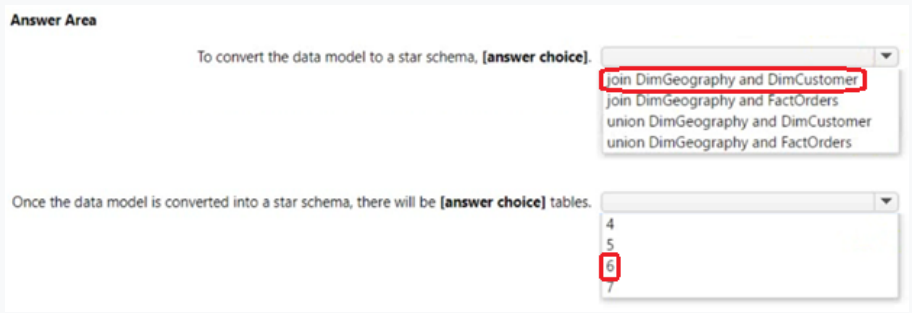
- join DimGeography and DimCustomer
- join DimGeography and FactOrders
- union DimGeography and DimCustomer
- union DimGeography and FactOrders

Once the data model is converted into a star schema, there will be **[answer choice]** tables.

- 4
- 5
- 6
- 7

Suggested Answer:





blenau Highly Voted 3 years, 4 months ago

Correct answer should be join DimGeography and DimCustomer and 5 tables.

You also need to combine ProductLine and Product in order for the schema to be considered a star schema. This would result in 5 remaining tables: DimCustomer (DimCustomer JOIN DimGeography), DimStore, Date, Product (Product JOIN ProductLine) and FactOrders.

upvoted 95 times

blenau Highly Voted 3 years, 4 months ago

Correct answer should be join DimGeography and DimCustomer and 5 tables. You also need to combine ProductLine and Product in order for the schema to be considered a star schema. This would result in 5 remaining tables: DimCustomer (DimCustomer JOIN DimGeography), DimStore, Date, Product (Product JOIN ProductLine) and FactOrders.It is a conversion of snow flake schema to star schema.AllThis is true, but we are limited by what possible answers the question provides us. So the ExamTopics selected answers are correct.

upvoted 95 times

moneytime 2 years, 6 months ago

It is a conversion of snow flake schema to star schema.All

upvoted 2 times

7082935 2 years ago

This is true, but we are limited by what possible answers the question provides us. So the ExamTopics selected answers are correct.

upvoted 2 times

Dataminer 3 years, 4 months ago

Agree with explanation. It will still be snowflake if Product and ProductLine is not combined

upvoted 20 times

renan_ineu 1 year, 11 months ago

I made the exam this week (09/09/2024) and this question was present but slightly different.

upvoted 3 times

Nadine_nm 2 years, 1 month ago

We should join the FacOrders with the DimGeography In a star schema we have a fact table and dimensions, and dimensions shouldn't be joined also since we are making a join we are keeping the 6 dimensions + the fact table = 7

upvoted 3 times

Alongi 2 years, 3 months ago

I found this question on my exam 30/04/2024, and I put - join dimGeography and dimCustomer - 5 I passed the exam with a high score, but I'm not sure if the answer is correct.

upvoted 1 times

dark_one 2 years, 4 months ago

If 2 tables are joined there are still 2 tables, we're not UNIONing them so still 7 tables

upvoted 1 times

Monadire 2 years, 4 months ago

Provided answer is correct: why not 5 tables? Please, look at schema, and count once more, how many tables are at all including fact as well :) So Join Geography with Customers = overall 6 tablesThe answer should be "join DimGeography and DimCustomer", as a star schema has no 2nd grade dimensions. The second question is tricky, it asks `Once the data model is converted to a STAR schema..' which insinuates that no 2nd grade dimension tables are left. As said earlier, this means the DimProduct and DimProductCategories should and would also be joined. This would leave us with the 4 dimensions and the fact table, which is 5 tables in total.Sorry, the correct is 5, bcs question is tricky and we still have to union ProductLine with Product to have at the end star schema

upvoted 1 times

Monadire 2 years, 4 months ago

Provided answer is correct: why not 5 tables? Please, look at schema, and count once more, how many tables are at all including fact as well :) So Join Geography with Customers = overall 6 tables

upvoted 1 times

chefDE Most Recent 1 year, 8 months ago

The answer should be "join DimGeography and DimCustomer", as a star schema has no 2nd grade dimensions. The second question is tricky, it asks `Once the data model is converted to a STAR schema..' which insinuates that no 2nd grade dimension tables are left. As said earlier, this means the DimProduct and DimProductCategories should and would also be joined. This would leave us with the 4 dimensions and the fact table, which is 5 tables in total.

upvoted 1 times

Monadire 2 years, 4 months ago

Sorry, the correct is 5, bcs question is tricky and we still have to union ProductLine with Product to have at the end star schema

upvoted 2 times

ChrisGe1234 2 years, 6 months ago

If you design DimGeography as Role Playing Dimension (How its usually done), then join Geography and Fact. Afterwards you could merge the Product Group and Product Table which leads to 6 Tables. So I would argue correct would be A: Join DimGeography and Fact B: 6Wrong and wrong

upvoted 2 times

ChrisGe1234 2 years, 6 months ago

If you design DimGeography as Role Playing Dimension (How its usually done), then join Geography and Fact. Afterwards you could merge the Product Group and Product Table which leads to 6 Tables. So I would argue correct would be A: Join DimGeography and Fact B: 6

upvoted 2 times

Fusejonny1 2 years, 6 months ago

Wrong and wrong

upvoted 3 times

sdg2844 2 years, 7 months ago

I think maybe we all missed something here. If it's star schema, there is nothing hanging off the outside of the outside tables. DimGeography should be joined to FactSales, with the geography placed in the FactSales Table. However, it doesn't solve the problem of Product and ProductCategory, which need to be combined. So there is just part of the answer missing. Once those two items are done, then there are 5 tables remainin.

upvoted 2 times

sdg2844 2 years, 7 months ago

I think maybe we all missed something here. If it's star schema, there is nothing hanging off the outside of the outside tables. DimGeography should be joined to FactSales, with the geography placed in the FactSales Table. However, it doesn't solve the problem of Product and ProductCategory, which need to be combined. So there is just part of the answer missing. Once those two items are done, then there are 5 tables remainin.There would still be 6 tables

upvoted 2 times

s_unsworth 2 years, 5 months ago

There would still be 6 tables

upvoted 2 times

surajpdh 2 years, 9 months ago

it's a tricky question , it says once we join DimGeography and DimCustomer then how many tables will remain in data model. Answer is 6.

upvoted 2 times

matiandal 2 years, 10 months ago

num of tables (dims + facts) == aka ==> 6

upvoted 4 times

ellala 2 years, 10 months ago

Correct would be: 1) join Geography with customer 2) (then join productline and product - this is not in the question, but must be done to transform into a star schema) 3) then we have 5 tables since Geography and ProductLine are no longer needed.

upvoted 8 times

hassexat 2 years, 11 months ago

join DimGeography and DimCustomer 5 tables

upvoted 3 times

AvSUN 2 years, 11 months ago

shouldn't it be 5 tables?

upvoted 3 times

kkk5566 2 years, 11 months ago

1 DimGeography and DimCustomer 2. 5 tables.

upvoted 3 times

ccesarrg 3 years ago

This question is really messy. It doesn't explicit say that by joining or unioning the tables this means they will be combined into a single table, to be it seems like we'll still have 2 tables (DimGeography and DimCustomer) in both options, besides the fact that just fixing DimGeography and DImCustomer won't generate a Star Schema

upvoted 1 times

Deeksha1234 3 years ago

join DimGeography and DimCustomer and 5 tables

upvoted 2 times

Question #76 *Topic 1*

You have an Azure Databricks workspace and an Azure Data Lake Storage Gen2 account named storage1.
New files are uploaded daily to storage1.
You need to recommend a solution that configures storage1 as a structured streaming source. The solution must meet the following requirements:
• Incrementally process new files as they are uploaded to storage1.
• Minimize implementation and maintenance effort.
• Minimize the cost of processing millions of files.
• Support schema inference and schema drift.
Which should you include in the recommendation?

A. COPY INTO

B. Azure Data Factory

C. Auto Loader

D. Apache Spark FileStreamSource

Suggested Answer: *C*
Official Answer: C

Community vote distribution
C (91%) D (9%)

Nikiboy Highly Voted 3 years, 4 months ago
Auto Loader provides a Structured Streaming source called cloudFiles. Plus, it supports schema drift. Hence, Auto Loader is the correct answer. https://learn.microsoft.com/en-us/azure/databricks/ingestion/auto-loader/Auto Loader does not support Azure Data Lake Storage Gen2It does. Refer this link https://learn.microsoft.com/en-us/azure/databricks/ingestion/auto-loader/Auto Loader can load data files from AWS S3 (s3://), Azure Data Lake Storage Gen2 (ADLS Gen2, abfss://), Google Cloud Storage (GCS, gs://), Azure Blob Storage (wasbs://), ADLS Gen1 (adl://), and Databricks File System (DBFS, dbfs:/).
upvoted 18 times

Nikiboy Highly Voted 3 years, 4 months ago
Auto Loader provides a Structured Streaming source called cloudFiles. Plus, it supports schema drift. Hence, Auto Loader is the correct answer. https://learn.microsoft.com/en-us/azure/databricks/ingestion/auto-loader/
upvoted 18 times

mr_examers 3 years, 3 months ago
Auto Loader does not support Azure Data Lake Storage Gen2
upvoted 1 times

mr_examers 3 years, 3 months ago
Auto Loader does not support Azure Data Lake Storage Gen2It does. Refer this link https://learn.microsoft.com/en-us/azure/databricks/ingestion/auto-loader/Auto Loader can load data files from AWS S3 (s3://), Azure Data Lake Storage Gen2 (ADLS Gen2, abfss://), Google Cloud Storage (GCS, gs://), Azure Blob Storage (wasbs://), ADLS Gen1 (adl://), and Databricks File System (DBFS, dbfs:/).
upvoted 1 times

cloud_lady 3 years, 3 months ago
It does. Refer this link https://learn.microsoft.com/en-us/azure/databricks/ingestion/auto-loader/
upvoted 1 times

vctrhugo 3 years, 1 month ago
Auto Loader can load data files from AWS S3 (s3://), Azure Data Lake Storage Gen2 (ADLS Gen2, abfss://), Google Cloud Storage (GCS, gs://), Azure Blob Storage (wasbs://), ADLS Gen1 (adl://), and Databricks File System (DBFS, dbfs:/).
upvoted 3 times

samianae Most Recent 1 year, 6 months ago
Auto Loader
upvoted 1 times

moize 1 year, 8 months ago

Auto Loader (option C) est la solution recommandée pour configurer storage1 comme source de streaming structurée dans Azure Databricks.

upvoted 1 times

EmnCours 1 year, 8 months ago

Auto Loader is correc

upvoted 1 times

ahana1074 1 year, 11 months ago

autoloader is correct-:Incremental processing: Auto Loader can automatically detect and incrementally process new files as they are uploaded to Azure Data Lake Storage Gen2. Minimize implementation and maintenance effort: Auto Loader is designed for simplicity, requiring minimal setup and automatically handling file management. It reduces operational overhead by automating many of the tasks required to manage a streaming source. Minimize the cost of processing millions of files: Auto Loader efficiently scales to handle millions of files and minimizes costs by using a directory listing mode or a more optimized file notification mode with Azure Event Grid. Support schema inference and schema drift: Auto Loader automatically infers schema and can handle schema drift, which allows it to dynamically adapt to changes in the file structure without requiring constant updates to the processing logic.

upvoted 1 times

Alongi 2 years, 4 months ago

Auto Loader is correct

upvoted 3 times

Homer23 2 years, 5 months ago

Reference: https://learn.microsoft.com/en-us/azure/databricks/ingestion/auto-loader/#incremental-ingestion-using-auto-loader-with-delta-live-tables

upvoted 1 times

Bill_Walker 2 years, 6 months ago

Auto Loader seems more correct. Copy Into focuses on loading from storage to a Delta table

upvoted 1 times

Charley92 2 years, 6 months ago

To configure storage1 as a structured streaming source that incrementally processes new files as they are uploaded minimizes implementation and maintenance effort, minimizes the cost of processing millions of files, and supports schema inference and schema drift, you should use Apache Spark FileStreamSource

upvoted 1 times

ellala 2 years, 10 months ago

Bing explains the following: The best option is C. Auto Loader. Auto Loader is a feature in Azure Databricks that uses a cloudFiles data source to incrementally and efficiently process new data files as they arrive in Azure Data Lake Storage Gen2. It supports schema inference and schema evolution (drift). It also minimizes implementation and maintenance effort, as it simplifies the ETL pipeline by reducing the complexity of identifying new files for processing. Other options do not meet the requirements because: A. COPY INTO: does not incrementally process new files as they are uploaded, which is one of your requirements. B. Azure Data Factory: does not natively support schema inference and schema drift. The incremental processing of new files would need to be manually implemented, which could increase implementation and maintenance effort. D. Apache Spark FileStreamSource: requires manual setup and does not natively support schema inference or schema drift. It also may not minimize the cost of processing millions of files as efficiently as Auto Loader.

upvoted 4 times

kkk5566 2 years, 11 months ago

Auto Loader

upvoted 1 times

Deeksha1234 3 years ago

C is correct

upvoted 1 times

vctrhugo 3 years, 2 months ago

To configure Azure Data Lake Storage Gen2 account (storage1) as a structured streaming source in Azure Databricks workspace, while meeting the given requirements, you should include the following in the recommendation: C. Auto Loader Auto Loader is a feature provided by Azure Databricks that automatically discovers and processes new files as they are uploaded to a specified directory in Azure Data Lake Storage Gen2. It provides an efficient and cost-effective way to incrementally process new files without the need for manual intervention. Auto Loader also supports schema inference and schema drift, allowing you to handle changes in the file schema over time. By using Auto Loader, you can minimize implementation and maintenance effort as it takes care of monitoring the storage directory for new files and processing them in an optimized manner. It also helps to minimize the cost of processing millions of files as it leverages the efficient processing capabilities of Databricks. Therefore, the correct answer is C. Auto Loader.

upvoted 4 times

rocky48 3 years, 3 months ago

Auto LoaderI recommend using Auto Loader. Here's why: Incremental Processing: Auto Loader in Azure Databricks allows you to process new files incrementally as they are uploaded to your storage account. It efficiently identifies and processes only the new data, reducing the need to reprocess entire datasets. Low Implementation and Maintenance Effort: Auto Loader simplifies the setup process. You can configure it easily within your Databricks workspace, and it automatically handles file discovery, partitioning, and schema inference. Cost-Effective: Auto Loader optimizes resource usage by processing only the necessary data. It avoids unnecessary scans of existing files, which helps minimize costs when dealing with millions of files. Schema Inference and Schema Drift Support: Auto Loader automatically infers the schema from the data and adapts to schema changes over time (schema drift). This flexibility ensures smooth processing even when the structure of incoming files evolves. Therefore, choose C.

upvoted 1 times

rocky48 3 years, 3 months ago

Selected Answer: C

Auto Loader

upvoted 1 times

rocky48 2 years, 6 months ago

I recommend using Auto Loader. Here's why: Incremental Processing: Auto Loader in Azure Databricks allows you to process new files incrementally as they are uploaded to your storage account. It efficiently identifies and processes only the new data, reducing the need to reprocess entire datasets. Low Implementation and Maintenance Effort: Auto Loader simplifies the setup process. You can configure it easily within your Databricks workspace, and it automatically handles file discovery, partitioning, and schema inference. Cost-Effective: Auto Loader optimizes resource usage by processing only the necessary data. It avoids unnecessary scans of existing files, which helps minimize costs when dealing with millions of files. Schema Inference and Schema Drift Support: Auto Loader automatically infers the schema from the data and adapts to schema changes over time (schema drift). This flexibility ensures smooth processing even when the structure of incoming files evolves. Therefore, choose C.

upvoted 1 times

nicololmen 3 years, 3 months ago

D according to ChatGPT

upvoted 1 times

AHUI 3 years, 4 months ago

Ans : B DF supports Schema Drift - https://learn.microsoft.com/en-us/azure/data-factory/concepts-data-flow-schema-driftAuto Loader is lower cost.

upvoted 1 times

AHUI 3 years, 4 months ago

Ans : B

DF supports Schema Drift - https://learn.microsoft.com/en-us/azure/data-factory/concepts-data-flow-schema-drift

upvoted 1 times

frankanalysis 3 years, 4 months ago

Auto Loader is lower cost.

upvoted 2 times

Question #77 *Topic 1*

You have an Azure subscription that contains the resources shown in the following table.
You need to read the TSV files by using ad-hoc queries and the OPENROWSET function. The solution must assign a name and override the inferred data type of each column.
What should you include in the OPENROWSET function?

Name	Type	Description
storage1	Azure Blob storage account	Contains publicly accessible TSV files that do **NOT** have a header row
WS1	Azure Synapse Analytics workspace	Contains a serverless SQL pool

A. the WITH clause

B. the ROWSET_OPTIONS bulk option

C. the DATAFILETYPE bulk option

D. the DATA_SOURCE parameter

Suggested Answer: *D*
Official Answer: D

Community vote distribution
A (75%) D (25%)

henryphchan Highly Voted 3 years, 3 months ago
In the Question "The solution must assign a name and override the inferred data type of each column", so we must need a WITH Clause to define the column names and data types.
upvoted 18 times

19c1ee5 Highly Voted 3 years, 3 months ago
I think it's A. WITH CLAUSE https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-openrowset
upvoted 9 times

19c1ee5 3 years, 3 months ago
I think it's A. WITH CLAUSE https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-openrowsetAgreed - Should be A. "To specify explicit column names and data types, you can override the default column names and inferred data types by providing a schema definition in a WITH clause" (https://learn.microsoft.com/en-us/training/modules/query-data-lake-using-azure-synapse-serverless-sql-pools/3-query-files)
upvoted 9 times

jlad26 3 years, 3 months ago
Agreed - Should be A. "To specify explicit column names and data types, you can override the default column names and inferred data types by providing a schema definition in a WITH clause" (https://learn.microsoft.com/en-us/training/modules/query-data-lake-using-azure-synapse-serverless-sql-pools/3-query-files)
upvoted 6 times

EmnCours Most Recent 1 year, 8 months ago
Selected Answer: A
upvoted 1 times

ahana1074 1 year, 11 months ago
A is correct because-:SELECT * FROM OPENROWSET(BULK 'https://storage1.blob.core.windows.net/container/file.tsv', FORMAT='CSV', PARSER_VERSION='2.0', FIELDTERMINATOR = '\t', ROWTERMINATOR = '\n') WITH (Column1Name datatype1, Column2Name datatype2, Column3Name datatype3) AS result; B. ROWSET_OPTIONS bulk option: Configures aspects of data processing like buffering or parallelism, but it doesn't handle column names or types. C. DATAFILETYPE bulk option: Used to specify the file type but not for defining column names or types. D. DATA_SOURCE parameter: Points to an external data source, but it doesn't help with naming or overriding data types. Thus, the WITH clause is the correct option to include in the OPENROWSET function to assign column names and override inferred data types.
upvoted 5 times

DanielCruz 2 years, 1 month ago

DATA_SOURCE is not even an existing parameter option in th OPENROWSET function, the right existing option is DATASOURCE, without underscore, but it is not needed here, so A is the right option, see doc: https://learn.microsoft.com/en-us/sql/t-sql/functions/openrowset-transact-sql?view=sql-server-ver16

upvoted 1 times

MBRSDG 2 years, 4 months ago

Question insists on the point that the file has not the header row! specifying a datasource is not strictly required, since OPENROWSET can connect to storage even with a simple link and managed identity. The WITH clause is the most reasonable recommendation to include in the solution.

upvoted 1 times

Alongi 2 years, 4 months ago

There isn't header, so you have to specify columns name manually in WITH clause

upvoted 2 times

gplusplus 2 years, 4 months ago

Selected Answer: D

With clause can be omitted for csv files see automatic schema discovery at https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-openrowset#automatic-schema-discovery Generic colnames will be assigned

upvoted 1 times

gplusplus 2 years, 4 months ago

With clause can be omitted for csv files see automatic schema discovery at https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-openrowset#automatic-schema-discovery Generic colnames will be assignedon secodn thought, even if the types CAN be inferred, the question clearly states that the inferred types MUST be overwritten. Therefore WITH

upvoted 1 times

gplusplus 2 years, 4 months ago

on secodn thought, even if the types CAN be inferred, the question clearly states that the inferred types MUST be overwritten. Therefore WITH

upvoted 2 times

s_unsworth 2 years, 5 months ago

There is no header row, therefore you should define one using the WITH Clause

upvoted 1 times

Charley92 2 years, 6 months ago

The WITH clause allows you to specify the schema of the data source and override the inferred data type of each column. You can also use the WITH clause to specify the name of the external table and the location of the data source

upvoted 1 times

prshntdxt7 2 years, 6 months ago

WITH CLAUSE

upvoted 1 times

Momoanwar 2 years, 8 months ago

To read TSV files without a header row using the `OPENROWSET` function and to assign a name and specify the data type for each column, you should use: A. the WITH clause The WITH clause is used in the `OPENROWSET` function to define the format file or to directly define the structure of the file by specifying the column names and data types.

upvoted 1 times

hcq31818 2 years, 9 months ago

A - WITH Clause is the correct answer.

upvoted 1 times

Runaj 2 years, 9 months ago

D is right.

upvoted 1 times

jhargett1 2 years, 9 months ago

To read TSV (Tab-Separated Values) files using ad-hoc queries and the OPENROWSET function in Azure Synapse Analytics, and to assign a name and override the inferred data type of each column, you should include the following in the OPENROWSET function: A. the WITH clause The WITH clause allows you to specify options for reading the data, including defining the column names and data types. You can use the WITH clause to provide column definitions and specify the data type for each column in the TSV file, which allows you to override the inferred data types.

upvoted 2 times

Metaalverf 2 years, 9 months ago

They ask for "in the function", not "in the query"

upvoted 5 times

ellala 2 years, 10 months ago

Correct is A (explained in another comment)

upvoted 1 times

Question #78 *Topic 1*

You have an Azure Synapse Analytics dedicated SQL pool.
You plan to create a fact table named Table1 that will contain a clustered columnstore index.
You need to optimize data compression and query performance for Table1.
What is the minimum number of rows that Table1 should contain before you create partitions?

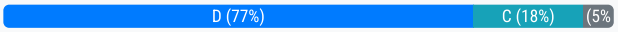
A. 100,000

B. 600,000

C. 1 million

D. 60 million

Suggested Answer: *A*
Official Answer: A

Community vote distribution


Ankit_Az Highly Voted 3 years, 2 months ago
Selected Answer: D
Clustered Column Store will by default have 60 partitions. And to achieve best compression we need at least 1 Million rows per partition, hence Option D 60 Millions (1M per partition)
upvoted 23 times

Ankit_Az Highly Voted 3 years, 2 months ago
Clustered Column Store will by default have 60 partitions. And to achieve best compression we need at least 1 Million rows per partition, hence Option D 60 Millions (1M per partition)You mean the dedicated SQL pool has 60 distributions "by default"?60 Million is correctNot by default, but always
upvoted 23 times

Vanq69 2 years, 10 months ago
You mean the dedicated SQL pool has 60 distributions "by default"?
upvoted 2 times

Vanq69 2 years, 10 months ago
You mean the dedicated SQL pool has 60 distributions "by default"?60 Million is correctNot by default, but always
upvoted 2 times

Lscranio 2 years, 8 months ago
60 Million is correct
upvoted 2 times

AkosL 2 years, 4 months ago
Not by default, but always
upvoted 1 times

EmnCours Most Recent 1 year, 8 months ago
Correct Answer: D
upvoted 1 times

ahana1074 1 year, 11 months ago
D is correct-:Partitioning in a dedicated SQL pool in Azure Synapse Analytics is typically used to manage very large tables, and the recommendation is to start considering partitions when the table contains 60 million rows or more. Partitioning helps optimize both data compression and query performance by allowing the system to process smaller subsets of data more efficiently. However, partitioning comes with overhead, and if you partition tables that are too small, it can actually degrade performance. Clustered columnstore indexes are designed to provide efficient compression and query performance for large datasets. Partitioning further helps with large fact tables, but for smaller tables (e.g., fewer than 60 million rows), partitioning is usually not necessary and might even be counterproductive.
upvoted 1 times

renan_ineu 1 year, 11 months ago

"When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributions." https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition#partition-sizing

upvoted 1 times

jaco_0930 2 years ago

1M * 60 = 60M

upvoted 1 times

AA9292 2 years, 1 month ago

it is 1 Million. the question is clearly asking for minimum number of rows before creating partitions. and in Microsoft document they stated that 1 million rows is the minimum number of rows before partitioning a tableThe question is asking the minimum number of rows in the whole table and not the number per partition. So, in order to have at least 1 million rows per partition, you must have 60 million rows in the fact table.

upvoted 2 times

AA9292 2 years, 1 month ago

it is 1 Million. the question is clearly asking for minimum number of rows before creating partitions. and in Microsoft document they stated that 1 million rows is the minimum number of rows before partitioning a table

upvoted 2 times

renan_ineu 1 year, 11 months ago

The question is asking the minimum number of rows in the whole table and not the number per partition. So, in order to have at least 1 million rows per partition, you must have 60 million rows in the fact table.

upvoted 1 times

Siva_Jsn_23 2 years, 2 months ago

1 Million (Option C) is correct. You need the minimum number of rows to create a optimized partition. A single optimal partition requires 1 Million rows.

upvoted 4 times

KarlGardnerDataEngineering 2 years, 3 months ago

I think the correct answer should be 120 million rows. Since splitting the data up into 2 partitions would result in 1 million rows per distribution and partition for 120 million rows

upvoted 1 times

MBRSDG 2 years, 4 months ago

even without knowing how many partitions you're going to create, you know that each partition should contain at least 1million rows. 60million rows are the only case enabling to use partitions.

upvoted 1 times

s_unsworth 2 years, 5 months ago

Cluster columnstore tables begin to achieve optimal compression once there is more than 60 million rows. For small lookup tables, less than 60 million rows, consider using HEAP or clustered index for faster query performance. -- Microsoft

upvoted 3 times

Charley92 2 years, 6 months ago

To optimize data compression and query performance for Table in Azure Synapse Analytics dedicated SQL pool, you should create partitions when the table contains at least 60 million rows. Partitioning tables can improve query performance by reducing the amount of data that needs to be scanned. It can also improve data compression by allowing each partition to be compressed separately. In general, you should consider partitioning a table when it contains a large amount of data and queries frequently filter on a specific column or set of columns

upvoted 2 times

58d2382 2 years, 7 months ago

Question says "What is the minimum number of rows that Table1 should contain before you create (add/new/extra) partitions?" As per microsoft documentation, each partition will contain 1Million records. So, if there atleast 1million records, we can go for partitioning. Here is the link for documentation https://learn.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-overview?view=sql-server-ver16

upvoted 3 times

6d954df 2 years, 7 months ago

60m, see https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition

upvoted 2 times

blazy001 2 years, 8 months ago

A 100mil is correct >> From the answers to this Q, I see that MS has done a bad job because people don't understand what distributions or partitions are. My explanation: Each table with column store index is auto divided into 60 distributions, on each of these distributions there is auto 1 partition. For good performance (with column store) each partition must have at least 1Mil rows. The question was: "What is the minimum number of rows that

Table1 should contain before you create (add/new/extra) partitions?" So there is no point in creating partitions with 60M rows, because then you divide this into 0.5Mil per partition. At least 120Mil would be ideal, but 100Mil already starts.

upvoted 1 times

hassexat 2 years, 11 months ago

60 million

upvoted 2 times

kkk5566 2 years, 11 months ago

is correct

upvoted 2 times

Deeksha1234 3 years ago

should be D

upvoted 3 times

Question #79 *Topic 1*

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named DimSalesPerson. DimSalesPerson contains the following columns:

- RepSourceID
- SalesRepID
- FirstName
- LastName
- StartDate
- EndDate
- Region

You are developing an Azure Synapse Analytics pipeline that includes a mapping data flow named Dataflow1. Dataflow1 will read sales team data from an external source and use a Type 2 slowly changing dimension (SCD) when loading the data into DimSalesPerson.

You need to update the last name of a salesperson in DimSalesPerson.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Update three columns of an existing row.

B. Update two columns of an existing row.

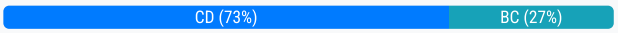
C. Insert an extra row.

D. Update one column of an existing row.

Suggested Answer: *C, D*

Official Answer: CD

Community vote distribution



bakamon Highly Voted 3 years, 2 months ago

Selected Answer: BC

1) Insert an extra row with the updated last name and the current date as the StartDate. 2) Update two columns of an existing row: set the EndDate of the previous row for that salesperson to the current date and set the current value of the SalesRepID column to inactive.

upvoted 23 times

bakamon Highly Voted 3 years, 2 months ago

1) Insert an extra row with the updated last name and the current date as the StartDate. 2) Update two columns of an existing row: set the EndDate of the previous row for that salesperson to the current date and set the current value of the SalesRepID column to inactive.SalesRepID is either business key or surrogate key, and you don't change it... there is no column in this table that acts as active flag, so you only need to change EndDate column.This will preserve the history of changes to the salesperson's last name while keeping the most current information in the tableExactly. SCD TYPE2 Needs three columns: StartDate, EndDate and isActive flag.active flag is optional for type 2 and in this scenario does not exist

upvoted 23 times

hypersam 1 year, 7 months ago

SalesRepID is either business key or surrogate key, and you don't change it... there is no column in this table that acts as active flag, so you only need to change EndDate column.

upvoted 2 times

bakamon 3 years, 2 months ago

This will preserve the history of changes to the salesperson's last name while keeping the most current information in the table

upvoted 2 times

bakamon 3 years, 2 months ago

This will preserve the history of changes to the salesperson's last name while keeping the most current information in the tableExactly. SCD TYPE2 Needs three columns: StartDate, EndDate and isActive flag.active flag is optional for type 2 and in this scenario does not exist

upvoted 2 times

robmz 2 years, 4 months ago

Exactly.

SCD TYPE2 Needs three columns: StartDate, EndDate and isActive flag.

upvoted 2 times

robmz 2 years, 4 months ago

Exactly. SCD TYPE2 Needs three columns: StartDate, EndDate and isActive flag.active flag is optional for type 2 and in this scenario does not exist

upvoted 2 times

Dusica 2 years, 3 months ago

active flag is optional for type 2 and in this scenario does not exist

upvoted 6 times

Ankit_Az 3 years, 2 months ago

CD is correct

upvoted 21 times

Stuartsen Most Recent 1 year, 5 months ago

Correct C, D

upvoted 1 times

samianae 1 year, 6 months ago

correct answer B and C

upvoted 1 times

EmnCours 1 year, 8 months ago

Correct Answer: CD

upvoted 1 times

Dusica 2 years, 3 months ago

C - extra row with new surname and new start date D-set end date in the existing column That makes it type 2 - one of the variations off

upvoted 2 times

s_unsworth 2 years, 5 months ago

I originally thought it would be BC, but then I realised you WOULDN'T update the Last Name, you want to retain the history of the LastName. You would update the EndDate column for that record to mark it as historic and then insert a new row with the new LastName.

upvoted 7 times

dakku987 2 years, 7 months ago

CD is correct as in scd2 we need startdate,enddate that is already present what we need is "ISActive/flag" and one more row thats all its takes to make scd2

upvoted 1 times

hassexat 2 years, 11 months ago

C & D are correct

upvoted 3 times

tankwayep 2 years, 11 months ago

- Update one column: EndDate to the change date - Insert a new record with the new value of LastName, StartDate as the change date.

upvoted 10 times

kkk5566 2 years, 11 months ago

correct

upvoted 2 times

lfss 3 years ago

cd is correct

upvoted 2 times

Deeksha1234 3 years ago

answer should be CD, since activeRow flag is not present, we need to update only end date.

upvoted 9 times

Rob77 3 years, 2 months ago

It's SCD Type 2 - you need to update at least three columns in the original raw: Surname, StartDate and EndDate. (IsActive if one exists). Then insert new record. A and C* "original row"but if you update the surname on the original row, don't you lose the previous value?No the StartDate stays, you only need to update the EndDate in the original row, the old name also stays to track which names he had, only the new row should have the new name. So you would only need to edit the EndDate column on the old row and since there is no "IsActive" flag you ignore it, maybe it's just queried by date and sorted by date and you take the last row which is the newest.

upvoted 1 times

Rob77 3 years, 2 months ago

It's SCD Type 2 - you need to update at least three columns in the original raw: Surname, StartDate and EndDate. (IsActive if one exists). Then insert new record. A and C

upvoted 1 times

Rob77 3 years, 2 months ago

* "original row"

upvoted 1 times

Rob77 3 years, 2 months ago

* "original row"but if you update the surname on the original row, don't you lose the previous value?

upvoted 1 times

peches 3 years, 2 months ago

but if you update the surname on the original row, don't you lose the previous value?

upvoted 7 times

Vanq69 2 years, 10 months ago

No the StartDate stays, you only need to update the EndDate in the original row, the old name also stays to track which names he had, only the new row should have the new name. So you would only need to edit the EndDate column on the old row and since there is no "IsActive" flag you ignore it, maybe it's just queried by date and sorted by date and you take the last row which is the newest.

upvoted 3 times

laurasscastro 3 years, 3 months ago

For me this is a little dubious since besides the end date update for the record we could have flg_is_active as well. Making B a possible answer in my opinion

upvoted 4 times

laurasscastro 3 years, 3 months ago

For me this is a little dubious since besides the end date update for the record we could have flg_is_active as well. Making B a possible answer in my opinionIt's saying "update on column of an EXISTING row". AKA you're just changing the IsCurrent part of the existing row, that's it.

upvoted 4 times

ajhak 3 years, 3 months ago

It's saying "update on column of an EXISTING row". AKA you're just changing the IsCurrent part of the existing row, that's it.

upvoted 1 times

henryphchan 3 years, 3 months ago

The answer is correct

upvoted 4 times

jeroenmouse 3 years, 3 months ago

SCD Type 2 will have historical changes hence we will have new row and we need to update the existing row's end date. Hence - CD
https://www.sqlshack.com/implementing-slowly-changing-dimensions-scds-in-data-warehouses/

upvoted 6 times

Question #80 *Topic 1*

HOTSPOT

-

You plan to use an Azure Data Lake Storage Gen2 account to implement a Data Lake development environment that meets the following requirements:

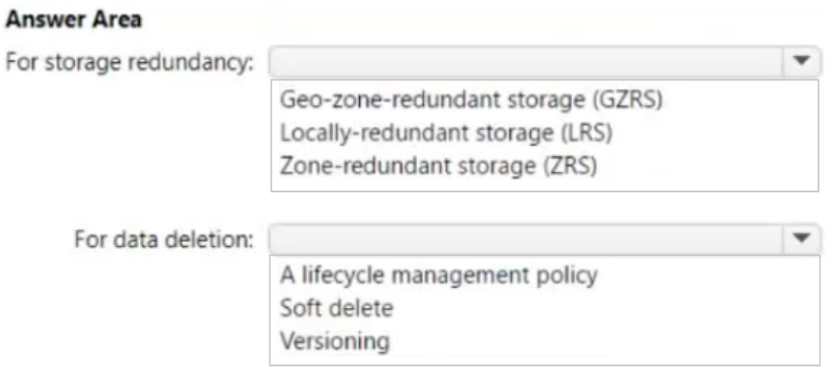
• Read and write access to data must be maintained if an availability zone becomes unavailable.

• Data that was last modified more than two years ago must be deleted automatically.

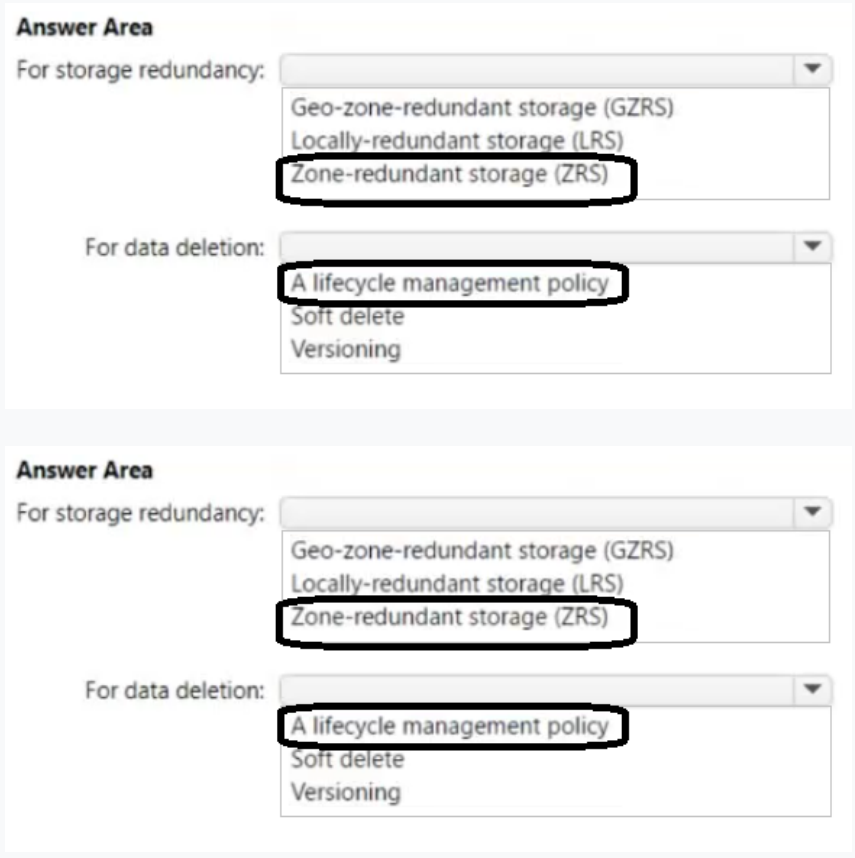
• Costs must be minimized.

What should you configure? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Suggested Answer:



bakamon Highly Voted 2 years, 8 months ago

Statement 1: For Storage redundancy, you should select ZRS (Zone-redundant storage). This will maintain read and write access to data even if an availability zone becomes unavailable. Statement 2: For data deletion, you should select A lifecycle management policy. This will allow you to automatically delete data that was last modified more than two years ago

upvoted 24 times

henryphchan Highly Voted 2 years, 9 months ago

Zone-redundant storage (ZRS) synchronously replicates your Azure managed disk across three Azure availability zones in the region you select. Each availability zone is a separate physical location with independent power, cooling, and networking

upvoted 6 times

Alongi Most Recent 1 year, 10 months ago

Correct guys

upvoted 1 times

Tusharsahoo 1 year, 12 months ago

correct

upvoted 1 times

AvSUN 2 years, 5 months ago

Correct

upvoted 1 times

kkk5566 2 years, 5 months ago

correct

upvoted 1 times

Deeksha1234 2 years, 6 months ago

correct answer

upvoted 2 times

Rajan191083 2 years, 9 months ago

Confusion is how the write access will be maintained in case primary zone failure?

upvoted 1 times

Rajan191083 2 years, 9 months ago

Confusion is how the write access will be maintained in case primary zone failure?With ZRS, your data is still accessible for both read and write operations even if a zone becomes unavailable. https://learn.microsoft.com/en-us/azure/storage/common/storage-redundancy

upvoted 1 times

Debasish93 2 years, 8 months ago

With ZRS, your data is still accessible for both read and write operations even if a zone becomes unavailable. https://learn.microsoft.com/en-us/azure/storage/common/storage-redundancy

upvoted 2 times

rocky48 2 years, 9 months ago

Zone-redundant storage (ZRS) & Lifecycle Policy

upvoted 4 times

makkelijkzat 2 years, 9 months ago

correct

upvoted 2 times

dksks 2 years, 9 months ago

correct

upvoted 1 times

HOTSPOT

-

You are designing an Azure Data Lake Storage Gen2 container to store data for the human resources (HR) department and the operations department at your company.

You have the following data access requirements:

• After initial processing, the HR department data will be retained for seven years and rarely accessed.

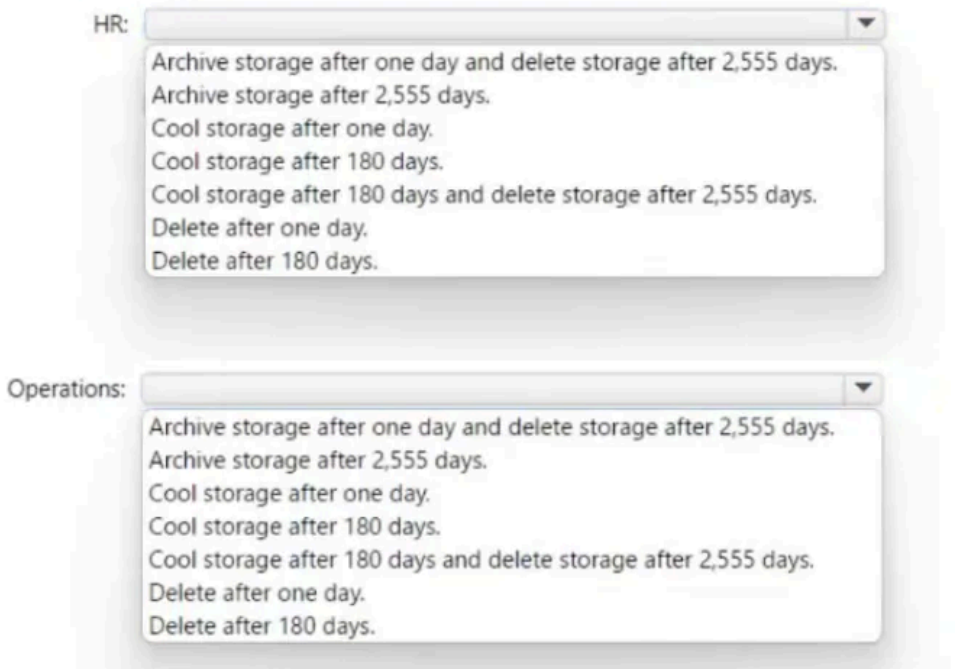
• The operations department data will be accessed frequently for the first six months, and then accessed once per month.

You need to design a data retention solution to meet the access requirements. The solution must minimize storage costs.

What should you include in the storage policy for each department? To answer, select the appropriate options in the answer area.

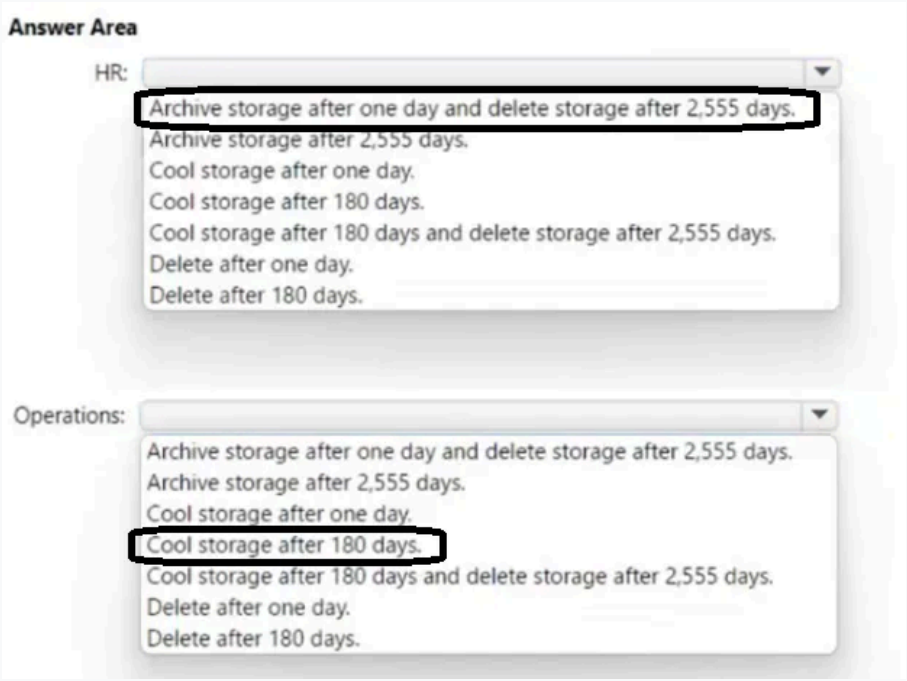
NOTE: Each correct selection is worth one point.

Answer Area

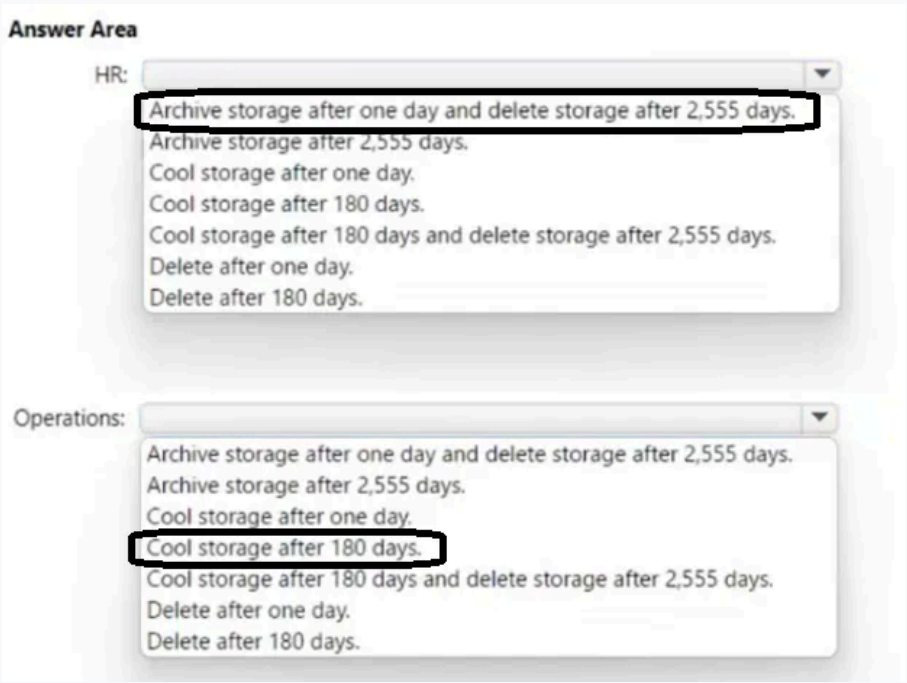


Suggested Answer:

Answer Area



Answer Area



dksks Highly Voted 2 years, 9 months ago

correct

upvoted 13 times

Alongi Most Recent 1 year, 10 months ago

Correct

upvoted 1 times

Charley92 2 years ago

After initial processing, the HR department data will be retained for seven years and rarely accessed That means 365*7 = 2,555 The operations department data will be accessed frequently for the first six months, and then accessed once per month. accessed frequently for the first six months - Cool 30*6 =180

upvoted 3 times

AvSUN 2 years, 5 months ago

I had to reread the question but the answer is correct, it would have been better if they mentioned what "rarely" means. Issue will arise if data needs to be accessed within 180 days of moving to archive.

upvoted 1 times

kkk5566 2 years, 5 months ago

correct

upvoted 2 times

Ankit_Az 2 years, 8 months ago

Correct

upvoted 2 times

Rob77 2 years, 8 months ago

You can't access data that was archived without rehydration. Rehydration requires either amending blob tier to hot or cold and is likely to incur a fee if stored less than 180 day or copying blob to another location... therefore "rarely" is unlikely a good option...

upvoted 1 times

rocky48 2 years, 9 months ago

Answer is correct

upvoted 3 times

rocky48 2 years, 9 months ago

Answer is correctAfter initial processing, the HR department data will be retained for seven years and rarely accessed That means 365*7 = 2,555 The operations department data will be accessed frequently for the first six months, and then accessed once per month. accessed frequently for the first six months - Cool 30*6 =180

upvoted 3 times

rocky48 2 years ago

After initial processing, the HR department data will be retained for seven years and rarely accessed That means 365*7 = 2,555 The operations department data will be accessed frequently for the first six months, and then accessed once per month. accessed frequently for the first six months - Cool 30*6 =180

upvoted 1 times

henryphchan 2 years, 9 months ago

the answer is correct

upvoted 4 times

shakes103 2 years, 9 months ago

Answer is correct

upvoted 4 times

Question #82 *Topic 1*

HOTSPOT

-

You are developing an Azure Synapse Analytics pipeline that will include a mapping data flow named Dataflow1. Dataflow1 will read customer data from an external source and use a Type 1 slowly changing dimension (SCD) when loading the data into a table named DimCustomer in an Azure Synapse Analytics dedicated SQL pool.

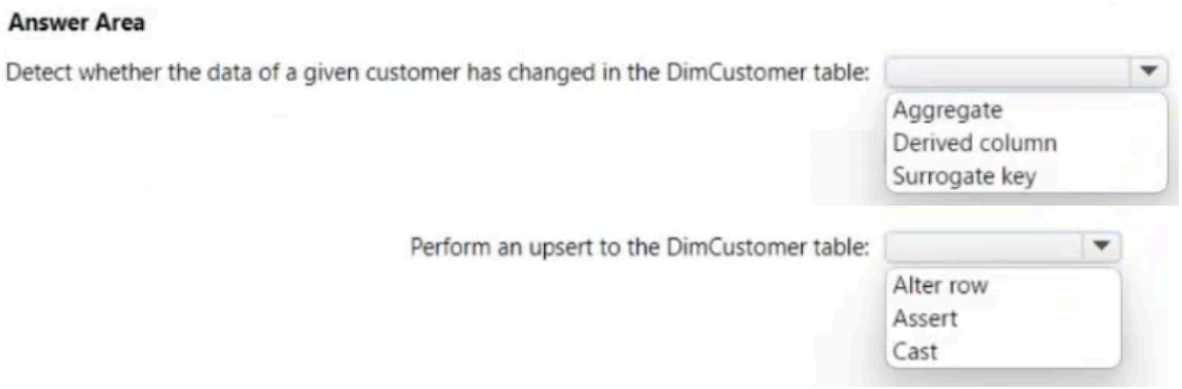
You need to ensure that Dataflow1 can perform the following tasks:

• Detect whether the data of a given customer has changed in the DimCustomer table.

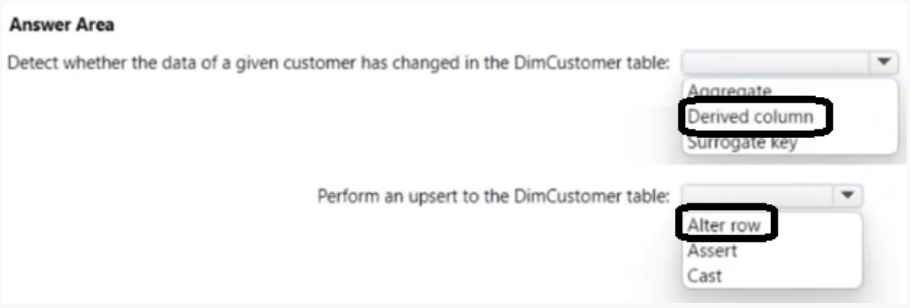
• Perform an upsert to the DimCustomer table.

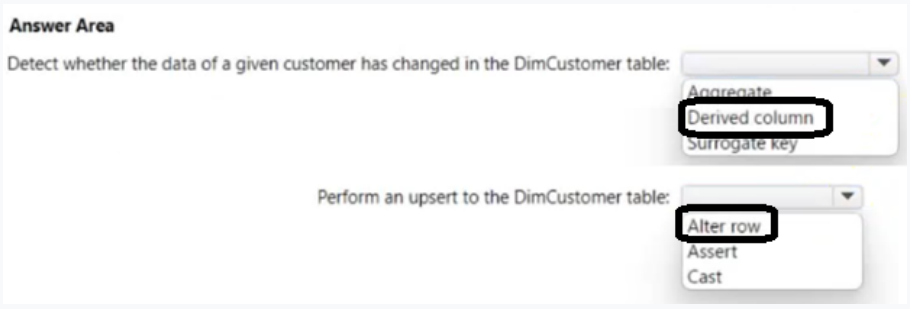
Which type of transformation should you use for each task? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.



Suggested Answer:





aemilka Highly Voted 2 years, 9 months ago

The answer is correct. Check "Exercise - Design and implement a Type 1 slowly changing dimension with mapping data flows", there is described implementation of the dataflow mentioned in this question. https://learn.microsoft.com/en-us/training/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/4-exercise-design-implement-type-1-dimension In the exercise 'Derived column' transformation is used to add InsertedDate and ModifiedDate columns. ModifiedDate column can be used to detect whether the customer data has changed. For Upsert 'Alter row' tranformation is used. The answer is definitely correct.

upvoted 29 times

Alongi Highly Voted 1 year, 9 months ago

I found this question on my exam 30/04/2024, and I put: - derived column - alter row I passed the exam with a high score, but I'm not sure if the answer is correct.

upvoted 8 times

moneytime 2 years ago

Aggregate and alter row. Aggregate will give table statistics which reveals information on max.minimum,sum ,average etc. While alter row encompasses ,update & insert which are upset operations

upvoted 3 times

kkk5566 2 years, 5 months ago

'Derived column' &'Alter row'

upvoted 1 times

Deeksha1234 2 years, 6 months ago

answer is correct

upvoted 1 times

bakamon 2 years, 8 months ago

Answer is correct.. 1) we don't need a surrogate key in SCD type 1. you can use a Derived Column transformation to compare the incoming data with the existing data in the DimCustomer table and detect changes.. Statement 2: To perform an upsert to the DimCustomer table, you should use an Alter Row transformation. This transformation can be used to specify the actions to take for each row of data, such as inserting new rows or updating existing rows. The current web page context is empty.

upvoted 7 times

rocky48 2 years, 9 months ago

'Derived column' & 'Alter row'

upvoted 2 times

haythemsi 2 years, 9 months ago

surrogate key and assert

upvoted 2 times

nicololmen 2 years, 9 months ago

surrogate key and alter row according to chatgpt

upvoted 2 times

Question #83 *Topic 1*

DRAG DROP

-

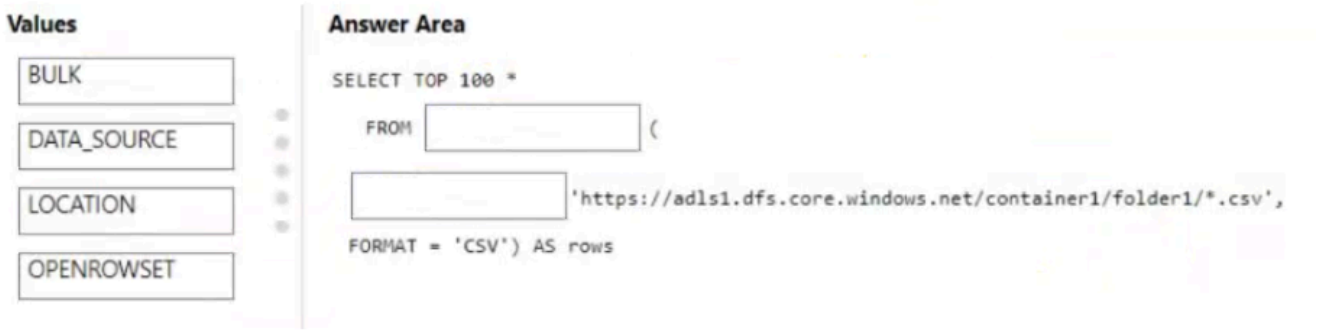
You have an Azure Synapse Analytics serverless SQL pool.

You have an Azure Data Lake Storage account named adls1 that contains a public container named container1. The container1 container contains a folder named folder1.

You need to query the top 100 rows of all the CSV files in folder1.

How should you complete the query? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point



Suggested Answer:



rocky48 Highly Voted 2 years, 9 months ago

The provided query is correct for Azure Synapse Analytics serverless SQL pool. It selects the top 100 rows from the data in CSV format located at the specified URL: https://adls1.dfs.core.windows.net/container1/folde1/*.csv. The results are returned under the alias rows. Answer is correct.

upvoted 8 times

hassexat Highly Voted 2 years, 5 months ago

OPENROWSET & Bulk

upvoted 6 times

quatrz Most Recent 1 year, 5 months ago

"the container 1 container contains"

upvoted 2 times

Alongi 1 year, 9 months ago

I found this question on my exam 30/04/2024, and I put - openrowset - bulk I passed the exam with a high score, but I'm not sure if the answer is correct.

upvoted 3 times

Alongi 1 year, 10 months ago

Correct

upvoted 1 times

kkk5566 2 years, 5 months ago

openrowset..bulk

upvoted 2 times

Deeksha1234 2 years, 6 months ago

ans is correct

upvoted 2 times

Ankit_Az 2 years, 8 months ago

Correct

upvoted 2 times

henryphchan 2 years, 9 months ago

The answer is correct

upvoted 3 times

shakes103 2 years, 9 months ago

Answer is correct

upvoted 3 times

Question #84 *Topic 1*

You have an Azure Synapse Analytics workspace named WS1 that contains an Apache Spark pool named Pool1.
You plan to create a database named DB1 in Pool1.
You need to ensure that when tables are created in DB1, the tables are available automatically as external tables to the built-in serverless SQL pool.
Which format should you use for the tables in DB1?

A. Parquet

B. ORC

C. JSON

D. HIVE

Suggested Answer: *A*
Official Answer: A

Community vote distribution
A (100%)

kam1122 Highly Voted 2 years, 8 months ago
always pick parquet first
upvoted 5 times

EmnCours Most Recent 1 year, 8 months ago
Selected Answer: A
upvoted 1 times

kkk5566 2 years, 11 months ago
is correct
upvoted 1 times

Deeksha1234 3 years ago
parquet. CSV , delta also possible but not an option here.
upvoted 2 times

akhil5432 3 years ago
parquet
upvoted 1 times

Ankit_Az 3 years, 2 months ago
Correct
upvoted 1 times

rocky48 3 years, 3 months ago
Parquet is the correct answer
upvoted 1 times

aemilka 3 years, 3 months ago
https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop Supported formats for serverless pool: Delimited/CSV, Parquet, Delta Lake So Parquet is the correct answer
upvoted 2 times

henryphchan 3 years, 3 months ago
Parquet is supported by serverless SQL pool https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-parquet-files
upvoted 2 times

makkelijkzat 3 years, 3 months ago
correct
upvoted 1 times

Question #85 *Topic 1*

You have an Azure Data Lake Storage Gen2 account named storage1.
You plan to implement query acceleration for storage1.
Which two file types support query acceleration? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

A. JSON

B. Apache Parquet

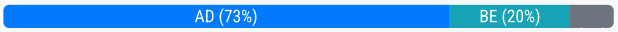
C. XML

D. CSV

E. Avro

Suggested Answer: *A, D*
Official Answer: AD

Community vote distribution



orionduo Highly Voted 3 years, 1 month ago
Correct. Query acceleration supports CSV and JSON formatted data as input to each request. https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-query-acceleration
upvoted 11 times

vctrhugo Highly Voted 3 years, 2 months ago
Query acceleration supports CSV and JSON formatted data as input to each request.
upvoted 6 times

sachin_mt Most Recent 1 year, 5 months ago
B & E options
upvoted 1 times

Moli62 1 year, 5 months ago
https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-query-acceleration
upvoted 1 times

jxs221 1 year, 5 months ago
Parquet is a columnar format, allowing efficient data retrieval and compression. Avro is optimized for efficient binary serialization, making it suitable for query acceleration.
upvoted 1 times

angelvenkovicch 1 year, 5 months ago
is this related to DP-300 or not
upvoted 1 times

samirarian 1 year, 6 months ago
These file formats are optimized for analytics and support features like columnar storage, which is beneficial for query acceleration, making them more efficient for querying in big data analytics workloads. Parquet: A columnar storage format that is widely used for analytics, providing fast query performance. Avro: A row-based storage format that also supports efficient querying, particularly with its schema support and compact data format. So, Parquet and Avro are the correct answers.
upvoted 1 times

EmnCours 1 year, 8 months ago
Correct Answer: AD
upvoted 1 times

MohamedBI12 2 years, 5 months ago
A and D
upvoted 1 times

moize 2 years, 7 months ago
https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-query-acceleration

upvoted 3 times

OldSchool 2 years, 11 months ago

Query acceleration supports CSV and JSON formatted data as input to each request.

upvoted 1 times

kkk5566 2 years, 11 months ago

CSV and JSON

upvoted 1 times

aga444 3 years, 2 months ago

Parquet and CSV

upvoted 1 times

IanKwok81 3 years, 2 months ago

Correct. https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-query-acceleration

upvoted 4 times

Question #86 *Topic 1*

You have an Azure subscription that contains the resources shown in the following table.
You need to read the files in storage1 by using ad-hoc queries and the OPENROWSET function. The solution must ensure that each rowset contains a single JSON record.
To what should you set the FORMAT option of the OPENROWSET function?

Name	Type	Description
storage1	Azure Blob storage account	Contains publicly accessible JSON files
WS1	Azure Synapse Analytics workspace	Contains a serverless SQL pool

A. JSON

B. DELTA

C. PARQUET

D. CSV

Suggested Answer: *A*
Official Answer: A

Community vote distribution
D (90%) A (10%)

Zak_Zakaria Highly Voted 3 years, 1 month ago
It should be D normally. I'll appreciate it if the one who curates the right answer and when the majority doesn't agree with his choice, brings some explanation so we can discuss and understand why he chooses it. It's not the first time I see a total disagreement without any explanation from the ones who select the right answers.To return product data from a folder containing multiple JSON files in this format, you could use the following SQL query: SQL Copy SELECT doc FROM OPENROWSET(BULK 'https://mydatalake.blob.core.windows.net/data/files/*.json', FORMAT = 'csv', FIELDTERMINATOR ='0x0b', FIELDQUOTE = '0x0b', ROWTERMINATOR = '0x0b') WITH (doc NVARCHAR(MAX)) as rows OPENROWSET has no specific format for JSON files, so you must use csv format with FIELDTERMINATOR, FIELDQUOTE, and ROWTERMINATOR set to 0x0b, and a schema that includes a single NVARCHAR(MAX) column. The result of this query is a rowset containing a single column of JSON documents, like this:
upvoted 13 times

Zak_Zakaria Highly Voted 3 years, 1 month ago
Selected Answer: D
It should be D normally. I'll appreciate it if the one who curates the right answer and when the majority doesn't agree with his choice, brings some explanation so we can discuss and understand why he chooses it. It's not the first time I see a total disagreement without any explanation from the ones who select the right answers.
upvoted 13 times

humma 1 year, 9 months ago
To return product data from a folder containing multiple JSON files in this format, you could use the following SQL query: SQL Copy SELECT doc FROM OPENROWSET(BULK 'https://mydatalake.blob.core.windows.net/data/files/*.json', FORMAT = 'csv', FIELDTERMINATOR ='0x0b', FIELDQUOTE = '0x0b', ROWTERMINATOR = '0x0b') WITH (doc NVARCHAR(MAX)) as rows OPENROWSET has no specific format for JSON files, so you must use csv format with FIELDTERMINATOR, FIELDQUOTE, and ROWTERMINATOR set to 0x0b, and a schema that includes a single NVARCHAR(MAX) column. The result of this query is a rowset containing a single column of JSON documents, like this:
upvoted 3 times

phydev 2 years, 9 months ago
Was on my exam today (31.10.2023).
upvoted 12 times

JT16 Most Recent 1 year, 4 months ago
Since the files in storage1 are publicly accessible JSON files, setting the FORMAT option to JSON allows Azure Synapse Analytics to correctly parse and process each file as individual JSON records. This ensures that each rowset corresponds to a single JSON record, aligning with the goal of ad-hoc queries via OPENROWSET.
upvoted 1 times

Moli62 1 year, 5 months ago

https://learn.microsoft.com/en-us/training/modules/query-data-lake-using-azure-synapse-serverless-sql-pools/3-query-files

upvoted 1 times

samirarian 1 year, 6 months ago

SELECT * FROM OPENROWSET(BULK 'https://mydatalake.blob.core.windows.net/data/files/*.json', FORMAT = 'JSON') AS doc

upvoted 1 times

moize 1 year, 8 months ago

Selected Answer: A

Pour garantir que chaque ensemble de lignes contient un seul enregistrement JSON lorsque vous utilisez la fonction OPENROWSET pour lire des fichiers dans Azure Data Lake Storage Gen2, vous devez définir l'option FORMAT sur A. JSON.

L'option D. CSV ne serait pas correcte dans ce contexte, car elle est utilisée pour lire des fichiers CSV, et non des fichiers JSON. Utiliser FORMAT = 'JSON' permet de s'assurer que chaque ligne de la sortie correspond à un enregistrement JSON unique. Voici un exemple de requête : SELECT * FROM OPENROWSET(BULK 'https://storage1.dfs.core.windows.net/container1/folder1/*.json', FORMAT = 'JSON') AS rows;

upvoted 2 times

moize 1 year, 8 months ago

Pour garantir que chaque ensemble de lignes contient un seul enregistrement JSON lorsque vous utilisez la fonction OPENROWSET pour lire des fichiers dans Azure Data Lake Storage Gen2, vous devez définir l'option FORMAT sur A. JSON. L'option D. CSV ne serait pas correcte dans ce contexte, car elle est utilisée pour lire des fichiers CSV, et non des fichiers JSON. Utiliser FORMAT = 'JSON' permet de s'assurer que chaque ligne de la sortie correspond à un enregistrement JSON unique. Voici un exemple de requête : SELECT * FROM OPENROWSET(BULK 'https://storage1.dfs.core.windows.net/container1/folder1/*.json', FORMAT = 'JSON') AS rows;This is wrong, just try it using serverless sql pool and it gives error: Invalid or unknown format type 'JSON'.

upvoted 2 times

hypersam 1 year, 7 months ago

This is wrong, just try it using serverless sql pool and it gives error: Invalid or unknown format type 'JSON'.

upvoted 2 times

EmnCours 1 year, 8 months ago

Correct Answer: D

upvoted 1 times

ahana1074 1 year, 11 months ago

This allows the SQL pool to treat each JSON object as a row and parse the file accordingly.

The correct answer is: A. JSON

upvoted 2 times

ahana1074 1 year, 11 months ago

This allows the SQL pool to treat each JSON object as a row and parse the file accordingly. The correct answer is: A. JSONOPENROWSET has no specific format for JSON files, so you must use csv format with FIELDTERMINATOR, FIELDQUOTE, and ROWTERMINATOR set to 0x0b, and a schema that includes a single NVARCHAR(MAX) column. The result of this query is a rowset containing a single column of JSON documents, like this: https://learn.microsoft.com/en-us/training/modules/query-data-lake-using-azure-synapse-serverless-sql-pools/3-query-files

upvoted 2 times

humma 1 year, 9 months ago

OPENROWSET has no specific format for JSON files, so you must use csv format with FIELDTERMINATOR, FIELDQUOTE, and ROWTERMINATOR set to 0x0b, and a schema that includes a single NVARCHAR(MAX) column. The result of this query is a rowset containing a single column of JSON documents, like this: https://learn.microsoft.com/en-us/training/modules/query-data-lake-using-azure-synapse-serverless-sql-pools/3-query-files

upvoted 2 times

AAhmadH 2 years ago

A OPENROWSET support JSON format in Synapse. In traditional SQL Server on-prem JSON format wasn't supported so we used CSV method. Synapse support JSON OPENROWSET, json object is treated as separate row, no need for CSV workaround, it would complicate the process requiring additional parsing record

upvoted 2 times

vik_sahani 2 years ago

A. OPENROWSET(BULK) is a table-valued function that can read data from any file on the local drive or network, if SQL Server has read access to that location. It returns a table with a single column that contains the contents of the file. Eg.. SELECT BulkColumn FROM OPENROWSET(BULK 'C:\JSON\Books\book.json', SINGLE_CLOB) as j;

upvoted 1 times

Prt1nov89 2 years, 1 month ago

You have three choices for input files that contain the target data for querying. Valid values are: 'CSV' - Includes any delimited text file with row/column separators. Any character can be used as a field separator, such as TSV: FIELDTERMINATOR = tab. 'PARQUET' - Binary file in Parquet format 'DELTA' - A set of Parquet files organized in Delta Lake (preview) format Values with blank spaces are not valid, e.g. 'CSV ' is not a valid value. So, its D

upvoted 1 times

ageorgieva 2 years, 1 month ago

Gemini says it is a CSV: Based on the information provided in the image, the format option of the OPENROWSET function should be set to D. CSV. Here's why: The data files in Azure Blob storage are JSON files. However, SQL Server doesn't natively support the JSON format for the OPENROWSET function. As a workaround, you can specify the format option as CSV and configure the field terminator and fieldquote options appropriately to process each line of the JSON file as a single record. Even though the data is in JSON format, choosing CSV as the format option allows you to read the data into a SQL table using OPENROWSET.

upvoted 5 times

qadeeros 2 years, 3 months ago

chatgpt says option A

upvoted 3 times

Elanche 2 years, 4 months ago

The easiest way to see to the content of your JSON file is to provide the file URL to the OPENROWSET function, specify csv FORMAT, and set values 0x0b for fieldterminator and fieldquote. If you need to read line-delimited JSON files, then this is enough.

upvoted 1 times

Elanche 2 years, 4 months ago

select top 10 * from openrowset(bulk 'https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/ecdc_cases/latest/ecdc_cases.jsonl', format = 'csv', fieldterminator ='0x0b', fieldquote = '0x0b') with (doc nvarchar(max)) as rows go select top 10 * from openrowset(bulk 'https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/ecdc_cases/latest/ecdc_cases.json', format = 'csv', fieldterminator ='0x0b', fieldquote = '0x0b', rowterminator = '0x0b' --> You need to override rowterminator to read classic JSON) with (doc nvarchar(max)) as rows

upvoted 1 times

s_unsworth 2 years, 5 months ago

OPENROWSET doesn't have a JSON option

upvoted 1 times

falzar 2 years, 6 months ago

I checked this on another dump site and chatgpt. Both said the answer is A.

upvoted 2 times

Question #87 *Topic 1*

HOTSPOT

-

You have an Azure subscription that contains the Azure Synapse Analytics workspaces shown in the following table.

Each workspace must read and write data to datalake1.

Each workspace contains an unused Apache Spark pool.

You plan to configure each Spark pool to share catalog objects that reference datalake1.

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Name	Primary storage account
workspace1	datalake1
workspace2	datalake2
workspace3	datalake1

Answer Area

Statements	Yes	No
The shared catalog objects can be stored in Azure Database for MySQL.	○	○
For the Apache Hive Metastore of each workspace, you must configure a linked service that uses user-password authentication.	○	○
The users of workspace1 must be assigned the Storage Blob Contributor role for datalake1.	○	○

Suggested Answer:

Answer Area

Statements	Yes	No
The shared catalog objects can be stored in Azure Database for MySQL.	◉	○
For the Apache Hive Metastore of each workspace, you must configure a linked service that uses user-password authentication.	◉	○
The users of workspace1 must be assigned the Storage Blob Contributor role for datalake1.	○	◉

Answer Area

Statements	Yes	No
The shared catalog objects can be stored in Azure Database for MySQL.	◉	○
For the Apache Hive Metastore of each workspace, you must configure a linked service that uses user-password authentication.	◉	○
The users of workspace1 must be assigned the Storage Blob Contributor role for datalake1.	○	◉

auwia Highly Voted 3 years, 1 month ago

Provided answers are correct: 1. Yes: Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog. When customers want to persist the Hive catalog metadata outside of the workspace, and share catalog objects with other computational engines outside of the workspace, such as HDInsight and Azure Databricks, they can connect to an external Hive Metastore. Only Azure SQL Database and Azure Database for MySQL are supported as an external Hive Metastore. 2. Yes: And currently we only support User-Password authentication. 3. No: And currently we only support User-Password authentication. ==> STORAGE BLOB CONTRIBUTOR is an

Azure RBAC (Role-Based Access Control) ==> NOT COMPATIBLE (it is supported User-Password authentication ONLY). ref. https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore

upvoted 18 times

aga444 Highly Voted 3 years, 2 months ago

No-Yes-Yes

upvoted 12 times

aga444 3 years, 2 months ago

No-Yes-YesI confirm the first No: The first statement, "The shared catalog objects can be stored in Azure Database for MySQL," is not true because Azure Database for MySQL is not the appropriate storage option for shared catalog objects in Azure Synapse Analytics. The shared catalog objects, which include metadata and schema information, are typically stored in a centralized metadata store such as the Apache Hive Metastore. Azure Synapse Analytics supports using an Azure SQL Database or an Azure SQL Data Warehouse (now called Azure Synapse SQL) as the metadata store, but Azure Database for MySQL is not a supported option for this purpose.Sorry I was wrong.

upvoted 12 times

auwia 3 years, 1 month ago

I confirm the first No: The first statement, "The shared catalog objects can be stored in Azure Database for MySQL," is not true because Azure Database for MySQL is not the appropriate storage option for shared catalog objects in Azure Synapse Analytics. The shared catalog objects, which include metadata and schema information, are typically stored in a centralized metadata store such as the Apache Hive Metastore. Azure Synapse Analytics supports using an Azure SQL Database or an Azure SQL Data Warehouse (now called Azure Synapse SQL) as the metadata store, but Azure Database for MySQL is not a supported option for this purpose.

upvoted 3 times

auwia 3 years, 1 month ago

I confirm the first No: The first statement, "The shared catalog objects can be stored in Azure Database for MySQL," is not true because Azure Database for MySQL is not the appropriate storage option for shared catalog objects in Azure Synapse Analytics. The shared catalog objects, which include metadata and schema information, are typically stored in a centralized metadata store such as the Apache Hive Metastore. Azure Synapse Analytics supports using an Azure SQL Database or an Azure SQL Data Warehouse (now called Azure Synapse SQL) as the metadata store, but Azure Database for MySQL is not a supported option for this purpose.Sorry I was wrong.

upvoted 3 times

auwia 3 years, 1 month ago

Sorry I was wrong.

upvoted 10 times

seranvijay Most Recent 1 year, 9 months ago

3.No If the underlying data of your Hive tables is stored in the workspace primary storage account, you don't need to do extra settings. It will just work as long as you followed storage setting up instructions during workspace creation. https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore#workspace-primary-storage-account

upvoted 1 times

rocky48 2 years, 6 months ago

Yes, Yes , No 1. Yes: Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog. When customers want to persist the Hive catalog metadata outside of the workspace, and share catalog objects with other computational engines outside of the workspace, such as HDInsight and Azure Databricks, they can connect to an external Hive Metastore. Only Azure SQL Database and Azure Database for MySQL are supported as an external Hive Metastore. 2. Yes: And currently we only support User-Password authentication. 3. No: And currently we only support User-Password authentication. ==> STORAGE BLOB CONTRIBUTOR is an Azure RBAC (Role-Based Access Control) ==> NOT COMPATIBLE (it is supported User-Password authentication ONLY). ref. https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore

upvoted 2 times

wel_fardeheb 2 years, 6 months ago

I got this question yesterday on my exam

upvoted 4 times

Bill_Walker 2 years, 6 months ago

Looks like the Storage Blob Data Contributor role does not need to be assigned. Workspace primary storage account If the underlying data of your Hive tables is stored in the workspace primary storage account, you don't need to do extra settings. It will just work as long as you followed storage setting up instructions during workspace creation. Other ADLS Gen 2 account If the underlying data of your Hive catalogs is stored in another ADLS Gen 2 account, you need to make sure the users who run Spark queries have Storage Blob Data Contributor role on the ADLS Gen2 storage account.

upvoted 1 times

ELJORDAN23 2 years, 7 months ago

Got this question on my exam on january 17, answers are correct.

upvoted 5 times

ExamDestroyer69 2 years, 8 months ago

Confusing discussion section, no consensus

upvoted 4 times

kkk5566 2 years, 11 months ago

correct

upvoted 1 times

kkk5566 2 years, 11 months ago

correctShouldn't WS1 have blob data contributor access?

upvoted 1 times

ahmadsayeed 2 years, 9 months ago

Shouldn't WS1 have blob data contributor access?

upvoted 1 times

Deeksha1234 3 years ago

given answer is correct https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore

upvoted 2 times

pavankr 3 years, 1 month ago

Correct order should be Yes, No, Yes

upvoted 2 times

DataSaM 3 years, 1 month ago

https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore

upvoted 1 times

Albeeliu 3 years, 1 month ago

What are the correct answers???

upvoted 1 times

Paulkuzzio 3 years, 1 month ago

Only Azure SQL Database and Azure Database for MySQL are supported as an external Hive Metastore. And currently we only support User-Password authentication. https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore

upvoted 6 times

sridat 3 years, 2 months ago

A "Blob Data Contributor" role must be assigned to user in order to access the files in blob storage. So it's a "Yes" https://learn.microsoft.com/en-us/azure/storage/blobs/assign-azure-role-data-access?tabs=portal

upvoted 6 times

Question #88 *Topic 1*

DRAG DROP

-

You have a data warehouse.

You need to implement a slowly changing dimension (SCD) named Product that will include three columns named ProductName, ProductColor, and ProductSize. The solution must meet the following requirements:

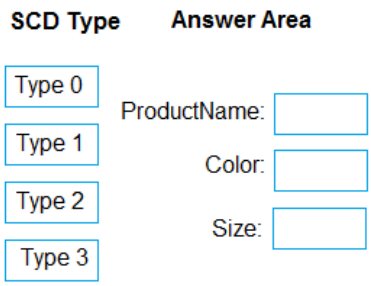
• Prevent changes to the values stored in ProductName.

• Retain only the current and the last values in ProductSize.

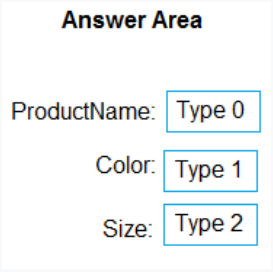
• Retain all the current and previous values in ProductColor.

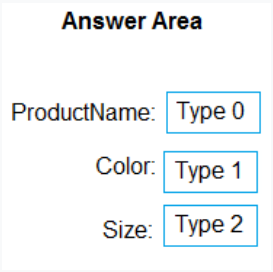
Which type of SCD should you implement for each column? To answer, drag the appropriate types to the correct columns. Each type may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.



Suggested Answer:





Ram0202 Highly Voted 3 years, 2 months ago

Product name -type 0 color -type 2 size -type 3

upvoted 116 times

andjurovicela Highly Voted 3 years, 2 months ago

ProductName - type 0, as no changes are done. Color - type 3, as with type 3 we have one column for the current value and one for the previous so only these two are preserved. Size - type 2, as it inserts a new row for every change, so we get all historical values.

upvoted 41 times

andjurovicela 3 years, 2 months ago

ProductName - type 0, as no changes are done. Color - type 3, as with type 3 we have one column for the current value and one for the previous so only these two are preserved. Size - type 2, as it inserts a new row for every change, so we get all historical values.Agree. beware that the order of ProductSize, ProductColor in the question, not same as in the graph. Product name -type 0 color -type 3 size -type 2

upvoted 41 times

hiyoww 3 years, 1 month ago

Agree. beware that the order of ProductSize, ProductColor in the question, not same as in the graph. Product name -type 0 color -type 3 size -type 2

upvoted 21 times

samirarian Most Recent 1 year, 6 months ago

Column Requirement SCD Type ProductName Prevent changes (No updates allowed) SCD Type 0 ProductSize Retain only the current and last value SCD Type 3 ProductColor Retain all historical values SCD Type 2

upvoted 2 times

Piantoni 1 year, 9 months ago

ProductName: Type 0 Color: Type 2 Size: Type3

upvoted 1 times

j888 2 years, 6 months ago

Type 1: Overwrite the existing data with new data, without maintaining a history. This is suitable when you want to prevent changes to the values stored in a column. Type 2: Keep a history of changes by adding new rows for each change. This is suitable for retaining all the current and previous values in a column. Type 3: Keep a limited history, typically by adding new columns to the dimension table to store both the current and previous values. This allows you to track changes but in a more compact way compared to Type 2. Therefore I believe the answer is 0 3 2I made a typo. My answer should be as follow ProductName: 0 Color: 2 Size: 3

upvoted 2 times

j888 2 years, 6 months ago

Type 1: Overwrite the existing data with new data, without maintaining a history. This is suitable when you want to prevent changes to the values stored in a column.

Type 2: Keep a history of changes by adding new rows for each change. This is suitable for retaining all the current and previous values in a column.

Type 3: Keep a limited history, typically by adding new columns to the dimension table to store both the current and previous values. This allows you to track changes but in a more compact way compared to Type 2. Therefore I believe the answer is 0 3 2

upvoted 2 times

j888 2 years, 6 months ago

I made a typo. My answer should be as follow ProductName: 0 Color: 2 Size: 3

upvoted 5 times

rocky48 2 years, 6 months ago

ProductName: To prevent changes to the values stored in ProductName, we should use Type 0. This means that the value remains static and does not change over time. ProductColor: Since we need to retain all the current and previous values in ProductColor, we should use Type 2. This type allows us to track historical changes by creating new rows for each change and maintaining a history of values. ProductSize: To retain only the current and last values in ProductSize, we should use Type 3. This type keeps track of the current value and the previous value in separate columns. Therefore, the correct SCD types for the three columns are as follows: ProductName: Type 0 ProductColor: Type 2 ProductSize: Type 3

upvoted 5 times

moneytime 2 years, 6 months ago

Product : Type 0 :No change (This eliminates possibilities of Type 1,2 and other higher dimensions Colour : Type 1 or 2 ;There is a change which can be rowise(type 1) or column wise(Type 2) Size :Type 3 ; All changes means both rowise and column wise

upvoted 1 times

Momoanwar 2 years, 8 months ago

For the given requirements: - **Product Name**: Since changes must be prevented, this would be a Type 0 SCD, as it maintains the original value without any changes. - **Product Size**: To retain only the current and the last values, you would use a Type 3 SCD, which keeps the original value and adds a new column for the current value. - **Product Color**: To retain all the current and previous values, a Type 2 SCD is used, as it tracks historical data by creating a new record for each change. So, you would apply: - Type 0 for ProductName - Type 3 for ProductSize - Type 2 for ProductColor

upvoted 6 times

Vanq69 2 years, 10 months ago

Name: 0 Fixed Dimension Color: 2 Row Versioning (current + last value) Size: 3 Previous Value column (current value + all previous values in extra column)I meant Name: 0 Color: 3 Size: 2

upvoted 5 times

Vanq69 2 years, 10 months ago

Name: 0 Fixed Dimension Color: 2 Row Versioning (current + last value) Size: 3 Previous Value column (current value + all previous values in extra column)

upvoted 5 times

Vanq69 2 years, 10 months ago

I meant Name: 0 Color: 3 Size: 2

upvoted 5 times

kkk5566 2 years, 11 months ago

SCD0-1-2

upvoted 3 times

hassexat 2 years, 11 months ago

Product Name - Type 0 Color - Type 2 Size - Type 3

upvoted 2 times

AvSUN 2 years, 11 months ago

The answers are 0, 3, 2ProductName - 0 ProductColor - 3 ProductSize - 2The requirements and blanks are out of order in the questionProductColor cannot be Type3 since it indicates that it has to preserve all current and previous values so Type3 It only preserves the current and previous value without maintaining a complete history. For me it is: ProductName: Type0 Color: Type2 Size: Type3

upvoted 2 times

AvSUN 2 years, 11 months ago

The answers are 0, 3, 2

upvoted 2 times

AvSUN 2 years, 11 months ago

ProductName - 0 ProductColor - 3 ProductSize - 2The requirements and blanks are out of order in the questionProductColor cannot be Type3 since it indicates that it has to preserve all current and previous values so Type3 It only preserves the current and previous value without maintaining a complete history. For me it is: ProductName: Type0 Color: Type2 Size: Type3

upvoted 2 times

AvSUN 2 years, 11 months ago

ProductName - 0 ProductColor - 3 ProductSize - 2

upvoted 2 times

AvSUN 2 years, 11 months ago

The requirements and blanks are out of order in the question

upvoted 2 times

Lucasmh 2 years, 8 months ago

ProductColor cannot be Type3 since it indicates that it has to preserve all current and previous values so Type3 It only preserves the current and previous value without maintaining a complete history. For me it is: ProductName: Type0 Color: Type2 Size: Type3

upvoted 4 times

Deeksha1234 3 years ago

correct answer is - type 0, type 3,type 2

upvoted 4 times

akhil5432 3 years ago

type 0 type 3 type 2

upvoted 3 times

ravigolu 3 years, 1 month ago

Answer is Product name -type 0 color -type 2 size -type 3 Type 0 – Fixed Dimension No changes allowed, dimension never changes A Type 1 SCD always reflects the latest values, and when changes in source data are detected, the dimension table data is overwritten. Type 2 SCD A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. Type 3 SCD A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD.

upvoted 9 times

auwia 3 years, 1 month ago

Answer are corrects as reported in the solution: 0, 1, and 2.

upvoted 3 times

auwia 3 years, 1 month ago

• Prevent changes to the values stored in ProductName. => TYPE 0 • Retain only the current and the last values in ProductSize. => TYPE 2 (current and last, menas all the history) • Retain all the current and previous values in ProductColor. => TYPE 3 (it includes a column for the previous value)

upvoted 4 times

auwia 3 years, 1 month ago

• Prevent changes to the values stored in ProductName. => TYPE 0 • Retain only the current and the last values in ProductSize. => TYPE 2 (current and last, menas all the history) • Retain all the current and previous values in ProductColor. => TYPE 3 (it includes a column for the previous value)• Retain only the current and the last values in ProductSize. => TYPE 1 (current and last, DOESN'T means all the history, but as written only the last ... meaning the current). https://learn.microsoft.com/en-us/training/modules/load-optimize-data-into-relational-data-warehouse/5-load-slowly-changing-dimensionsRetain all the current and previous values in ProductColor. => TYPE 2 (because the plural in the requirement: previous values, probably it means all the history). Concluding the answers provided are correct :)

upvoted 4 times

auwia 3 years, 1 month ago

• Retain only the current and the last values in ProductSize. => TYPE 1 (current and last, DOESN'T means all the history, but as written only the last ... meaning the current). https://learn.microsoft.com/en-us/training/modules/load-optimize-data-into-relational-data-warehouse/5-load-slowly-changing-dimensions

upvoted 2 times

auwia 3 years, 1 month ago

Retain all the current and previous values in ProductColor. => TYPE 2 (because the plural in the requirement: previous values, probably it means all the history). Concluding the answers provided are correct :)

upvoted 2 times

Question #89 *Topic 1*

HOTSPOT

-

You are incrementally loading data into fact tables in an Azure Synapse Analytics dedicated SQL pool.

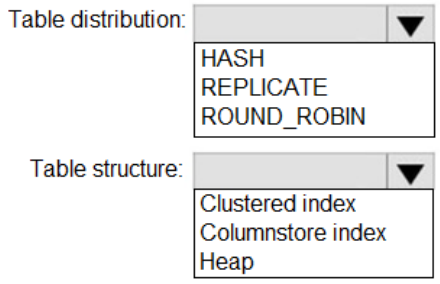
Each batch of incoming data is staged before being loaded into the fact tables.

You need to ensure that the incoming data is staged as quickly as possible.

How should you configure the staging tables? To answer, select the appropriate options in the answer area.

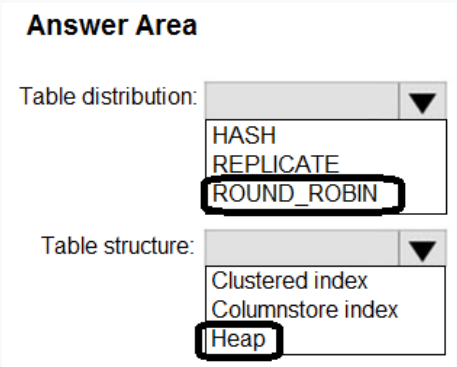
NOTE: Each correct selection is worth one point.

Answer Area

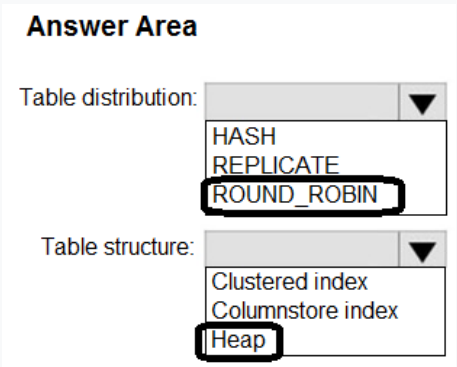


Suggested Answer:

Answer Area



Answer Area



orionduo Highly Voted 2 years, 1 month ago

Correct! The ROUND_ROBIN distribution distributes the data evenly across all distribution nodes in the SQL pool. This distribution type is suitable for loading data quickly into the staging tables because it minimizes the data movement during the loading process. Use a HEAP table: Instead of creating a clustered index on the staging table, it is recommended to create a HEAP table. A HEAP table does not have a clustered index, which eliminates the need for maintaining the index and improves the data loading performance. It allows for faster insert operations.

upvoted 23 times

kkk5566 1 year, 11 months ago

correct

upvoted 1 times

kkk5566 1 year, 11 months ago

correctFor the staging data

upvoted 1 times

kkk5566 Most Recent 1 year, 11 months ago

For the staging data

upvoted 1 times

Deeksha1234 2 years ago

ans is correct

upvoted 2 times

mehroosali 2 years, 2 months ago

correct

upvoted 3 times

abdallaissa 2 years, 2 months ago

Correct

upvoted 2 times

IanKwok81 2 years, 2 months ago

Correct

upvoted 2 times

Question #90 *Topic 1*

You have an Azure subscription that contains an Azure Synapse Analytics workspace named ws1 and an Azure Cosmos DB database account named Cosmos1. Cosmos1 contains a container named container1 and ws1 contains a serverless SQL pool.
You need to ensure that you can query the data in container1 by using the serverless SQL pool.
Which three actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Enable Azure Synapse Link for Cosmos1.

B. Disable the analytical store for container1.

C. In ws1, create a linked service that references Cosmos1.

D. Enable the analytical store for container1.

E. Disable indexing for container1.

Suggested Answer: *A, C, D*
Official Answer: ACD

Community vote distribution
ACD (100%)

pramod4lk Highly Voted 3 years ago
The answer is correct. We need to enable an analytical store in container1.
upvoted 5 times

owl29208 Most Recent 1 year, 6 months ago
A C D are correct
upvoted 1 times

moize 1 year, 8 months ago
Activez Azure Synapse Link pour Cosmos1 : A. Activez Azure Synapse Link pour Cosmos1. Cela permet d'intégrer Azure Synapse Analytics avec Azure Cosmos DB pour l'analyse des donnée Activez le magasin analytique pour le conteneur 1 : D. Activez le magasin analytique pour le conteneur 1. Le magasin analytique doit être activé pour permettre l'interrogation des données à l'aide de Synapse SQL Dans ws1, créez un service lié qui référence Cosmos1 : C. Dans ws1, créez un service lié qui référence Cosmos1. Cela permet de configurer la connexion entre Azure Synapse Analytics et Azure Cosmos DB
upvoted 1 times

EmnCours 1 year, 8 months ago
Correct Answer: ACD
upvoted 1 times

rocky48 2 years, 6 months ago
To enable querying data in container1 using the serverless SQL pool, you should perform the following actions: Enable Azure Synapse Link for Cosmos1: This allows seamless integration between Azure Synapse Analytics and Cosmos DB, enabling efficient querying of data stored in Cosmos DB containers 1. Create a linked service in ws1 that references Cosmos1: By creating a linked service, you establish a connection to your Cosmos DB account, allowing the serverless SQL pool to access data from Cosmos1 2. Disable the analytical store for container1: Since we want to query data using the serverless SQL pool, we don't need the analytical store feature for this specific scenario 3. Therefore, the correct actions are A, C, and B."Make sure that you have prepared Analytical store" is a prerequisite. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-cosmos-db-analytical-store?tabs=openrowset-key#prerequisites
upvoted 1 times

rocky48 2 years, 6 months ago
To enable querying data in container1 using the serverless SQL pool, you should perform the following actions:

Enable Azure Synapse Link for Cosmos1: This allows seamless integration between Azure Synapse Analytics and Cosmos DB, enabling efficient querying of data stored in Cosmos DB containers 1. Create a linked service in ws1 that references Cosmos1: By creating a linked service, you establish a connection to your Cosmos DB account, allowing the serverless SQL pool to access data from Cosmos1 2. Disable the analytical store for container1: Since we want to query data using the serverless SQL pool, we don't need the analytical store feature for this specific scenario 3. Therefore, the correct actions are A, C, and B.

upvoted 1 times

lcss27 2 years, 4 months ago

"Make sure that you have prepared Analytical store" is a prerequisite. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-cosmos-db-analytical-store?tabs=openrowset-key#prerequisites

upvoted 2 times

ChrisGe1234 2 years, 6 months ago

There are only 2 Correct answers. A & D. You don't need a linked Service to query Cosmos DB from Serverless Pool. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-cosmos-db-analytical-store?tabs=openrowset-key#prerequisites

upvoted 1 times

ChrisGe1234 2 years, 6 months ago

There are only 2 Correct answers. A & D. You don't need a linked Service to query Cosmos DB from Serverless Pool. https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-cosmos-db-analytical-store?tabs=openrowset-key#prerequisitesIf you need also synapse pipelines, adding a linked service is a requirement. It is a little trap, I think.

upvoted 1 times

MBRSDG 2 years, 4 months ago

If you need also synapse pipelines, adding a linked service is a requirement. It is a little trap, I think.

upvoted 1 times

AvSUN 2 years, 11 months ago

Correct

upvoted 1 times

kkk5566 2 years, 11 months ago

correct

upvoted 1 times

gusztimm 2 years, 12 months ago

Correct

upvoted 1 times

Deeksha1234 3 years ago

The answer is correct

upvoted 2 times

HOTSPOT

-

You have an Azure subscription that contains the resources shown in the following table.

The storage1 account contains a container named container1. The container1 container contains the following files.

In Pool1, you run the following script.

In the Built-in serverless SQL pool, you run the following script.

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Name	Type	Description
Workspace1	Azure Synapse workspace	Contains the Built-in serverless SQL pool
Pool1	Azure Synapse Analytics dedicated SQL pool	Deployed to Workspace1
storage1	Storage account	Hierarchical namespace enabled

```
Webdata <root folder>
    Monthly <folder>
            _monthly.csv
            Monthly.csv
    .testdata.csv
    testdata.csv
```

```
CREATE EXTERNAL DATA SOURCE Ds1
WITH
    ( LOCATION = 'abfss://container1@storage1.dfs.core.windows.net' ,
    CREDENTIAL = credential1,
    TYPE = HADOOP
  ) ;
```

```
CREATE EXTERNAL DATA SOURCE Ds2
WITH (
    LOCATION = 'https://storage1.blob.core.windows.net/container1/Webdata/',
    CREDENTIAL = credential2
);
```

Answer Area

Statements	Yes	No
An external table that uses Ds1 can read the _monthly.csv file.	○	○
An external table that uses Ds1 can read the Monthly.csv file.	○	○
An external table that uses Ds2 can read the .testdata.csv file.	○	○

Suggested Answer:

Answer Area

Statements	Yes	No
An external table that uses Ds1 can read the _monthly.csv file.	○	●
An external table that uses Ds1 can read the Monthly.csv file.	●	○
An external table that uses Ds2 can read the .testdata.csv file.	●	○

Answer Area

Statements	Yes	No
An external table that uses Ds1 can read the _monthly.csv file.	○	●
An external table that uses Ds1 can read the Monthly.csv file.	●	○
An external table that uses Ds2 can read the .testdata.csv file.	●	○

pramod4lk Highly Voted 3 years ago

The answer is No, Yes, No It will ignore "_" and "."

upvoted 54 times

pc1337xd Highly Voted 3 years ago

Both Hadoop(dedicated) and native(serverless) external tables will skip the files with the names that begin with an underline (_) or a period (.).

upvoted 12 times

AMJB Most Recent 1 year, 5 months ago

NO YES NO You can't read hidden files on both hadoop and native query tables . i.e _xxxx or .xxxxx will be hidden

upvoted 2 times

CezarioAbrantesPP 2 years, 1 month ago

Both Hadoop and native external tables will skip the files with the names that begin with an underline (_) or a period (.)

upvoted 1 times

Dusica 2 years, 3 months ago

No No No

upvoted 2 times

shadow2332 2 years, 6 months ago

No, No, No Because names that begin with an underline (_) or a period (.) are ignored and in the mentioned location, it will not be going to read subfolders, until '/**' is mentioned.

upvoted 6 times

shadow2332 2 years, 6 months ago

No, No, No Because names that begin with an underline (_) or a period (.) are ignored and in the mentioned location, it will not be going to read subfolders, until '/**' is mentioned.when you use DATA SOURCE to create an External Table you also use LOCATION (where you specified the relative path e.g. '/directory/**') https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#b-import-data-from-parquet-into-No, Yes, No In serverless SQL pools must be specified /** at the end of the location path. In Dedicated pool the folders are always scanned recursively.

upvoted 6 times

lcss27 2 years, 4 months ago

when you use DATA SOURCE to create an External Table you also use LOCATION (where you specified the relative path e.g. '/directory/**') https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#b-import-data-from-parquet-into-

upvoted 1 times

5a939c2 2 years ago

No, Yes, No In serverless SQL pools must be specified /** at the end of the location path. In Dedicated pool the folders are always scanned recursively.

upvoted 3 times

moize 2 years, 7 months ago

Réponse trouvée : Non - Oui - Non. Lien de référence est : https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1

upvoted 1 times

hcq31818 2 years, 8 months ago

No, Yes, No

upvoted 1 times

AvSUN 2 years, 11 months ago

NO, YES, NO https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1

upvoted 5 times

kkk5566 2 years, 11 months ago

NO YES NO see previous quizs.

upvoted 4 times

subhraz 2 years, 11 months ago

NO YES NO

upvoted 3 times

g2000 3 years ago

The last one is No. File is prefxied with a period and therefore can't be returned. https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#location--folder_or_filepath-1

upvoted 4 times

Question #92 *Topic 1*

DRAG DROP

-

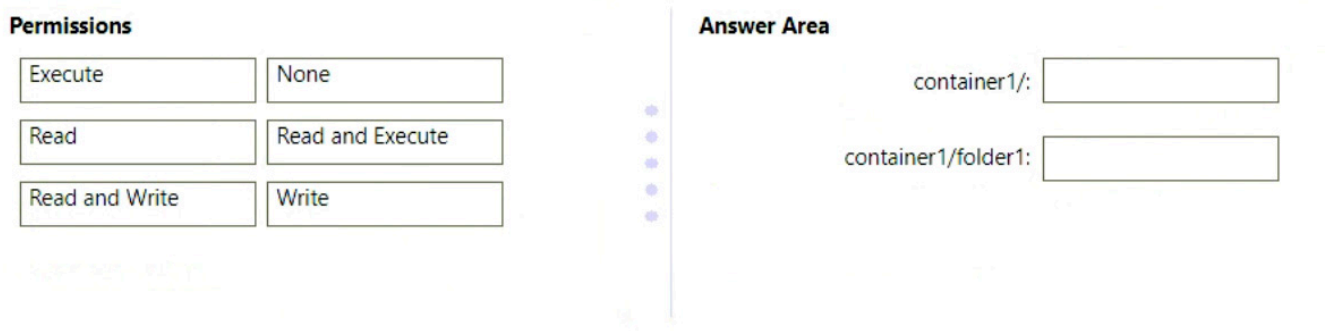
You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named account1 and a user named User1.

In account1, you create a container named container1. In container1, you create a folder named folder1.

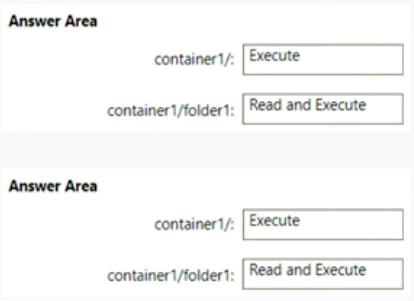
You need to ensure that User1 can list and read all the files in folder1. The solution must use the principle of least privilege.

How should you configure the permissions for each folder? To answer, drag the appropriate permissions to the correct folders. Each permission may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.



Suggested Answer:



g2000 Highly Voted 2 years, 6 months ago

correct! https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control#levels-of-permission

upvoted 21 times

Alongi Most Recent 1 year, 10 months ago

Correct, execute is to navigate folders

upvoted 1 times

TuxBingo 2 years, 1 month ago

Correct

upvoted 1 times

kkk5566 2 years, 5 months ago

correct

upvoted 2 times

Question #93 *Topic 1*

You have an Azure Data Factory pipeline named pipeline1.
You need to execute pipeline1 at 2 AM every day. The solution must ensure that if the trigger for pipeline1 stops, the next pipeline execution will occur at 2 AM, following a restart of the trigger.
Which type of trigger should you create?

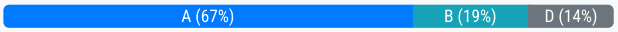
A. schedule

B. tumbling

C. storage event

D. custom event

Suggested Answer: *A*
Official Answer: A

Community vote distribution


yassine70 Highly Voted 2 years, 11 months ago
Answer is Tumbling : Link : https://learn.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers#trigger-type-comparison "Retry capability Supported. Failed pipeline runs have a default retry policy of 0, or a policy that's specified by the user in the trigger definition. Automatically retries when the pipeline runs fail due to concurrency/server/throttling limits (that is, status codes 400: User Error, 429: Too many requests, and 500: Internal Server error)" Retry capability is not supported on Schedule triggerI don't see 'Retry' as a requirement, they mentioned only about the next pipeline execution. 'A' seems to be more appropriate.
upvoted 10 times

yassine70 Highly Voted 2 years, 11 months ago
Answer is Tumbling : Link : https://learn.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers#trigger-type-comparison "Retry capability Supported. Failed pipeline runs have a default retry policy of 0, or a policy that's specified by the user in the trigger definition. Automatically retries when the pipeline runs fail due to concurrency/server/throttling limits (that is, status codes 400: User Error, 429: Too many requests, and 500: Internal Server error)"

Retry capability is not supported on Schedule trigger
upvoted 10 times

p_mks 2 years, 11 months ago
I don't see 'Retry' as a requirement, they mentioned only about the next pipeline execution. 'A' seems to be more appropriate.
upvoted 4 times

17lan Most Recent 1 year, 8 months ago
Tumbling because "following a restart of the trigger." Schedule does not support restart capability https://learn.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers#trigger-type-comparison
upvoted 1 times

EmnCours 1 year, 8 months ago
B. tumbling
upvoted 1 times

Harteko 1 year, 10 months ago
A tumbling window trigger runs at a specified interval, and if the trigger stops, it maintains its schedule when restarted. This ensures the pipeline executes exactly at 2 AM daily, even after a trigger restart. Tumbling windows are useful for consistent execution without overlap or gaps, which makes them ideal for daily fixed scheduling like in this scenario.
upvoted 1 times

Harteko 1 year, 10 months ago
A tumbling window trigger runs at a specified interval, and if the trigger stops, it maintains its schedule when restarted. This ensures the pipeline executes exactly at 2 AM daily, even after a trigger restart. Tumbling windows are useful for consistent execution without overlap or gaps, which makes them ideal for daily fixed scheduling like in your scenario.
upvoted 1 times

KarlGardnerDataEngineering 2 years, 2 months ago

This question is not clear enough. "If a pipeline stops" can mean many different things.

upvoted 2 times

Dusica 2 years, 3 months ago

Schedule at 1 , repeat once, delay 60 minutes

upvoted 1 times

lcss27 2 years, 4 months ago

I'll go for A based on the use cases provided Microsoft Data Platform MVP: https://www.cathrinewilhelmsen.net/triggers-azure-data-factory/

upvoted 1 times

Azure_2023 2 years, 6 months ago

You set up a tumbling window trigger with a window size of 1 hour and a start time of 2 AM. The trigger fires automatically at 2 AM every hour. If the trigger encounters an outage and misses a window (e.g., cannot execute at 2 AM), it automatically resumes at the beginning of the next window (i.e., at 3 AM). This ensures that the pipeline executes at 2 AM after the restart, effectively catching up for the missed execution.

upvoted 1 times

Khadija10 2 years, 6 months ago

in this scenario, the requirement is to execute the pipeline at a specific time every day (2 AM) and ensure that if the trigger stops, the next pipeline execution occurs at 2 AM following a restart of the trigger. A tumbling window trigger might not guarantee that the pipeline always runs at exactly 2 AM daily, especially if the trigger stops and restarts at different times. https://learn.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers#tumbling-window-trigger https://learn.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers#schedule-trigger-with-json

upvoted 1 times

moneytime 2 years, 6 months ago

The answer is B Tumbling window trigger is more specific .The pipeline is trigger every 24hrs which is how a tumbling window funtion works. Scheduler is general . One can say that a schedule trigger is called Tumbling trigger when it works at non overlapping equal intervals of time just as in the pipeline case study

upvoted 1 times

moneytime 2 years, 6 months ago

A Tumbling widow trigger is more specific. Scheduler is general . I can say this Schedule trigger is called Tumbling trigger..

upvoted 1 times

jongert 2 years, 7 months ago

The tumbling window trigger run waits for the triggered pipeline run to finish. Its run state reflects the state of the triggered pipeline run. For example, if a triggered pipeline run is cancelled, the corresponding tumbling window trigger run is marked cancelled. This is different from the "fire and forget" behavior of the schedule trigger, which is marked successful as long as a pipeline run started. https://learn.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers#trigger-type-comparison

upvoted 3 times

mav2000 2 years, 8 months ago

I believe it's tumbling, It doesn't say that if the trigger fails from the get-go, but if it stops, and in schedule, the state of a pipe is successful if the pipeline ran at first "The tumbling window trigger run waits for the triggered pipeline run to finish. Its run state reflects the state of the triggered pipeline run. For example, if a triggered pipeline run is cancelled, the corresponding tumbling window trigger run is marked cancelled. This is different from the "fire and forget" behavior of the schedule trigger, which is marked successful as long as a pipeline run started." https://learn.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers#trigger-type-comparison

upvoted 1 times

jsdsadfsad 2 years, 9 months ago

Looks to me like Tumbling

upvoted 3 times

madhubanti123 2 years, 10 months ago

Schedule triggers allow you to schedule pipelines to run at specific times and intervals. They are also idempotent, which means that if a pipeline execution fails due to a trigger failure, the next pipeline execution will still occur at the scheduled time.

upvoted 4 times

Vanq69 2 years, 10 months ago

"following a restart of the trigger" sounds a lot like a "retry" and only tumbling would offer a retry not schedule.

upvoted 1 times

HOTSPOT

-

You have an Azure data factory named adf1 that contains a pipeline named ExecProduct. ExecProduct contains a data flow named Product.

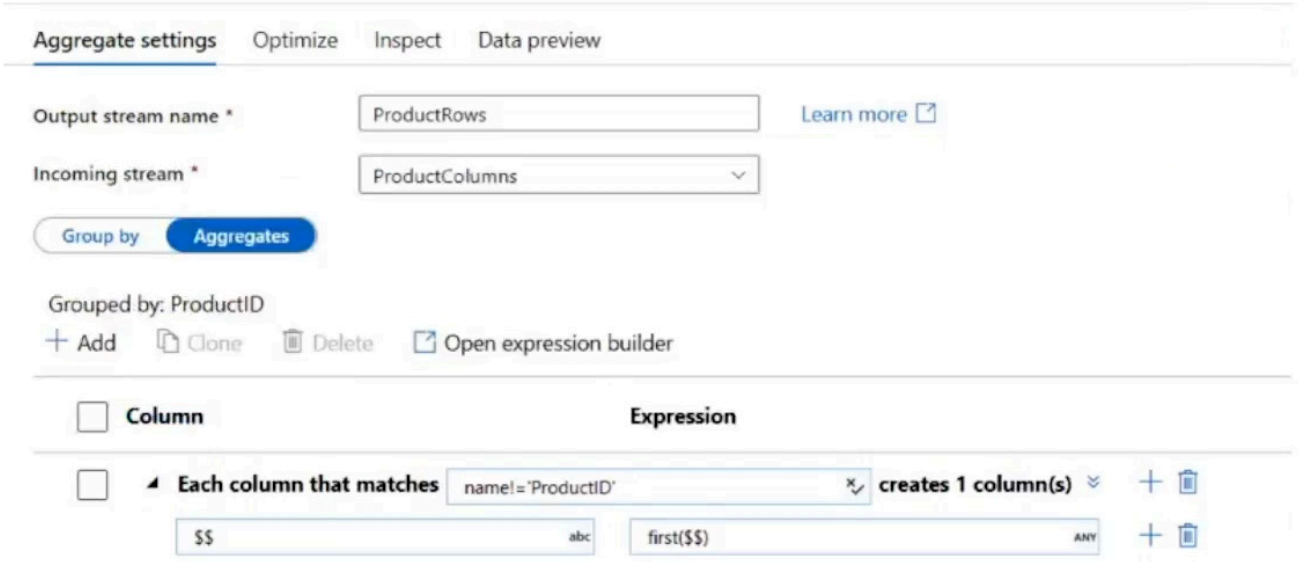
The Product data flow contains the following transformations:

1. WeeklyData: A source that points to a CSV file in an Azure Data Lake Storage Gen2 account with 20 columns
2. ProductColumns: A select transformation that selects from WeeklyData six columns named ProductID, ProductDescr, ProductSubCategory, ProductCategory, ProductStatus, and ProductLastUpdated
3. ProductRows: An aggregate transformation
4. ProductList: A sink that outputs data to an Azure Synapse Analytics dedicated SQL pool

The Aggregate settings for ProductRows are configured as shown in the following exhibit.

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

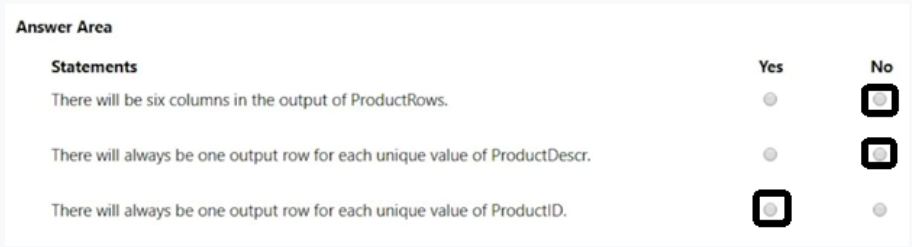
NOTE: Each correct selection is worth one point.

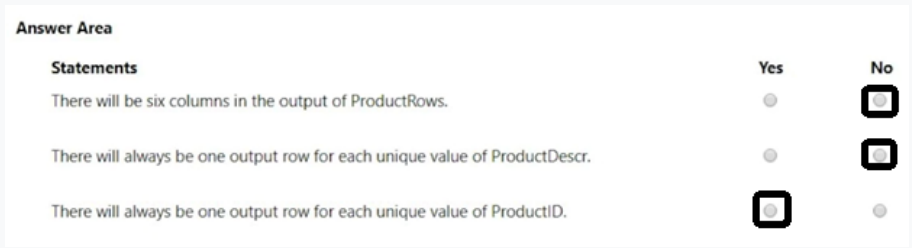


Answer Area

Statements	Yes	No
There will be six columns in the output of ProductRows.	○	○
There will always be one output row for each unique value of ProductDescr.	○	○
There will always be one output row for each unique value of ProductID.	○	○

Suggested Answer:





aeab260 Most Recent 1 year, 6 months ago

Y,N,Y. Almost exactly same solution on Microsoft https://learn.microsoft.com/en-us/azure/data-factory/data-flow-aggregate#removing-duplicate-rows

upvoted 1 times

Aurangzaib 2 years ago

No, No, Yes.

upvoted 2 times

Dusica 2 years, 3 months ago

Yes, No, Yes

upvoted 1 times

Alongi 2 years, 4 months ago

Yes, no, yes. You can take a look on https://learn.microsoft.com/en-us/azure/data-factory/data-flow-aggregate

upvoted 2 times

mariano2 2 years, 6 months ago

How come we might have multiple product description for unique product id if we group by product id ? It must be yes yes yes

upvoted 4 times

mariano2 2 years, 6 months ago

How come we might have multiple product description for unique product id if we group by product id ? It must be yes yes yesI also think it's yes, yes, yes. We're grouping by ProductID and select the first/max ProductDescr per ProductID, like so: select ProductID, first/max(ProductDescr) as FirstProductDescr from ProductColumns group by ProductID

upvoted 4 times

lola_mary5 2 years, 6 months ago

I also think it's yes, yes, yes. We're grouping by ProductID and select the first/max ProductDescr per ProductID, like so: select ProductID, first/max(ProductDescr) as FirstProductDescr from ProductColumns group by ProductID

upvoted 2 times

Bharatkumarc 2 years, 6 months ago

What if product id is duplicate i.e. for different combinations of the other columns we have the same product ID.. in that case, we will have multiple records for same product id even after aggregation and so answer to third Q would be No

upvoted 1 times

Charley92 2 years, 6 months ago

Yes / No / Yes

upvoted 1 times

Anto____ 2 years, 6 months ago

YES. NO. YES

upvoted 2 times

be8a152 2 years, 6 months ago

No - Input Schema is same as output schema, since there is no derived column to hold the aggregate value (Function "first" -returns the fist row among duplicate grouping key ie thge first unique row Product Id in this example). Input Schema will be same as Output Schema as the aggregator in this case is only used for removing duplicate rows by Product Id No - Product Description is not a group by column, hence there is no guarantee it will be unique. Yes - Product Id is the grouping Key hence the out will only have 1 unique row per Product IdYour explanation is correct for the first but you misscounted the columns - there are 6; Y N YWhat about the column qualifier !=ProductID ? I think that makes it 5 in the return

upvoted 2 times

be8a152 2 years, 6 months ago

No - Input Schema is same as output schema, since there is no derived column to hold the aggregate value (Function "first" -returns the fist row among duplicate grouping key ie thge first unique row Product Id in this example). Input Schema will be same as Output Schema as the aggregator in this case is only used for removing duplicate rows by Product Id No - Product Description is not a group by column, hence there is no guarantee it will be unique. Yes - Product Id is the grouping Key hence the out will only have 1 unique row per Product Id

upvoted 2 times

Dusica 2 years, 3 months ago

Your explanation is correct for the first but you misscounted the columns - there are 6; Y N YWhat about the column qualifier !=ProductID ? I think that makes it 5 in the return

upvoted 1 times

Dusica 2 years, 3 months ago

Your explanation is correct for the first but you misscounted the columns - there are 6; Y N Y

upvoted 1 times

7f668ce 2 years, 3 months ago

What about the column qualifier !=ProductID ? I think that makes it 5 in the return

upvoted 1 times

jongert 2 years, 7 months ago

1 - Yes, group by productId and then column pattern match the others results in 1 + 5 = 6 columns 2 - No, for aggregation using a group by clause we get one row for each unique value that we group by. 3 - Yes, the opposite compared to (2), since we are actually grouping by productId now.

upvoted 1 times

hcq31818 2 years, 8 months ago

1- Yes 2- No 3- Yes we are not creating new column, using aggregate function for deduplication https://learn.microsoft.com/en-us/azure/data-factory/data-flow-aggregate

upvoted 2 times

Tincox 2 years, 8 months ago

The syntax mentions name!= which means "not equal to" ProductID, so my answer would be: yes, yes, noUpon second glance my assumption was incorrect; the "first" pattern applies to all rows who's name is not ProductID. ProductID itself is the grouping column, so will return one output row for each unique value. ProductDescr., however, is part of the other columns so here it's the combination of these columns that has to be unique (first row of this combination of values is returned, the rest is dropped). The ProductDescr. column itself can generate more rows per unique value. So answer should be Yes, No, Yes.

upvoted 1 times

Tincox 2 years, 8 months ago

The syntax mentions name!= which means "not equal to" ProductID, so my answer would be: yes, yes, no

upvoted 1 times

Tincox 2 years, 8 months ago

Upon second glance my assumption was incorrect; the "first" pattern applies to all rows who's name is not ProductID. ProductID itself is the grouping column, so will return one output row for each unique value. ProductDescr., however, is part of the other columns so here it's the combination of these columns that has to be unique (first row of this combination of values is returned, the rest is dropped). The ProductDescr. column itself can generate more rows per unique value. So answer should be Yes, No, Yes.

upvoted 2 times

OldSchool 2 years, 11 months ago

1. No There will be 7 in output (6 are at input) 2. No We haven't tested anything by ProductDescr 3. Yes We have grouped by ProductID and added new columnDid we add a new column tho? $$ -> first($$) would just remove duplicate rows in ProductID and only keep the first value that is encountered. So it should be yes, no, yes. https://learn.microsoft.com/en-us/azure/data-factory/data-flow-aggregate

upvoted 4 times

OldSchool 2 years, 11 months ago

1. No There will be 7 in output (6 are at input) 2. No We haven't tested anything by ProductDescr 3. Yes We have grouped by ProductID and added new column

upvoted 4 times

Vanq69 2 years, 10 months ago

Did we add a new column tho? $$ -> first($$) would just remove duplicate rows in ProductID and only keep the first value that is encountered. So it should be yes, no, yes. https://learn.microsoft.com/en-us/azure/data-factory/data-flow-aggregate

upvoted 3 times

kkk5566 2 years, 11 months ago

yes, no, yes

upvoted 4 times

kkk5566 2 years, 11 months ago

Yes , no, yes

upvoted 3 times

mmoayed 3 years ago

I have notice that some answers might be wrong. What does this mean? who is confirming the correct answers ?

upvoted 4 times

mmoayed 3 years ago

yes, no, yes

upvoted 2 times

Question #95 *Topic 1*

You manage an enterprise data warehouse in Azure Synapse Analytics.
Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries.
You need to monitor resource utilization to determine the source of the performance issues.
Which metric should you monitor?

A. DWU limit

B. Cache hit percentage

C. Local tempdb percentage

D. Data IO percentage

Suggested Answer: *B*
Official Answer: B

Community vote distribution
B (100%)

MSExpert Highly Voted 2 years, 6 months ago
Correct
upvoted 6 times

ELJORDAN23 Most Recent 2 years, 1 month ago
Got this question on my exam on january 17, answer B is correct.
upvoted 1 times

ndangalasi 2 years, 4 months ago
https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monitor-cache
upvoted 3 times

Vanq69 2 years, 4 months ago
B. Cache hit percentage should be correct since it only affects common used queries, which should be saved and loaded from cache.
upvoted 3 times

kkk5566 2 years, 5 months ago
B is correct
upvoted 2 times

Question #96 *Topic 1*

HOTSPOT

-

You have an Azure Synapse Analytics serverless SQL pool.

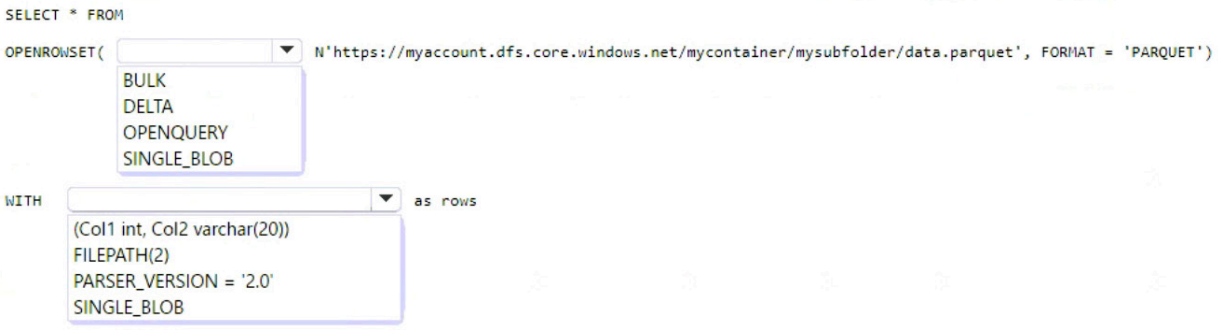
You have an Apache Parquet file that contains 10 columns.

You need to query data from the file. The solution must return only two columns.

How should you complete the query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area



Suggested Answer:





g2000 Highly Voted 2 years ago

correct! https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-openrowset#data-source

upvoted 15 times

kkk5566 Most Recent 1 year, 11 months ago

SELECT * FROM OPENROWSET(BULK 'http:// .dfs.core.windows.net/container/folder/*.parquet', FORMAT = 'PARQUET') AS [file]

upvoted 4 times

pramod4lk 2 years ago

Correct, Serverless SQL pool uses BULK.

upvoted 4 times

Question #97 *Topic 1*

You have an Azure Synapse Analytics workspace that contains an Apache Spark pool named SparkPool1. SparkPool1 contains a Delta Lake table named SparkTable1.
You need to recommend a solution that supports Transact-SQL queries against the data referenced by SparkTable1. The solution must ensure that the queries can use partition elimination.
What should you include in the recommendation?

A. a partitioned table in a dedicated SQL pool

B. a partitioned view in a dedicated SQL pool

C. a partitioned index in a dedicated SQL pool

D. a partitioned view in a serverless SQL pool

Suggested Answer: *D*
Official Answer: D

Community vote distribution



fahfouhi94 Most Recent 2 years ago
from this link : https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop Partition elimination is available only in the partitioned tables created on Parquet or CSV formats that are synchronized from Apache Spark pools. You might create external tables on Parquet partitioned folders, but the partitioning columns are inaccessible and ignored, while the partition elimination won't be applied. Don't create external tables on Delta Lake folders because they aren't supported. Use Delta partitioned views if you need to query partitioned Delta Lake data.
upvoted 3 times

Dusica 2 years, 3 months ago
D is correct the keyword here is: "Delta Lake table named SparkTable1" here is how partitioned view works: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-views#partitioned-views And this link shows that dedicated pool does not talk to delta table: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop
upvoted 1 times

mahmoud_salah30 2 years, 4 months ago
Partition elimination is available only in the partitioned tables created on Parquet or CSV formats that are synchronized from Apache Spark pools."https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop "
upvoted 1 times

Alongi 2 years, 4 months ago
Only partitioned Table can support partition elimination, so the answer should be A
upvoted 4 times

mariano2 2 years, 5 months ago
The folder partition elimination is available in the native external tables that are synchronized from the Synapse Spark pools. If you have partitioned data set and you would like to leverage the partition elimination with the external tables that you create, use the partitioned views instead of the external tables.
upvoted 1 times

moneytime 2 years, 6 months ago
D is correct. The solution is achieved by storing the delta lake table in sql spool ,so that transact-SQL query can be used on it..The choice of sql pool to use is determined by the format of the data .which is Delta.This format (Delta) is only supported by severless sql pool,hence it will be chosen for the job.Yes,it is important to state thay it can use the transct sql query through the openrowset function. The next case in the solution is the partitioning part. As a precaution from azure documentation ,external tables in severless sql pool with delta format should not be partitioned .If a partition must be used for query optimization,then use , a partition view .This view also contain partition folders or rather support partition folder elimination which is need for query optimization
upvoted 4 times

arihant_jain 2 years, 8 months ago
Everything is mentioned here: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop

upvoted 2 times

arihant_jain 2 years, 8 months ago

Everything is mentioned here: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoopQuote from link provided suggesting D is correct “use the partitioned views instead of the external tables.”

upvoted 2 times

ExamDestroyer69 2 years, 8 months ago

Quote from link provided suggesting D is correct “use the partitioned views instead of the external tables.”

upvoted 1 times

metiii 2 years, 9 months ago

D is correct. "The OPENROWSET function is not supported in dedicated SQL pools in Azure Synapse." so it eliminates A,B and C. Ref: https://learn.microsoft.com/en-us/sql/t-sql/functions/openrowset-transact-sql?view=sql-server-ver16 Only the partitioned view in the serverless sql pool is correct since "External tables in serverless SQL pools do not support partitioning on Delta Lake format. Use Delta partitioned views instead of tables if you have partitioned Delta Lake data sets." Ref: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables#delta-tables-on-partitioned-folders

upvoted 4 times

Question #98 *Topic 1*

You are designing a sales transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will contain approximately 60 million rows per month and will be partitioned by month. The table will use a clustered column store index and round-robin distribution.
Approximately how many rows will there be for each combination of distribution and partition?

A. 1 million

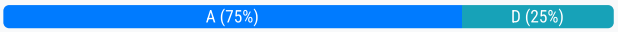
B. 5 million

C. 20 million

D. 60 million

Suggested Answer: *A*
Official Answer: A

Community vote distribution



Blablatest123 Highly Voted 2 years, 9 months ago
Partitioned by month and with 60 nodes, means it's 1M per combination
upvoted 14 times

moneytime Highly Voted 2 years, 6 months ago
Answer is A. Each partition has one million rows(minimum). *A partition with 1 distribution will have 1M rows * Partition with 2 distribution will have 2M rows ----' Partition with 60 distributions will have 60M rows ....Partition with n-distribution will have n ×1M rows = n M rows So in general, if there is p-partitions and d--distributions .then the combination of P and d describes the total rows or records (Q) which can be mathematical represented below Q = 1000000 * p *d For each combination of partition and distribution, P = 1 and d = 1 Therefore , The total rows,Q = 1000000*1*1= 1M. N.B .1M rows is the minimum per partion and distribution required for a table compression and optimization in table partitioning
upvoted 5 times

ff5037f Most Recent 1 year, 9 months ago
answer is A, as there will be 60 nodes for every distrubution
upvoted 1 times

Alongi 2 years, 4 months ago
Should be A
upvoted 2 times

d046bc0 2 years, 8 months ago
60 nodes so 1M per distribiution and partition
upvoted 1 times

BitacTeam 2 years, 8 months ago
1 Mio per combination
upvoted 1 times

SimonQBDS 2 years, 8 months ago
same as Blablatest123 said
upvoted 1 times

mishoka23 2 years, 9 months ago
Duplicate Question
upvoted 2 times

Question #99 *Topic 1*

You have an Azure Synapse Analytics workspace.
You plan to deploy a lake database by using a database template in Azure Synapse.
Which two elements are included in the template? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. relationships

B. data formats

C. linked services

D. table permissions

E. table definitions

Suggested Answer: *A, E*
Official Answer: AE

Community vote distribution
AE (100%)

metiii Highly Voted 2 years, 3 months ago
Correct, AE. Only table definition and their relationship is included in the template. The rest of the options should be configured Ref: https://learn.microsoft.com/en-us/azure/synapse-analytics/database-designer/create-lake-database-from-lake-database-templatesPlease can you explain this part of the template requirements. Prerequisites At least Synapse User role permissions are required for exploring a lake database template from Gallery. Synapse Administrator, or Synapse Contributor permissions are required on the Synapse workspace for creating a lake database. Storage Blob Data Contributor permissions are required on data lake when using the create table From data lake option. .I copied directly from the link you dropped .
upvoted 9 times

metiii Highly Voted 2 years, 3 months ago
Selected Answer: AE
Correct, AE. Only table definition and their relationship is included in the template. The rest of the options should be configured Ref: https://learn.microsoft.com/en-us/azure/synapse-analytics/database-designer/create-lake-database-from-lake-database-templates
upvoted 9 times

moneytime 2 years ago
Please can you explain this part of the template requirements. Prerequisites At least Synapse User role permissions are required for exploring a lake database template from Gallery. Synapse Administrator, or Synapse Contributor permissions are required on the Synapse workspace for creating a lake database. Storage Blob Data Contributor permissions are required on data lake when using the create table From data lake option. .I copied directly from the link you dropped .
upvoted 1 times

Dusica Most Recent 1 year, 9 months ago
B and E 100% Relationships you have to define
upvoted 2 times

Dusica 1 year, 9 months ago
B E ; relationships have to be added in the designer after tables are added
upvoted 2 times

Charley92 2 years ago
According to the Microsoft Learn documentation1, a lake database template in Azure Synapse includes the following elements: Table definitions: These are the schemas for the tables that store the data in the lake database. They include the table name, column names, data types, and descriptions. Relationships: These are the associations between the tables that define how they are related to each other. They include the primary and foreign keys, cardinality, and referential integrity rules. Data formats: These are the specifications for how the data is stored and accessed in the lake database. They include the file format, compression type, delimiter, encoding, and schema inference options. Therefore, the correct answers are A and E.
upvoted 1 times

ELJORDAN23 2 years, 1 month ago

Correct

upvoted 1 times

Question #100 *Topic 1*

You are implementing a star schema in an Azure Synapse Analytics dedicated SQL pool.
You plan to create a table named DimProduct.
DimProduct must be a Type 3 slowly changing dimension (SCD) table that meets the following requirements:
• The values in two columns named ProductKey and ProductSourceID will remain the same.
• The values in three columns named ProductName, ProductDescription, and Color can change.
You need to add additional columns to complete the following table definition.
Which three columns should you add? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

```
CREATE TABLE [dbo].[dimproduct]
  (
     [ProductKey]          INT NOT NULL,
     [ProductSourceID]     INT NOT NULL,
     [ProductName]         NVARCHAR(100) NOT NULL,
     [ProductDescription]  NVARCHAR(2000) NOT NULL,
     [Color]               NVARCHAR(50) NOT NULL
  )
WITH
(
    DISTRIBUTION = REPLICATE,
    CLUSTERED COLUMNSTORE INDEX
);
```

A. [EffectiveStartDate] [datetime] NOT NULL

B. [EffectiveEndDate] [datetime] NOT NULL

C. [OriginalProductDescription] NVARCHAR(2000) NOT NULL

D. [IsCurrentRow] [bit] NOT NULL

E. [OriginalColor] NVARCHAR(50) NOT NULL

F. [OriginalProductName] NVARCHAR(100) NULL

Suggested Answer: *C, E, F*
Official Answer: CEF

Community vote distribution
CEF (94%)

dakku987 Highly Voted 2 years, 8 months ago
CEF IS CORRECT
upvoted 5 times

nockda Most Recent 1 year, 8 months ago
A Type 3 Slowly Changing Dimension (SCD) table stores both historical and current values. When a change occurs, it requires effective dates (start/end) and a flag column to indicate the current state.
upvoted 1 times

ff5037f 1 year, 9 months ago
date columns are not mandatory in type 3. we have to store old and new values in type 3 scenario
upvoted 1 times

ivarozelin 2 years, 1 month ago
The key here is that the 3 columns can change. In Type SCD tables, we need to track the old and the new values by adding a new column that explicitly says it is the new value. For example, we will leave column Color and will add additional column with name NewColor.
upvoted 2 times

Dusica 2 years, 3 months ago
C D F is correct for type 3 A B D would be type 2

upvoted 1 times

Khadija10 2 years, 6 months ago

correct answer

upvoted 2 times

jimoskar 2 years, 6 months ago

yes cef!

upvoted 1 times

d046bc0 2 years, 8 months ago

correct

upvoted 4 times

Lucasmh 2 years, 8 months ago

The proposed solution is incorrect and corresponds to a type 2 scd, not 3. For it to be a type 3 scd, the start and end date of the change is needed along with the current value.

upvoted 4 times

Lucasmh 2 years, 8 months ago

The proposed solution is incorrect and corresponds to a type 2 scd, not 3. For it to be a type 3 scd, the start and end date of the change is needed along with the current value.You dont need start and end date for a type 3 scd In Type 3 Slowly Changing Dimension, there will be two columns to indicate the particular attribute of interest, one indicating the original value, and one indicating the current value. CEF is correct

upvoted 4 times

Kapello10 Highly Voted 2 years, 8 months ago

You dont need start and end date for a type 3 scd In Type 3 Slowly Changing Dimension, there will be two columns to indicate the particular attribute of interest, one indicating the original value, and one indicating the current value. CEF is correct

upvoted 7 times

metiii 2 years, 9 months ago

Correct. The other three options are needed for a scd type 2 table.

upvoted 3 times

Question #101 *Topic 1*

HOTSPOT

-

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Table1. Table1 contains sales data. Sixty-five million rows of data are added to Table1 monthly.

At the end of each month, you need to remove data that is older than 36 months. The solution must minimize how long it takes to remove the data.

How should you partition Table1, and how should you remove the old data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Partition the data:
- Partition by date with one partition per day.
- Partition by date with one partition per month.
- Partition by product.

Remove the data:
- Delete the old data from Table1 by using a WHERE clause.
- Delete the old data from Table1 by using a JOIN.
- Switch the oldest partition to another table named Table2 and drop Table2.
- Truncate the oldest partition.

Suggested Answer:

Answer Area

Partition the data: Partition by date with one partition per month.

Remove the data: Switch the oldest partition to another table named Table2 and drop Table2.

Answer Area

Partition the data: Partition by date with one partition per month.

Remove the data: Switch the oldest partition to another table named Table2 and drop Table2.

jongert Highly Voted 2 years, 7 months ago

Correct, dedicated SQL pool divides partitions into 60 databases and should aim for at least 1M rows per distribution (although not mentioned in the question, clustered columnstore compression becomes efficient at this scale). Therefore, partitioning by day would result in too small partitions. Having partitioned by months, we can use switch to move it using metadata operations only, which makes the operation extremely efficient.

upvoted 10 times

DataSturdy Most Recent 1 year, 11 months ago

why not just truncate the oldest partition

upvoted 1 times

Kryor 2 years, 1 month ago

Should you migrate the partition another table, delete the data from the main table, migrate again the partition and the drop table2?

upvoted 1 times

Dusica 2 years, 3 months ago

correct

upvoted 1 times

swathi_rs 2 years, 3 months ago

correct !

upvoted 1 times

Alongi 2 years, 4 months ago

Correct

upvoted 2 times

vernillen 2 years, 6 months ago

Correct

upvoted 3 times

Question #102 *Topic 1*

You have an Azure subscription that contains an Azure Synapse Analytics serverless SQL pool.
You execute the following query.
Where will the rows returned by the query be stored?

```
CREATE EXTERNAL TABLE Orders
WITH
(
     LOCATION = 'orders/',
     DATA_SOURCE = sales,
     FILE_FORMAT = SalesOrders
)
AS
SELECT OrderID, CustomerName, OrderTotal
FROM OPENROWSET
(
     BULK 'sales_orders/*.csv',
     DATA_SOURCE = 'sales',
     FORMAT = 'CSV',
     PARSER_VERSION = '2.0',
     HEADER_ROW = TRUE
) AS source_data
WHERE OrderType = 'Customer Order';
```

A. in a file in a data lake

B. in a relational database

C. in a global temporary table

D. in a session temporary table

Suggested Answer: *A*
Official Answer: A

Community vote distribution
A (100%)

jongert Highly Voted 2 years, 7 months ago
Correct: Serverless SQL pools can query, import, and store data from Azure Blob Storage, Azure Data Lake Storage Gen1 and Gen2. Serverless does not support TYPE=Hadoop. https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#overview-azure-synapse-analytics
upvoted 5 times

mavdplas Most Recent 1 year, 8 months ago
Answer A
upvoted 1 times

KarlGardnerDataEngineering 2 years, 2 months ago
How do we know this is ADLS (data lake) storage? Can't it be plain blob storage as well?
upvoted 1 times

swathi_rs 2 years, 3 months ago
Correct
upvoted 1 times

MohamedBI12 2 years, 5 months ago
their is no Tables with servless poolYou are right. We can't create the table in serverless SQL pool.
upvoted 2 times

MohamedBI12 2 years, 5 months ago
Selected Answer: A
their is no Tables with servless pool
upvoted 2 times

Bakhtiyor 2 years, 4 months ago

You are right. We can't create the table in serverless SQL pool.

upvoted 1 times

mav2000 2 years, 6 months ago

Correct "DATA_SOURCE = external_data_source_name Specifies the name of the external data source that contains the location of the external data. This location is in Azure Data Lake" https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#overview-azure-synapse-analytics

upvoted 4 times

vernillen 2 years, 6 months ago

A is the correct one

upvoted 1 times

jsav1 2 years, 7 months ago

correct

upvoted 1 times

moize 2 years, 7 months ago

Bonne réponse : A. Lien est : https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=dedicated#overview-azure-synapse-analytics

upvoted 1 times

Question #103 *Topic 1*

You are deploying a lake database by using an Azure Synapse database template.
You need to add additional tables to the database. The solution must use the same grouping method as the template tables.
Which grouping method should you use?

A. partition style

B. business area

C. size

D. facts and dimensions

Suggested Answer: *B*
Official Answer: B

Community vote distribution
B (91%) D (9%)

vernillen Highly Voted 2 years ago
Azure Synapse database templates are organized by "business areas," which are broader groupings that include various tables related to specific aspects of a business or industry. While "facts and dimensions" are a part of the overall data modeling process within these business areas, the question specifically asks about the grouping method used in the template tables. Therefore, the grouping method for adding additional tables to an Azure Synapse database template would align with the business areas defined within the template, making option B, "business area," the most appropriate choice given the context of the question and the structure of the Synapse database templates.top explanation, thanks
upvoted 16 times

vernillen Highly Voted 2 years ago
Selected Answer: B
Azure Synapse database templates are organized by "business areas," which are broader groupings that include various tables related to specific aspects of a business or industry.

While "facts and dimensions" are a part of the overall data modeling process within these business areas, the question specifically asks about the grouping method used in the template tables. Therefore, the grouping method for adding additional tables to an Azure Synapse database template would align with the business areas defined within the template, making option B, "business area," the most appropriate choice given the context of the question and the structure of the Synapse database templates.
upvoted 16 times

the_frix 1 year, 11 months ago
top explanation, thanks
upvoted 3 times

ELJORDAN23 2 years, 1 month ago
The answer is right below the first image in the documentation: https://learn.microsoft.com/en-us/azure/synapse-analytics/database-designer/overview-database-templates
upvoted 7 times

Dusica 1 year, 9 months ago
B is correct
upvoted 1 times

Alongi 1 year, 10 months ago
It's B
upvoted 1 times

sdg2844 2 years, 1 month ago
I think it should be D also.
upvoted 1 times

moize 2 years, 1 month ago
Réponse : B Ref : https://learn.microsoft.com/en-us/azure/synapse-analytics/database-designer/overview-database-templates
upvoted 1 times

Filda123 2 years, 1 month ago

jongert is right, should be B

upvoted 3 times

dakku987 2 years, 1 month ago

B. business area This is a common approach where tables are grouped based on the business areas or domains they belong to, making it easier to manage, organize, and understand the structure of the database.

upvoted 2 times

jongert 2 years, 1 month ago

Correct https://learn.microsoft.com/en-us/azure/synapse-analytics/database-designer/overview-database-templates

upvoted 4 times

mrplmcc 2 years, 1 month ago

Selected Answer: D

Why? Shouldn't it be D?

upvoted 2 times

mrplmcc 2 years, 1 month ago

Why? Shouldn't it be D?This grouping applies to data warehouse star/snowflake schemas, not lake databases

upvoted 2 times

ihmot **Most Recent** 1 year, 6 months ago

This grouping applies to data warehouse star/snowflake schemas, not lake databases

upvoted 1 times

Question #104 *Topic 1*

You have an Azure data factory connected to a Git repository that contains the following branches:
• main: Collaboration branch
• abc: Feature branch
• xyz: Feature branch
You save changes to a pipeline in the xyz branch.
You need to publish the changes to the live service.
What should you do first?

A. Publish the data factory.

B. Create a pull request to merge the changes into the main branch.

C. Create a pull request to merge the changes into the abc branch.

D. Push the code to a remote origin.

Suggested Answer: *B*
Official Answer: B

Community vote distribution
B (88%) A (12%)

NAWRESS96 Most Recent 2 years ago
Correct . C'est la première étape appropriée. Vous devez créer une pull request pour fusionner vos modifications de la branche xyz dans la branche main. Une fois que cette PR est approuvée et fusionnée, vous pourrez publier les changements dans le service en direct.
upvoted 1 times

606a82e 2 years ago
correct
upvoted 1 times

Danweo 2 years, 1 month ago
B is correct
upvoted 1 times

princepark 2 years, 4 months ago
Save changes to a pipeline in the xyz branch means that the changes are Published in the data factory
upvoted 1 times

Alongi 2 years, 4 months ago
Why not A? You should Publish the save changes before every pull requests
upvoted 1 times

DiLsH 2 years, 6 months ago
Correct
upvoted 1 times

jsav1 2 years, 7 months ago
Correct, always merge the feature into to the main branch and then publish the main branch
upvoted 3 times

moize 2 years, 7 months ago
Bonne réponse : B.
upvoted 1 times

jongert 2 years, 7 months ago
Correct, simply best practices for version control. Each feature branch develops changes, then should create pull requests to merge with main branch before publishing.
upvoted 4 times

Question #105 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure subscription that contains an Azure data factory named ADF1.
From Azure Data Factory Studio, you build a complex data pipeline in ADF1.
You discover that the Save button is unavailable, and there are validation errors that prevent the pipeline from being published.
You need to ensure that you can save the logic of the pipeline.
Solution: You enable Git integration for ADF1.
Does this meet the goal?

A. Yes

B. No

Suggested Answer: *A*
Official Answer: A

Community vote distribution



lola_mary5 Highly Voted 2 years, 6 months ago
Partial saves: "Whether your pipelines are not finished or you simply don't want to lose changes if your computer crashes, git integration allows for incremental changes of data factory resources regardless of what state they are in." https://learn.microsoft.com/en-us/azure/data-factory/source-control
upvoted 10 times

ddhoogduin Highly Voted 2 years, 7 months ago
Selected Answer: B
Git configuration is already enabled otherwise the "Save" option would be a "Publish" option.
upvoted 6 times

ddhoogduin 2 years, 7 months ago
Git configuration is already enabled otherwise the "Save" option would be a "Publish" option.It says the Save option is unavailable, which can mean its not present. In that case, it implies that Git is not configuredMan, I feel like this is a trick question with the wording lol. Depends on how you interpret the word "unavailable". I think_Ahan_ might be right with the word "unavailable" meaning not present. Anyways, gonna do some more research on it!
upvoted 6 times

Ahan 2 years, 3 months ago
It says the Save option is unavailable, which can mean its not present. In that case, it implies that Git is not configured
upvoted 1 times

Ahan 2 years, 3 months ago
It says the Save option is unavailable, which can mean its not present. In that case, it implies that Git is not configuredMan, I feel like this is a trick question with the wording lol. Depends on how you interpret the word "unavailable". I think_Ahan_ might be right with the word "unavailable" meaning not present. Anyways, gonna do some more research on it!
upvoted 1 times

KarlGardnerDataEngineering 2 years, 2 months ago
Man, I feel like this is a trick question with the wording lol. Depends on how you interpret the word "unavailable". I think_Ahan_ might be right with the word "unavailable" meaning not present. Anyways, gonna do some more research on it!
upvoted 1 times

renan_ineu Most Recent 1 year, 11 months ago
The answer is YES: "When authoring against the data factory service, you can't save changes as a draft and all publishes must pass data factory validation. Whether your pipelines are not finished or you simply don't want to lose changes if your computer crashes, git integration allows for incremental changes of data factory resources regardless of what state they are in. Configuring a git repository allows you to save changes, letting you only publish when you have tested your changes to your satisfaction." https://learn.microsoft.com/en-us/answers/questions/391324/save-option-not-enabled
upvoted 1 times

Danweo 2 years, 1 month ago

Don't overthink it, you enable git so you can save it. This is not about fixing the problem just making sure we can save what we have so far.

upvoted 1 times

Sr18 2 years, 2 months ago

Answer is A 100%. Reason: Save button is unavailable means git is not enabled or I say you are in live mode you will see only Publish icon. In that case there is no way you can Publish the code. So you need to enable the GIT to save the code, but switching to git mode from live mode with remove the changes in live mode, so go to json view and copy the code. Discard changes switch to git mode, create new pipeline and paste the copied json in the json view and you can save the code now

upvoted 4 times

Sr18 2 years, 2 months ago

Answer is A 100%. Reason: Save button is unavailable means git is not enabled or I say you are in live mode you will see only Publish icon. In that case there is no way you can Publish the code. So you need to enable the GIT to save the code, but switching to git mode from live mode with remove the changes in live mode, so go to json view and copy the code. Discard changes switch to git mode, create new pipeline and paste the copied json in the json view and you can save the code nowExplained very well.

upvoted 4 times

learnwell 2 years, 1 month ago

Explained very well.

upvoted 1 times

jpgsa11 2 years, 3 months ago

Enabling Git Integration is just the first step. Doesn't necessarily allow to have a backup. JSON file on the other hand its different

upvoted 1 times

Alongi 2 years, 4 months ago

It's A

upvoted 1 times

poesklap 2 years, 4 months ago

Enabling Git integration for Azure Data Factory (ADF) does not directly address the issue of validation errors preventing the pipeline from being published. Git integration allows you to manage your ADF resources using source control and enables collaboration among multiple developers. It doesn't inherently fix validation errors.

upvoted 1 times

MBRSDG 2 years, 5 months ago

As stated, this question is a bit misleading. "You need to ensure that you can save the logic of the pipeline" means "you have to prevent the situation in which already working pipelines are damaged by buggy implementations". If you have the working, not buggy, code already inside the repository, you can create a new branch and do the updates there. If there's a error, it is contained inside the feature branch and not inside the main branch, which keeps working fine.

upvoted 2 times

Delphin_8150 2 years, 5 months ago

Answer should be A

upvoted 1 times

Alongi 2 years, 6 months ago

No, enabling Git integration alone does not directly address the immediate goal of saving the logic of the pipeline and resolving validation errors in Azure Data Factory Studio. You need to fix the validation errors and save the changes locally within the studio before enabling Git integration for long-term version control.

upvoted 2 times

ELJORDAN23 2 years, 7 months ago

Although the validation might be wrong, enablng the git integration allows you to save your datasets, pipelines and everything else as json files in your git repo, so A) is correct.

upvoted 4 times

jongert 2 years, 7 months ago

Safe to assume that enabling git integration also means setting up the repo and branches, then it would allow saving. Would answer yes.

upvoted 3 times

Question #106 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure subscription that contains an Azure data factory named ADF1.
From Azure Data Factory Studio, you build a complex data pipeline in ADF1.
You discover that the Save button is unavailable, and there are validation errors that prevent the pipeline from being published.
You need to ensure that you can save the logic of the pipeline.
Solution: You view the JSON code representation of the resource and copy the JSON to a file.
Does this meet the goal?

A. Yes

B. No

Suggested Answer: *B*
Official Answer: B

Community vote distribution
A (64%) B (36%)

dakku987 Highly Voted 2 years, 7 months ago
B. No The Save button being unavailable and validation errors preventing the pipeline from being published indicate issues with the current configuration or logic of the pipeline within Azure Data Factory Studio. Copying the JSON code to a file won't resolve the validation errors or allow you to save the pipeline.Please ignore my comment its wrong
upvoted 6 times

dakku987 2 years, 7 months ago
Selected Answer: B
B. No

The Save button being unavailable and validation errors preventing the pipeline from being published indicate issues with the current configuration or logic of the pipeline within Azure Data Factory Studio. Copying the JSON code to a file won't resolve the validation errors or allow you to save the pipeline.
upvoted 6 times

dakku987 Highly Voted 2 years, 7 months ago
Please ignore my comment its wrong
upvoted 11 times

Danweo Most Recent 2 years, 1 month ago
Horrible question, but yes you technically can store it as JSON to keep the logic and you're work.
upvoted 2 times

gplusplus 2 years, 4 months ago
Selected Answer: B
"You view the JSON code representation of the RESOURCE and copy the JSON to a file". what "resource" are we talking about? If the JSON of the "pipeline resource" was specified, I would be 50/50% as saving a JSON achives this but isn't best practice. However here the proposed solution is intentionally vague, doesn't mention that the full json pipeline will be saved, could be another "resource" within the pipeline. Going for Nope
upvoted 1 times

gplusplus 2 years, 4 months ago
"You view the JSON code representation of the RESOURCE and copy the JSON to a file". what "resource" are we talking about? If the JSON of the "pipeline resource" was specified, I would be 50/50% as saving a JSON achives this but isn't best practice. However here the proposed solution is intentionally vague, doesn't mention that the full json pipeline will be saved, could be another "resource" within the pipeline. Going for NopeMoreover, the first correct variation is much more specific in its solution, mentioning the correct name of the resource: "Solution: You enable Git integration for ADF1".
upvoted 1 times

gplusplus 2 years, 4 months ago

Moreover, the first correct variation is much more specific in its solution, mentioning the correct name of the resource: "Solution: You enable Git integration for ADF1".

upvoted 1 times

poesklap 2 years, 4 months ago

Viewing the JSON code representation of the pipeline and copying it to a file can help preserve the logic of the pipeline, even if the Save button is unavailable due to validation errors. This allows you to retain the pipeline configuration and logic for future reference or for manual editing to address the validation errors. While it doesn't directly fix the validation errors, it ensures that you have a backup of the pipeline definition.

upvoted 2 times

moneytime 2 years, 6 months ago

A is correct. The solution only aims at preserving the logic of the code .So viewing and copying the JSON code to another file will support versioning through partial saves which is required for securing the logic. of the code. N.B The acceptable solution in Azure is through the provisioning of the git repository which helps in source control,versioning ,collaboration etc.

upvoted 1 times

Alongi 2 years, 6 months ago

Yes, it works fine

upvoted 1 times

ChrisGe1234 2 years, 6 months ago

Question asks how to save logic. This would work.

upvoted 1 times

ELJORDAN23 2 years, 7 months ago

Maybe you can save manually your json by copying the content to your local machine or something like that, but in an Azure context, I think that the question implies that we are using a solution involving Azure technology. Can you copy your json contents to a file? Yes of course. Does that enable the Save button? No, it doesn't. It is not the best practice, so I'm going with a No.

upvoted 2 times

jsav1 2 years, 7 months ago

Yes, it would theoretically work, but it is not a good idea.

upvoted 1 times

vernillen 2 years, 7 months ago

Anwser should be "Yes", and because of the phrase: "You need to ensure that you can save the logic of the pipeline.". This means you have to save the logic of the pipeline, and not the pipeline itself. This won't, however, resolve the issues and errors, but it will provide you with a back-up of your work so far.

upvoted 3 times

moize 2 years, 7 months ago

Bonne réponse : B-----> NON

upvoted 1 times

dakku987 2 years, 7 months ago

Yes, viewing the JSON code representation of the resource and copying the JSON to a file is a valid approach to save the logic of the pipeline. The JSON code represents the configuration and structure of the data pipeline, and you can store it in a file for future use or reference. While it doesn't directly address the validation errors preventing publishing, it allows you to preserve the logic of the pipeline.

upvoted 1 times

jongert 2 years, 7 months ago

The JSON file contains the logic of the pipeline and configurations such as paths. It should achieve the goal, although it would not be best practice.

upvoted 3 times

mrplmcc 2 years, 7 months ago

Yes it should work

upvoted 2 times

Question #107 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure subscription that contains an Azure data factory named ADF1.
From Azure Data Factory Studio, you build a complex data pipeline in ADF1.
You discover that the Save button is unavailable, and there are validation errors that prevent the pipeline from being published.
You need to ensure that you can save the logic of the pipeline.
Solution: You export ADF1 as an Azure Resource Manager (ARM) template.
Does this meet the goal?

A. Yes

B. No

Suggested Answer: *B*
Official Answer: B

Community vote distribution
B (61%) A (39%)

[Removed] Highly Voted 2 years, 6 months ago
Selected Answer: B
This has to be B. If you have experience working with Azure before you will know that it is impossible to export anything as an ARM template with validation errors.

https://learn.microsoft.com/en-us/azure/azure-resource-manager/troubleshooting/error-invalid-template?tabs=bicep

Scroll down to "Solution 3".
upvoted 8 times

[Removed] Highly Voted 2 years, 6 months ago
This has to be B. If you have experience working with Azure before you will know that it is impossible to export anything as an ARM template with validation errors. https://learn.microsoft.com/en-us/azure/azure-resource-manager/troubleshooting/error-invalid-template?tabs=bicep Scroll down to "Solution 3".Solution 3 is about an error during deployment. The question is about a validation error so, I don't see how this comment applies to the question.
upvoted 8 times

lola_mary5 2 years, 6 months ago
Solution 3 is about an error during deployment. The question is about a validation error so, I don't see how this comment applies to the question.
upvoted 2 times

ChrisGe1234 2 years, 6 months ago
You can't export an ARM Template with validation errors. I don't know why some people voted yes on this one.
upvoted 7 times

Danweo Most Recent 2 years, 1 month ago
From Azure documentation: When exporting from a resource group or resource, the exported template is generated from the published schema We can't publish with errors so this cannot be done with ARM template export. https://learn.microsoft.com/en-us/azure/azure-resource-manager/templates/export-template-portal
upvoted 2 times

SarathChandra 2 years, 2 months ago
B is correct. ARM template will only have latest Pipeline work only when published. Since the pipeline has validation errors it can't be published.
upvoted 1 times

tadenet 2 years, 4 months ago
Yes, you can export Azure Data Factory (ADF1) pipelines, datasets, linked services, and other artifacts as Azure Resource Manager (ARM) templates. This process allows you to capture the configuration of your data factory in a JSON-based template format. You can then use this template to automate deployment, manage version control, or replicate your data factory across different environments. Exporting an ADF1 data factory as an

ARM template can be done from the Azure portal. Simply navigate to your data factory, select the "Author" tab, and then click on "Export ARM template" from the menu. This will generate an ARM template containing the definition of your data factory and its components, which you can download and use as needed.

upvoted 1 times

poesklap 2 years, 4 months ago

Exporting the Azure Data Factory (ADF) as an Azure Resource Manager (ARM) template would capture the logic of the pipeline along with other ADF resources. This allows you to save the configuration and logic of the pipeline in a structured format, even if the Save button is unavailable due to validation errors. While it doesn't directly fix the validation errors, it ensures that you have a backup of the pipeline definition in an ARM template, which can be modified and redeployed later.

upvoted 1 times

moneytime 2 years, 6 months ago

Answer is B. The ARM template is disabled or rather it is not available in ADF ,hence there is no template to export .This means that there no template to hold the logic of the program. However, when it is connected to the git,the template is enabled in there (lol..They are good marketers)

upvoted 1 times

Charley92 2 years, 6 months ago

This will allow you to save the logic of the pipeline and make changes to it as needed

upvoted 2 times

AzurePart 2 years, 6 months ago

No, exporting ADF1 as an Azure Resource Manager (ARM) template does not meet the goal of ensuring that you can save the logic of the pipeline. ARM templates are used to deploy resources, not to save the logic of a data pipeline. To save the logic of the pipeline, you need to resolve the validation errors that are preventing the pipeline from being published. Once the errors are resolved, the Save button will become available, and you can save the pipeline.

upvoted 1 times

jsav1 2 years, 7 months ago

answer is A as it is possible to use an ARM template to save the logic, but it's not necessarily best practice when you could use a git repo instead

upvoted 1 times

dakku987 2 years, 7 months ago

A. Yes Exporting ADF1 as an Azure Resource Manager (ARM) template will capture the logic of the pipeline in JSON format. Even if the Save button is unavailable in the Azure Data Factory Studio due to validation errors, exporting the ARM template allows you to save the pipeline logic in a file. You can then review and edit the JSON code to correct the validation errors and redeploy the updated ARM template to resolve the issues.

upvoted 3 times

mrplmcc 2 years, 7 months ago

Selected Answer: A

From chat gpt: Yes, exporting the Azure Data Factory (ADF1) as an Azure Resource Manager (ARM) template can meet the goal of ensuring that you can save the logic of the pipeline, even when the Save button is unavailable due to validation errors.

When you export the Azure Data Factory as an ARM template, it captures the entire structure and configuration of the Data Factory, including pipelines, datasets, linked services, triggers, and other artifacts in JSON format. This exported ARM template serves as a backup or snapshot of your Data Factory configuration.

Therefore, by exporting ADF1 as an ARM template, you create a backup of the entire Data Factory structure, including the complex data pipeline that you built. This allows you to save the logic of the pipeline, despite the Save button being unavailable due to validation errors. Later, you can rectify the issues causing validation errors and re-import the updated ARM template to restore the logic of the pipeline.

upvoted 4 times

mrplmcc 2 years, 7 months ago

From chat gpt: Yes, exporting the Azure Data Factory (ADF1) as an Azure Resource Manager (ARM) template can meet the goal of ensuring that you can save the logic of the pipeline, even when the Save button is unavailable due to validation errors. When you export the Azure Data Factory as an ARM template, it captures the entire structure and configuration of the Data Factory, including pipelines, datasets, linked services, triggers, and other artifacts in JSON format. This exported ARM template serves as a backup or snapshot of your Data Factory configuration. Therefore, by exporting ADF1 as an ARM template, you create a backup of the entire Data Factory structure, including the complex data pipeline that you built. This allows you to save the logic of the pipeline, despite the Save button being unavailable due to validation errors. Later, you can rectify the issues causing validation errors and re-import the updated ARM template to restore the logic of the pipeline.Agree, also the MS documentation for it. https://learn.microsoft.com/en-us/azure/data-factory/continuous-integration-delivery-manual-promotionchatgpt B. No Exporting ADF1 as an Azure Resource Manager (ARM) template is not a direct solution to saving the logic of the pipeline in the Azure Data Factory Studio when the Save button is unavailable due to validation errors. Exporting as an ARM template is typically done for versioning, source control, or deployment purposes, and it

does not directly address the issue of saving the pipeline logic within the Data Factory Studio interface. The suggested approach in the scenario would be to address and resolve the validation errors preventing the pipeline from being published, allowing you to save the changes within the Azure Data Factory Studio. Once the validation errors are fixed, you should be able to save and publish the pipeline without exporting it as an ARM template.It's correct, you shouldn't be using ARM, but it works if you want to save the state of the p

upvoted 4 times

jongert 2 years, 7 months ago

Agree, also the MS documentation for it. https://learn.microsoft.com/en-us/azure/data-factory/continuous-integration-delivery-manual-promotion

upvoted 3 times

dakku987 2 years, 7 months ago

chatgpt B. No

Exporting ADF1 as an Azure Resource Manager (ARM) template is not a direct solution to saving the logic of the pipeline in the Azure Data Factory Studio when the Save button is unavailable due to validation errors. Exporting as an ARM template is typically done for versioning, source control, or deployment purposes, and it does not directly address the issue of saving the pipeline logic within the Data Factory Studio interface.

The suggested approach in the scenario would be to address and resolve the validation errors preventing the pipeline from being published, allowing you to save the changes within the Azure Data Factory Studio. Once the validation errors are fixed, you should be able to save and publish the pipeline without exporting it as an ARM template.

upvoted 1 times

dakku987 2 years, 7 months ago

chatgpt B. No Exporting ADF1 as an Azure Resource Manager (ARM) template is not a direct solution to saving the logic of the pipeline in the Azure Data Factory Studio when the Save button is unavailable due to validation errors. Exporting as an ARM template is typically done for versioning, source control, or deployment purposes, and it does not directly address the issue of saving the pipeline logic within the Data Factory Studio interface. The suggested approach in the scenario would be to address and resolve the validation errors preventing the pipeline from being published, allowing you to save the changes within the Azure Data Factory Studio. Once the validation errors are fixed, you should be able to save and publish the pipeline without exporting it as an ARM template.It's correct, you shouldn't be using ARM, but it works if you want to save the state of the pipeline. Gpt is saying that it's not the optimal way, and it should be done with Git Integration

upvoted 1 times

mav2000 2 years, 6 months ago

It's correct, you shouldn't be using ARM, but it works if you want to save the state of the pipeline. Gpt is saying that it's not the optimal way, and it should be done with Git Integration

upvoted 2 times

Question #108 *Topic 1*

HOTSPOT

-

You have an Azure Databricks workspace.

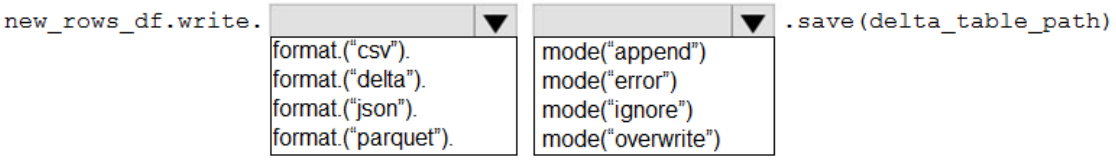
You read data from a CSV file by using a notebook, and then load the data to a DataFrame.

You need to add rows from the DataFrame to an existing Delta table by using Python code.

How should you complete the code? To answer, select the appropriate options in the answer area.

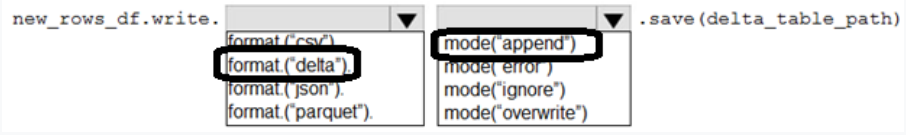
NOTE: Each correct selection is worth one point.

Answer Area

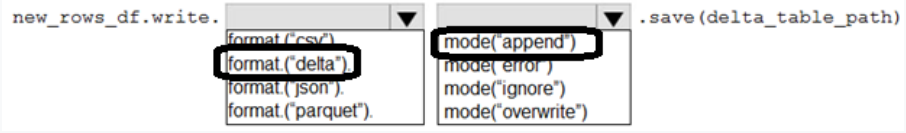


Suggested Answer:

Answer Area



Answer Area



jongert Highly Voted 2 years, 1 month ago

Correct: Add to a delta table => format is delta Add the rows to existing table => appendAgreed

upvoted 8 times

jongert Highly Voted 2 years, 1 month ago

Correct: Add to a delta table => format is delta Add the rows to existing table => append

upvoted 8 times

vernillen 2 years ago

Agreed

upvoted 1 times

Khadija10 Most Recent 2 years ago

correct https://learn.microsoft.com/en-us/training/modules/use-delta-lake-azure-synapse-analytics/3-create-delta-tables

upvoted 3 times

Alongi 2 years ago

Correct

upvoted 1 times

DRAG DROP

-

You have an Azure subscription that contains an Azure Cosmos DB for NoSQL account named account1. The account1 account contains a container named Container1 that has the following configurations:

• Analytical store: On

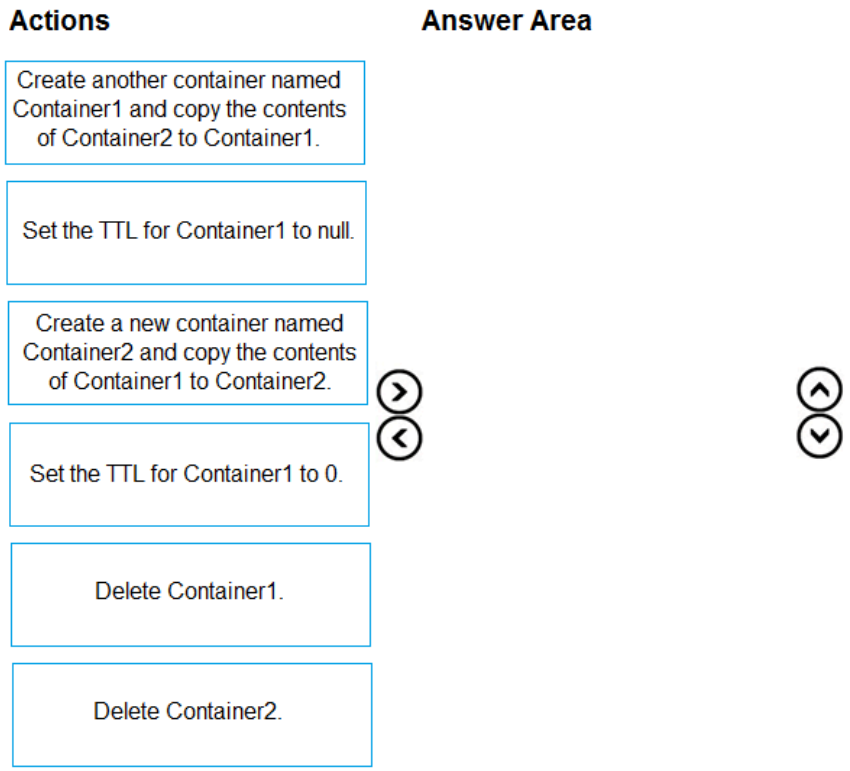
• TTL: 3600

You need to remove analytical store support from Container1. The solution must meet the following requirements:

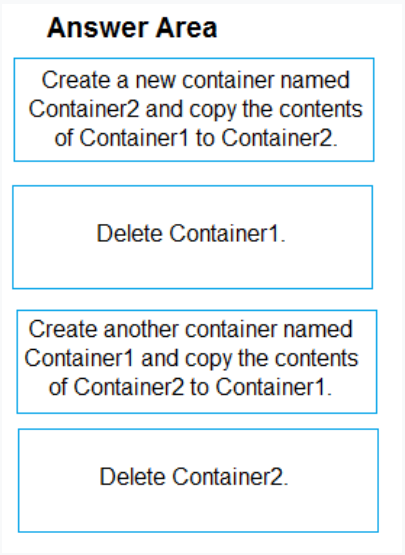
• Minimize the impact on the apps that reference Container1.

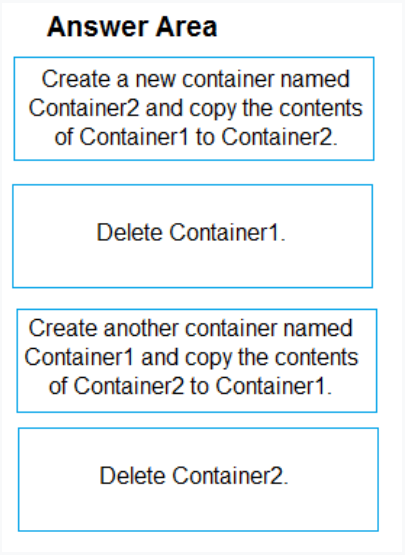
• Minimize storage usage.

Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



Suggested Answer:





jongert Highly Voted 2 years, 7 months ago

CorrectAgreedagreedAgreedagreedagreedAgreed

upvoted 11 times

jongert Highly Voted 2 years, 7 months ago

Correct

upvoted 11 times

vernillen 2 years, 6 months ago

AgreedagreedAgreedagreedagreed

upvoted 3 times

vernillen 2 years, 6 months ago

Agreed

upvoted 3 times

efeee333 2 years, 5 months ago

agreedAgreedagreedagreed

upvoted 3 times

efeee333 2 years, 5 months ago

agreed

upvoted 3 times

Sumanth567 2 years, 5 months ago

Agreedagreedagreed

upvoted 3 times

Sumanth567 2 years, 5 months ago

Agreed

upvoted 3 times

ItsNikMon 2 years, 4 months ago

agreed

upvoted 2 times

ItsNikMon 2 years, 4 months ago

agreedagreed

upvoted 2 times

paffy 2 years ago

agreed

upvoted 2 times

seranvijay Most Recent 1 year, 9 months ago

Agreed

upvoted 1 times

KarlGardnerDataEngineering 2 years, 2 months ago

Why couldn't we just simply turn off the analytical store on Container1?

upvoted 2 times

KarlGardnerDataEngineering 2 years, 2 months ago

Why couldn't we just simply turn off the analytical store on Container1?Analytical store cannot be turned off after it has been enabled. A new container is needed to transfer

upvoted 2 times

Danweo 2 years, 1 month ago

Analytical store cannot be turned off after it has been enabled. A new container is needed to transfer

upvoted 2 times

Alongi 2 years, 4 months ago

Correct way

upvoted 1 times

Ekueaf 2 years, 5 months ago

Correct!

upvoted 1 times

Question #110 *Topic 1*

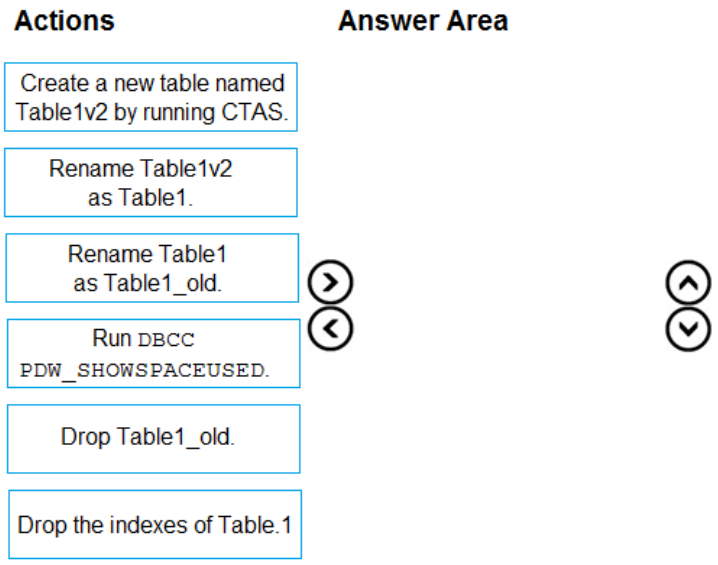
DRAG DROP

-

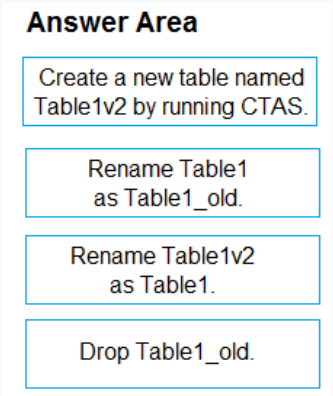
You have an Azure Synapse Analytics dedicated SQL pool named SQL1 that contains a hash-distributed fact table named Table1.

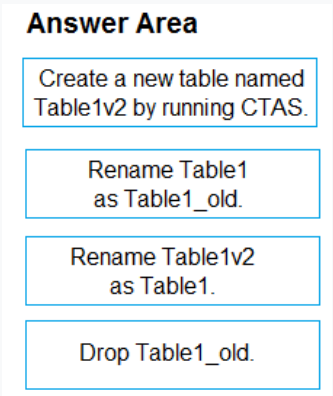
You need to recreate Table1 and add a new distribution column. The solution must maximize the availability of data.

Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



Suggested Answer:





Gman1986 Highly Voted 2 years, 1 month ago

Answers are correct. Dropping the index or running pdw_showspaceused is not required.

upvoted 8 times

Gman1986 Highly Voted 2 years, 1 month ago

Answers are correct. Dropping the index or running pdw_showspaceused is not required.Agreed!https://learn.microsoft.com/en-us/azure/search/search-howto-reindexAgreed!

upvoted 8 times

vernillen 2 years ago

Agreed!https://learn.microsoft.com/en-us/azure/search/search-howto-reindexAgreed!

upvoted 2 times

vernillen 2 years ago

Agreed!

upvoted 2 times

moneytime 2 years ago

https://learn.microsoft.com/en-us/azure/search/search-howto-reindexAgreed!

upvoted 1 times

moneytime 2 years ago

https://learn.microsoft.com/en-us/azure/search/search-howto-reindex

upvoted 1 times

efeee333 1 year, 11 months ago

Agreed!

upvoted 1 times

Sathya_sree Most Recent 1 year, 4 months ago

Create old, rename old

upvoted 1 times

Question #111 *Topic 1*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure subscription that contains an Azure data factory named ADF1.
From Azure Data Factory Studio, you build a complex data pipeline in ADF1.
You discover that the Save button is unavailable, and there are validation errors that prevent the pipeline from being published.
You need to ensure that you can save the logic of the pipeline.
Solution: You disable all the triggers for ADF1.
Does this meet the goal?

A. Yes

B. No

Suggested Answer: *B*
Official Answer: B

Community vote distribution
B (100%)

vernillen Most Recent 2 years ago
Agreed with the others: Git is needed to save the changes, and triggers have nothing to do with saving the pipeline logic. Answer is NO.
upvoted 3 times

dakku987 2 years, 1 month ago
IT NEED GIT CONFIG TO SAVE THE CHANGES
upvoted 1 times

jongert 2 years, 1 month ago
Correct, triggers have nothing to do with saving the pipeline logic.
upvoted 2 times

Question #112 *Topic 1*

You have an Azure subscription that contains the resources shown in the following table.
You need to implement Azure Synapse Link for Azure SQL Database.
Which two actions should you perform on sql1? Each correct answer presents a part of the solution.
NOTE: Each correct selection is worth one point.

Name	Type	Description
workspace1	Azure Synapse Analytics workspace	*None*
sql1	Azure SQL Database server	*None*
SQLDb1	Azure SQL database	Hosted on sql1

A. Update the firewall rules to allow Azure services to access sql1.

B. Enable the system-assigned managed identity.

C. From the Access control (IAM) settings, assign the Contributor role to the system-assigned managed identity of workspace1.

D. Disable Transparent Data Encryption (TDE).

Suggested Answer: *A, B*
Official Answer: AB

Community vote distribution
AB (71%) AC (29%)

9b71f40 Highly Voted 2 years, 2 months ago
I think the given answers are correct, Please refer the below link https://learn.microsoft.com/en-us/azure/synapse-analytics/synapse-link/connect-synapse-link-sql-database
upvoted 10 times

9b71f40 Highly Voted 2 years, 2 months ago
I think the given answers are correct, Please refer the below link https://learn.microsoft.com/en-us/azure/synapse-analytics/synapse-link/connect-synapse-link-sql-databaseIndeed it will be AB.
upvoted 10 times

Sr18 2 years, 2 months ago
Indeed it will be AB.
upvoted 4 times

17lan Most Recent 1 year, 8 months ago
Agreed
upvoted 2 times

ff5037f 1 year, 9 months ago
AB ARE correct. for reference please check https://learn.microsoft.com/en-us/azure/synapse-analytics/synapse-link/connect-synapse-link-sql-database
upvoted 3 times

babaksalarvand 1 year, 11 months ago
To implement Azure Synapse Link for Azure SQL Database, you'll need to take the following actions on sql1: A. Update the firewall rules to allow Azure services to access sql1. This is necessary because Azure Synapse Link requires the Azure Synapse workspace to be able to connect to the Azure SQL Database. Updating the firewall rules ensures that Azure services can access sql1. B. Enable the system-assigned managed identity. This is important for security and authentication. The system-assigned managed identity allows Azure Synapse to authenticate to Azure SQL Database securely without needing to store credentials. So, the correct actions are A and B.
upvoted 1 times

a85becd 1 year, 11 months ago

To implement Azure Synapse Link for Azure SQL Database, you need to perform the following actions on sql1: Update the firewall rules to allow Azure services to access sql1. Assign the Contributor role to the system-assigned managed identity of workspace1 from the Access control (IAM) settings.

upvoted 1 times

NAWRESS96 2 years ago

i think the answer is : A. Update the firewall rules to allow Azure services to access sql1. C. From the Access control (IAM) settings, assign the Contributor role to the system-assigned managed identity of workspace1.

upvoted 1 times

Question #113 *Topic 1*

You have an Azure subscription that contains an Azure Cosmos DB database. Azure Synapse Link is implemented on the database.
You configure a full fidelity schema for the analytical store.
You perform the following actions:
• Insert {"customerID": 12, "customer": "Tailspin Toys"} as the first document in the container.
• Insert {"customerID": "14", "customer": "Contoso"} as the second document in the container.
How many columns will the analytical store contain?

A. 1

B. 2

C. 3

D. 4

Suggested Answer: *B*
Official Answer: B

Community vote distribution
C (100%)

JamieMcD Highly Voted 2 years, 2 months ago
C - With a full fidelity schema, the analytical store will track both the data and their types accurately. This means different types for the same field will be stored in separate columns. Specifically: There will be one column for customerID as an integer. There will be another column for customerID as a string. There will be one column for customer as a string.The above explained is correct answer must be C. 3. 12 and "14" being different datatypes will have separate columns in full fidelity schema.Answer C
upvoted 10 times

JamieMcD Highly Voted 2 years, 2 months ago
C - With a full fidelity schema, the analytical store will track both the data and their types accurately. This means different types for the same field will be stored in separate columns. Specifically:

There will be one column for customerID as an integer. There will be another column for customerID as a string. There will be one column for customer as a string.
upvoted 10 times

Sleuth 2 years, 2 months ago
The above explained is correct answer must be C. 3. 12 and "14" being different datatypes will have separate columns in full fidelity schema.
upvoted 3 times

Sr18 2 years, 2 months ago
Answer C
upvoted 1 times

shinypriti23 Most Recent 1 year, 8 months ago
Total Columns: customerID_Number: For the Integer type in the first document. customerID_String: For the String type in the second document. customer_String: For the consistent String type. Total columns = 3
upvoted 1 times

Okea 1 year, 9 months ago
Spark will manage each datatype as a column when loading into a DataFrame https://learn.microsoft.com/en-us/azure/cosmos-db/analytical-store-introduction
upvoted 1 times

17c5d1e 1 year, 9 months ago
Whats the actual answer? I thought B as two columns will be there, CustomerID and Customer. But I guess that C could be right as a third column is added for the Text "14" ??
upvoted 1 times

NAWRESS96 2 years ago

Because the customerID field appears with different data types (integer in the first and string in the second, the analytical store will treat these as separate columns to maintain the schema's fidelity.

upvoted 1 times

Danweo 2 years, 1 month ago

C, a third column will be added for the change in datatype.

upvoted 1 times

HOTSPOT

-

You have an Azure Data Lake Storage account that contains CSV files. The CSV files contain sales order data and are partitioned by using the following format.

/data/salesorders/year=xxxx/month=y

You need to retrieve only the sales orders from January 2023 and February 2023.

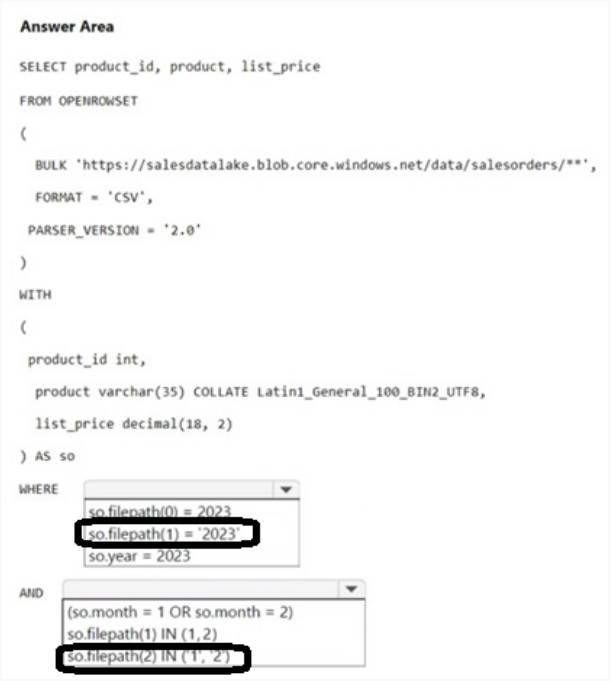
How should you complete the query? To answer, select the appropriate options in the answer area.

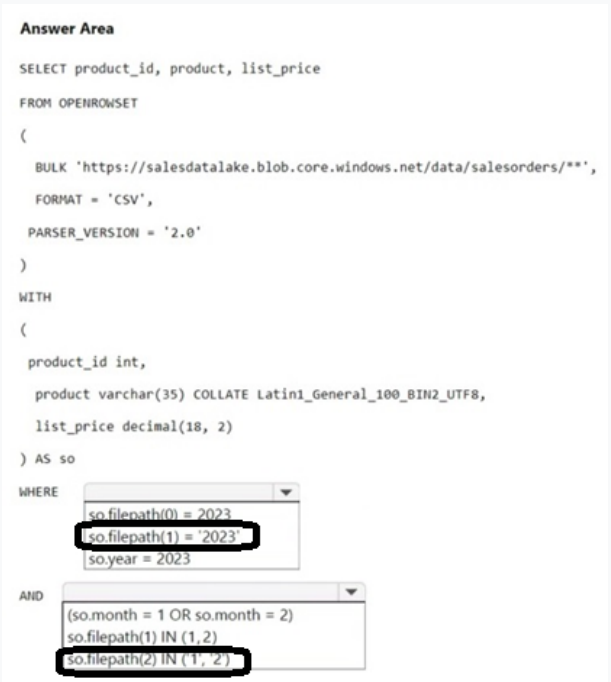
NOTE: Each correct selection is worth one point.

Answer Area

```
SELECT product_id, product, list_price
FROM OPENROWSET
(
  BULK 'https://salesdatalake.blob.core.windows.net/data/salesorders/**',
  FORMAT = 'CSV',
 PARSER_VERSION = '2.0'
)
WITH
(
 product_id int,
  product varchar(35) COLLATE Latin1_General_100_BIN2_UTF8,
  list_price decimal(18, 2)
) AS so
WHERE [ so.filepath(0) = 2023 | so.filepath(1) = '2023' | so.year = 2023 ]
AND [ (so.month = 1 OR so.month = 2) | so.filepath(1) IN (1,2) | so.filepath(2) IN ('1', '2') ]
```

Suggested Answer:





Sleuth Highly Voted 2 years, 2 months ago

Answer is correct see below link for more details on filepath() function https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-specific-files#filepath

upvoted 5 times

CezarioAbrantesPP Most Recent 2 years, 1 month ago

SELECT r.filepath() AS filepath ,r.filepath(1) AS [year] ,r.filepath(2) AS [month] ,COUNT_BIG(*) AS [rows] FROM OPENROWSET(BULK 'csv/taxi/yellow_tripdata_*-*.csv', DATA_SOURCE = 'SqlOnDemandDemo', FORMAT = 'CSV', PARSER_VERSION = '2.0', FIRSTROW = 2) WITH (vendor_id INT) AS [r] WHERE r.filepath(1) IN ('2017') AND r.filepath(2) IN ('10', '11', '12') GROUP BY r.filepath() ,r.filepath(1) ,r.filepath(2) ORDER BY filepath;

upvoted 2 times

Question #115 *Topic 1*

HOTSPOT

-

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named adlsv2. adlsv2 contains a container named logs and an Azure Synapse Analytics dedicated SQL pool named Sqlpool.

You plan to use PolyBase to load external data into Sqlpool.

You need to define the external data source.

How should you complete the script? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```
...
CREATE EXTERNAL DATA SOURCE logs
with (
  TYPE = HADOOP,
  LOCATION = [ ] [ ] .dfs.core.windows.net',
  CREDENTIAL = dbcred
);
GO
...
```

('abfss://, 'adl://, 'wasbs://) (logs@adlsv2, sqlpool@adlsv2, dbcred@adlsv2)

Suggested Answer:

Answer Area

```
...
CREATE EXTERNAL DATA SOURCE logs
with (
  TYPE = HADOOP,
  LOCATION = [adl://] [logs@adlsv2] .dfs.core.windows.net',
  CREDENTIAL = dbcred
);
GO
...
```

dbcred@adlsv2

Answer Area

```
...
CREATE EXTERNAL DATA SOURCE logs
with (
  TYPE = HADOOP,
  LOCATION = [adl://] [logs@adlsv2] .dfs.core.windows.net',
  CREDENTIAL = dbcred
);
GO
...
```

dbcred@adlsv2

Rohit_0 Most Recent 1 year, 5 months ago

adl:// is the old protocol used by the Azure Data Lake Store (ADLS) Gen1. It was primarily used for accessing Azure Data Lake Store Gen1. So it should be abfss://

upvoted 1 times

Pey1nkh 1 year, 6 months ago

should be abfss:// ---> (Azure Blob File System Secure)

upvoted 2 times

phant0mw0lf 1 year, 6 months ago

abfss ist correct due to the fact that it is a gen2 storage account and .dfs.core.windows.net is used as the location

upvoted 1 times

Question #116 *Topic 1*

You are developing a solution that will stream to Azure Stream Analytics. The solution will have both streaming data and reference data.
Which input type should you use for the reference data?

A. Azure Service Bus

B. Azure Blob storage

C. Azure IoT Hub

D. Azure Event Hubs

Suggested Answer: *B*
Official Answer: B

Community vote distribution
B (100%)

HA29 Most Recent 1 year, 4 months ago
Blob storage and SQL Database are valid reference types
upvoted 1 times

Pey1nkh 1 year, 6 months ago
Azure Stream Analytics supports reference data from Azure Blob Storage, and the reference data is typically stored in CSV, JSON, or Parquet format.
upvoted 2 times

Question #117 *Topic 1*

You have an Azure subscription that contains an Azure Synapse Analytics workspace and a user named User1.
You need to ensure that User1 can create a new lake database by using an Azure Synapse database template from Gallery. The solution must follow the principle of least privilege.
Which role should you assign to User1?

A. Synapse Contributor

B. Synapse Administrator

C. Synapse User

D. Storage Blob Data Contributor

Suggested Answer: *A*
Official Answer: A

Community vote distribution
A (100%)

Moli62 Most Recent 1 year, 5 months ago
To allow User1 to create a lake database using an Azure Synapse database template from the Gallery, the user needs permission to manage Synapse resources, including databases. Why Synapse Contributor? The Synapse Contributor role allows users to create and manage Synapse resources, including lake databases. It follows the principle of least privilege by granting only the necessary permissions without giving full administrative control.
upvoted 1 times

Question #118 *Topic 1*

HOTSPOT

-

You are designing the dimensions for a data warehouse in an Azure Synapse Analytics dedicated SQL pool.

You build a table by using the statement shown in the following exhibit.

Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the exhibit.

NOTE: Each correct selection is worth one point.

```
Create Table [dbo].[DimSalesRep] (
	SalesRepID bigint IDENTITY(1,1) NOT NULL,
	RepSourceID bigint NOT NULL,
	FirstName nvarchar(100) NOT NULL,
	LastName nvarchar(100) NOT NULL,
	Region nvarchar(50) NOT NULL,
	EffectiveStart datetime NOT NULL,
	EffectiveEnd datetime NOT NULL,
	IsCurrentRow bit NOT NULL,
	ETLAuditID bigint NOT NULL
)
```

Answer Area

DimSalesRep is a **[answer choice]** slowly changing dimension (SCD).

- Type 1
- Type 2
- Type 3

SalesRepID is **[answer choice]**.

- a business key
- a surrogate key
- an audit column

Suggested Answer:

Answer Area

DimSalesRep is a **[answer choice]** slowly changing dimension (SCD). — Type 2

SalesRepID is **[answer choice]**. — a surrogate key

Answer Area

DimSalesRep is a **[answer choice]** slowly changing dimension (SCD). — Type 2

SalesRepID is **[answer choice]**. — a surrogate key

Podavenna Highly Voted 1 month, 1 week ago

Correct solution!

upvoted 52 times

MrWood47 Highly Voted 1 month, 1 week ago

Answers provided are correct! The input type for the Stream Analytics job should be Stream, as it will be processing real-time data from devices. The function to include in the Stream Analytics job should be Geospatial, which allows you to perform calculations on geographic data and make spatial queries, such as determining the distance between two points. This is necessary to determine if a device has traveled more than 200 meters away from a designated location.

upvoted 19 times

Amnoyana 1 month, 1 week ago

Good application and need to learn more about it!

upvoted 4 times

Deeksha1234 1 month, 1 week ago

solution is correct

upvoted 4 times

dheeraj7654 Most Recent 1 month, 1 week ago

Correct

upvoted 3 times

Question #119 *Topic 1*

HOTSPOT

-

You have an Azure subscription that contains an Azure Synapse Analytics serverless SQL pool.

You run the following query in the pool.

```
SELECT *
FROM OPENROWSET(BULK 'https://mydatalake.blob.core.windows.net/data/**',
FORMAT = 'CSV') AS datarows;
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Answer Area

Statements	Yes	No
The query will create a table named `datarows`.	○	○
The `https://mydatalake.blob.core.windows.net/data/file1` file will be read.	○	○
The `https://mydatalake.blob.core.windows.net/data/folder1/file1.csv` file will be read.	○	○

Suggested Answer:

Answer Area

Statements	Yes	No
The query will create a table named `datarows`.	○	○
The `https://mydatalake.blob.core.windows.net/data/file1` file will be read.	○	○
The `https://mydatalake.blob.core.windows.net/data/folder1/file1.csv` file will be read.	○	○

Answer Area

Statements	Yes	No
The query will create a table named `datarows`.	○	○
The `https://mydatalake.blob.core.windows.net/data/file1` file will be read.	○	○
The `https://mydatalake.blob.core.windows.net/data/folder1/file1.csv` file will be read.	○	○

Podavenna Highly Voted 1 month, 1 week ago

Correct solution!

upvoted 52 times

MrWood47 Highly Voted 1 month, 1 week ago

Answers provided are correct! The input type for the Stream Analytics job should be Stream, as it will be processing real-time data from devices. The function to include in the Stream Analytics job should be Geospatial, which allows you to perform calculations on geographic data and make spatial queries, such as determining the distance between two points. This is necessary to determine if a device has traveled more than 200 meters away from a designated location.

upvoted 19 times

Amnoyana 1 month, 1 week ago

Good application and need to learn more about it!

upvoted 4 times

Deeksha1234 Most Recent 1 month, 1 week ago

solution is correct

upvoted 4 times

dheeraj7654 1 month, 1 week ago

Correct

upvoted 3 times

Question #1 *Topic 2*

HOTSPOT -

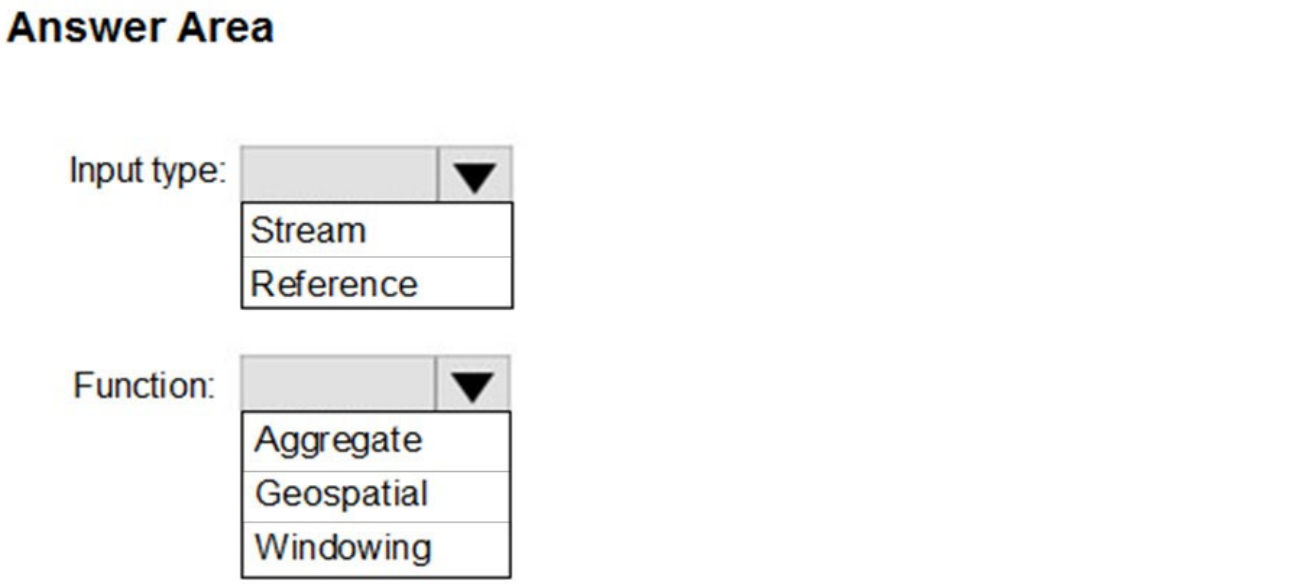
You plan to create a real-time monitoring app that alerts users when a device travels more than 200 meters away from a designated location.

You need to design an Azure Stream Analytics job to process the data for the planned app. The solution must minimize the amount of code developed and the number of technologies used.

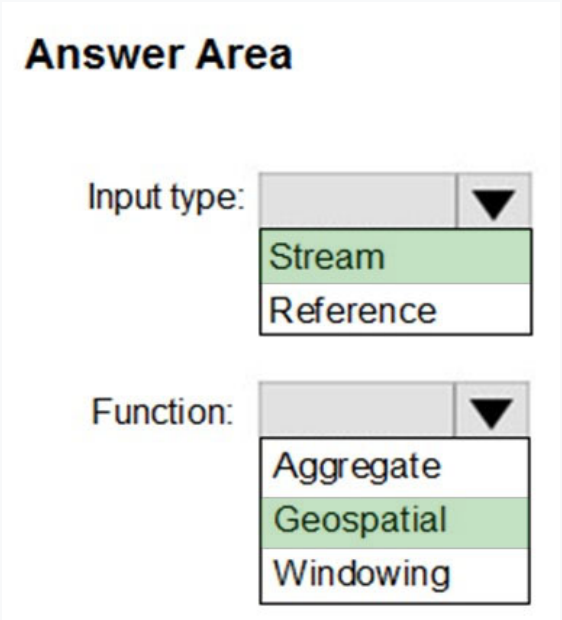
What should you include in the Stream Analytics job? To answer, select the appropriate options in the answer area.

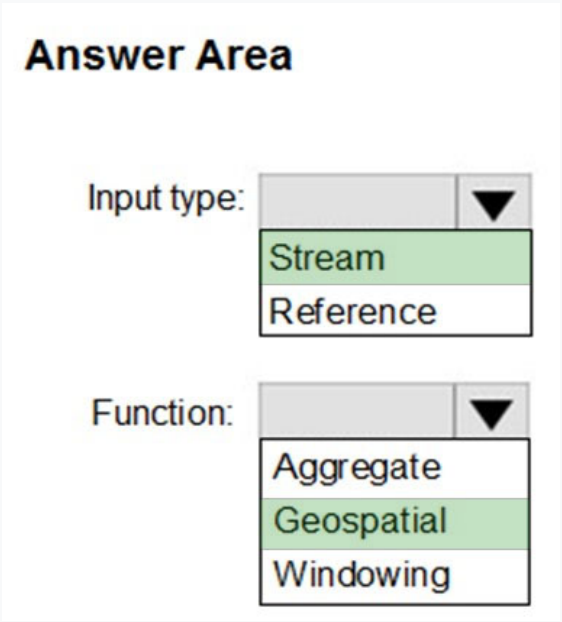
NOTE: Each correct selection is worth one point.

Hot Area:



Suggested Answer:





Input type: Stream - You can process real-time IoT data streams with Azure Stream Analytics.

Function: Geospatial - With built-in geospatial functions, you can use Azure Stream Analytics to build applications for scenarios such as fleet management, ride sharing, connected cars, and asset tracking. Note: In a real-world scenario, you could have hundreds of these sensors generating events as a stream. Ideally, a gateway device would run code to push these events to Azure Event Hubs or Azure IoT Hubs.
Reference: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-get-started-with-azure-stream-analytics-to-process-data-from-iot-devices https://docs.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios

Podavenna Highly Voted 4 years, 5 months ago
Correct solution!
upvoted 52 times

Podavenna Highly Voted 4 years, 5 months ago
Correct solution!Agree!
upvoted 52 times

jpgsa11 Most Recent 1 year, 9 months ago
Agree!
upvoted 1 times

MrWood47 3 years, 1 month ago

Answers provided are correct! The input type for the Stream Analytics job should be Stream, as it will be processing real-time data from devices. The function to include in the Stream Analytics job should be Geospatial, which allows you to perform calculations on geographic data and make spatial queries, such as determining the distance between two points. This is necessary to determine if a device has traveled more than 200 meters away from a designated location.Very clear ,thank you for the explanation !

upvoted 19 times

MrWood47 3 years, 1 month ago

Answers provided are correct! The input type for the Stream Analytics job should be Stream, as it will be processing real-time data from devices. The function to include in the Stream Analytics job should be Geospatial, which allows you to perform calculations on geographic data and make spatial queries, such as determining the distance between two points. This is necessary to determine if a device has traveled more than 200 meters away from a designated location.

upvoted 19 times

Little_Soap 1 year, 10 months ago

Very clear ,thank you for the explanation !

upvoted 1 times

Dusica 1 year, 9 months ago

Stream it is But streaming must use windowing function, I think

upvoted 1 times

kkk5566 2 years, 5 months ago

1-stream 2-geospatial

upvoted 1 times

akhil5432 2 years, 6 months ago

1-stream 2-geospatial

upvoted 2 times

hiyoww 2 years, 6 months ago

this is microsoft link for reference: https://learn.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios

upvoted 1 times

dheeraj7654 2 years, 8 months ago

Correct

upvoted 3 times

Ankit_Az 2 years, 8 months ago

Stream & Geospatial is Correct

upvoted 2 times

rocky48 2 years, 9 months ago

Stream & Geospatial

upvoted 2 times

hanzocuk 3 years, 1 month ago

I doubt that given solution is correct. No reason to stream the designated location, that is used for lookup. Think of it as a dimension table. Input type: Reference Function: Geospatial

upvoted 2 times

vigilante89 3 years, 2 months ago

Input Type: Stream Function: Geospatial

upvoted 1 times

Amnoyana 3 years, 3 months ago

Good application and need to learn more about it!

upvoted 4 times

Deeksha1234 3 years, 6 months ago

solution is correct

upvoted 4 times

ClassMistress 3 years, 9 months ago

Correct

upvoted 2 times

NewTuanAnh 3 years, 10 months ago

Correct!

upvoted 2 times

PallaviPatel 4 years ago

Correct

upvoted 2 times

Question #2 *Topic 2*

A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream
Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU).
You need to optimize performance for the Azure Stream Analytics job.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Implement event ordering.

B. Implement Azure Stream Analytics user-defined functions (UDF).

C. Implement query parallelization by partitioning the data output.

D. Scale the SU count for the job up.

E. Scale the SU count for the job down.

F. Implement query parallelization by partitioning the data input.

Suggested Answer: *D, F*
D: Scale out the query by allowing the system to process each input partition separately. F: A Stream Analytics job definition includes inputs, a query, and output. Inputs are where the job reads the data stream from. Reference: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization

Community vote distribution
CF (50%) DF (38%) CD (9%) (3

manquak Highly Voted 4 years, 11 months ago
Partition input and output. REF: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization
upvoted 70 times

manquak Highly Voted 4 years, 11 months ago
Partition input and output. REF: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelizationAgree. And partitioning Input and output with same number of partitions gives the best performance optimization..
upvoted 70 times

kolakone 4 years, 11 months ago
Agree. And partitioning Input and output with same number of partitions gives the best performance optimization..
upvoted 14 times

Lio95 4 years, 11 months ago
No event consumer was mentioned. Therefore, partitioning output is not relevant. Answer is correctThe stream analytics job is the consumer.Stream analytics ALWAYS has at least one output. There is no need to mention that. So correct answer is input and output
upvoted 15 times

Lio95 4 years, 11 months ago
No event consumer was mentioned. Therefore, partitioning output is not relevant. Answer is correct
upvoted 15 times

Boompiee 4 years, 3 months ago
The stream analytics job is the consumer.
upvoted 2 times

nicolas1999 4 years, 9 months ago
Stream analytics ALWAYS has at least one output. There is no need to mention that. So correct answer is input and output
upvoted 4 times

sakis213 Most Recent 1 year, 6 months ago
Partitioning the output data can improve write performance or manage how results are distributed, but it doesn't directly impact the performance of data ingestion or processing within the job itself.
upvoted 1 times

GiuseppeTanda 1 year, 7 months ago

As suggested in the following link: https://techcommunity.microsoft.com/blog/analyticsonazure/optimize-your-stream-analytics-job%E2%80%99s-performance-using-job-diagram-simulator/3652303 "One way to optimize a Stream Analytics job's performance is to leverage parallelism in query" and "For a job to be parallel, you need to align partition keys between all inputs, query steps, and outputs"

upvoted 2 times

de_examtopics 1 year, 8 months ago

C. Implement query parallelization by partitioning the data output. By partitioning the data output, you can enable query parallelization. This allows the queries to be distributed across multiple partitions, which can significantly enhance performance. F. Implement query parallelization by partitioning the data input.

upvoted 1 times

Danweo 2 years, 1 month ago

Partitioning on both input and output can help, but we don't know if the output is a service that doesn't support partitioning like Power BI. Scaling up will always assign more resources at least.

upvoted 1 times

e56bb91 2 years, 1 month ago

ChatGPT 4o C. Implement query parallelization by partitioning the data output: Output Partitioning: By partitioning the data output, you can ensure that the processing load is distributed evenly across multiple nodes, which can significantly improve performance by reducing bottlenecks in data writing. F. Implement query parallelization by partitioning the data input: Input Partitioning: Partitioning the data input allows the Stream Analytics job to process different partitions in parallel, leading to better utilization of the available streaming units and improved throughput.

upvoted 2 times

Dusica 2 years, 3 months ago

C and F > same partitions > embarrassingly parallel processing

upvoted 1 times

Dusica 2 years, 3 months ago

It says optimize performance, does not say that it is bad so adding SU may be unneccesary cost increase. Parallelization and embarrassingly parallel job is correct

upvoted 1 times

Bhargava12 2 years, 4 months ago

Answer is D & F

upvoted 1 times

Elanche 2 years, 5 months ago

D. Scale the SU count for the job up: Increasing the number of Streaming Units (SUs) can improve the performance of the Stream Analytics job by providing more processing power to handle the incoming data stream. C. Implement query parallelization by partitioning the data output: Partitioning the data output can help distribute the processing load across multiple partitions, allowing for parallel execution of queries and enhancing performance.

upvoted 1 times

Alongi 2 years, 6 months ago

C and D

upvoted 1 times

prshntdxt7 2 years, 6 months ago

C. Implement query parallelization by partitioning the data output: "Partitioning lets you divide data into subsets based on a partition key. If your input (for example Event Hubs) is partitioned by a key, it's highly recommended to specify this partition key when adding input to your Stream Analytics job. Scaling a Stream Analytics job takes advantage of partitions in the input and output. A Stream Analytics job can consume and write different partitions in parallel, which increases throughput." D. Scale the SU count for the job up: "The total number of streaming units that can be used by a Stream Analytics job depends on the number of steps in the query defined for the job and the number of partitions for each step... All non-partitioned steps together can scale up to one streaming unit (SU V2s) for a Stream Analytics job. In addition, you can add 1 SU V2 for each partition in a partitioned step."

upvoted 1 times

sdg2844 2 years, 7 months ago

As there is no indication of any query parallelization currently, we have to choose to parallelize for both input and output as the first/correct answers.

upvoted 2 times

Khadija10 2 years, 7 months ago

Partitioning lets you divide data into subsets based on a partition key. If your input (for example Event Hubs) is partitioned by a key, it's highly recommended to specify this partition key when adding input to your Stream Analytics job. Scaling a Stream Analytics job takes advantage of partitions in the input and output. A Stream Analytics job can consume and write different partitions in parallel, which increases throughput. Ref: https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization

upvoted 2 times

jongert 2 years, 7 months ago

An embarrassingly parallel job allows the highest degree of parallelization. Looking at how the max number of stream units is calculated, it would not be useful to scale them up if you keep a bottleneck at the output. Unsure what a good reference value would be for the number of SUs, but 120 does not seem very low to me.

upvoted 2 times

d046bc0 2 years, 8 months ago

Scale the SU count for the job up - (ChatGPT) This will not necessarily improve the performance of your job, unless your query is CPU-bound or memory-bound. Scaling up the SU count will increase the amount of resources available for your job, but it will also increase the cost. You should first try to optimize your query by using parallelization and repartitioning techniques, and then scale up the SU count only if needed1

upvoted 1 times

Question #3 *Topic 2*

You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container.
Which resource provider should you enable?

A. Microsoft.Sql

B. Microsoft.Automation

C. Microsoft.EventGrid

D. Microsoft.EventHub

Suggested Answer: *C*
Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure Blob Storage account. Data Factory natively integrates with Azure Event Grid, which lets you trigger pipelines on such events. Reference: https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers

Community vote distribution
C (100%)

jv2120 Highly Voted 4 years, 8 months ago
Correct. C Azure Event Grids – Event-driven publish-subscribe model (think reactive programming) Azure Event Hubs – Multiple source big data streaming pipeline (think telemetry data) In this case its more suitable vs Event Hubs.
upvoted 34 times

medsimus Highly Voted 4 years, 10 months ago
Correct https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger?tabs=data-factory
upvoted 12 times

GiuseppeTanda Most Recent 1 year, 7 months ago
https://learn.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger?tabs=data-factory
upvoted 1 times

CezarioAbrantesPP 2 years, 1 month ago
Data Factory and Synapse pipelines natively integrate with Azure Event Grid, which lets you trigger pipelines on such events. the arrival or deletion of a file in Azure Blob Storage account
upvoted 1 times

lisa710 2 years, 7 months ago
c is correct
upvoted 1 times

kkk5566 2 years, 11 months ago
custom trigger use Event-grid
upvoted 1 times

kkk5566 2 years, 11 months ago
Correct. C
upvoted 1 times

kkk5566 2 years, 11 months ago
Correct
upvoted 1 times

Ankit_Az 3 years, 2 months ago
Correct
upvoted 2 times

esaade 3 years, 5 months ago
To trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container, you should enable the Microsoft.EventGrid resource provider. Microsoft.EventGrid is an event-based publish/subscribe service that allows you to easily route events between different Azure

services. By subscribing to the blob-created event in an Azure Data Lake Storage Gen2 container, you can trigger an Azure Data Factory pipeline whenever a file arrives in the container. Therefore, the correct answer is C. Microsoft.EventGrid.

upvoted 2 times

vigilante89 3 years, 8 months ago

Usually while triggering an event using ADF, there is event-based trigger. Apart from that, ADF is well integrated with Azure Event Grid, which lets us trigger pipelines on an event.

upvoted 2 times

Selma97 3 years, 9 months ago

Why it's not Microsoft.Automation?Would that be part of the Power Platform instead?

upvoted 1 times

Selma97 3 years, 9 months ago

Why it's not Microsoft.Automation?

upvoted 1 times

Amnoyana 3 years, 9 months ago

Would that be part of the Power Platform instead?

upvoted 1 times

Deeksha1234 4 years ago

correct

upvoted 1 times

PallaviPatel 4 years, 6 months ago

Correct.

upvoted 2 times

romanzdk 4 years, 7 months ago

But EventHub does not support ADLS, only Blob storage

upvoted 1 times

romanzdk 4 years, 7 months ago

But EventHub does not support ADLS, only Blob storagehttps://docs.microsoft.com/en-us/azure/event-grid/overview

upvoted 1 times

romanzdk 4 years, 7 months ago

https://docs.microsoft.com/en-us/azure/event-grid/overview

upvoted 2 times

Swagat039 4 years, 7 months ago

C. is correct. You need storage event trigger (for this Microsoft.EventGrid service needs to be enabled).

upvoted 1 times

Vardhan_Brahmanapally 4 years, 9 months ago

Why not eventhub?

upvoted 4 times

Question #4 *Topic 2*

You plan to perform batch processing in Azure Databricks once daily.
Which type of Databricks cluster should you use?

A. High Concurrency

B. automated

C. interactive

Suggested Answer: *B*
Automated Databricks clusters are the best for jobs and automated batch processing. Note: Azure Databricks has two types of clusters: interactive and automated. You use interactive clusters to analyze data collaboratively with interactive notebooks. You use automated clusters to run fast and robust automated jobs. Example: Scheduled batch workloads (data engineers running ETL jobs) This scenario involves running batch job JARs and notebooks on a regular cadence through the Databricks platform. The suggested best practice is to launch a new cluster for each run of critical jobs. This helps avoid any issues (failures, missing SLA, and so on) due to an existing workload (noisy neighbor) on a shared cluster. Reference: https://docs.microsoft.com/en-us/azure/databricks/clusters/create https://docs.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-scheduled-batch-workloads-data-engineers-running-etl-jobs

Community vote distribution
B (100%)

Podavenna 3 years, 11 months ago
Correct!
upvoted 16 times

kkk5566 Most Recent 1 year, 11 months ago
Automatic/Jobs - best for jobs and automated batch processing.
upvoted 1 times

Ankit_Az 2 years, 2 months ago
Given Answer and explanation is correct
upvoted 2 times

Lestrang 2 years, 6 months ago
Databricks makes a distinction between all-purpose clusters and job clusters. You use all-purpose clusters to analyze data collaboratively using interactive notebooks. You use job clusters to run fast and robust automated jobs. You can create an all-purpose cluster using the UI, CLI, or REST API. You can manually terminate and restart an all-purpose cluster. Multiple users can share such clusters to do collaborative interactive analysis. The Databricks job scheduler creates a job cluster when you run a job on a new job cluster and terminates the cluster when the job is complete. You cannot restart a job cluster.
upvoted 4 times

NintyFour 2 years, 10 months ago
https://learn.microsoft.com/en-us/azure/databricks/clusters/configure as per above link: Datablocks has 3 modes of cluster
upvoted 2 times

Deeksha1234 3 years ago
right answer
upvoted 1 times

HebaN 3 years, 2 months ago
Its correct answer
upvoted 2 times

necktru 3 years, 3 months ago
correct
upvoted 1 times

PallaviPatel 3 years, 6 months ago
correct.
upvoted 1 times

satyamkishoresingh 3 years, 11 months ago

What is automated cluster ?

upvoted 2 times

satyamkishoresingh 3 years, 11 months ago

What is automated cluster ?Job clusterThere are 2 types of databricks clusters: 1) Standard/Interactive - best for querying and processing data by users. 2) Automatic/Jobs - best for jobs and automated batch processing.hi, may I ask where I can find these words?

upvoted 2 times

Swagat039 3 years, 7 months ago

Job cluster

upvoted 3 times

wijaz789 Highly Voted 3 years, 11 months ago

There are 2 types of databricks clusters: 1) Standard/Interactive - best for querying and processing data by users. 2) Automatic/Jobs - best for jobs and automated batch processing.hi, may I ask where I can find these words?

upvoted 20 times

wijaz789 Highly Voted 3 years, 11 months ago

There are 2 types of databricks clusters: 1) Standard/Interactive - best for querying and processing data by users. 2) Automatic/Jobs - best for jobs and automated batch processing.

upvoted 20 times

youngbug 3 years ago

hi, may I ask where I can find these words?

upvoted 2 times

Question #5 *Topic 2*

HOTSPOT -

You are processing streaming data from vehicles that pass through a toll booth.

You need to use Azure Stream Analytics to return the license plate, vehicle make, and hour the last vehicle passed during each 10-minute window.

How should you complete the query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
WITH LastInWindow AS
(
        SELECT
                [ ▼ ] (Time) AS LastEventTime
                  COUNT
                  MAX
                  MIN
                  TOPONE
        FROM
                Input TIMESTAMP BY Time
        GROUP BY
                [ ▼ ] (minute, 10)
                  HoppingWindow
                  SessionWindow
                  SlidingWindow
                  TumblingWindow
)
SELECT
        Input.License_plate,
        Input.Make,
        Input.Time
FROM
        Input TIMESTAMP BY Time
        INNER JOIN LastInWindow
        ON [ ▼ ] (minute, Input, LastInWindow) BETWEEN 0 AND 10
                  DATEADD
                  DATEDIFF
                  DATENAME
                  DATEPART
AND Input.Time = LastInWindow.LastEventTime
```

Suggested Answer:

Answer Area

```
WITH LastInWindow AS
(
        SELECT
                [ COUNT | MAX | MIN | TOPONE ] (Time) AS LastEventTime
        FROM
                Input TIMESTAMP BY Time
        GROUP BY
                [ HoppingWindow | SessionWindow | SlidingWindow | TumblingWindow ] (minute, 10)
)
SELECT
        Input.License_plate,
        Input.Make,
        Input.Time
FROM
        Input TIMESTAMP BY Time
        INNER JOIN LastInWindow
        ON [ DATEADD | DATEDIFF | DATENAME | DATEPART ] (minute, Input, LastInWindow) BETWEEN 0 AND 10
AND Input.Time = LastInWindow.LastEventTime
```

Answer Area

```
WITH LastInWindow AS
(
        SELECT
                [ COUNT | MAX | MIN | TOPONE ] (Time) AS LastEventTime
        FROM
                Input TIMESTAMP BY Time
        GROUP BY
                [ HoppingWindow | SessionWindow | SlidingWindow | TumblingWindow ] (minute, 10)
)
SELECT
        Input.License_plate,
        Input.Make,
        Input.Time
FROM
        Input TIMESTAMP BY Time
        INNER JOIN LastInWindow
        ON [ DATEADD | DATEDIFF | DATENAME | DATEPART ] (minute, Input, LastInWindow) BETWEEN 0 AND 10
AND Input.Time = LastInWindow.LastEventTime
```

Box 1: MAX - The first step on the query finds the maximum time stamp in 10-minute windows, that is the time stamp of the last event for that window. The second step joins the results of the first query with the original stream to find the event that match the last time stamps in each window. Query:

WITH LastInWindow AS - (

SELECT -

MAX(Time) AS LastEventTime -

FROM -

Input TIMESTAMP BY Time -

GROUP BY - TumblingWindow(minute, 10))

SELECT - Input.License_plate, Input.Make,

Input.Time -

FROM -

Input TIMESTAMP BY Time -

INNER JOIN LastInWindow - ON DATEDIFF(minute, Input, LastInWindow) BETWEEN 0 AND 10 AND Input.Time = LastInWindow.LastEventTime

Box 2: TumblingWindow - Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

Box 3: DATEDIFF - DATEDIFF is a date-specific function that compares and returns the time difference between two DateTime fields, for more information, refer to date functions. Reference: https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

rikku33 Highly Voted 4 years, 4 months ago

correct

upvoted 31 times

rikku33 Highly Voted 4 years, 4 months ago

correctWhy not Hopping Window??Because a hopping window can overlap, and we need the data from 10 minute time frames that DON'T overlapit needs 3 parameters in input.

upvoted 31 times

Jerrylolu 4 years, 2 months ago

Why not Hopping Window??Because a hopping window can overlap, and we need the data from 10 minute time frames that DON'T overlapit needs 3 parameters in input.

upvoted 1 times

Jerrylolu 4 years, 2 months ago

Why not Hopping Window??

upvoted 1 times

Wijn4nd 4 years, 1 month ago

Because a hopping window can overlap, and we need the data from 10 minute time frames that DON'T overlap

upvoted 12 times

auwia 2 years, 7 months ago

it needs 3 parameters in input.

upvoted 5 times

GodfreyMbizo 3 years ago

answer is here https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns#return-the-last-event-in-a-window

upvoted 20 times

GodfreyMbizo 3 years ago

answer is here https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns#return-the-last-event-in-a-windowYou are great!The referenced link shows the provided answer is correct

upvoted 20 times

goldy29 2 years, 7 months ago

You are great!

upvoted 1 times

ExamDestroyer69 2 years, 2 months ago

The referenced link shows the provided answer is correct

upvoted 2 times

sdg2844 2 years, 1 month ago

I agree with the person who said this says to define the last event in that HOUR for the 10 minute window increments. So it should only return one value. This will return more than one value, as it returns the last values in EACH 10 minute window, not the maximum timestamp for a defined HOUR.

upvoted 1 times

blazy001 2 years, 2 months ago

error in code: wrong: DATEDIFF(minut, Input, LastInWindow) correct: DATEDIFF(minut, Input, LastEventTime)Not according to Microsoft: https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns#return-the-last-event-in-a-window Special thanks to @GodfreyMbizo

upvoted 3 times

blazy001 2 years, 2 months ago

error in code: wrong: DATEDIFF(minut, Input, LastInWindow) correct: DATEDIFF(minut, Input, LastEventTime)

upvoted 3 times

lola_mary5 Most Recent 1 year, 11 months ago

Not according to Microsoft: https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns#return-the-last-event-in-a-window Special thanks to @GodfreyMbizo

upvoted 2 times

gggqqqqq 2 years, 4 months ago

DATEDIFF used in the SELECT statement uses the general syntax where a datetime column or expression is passed in as the second and third parameter. However, when the DATEDIFF function is used inside the JOIN condition, the input_source name or its alias is used. Internally the timestamp associated for each event in that source is picked.

upvoted 1 times

kkk5566 2 years, 5 months ago

correct

upvoted 1 times

mamahani 2 years, 9 months ago

max / tumblingwindow/datediff

upvoted 3 times

AZLearn111 3 years ago

May be a dumb question but why the datediff condition when the other condition is exactly matching on timestamp. Is it not unnecessary?

upvoted 6 times

AZLearn111 3 years ago

May be a dumb question but why the datediff condition when the other condition is exactly matching on timestamp. Is it not unnecessary?exactly my thought too; we already have the unique timestamp per 10 min windows, so why simply not match the car's (event) timestamp with the max? what is the value added of the datediffThe reason for using DATEDIFF in this context is to create a condition for joining the two streams based on a time window. If you directly join on Input.Time = LastInWindow.LastEventTime, you are essentially checking for an exact match of timestamps. This might not capture events that are close to each other in time but are not exactly equal. By using DATEDIFF, you allow for a time range (0 to 10 minutes) within which events will be considered as part of the same window. This ensures that events occurring slightly before or after the last event in the window are included in the result.it's AND Input.Time = LastInWindow.LastEventTime...i'm thinking this condition would invalidate anything that's only close in time, but not an exact match

upvoted 6 times

mamahani 2 years, 10 months ago

exactly my thought too; we already have the unique timestamp per 10 min windows, so why simply not match the car's (event) timestamp with the max? what is the value added of the datediff

upvoted 2 times

alphilla 2 years, 2 months ago

The reason for using DATEDIFF in this context is to create a condition for joining the two streams based on a time window. If you directly join on Input.Time = LastInWindow.LastEventTime, you are essentially checking for an exact match of timestamps. This might not capture events that are close to each other in time but are not exactly equal.

By using DATEDIFF, you allow for a time range (0 to 10 minutes) within which events will be considered as part of the same window. This ensures that events occurring slightly before or after the last event in the window are included in the result.

upvoted 1 times

alphilla 2 years, 2 months ago

The reason for using DATEDIFF in this context is to create a condition for joining the two streams based on a time window. If you directly join on Input.Time = LastInWindow.LastEventTime, you are essentially checking for an exact match of timestamps. This might not capture events that are close to each other in time but are not exactly equal. By using DATEDIFF, you allow for a time range (0 to 10 minutes) within which events will be considered as part of the same window. This ensures that events occurring slightly before or after the last event in the window are included in the result.it's AND Input.Time = LastInWindow.LastEventTime...i'm thinking this condition would invalidate anything that's only close in time, but not an exact match

upvoted 1 times

meatpoof 2 years, 1 month ago

it's AND Input.Time = LastInWindow.LastEventTime...i'm thinking this condition would invalidate anything that's only close in time, but not an exact match

upvoted 1 times

Bro111 3 years, 2 months ago

Are "Input" and "LastInWindow" DateTime fields, to be comapred with datediff???I have the answer here: https://learn.microsoft.com/en-us/stream-analytics-query/join-azure-stream-analytics

upvoted 2 times

Bro111 3 years, 2 months ago

Are "Input" and "LastInWindow" DateTime fields, to be comapred with datediff???

upvoted 2 times

Bro111 3 years, 2 months ago

I have the answer here: https://learn.microsoft.com/en-us/stream-analytics-query/join-azure-stream-analytics

upvoted 4 times

anks84 3 years, 5 months ago

CORRECT

upvoted 1 times

Deeksha1234 3 years, 6 months ago

correct

upvoted 2 times

y203 3 years, 6 months ago

The full example with the answer is here: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns#return-the-last-event-in-a-window

upvoted 3 times

Sriramiyer92 3 years, 7 months ago

Correct! The keyword "each" in "last vehicle passed during each 10-minute window" pretty much makes it Clear!

upvoted 1 times

Revave2 3 years, 7 months ago

I understand the first two, but why datediff? They ask for the hour that the last vehicle went through, shouldn't that be datepart?In order to match the correct inputs with the last event in window, the difference between both times should not exceed 10 minutes.

upvoted 3 times

Revave2 3 years, 7 months ago

I understand the first two, but why datediff? They ask for the hour that the last vehicle went through, shouldn't that be datepart?

upvoted 3 times

sensaint 3 years, 3 months ago

In order to match the correct inputs with the last event in window, the difference between both times should not exceed 10 minutes.

upvoted 2 times

PallaviPatel 4 years ago

correct.

upvoted 1 times

BusinessApps 4 years ago

HoppingWindow has a minimum of three arguments whereas TumblingWindow only takes two so considering the solution only has two arguments it has to be Tumbling https://docs.microsoft.com/en-us/stream-analytics-query/hopping-window-azure-stream-analytics https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

upvoted 4 times

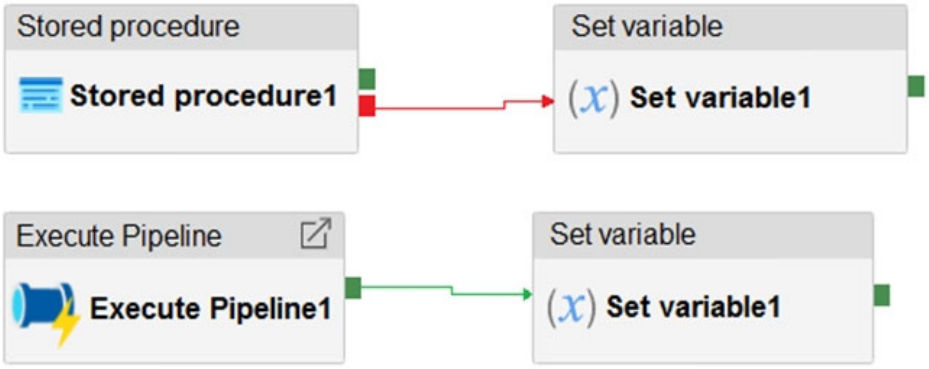
DrTaz 4 years, 1 month ago

Answer is 100% correct.

upvoted 2 times

Question #6 *Topic 2*

You have an Azure Data Factory instance that contains two pipelines named Pipeline1 and Pipeline2.
Pipeline1 has the activities shown in the following exhibit.
Pipeline2 has the activities shown in the following exhibit.
You execute Pipeline2, and Stored procedure1 in Pipeline1 fails.
What is the status of the pipeline runs?



A. Pipeline1 and Pipeline2 succeeded.

B. Pipeline1 and Pipeline2 failed.

C. Pipeline1 succeeded and Pipeline2 failed.

D. Pipeline1 failed and Pipeline2 succeeded.

Suggested Answer: *A*



Activities are linked together via dependencies. A dependency has a condition of one of the following: Succeeded, Failed, Skipped, or Completed. Consider Pipeline1: If we have a pipeline with two activities where Activity2 has a failure dependency on Activity1, the pipeline will not fail just because Activity1 failed. If Activity1 fails and Activity2 succeeds, the pipeline will succeed. This scenario is treated as a try-catch block by Data Factory.

The failure dependency means this pipeline reports success. Note: If we have a pipeline containing Activity1 and Activity2, and Activity2 has a success dependency on Activity1, it will only execute if Activity1 is successful. In this scenario, if Activity1 fails, the pipeline will fail. Reference: https://datasavvy.me/category/azure-data-factory/

Community vote distribution

A (87%) B (13%)

SaferSephy Highly Voted 4 years, 11 months ago

Correct answer is A. The trick is the fact that pipeline 1 only has a Failure dependency between de activity's. In this situation this results in a Succeeded pipeline if the Stored procedure failed. If also the success connection was linked to a follow up activity, and the SP would fail, the pipeline would be indeed marked as failed. So A.well explained! A is rightcorrect ,Execute pipeline explained here https://youtu.be/Jkz1dtLrBE4

upvoted 51 times

SaferSephy Highly Voted 4 years, 11 months ago

Correct answer is A. The trick is the fact that pipeline 1 only has a Failure dependency between de activity's. In this situation this results in a Succeeded pipeline if the Stored procedure failed.

If also the success connection was linked to a follow up activity, and the SP would fail, the pipeline would be indeed marked as failed.

So A.

upvoted 51 times

BK10 4 years, 6 months ago

well explained! A is right

upvoted 1 times

Ram0202 3 years, 2 months ago

correct ,Execute pipeline explained here https://youtu.be/Jkz1dtLrBE4

upvoted 5 times

echerish 4 years, 11 months ago

Pipeline 2 executes Pipeline 1 if success set variable. Since Pipeline 1 exists it's a success Pipeline 1 Stored procedure fails. If fails set variable. Since the expected outcome is fail the job runs successfully and sets variable1. At least that's how I understand it

upvoted 28 times

Danweo Most Recent 2 years, 1 month ago

A is correct

upvoted 1 times

moneytime 2 years, 6 months ago

I chose D There are two paths defined in the question. For pipeline1, upon failure For pipeline2; Upon success. Now ,for pipeline1, to respond to the run of pipeline 2 simply showed that they were connected. This is a case of Upon if else block condition. When previous activity (procedure)failed : node Upon Success is skipped and its parent node failed; overall pipeline1 failed When previous activity succeeds: node Upon Success succeeded and node Upon Failure is skipped (and its parent node succeeds); overall pipeline succeeds. Hence ,pipeline 2 succeeds The pipeline2 can be visualized as thus P2->P1

upvoted 1 times

kkk5566 2 years, 11 months ago

Correct answer is A.

upvoted 1 times

azaspirant 3 years ago

How do we know which is the default behaviour? Its not mentioned anywhere wether there is a success dependency or failure dependency

upvoted 3 times

akhil5432 3 years ago

OPTION "A"

upvoted 1 times

joponlu 3 years, 2 months ago

Correct!!!

upvoted 2 times

mamahani 3 years, 3 months ago

A is correct answer

upvoted 2 times

Billybob0604 3 years, 8 months ago

So the default behaviour is Failed dependency ? If so the answer is A. But it doesn't say this anywhere in the question.

upvoted 1 times

NoobTester 3 years, 10 months ago

Answer is correct. This article helped: https://www.sqlshack.com/dependencies-in-azure-data-factory/

upvoted 1 times

Deeksha1234 4 years ago

A is correct

upvoted 1 times

SebK 4 years, 4 months ago

Correct

upvoted 1 times

AngelJP 4 years, 5 months ago

Selected Answer: A

A correct: Pipeline 1 is in try catch sentence --> Success Pipeline 2 --> Success https://docs.microsoft.com/en-us/azure/data-factory/tutorial-pipeline-failure-error-handling#try-catch-block

upvoted 6 times

AngelJP 4 years, 5 months ago

A correct: Pipeline 1 is in try catch sentence --> Success Pipeline 2 --> Success https://docs.microsoft.com/en-us/azure/data-factory/tutorial-pipeline-failure-error-handling#try-catch-blockThanks

upvoted 6 times

AvSUN 2 years, 11 months ago

Thanks

upvoted 1 times

PallaviPatel 4 years, 6 months ago

A correct. I agree with SaferSephy's comments below.

upvoted 3 times

dev2dev 4 years, 7 months ago

A is correct. Pipeline 1 is connected to Set variable to Failure node/event. Its like handling exceptions/errors in programming language. Without Failure node, it would be treated as failed.

upvoted 2 times

VeroDon 4 years, 7 months ago

Correct

upvoted 1 times

Question #7 *Topic 2*

HOTSPOT -

A company plans to use Platform-as-a-Service (PaaS) to create the new data pipeline process. The process must meet the following requirements:

Ingest:

- ✑ Access multiple data sources.
- ✑ Provide the ability to orchestrate workflow.
- ✑ Provide the capability to run SQL Server Integration Services packages.

Store:

- ✑ Optimize storage for big data workloads.
- ✑ Provide encryption of data at rest.
- ✑ Operate with no size limits.

Prepare and Train:

- ✑ Provide a fully-managed and interactive workspace for exploration and visualization.
- ✑ Provide the ability to program in R, SQL, Python, Scala, and Java.

Provide seamless user authentication with Azure Active Directory.

Model & Serve:

- ✑ Implement native columnar storage.
- ✑ Support for the SQL language
- ✑ Provide support for structured streaming.

You need to build the data integration pipeline.

Which technologies should you use? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

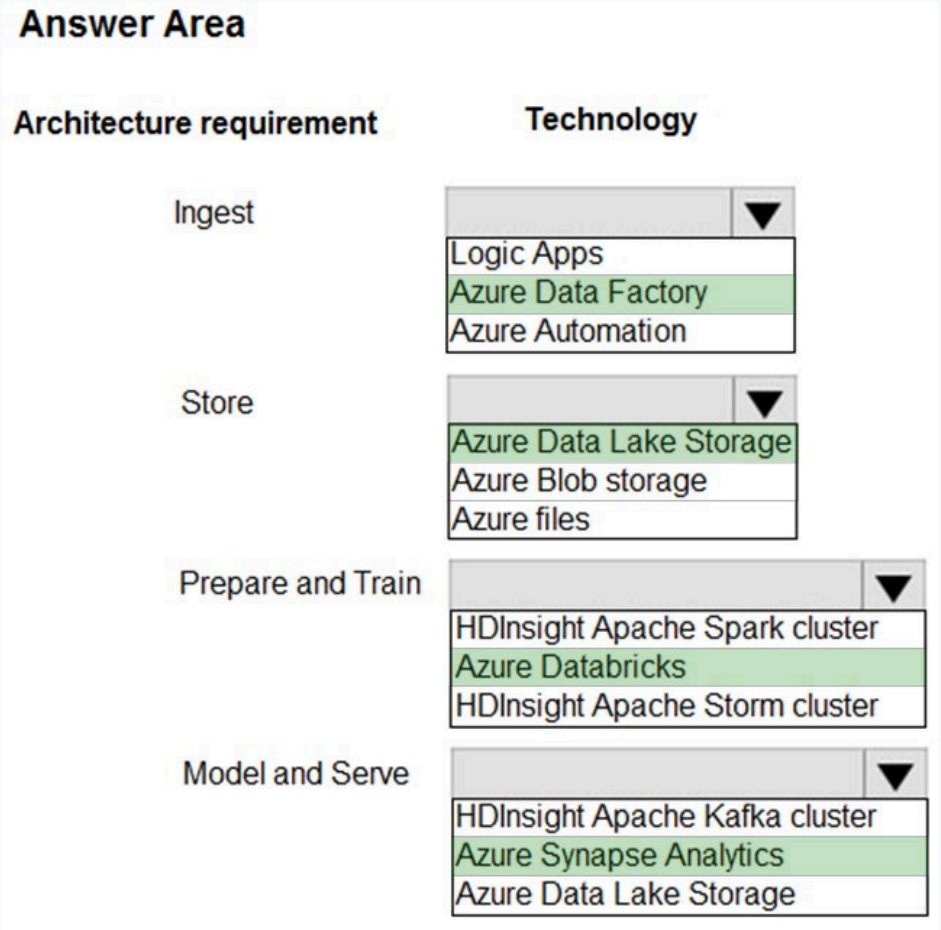
Hot Area:

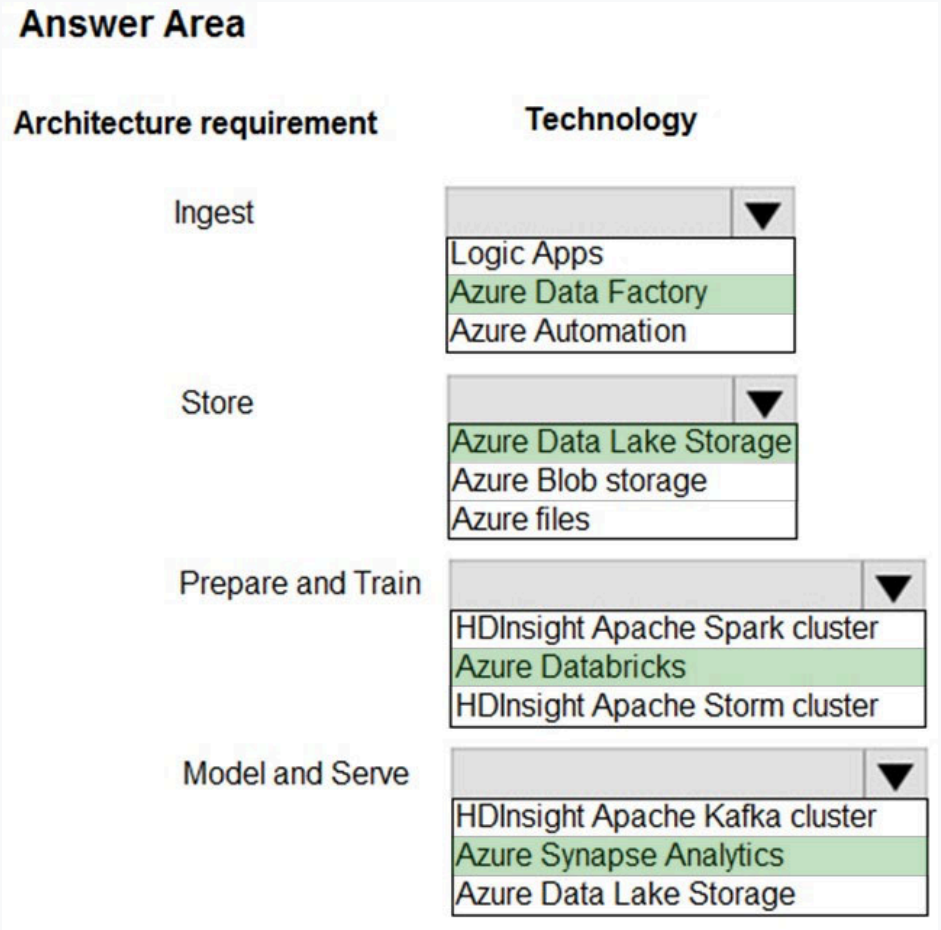
.

Answer Area

Architecture requirement	Technology
Ingest	Logic Apps / Azure Data Factory / Azure Automation
Store	Azure Data Lake Storage / Azure Blob storage / Azure files
Prepare and Train	HDInsight Apache Spark cluster / Azure Databricks / HDInsight Apache Storm cluster
Model and Serve	HDInsight Apache Kafka cluster / Azure Synapse Analytics / Azure Data Lake Storage

Suggested Answer:





Ingest: Azure Data Factory - Azure Data Factory pipelines can execute SSIS packages. In Azure, the following services and tools will meet the core requirements for pipeline orchestration, control flow, and data movement: Azure Data Factory, Oozie on HDInsight, and SQL Server Integration Services (SSIS).

Store: Data Lake Storage - Data Lake Storage Gen1 provides unlimited storage. Note: Data at rest includes information that resides in persistent storage on physical media, in any digital format. Microsoft Azure offers a variety of data storage solutions to meet different needs, including file, disk, blob, and table storage. Microsoft also provides encryption to protect Azure SQL Database, Azure Cosmos DB, and Azure Data Lake. Prepare and Train: Azure Databricks Azure Databricks provides enterprise-grade Azure security, including Azure Active Directory integration. With Azure Databricks, you can set up your Apache Spark environment in minutes, autoscale and collaborate on shared projects in an interactive workspace. Azure Databricks supports Python, Scala, R, Java and SQL, as well as data science frameworks and libraries including TensorFlow, PyTorch and scikit-learn. Model and Serve: Azure Synapse Analytics Azure Synapse Analytics/ SQL Data Warehouse stores data into relational tables with columnar storage. Azure SQL Data Warehouse connector now offers efficient and scalable structured streaming write support for SQL Data Warehouse. Access SQL Data Warehouse from Azure Databricks using the SQL Data Warehouse connector. Note: As of November 2019, Azure SQL Data Warehouse is now Azure Synapse Analytics. Reference: https://docs.microsoft.com/bs-latn-ba/azure/architecture/data-guide/technology-choices/pipeline-orchestration-data-movement https://docs.microsoft.com/en-us/azure/azure-databricks/what-is-azure-databricks

datachamp Highly Voted 4 years, 11 months ago

Is this an ad?

upvoted 78 times

datachamp Highly Voted 4 years, 11 months ago

Is this an ad?MS is a MKTcompany, tech is just an excusesounds like one, replace databricks with Azure Fabric and then it will be for sureYou mean like all the Azure certification exams?

upvoted 78 times

Fagner85 Most Recent 1 year, 6 months ago

MS is a MKTcompany, tech is just an excuse

upvoted 1 times

Bhabani83 3 years, 1 month ago

sounds like one, replace databricks with Azure Fabric and then it will be for sure

upvoted 5 times

phydev 2 years, 9 months ago

You mean like all the Azure certification exams?

upvoted 8 times

Podavenna 4 years, 11 months ago

Correct solution!

upvoted 49 times

renan_ineu 1 year, 11 months ago

"I don't believe the storage solution is Azure Data Lake. Given that the question mentions PaaS, the Azure Data Encryption at Rest doc from Microsoft states: 'Platform as a Service (PaaS) customer's data typically resides in a storage service such as Blob Storage but may also be cached or stored in the application execution environment, such as a virtual machine.' To see the encryption at rest options available to you, examine the Data encryption models: supporting services table for the storage and application platforms that you use. So I believe the options are: Data Factory, Blob Storage, Databricks, and Synapse." https://learn.microsoft.com/en-us/azure/security/fundamentals/encryption-atrest#encryption-at-rest-for-paas-customers

upvoted 1 times

Alongi 2 years, 4 months ago

All correct, congrats!

upvoted 1 times

positivitypeople 2 years, 8 months ago

Got this question today on the exam

upvoted 7 times

kkk5566 2 years, 11 months ago

Correct

upvoted 1 times

anks84 3 years, 11 months ago

Given answer is correct !

upvoted 5 times

Deeksha1234 4 years ago

Correct

upvoted 2 times

SebK 4 years, 4 months ago

Correct

upvoted 1 times

Massy 4 years, 5 months ago

for the store, couldn't we use also Azure Blob Storage? It supports all the three requisites

upvoted 3 times

Massy 4 years, 5 months ago

for the store, couldn't we use also Azure Blob Storage? It supports all the three requisitesBecause ADLS Gen2 support Big Data Workload betterADLSG2 supports big data and large data storage as compared with blob

upvoted 3 times

NewTuanAnh 4 years, 4 months ago

Because ADLS Gen2 support Big Data Workload better

upvoted 4 times

kreative_feco 2 years, 7 months ago

ADLSG2 supports big data and large data storage as compared with blob

upvoted 1 times

paras_gadhiya 4 years, 6 months ago

Correct

upvoted 1 times

PallaviPatel 4 years, 6 months ago

Correct solution.

upvoted 1 times

joeljohnrm 4 years, 7 months ago

Correct Solution

upvoted 1 times

corebit 4 years, 8 months ago

Would be best if people including answers that go against the popular responses provide some reference instead of blinding saying false

upvoted 5 times

Akash0105 4 years, 9 months ago

Answer is correct. Azure Databricks supports java: https://azure.microsoft.com/en-us/services/databricks/#overview

upvoted 2 times

Pratikh 4 years, 9 months ago

Databricks doesn't support Java so in the Prep and Train should be HDInsight Apache Spark Cluster

upvoted 5 times

Pratikh 4 years, 9 months ago

Databricks doesn't support Java so in the Prep and Train should be HDInsight Apache Spark ClusterAzure Databricks supports Python, Scala, R, Java, and SQL, as well as data science frameworks and libraries including TensorFlow, PyTorch, and scikit-learn.

upvoted 5 times

KOSTA007 4 years, 9 months ago

Azure Databricks supports Python, Scala, R, Java, and SQL, as well as data science frameworks and libraries including TensorFlow, PyTorch, and scikit-learn.

upvoted 11 times

Aslam208 4 years, 9 months ago

Databricks does not support Java, Prepare and Train should be Azure HDInsight Apache spark clusterI would like to correct my answer here... java is supported in Azure Databricks, therefore Prepare and Train can be done with Azure Databricks

upvoted 1 times

Aslam208 4 years, 9 months ago

Databricks does not support Java, Prepare and Train should be Azure HDInsight Apache spark cluster

upvoted 1 times

Aslam208 4 years, 8 months ago

I would like to correct my answer here... java is supported in Azure Databricks, therefore Prepare and Train can be done with Azure Databricks

upvoted 5 times

DRAG DROP -

You have the following table named Employees.

You need to calculate the employee_type value based on the hire_date value.

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

first_name	last_name	hire_date	employee_type
Jane	Doe	2019-08-23	new
Ben	Smith	2017-12-15	Standard

Values

CASE

ELSE

OVER

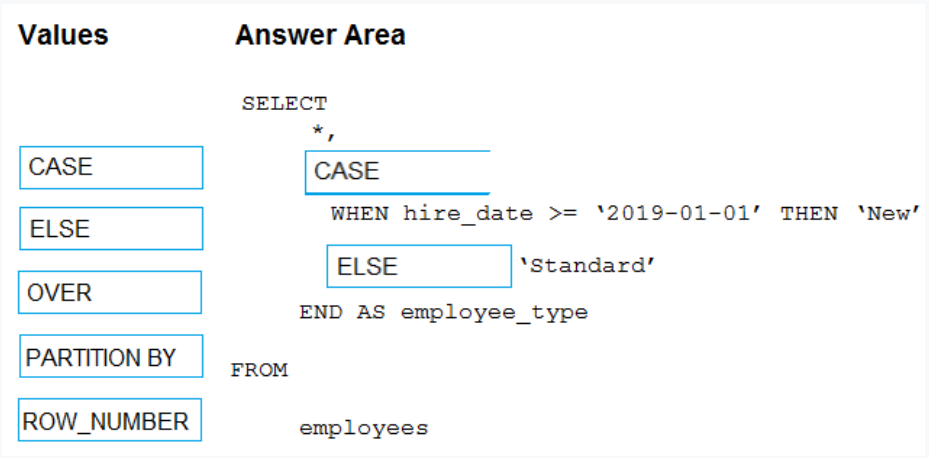
PARTITION BY

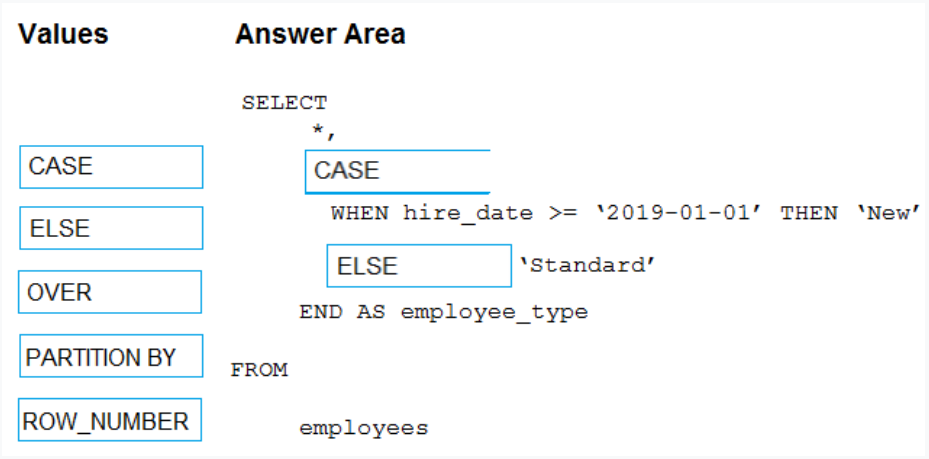
ROW_NUMBER

Answer Area

```
SELECT
     *,
     [          ]
       WHEN hire_date >= '2019-01-01' THEN 'New'
       [          ] 'Standard'
     END AS employee_type
FROM
     employees
```

Suggested Answer:





Box 1: CASE - CASE evaluates a list of conditions and returns one of multiple possible result expressions. CASE can be used in any statement or clause that allows a valid expression. For example, you can use CASE in statements such as SELECT, UPDATE, DELETE and SET, and in clauses such as select_list, IN, WHERE, ORDER BY, and HAVING. Syntax: Simple CASE expression:

CASE input_expression - WHEN when_expression THEN result_expression [...n] [ELSE else_result_expression]

END -

Box 2: ELSE - Reference: https://docs.microsoft.com/en-us/sql/t-sql/language-elements/case-transact-sql

MoDar Highly Voted 4 years, 11 months ago

Correct

upvoted 43 times

[Removed] 4 years, 12 months ago

The answer is correct. But, is this in the scope of this exam?it seemsGot this question yesterday so yes.correct without reading the question as this is basic sqlmake sense to me , data engineer should be able to write Queries

upvoted 9 times

[Removed] 4 years, 12 months ago

The answer is correct. But, is this in the scope of this exam?

upvoted 9 times

anto69 4 years, 7 months ago

it seems

upvoted 1 times

mkutts Highly Voted 4 years, 9 months ago

Got this question yesterday so yes.

upvoted 13 times

f7c717f Most Recent 1 year, 8 months ago

correct without reading the question as this is basic sql

upvoted 1 times

parwa 4 years, 11 months ago

make sense to me , data engineer should be able to write Queries

upvoted 11 times

606a82e 2 years ago

Correct

upvoted 1 times

kkk5566 2 years, 11 months ago

Correct

upvoted 1 times

Ankit_Az 3 years, 2 months ago

Correct

upvoted 1 times

GodfreyMbizo 3 years, 6 months ago

they cant bring hard questions only for exam to balance

upvoted 1 times

DindaS 3 years, 6 months ago

yes the answer is correct but not sure if they will be included the DP-203 exam

upvoted 1 times

Igor85 3 years, 9 months ago

i wish all the questions in the exam were like that one :)

upvoted 6 times

Deeksha1234 4 years ago

correct

upvoted 1 times

NewTuanAnh 4 years, 4 months ago

the answer is correct CASE ... WHEN ... THEN... ELSE ...

upvoted 2 times

PallaviPatel 4 years, 6 months ago

correct

upvoted 2 times

Question #9 *Topic 2*

DRAG DROP -

You have an Azure Synapse Analytics workspace named WS1.

You have an Azure Data Lake Storage Gen2 container that contains JSON-formatted files in the following format.

You need to use the serverless SQL pool in WS1 to read the files.

How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:


```
{
     "id": "66532691-ab20-11ea-8b1d-936b3ec64e54",
     "context": {
          "data": {
               "eventTime": "2020-06-10T13:43:34.553Z",
               "samplingRate": "100.0",
               "isSynthetic": "false"
          },
          "session": {
               "isFirst": "false",
               "id": "38619c14-7a23-4687-8268-95862c5326b1"
          },
          "custom": {
               "dimensions": [
                    {
                         "customerInfo": {
                              "ProfileType": "ExpertUser",
                              "RoomName": "",
                              "CustomerName": "diamond",
                              "UserName": "XXXX@yahoo.com"
                    }
          },
          {
                         "customerInfo" {
                              "ProfileType": "Novice",
                              "RoomName": "",
                              "CustomerName": "topaz",
                              "UserName": "XXXX@outlook.com"
                         }
                    }
               ]
          }
     }
}
```

Values

- opendatasource
- openjson
- openquery
- openrowset

Answer Area

```
select*

FROM
[                ] (
     BULK 'https://contoso.blob.core.windows.net/contosodw',
     FORMAT= 'CSV',
     fieldterminator = '0x0b',
     fieldquote = '0x0b',
     rowterminator = '0x0b'
)
with (id varchar(50),
     contextdateventTime varchar(50) '$.context.data.eventTime',
     contextdatasamplingRate varchar(50) '$.context.data.samplingRate',
     contextdataisSynthetic varchar(50) '$.context.data.isSynthetic'.
     contextsessionisFirst varchar(50) '$.context.session.isFirst',
     contextsession varchar(50) '$.context.session.id',
     contextcustomdimensions varchar(max) '$.context.custom.dimensions'

) as q
cross apply [                ](contextcustomdimensions)

with ( ProfileType varchar(50) '$.customerInfo.ProfileType',
          RoomName varchar(50) '$.customerInfo.RoomName',
          CustomerName varchar(50) '$.customerInfo.CustomerName',
          UserName varchar(50) '$.customerInfo.UserName'
     )
```

Suggested Answer:

Values

opendatasource

openquery

Answer Area

```
select*

FROM
   openrowset (

      BULK 'https://contoso.blob.core.windows.net/contosodw',
      FORMAT= 'CSV',
      fieldterminator = '0x0b',
      fieldquote = '0x0b',
      rowterminator = '0x0b'
)
with (id varchar(50),
      contextdateventTime varchar(50) '$.context.data.eventTime',
      contextdatasamplingRate varchar(50) '$.context.data.samplingRate',
      contextdataisSynthetic varchar(50) '$.context.data.isSynthetic'.
      contextsessionisFirst varchar(50) '$.context.session.isFirst',
      contextsession varchar(50) '$.context.session.id',
      contextcustomdimensions varchar(max) '$.context.custom.dimensions'

) as q
cross apply   openjson   (contextcustomdimensions)

with ( ProfileType varchar(50) '$.customerInfo.ProfileType',
          RoomName varchar(50) '$.customerInfo.RoomName',
          CustomerName varchar(50) '$.customerInfo.CustomerName',
          UserName varchar(50) '$.customerInfo.UserName'
      )
```

Values

opendatasource

openquery

Answer Area

```
select*

FROM
   openrowset (

      BULK 'https://contoso.blob.core.windows.net/contosodw',
      FORMAT= 'CSV',
      fieldterminator = '0x0b',
      fieldquote = '0x0b',
      rowterminator = '0x0b'
)
with (id varchar(50),
      contextdateventTime varchar(50) '$.context.data.eventTime',
      contextdatasamplingRate varchar(50) '$.context.data.samplingRate',
      contextdataisSynthetic varchar(50) '$.context.data.isSynthetic'.
      contextsessionisFirst varchar(50) '$.context.session.isFirst',
      contextsession varchar(50) '$.context.session.id',
      contextcustomdimensions varchar(max) '$.context.custom.dimensions'

) as q
cross apply   openjson   (contextcustomdimensions)

with ( ProfileType varchar(50) '$.customerInfo.ProfileType',
          RoomName varchar(50) '$.customerInfo.RoomName',
          CustomerName varchar(50) '$.customerInfo.CustomerName',
          UserName varchar(50) '$.customerInfo.UserName'
      )
```

Box 1: openrowset - The easiest way to see to the content of your CSV file is to provide file URL to OPENROWSET function, specify csv FORMAT.
Example: SELECT * FROM OPENROWSET(BULK 'csv/population/population.csv', DATA_SOURCE = 'SqlOnDemandDemo', FORMAT = 'CSV', PARSER_VERSION = '2.0', FIELDTERMINATOR =',', ROWTERMINATOR = '\n'

Box 2: openjson - You can access your JSON files from the Azure File Storage share by using the mapped drive, as shown in the following example:

SELECT book.* FROM - OPENROWSET(BULK N't:\books\books.json', SINGLE_CLOB) AS json CROSS APPLY OPENJSON(BulkColumn) WITH(id nvarchar(100), name nvarchar(100), price float, pages_i int, author nvarchar(100)) AS book Reference: https://docs.microsoft.com/en-

us/azure/synapse-analytics/sql/query-single-csv-file https://docs.microsoft.com/en-us/sql/relational-databases/json/import-json-documents-into-sql-server

Maunik Highly Voted 4 years, 11 months ago

Answer is correct https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files

upvoted 47 times

Maunik Highly Voted 4 years, 11 months ago

Answer is correct https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-filesanswer is correct as per this link

upvoted 47 times

Lrng15 4 years, 11 months ago

answer is correct as per this link

upvoted 2 times

gf2tw 4 years, 11 months ago

The question and answer seem out of place, there was no mention of the CSV and the query in the answer doesn't match up with openjson at all

upvoted 11 times

gf2tw 4 years, 11 months ago

The question and answer seem out of place, there was no mention of the CSV and the query in the answer doesn't match up with openjson at allagree with you, very misleadingLook at the WITH statement, the csv column can contain json data.The easiest way to see to the content of your JSON file is to provide the file URL to the OPENROWSET function, specify csv FORMAT, and set values 0x0b for fieldterminator and fieldquote.Actually, the csv format is specified if you're using OPENROWSET to read json files in Synapse. The OPENJSON is required if you want to parse data from every array in the document. See the OPENJSON example in this link: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files#query-json-files-using-openjsonThanks, you're right: "The easiest way to see to the content of your JSON file is to provide the file URL to the OPENROWSET function, specify csv FORMAT, and set values 0x0b for fieldterminator and fieldquote."

upvoted 11 times

gssd4scoder 4 years, 9 months ago

agree with you, very misleading

upvoted 1 times

dev2dev 4 years, 7 months ago

Look at the WITH statement, the csv column can contain json data.

upvoted 1 times

vctrhugo 3 years, 1 month ago

The easiest way to see to the content of your JSON file is to provide the file URL to the OPENROWSET function, specify csv FORMAT, and set values 0x0b for fieldterminator and fieldquote.

upvoted 3 times

dead_SQL_pool 4 years, 9 months ago

Actually, the csv format is specified if you're using OPENROWSET to read json files in Synapse. The OPENJSON is required if you want to parse data from every array in the document. See the OPENJSON example in this link: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files#query-json-files-using-openjson

upvoted 12 times

dead_SQL_pool 4 years, 9 months ago

Actually, the csv format is specified if you're using OPENROWSET to read json files in Synapse. The OPENJSON is required if you want to parse data from every array in the document. See the OPENJSON example in this link: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files#query-json-files-using-openjsonThanks, you're right: "The easiest way to see to the content of your JSON file is to provide the file URL to the OPENROWSET function, specify csv FORMAT, and set values 0x0b for fieldterminator and fieldquote."

upvoted 12 times

gf2tw 4 years, 8 months ago

Thanks, you're right: "The easiest way to see to the content of your JSON file is to provide the file URL to the OPENROWSET function, specify csv FORMAT, and set values 0x0b for fieldterminator and fieldquote."

upvoted 5 times

evangelist 2 years, 1 month ago

To complete the Transact-SQL statement for reading JSON-formatted files using the serverless SQL pool in WS1, you should use OPENROWSET to access the data and OPENJSON to parse the JSON content. Here is the correct completion of the statement:

upvoted 1 times

KarlGardnerDataEngineering 2 years, 2 months ago

This took me about 10 hours to understand this query

upvoted 1 times

KarlGardnerDataEngineering 2 years, 2 months ago

This took me about 10 hours to understand this queryMe too, dammit!

upvoted 1 times

drosen Most Recent 1 year, 10 months ago

Me too, dammit!

upvoted 1 times

ELJORDAN23 2 years, 7 months ago

Answer is correct: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files#query-json-files-using-openjson

upvoted 2 times

kkk5566 2 years, 11 months ago

Answer is correct

upvoted 1 times

mamahani 3 years, 3 months ago

openrowset / openjson

upvoted 1 times

zorko10 3 years, 10 months ago

does openjson do the same thing as jsoncontent ? I tried running a query on a json file and the auto filled code used jsoncontent instead of openjson

upvoted 1 times

Deeksha1234 4 years ago

correct

upvoted 1 times

SebK 4 years, 4 months ago

Correct

upvoted 1 times

PallaviPatel 4 years, 6 months ago

correct

upvoted 1 times

DRAG DROP -

You have an Apache Spark DataFrame named temperatures. A sample of the data is shown in the following table.

You need to produce the following table by using a Spark SQL query.

How should you complete the query? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all.

You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

Date	Temp
...	...
18-01-2021	3
19-01-2021	4
20-01-2021	2
21-01-2021	2
...	...

Year	JAN	FEB	MAR	APR	MAY
2019	2.3	4.1	5.2	7.6	9.2
2020	2.4	4.2	4.9	7.8	9.1
2021	2.6	5.3	3.4	7.9	9.5

Values

CAST

COLLATE

CONVERT

FLATTEN

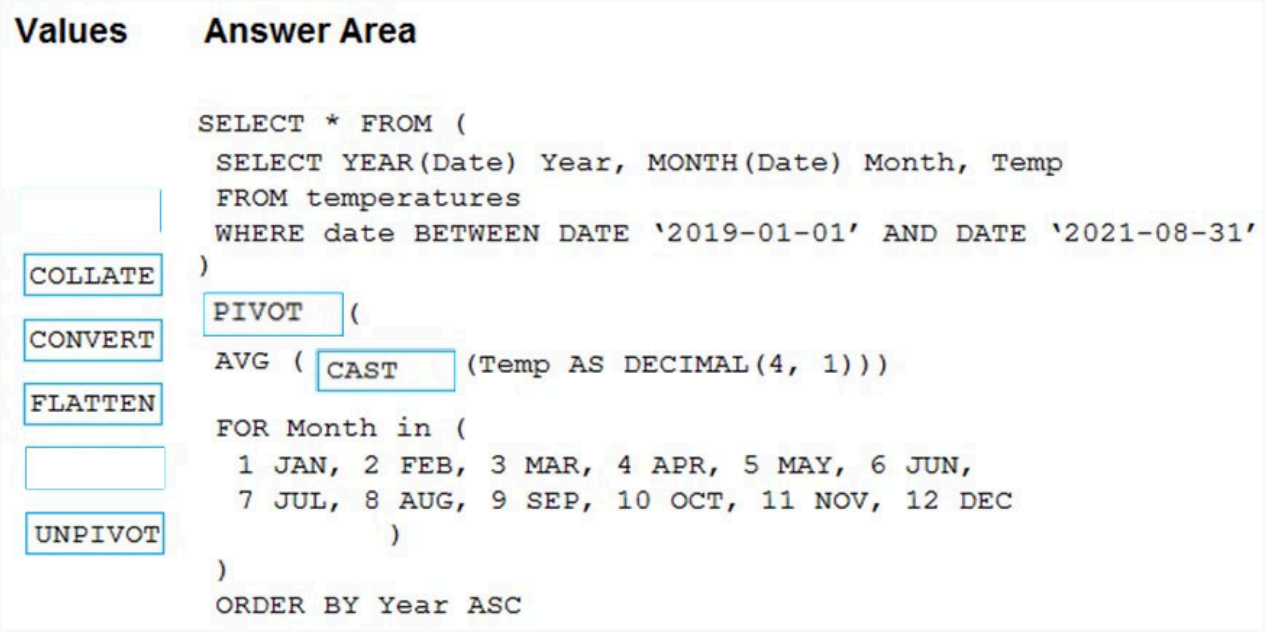
PIVOT

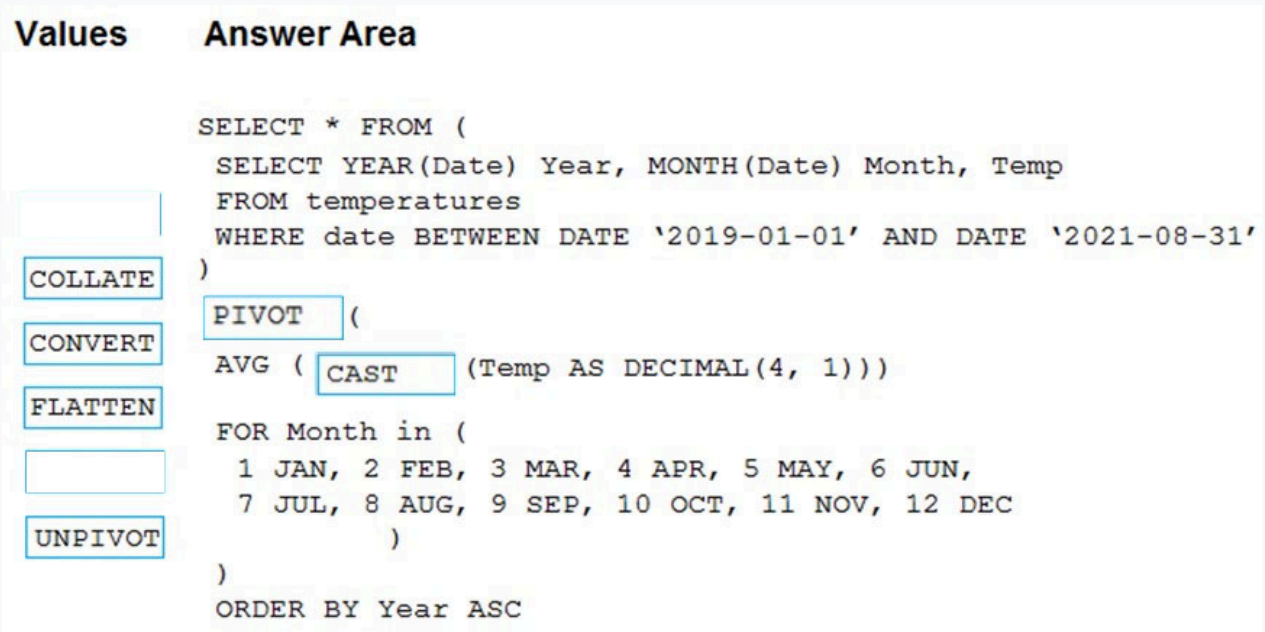
UNPIVOT

Answer Area

```
SELECT * FROM (
 SELECT YEAR(Date) Year, MONTH(Date) Month, Temp
 FROM temperatures
 WHERE date BETWEEN DATE '2019-01-01' AND DATE '2021-08-31'
)
 [          ] (
 AVG ( [          ] (Temp AS DECIMAL(4, 1)))
 FOR Month in (
  1 JAN, 2 FEB, 3 MAR, 4 APR, 5 MAY, 6 JUN,
  7 JUL, 8 AUG, 9 SEP, 10 OCT, 11 NOV, 12 DEC
          )
 )
 ORDER BY Year ASC
```

Suggested Answer:





Box 1: PIVOT - PIVOT rotates a table-valued expression by turning the unique values from one column in the expression into multiple columns in the output. And PIVOT runs aggregations where they're required on any remaining column values that are wanted in the final output. Incorrect Answers: UNPIVOT carries out the opposite operation to PIVOT by rotating columns of a table-valued expression into column values.

Box 2: CAST - If you want to convert an integer value to a DECIMAL data type in SQL Server use the CAST() function. Example:

SELECT - CAST(12 AS DECIMAL(7,2)) AS decimal_value; Here is the result: decimal_value 12.00 Reference: https://learnsql.com/cookbook/how-to-convert-an-integer-to-a-decimal-in-sql-server/ https://docs.microsoft.com/en-us/sql/t-sql/queries/from-using-pivot-and-unpivot

SujithaVulchi Highly Voted 4 years, 4 months ago
correct answer, pivot and cast
upvoted 46 times

ggggyyyyy Highly Voted 4 years, 5 months ago
correct. cast not convert
upvoted 5 times

Alongi Most Recent 1 year, 10 months ago
Correct

upvoted 1 times

MarkJoh 2 years, 2 months ago

Pivot and Cast are correct. There is an issue with the problem though. It should be CAST(avg(temp)... and not Avg (Cast(temp)...

upvoted 1 times

MarkJoh 2 years, 2 months ago

Pivot and Cast are correct. There is an issue with the problem though. It should be CAST(avg(temp)... and not Avg (Cast(temp)...No this is not correct, if you CAST(AVG(temp)) then you will first get AVG(temp) as an int. Casting it then results in the decimal value being 0 (like 2.0, 3.0...). Therefore, we have to AVG(CAST(temp)).

upvoted 1 times

jongert 2 years, 1 month ago

No this is not correct, if you CAST(AVG(temp)) then you will first get AVG(temp) as an int. Casting it then results in the decimal value being 0 (like 2.0, 3.0...). Therefore, we have to AVG(CAST(temp)).

upvoted 2 times

ellala 2 years, 4 months ago

Answer is correct

upvoted 1 times

kkk5566 2 years, 5 months ago

pivot and cast

upvoted 1 times

Ankit_Az 2 years, 8 months ago

Correct

upvoted 1 times

jainparag1 3 years, 7 months ago

correct explanation

upvoted 2 times

Question #11 *Topic 2*

You have an Azure Data Factory that contains 10 pipelines.
You need to label each pipeline with its main purpose of either ingest, transform, or load. The labels must be available for grouping and filtering when using the monitoring experience in Data Factory.
What should you add to each pipeline?

A. a resource tag

B. a correlation ID

C. a run group ID

D. an annotation

Suggested Answer: *D*
Annotations are additional, informative tags that you can add to specific factory resources: pipelines, datasets, linked services, and triggers. By adding annotations, you can easily filter and search for specific factory resources. Reference: https://www.cathrinewilhelmsen.net/annotations-user-properties-azure-data-factory/

Community vote distribution
D (100%)

umeshkd05 Highly Voted 3 years, 11 months ago
Annotation
upvoted 22 times

umeshkd05 Highly Voted 3 years, 11 months ago
AnnotationCause ADF pipelines are not first class resources
upvoted 22 times

anto69 3 years, 7 months ago
Cause ADF pipelines are not first class resources
upvoted 3 times

AhmedDaffaie 3 years, 5 months ago
What is the difference between resource tags and annotations?
upvoted 13 times

kkk5566 Most Recent 1 year, 11 months ago
D -Annotation
upvoted 1 times

akhil5432 2 years ago
OPTION -D
upvoted 1 times

Ankit_Az 2 years, 2 months ago
Correct
upvoted 1 times

joponlu 2 years, 2 months ago
D is correct!!
upvoted 1 times

esaade 2 years, 5 months ago
To label each pipeline with its main purpose of either ingest, transform, or load and make the labels available for grouping and filtering when using the monitoring experience in Data Factory, you should add an annotation to each pipeline. Therefore, the correct answer is D. an annotation.
Annotations are key-value pairs that you can add to pipelines, datasets, and activities to help you organize and categorize them. They can be used for a variety of purposes, including labeling pipelines with their main purpose of either ingest, transform, or load. Annotations can also be used for filtering, grouping, and searching for resources in the Data Factory monitoring experience.
upvoted 3 times

DindaS 2 years, 6 months ago
D -Annotation

upvoted 3 times

vigilante89 2 years, 8 months ago

ADF annotations are tags that you can add to your Azure Data Factory components to identify them. A tag allows you to classify or group different objects in order to easily monitor them after an execution. You can create multiple Azure Data Factory annotations.

upvoted 5 times

arunesh789 2 years, 10 months ago

hjghfgh

upvoted 2 times

arunesh789 2 years, 10 months ago

hjghfghKIndly delete above comment. Answer is D.

upvoted 2 times

arunesh789 2 years, 10 months ago

KIndly delete above comment. Answer is D.

upvoted 4 times

Deeksha1234 3 years ago

correct

upvoted 1 times

paras_gadhiya 3 years, 6 months ago

Correct!

upvoted 1 times

PallaviPatel 3 years, 6 months ago

correct

upvoted 1 times

huesazo 3 years, 7 months ago

Anotacion

upvoted 1 times

aarthy2 3 years, 10 months ago

yes correct, annotation provides label functionality than show in pipeline monitoring.

upvoted 2 times

Question #12 *Topic 2*

HOTSPOT -

The following code segment is used to create an Azure Databricks cluster.

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

Hot Area:

```
{
     "num_workers": null,
     "autoscale": {
          "min_workers": 2,
          "max_workers": 8
     },
     "cluster_name": "MyCluster",
     "spark_version": "latest-stable-scala2.11",
     "spark_conf": {
          "spark.databricks.cluster.profile": "serverless",
          "spark.databricks.repl.allowedLanguages": "sql,python,r"
     },
     "node_type_id": "Standard_DS13_v2",
     "ssh_public_keys": [],
     "custom_tags": {
          "ResourceClass": "Serverless"
     },
     "spark_env_vars": {
          "PYSPARK_PYTHON": "/databricks/python3/bin/python3"
     },
     "autotermination_minutes": 90,
     "enable_elastic_disk": true,
     "init_scripts": []
}
```

Answer Area

Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	○	○
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	○	○
The Databricks cluster supports the creation of a Delta Lake table.	○	○

Suggested Answer:

Answer Area

Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	○	○
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	○	○
The Databricks cluster supports the creation of a Delta Lake table.	○	○

Answer Area

Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	○	○
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	○	○
The Databricks cluster supports the creation of a Delta Lake table.	○	○

Box 1: Yes - A cluster mode of 'High Concurrency' is selected, unlike all the others which are 'Standard'. This results in a worker type of Standard_DS13_v2.

Box 2: No - When you run a job on a new cluster, the job is treated as a data engineering (job) workload subject to the job workload pricing. When you run a job on an existing cluster, the job is treated as a data analytics (all-purpose) workload subject to all-purpose workload pricing.

Box 3: Yes - Delta Lake on Databricks allows you to configure Delta Lake based on your workload patterns. Reference: https://adatis.co.uk/databricks-cluster-sizing/ https://docs.microsoft.com/en-us/azure/databricks/jobs https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html https://docs.databricks.com/delta/index.html

GameLift Highly Voted 4 years, 4 months ago

FROM DP-201, thanks to rmk4ever :: 1. Yes A cluster mode of 'High Concurrency' is selected, unlike all the others which are 'Standard'. This results in a worker type of Standard_DS13_v2. ref: https://adatis.co.uk/databricks-cluster-sizing/ 2. NO recommended: New Job Cluster. When you run a job on a new cluster, the job is treated as a data engineering (job) workload subject to the job workload pricing. When you run a job on an existing cluster, the job is treated as a data analytics (all-purpose) workload subject to all-purpose workload pricing. ref: https://docs.microsoft.com/en-us/azure/databricks/jobs Scheduled batch workload- Launch new cluster via job ref: https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html#plan-capacity-and-control-cost 3.YES Delta Lake on Databricks allows you to configure Delta Lake based on your workload patterns. ref: https://docs.databricks.com/delta/index.htmlagree on this oneFor 1, where do you see high concurrency?I hadn't read the other comments yet, apparently it's in 'serverless' :)

upvoted 56 times

GameLift Highly Voted 4 years, 4 months ago

FROM DP-201, thanks to rmk4ever ::

1. Yes A cluster mode of 'High Concurrency' is selected, unlike all the others which are 'Standard'. This results in a worker type of Standard_DS13_v2. ref: https://adatis.co.uk/databricks-cluster-sizing/

2. NO recommended: New Job Cluster. When you run a job on a new cluster, the job is treated as a data engineering (job) workload subject to the job workload pricing. When you run a job on an existing cluster, the job is treated as a data analytics (all-purpose) workload subject to all-purpose workload pricing. ref: https://docs.microsoft.com/en-us/azure/databricks/jobs Scheduled batch workload- Launch new cluster via job ref:

https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html#plan-capacity-and-control-cost

3.YES Delta Lake on Databricks allows you to configure Delta Lake based on your workload patterns. ref: https://docs.databricks.com/delta/index.html

upvoted 56 times

Egocentric 3 years, 10 months ago

agree on this one

upvoted 3 times

semauni 2 years, 6 months ago

For 1, where do you see high concurrency?

upvoted 3 times

semauni 2 years, 6 months ago

For 1, where do you see high concurrency?I hadn't read the other comments yet, apparently it's in 'serverless' :)

upvoted 3 times

semauni 2 years, 6 months ago

I hadn't read the other comments yet, apparently it's in 'serverless' :)

upvoted 1 times

Canary_2021 4 years, 1 month ago

Answer is Correct. Box 1: Yes "spark.databricks.cluster.profile": "serverless" means that the cluster is a High Concurrency Cluster, which support multi-users. Box 2: No Scheduled jobs should run in standard cluster. High Concurrency clusters are intended for multi-users and won't benefit a cluster running a single job. Box 3:Yes

upvoted 39 times

Software_One Most Recent 1 year, 11 months ago

N, N, Y. 1 : N because that's for single user, shared not support R

upvoted 1 times

kkk5566 2 years, 5 months ago

the answer is Yes, No, Yes.

upvoted 1 times

hiyoww 2 years, 6 months ago

the naming of the clusters are changed in recent UI: https://docs.databricks.com/en/archive/compute/cluster-ui-preview.html

upvoted 1 times

mamahani 2 years, 9 months ago

yes / no / yes

upvoted 3 times

Igor85 3 years, 3 months ago

i guess this question won't be relevant anymore, since cluster creation UI has changedThey still use this question, I had this one on my exam this week

upvoted 3 times

Igor85 3 years, 3 months ago

i guess this question won't be relevant anymore, since cluster creation UI has changed

upvoted 3 times

WielK 2 years, 10 months ago

They still use this question, I had this one on my exam this week

upvoted 3 times

US007 3 years, 6 months ago

1. should be 'No'. Its a standard cluster and it also has scala which is not supported on High Concurrence cluster.

upvoted 4 times

Deeksha1234 3 years, 6 months ago

Yes, No, Yes

upvoted 2 times

PallaviPatel 4 years ago

Correct Answer. I agree with Canary_2021

upvoted 4 times

edba 4 years, 2 months ago

I would say the answer is Yes, No, Yes. Delta lake was supported starting from Azure Databricks Runtime 6.0 with Scala 2.11.12. https://docs.microsoft.com/en-us/azure/databricks/release-notes/runtime/6.0#system-environment

upvoted 3 times

thuggie300 4 years, 4 months ago

what is the answer lol

upvoted 3 times

aarthy2 4 years, 4 months ago

the same question is in DP-201 with the same answer. https://www.examtopics.com/discussions/microsoft/view/16875-exam-dp-201-topic-2-question-11-discussion/

upvoted 1 times

rav009 4 years, 4 months ago

IMO NO, YES, YES

upvoted 1 times

rav009 4 years, 4 months ago

IMO NO, YES, YESSorry, it should be NO,NO,YES. For Box 2, the cheapest way is creating the cluster when it's time to execute the job and terminate immediately after the task completes. This is called New Job Clusters . https://docs.microsoft.com/en-us/azure/databricks/jobs

upvoted 1 times

rav009 4 years, 4 months ago

Sorry, it should be NO,NO,YES. For Box 2, the cheapest way is creating the cluster when it's time to execute the job and terminate immediately after the task completes. This is called New Job Clusters . https://docs.microsoft.com/en-us/azure/databricks/jobs

upvoted 3 times

parwa 4 years, 5 months ago

what is correct answer here please?

upvoted 2 times

parwa 4 years, 5 months ago

what is correct answer here please?Yes No No

upvoted 2 times

amma 4 years, 5 months ago

Yes No No

upvoted 10 times

Amyqwertyu 4 years, 5 months ago

High Concurrency clusters are intended for use by multiple users. hence correct answer

upvoted 2 times

Amalbenrebai 4 years, 5 months ago

NO, NO, YES

upvoted 3 times

Question #13 *Topic 2*

You are designing a statistical analysis solution that will use custom proprietary Python functions on near real-time data from Azure Event Hubs.
You need to recommend which Azure service to use to perform the statistical analysis. The solution must minimize latency.
What should you recommend?

A. Azure Synapse Analytics

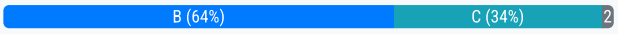
B. Azure Databricks

C. Azure Stream Analytics

D. Azure SQL Database

Suggested Answer: *C*
Reference: https://docs.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics

Community vote distribution



kolakone Highly Voted 4 years, 11 months ago
My answer will be B Stream Analytics supports "extending SQL language with JavaScript and C# user-defined functions (UDFs)". There is no mention of Python support; hence Stream Analytics is not correct. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction Azure Databricks supports near real-time data from Azure Event Hubs. And includes support for R, SQL, Python, Scala, and Java. So I will go for option B.But Python runs on Event Hubs why the other service does should support Python too?It's mentioned that "python runs on real time data from event hubs not on event hubs". Also event hub is to gather that data and after that it is analyzed by either databricks stream analytics. And since stream analytics doesn't support python so the answer is databrickstherefore i agree wih ASApython can run Event Hubs libraries real time, it doesn't have to be supported by ASA, it just needs to send data to analytics service@RoyP654, the question asks which service to perform the statistical analysis (e.g. execute the python) suggesting that the python has not/will not be ran in events hubs
upvoted 87 times

kolakone Highly Voted 4 years, 11 months ago
My answer will be B Stream Analytics supports "extending SQL language with JavaScript and C# user-defined functions (UDFs)". There is no mention of Python support; hence Stream Analytics is not correct. https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction

Azure Databricks supports near real-time data from Azure Event Hubs. And includes support for R, SQL, Python, Scala, and Java. So I will go for option B.
upvoted 87 times

anto69 4 years, 6 months ago
But Python runs on Event Hubs why the other service does should support Python too?
upvoted 2 times

anto69 4 years, 6 months ago
But Python runs on Event Hubs why the other service does should support Python too?It's mentioned that "python runs on real time data from event hubs not on event hubs". Also event hub is to gather that data and after that it is analyzed by either databricks stream analytics. And since stream analytics doesn't support python so the answer is databrickstherefore i agree wih ASApython can run Event Hubs libraries real time, it doesn't have to be supported by ASA, it just needs to send data to analytics service@RoyP654, the question asks which service to perform the statistical analysis (e.g. execute the python) suggesting that the python has not/will not be ran in events hubs
upvoted 2 times

Aditya0891 4 years, 2 months ago
It's mentioned that "python runs on real time data from event hubs not on event hubs". Also event hub is to gather that data and after that it is analyzed by either databricks stream analytics. And since stream analytics doesn't support python so the answer is databrickstherefore i agree wih ASApython can run Event Hubs libraries real time, it doesn't have to be supported by ASA, it just needs to send data to analytics service@RoyP654, the question asks which service to perform the statistical analysis (e.g. execute the python) suggesting that the python has not/will not be ran in events hubs
upvoted 2 times

Aditya0891 4 years, 2 months ago

It's mentioned that "python runs on real time data from event hubs not on event hubs". Also event hub is to gather that data and after that it is analyzed by either databricks stream analytics. And since stream analytics doesn't support python so the answer is databricks

upvoted 2 times

RoyP654 3 years, 2 months ago

therefore i agree wih ASA

upvoted 2 times

RoyP654 3 years, 2 months ago

therefore i agree wih ASApython can run Event Hubs libraries real time, it doesn't have to be supported by ASA, it just needs to send data to analytics service@RoyP654, the question asks which service to perform the statistical analysis (e.g. execute the python) suggesting that the python has not/will not be ran in events hubs

upvoted 2 times

RoyP654 3 years, 2 months ago

python can run Event Hubs libraries real time, it doesn't have to be supported by ASA, it just needs to send data to analytics service

upvoted 1 times

RoyP654 3 years, 2 months ago

python can run Event Hubs libraries real time, it doesn't have to be supported by ASA, it just needs to send data to analytics service@RoyP654, the question asks which service to perform the statistical analysis (e.g. execute the python) suggesting that the python has not/will not be ran in events hubs

upvoted 1 times

ExamDestroyer69 2 years, 8 months ago

@RoyP654, the question asks which service to perform the statistical analysis (e.g. execute the python) suggesting that the python has not/will not be ran in events hubs

upvoted 1 times

anto69 4 years, 6 months ago

I'm sure it's Stream Analytics cause Event Hubs already supports Python (https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-python-get-started-send). We don't need the other service to support it. We just need to lower costs. Hence ASA is the correct solutionthe question does not ask which service can run Python, it's asking where to send the data for analytics since Python can run with Event Hubs libraries

upvoted 16 times

anto69 4 years, 6 months ago

I'm sure it's Stream Analytics cause Event Hubs already supports Python (https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-python-get-started-send). We don't need the other service to support it. We just need to lower costs. Hence ASA is the correct solution

upvoted 16 times

RoyP654 3 years, 2 months ago

the question does not ask which service can run Python, it's asking where to send the data for analytics since Python can run with Event Hubs libraries

upvoted 1 times

fbc11c2 Most Recent 1 year, 11 months ago

Should be B. See the link below. Under the heading "When to use other technologies", it mentions "Azure Stream Analytics supports user-defined functions (UDF) or user-defined aggregates (UDA) in JavaScript for cloud jobs and C# for IoT Edge jobs. C# user-defined deserializers are also supported. If you want to implement a deserializer, a UDF, or a UDA in other languages, such as Java or Python, you can use Spark Structured Streaming. You can also run the Event Hubs EventProcessorHost on your own virtual machines to do arbitrary streaming processing." As the question mentions user-defined-function (UDF) in Python, ASA seems not support UDF in Python. Should use Spark Structured Streaming, which in this case here is Azure Databricks. https://learn.microsoft.com/en-us/azure/stream-analytics/streaming-technologies

upvoted 1 times

evangelist 2 years, 1 month ago

C is wrong, stream analytics is a SQL based analytics

upvoted 1 times

e56bb91 2 years, 1 month ago

ChatGPT 4o Azure Databricks is well-suited for real-time data processing and analytics. It provides a collaborative environment for working with Apache Spark, which is ideal for performing complex statistical analyses and machine learning tasks in real-time.

upvoted 1 times

KarlGardnerDataEngineering 2 years, 2 months ago

It says we are using python after it is sent to event hubs: "custom proprietary Python functions on near real-time data FROM Azure Event Hubs". Yes, we can send events to Event Hub with python but it says that we are running statistical analysis AFTER we send it to Event Hub. Therefore, my answer is Databricks

upvoted 1 times

eb36a01 2 years, 2 months ago

Azure Stream Analytics: Azure Stream Analytics is designed for real-time data processing and can directly ingest data from Azure Event Hubs. However, it has limited support for custom Python functions. It is more suitable for simple real-time analytics and transformations rather than complex statistical analysis with custom code. correct answer: Azure Databricks, we have custom python function

upvoted 2 times

Dusica 2 years, 3 months ago

B is correct

upvoted 1 times

poesklap 2 years, 4 months ago

Azure Databricks provides a fast and scalable Apache Spark-based analytics platform that supports Python, among other programming languages. It allows you to perform near real-time data processing and analysis efficiently, making it ideal for scenarios where low latency is a priority. Additionally, it offers seamless integration with Azure Event Hubs, enabling you to ingest data in real-time and apply custom Python functions for statistical analysis.

upvoted 1 times

Elanche 2 years, 5 months ago

B. Azure Databricks Azure Databricks provides a fully managed Apache Spark-based analytics platform that is well-suited for processing and analyzing real-time streaming data. It offers native integration with Azure Event Hubs, allowing you to ingest data in real-time and apply custom Python functions for statistical analysis with minimal latency. Additionally, Databricks provides scalable compute resources, optimized processing capabilities, and support for various programming languages, making it an ideal choice for near real-time data analysis scenarios.

upvoted 1 times

moneytime 2 years, 6 months ago

C is correct. At near realtime ,the window functions in azure stream analytics can be employed in compute some statical values (e.g count,maximum,min,avg. etc) of the data streaming from the even hub.

upvoted 1 times

Azure_2023 2 years, 6 months ago

Corrected!!! FROM Azure Event Hubs, not ON Azure Databricks

upvoted 1 times

Azure_2023 2 years, 6 months ago

FROM Azure Event Hubs, not ON

upvoted 1 times

prshntdxt7 2 years, 6 months ago

chatGPT explains - Azure Stream Analytics is designed for real-time data stream processing and analytics. It can ingest data from various sources, including Azure Event Hubs, and allows you to run near real-time analytics using a SQL-like language. With Stream Analytics, you can easily apply custom Python functions using user-defined functions (UDFs) and achieve low-latency processing. Azure Synapse Analytics and Azure Databricks are powerful analytics services, but they are more suitable for complex analytics and big data processing rather than near real-time, low-latency scenarios. Azure SQL Database is a relational database service and is not specifically designed for real-time stream processing. Therefore, in this case, Azure Stream Analytics is the recommended choice for minimizing latency in statistical analysis on near real-time data from Azure Event Hubs.

upvoted 3 times

sdg2844 2 years, 7 months ago

Simply, they always want the Stream Analytics answer. It's the most straightforward.

upvoted 3 times

maxCarter 2 years, 7 months ago

Azure Databricks

upvoted 1 times

HSZ 2 years, 11 months ago

Selected Answer: C

From ChatGPT, To minimize latency for statistical analysis on near real-time data from Azure Event Hubs, I recommend using Azure Stream Analytics (Option C). Azure Stream Analytics is designed for real-time data processing and can ingest and analyze data from Event Hubs with low latency, making it a suitable choice for this scenario.

upvoted 3 times

HSZ 2 years, 11 months ago

From ChatGPT, To minimize latency for statistical analysis on near real-time data from Azure Event Hubs, I recommend using Azure Stream Analytics (Option C). Azure Stream Analytics is designed for real-time data processing and can ingest and analyze data from Event Hubs with low latency, making it a suitable choice for this scenario.Also from ChatGPT (GPT4) lol: For processing near real-time data with custom proprietary Python

functions and minimizing latency, the best service would be: B. Azure Databricks Here's why: Azure Databricks is an Apache Spark-based analytics service that integrates smoothly with Azure services such as Azure Event Hubs. It supports real-time streaming data processing and can execute custom Python code, which is necessary for your custom statistical analysis functions. Databricks is designed to handle large-scale data processing and analytics with low latency, making it suitable for near real-time scenarios. The other services have their uses but may not be the optimal choice for this particular scenario

upvoted 3 times

mav2000 2 years, 8 months ago

Also from ChatGPT (GPT4) lol: For processing near real-time data with custom proprietary Python functions and minimizing latency, the best service would be: B. Azure Databricks Here's why: Azure Databricks is an Apache Spark-based analytics service that integrates smoothly with Azure services such as Azure Event Hubs. It supports real-time streaming data processing and can execute custom Python code, which is necessary for your custom statistical analysis functions. Databricks is designed to handle large-scale data processing and analytics with low latency, making it suitable for near real-time scenarios. The other services have their uses but may not be the optimal choice for this particular scenario

upvoted 1 times

Question #14 *Topic 2*

HOTSPOT -

You have an enterprise data warehouse in Azure Synapse Analytics that contains a table named FactOnlineSales. The table contains data from the start of 2009 to the end of 2012.

You need to improve the performance of queries against FactOnlineSales by using table partitions. The solution must meet the following requirements:

✑ Create four partitions based on the order date.

✑ Ensure that each partition contains all the orders placed during a given calendar year.

How should you complete the T-SQL command? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE TABLE [dbo].FactOnlineSales
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime] NOT NULL,
[StoreKey] [int] NOT NULL,
[ProductKey] [int] NOT NULL,
[CustomerKey] [int] NOT NULL,
[SalesOrderNumber] [nvarchar](20) NOT NULL,
[SalesQuantity] [int] NOT NULL,
[SalesAmount] [money] NOT NULL,
[UnitPrice] [money] NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey] RANGE [ ▼ ] FOR VALUES
    RIGHT
    LEFT
( [ ▼ ] )
    20090101,20121231
    20100101,20110101,20120101
    20090101,20100101,20110101,20120101
```

Suggested Answer:

Answer Area

```
CREATE TABLE [dbo].FactOnlineSales
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime] NOT NULL,
[StoreKey] [int] NOT NULL,
[ProductKey] [int]    NOT NULL,
[CustomerKey] [int]   NOT NULL,
[SalesOrderNumber] [nvarchar](20) NOT NULL,
[SalesQuantity] [int] NOT NULL,
[SalesAmount] [money] NOT NULL,
[UnitPrice] [money]   NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey] RANGE [ RIGHT / LEFT ] FOR VALUES
    ( [ 20090101,20121231 / 20100101,20110101,20120101 / 20090101,20100101,20110101,20120101 ] )
```

Answer Area

```
CREATE TABLE [dbo].FactOnlineSales
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime] NOT NULL,
[StoreKey] [int] NOT NULL,
[ProductKey] [int]    NOT NULL,
[CustomerKey] [int]   NOT NULL,
[SalesOrderNumber] [nvarchar](20) NOT NULL,
[SalesQuantity] [int] NOT NULL,
[SalesAmount] [money] NOT NULL,
[UnitPrice] [money]   NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey] RANGE [ RIGHT / LEFT ] FOR VALUES
    ( [ 20090101,20121231 / 20100101,20110101,20120101 / 20090101,20100101,20110101,20120101 ] )
```

Range Left or Right, both are creating similar partition but there is difference in comparison For example: in this scenario, when you use LEFT and 20100101,20110101,20120101 Partition will be, datecol 20100101 and datecol 20110101 and datecol 20120101 But if you use range RIGHT and 20100101,20110101,20120101 Partition will be, datecol =20100101 and datecol =20110101 and datecol =20120101 In this example,

Range RIGHT will be suitable for calendar comparison Jan 1st to Dec 31st Reference: https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql?view=sql-server-ver15

Canary_2021 Highly Voted 4 years, 8 months ago

Answer is correct. https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql?view=sql-server-ver15

upvoted 27 times

victor90 4 years, 8 months ago

I think the box 2 should be 20090101,2010101,20110101,20120101 since the question asked about 4 partitions. https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql?view=sql-server-ver15#c-creating-a-range-right-partition-function-on-a-datetime-column

upvoted 18 times

victor90 4 years, 8 months ago

I think the box 2 should be 20090101,2010101,20110101,20120101 since the question asked about 4 partitions. https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql?view=sql-server-ver15#c-creating-a-range-right-partition-function-on-a-datetime-columnChoosing box 2 with range right would create five partitions. The first partition would be <20090101. So the provided answer is correctIncorrect you read the requirement: -Ensure that each partition contains all the orders placed during a given calendar year.- now as you said [time<2009-01-01] goes to the first partition and for sure in the second partition we will have only 2009No! That's wrong! Number of partitions created = Number of partition boundaries specified + 1.

upvoted 18 times

onyerleft 4 years, 8 months ago

Choosing box 2 with range right would create five partitions. The first partition would be <20090101. So the provided answer is correct

upvoted 6 times

onyerleft 4 years, 8 months ago

Choosing box 2 with range right would create five partitions. The first partition would be <20090101. So the provided answer is correctIncorrect you read the requirement: -Ensure that each partition contains all the orders placed during a given calendar year.- now as you said [time<2009-01-01] goes to the first partition and for sure in the second partition we will have only 2009

upvoted 6 times

rlnd2000 Most Recent 2 years ago

Incorrect you read the requirement: -Ensure that each partition contains all the orders placed during a given calendar year.- now as you said [time<2009-01-01] goes to the first partition and for sure in the second partition we will have only 2009

upvoted 1 times

TestMitch Highly Voted 4 years, 8 months ago

No! That's wrong! Number of partitions created = Number of partition boundaries specified + 1.

upvoted 30 times

rlnd2000 2 years ago

The given answer does not fully meet the requirement: ""Ensure"" that each partition contains all the orders placed during a given calendar year. Using the values 2010101, 20110101, 20120101 means that all dates before 2010101 will go to the first partition. If, by any chance, a new date from the year 2008 is inserted into the table, it will go to the first partition together with 2009. Therefore, I believe 20090101, 2010101, 20110101, 20120101 is a better answer.

upvoted 2 times

Bakhtiyor 2 years, 4 months ago

RIGHT - [time<2010-01-01] [2010-01-01<=time<2011-01-01] [2011-01-01<=time<2012-01-01] [time<=2012-01-01] LEFT - [time<=2010-01-01] [2010-01-01<time<=2011-01-01] [2011-01-01<time<=2012-01-01] [time<2012-01-01]

upvoted 2 times

moneytime 2 years, 6 months ago

I stand with the chosen answer. For me it is 100% correct If you chose LEFT partition with 20090101,2010101,20110101, then be aware that the fourth partition will not include 1st January 2012 which means the partition doesn't actually supports all the 12 calendar months is a particular year.

upvoted 1 times

kkk5566 2 years, 11 months ago

RIGHT : t 20120101

upvoted 2 times

Maddhy 3 years, 8 months ago

Answer is 1.right 2. Last option becoz there they mentioned 4 partitions (I'm sure that it is guarenteed)

upvoted 4 times

Maddhy 3 years, 8 months ago

Answer is 1.right 2. Last option becoz there they mentioned 4 partitions (I'm sure that it is guarenteed)The reason we are using right that is here the values are not null

upvoted 4 times

Maddhy 3 years, 8 months ago

The reason we are using right that is here the values are not null

upvoted 1 times

rzeng 3 years, 9 months ago

IF use [RIGHT], it means : [time<20100101], [20100101<=time<20110101] and so on IF use [LEFT], it means: [time<=20100101], [20100101<time <= 20110101] and so on See if you choose [LEFT] then will NOT include the 0101 value of current calendar year into the query, so GO with [RIGHT]

upvoted 9 times

Deeksha1234 4 years ago

Answer is correct.. given1st boundary value will be included in the 2nd partition (since right) so 1st partition will end at 20091231 and 2nd will start at 20100101 and end at 20101231 and so on..

upvoted 2 times

Deeksha1234 4 years ago

Answer is correct

upvoted 1 times

dsp17 4 years, 1 month ago

How to remember LEFT / RIGHT concept, it's confusing, pls helpI do that by initials of L(eft) means starting as "Less than equal to first boundary value, there onwards greater than " and for right, other way around.

upvoted 2 times

dsp17 4 years, 1 month ago

How to remember LEFT / RIGHT concept, it's confusing, pls help

upvoted 2 times

monibun 3 years, 11 months ago

I do that by initials of L(eft) means starting as "Less than equal to first boundary value, there onwards greater than " and for right, other way around.

upvoted 7 times

PallaviPatel 4 years, 6 months ago

correct Answer.

upvoted 1 times

dev2dev 4 years, 7 months ago

where does 2009 year stored? i think the 1st choice should be LEFT so ththe answer is correct, check the requirements: ✑ Create four partitions based on the order date. ✑ Ensure that each partition contains all the orders placed during a given calendar year. both right an d left results in 4 partitions, but left will have mixed values from 2 years, check the clarification in the answer. so right will result in 4 partions each partition contains all the orders placed during a given calendar year.

upvoted 4 times

dev2dev 4 years, 7 months ago

where does 2009 year stored? i think the 1st choice should be LEFT so th

upvoted 4 times

allagowf 3 years, 10 months ago

the answer is correct, check the requirements: ✑ Create four partitions based on the order date. ✑ Ensure that each partition contains all the orders placed during a given calendar year. both right an d left results in 4 partitions, but left will have mixed values from 2 years, check the clarification in the answer. so right will result in 4 partions each partition contains all the orders placed during a given calendar year.

upvoted 3 times

VeroDon 4 years, 7 months ago

correct

upvoted 2 times

alexleonvalencia 4 years, 8 months ago

Respuesta correcta. RIGTH, [3 VALORES].

upvoted 2 times

Question #15 *Topic 2*

You need to implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool.
You have a table that was created by using the following Transact-SQL statement.
Which two columns should you add to the table? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

```
CREATE TABLE [DBO].[DimProduct] (
[ProductKey] [int] IDENTITY(1,1) NOT NULL,
[ProductSourceID] [int] NOT NULL,
[ProductName] [nvarchar] (100) NULL,
[Color] [nvarchar] (15) NULL,
[SellStartDate] [date] NOT NULL,
[SellEndDate] [date] NULL,
[RowInsertedDateTime] [datetime] NOT NULL,
[RowUpdatedDateTime] [datetime] NOT NULL,
[ETLAuditID] [int] NOT NULL
)
```

A. [EffectiveStartDate] [datetime] NOT NULL,

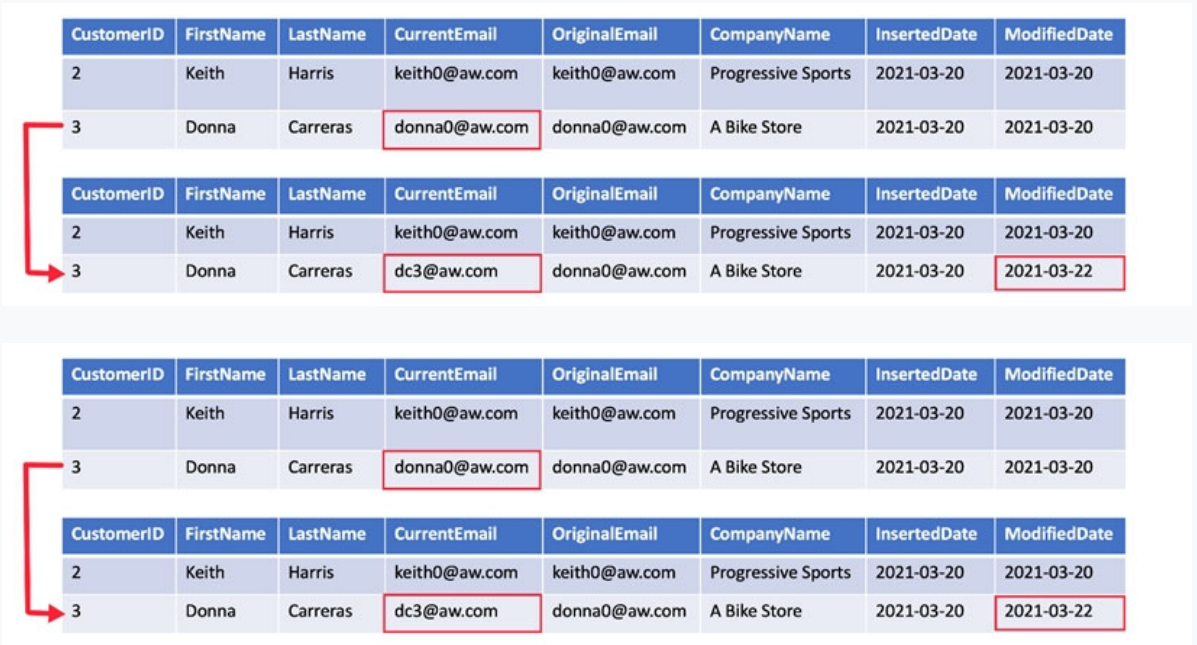
B. [CurrentProductCategory] [nvarchar] (100) NOT NULL,

C. [EffectiveEndDate] [datetime] NULL,

D. [ProductCategory] [nvarchar] (100) NOT NULL,

E. [OriginalProductCategory] [nvarchar] (100) NOT NULL,

Suggested Answer: *B, E*



A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD. This type of tracking may be used for one or two columns in a dimension table. It is not common to use it for many members of the same table. It is often used in combination with Type 1 or Type 2 members.

Reference: https://k21academy.com/microsoft-azure/azure-data-engineer-dp203-q-a-day-2-live-session-review/

Community vote distribution

BE (100%)

Maunik Highly Voted 4 years, 5 months ago

Correct answer

upvoted 29 times

lukeonline Highly Voted 4 years, 1 month ago

Why can't the name of the current ProductCategory be just "ProductCategory"? I would say that D and E could be also correct.because if you look at Type 3 examples, usually there are "original" and "current".

upvoted 14 times

lukeonline 4 years, 1 month ago

Selected Answer: BE

Why can't the name of the current ProductCategory be just "ProductCategory"? I would say that D and E could be also correct.

upvoted 14 times

Dicer 3 years, 6 months ago

because if you look at Type 3 examples, usually there are "original" and "current".

upvoted 9 times

sdg2844 Most Recent 2 years, 1 month ago

This is correct. The ProductCategory may represent a current product category or an original one, but we don't know. Wording on this question is a little off.

upvoted 1 times

kkk5566 2 years, 5 months ago

BE is Correct

upvoted 1 times

Ankit_Az 2 years, 8 months ago

Correct

upvoted 1 times

esaade 2 years, 11 months ago

To implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool and add the required columns to the existing table, you should add the following columns: A. [EffectiveStartDate] [datetime] NOT NULL, to track the start date of the current product category value. D. [ProductCategory] [nvarchar] (100) NOT NULL, to store the current product category value. Explanation: A Type 3 SCD tracks both the current and previous values of a column. For the product category data, you need to store the current product category value, as well as the previous/original value. To achieve this, you need to add the following columns to the existing table: A. [EffectiveStartDate] [datetime] NOT NULL, to track the start date of the current product category value. This column will store the date when the current product category value became effective. D. [ProductCategory] [nvarchar] (100) NOT NULL, to store the current product category value. This column will contain the most recent value for the product category.

upvoted 1 times

UristMcFarmer 3 years, 1 month ago

BE is how you should answer, but in reality DE is better. The end user / analyst shouldn't have to remember "oh, this is a Type 3 SCD field so I need to look under 'C' for CurrentProductCategory instead of 'P' for ProductCategory."

upvoted 1 times

vrodriguesp 3 years, 1 month ago

Microsoft documentation is pretty confusing (https://learn.microsoft.com/en-us/training/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types) pictures have no sense. I find at this link a good explanation: https://medium.com/geekculture/6-different-types-of-slowly-changing-dimensions-and-how-to-apply-them-b152ef908d4e

upvoted 1 times

SomethingRight100 3 years, 2 months ago

This is the same question as Question #59Topic 1

upvoted 5 times

anks84 3 years, 5 months ago

Given Answer is correct !!

upvoted 1 times

Deeksha1234 3 years, 6 months ago

BE is correct

upvoted 1 times

Sriramiyer92 3 years, 7 months ago

Right Answer!

upvoted 1 times

Fishy_Marcy 3 years, 7 months ago

the explanation is confusing as email is more likely to be a type2 change, it is also confusing as we need to add the "original" category (it should already be there as ProductCategory right?). So my first guess was only to add ProductCategory, but that would not be a scd change, only an initialization of a variable that could be subject to a scd-3 change. SO I guess the only viable option is adding "original" and "new" categories. I do not like scd3 type changes anyway, maybe there should be a special scd type2 historized scdtype3 dimension ? Okay, too much, lets stick with this answer BE (It says that the change is also metioning the Previous ProductCategory key wasn't there yet, but should have been)

upvoted 1 times

Egocentric 3 years, 10 months ago

BE is correct

upvoted 1 times

SebK 3 years, 11 months ago

Correct

upvoted 1 times

VeroDon 4 years, 1 month ago

correct

upvoted 1 times

dija123 4 years, 2 months ago

B and E

upvoted 4 times

Question #16 *Topic 2*

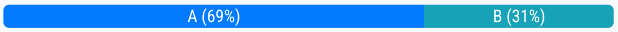
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an Azure Stream Analytics solution that will analyze Twitter data.
You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.
Solution: You use a hopping window that uses a hop size of 10 seconds and a window size of 10 seconds.
Does this meet the goal?

A. Yes

B. No

Suggested Answer: *B*
Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. Reference: https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

Community vote distribution



medsimus Highly Voted 4 years, 10 months ago
Unlike tumbling windows, hopping windows model scheduled overlapping windows. A hopping window specification consist of three parameters: the timeunit, the windowsize (how long each window lasts) and the hopsize (by how much each window moves forward relative to the previous one). Additionally, offsetsize may be used as an optional fourth parameter. Note that a tumbling window is simply a hopping window whose 'hop' is equal to its 'size'.
upvoted 19 times

Databrick Highly Voted 3 years, 10 months ago
Note that a tumbling window is simply a hopping window whose 'hop' is equal to its 'size'. https://learn.microsoft.com/en-us/stream-analytics-query/hopping-window-azure-stream-analytics
upvoted 6 times

639b1cb Most Recent 1 year, 8 months ago
Correct Answer: 2. No To meet the goal, you should use a tumbling window with a window size of 10 seconds, which would guarantee that each tweet is counted exactly once within each 10-second period.
upvoted 1 times

AzureSQLDatabase 1 year, 8 months ago
This hopping is just gonna work like a regular tumbling. https://learn.microsoft.com/en-us/stream-analytics-query/hopping-window-azure-stream-analytics
upvoted 1 times

Danweo 2 years, 1 month ago
Overlapping only applies if the window size is larger than the hop size. Hopping is the same as tumbling with an equal hop to window size.
upvoted 1 times

ageorgieva 2 years, 2 months ago
The answer is B:No. Tumbling window should be used. With hopping window there is an overlap and an event can be counted more than once.
upvoted 1 times

Dusica 2 years, 3 months ago
Yes is correct
upvoted 1 times

Alongi 2 years, 4 months ago
A. Yes A tumbling window is the same as an hopping window with the window size equal to the hop size.
upvoted 1 times

Elanche 2 years, 5 months ago
B. No Using a hopping window with a hop size of 10 seconds and a window size of 10 seconds wouldn't ensure that each tweet is counted only once. A hopping window with these parameters would result in overlapping windows, which means that a tweet might fall into multiple windows and could potentially be counted multiple times. To ensure that each tweet is counted only once within a specific 10-second window, you should use a tumbling

window with a size of 10 seconds. Tumbling windows are non-overlapping and fixed-size, ensuring that each tweet is counted within a single, distinct window.

upvoted 2 times

moneytime 2 years, 6 months ago

Answer is Yes. The solution actually defined Tumbling window using hopping window. The condition is ; if the hopping size equals window size,then the window is also Tumbling.

upvoted 1 times

Azure_2023 2 years, 6 months ago

To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.

upvoted 2 times

sdg2844 2 years, 7 months ago

While a tumbling window would be more straightforward, a Hopping window with 10/10 parameters will produce the same result, as it is set not to overlap.

upvoted 2 times

sr0017 2 years, 7 months ago

risposta B corretta

upvoted 1 times

Amir_Abed 2 years, 7 months ago

Answer should be B

upvoted 1 times

6d954df 2 years, 7 months ago

A. Yes The proposed solution meets the goal. In Azure Stream Analytics, a hopping window generates output every hop size interval, and it aggregates events for the window size period. If both the hop size and window size are set to 10 seconds, the system will count the tweets in each 10-second window, ensuring each tweet is counted only once. This is because the window “hops” forward by the specified hop size (10 seconds in this case) and does not overlap with the next window. Therefore, each tweet will fall into exactly one window and will be counted once. This makes the hopping window suitable for this scenario.

upvoted 1 times

positivitypeople 2 years, 8 months ago

Got this question today on the exam

upvoted 2 times

Momoanwar 2 years, 8 months ago

Chatgpt : No The answer is **B. No**, this does not meet the goal. Using a hopping window with both the hop size and the window size set to 10 seconds would result in each tweet being counted multiple times as each tweet could appear in several windows. To ensure each tweet is counted only once, the hop size should be equal to the window size or a different type of windowing, such as tumbling windows, should be used.

upvoted 2 times

Question #17 *Topic 2*

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an Azure Stream Analytics solution that will analyze Twitter data.
You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.
Solution: You use a hopping window that uses a hop size of 5 seconds and a window size 10 seconds.
Does this meet the goal?

A. Yes

B. No

Suggested Answer: *B*
Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. Reference: https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

Community vote distribution
B (100%)

Ati1362 Highly Voted 4 years, 8 months ago
answer is correct
upvoted 20 times

allagowf Highly Voted 3 years, 4 months ago
if the hop size is equivalent to the window size then it can be true, but because the hop size is smaller, then each tweet can be count more than one and the windows will overlap with each others.
upvoted 5 times

ELJORDAN23 Most Recent 2 years, 1 month ago
Got this question on my exam on january 17, answer B is correct.
upvoted 1 times

sdg2844 2 years, 1 month ago
answer is correct.
upvoted 1 times

positivitypeople 2 years, 1 month ago
Got this question today on the exam
upvoted 1 times

MioHaze 2 years, 5 months ago
correct
upvoted 1 times

kkk5566 2 years, 5 months ago
B is correct
upvoted 1 times

Ankit_Az 2 years, 8 months ago
Correct
upvoted 1 times

DindaS 3 years ago
definitely No. as the twits will be counted more than once in this solution
upvoted 2 times

Deeksha1234 3 years, 6 months ago
CORRECR
upvoted 1 times

Dicer 3 years, 6 months ago
No If solution is: You use a hopping window that uses a hop size of 10 seconds and a window size 10 seconds, it is correct.

upvoted 1 times

Dicer 3 years, 6 months ago

Answer is B. The solution must ensure that each tweet is counted only once. Therefore, it means no overlapping. Hence, the better solution is tumbling-window. The answer is B

upvoted 1 times

rl_1871 3 years, 7 months ago

B is the correct answer.

upvoted 1 times

Egocentric 3 years, 10 months ago

hop size must be the same as window size... so B is correct

upvoted 1 times

Yohannesmulu 3 years, 11 months ago

B is the correct answer.

upvoted 2 times

scass 3 years, 12 months ago

Correct answer is B

upvoted 2 times

jh777 4 years ago

correct B

upvoted 1 times

Question #18 *Topic 2*

HOTSPOT -

You are building an Azure Stream Analytics job to identify how much time a user spends interacting with a feature on a webpage.

The job receives events based on user actions on the webpage. Each row of data represents an event. Each event has a type of either 'start' or 'end'.

You need to calculate the duration between start and end events.

How should you complete the query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

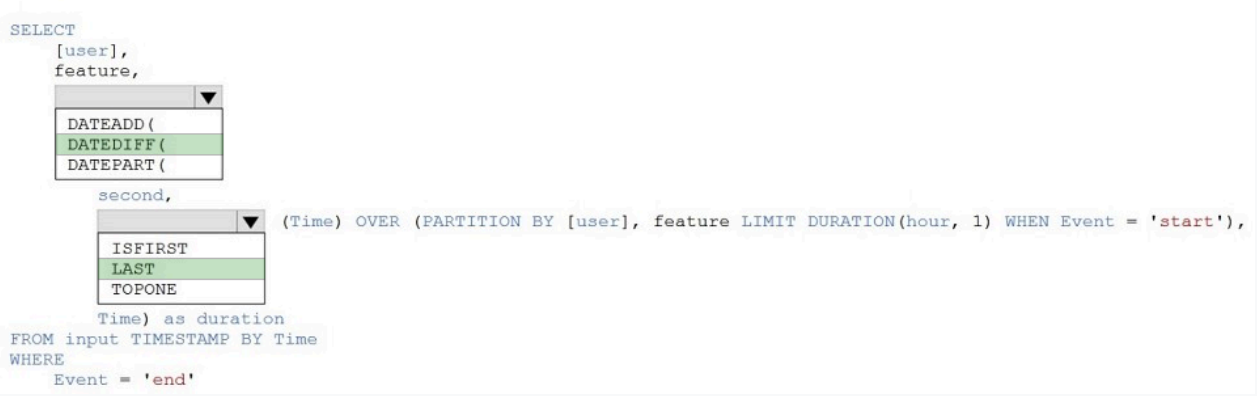
Hot Area:

Answer Area

```
SELECT
    [user],
    feature,
    [ DATEADD( | DATEDIFF( | DATEPART( ]
        second,
        [ ISFIRST | LAST | TOPONE ] (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
        Time) as duration
FROM input TIMESTAMP BY Time
WHERE
    Event = 'end'
```

Suggested Answer:

Answer Area



Answer Area



Box 1: DATEDIFF - DATEDIFF function returns the count (as a signed integer value) of the specified datepart boundaries crossed between the specified startdate and enddate. Syntax: DATEDIFF (datepart , startdate, enddate)

Box 2: LAST - The LAST function can be used to retrieve the last event within a specific condition. In this example, the condition is an event of type Start, partitioning the search by PARTITION BY user and feature. This way, every user and feature is treated independently when searching for the Start event. LIMIT DURATION limits the search back in time to 1 hour between the End and Start events. Example:

SELECT - [user], feature, DATEDIFF(second, LAST(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),

Time) as duration -

FROM input TIMESTAMP BY Time -

WHERE - Event = 'end' Reference: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns

Miris Highly Voted 4 years, 8 months ago

correct

upvoted 44 times

ohana Highly Voted 4 years, 3 months ago

Took the exam today. This question came out. Ans: DateDiff, Lasthow do you know?https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns#detect-the-duration-between-events

upvoted 27 times

ohana 4 years, 3 months ago

Took the exam today. This question came out. Ans: DateDiff, Last

upvoted 27 times

romanzdk 4 years, 1 month ago

how do you know?https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns#detect-the-duration-between-events

upvoted 2 times

romanzdk 4 years, 1 month ago

how do you know?

upvoted 2 times

assU2 4 years ago

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns#detect-the-duration-between-events

upvoted 13 times

Gikan Most Recent 2 years ago

Answer should be correct, because the link is there, but what a freak SQL is it? The original T-SQL is LAST_VALUE(COLUMN) OVER(ORDER BY ...)! If I search the internet for this solution, the only page where it exists is the link in the answer!

upvoted 2 times

kkk5566 2 years, 5 months ago

Ans: DateDiff, Last

upvoted 1 times

semauni 2 years, 6 months ago

Why is the answer to the second blank LAST() instead of TOPONE(), when this is about the startdate?

upvoted 1 times

Deeksha1234 3 years, 6 months ago

answer is correct

upvoted 2 times

RajeshAzure 3 years, 6 months ago

When Event ='Start' should not be there in this question

upvoted 1 times

PallaviPatel 4 years ago

correct answer.

upvoted 1 times

hugoborda 4 years, 4 months ago

The answer is correct

upvoted 3 times

mdalorso 4 years, 6 months ago

This is Stream Analytics Query Language, a little different than tsql https://docs.microsoft.com/en-us/stream-analytics-query/last-azure-stream-analytics

upvoted 5 times

mdalorso 4 years, 6 months ago

This is Stream Analytics Query Language, a little different than tsql https://docs.microsoft.com/en-us/stream-analytics-query/last-azure-stream-analyticsso is the answer DATEDIFF+LAST incorrect then?DATEDIFF and LAST are correct, please refer to: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns ("Detect the duration between events")

upvoted 5 times

AvithK 4 years, 6 months ago

so is the answer DATEDIFF+LAST incorrect then?DATEDIFF and LAST are correct, please refer to: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns ("Detect the duration between events")

upvoted 1 times

AvithK 4 years, 6 months ago

so is the answer DATEDIFF+LAST incorrect then?

upvoted 1 times

Juan27 4 years, 5 months ago

DATEDIFF and LAST are correct, please refer to: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns ("Detect the duration between events")

upvoted 4 times

vlad888 4 years, 7 months ago

The query has no sense, at least if it is T-SQL. Look: each row is end event or start event. How window function (Last() over partition) can get start event if there is where condition that filter out end event only???

upvoted 9 times