

DP-203 Data Engineering on Microsoft Azure

店铺：IT认证考试服务

店铺：IT认证考试服务

Topic 1 - Question Set 1

 Custom View Settings

You have a table in an Azure Synapse Analytics dedicated SQL pool. The table was created by using the following Transact-SQL statement.

```
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
)
```

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?

- A. [ManagerEmployeeID] [smallint] NULL
- B. [ManagerEmployeeKey] [smallint] NULL
- C. [ManagerEmployeeKey] [int] NULL
- D. [ManagerName] [varchar](200) NULL

Correct 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 (100%)

✉  **jasonsmithss** Highly Voted 1 month, 2 weeks ago

itexamslab.com

c is correct

upvoted 75 times

✉  **gihey13844** 1 month ago

The success that I achieved in passing my Microsoft DP-203 exam was a huge career milestone for me. I couldn't have done it without the help of this exam dumps and practice tests

upvoted 1 times

✉  **mageyi7032** 1 month, 1 week ago

this website is valid i pass from it will get more exam next January

upvoted 19 times

✉  **nateb27235** 1 month ago

iam also taken this wish me luck

upvoted 1 times

✉  **Ycombo** 1 month ago

How was it and how many from here still in Dec 2023 end ? Hope you passed ?

upvoted 1 times

✉  **buxalo** Highly Voted 1 month ago

Got certified today. Dumps are still viable as of September 2023

<https://www.dumps4azure.com/>

upvoted 50 times

✉  **Ashishprajapati** Most Recent 1 week, 1 day ago

Do i need to buy paid subscription to prepare from the dumps ?

upvoted 1 times

 **justin_red** 1 month ago

itexamstest.com

Answer is correct

upvoted 13 times

 **xidowi9102** 1 month, 1 week ago

I gave the Microsoft Azure DP-203 exam and studied from this dump as it has authentic and valid practice questions available which helped me pass the exam by 895/1000. Thanks a lot: <https://rb.gy/ly3s01>

upvoted 21 times

 **Yiworo** 1 month ago

Really thanks for your suggestion. I am glad that I selected this source and get 92%. I would recommend it a 100% Thanks again

upvoted 1 times

 **Anwana** 2 months, 3 weeks ago

Selected Answer: C
You need a key to link the Manager to an employee

upvoted 1 times

 **74gjd_37** 4 months ago

Selected Answer: C
This is implied from the schema

upvoted 1 times

 **gingerbread_person** 4 months, 1 week ago

Selected Answer: C
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 3 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: C
c is correct

upvoted 1 times

 **cloudrain** 5 months, 4 weeks ago

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

upvoted 1 times

 **steveo123** 6 months, 3 weeks ago

Selected Answer: C
C correct

upvoted 1 times

 **DindaS** 11 months, 3 weeks ago

C as the data types of the primary key should be same for the manager.

upvoted 1 times

 **gfssss** 11 months, 4 weeks ago

Selected Answer: C
CCCCCC

upvoted 1 times

 **rajesh20200904** 1 year ago

Selected Answer: C
Both EmployeeKey and EmployeeId are int, so the ManagerEmployeeKey or ManagerEmployeeId also should be int. So C is correct

upvoted 1 times

 **vigilante89** 1 year ago

Selected Answer: C
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 4 times

 **vigilante89** 1 year ago

Selected Answer: C

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

 Mansoorsheikh99 1 year, 1 month ago

Selected Answer C

upvoted 1 times

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.

EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

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?

- A. 24
- B. an error
- C. a null value

Correct 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

B (73%)

A (27%)

✉️  **dmitriypo** Highly Voted 1 year, 2 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 46 times

✉️  **SenMia** 3 weeks, 2 days ago

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

upvoted 1 times

✉️  **devnginx** 1 month, 2 weeks ago

i think B option is the correct too

upvoted 1 times

✉️  **vinicius_cb** 10 months, 4 weeks ago

Actually the error it is because of the lower case.

"Table names will be converted to lower case and need to be queried using the lower case".

There is nothing wrong with "dbo", since the docs shows this exact same example and contains the "dbo" in the query.

Source: <https://learn.microsoft.com/en-us/azure/synapse-analytics/metadata/table>

upvoted 15 times

✉️  **Shaik_Shahul** 2 months, 3 weeks 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 4 times

 **gerrie1979** Highly Voted 1 year, 2 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 41 times

 **maximilianogarcia6** 1 year, 1 month ago

Did you tried the same query that is presented here? with "mytestdb.dbo.myParquetTable"??

upvoted 4 times

 **Virul** 11 months ago

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

Answer is A

upvoted 3 times

 **yogiazzaad** 11 months, 2 weeks ago

The table and Column names are case insensitive.

upvoted 3 times

 **sdg2844** Most Recent 1 week ago

Selected Answer: B

There are multiple issues here. In the select line, there is a dash after EmployeeID. I don't see any case issues with this, but still have to think it would error out with that dash... although I'll bet that's a typo.

B is the safest guess here.

upvoted 1 times

 **lisa710** 2 weeks, 6 days ago

the answer is B. The query filters on name = 'Alice', but the actual column name in the table is EmployeeName (case-sensitive).

upvoted 1 times

 **dakku987** 3 weeks, 2 days ago

Selected Answer: B

yes getting error bcz of dbo so ans is b

upvoted 1 times

 **ChrisGe1234** 1 month ago

Selected Answer: A

dbo is the default schema for replicated dbs

upvoted 1 times

 **AlphaBoy79** 1 month, 1 week ago

The reason is that the table mytestdb.myParquetTable is created using Apache Spark and the Parquet format, but you are trying to query it using serverless SQL pool syntax (mytestdb.dbo.myParquetTable).

upvoted 2 times

 **Abdulwahab1983** 1 month, 3 weeks ago

Once a database has been created by a Spark job, you can create tables in it with Spark that use Parquet, Delta, or CSV 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.

upvoted 1 times

 **EduardPaul** 1 month, 3 weeks ago

A is correct, see here (same sample from MS):

<https://learn.microsoft.com/en-us/azure/synapse-analytics/metadata/table#create-a-managed-table-in-spark-and-query-from-serverless-sql-pool>

upvoted 2 times

 **fadaei** 4 days, 22 hours ago

I used also this example in MS and got the same answer(A. 24)

upvoted 1 times

 **jhargett1** 2 months, 2 weeks ago

The given query is trying to select the EmployeeID from the Parquet table myParquetTable in the mytestdb database, where the EmployeeName is 'Alice'. However, it seems like there's a minor typo in the query. The query should be like this:

```
SELECT EmployeeID  
FROM mytestdb.myParquetTable  
WHERE EmployeeName = 'Alice';
```

The given query is trying to select the EmployeeID from the Parquet table myParquetTable in the mytestdb database, where the EmployeeName is 'Alice'. However, it seems like there's a minor typo in the query. The query should be like this:

sql

Copy code

```
SELECT EmployeeID
```

FROM mytestdb.myParquetTable
WHERE EmployeeName = 'Alice';
The corrected query would return:

A. 24

So, the correct answer is A. The query will return the EmployeeID, which is 24.

upvoted 3 times

 **alphilla** 2 months, 2 weeks ago

Answer is A guys if we ignore the probably type "-". <https://learn.microsoft.com/en-us/azure/synapse-analytics/metadata/table> There is an exact example from microsoft

upvoted 3 times

 **Shaik_Shahul** 2 months, 3 weeks ago

No it is not a error, the correct answer is A=24

upvoted 3 times

 **Katiane** 2 months, 3 weeks ago

Right answer is B.

The query must run into Serverless SQL poll, not into Apache spark.

"One minute later, you execute the following query from a serverless SQL pool in MyWorkspace."

If we run that query into apache spark pool, using notebook for example, we must use "SELECT database.table".

So, according the question, we must use serverless sql pool and, cause of that, we have to use "SELECT database.dbo.table" OR "use database; select table"

upvoted 1 times

 **AlejandroU** 3 months ago

It seems A, if it is created as a "managed table" and ignoring the probable typing error [-] in the select statement. A similar example of creating a "managed table" is in the section "Create a managed table in Spark and query from serverless SQL pool" in the link below:

<https://learn.microsoft.com/en-us/azure/synapse-analytics/metadata/table#expose-a-spark-table-in-sql>

upvoted 1 times

 **tpositive** 3 months ago

A is the answer

upvoted 1 times

 **74gjd_37** 4 months ago

Selected Answer: B

First, the name "MyWorkspace" is invalid; it cannot contain uppercase character; it could only have been "myworkspace"

upvoted 1 times

 **MBRSDG** 4 months ago

did anyone notice the dash before the col name? It returns a syntax error, hence B is the correct answer.

upvoted 3 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).

Correct 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).

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>

Given answer D A C is correct.

upvoted 42 times

 **svik** 2 years, 4 months ago

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

upvoted 2 times

 **vigilante89**  1 year 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 <source table> SWITCH PARTITION <partition number> to <destination table>

Step 3:

Delete the SalesFact_Work table.

upvoted 8 times

 **the_frix**  1 month, 1 week 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 1 times

 **kkk5566** 4 months, 1 week ago

correct

upvoted 1 times

 **rocky48** 6 months ago

Given answer - Options D A C are correct.

upvoted 1 times

 **UzairMir** 8 months, 2 weeks ago

Hi

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

Thanks

upvoted 3 times

 **Xarvastia** 8 months ago

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

upvoted 3 times

 **Aaron1234** 8 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

 **vrodriguesp** 11 months, 3 weeks ago

Given answer is correct:

Step 1:

Create an empty table SalesFact_Work with same schema as SalesFact (that it will contains records 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 3 times

 **Dusica** 12 months ago

D A C IS CORRECT

upvoted 1 times

 **Jawidkaderi** 1 year, 3 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** 1 year, 4 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

□ **supriyako** 1 year, 2 months ago

F seems wrong as it says CTAS to copy the data

upvoted 1 times

□ **pmc08** 1 year, 4 months ago

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

upvoted 1 times

□ **Deeksha1234** 1 year, 4 months ago

correct ans.

upvoted 2 times

□ **mkthoma3** 1 year, 7 months ago

D,A,C

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

upvoted 1 times

□ **Dothy** 1 year, 8 months ago

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.

Step 3: Drop the SalesFact_Work table.

upvoted 1 times

□ **JJdeWit** 1 year, 9 months ago

D A C is the right option.

For more information, this doc discusses exactly this example: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition>

upvoted 1 times

□ **theezin** 1 year, 9 months ago

Why not included deleting sales data older than 36 months which is mentioned in question?

upvoted 1 times

□ **RamGhase** 1 year, 11 months ago

i could not understand how answer handled to remove data before 36 month

upvoted 2 times

□ **gerard** 1 year, 11 months ago

you have to move the partitions that contains the date before 36 months

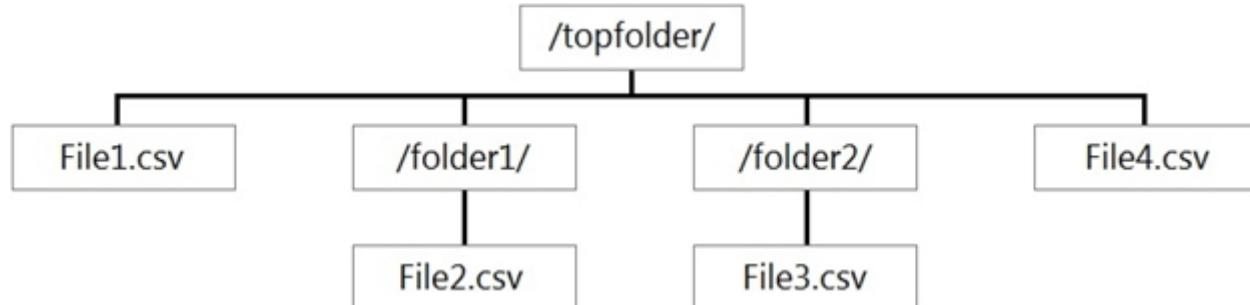
upvoted 4 times

□ **PallaviPatel** 1 year, 11 months ago

D A C is correct.

upvoted 1 times

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

Correct 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 (70%)

C (28%)

✉ **Chillem1900** Highly Voted 2 years, 8 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 115 times

✉ **captainpike** 2 years, 2 months ago

I tested and prove 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 36 times

✉ **alain2** Highly Voted 2 years, 7 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 27 times

✉ **Preben** 2 years, 7 months ago

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

upvoted 21 times

✉ **captainpike** 2 years, 2 months ago

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

upvoted 3 times

✉ **bomafrique** Most Recent 5 days, 17 hours ago

Answer B is correct for me too.

upvoted 1 times

✉ **sdg2844** 1 week ago

Selected Answer: B

Agree it should be B. The question is a little off, because they don't specify whether using or not using a wildcard to do so. Assuming by default then, no wildcard is used, only those top-level files will be returned.

upvoted 1 times

✉ **lisa710** 2 weeks, 6 days ago

answer c is correct

upvoted 1 times

✉ **bInak32** 1 month, 1 week ago

Selected Answer: B

Strongly B: (solid reason with reference)

1. This query uses Serverless Pool and it is only available for native External Table

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#arguments-create-external-table>

2. "native external tables don't return subfolders unless you specify /** at the end of path"

https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=serverless#location--folder_or_filepath-1

upvoted 1 times

✉ **Shanuramasubbu** 1 month, 2 weeks ago

The answer is B based on the below doc

<https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=sql-server-ver16&tabs=dedicated>

upvoted 1 times

✉ **y154707** 2 months ago

Question says "You create an external table named ExtTable that has LOCATION='/topfolder/'. "

Based on this link: https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=serverless#location--folder_or_filepath-1

"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 (.)."

So based on this, the answer is correct. When created, the ExtTable get data from files on the topfolder and all of its subfolders, thus when queried it would return the data from all the files.

upvoted 1 times

✉ **phydev** 2 months, 1 week ago

Was on my exam today (31.10.2023).

upvoted 1 times

✉ **74gjd_37** 4 months ago

Selected Answer: B

Since there were no wildcards (**) no nested folders are used.

upvoted 1 times

✉ **Ram9198** 4 months ago

Selected Answer: A

Native external table

upvoted 1 times

✉ **kkk5566** 4 months, 1 week ago

Selected Answer: B

correct

upvoted 1 times

✉ **lfss** 4 months, 4 weeks ago

the correct answer is B

upvoted 1 times

✉ **akhil5432** 5 months, 1 week ago

Selected Answer: B

MOST SUITED ans is option B

upvoted 1 times

✉ **Deeksha1234** 7 months ago

Selected Answer: B

B should be the answer

upvoted 1 times

✉ **VikkiC** 7 months ago

This documentation confirmed the answer B is correct.

https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=azure-sqldw-latest&tabs=serverless#location--folder_or_filepath-1

upvoted 2 times

✉ **bakamon** 7 months, 3 weeks ago

When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, only File1.csv and File4.csv will be returned [B].

When you create an external table with LOCATION='/topfolder/', only the files that are directly under the specified folder will be returned. In this case, only File1.csv and File4.csv are directly under the /topfolder/ directory and will be returned when querying ExtTable.

upvoted 1 times

店铺: IT认证考试服务

店铺: IT认证考试服务

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

Report1:

Avro
CSV
Parquet
TSV

Report2:

Avro
CSV
Parquet
TSV

Answer Area

Report1:

Avro
CSV
Parquet
TSV

Correct Answer:

Report2:

Avro
CSV
Parquet
TSV

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 2 years, 7 months ago

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

upvoted 190 times

 **terajuana** 2 years, 7 months ago

the web is full of old information. timestamp support has been added to parquet
upvoted 8 times

✉  **vlad888** 2 years, 6 months ago

Ok, but in 1st case we need only 3 of 50 columns. Parquet is columnar format. In 2nd Avro because ideal for read full row
upvoted 24 times

✉  **XiltroX** 1 year, 1 month ago

Thanks for the video share, this really helps. Cheers.
upvoted 2 times

✉  **azurestudent1498** 1 year, 8 months ago

this is correct.
upvoted 2 times

✉  **Himlo24**  2 years, 8 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 29 times

✉  **sdg2844**  1 week ago

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

✉  **lisa710** 2 weeks, 6 days 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

✉  **matiandal** 3 months ago

--> TLDR <--
AVRO PARQUET ORC
Anal. Queries v
Write Ops (ETL ops) v
Nested Data v
ACID Properties v
Sch.Flexibility v
upvoted 1 times

✉  **hassexat** 4 months ago

1. Parquet --> Column format
2. AVRO --> Row format with timestamp type
upvoted 1 times

✉  **kkk5566** 4 months, 1 week ago

1- Parquet 2- Parquet
upvoted 1 times

✉  **kkk5566** 4 months ago

sorry 2 avro
upvoted 1 times

✉  **endeesa** 8 months, 1 week ago

Report 1: parquet
Report 2: Avro
upvoted 2 times

✉  **rocky48** 8 months, 2 weeks ago

1: Parquet - column-oriented binary file format
2: AVRO - Row based format
upvoted 3 times

✉  **DipikaChavan** 8 months, 4 weeks ago

It should be Parquet and Avro
upvoted 2 times

✉  **deutscher** 9 months ago

I completely agree,
Parquet is column based
AVRO is row based
upvoted 1 times

 **bubby248** 11 months ago

Parquet,Avro
upvoted 2 times

 **Venub28** 12 months ago

Parquet
AVRO
upvoted 2 times

 **akk_1289** 1 year 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

 **Yamarh** 1 year ago

Parquet - Avro
upvoted 1 times

 **vigilante89** 1 year ago

Report 1: PARQUET
Because parquet is a columnar file format file.

Report 2: AVRO

Because Avro is a row-based file format (as Json) which is connected to logical timestamp.

upvoted 2 times

 **Deeksha1234** 1 year, 4 months ago

1. Parquet
2. AVRO
upvoted 3 times

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

Correct 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 (100%)

 **sagga** Highly Voted 2 years, 8 months ago

D is correct

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices#batch-jobs-structure>

upvoted 60 times

 **sdg2844** Most Recent 1 week ago

Selected Answer: D

Definitely D.

upvoted 1 times

 **dakku987** 3 weeks, 2 days ago

Selected Answer: D

bcz i said so

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: D

is correct

upvoted 1 times

 **akhil5432** 5 months, 1 week ago

Selected Answer: D

D is correct

upvoted 1 times

 **tbluhm** 8 months, 1 week 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** 8 months, 2 weeks ago

Selected Answer: D

Should be D.

upvoted 2 times

 **pedrOliveira** 8 months, 3 weeks ago

Selected Answer: D

D is correct

upvoted 1 times

DindaS 11 months, 3 weeks ago

yes its D :)

upvoted 1 times

vigilante89 1 year ago

Selected Answer: D

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 1 year, 1 month ago

Selected Answer: D

D is correct

upvoted 1 times

Deeksha1234 1 year, 4 months ago

Selected Answer: D

correct

upvoted 1 times

examtopicscap 1 year, 5 months ago

Selected Answer: D

D is correct

upvoted 1 times

StudentFromAus 1 year, 6 months ago

Selected Answer: D

Correct

upvoted 1 times

Dothy 1 year, 8 months ago

D is correct

upvoted 1 times

Olukunmi 1 year, 8 months ago

Selected Answer: D

D is correct

upvoted 1 times

Egocentric 1 year, 8 months 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

Answer Area

Columnar format:

Avro
GZip
Parquet
TXT

Correct Answer:

JSON with a timestamp:

Avro
GZip
Parquet
TXT

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 2 years ago

Parquet and AVRO is correct option
upvoted 39 times

↪  **h4y** Most Recent 1 month ago

Parquet and Avro are correct
upvoted 1 times

↪  **shaneWatson311** 2 months ago

I am feeling extremely happy, today I just passed my DP-203 exam by scoring 927/1000. Found only 5 new questions. One of them was the question about case study. Thanks to ET and pass4surehub for helping me out. <https://rb.gy/j74a3n>
upvoted 1 times

↪  **hassexat** 4 months ago

The provided answer is correct: Parquet & AVRO
upvoted 2 times

↪  **kkk5566** 4 months, 1 week ago

1.parquet 2avro
upvoted 1 times

↪  **rocky48** 8 months, 2 weeks 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** 11 months ago

Parquet,Avro
upvoted 1 times

↪  **vigilante89** 1 year 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** 1 year, 4 months ago

Parquet and AVRO is correct.
upvoted 1 times

↪  **Deeksha1234** 1 year, 4 months ago

correct
upvoted 1 times

↪  **Raza12** 1 year, 5 months ago

Parquet and AVRO is correct option
upvoted 3 times

↪  **Dothy** 1 year, 8 months ago

agree with the answer
upvoted 3 times

↪  **RalphLiang** 1 year, 10 months ago

Parquet and AVRO is correct option
upvoted 2 times

↪  **PallaviPatel** 1 year, 11 months ago

correct
upvoted 1 times

↪  **Skyrocket** 1 year, 11 months ago

Parquet and AVRO is right.
upvoted 2 times

↪  **edba** 2 years 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** 2 years ago

agree with the answer

upvoted 2 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

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

Copy behavior:

- Flatten hierarchy
- Merge files
- Preserve hierarchy

Sink file type:

- CSV
- JSON
- Parquet
- TXT

Answer Area

Copy behavior:

- Flatten hierarchy
- Merge files
- Preserve hierarchy

Correct Answer:

Sink file type:

- CSV
- JSON
- Parquet
- TXT

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 2 years, 7 months ago

1. Merge Files

2. Parquet

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-performance-tuning-guidance>

upvoted 150 times

- **Ram0202** 7 months, 3 weeks ago
Copy behaviour types explained
<https://youtu.be/SUX1NiFSPkM>
upvoted 4 times
- **Ameenymous** 2 years, 7 months ago
The smaller the files, the negative the performance so Merge and Parquet seems to be the right answer.
upvoted 23 times
- **edba** 2 years 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 10 times
- **kilowd** 2 years, 2 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 8 times
- **captainbee** Highly Voted 2 years, 6 months ago
It's frustrating just how many questions ExamTopics get wrong. Can't be helpful
upvoted 56 times
- **gssd4scoder** 2 years, 2 months ago
Trying to understand if an answer is correct will help learn more
upvoted 11 times
- **RyuHayabusa** 2 years, 5 months ago
At least it helps in learning, as you have to research and think for yourself. Another big topic is having this questions in the first place is immensely helpful
upvoted 47 times
- **flaviодиасps** 1 year, 7 months ago
it is misleading, they should not give any answer at all
upvoted 5 times
- **SebK** 1 year, 9 months ago
Agree.
upvoted 2 times
- **lisa710** Most Recent 2 weeks, 6 days 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 2 times
- **phydev** 2 months, 1 week ago
Was on my exam today (31.10.2023).
upvoted 1 times
- **survivingtech** 1 month, 3 weeks ago
please how helpful was exam topics for your exam? any other resources that helped you?
upvoted 1 times
- **moumen** 2 months, 1 week ago
Hello, How many questions do you have in exam? Most of the questions are in exam topics?
upvoted 1 times
- **Chemmangat** 3 months, 3 weeks 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** 4 months ago
MERGE FILES since you need to make transformation and data have the same attributes
PARQUET because is the most efficient file format
upvoted 1 times
- **kkk5566** 4 months, 1 week ago
- Merge - Parquet
upvoted 2 times
- **ladistar** 4 months, 4 weeks ago

ChatGPT confirms it's 1. Merge, 2. Parquet
upvoted 3 times

✉️ **rocky48** 8 months, 2 weeks ago

1. Merge Files
2. Parquet
<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-performance-tuning-guidance>
upvoted 5 times

✉️ **Honour** 9 months, 2 weeks ago

Hey, the answer here should be "Merge Files" and "Parquet".

The question said nothing about hierarchies.

upvoted 2 times

✉️ **bubby248** 11 months ago

Merge small files will be best for fast retrieving. Parquet for sink file type
upvoted 1 times

✉️ **INDEAVR** 11 months ago

Either preserving or flattening hierarchy has little to no performance overhead, whereas merging files causes additional performance overhead. It's perverse
upvoted 1 times

✉️ **vigilante89** 1 year ago

Copy Behaviour: MERGE FILES
Because the small files already have same data attributes i.e. same schema. So merging all the data into one single file and converting the file to parquet makes more sense to make the query time, space and cost efficient.

Sink/Destination File Type: PARQUET

This is a no-brainer because parquet is the most efficient file format in this case in terms of time, space and cost efficiency.
upvoted 4 times

✉️ **Selma97** 1 year, 2 months ago

I think "Automatically infer the schema from the underlying files" means we should keep the same hierarchy and not merge all the data into a single file. So I would say that the first one is Preserve Hierarchy.
upvoted 1 times

✉️ **AhmedDaffaie** 11 months, 2 weeks ago

If you preserve the hierarchy, you will keep them in small files. This will affect the performance negatively. That is why Merge is better
upvoted 1 times

✉️ **OldSchool** 1 year, 1 month ago

Q:"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."
Since all have same data and attributes we can Merge them in one file and automatically infer the schema from the underlying 10 small files.
- Merge
- Parquet
upvoted 1 times

✉️ **Azure_Kraken** 1 year, 4 months ago

So many answers are misleading, totally fed-up by checking the comment section for each and every question
upvoted 4 times

✉️ **tomras** 1 year, 2 months ago

This site isn't a brain dump, it's a study tool and it very often has some insights that are actually useful in my job beyond the tests. If all you want is answers I'm sure that you can find something else that fits the bill.
upvoted 4 times

✉️ **tomras** 1 year, 2 months ago

It's more along the lines of being a Stackoverflow for the Microsoft exams.
upvoted 1 times

✉️ **Deeksha1234** 1 year, 4 months ago

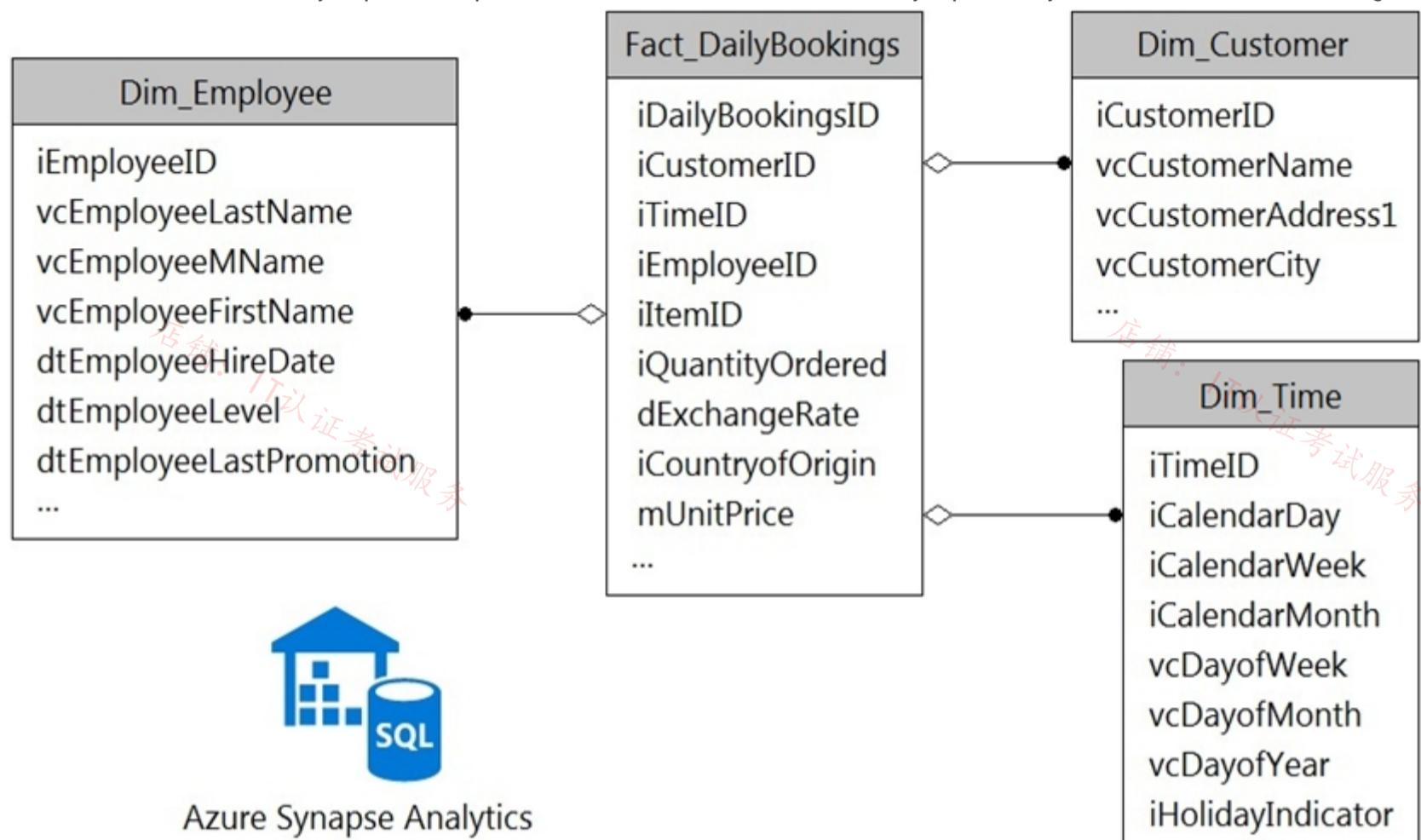
1. Merge Files
2. Parquet
upvoted 3 times

✉️ **Fiddi** 1 year, 5 months ago

I would say Merge might not be possible, because we do not know if the Json contains enough information about the subsidiary, which we would need if we merge.
upvoted 1 times

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

Answer Area

Dim_Customer:

Hash distributed
Round-robin
Replicated

Dim_Employee:

Hash distributed
Round-robin
Replicated

Correct Answer:

店铺：IT认证考试服务

Dim_Time:

Hash distributed
Round-robin
Replicated

Fact_DailyBookings:

Hash distributed
Round-robin
Replicated

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 2 years, 3 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 227 times

Tara123 3 weeks, 6 days 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

lisa710 2 weeks, 6 days ago

exceeding 10 gigabytes (GB) are often considered large
upvoted 3 times

✉ **virendrapsingh** 1 year, 7 months ago

This is a wonderful explanation. Worth giving a like.
upvoted 7 times

✉ **GameLift** 2 years, 3 months ago

Thanks, but where in the question does it indicate about Fact table has clustered columnstore index.?
upvoted 3 times

✉ **berserksap** 2 years, 2 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 2 times

✉ **berserksap** 2 years, 2 months ago

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-overview>
upvoted 1 times

✉ **ohana** Highly Voted 2 years, 2 months ago

Took the exam today, this question came out.
Ans: All the Dim tables --> Replicated
Fact Tables --> Hash Distributed
upvoted 46 times

✉ **gidemay237** Most Recent 3 weeks, 1 day ago

itexamslab.com

Given answer is correct
upvoted 1 times

✉ **xigaf50758** 3 weeks, 1 day ago

itexamslab.com

Given answer is correct
upvoted 1 times

✉ **dakku987** 3 weeks, 2 days ago

REplicated, REplicated, REplicated and hash
upvoted 1 times

✉ **hassexat** 4 months ago

Replicated / Replicated / Replicated / Hash
upvoted 1 times

✉ **kkk5566** 4 months, 1 week ago

Ans: All the Dim tables --> Replicated Fact Tables --> Hash Distributed
is correct
upvoted 1 times

✉ **DataEngDP** 6 months ago

<https://learn.microsoft.com/en-us/training/modules/design-multidimensional-schema-to-optimize-analytical-workloads/3-create-tables>
upvoted 1 times

✉ **DataEngDP** 6 months 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** 1 year 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** 1 year, 4 months ago

correct

upvoted 1 times

 **objecto** 1 year, 7 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** 1 year, 8 months ago

The answer is correct.

upvoted 1 times

 **PallaviPatel** 1 year, 11 months ago

correct answer

upvoted 2 times

 **Pritam85** 1 year, 11 months ago

Got this question on 2312/2021...answer is correct

upvoted 2 times

 **Mahesh_mm** 2 years ago

Ans is correct

upvoted 2 times

 **alfonsodisalvo** 2 years, 2 months ago

Dimension are Replicated :

"Since the table has multiple copies, replicated tables work best when the table size is less than 2 GB compressed."

"Replicated tables may not yield the best query performance when:

The table has frequent insert, update, and delete operations"

" We recommend using replicated tables instead of round-robin tables in most cases"

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-guidance-for-replicated-tables>

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.

Answer Area

Five-year-old data:

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

Correct Answer:

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.

	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

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

⊕  **yobllip** Highly Voted 2 years, 7 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 87 times

⊕  **r00s** 1 year, 7 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 3 times

⊕  **Justbu** Highly Voted 2 years, 3 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 12 times

⊕  **joshfry** 1 year, 5 months ago

"Costs must be minimized while maintaining the required availability." So cold and archive are the correct answers.
 upvoted 4 times

⊕  **lisa710** Most Recent 2 weeks, 6 days 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 1 times

⊕  **phydev** 2 months, 1 week ago

Was on my exam today (31.10.2023).

upvoted 5 times

⊕  **hassexat** 4 months ago

1. Cool --> You will access infrequently but data must be available in few time if you want to access them

2. Archive --> You will never access 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** 4 months, 1 week ago

1.cool and 2.archive
upvoted 1 times

 **akhil5432** 5 months, 1 week ago

cool and archive
upvoted 1 times

 **Andrew_Chen** 5 months, 2 weeks ago

Question: Why can't we just delete the blob to save more cost?
upvoted 2 times

 **janaki** 8 months ago

Answer: Cool & Archive
upvoted 2 times

 **aachao** 10 months, 3 weeks ago

answer is correct!
upvoted 2 times

 **vigilante89** 1 year 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** 1 year, 5 months ago

1.Cool, 2.Archive
upvoted 1 times

 **examtopicscap** 1 year, 5 months ago

1 - Hot
2 - Cold

There is a question with 'data over 5/7 years old' so it's very tricky

upvoted 2 times

 **Dothy** 1 year, 8 months ago

ans is correct
upvoted 1 times

 **PallaviPatel** 1 year, 11 months ago

ans is correct
upvoted 1 times

 **ANath** 2 years ago

1. Cool Storage
2. Archive Storage
upvoted 1 times

 **Mahesh_mm** 2 years ago

Answer is correct
upvoted 1 times

👤 **vigilante89** 1 year 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:

```
<table_option> ::=  
{  
    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** 1 year, 4 months ago

correct

upvoted 2 times

👤 **gursimran_s** 1 year, 7 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 2 times

👤 **topggggggg** 4 months, 2 weeks ago

SAVAGE XD

upvoted 1 times

👤 **Dothy** 1 year, 8 months ago

Answer is correct

upvoted 1 times

👤 **Egocentric** 1 year, 8 months ago

provided answer is correct

upvoted 1 times

👤 **vineet1234** 1 year, 9 months ago

D comes before P as in DP-203

upvoted 8 times

👤 **PallaviPatel** 1 year, 11 months ago

correct

upvoted 1 times

👤 **Jaws1990** 2 years ago

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

upvoted 1 times

👤 **ploer** 1 year, 11 months ago

Having three boundaries will lead to four partitions:

1. Partition for values < 1
2. Partition for values from 1 to 999999
3. Partition for values from 1000000 to 1999999
4. Partition for values >= 2000000

upvoted 3 times

👤 **nastyaaa** 1 year, 10 months ago

but only <= and >. it is range left for values, right

upvoted 2 times

👤 **Mahesh_mm** 2 years ago

Answer is correct

upvoted 1 times

👤 **hugoborda** 2 years, 3 months ago

Answer is correct

upvoted 1 times

👤 **hsetin** 2 years, 4 months ago

Indeed! Answer is correct

upvoted 1 times

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

Correct 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**  2 years, 1 month ago

Selected Answer: D

D is correct (voting comment that people dont have to open discussion always, please upvote to help others)
upvoted 151 times

✉  **bc5468521**  2 years, 7 months ago

Answer D; Temporal table is better than SCD2, but it is not supported in Synapse yet
upvoted 76 times

✉  **SolutionA** 5 months, 1 week 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

✉  **GodfreyMbizo** 11 months, 2 weeks ago

Temporarytables is completely offtopic ,the catch here is ...at a point in time hence SCD is the way to go
upvoted 2 times

✉  **Preben** 2 years, 7 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

✉  **mbravo** 2 years, 7 months ago

Temporal tables and Temporary tables are two very distinct concepts. Your link has absolutely nothing to do with this question.
upvoted 17 times

✉  **Vaishnav** 2 years, 6 months ago

<https://docs.microsoft.com/en-us/azure/azure-sql/temporal-tables>
Answer : A Temporal Tables
upvoted 1 times

✉  **Vaishnav** 2 years, 6 months 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** 1 year, 10 months ago

though this not something relative to this question. temporal tables looks alike to delta table.
upvoted 1 times

 **joxis4786** Most Recent ⓘ 3 weeks, 1 day ago
itexamslab.com

Given answer is correct
upvoted 1 times

 **dakku987** 3 weeks, 1 day ago

Selected Answer: D
scd 2 maintain the inactive current time end time like this in table
upvoted 1 times

 **Bail** 1 month, 2 weeks ago

Selected Answer: D
It's D
upvoted 1 times

 **hassexat** 4 months ago

D

Because is the way as you can get historical data for the employee
upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: D
correct
upvoted 1 times

 **Deeksha1234** 7 months ago

D is correct
upvoted 1 times

 **steveo123** 8 months, 2 weeks ago

Selected Answer: D
D is correct.
upvoted 2 times

 **haidebelognime** 11 months ago

Selected Answer: D
it is D 100%, stop messing around with answers like "temporal table"
upvoted 4 times

 **bubby248** 11 months ago

SCD type 2 which has latest boolean flag to retrieve
upvoted 1 times

 **vigilante89** 1 year 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 4 times

 **Fernando_Caemerer** 1 year, 1 month ago

Selected Answer: D
Answer D
upvoted 1 times

 **OldSchool** 1 year, 1 month ago

Selected Answer: D
SCD 2 - D
upvoted 1 times

 **Deeksha1234** 1 year, 4 months ago

correct D
upvoted 2 times

 **PugazhManohar** 1 year, 5 months ago

Maintains the latest employee information - SCD-Type2 (Ans-D)

upvoted 1 times

 **Dothy** 1 year, 8 months ago

Answer is correct

upvoted 2 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

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.

Correct Answer: ABF

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%)

✉️  **Diane** Highly Voted  2 years, 8 months ago

correct answer is ABF <https://www.examtopics.com/discussions/microsoft/view/41207-exam-dp-200-topic-1-question-56-discussion/>
upvoted 83 times

✉️  **AvithK** 2 years, 5 months ago

yes but the order is different it is FAB
upvoted 45 times

✉️  **KingIlo** 2 years, 4 months ago

The question didn't specify order or sequence
upvoted 12 times

✉️  **gssd4scoder** 2 years, 2 months ago

Exactly, agree with you
upvoted 2 times

✉️  **IDKoi** Highly Voted  2 years, 5 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 44 times

✉️  **kkk5566** Most Recent  4 months, 1 week ago

Selected Answer: ABF

FAB is correct
upvoted 1 times

✉️  **vigilante89** 1 year ago

Selected Answer: ABF

The sequence is: FAB
create managed identity --> add managed identity to the sales group --> use managed identity as credentials for data load process.
upvoted 6 times

⊕ **XiltroX** 1 year, 1 month ago

Selected Answer: ABF

First create a managed ID, Add the managed ID, use the managed ID
upvoted 1 times

⊕ **nahantai** 1 year, 3 months ago

Can anyone explain why we cannot useshared access signature in this case?
upvoted 1 times

⊕ **dmitriypo** 1 year, 2 months ago

Managed Identity authentication is required when your storage account is attached to a VNet.
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-examples>
upvoted 5 times

⊕ **coolin** 1 year, 4 months ago

ABF is the correct ans
upvoted 1 times

⊕ **Deeksha1234** 1 year, 4 months ago

Selected Answer: ABF

ABF is correct
upvoted 3 times

⊕ **Dothy** 1 year, 8 months ago

correct answer is ABF
upvoted 1 times

⊕ **Egocentric** 1 year, 8 months ago

ABF is correct
upvoted 1 times

⊕ **practicewizards** 1 year, 9 months ago

Selected Answer: ABF

FAB - create, add to group, use to load data
upvoted 2 times

⊕ **Backy** 2 years ago

Is answer A properly worded?
"Add the managed identity to the Sales group" should be "Add the Sales group to managed identity"
upvoted 6 times

⊕ **lukeonline** 2 years ago

Selected Answer: ABF

FAB should be correct
upvoted 5 times

⊕ **VeroDon** 2 years ago

Selected Answer: ABF

FAB is correct sequence
upvoted 2 times

⊕ **SabaJamal2010AtGmail** 2 years ago

1. Create a managed identity.
2. Add the managed identity to the Sales group.
3. Use the managed identity as the credentials for the data load process.
upvoted 2 times

⊕ **Mahesh_mm** 2 years ago

FAB is correct sequence
upvoted 1 times

⊕ **Lewistrick** 2 years ago

Would it even be a good idea to have the data load process be part of the Sales team? They have separate responsibilities, so should be part of another group. I know that's not possible in the answer list, but I'm trying to think best practices here.

upvoted 5 times

HOTSPOT -

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

Name	Role
User1	Server admin
User2	db_dreader

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

```

1  SELECT c.name,
2      tbl.name AS table_name,
3      typ.name AS datatype,
4      c.is_masked,
5      c.masking_function
6  FROM sys.masked_columns AS c
7  INNER JOIN sys.tables AS tbl ON c.[object_id] = tbl.[object_id]
8  INNER JOIN sys.types typ ON c.user_type_id = typ.user_type_id
9  WHERE is_masked = 1; 收录
10

```

店铺：IT认证考试服务

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()

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:

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

店铺：IT认证考试服务

Correct 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

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 2 years, 4 months ago

user 1 is admin, so he will see the value stored in dbms.

1. 0
 2. Value in database
- upvoted 98 times

✉  **examtopicsofyannick** 5 months, 1 week ago

Confirmed. Everything explained here <https://learn.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview?view=azuresql-mi>
upvoted 1 times

✉  **azurearmy** 2 years, 2 months ago

2 is wrong
upvoted 2 times

✉  **Aditya0891** 1 year, 6 months ago

azurearmy read the question properly and then answer, 2nd is queried by user1, masking doesn't apply to user1
upvoted 18 times

✉  **rjile** Highly Voted 2 years, 4 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 28 times

✉  **berserksap** 2 years, 2 months ago

The second question is queried by User 1 who is the admin
upvoted 21 times

✉  **Joanna0** Most Recent 1 day, 13 hours ago

* Use a zero value for numeric data types (bigint, bit, decimal, int, money, numeric, smallint, smallmoney, tinyint, float, real).
upvoted 1 times

✉  **phydev** 2 months, 1 week ago

Was on my exam today (31.10.2023).
upvoted 2 times

✉  **hassexat** 4 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** 4 months, 1 week ago

1. 0
 2. Value stored in database
- upvoted 2 times

□ **nicky87654** 12 months 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

□ **fuel** 11 months, 4 weeks 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** 1 year 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 birthdate column (is_masked=1) in the database 'sys'.

upvoted 4 times

□ **OldSchool** 1 year, 1 month 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** 1 year, 4 months ago

correct

upvoted 5 times

□ **azure900test** 1 year, 6 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** 1 year, 7 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 default to a date datetime smalldate should be 1900-01-01. Strangely there is no such option. Any ideas anyone?
upvoted 2 times

□ **objecto** 1 year, 7 months ago

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

upvoted 2 times

□ **Dothy** 1 year, 8 months ago

1. 0
 2. Value in database
- upvoted 1 times

□ **Egocentric** 1 year, 8 months ago

on this question its just about paying attention to detail

upvoted 5 times

□ **manan16** 1 year, 9 months ago

How user2 can access data as it is masked?

upvoted 1 times

 **manan16** 1 year, 9 months ago

Can Someone explain first option as in doc it says 0

upvoted 1 times

 **Mahesh_mm** 2 years 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

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.

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

B.

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

C.

```
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
);
```

D.

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

Correct Answer: C

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 2 years, 4 months ago

C is correct

<https://www.examtopics.com/discussions/microsoft/view/19469-exam-dp-200-topic-1-question-27-discussion/>

upvoted 52 times

□  **baiy** Highly Voted 10 months, 3 weeks 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 13 times

□  **phydev** Most Recent 2 months, 1 week ago

Was on my exam today (31.10.2023).

upvoted 5 times

□  **hassexat** 4 months ago

C is correct since you need to drop first and create again the external table

upvoted 2 times

□  **kkk5566** 4 months, 1 week ago

drop and create
upvoted 1 times

✉ **vigilante89** 1 year 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 4 times

✉ **OldSchool** 1 year, 1 month ago

C is Correct

upvoted 1 times

✉ **Selma97** 1 year, 2 months ago

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

upvoted 1 times

✉ **anto69** 1 year, 1 month ago

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

upvoted 6 times

✉ **Deeksha1234** 1 year, 4 months ago

correct

upvoted 2 times

✉ **dsp17** 1 year, 6 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** 1 year, 9 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** 1 year, 1 month 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** 1 year, 11 months ago

c is correct.

upvoted 1 times

✉ **hugoborda** 2 years, 3 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.

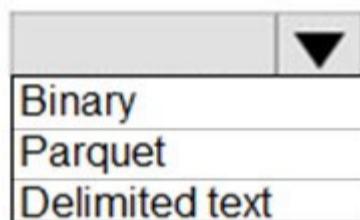
How should you configure the copy activity? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

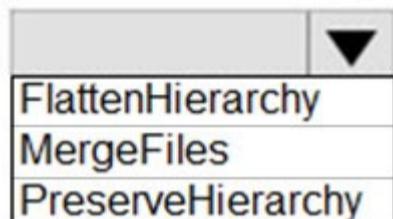
Hot Area:

Answer Area

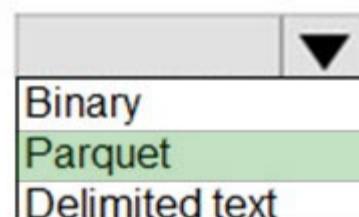
Source dataset type:



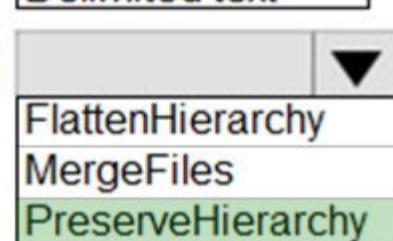
Copy activity copy behavior:

**Answer Area**

Source dataset type:



Correct 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 2 years, 4 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 answer.

upvoted 74 times

□ **GameLift** 2 years, 2 months ago

But the doc says "When using Binary dataset in copy activity, you can only copy from Binary dataset to Binary dataset." So I guess it's parquet then?

upvoted 10 times

□ **conscience** 2 months, 1 week ago

I have used Binary to copy entire folders with its subfolder and files which were csv & parquet both. So, IMO binary would be correct answer.

upvoted 1 times

□ **captainpike** 2 years, 2 months ago

This note is referring to the fact that, in the template, you have to specify "BinarySink" as the type for the target Sink; and that exactly what the Copy data tool does. (you can check this by editing the created copy pipeline and see the code). Choosing Binary and PreserveHierarchy copy all file as they are perfectly.

upvoted 4 times

□ **iooj** 1 year, 10 months ago

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

upvoted 4 times

□ **jed_elhak** 2 years, 3 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** 2 years, 4 months ago

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

upvoted 6 times

□ **AbhiGola** Highly Voted 2 years, 4 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** 1 year, 7 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

□ **anto69** 1 year 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

□ **mhi** 8 months, 1 week ago

when doing a binary copy, you're not doing any transformation!

upvoted 4 times

□ **phydev** Most Recent 2 months, 1 week ago

Was on my exam today (31.10.2023).

upvoted 4 times

□ **SenMia** 3 weeks, 2 days ago

mind helping with the right option? binary or parquet for the first box? thanks!! :)

upvoted 1 times

□ **kkk5566** 4 months, 1 week ago

Binary & PerserveHierarchy

upvoted 3 times

□ **tonyfig** 4 months, 3 weeks 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 3 times

□ **rocky48** 6 months 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

□ **klayytech** 6 months, 1 week 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** 6 months, 3 weeks ago

Massimo Manganiello <massimo.manganiello@gmail.com>

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

✉ **trantrongw** 9 months, 2 weeks ago

Agree. I've checked it.

upvoted 1 times

✉ **Lestrang** 11 months, 4 weeks 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 11 times

✉ **mtc9** 7 months ago

ChatGPT is plainly wrong, binary type retains the original parquet format, because 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

✉ **JustAnotherDBA** 1 year 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 10 times

✉ **mtc9** 7 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 1 times

✉ **JustAnotherDBA** 1 year ago

<https://learn.microsoft.com/en-us/azure/data-factory/format-binary>

upvoted 3 times

✉ **Rrk07** 1 year, 1 month ago

Answer is correct .

upvoted 1 times

✉ **temacc** 1 year, 1 month ago

Binary - copy files as is in fastest way.
PreserveHierarchy - for saving folder structure.
upvoted 2 times

□ **OldSchool** 1 year, 1 month ago

Answer is correct. No transformation and preserve hierarchy
upvoted 1 times

□ **RBKasemodel** 1 year, 1 month 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 find any information saying that parquet won't be parsed if the source and sink are parquet files. So I -think- it will parse, and we can understand that it is a transformation.

upvoted 2 times

□ **allagowf** 1 year, 3 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 6 times

□ **Deeksha1234** 1 year, 4 months ago

given ans is correct
upvoted 2 times

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)

Correct 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 (72%) B (28%)

  **meetj** Highly Voted 2 years, 4 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 104 times

  **dev2dev** 2 years ago

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

upvoted 31 times

  **BK10** 1 year, 11 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 29 times

  **AnonymousJhb** 8 months, 1 week 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 3 times

  **semauni** 5 months, 2 weeks 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 6 times

  **Billybob0604** 1 year, 1 month 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 6 times

  **kenmexam** 1 year, 2 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 10 times

 **dylan_t** 7 months, 2 weeks 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 2 times

 **Sasha_in_San_Francisco** Highly Voted  2 years, 2 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 70 times

 **SabaJamal2010AtGmail** 2 years 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

 **Joanna0** Most Recent  1 day, 12 hours ago

Selected Answer: B

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

When configured to use globally redundant storage (GRS, GZRS, and RA-GZRS), Azure copies your data asynchronously to a secondary geographic region located hundreds of miles away. This level of redundancy allows you to recover your data if there's an outage throughout the entire primary region.

Read-access geo-redundant storage (RA-GRS) and read-access geo-zone-redundant storage (RA-GZRS) also provide geo-redundant storage, but offer the added benefit of read access to the secondary endpoint. These options are ideal for applications designed for high availability business-critical applications. If the primary endpoint experiences an outage, applications configured for read access to the secondary region can continue to operate. Microsoft recommends RA-GZRS for maximum availability and durability of your storage accounts.

<https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance>

upvoted 1 times

 **lucassn_** 1 month, 3 weeks ago

Selected Answer: A

A is cheaper than B.

<https://learn.microsoft.com/en-us/azure/storage/common/storage-redundancy>

upvoted 1 times

 **conscience** 2 months, 1 week ago

Selected Answer: A

2 points need to note:

-> secondary regions need IF AN OUTAGE OCCURES in the primary region

-> Cost minimizes

RA-GRS cannot meet the second point hence correct answer is A

upvoted 2 times

 **jiriz** 3 months ago

Selected Answer: B

RA-GRS - It's clearly said there <https://learn.microsoft.com/cs-cz/azure/storage/common/storage-redundancy>

upvoted 1 times

 **mav2000** 3 months, 3 weeks ago

A) is correct because your priority is to minimize costs not to have instant read access to the second region, therefore GRS is the answer because if region1 fails, Microsoft will start a failover process to set the second region as the primary and you will get your data much cheaper (but slower) than RA-GRS

upvoted 1 times

 **AvSUN** 4 months ago

Since you need to minimize cost, there is not need to get confused.

GRS is cheaper than RA-GRS.

You can check the pricing here - <https://azure.microsoft.com/en-in/pricing/details/storage/blobs/>
upvoted 1 times

□ **KKK5566** 4 months, 1 week ago

Selected Answer: A

is correct

upvoted 1 times

□ **Amitj2625** 5 months, 1 week ago

The correct answer is:

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

With RA-GRS, your data is not only replicated to a secondary region (geo-redundant storage, GRS) but also allows read access to the data in the secondary region. This means that if there is an outage in the primary region, you can access and read the data from the secondary region, providing business continuity and reducing downtime.

While geo-redundant storage (GRS) on its own provides data redundancy across regions, it only allows read and write access in the primary region. To meet the requirement of having read access in the secondary region during an outage, RA-GRS is the appropriate option.

Zone-redundant storage (ZRS) and locally-redundant storage (LRS) do not provide the capability of data redundancy across regions, so they are not suitable for ensuring read access in a secondary region during a primary region outage.

upvoted 1 times

□ **TechieBloke** 5 months, 2 weeks ago

I would think it's A

The difference between GRS and RA GRS is fairly simple, GRS only allows to be read in the secondary zone in the even of a failover from the primary to secondary while RA GRS allows the option to read in the secondary whenever.

upvoted 2 times

□ **semauni** 5 months, 2 weeks ago

Selected Answer: A

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

□ **matiandal** 6 months, 1 week ago

vote , for sure , A as the correct answer.

notes:

--> Azure Storage offers two options for copying your data to a secondary region:

GRS | GZRS

--> With GRS or GZRS, the data in the secondary region isn't available for read or write access unless there's a failover to the primary region.

>> Q >>"ensure that the data in the container is available for read workloads in a secondary region IF an OUTAGE OCCURS in the primary region."
-- aka --> the Q does not imply something about read available without a system failure
-- so --> no need of a RA-GRS

MS DOC:

For read access to the secondary region, configure your storage account to use read-access geo-redundant storage (RA-GRS) or read-access geo-zone-redundant storage (RA-GZRS).

R: <https://learn.microsoft.com/en-us/azure/storage/common/storage-redundancy>

upvoted 1 times

□ **klayytech** 6 months, 1 week ago

The answer is B. RA-GRS.

The question explicitly specifies that you need to ensure that the data in the container is available for read workloads in a secondary region. This means that you need to be able to read data from the secondary region even if the primary region is unavailable.

RA-GRS (Read-Access Geo-Redundant Storage) is the only option that allows you to read data from the secondary region. With RA-GRS, your data is replicated to a secondary region in a different geographic location, and you can also read data from the secondary region.

upvoted 1 times

□ **klayytech** 6 months, 1 week ago

Selected Answer: B

However, the question specifically states that you need to ensure that the data in the container is available for read workloads in a secondary region. This means that you need to be able to read data from the secondary region even if the primary region is unavailable.

upvoted 1 times

□ **klayytech** 6 months, 1 week ago

In this case, the need for read access to the secondary region outweighs the cost of RA-GRS. This is because the application needs to be able to continue reading data even if the primary region is unavailable. RA-GRS provides this functionality, while GRS does not.

Therefore, the best option for this scenario is RA-GRS. This will ensure that the data is always available, while minimizing costs.

upvoted 1 times

 Ayman79 6 months, 2 weeks ago

Selected Answer: A

Read-access geo-redundant storage (RA-GRS) is a more expensive option that allows you to read data from the secondary region without having to initiate a failover. However, RA-GRS does not provide any additional protection against regional outages.

upvoted 2 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

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)

Correct 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

D (99%)

✉️  **MadEgg** Highly Voted 2 years 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 81 times

✉️  **DrTaz** 2 years ago

I agree.

Please give this comment a medal (or a cookie).

upvoted 6 times

✉️  **Ozren** 1 year, 9 months ago

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

upvoted 2 times

✉️  **Narasimhap** 1 year, 11 months ago

Well explained!

upvoted 2 times

✉️  **JohnMasipa** Highly Voted 2 years, 4 months ago

This can't be correct. Should be D.

upvoted 78 times

✉️  **JayBird** 2 years, 4 months ago

Why, LRS is cheaper?

upvoted 1 times

✉️  **Vitality** 2 years, 4 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** 2 years, 2 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

 **lisa710** Most Recent 2 weeks, 5 days ago

zone-redundant storage (ZRS)

upvoted 1 times

 **SillyChili** 1 month, 3 weeks 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 1 times

 **74gjd_37** 4 months ago

Selected Answer: D

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. Thereofe, option C is incorrect.

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: D

is correct

upvoted 1 times

 **Mani_V** 6 months, 2 weeks ago

LRS/ZRS doesn't come to picture if anything needs to available in other regions. so GRS is rite one.

upvoted 2 times

 **deutscher** 9 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** 11 months, 2 weeks 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** 11 months, 1 week ago

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

upvoted 1 times

 **nicky87654** 12 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** 1 year, 3 months ago

Selected Answer: D

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

upvoted 1 times

 **Deeksha1234** 1 year, 4 months ago

Selected Answer: D

D is correct

upvoted 1 times

 **Rrk07** 1 year, 7 months ago

D is correct as it talks about "a data center" means we can not use the LRS (LOCAL)

upvoted 1 times

 **olavrab8** 1 year, 8 months ago

Selected Answer: D

D -> Data is replicated synchronously

upvoted 1 times

 **Egocentric** 1 year, 8 months ago

D is correct

upvoted 2 times

 **ravi2931** 1 year, 9 months ago

it should be D

upvoted 1 times

 **ravi2931** 1 year, 9 months ago

see this explained clearly -

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)

upvoted 1 times

 **ASG1205** 1 year, 9 months ago

Selected Answer: D

Answer should be D, as LRS won't be helpfull in case of whole datacenter failure.

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:

Answer Area

Distribution:

Hash
Replicated
Round-robin

Indexing:

Clustered
Clustered columnstore
Heap

Partitioning:

Date
None

Answer Area

Distribution:

Hash
Replicated
Round-robin

Correct Answer:

Indexing:

Clustered
Clustered columnstore
Heap

Partitioning:

Date
None

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 2 years, 4 months ago

Round-Robin

Heap

None

upvoted 364 times

□  **Deepshikha1228** 1 year, 5 months ago

I agree, Round Robin, Heap and None is the correct option

upvoted 7 times

□  **anto69** 2 years ago

I agree too

upvoted 6 times

□  **Narasimhap** 1 year, 11 months ago

Round- Robin

Heap

None.

No brainer for this question.

upvoted 15 times

□  **gssd4scoder** 2 years, 2 months ago

Agree 100%.

All in paragraphs under this: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-overview>.

upvoted 9 times

□  **DrTaz** 2 years ago

Also agree 100%

upvoted 4 times

□  **laszek** Highly Voted 2 years, 4 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 60 times

□  **berserksap** 2 years, 2 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 1 times

□  **andy_g** 1 year, 11 months ago

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

upvoted 2 times

□  **Vardhan_Brahmanapally** 2 years, 2 months ago

Can you explain me why should we use heap?

upvoted 1 times

□  **DrTaz** 2 years 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 11 times

□  **SQLDev0000** 1 year, 10 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

□  **_Tom** Most Recent 1 month, 3 weeks ago

Round-Robin

Heap

None

upvoted 2 times

□  **74gjd_37** 4 months ago

Round-Robin, 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 4 times

□ **kkk5566** 4 months, 1 week ago

Round-Robin Heap None

upvoted 1 times

□ **kkk5566** 4 months, 2 weeks ago

Round-Robin

Heap

None

upvoted 1 times

□ **janaki** 7 months, 1 week ago

Never ever use date partitioning with hash distribution.

The correct answer is: Round robin, Heap and None

upvoted 2 times

□ **rocky48** 8 months, 2 weeks 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

□ **SHENO000** 11 months, 1 week ago

Round-Robin

Heap

None

upvoted 3 times

□ **akk_1289** 11 months, 2 weeks 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

□ **jhargett1** 11 months, 2 weeks 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

□ **DindaS** 11 months, 3 weeks ago

Round-robin - the default once. As for any staging table the Round Robin should be selected. if you don't select during table design this will be considered by default.

Heap - this is a staging table

No partitions - this is a staging table

upvoted 2 times

□ **JosephVishal** 11 months, 3 weeks ago

Answer is

- 1.) Round-Robin
- 2.) Heap
- 3.) None

upvoted 4 times

 **Dusica** 12 months ago

requirement is to optimize LOAD not QUERY performance

Round-Robin;Heap;None

upvoted 1 times

 **rj02** 1 year, 1 month ago

Round robin and heap sounds good but why not date partition as question states daily load to staging

upvoted 2 times

 **ToddW** 1 year ago

The 1 million row will be loaded daily, but nothing is said that these records are for one day. In addition, the table is truncated before each load so it is pointless.

upvoted 1 times

 **Rrk07** 1 year, 1 month ago

As it only ~~about~~ stage table so it should be

Round-Robin

Heap

None

upvoted 2 times

 **rohitbinnani** 1 year, 1 month ago

#1 Round-Robin -

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>.

#2 Heap - A heap table can be especially useful for loading transient data, such as a staging table, which is transformed into a final table.

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

#3 None - When you do full truncate and load daily, there is no point in partitioning.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition>

upvoted 4 times

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.

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

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?

- A. replicated
- B. hash-distributed on PurchaseKey
- C. round-robin
- D. hash-distributed on IsOrderFinalized

Correct 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

B (88%)

10%

✉️ FredNo Highly Voted 2 years, 1 month ago

Selected Answer: B

Correct

upvoted 43 times

✉️ Deepshikha1228 1 year, 5 months ago

B is correct

upvoted 2 times

□ **GameLift** Highly Voted 2 years, 4 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 25 times

□ **Podavenna** 2 years, 3 months ago

Yes, your logic is correct!

upvoted 8 times

□ **sdg2844** Most Recent 1 week ago

Selected Answer: B

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

□ **phydev** 2 months, 1 week ago

Selected Answer: B

Was on my exam today (31.10.2023).

upvoted 5 times

□ **pperf** 3 months ago

Selected Answer: B

<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** 3 months, 3 weeks ago

why the answer says it cannot be a date column?

upvoted 1 times

□ **74gjd_37** 4 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** 3 months, 3 weeks ago

why we cannot use data column?

upvoted 1 times

□ **kkk5566** 4 months, 1 week ago

Selected Answer: B

B is correct

upvoted 1 times

□ **SolutionA** 5 months, 1 week 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** 5 months, 1 week 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 1 times

□ **mamahani** 8 months ago

Selected Answer: B

B is correct

upvoted 1 times

□ **henryphchan** 8 months ago

Selected Answer: B

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

upvoted 1 times

□ **SHENO000** 11 months, 1 week ago

Selected Answer: B

B is the Correct Answer

upvoted 3 times

□ **astone42** 11 months, 2 weeks ago

Selected Answer: B

B is correct.

upvoted 2 times

 **DindaS** 11 months, 3 weeks ago

Ideally there should be an option to create partition DateKey. When we use the partition key column in the where condition , the unwanted partition's data will be eliminated automatically. that's the beauty of the partition and how it works in conjunction with the query. However, would like to know from the experts in the forum.

upvoted 5 times

 **Dusica** 12 months ago

Selected Answer: B

what about B plus (imaginary) partitioning on date ? Or is error in question because Purchase Key by itself would not be very helpful

upvoted 1 times

 **vigilante89** 1 year, 1 month ago

B is correct!!!

Because "hash-distributed on IsOrderFinalized" as a distribution column would only use 2 out of 60 distributions (for Yes, No) which is a waste of compute and time resources.

So "hash-distributed on PurchaseKey" with multiple unique values will utilize all 60 distributions and make the query process much faster and utilize all the compute efficiently.

upvoted 3 times

 **Rrk07** 1 year, 1 month ago

Correct answer

upvoted 1 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.

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

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:

Answer Area

EventCategory:

DimChannel
DimDate
DimEvent
FactEvents

ChannelGrouping:

DimChannel
DimDate
DimEvent
FactEvents

TotalEvents:

DimChannel
DimDate
DimEvent
FactEvents

Answer Area

EventCategory:

DimChannel
DimDate
DimEvent
FactEvents

Correct Answer: ChannelGrouping:

DimChannel
DimDate
DimEvent
FactEvents

TotalEvents:

DimChannel
DimDate
DimEvent
FactEvents

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 2 years, 2 months ago

It seems to be correct
upvoted 69 times

✉  **DingDongSingSong** Highly Voted 1 year, 9 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 22 times

✉  **sdegcp** 1 year, 6 months ago

Question says, including a dimTime table so we need to design only 3 tables.
upvoted 4 times

✉  **allagowf** 1 year, 3 months ago

the question is clear try to read it again, there is no need to design DIM_DATE
upvoted 1 times

✉  **hassexat** Most Recent 4 months ago

Correct provided answer
upvoted 1 times

✉  **kkk5566** 4 months, 1 week ago

EventCategory -> dimEvent
channelGrouping -> dimChannel
TotalEvents -> factEven
upvoted 1 times

✉  **mamahani** 8 months ago

DimEvent / DimChannel / FactEvents
upvoted 1 times

✉  **SHENO000** 11 months, 1 week ago

Category or Group will go with the table DIM, Total Events will go with the Fact
upvoted 2 times

✉  **Deeksha1234** 1 year, 4 months ago

correct
upvoted 1 times

□ **Rrk07** 1 year, 7 months ago

EventCategory -> dimEvent
channelGrouping -> dimChannel
TotalEvents -> factEven
upvoted 4 times

□ **Dothy** 1 year, 8 months ago

EventCategory -> dimEvent
channelGrouping -> dimChannel
TotalEvents -> factEven
upvoted 1 times

□ **JJdeWit** 1 year, 8 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 3 times

□ **Mahesh_mm** 2 years ago

Answer is correct
upvoted 4 times

□ **Iaszek** 2 years, 4 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** 1 year, 10 months ago

I mean if you think about it, ChannelName (facebook,google,youtube), ChannelType (paid media, free posts, ads), ChanneDelivery (chrome, etc etc). Just thinking out loud
upvoted 1 times

□ **manquak** 2 years, 4 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

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

Correct 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 (86%)

14%

✉  **Fahd92** Highly Voted 2 years, 3 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 16 times

✉  **ANath** 1 year, 12 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 4 times

✉  **Thij** 2 years, 3 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** 3 weeks ago

i think when compression is in question we should go for parquet/avro bcz only they give compression
upvoted 1 times

✉  **moneytime** Most Recent 2 months, 2 weeks 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 1 times

✉  **kkk5566** 4 months, 1 week ago

Selected Answer: A

Answer is yes
upvoted 1 times

✉  **akhil5432** 5 months, 1 week ago

Selected Answer: A

"a" is correct option
upvoted 1 times

✉  **SHENO000** 11 months, 1 week ago

Selected Answer: A

A will do the job
upvoted 3 times

✉  **Rrk07** 1 year, 1 month ago

Delimited text file is true,

upvoted 1 times

✉ **nahantai** 1 year, 3 months ago

how do you know this question is about PolyBase?

upvoted 3 times

✉ **greenlever** 1 year, 3 months ago

Selected Answer: A

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** 1 year, 4 months ago

Selected Answer: A

yes, answer is A

upvoted 1 times

✉ **Janisys** 1 year, 5 months 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

✉ **Deepshikha1228** 1 year, 5 months 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** 1 year, 7 months ago

Selected Answer: A

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

upvoted 3 times

✉ **Rrk07** 1 year, 7 months ago

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

upvoted 1 times

✉ **Muishkin** 1 year, 8 months ago

A text file seems to be too simple an answer however true as per the microsoft link.I was thinking of parquet/avro files

upvoted 1 times

✉ **Massy** 1 year, 10 months ago

Selected Answer: B

From the question: "75% of the rows contain description data that has an average length of 1.1 MB". You can't

From the documentation: "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."

So 75% of rows aren't good for a delimited text files... why you said answer is yes?

upvoted 3 times

✉ **kamil_k** 1 year, 10 months ago

I initially thought so too, however isn't this limit only relevant to PolyBase copy? It is not mentioned which method is used to transfer the data so you could fit more than 1mb into a column in the table if you want to, you just have to use something else e.g. COPY command.

upvoted 3 times

✉ **PallaviPatel** 1 year, 11 months ago

Selected Answer: A

correct answer.

upvoted 2 times

✉ **Mahesh_mm** 2 years ago

A is correct

upvoted 1 times

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

Correct 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%)

✉ **bhanuprasad9331** Highly Voted 1 year, 11 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** Highly Voted 2 years, 2 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 7 times

✉ **tbhttp** 1 year, 3 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 4 times

✉ **Julius7000** 2 years, 2 months ago

Can You explain this in more details?

upvoted 11 times

✉ **helly13** 2 years, 1 month ago

I really didn't understand this , can you explain?

upvoted 7 times

✉ **aurorafang** 1 year, 5 months ago

it's a virtualized chart that the guy want to simplified the question set.

upvoted 2 times

 **ML_Novice** 1 year, 4 months ago
but it becomes more complicated with this chart haha
upvoted 5 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: B
Answer is no ,u use HEAP idx
upvoted 1 times

 **youngbug** 11 months, 3 weeks ago
For fast loading to a table, using a staging table which is a heap table.
upvoted 1 times

 **DindaS** 11 months, 3 weeks 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** 1 year, 4 months ago
B is right
upvoted 1 times

 **Janisys** 1 year, 5 months 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 3 times

 **Deepshikha1228** 1 year, 5 months ago
B is right
upvoted 1 times

 **Amsterliese** 1 year, 9 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** 1 year, 8 months ago
Yes load to a table without indexes for faster load right?
upvoted 1 times

 **lionurag** 1 year, 10 months ago
Selected Answer: B
B is correct
upvoted 3 times

 **PallaviPatel** 1 year, 11 months ago
Selected Answer: B
B is correct.
upvoted 1 times

 **DE_Sanjay** 1 year, 12 months ago
NO is the answer.
upvoted 1 times

 **Mahesh_mm** 2 years ago
B is correct
upvoted 1 times

 **rashjan** 2 years, 1 month ago
Selected Answer: B
Correct Answer: No.
upvoted 2 times

 **sachabess79** 2 years, 3 months ago
No, The index will expand the time of insertion
upvoted 3 times

 **michals** 2 years, 4 months ago
<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/guidance-for-loading-data>. "For the fastest load, use compressed delimited text files."
upvoted 1 times

 **umeshkd05** 2 years, 4 months ago

But the row size also need to be < 1 MB
So, files need to be modified to make all rows < 1 MB
Answer: NO
upvoted 4 times

✉ **Julius7000** 2 years, 2 months ago

In other words, i think that 100GB is much to much for the columnstore index memorywise. The documentation is unclear with the context of this particular question, but i think the answer is NO, as the given answer is the wrong idea anyways.
upvoted 1 times

✉ **gk765** 2 years, 3 months ago

Correct answer should be NO
upvoted 2 times

✉ **Julius7000** 2 years, 2 months ago

Not Row size, row NUMBER have to be at maximum of 1,048,576 rows.
"When there is memory pressure, the columnstore index might not be able to achieve maximum compression rates. This effects query performance."
upvoted 1 times

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 ~~should~~ modify the files to ensure that each row is more than 1 MB.

Does this meet the goal?

A. Yes

B. No

Correct 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 2 years, 3 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 12 times

✉  **amarG1996** Highly Voted 2 years 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 5 times

✉  **amarG1996** 2 years ago

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/data-loading-best-practices#prepare-data-in-azure-storage>

upvoted 2 times

✉  **kamil_k** 1 year, 10 months ago

is it stated anywhere that we have to use PolyBase? What about COPY command?

upvoted 2 times

✉  **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: B

Answer is no

upvoted 1 times

✉  **vigilante89** 1 year, 1 month ago

Selected Answer: B

B is correct!!!

upvoted 2 times

✉  **Deeksha1234** 1 year, 4 months ago

B is correct, agree with explanation by Amar

upvoted 1 times

✉  **PallaviPatel** 1 year, 11 months ago

Selected Answer: B

B is correct.

upvoted 4 times

✉  **Mahesh_mm** 2 years ago

Answer is No

upvoted 1 times

✉  **rashjan** 2 years, 1 month ago

Selected Answer: B

Correct Answer: No.

upvoted 2 times

✉  **Odoxtoom** 2 years, 2 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

✉  **Aslam208** 2 years, 1 month ago

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

upvoted 5 times

✉  **Bishtu** 2 years ago

Yes
No
No
upvoted 2 times

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

Correct 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 (100%)

 **ANath** Highly Voted 1 year, 11 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 21 times

 **Canary_2021** Highly Voted 2 years ago

Selected Answer: B

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 10 times

 **Joanna0** Most Recent 11 hours, 36 minutes ago

Selected Answer: B

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 1 times

 **ll94** 6 days, 22 hours 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 months, 1 week ago

Selected Answer: B

Was on my exam today (31.10.2023).
upvoted 4 times

□ **kkk5566** 4 months, 1 week ago

Selected Answer: B

Materialized view
upvoted 1 times

□ **norbitek** 1 year 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

□ **Deepshikha1228** 1 year, 5 months ago

B is correct
upvoted 2 times

□ **SKN0865** 1 year, 6 months 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** 1 year, 8 months ago

B materialized view
upvoted 2 times

□ **Egocentric** 1 year, 8 months ago

B is correct without a doubt
upvoted 2 times

□ **DingDongSingSong** 1 year, 9 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 better query performance in case of a complex query where query isn't static.

upvoted 1 times

□ **uzairahm** 1 year, 6 months ago

I was thinking on the same level initially but there are multiple tables involved 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

□ **PallaviPatel** 1 year, 11 months ago

Selected Answer: B

B correct.
upvoted 1 times

□ **VeroDon** 2 years ago

Selected Answer: B

Correct
upvoted 3 times

□ **Mahesh_mm** 2 years ago

B is correct
upvoted 1 times

□ **bad_atitude** 2 years ago

B materialized view
upvoted 2 times

□ **alexleonvalencia** 2 years, 1 month ago

Respuesta Correcta B, Una vista materializada.
upvoted 5 times

店铺: IT认证考试服务

店铺: IT认证考试服务

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

Correct 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  2 years 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 20 times

 **ZIMARAKI** Highly Voted  1 year, 2 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** 1 year, 1 month ago

But they never asked about faster, so why it cant be A

upvoted 3 times

 **shakes103** 11 months, 3 weeks ago

In this business, time is money.

upvoted 17 times

 **phydev** Most Recent  2 months, 1 week ago

Selected Answer: D

Was on my exam today (31.10.2023).

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: D

Prefer to D

upvoted 2 times

 **ExamWiner** 6 months, 4 weeks ago

Selected Answer: D

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** 1 year, 4 months ago

Selected Answer: D

both A and D; D - Parquet is faster
upvoted 2 times

 **jainparag1** 1 year, 6 months ago

Option D as the explanation suggests. Parquet is always faster than CSV being columnar data store.
upvoted 3 times

 **Dicer** 1 year, 6 months ago

both A and D are okay, but Parquet is faster than CSV, so answer is D.
upvoted 2 times

 **Rrk07** 1 year, 7 months ago

Both A & D are correct , as the explanation also suggest same.
upvoted 2 times

 **RehanRajput** 1 year, 7 months ago

Both A and D
upvoted 2 times

 **RehanRajput** 1 year, 7 months ago

<https://docs.microsoft.com/en-us/azure/synapse-analytics/metadata/database>
upvoted 1 times

 **MatiCiri** 1 year, 8 months ago

Selected Answer: D
Looks correct to me
upvoted 2 times

 **AhmedDaffaie** 1 year, 10 months ago

JSON is also supported by Serverless SQL Pool but it is kinda complicated. Why is it not selected?
upvoted 2 times

 **Ajitzk27** 1 year, 10 months ago

Selected Answer: D
Looks correct to me
upvoted 2 times

 **VijayMore** 1 year, 10 months ago

Selected Answer: D
Correct
upvoted 1 times

 **PallaviPatel** 1 year, 11 months ago

Selected Answer: D
Both A and D are correct. as CSV and Parquet are correct answers.
upvoted 1 times

 **Mahesh_mm** 2 years ago

Parquet and CSV are correct
upvoted 4 times

 **Nifl91** 2 years, 1 month ago

I think A and D are both correct answers.
upvoted 3 times

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

Correct Answer: D

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

General capabilities -

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

Integration capabilities -

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

Reference:

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

Community vote distribution

□ **Nifl91** Highly Voted 2 years, 1 month ago

Correct!

upvoted 23 times

□ **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: D

DataBricks with Java lang

upvoted 1 times

□ **auwia** 6 months, 3 weeks ago

Selected Answer: D

D is correct.

upvoted 1 times

□ **Mohitsain** 10 months ago

D is the correct one for sure.

upvoted 2 times

□ **Fernando_Caemerer** 1 year, 1 month ago

Selected Answer: D

D is correct

upvoted 2 times

□ **Aaashu** 1 year, 3 months ago

correct

upvoted 2 times

□ **Deepshikha1228** 1 year, 5 months ago

D is correct

upvoted 3 times

□ **jainparag1** 1 year, 6 months ago

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

upvoted 3 times

□ **NewTuanAnh** 1 year, 9 months ago

why not C: Azure Stream Analytics?

upvoted 1 times

□ **NewTuanAnh** 1 year, 9 months ago

I see, Azure Stream Analytics does not associate with Java

upvoted 2 times

□ **sdokmak** 1 year, 7 months ago

or databricks

upvoted 1 times

□ **sdokmak** 1 year, 7 months ago

kafka*

upvoted 2 times

□ **Muishkin** 1 year, 8 months ago

Yes Azure stream Analytics for streaming data?

upvoted 1 times

□ **PallaviPatel** 1 year, 11 months ago

Selected Answer: D

correct.

upvoted 2 times

□ **Mahesh_mm** 2 years ago

Answer is correct

upvoted 3 times

□ **alexleonvalencia** 2 years, 1 month ago

Respuesta correcta Azure DataBricks.

upvoted 4 times

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

Correct 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

D (81%)	Other
---------	-------

👤 **VeroDon** Highly Voted 2 years 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 50 times

👤 **SenMia** 3 weeks, 2 days ago

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

upvoted 1 times

👤 **auwia** 6 months, 3 weeks 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 5 times

👤 **bhrz** 1 year, 3 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

👤 **Massy** 1 year, 8 months ago

I can understand why you say not merge, but why avro?

upvoted 4 times

👤 **anks84** 1 year, 4 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

👤 **Canary_2021** Highly Voted 2 years ago

Selected Answer: D

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 21 times

 **Joanna0** Most Recent 11 hours, 17 minutes ago

Selected Answer: B

Binary Serialization:

Avro uses a compact binary format, making it more efficient in terms of storage and transmission compared to plain text formats like CSV. This can be advantageous for batch processing scenarios, especially when dealing with large volumes of data.

Schema Evolution:

Avro supports schema evolution, allowing you to change the schema of your data without requiring modifications to the entire dataset or affecting backward compatibility. This flexibility is beneficial in scenarios where your data schema may evolve over time.

Compression:

While Avro itself is a binary format that provides some level of compression, you can further enhance compression by applying additional compression algorithms. This is particularly useful when dealing with large files, and it helps to reduce storage costs and improve data transfer efficiency.

upvoted 1 times

 **ll94** 6 days, 22 hours ago

Selected Answer: C

- A. Convert the files to JSON => no sense
- B. Convert the files to Avro => my understanding is that the format of the file csv is given, so no
- C. Compress the files => for batch processing it's a win and it's the only option that you can assume true given the available information
- D. Merge the files => this can be true but not knowing how many files there is big issue

upvoted 1 times

 **jongert** 2 weeks, 4 days ago

Selected Answer: B

AVRO is binary format, so it will be optimized for batch processing.

Problem with merging files is that it is still CSV, String typed which has to be parsed when processing later. Therefore, it would not qualify as being optimized for batch processing.

upvoted 1 times

 **lisa710** 2 weeks, 4 days ago

compress the files

upvoted 1 times

 **SenMia** 3 weeks, 2 days ago

the confusing point in this question is that we will not know how many files are expected in an hour, if that's the case. will merging files really be helpful?

upvoted 1 times

 **Moo925** 3 months, 3 weeks ago

Selected Answer: C

compressing the CSV files is the most practical and efficient way to optimize them for batch processing, especially when dealing with varying file sizes

upvoted 1 times

 **Ranjan6214** 4 months, 1 week ago

Selected Answer: D

Sometimes, data pipelines have limited control over the raw data, which has lots of small files. In general, we recommend that your system have some sort of process to aggregate small files into larger ones for use by downstream applications. If you're processing data in real time, you can use a real time streaming engine (such as Azure Stream Analytics or Spark Streaming) together with a message broker (such as Event Hubs or Apache Kafka) to store your data as larger files. As you aggregate small files into larger ones, consider saving them in a read-optimized format such as Apache Parquet for downstream processing.

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices#optimize-for-data-ingest>

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: D

Should be D

upvoted 1 times

 **sidh_r** 4 months, 3 weeks ago

Selected Answer: B

Merging might not be the best one since the size of files are varying, we wouldn't be able to come with a simple approach to merge. If we have to merge then again we would cause some overhead, hence converting to avro seems to be a best choice

upvoted 1 times

 **vijay007123** 4 months, 3 weeks ago

avro is correct ans. Merging might help when many small sized files are there however, it'll introduce one more challenge to handle schema variations. If too many large sized files need to merged that'll be difficult.

upvoted 1 times

 **akhil5432** 5 months, 1 week ago

Selected Answer: C

option c

upvoted 1 times

 **auwia** 6 months, 3 weeks ago

Selected Answer: D

D. Merge the files

To optimize the files stored in the Azure Data Lake Storage Gen2 container for batch processing, you should merge the files. Merging smaller files into larger files is a common optimization technique in data processing scenarios.

Having a large number of small files can introduce overhead in terms of file management, metadata processing, and data scanning. By merging the smaller files into larger files, you can reduce this overhead and improve the efficiency of batch processing operations.

Merging the files is especially beneficial when dealing with varying file sizes, as it helps to create a more balanced distribution of data across the files and reduces the impact of small files on processing performance.

Therefore, in this scenario, merging the files would be the recommended approach to optimize the files for batch processing.

upvoted 3 times

 **mamahani** 8 months ago

Selected Answer: D

D: merge the files

upvoted 1 times

 **janaki** 8 months ago

The best option is to compress the files. Azure Data Lake Storage Gen2 supports a variety of compression codecs, including GZIP, DEFLATE, and Snappy, which can help to reduce file sizes and improve batch processing times.

upvoted 1 times

 **steveo123** 8 months, 2 weeks ago

Selected Answer: D

"larger files lead to better performance"

upvoted 1 times

HOTSPOT -

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

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

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:

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

Correct Answer:

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	店铺: IT认证考试服务
container1/docs/contoso.json	店铺: IT认证考试服务
container1/mycontoso/contoso.csv	店铺: IT认证考试服务

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-sensitive 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 2 years ago

correct

upvoted 38 times

✉ **adfgasd** 2 years ago

why the .csv?

upvoted 3 times

✉ **Lewistrick** 2 years ago

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

upvoted 17 times

✉ **alexeonvalencia** Highly Voted 2 years, 1 month ago

Respuesta Cool Tier & Container1/contoso.csv

upvoted 7 times

✉ **kwokeric97** Most Recent 1 month, 2 weeks 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": {}
"daysAfterModificationGreaterThanOrEqual":

<https://stackoverflow.com/questions/62368999/skip-rule-creation-if-parameter-is-empty-in-azure-storage-life-cycle-management>
upvoted 1 times

✉ **data_guy** 1 month, 1 week ago

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

upvoted 1 times

✉ **kkk5566** 4 months, 1 week ago

Move to Cool Tier & Container1/contoso.csv

upvoted 2 times

✉ **kkk5566** 4 months, 2 weeks ago

correct

upvoted 1 times

✉ **darth_vader_007** 8 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** 1 year ago

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

upvoted 1 times

✉ **Deeksha1234** 1 year, 4 months ago

Cool Tier & Container1/contoso.csv
upvoted 2 times

✉ **Deepshikha1228** 1 year, 5 months ago

given answer is correct
upvoted 2 times

✉ **azure900test** 1 year, 6 months 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** 1 year, 6 months ago

sorry overlooked the cold tier policy, please ignore
upvoted 3 times

✉ **AJ01** 2 years ago

shouldn't the question be greater than 60 days?

upvoted 2 times

✉ **stunner85_** 1 year, 11 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** 2 years ago

correct

upvoted 3 times

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

Correct 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

D (96%)	4%
---------	----

✉ **Lewistrick** Highly Voted 2 years ago

Anyone else thinks this is a very badly explained situation?

upvoted 50 times

✉ **SillyChili** 3 weeks, 6 days ago

agree that it is a very badly explained situation.

i took Microsoft practice assessment, it has similar question, and the answer is not on date column because "Using date and partitioning by month, all sales for a month will be in the same partition, not providing parallelism."

upvoted 1 times

✉ **Tobestosis** 2 weeks, 5 days ago

I think in the practice assessment it's streaming data where you want to partition equally but I agree it didn't specify and caused lots of confusion

upvoted 1 times

✉ **Aditya0891** 1 year, 7 months ago

yes i read 5 times still couldn't figure out what's being explained

upvoted 8 times

✉ **Sophie_** 4 months, 3 weeks ago

Explanation seems self-confused between Distribution specific rules and partitioning key considerations...

upvoted 1 times

✉ **Canary_2021** Highly Voted 2 years ago

Selected Answer: D

Select D because analysts will most commonly analyze transactions for a given month,

upvoted 24 times

✉ **DooperMan** Most Recent 2 months ago

Selected Answer: D

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

✉ **phydev** 2 months, 1 week ago

Selected Answer: D

Was on my exam today (31.10.2023).

upvoted 2 times

✉ **arihant_jain** 1 month, 2 weeks ago

What did you answer? Was your answer right?

upvoted 1 times

✉ **hydmt07** 1 month ago

It's a bot posting the same comment on every answer.

upvoted 2 times

✉ **ellala** 3 months ago

Selected Answer: B

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 1 times

✉ **mav2000** 3 months, 3 weeks 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 1 times

✉ **hassexat** 4 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 1 times

✉ **kkk5566** 4 months, 1 week ago

Selected Answer: D

Partition by month is a good idea

upvoted 1 times

✉ **akhil5432** 5 months, 1 week ago

Selected Answer: D

transaction month

upvoted 1 times

✉ **Shanks111** 7 months, 1 week 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

✉ **henryphchan** 8 months ago

Selected Answer: D

D is correct. In most of the cases, partition the data by date.

upvoted 3 times

✉ **SinSS** 8 months, 2 weeks ago

Selected Answer: B

Account Type

upvoted 1 times

✉ **SinSS** 8 months, 2 weeks 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 6 times

✉ **Rob77** 7 months, 3 weeks ago

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

upvoted 2 times

✉ **semauni** 5 months, 2 weeks ago

The Microsoft question is also about partitioning.

upvoted 1 times

□ **shleemcgee** 10 months, 1 week ago

I think A.

If you partition over month, then when the reports are run for a given month all the data will just be contained in a single partition? For that type of query, you want the data split over the partitions. TransactionType is the most common report slicer with the largest number of rows per TransactionType (40M, compared to 4M for CustomerSegment). AccountType has 500M rows per value, but this is not used in all the queries (it says and/or in the question).

upvoted 3 times

□ **OfficeSaracus** 8 months, 1 week ago

I think you are confusing partition with distribution. Each partition will still be distributed over 60 distributions. Its just to have "smaller" tables separated (partitioned) by month. So instead of looking through a huge table your querys only browse one or two smaller ones, hence way faster queue times

upvoted 1 times

□ **semauni** 5 months, 2 weeks ago

I agree that that sounds logical, but there is a Microsoft practice exam question that looks like this one, also about partitioning, which also mentions parallel processing.

upvoted 1 times

□ **akshaynag95** 11 months, 1 week ago

Selected Answer: D

D is the right answer

upvoted 1 times

□ **DindaS** 11 months, 3 weeks ago

D should be correct answer. As analysts will be filtering based on the month hence the query will be executed on the relevant partitions :)

upvoted 3 times

□ **kim32** 12 months ago

CustomerSegment: Because the aim of a Partition column is to split the data into the smallest portions possible to improve query performance

upvoted 2 times

□ **OdogwuSaina** 11 months, 3 weeks ago

And the question says "Analysts will most commonly analyze transactions for a given MONTH" hence TransactionMonth is the right answer.

upvoted 1 times

HOTSPOT -

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

Type	Designated retention period
Application	360 days
Infrastructure	60 days

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:

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	▼

Correct 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	▼

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 2 years, 1 month 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 81 times

✉  **RoyP654** 6 months, 4 weeks 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

✉  **Anshuman_B** 11 months, 1 week ago

Thanks for sharing this info.

upvoted 1 times

✉  **dsp17** 1 year, 6 months ago

Thanks a ton for explaining.

upvoted 2 times

✉  **ANath**  2 years 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 19 times

✉  **kkk5566**  4 months, 1 week ago

correct

upvoted 1 times

✉  **Rashmi007** 7 months ago

Given answer is correct.

upvoted 3 times

✉  **maochi** 7 months ago

The answers are correct

upvoted 1 times

✉  **mamahani** 8 months ago

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

upvoted 1 times

✉  **mamahani** 9 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 1 times

✉  **NeerajGarg** 10 months, 2 weeks 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

✉  **SHENO000** 11 months, 1 week ago

Given Answers are correct

upvoted 2 times

✉  **zilvakas** 1 year 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 retrieval are removed) Cool storage for 1 month for 1TB costs 11.10 Eur for 1 month (that is $11.10 \times 6 = 66.6$ for 180 days) and archive storage costs 3,41 Eur per month (that is $3,41 \times 6 = 20,46$ Eur for 6 months). Therefore Archive tier is cheaper 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

✉  **patvn** 11 months, 1 week 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

✉  **zilvakas** 1 year ago

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

upvoted 2 times

□ **Deeksha1234** 1 year, 5 months ago

given answer is correct

upvoted 2 times

□ **georgiakon** 1 year, 6 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** 1 year, 8 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

□ **sdokmak** 1 year, 7 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** 1 year, 8 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

□ **JNunn** 1 year ago

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

upvoted 1 times

□ **Mahesh_mm** 2 years ago

Answers are correct

upvoted 2 times

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

Correct 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 1 year, 7 months ago

Selected Answer: B

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 28 times

✉  **hrastogi7** Highly Voted 2 years ago

Parquet can be quickly retrieved and maintain metadata in itself. Hence Parquet is correct answer.

upvoted 22 times

✉  **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: B

should be correct

upvoted 1 times

✉  **akhil5432** 5 months, 1 week ago

Selected Answer: B

Parquet

upvoted 1 times

✉  **Deeksha1234** 1 year, 5 months ago

Parquet is correct

upvoted 3 times

✉  **Rrk07** 1 year, 7 months ago

Parquet is correct

upvoted 2 times

✉  **Muishkin** 1 year, 8 months ago

Isnt JSON good for batch processing/streaming?

upvoted 1 times

✉  **RehanRajput** 1 year, 7 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** 1 year, 9 months ago

I am confused!

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

upvoted 5 times

✉  **Boompiee** 1 year, 8 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 months ago

"Polybase currently supports only delimited text, rcfie, orc and parquet formats."

R: <https://msdn.microsoft.com/en-us/library/dn935025.aspx>

upvoted 1 times

 **PallaviPatel** 1 year, 11 months ago

Selected Answer: B

correct.

upvoted 1 times

 **EmmettBrown** 1 year, 11 months ago

Selected Answer: B

Parquet is the correct answer

upvoted 1 times

 **alexleonvalencia** 2 years, 1 month ago

Respuesta correcta PARQUET

upvoted 1 times

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.

Correct 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

C (93%)	7%
---------	----

✉ **Aslam208** Highly Voted 2 years, 1 month ago

Selected Answer: C

The correct answer is C

upvoted 66 times

✉ **Nifl91** Highly Voted 2 years, 1 month ago

this must be C. since the need is to overwrite dbo.Sales with the content of stg.Sales.

SWITCH source TO target

upvoted 34 times

✉ **lisa710** Most Recent 2 weeks, 4 days ago

The most efficient approach to overwrite the content of the first partition in dbo.Sales with the content of the same partition in stg.Sales while minimizing load times is to switch the first partition from stg.Sales to dbo.Sales. This is option C.

upvoted 1 times

✉ **d046bc0** 1 month ago

ALTER TABLE stg.Sales SWITCH PARTITION 1 TO dbo.Sales PARTITION 1 WITH (TRUNCATE_TARGET = ON);

upvoted 1 times

✉ **ellala** 3 months ago

This is quite a weird situation because according to Microsoft documentation:

"When reassigning a table's data as a partition to an already-existing partitioned table, or switching a partition from one partitioned table to another, the target partition must exist and it MUST BE EMPTY." (https://learn.microsoft.com/en-us/sql/t-sql/statements/alter-table-transact-sql?view=azure-sqldw-latest&preserve-view=true#switch--partition-source_partition_number_expression--to--schema_name--target_table--partition-target_partition_number_expression-)

Therefore none of the options would be possible if considering that both tables are not empty on that partition. Then I have no idea what would be the correct answer, although I answered C.

upvoted 2 times

✉ **gggqqqqq** 3 months, 2 weeks ago

When reassigning a table's data as a partition to an already-existing partitioned table, or switching a partition from one partitioned table to another, the target partition must exist and it must be empty. Thus, B is correct because it is the first step to remove the data from dbo.Sales. Then we need to another step which is C: to move the partition with data from stg.Sales to dbo.Sales.

upvoted 1 times

□ **kkk5566** 4 months, 1 week ago

Selected Answer: C
ALTER TABLE stg.Sales
SWITCH PARTITION 1
TO dbo.Sales
PARTITION 1;
upvoted 3 times

□ **lfss** 4 months, 3 weeks ago

The correct answer is C
upvoted 1 times

□ **Pfffff** 6 months, 3 weeks ago

Selected Answer: B
from dbo to stg
upvoted 2 times

□ **auwia** 6 months, 3 weeks ago

Selected Answer: B
ALTER TABLE dbo.Sales SWITCH PARTITION 1 TO stg.Sales PARTITION 1;
upvoted 1 times

□ **auwia** 6 months, 3 weeks ago

I'm sorry I was wrong, the answer is C and this is the sql:

ALTER TABLE stg.Sales
SWITCH PARTITION 1
TO dbo.Sales
PARTITION 1;
upvoted 5 times

□ **varunlpu** 7 months ago

C is right
upvoted 1 times

□ **ajhak** 7 months, 2 weeks ago

This wording is horrific. The fact is the first partition in dbo.sales needs to be replaced with the stg.sales. The way it's worded makes both answers seem like it's being replaced depending how you read it.
upvoted 4 times

□ **mgastalho** 8 months, 2 weeks ago

Selected Answer: C
ALTER TABLE dbo.Sales SWITCH PARTITION 1 TO stg.Sales PARTITION 1;
This statement switches partition 1 of the dbo.Sales table with partition 1 of the stg.Sales table, effectively overwriting the contents of the partition in dbo.Sales with the contents of the partition in stg.Sales.
upvoted 2 times

□ **auwia** 6 months, 3 weeks ago

You wrote correctly the alter table, but you selected the wrong answer! C instead of B.
upvoted 1 times

□ **Maartjeee89** 8 months, 3 weeks ago

Selected Answer: C
<https://medium.com/@cocci.g/switch-partitions-in-azure-synapse-sql-dw-1e0e32309872>
upvoted 1 times

□ **mamahani** 9 months ago

the correct answer is C: from stg to dbo;
as per microsoft doc: "In dedicated SQL pool, the TRUNCATE_TARGET option is supported in the ALTER TABLE command. With TRUNCATE_TARGET the ALTER TABLE command overwrites existing data in the partition with new data." "<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition?context=%2Fazure%2Fsynapse-analytics%2Fcontext%2Fcontext#load-new-data-into-partitions-that-contain-data-in-one-step>
upvoted 2 times

□ **shubz2020** 9 months, 1 week ago

Selected Answer: C
From Source to target
upvoted 1 times

□ **esaade** 10 months ago

Selected Answer: C
The best option for overwriting the content of the first partition in dbo.Sales with the content of the same partition in stg.Sales and minimizing load times would be to switch the first partition from stg.Sales to dbo.Sales, option C.

Switching partitions is a common approach to efficiently manage large tables in SQL Server. By using the ALTER TABLE SWITCH statement, it is possible to quickly move data between tables with minimal overhead. In this scenario, switching the first partition from stg.Sales to dbo.Sales will replace the data in the first partition of dbo.Sales with the data from the corresponding partition in stg.Sales.

upvoted 3 times

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

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.

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-text description of the supplier company
SupplierCategory	Category of goods provided by the supplier company

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.

- A. surrogate primary key
- B. effective start date
- C. business key
- D. last modified date
- E. effective end date
- F. foreign key

Correct Answer: BCE

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 (88%)

12%

 **ItHYMeRish** Highly Voted 2 years, 1 month 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 109 times

 **Achu24** 1 year ago

Correct

upvoted 3 times

 **CHOPIN** Highly Voted 2 years 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 19 times

 **bhrz** 1 year, 3 months ago

@CHOPIN the correct answer is ABE. Read the reference here <https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types>

upvoted 2 times

 **bakstorage0001** 1 year, 3 months ago

@CHOPIN maybe, but the problem is that the question says "which three columns should you Add", and here the problem is Add. If the Business Key is already there should NOT be added, because is already there. You need a surrogate key.

upvoted 8 times

 **auwia** 6 months, 2 weeks ago

Look at the following Question #35 and you will understand quickly! ;-)

upvoted 1 times

 **jds0** 9 months, 2 weeks ago

"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 1 times

 **blazy001** Most Recent 1 month ago

I'm working now more than 13 years 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 2 times

 **hassexat** 4 months ago

Selected Answer: ABE

Surrogate Key

Effective Start Date

Effective End Date

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: BCE

BCE is answer

upvoted 1 times

 **subhraz** 4 months, 1 week ago

How come it can be anything other than ABE.

ABE is the correct answer.

upvoted 1 times

 **Amitj2625** 5 months, 1 week ago

To create a Type 2 Slowly Changing Dimension (SCD) in Azure Synapse Analytics dedicated SQL pool for supplier data, you would need to add the following three additional columns:

A. Surrogate Primary Key: This column is a unique identifier for each supplier record and is used as a primary key in the dimension table.

B. Effective Start Date: This column indicates the date when a particular version of the supplier data becomes effective or active.

E. Effective End Date: This column indicates the date when a particular version of the supplier data becomes obsolete or inactive. It is usually set to a specific value (e.g., '9999-12-31') to indicate the current active record.

With these three additional columns, you can effectively manage historical changes to the supplier data and track different versions of each supplier record over time.

upvoted 4 times

 **amirshaz** 6 months, 3 weeks ago

Selected Answer: ABE

The Supplier Key from the ERP system is the business key, hence we need a surrogate key

upvoted 1 times

SinSS 8 months, 2 weeks ago

Business key is already there, SupplierSystemId

upvoted 2 times

SinSS 8 months, 2 weeks ago

Selected Answer: ABE

ABE is correct

upvoted 1 times

mgastalho 8 months, 2 weeks ago

Selected Answer: ABE

To create a Type 2 SCD, you should add the following three additional columns to the data:

- A. surrogate primary key: This is a unique identifier for each record in the table.
- B. effective start date: This is the date and time when the record becomes effective.
- E. effective end date: This is the date and time when the record is no longer effective.

So the correct options are A, B, and E.

upvoted 2 times

divadbou 8 months, 3 weeks ago

To create a Type 2 slowly changing dimension (SCD) for supplier data in an Azure Synapse Analytics dedicated SQL pool, the following three additional columns should be added to the data:

- A. surrogate primary key: A surrogate key is used as a primary key to uniquely identify each record in the dimension table.
- B. effective start date: This column represents the date when the current version of the supplier data became effective.
- E. effective end date: This column represents the date when the current version of the supplier data will end, or become ineffective.

Therefore, the correct options are A, B, and E.

upvoted 1 times

esaade 10 months ago

Selected Answer: ABE

To create a Type 2 slowly changing dimension (SCD) for supplier data in an Azure Synapse Analytics dedicated SQL pool, in addition to the existing columns, the following three additional columns should be added:

- B. Effective start date - This column specifies the date and time when the supplier record becomes active or effective.
- E. Effective end date - This column specifies the date and time when the supplier record is no longer active or effective.

A. Surrogate primary key - This column is used as a unique identifier for each supplier record and can be used as a foreign key in other tables.

Option C (business key) is not mandatory for creating a Type 2 SCD, but it can be used to ensure the uniqueness of each supplier record.

upvoted 1 times

Anshuman_B 11 months, 1 week ago

Selected Answer: BCE

If option C - Business Key is a composite key that needs to be added then answer is BCE.

Otherwise ABE. In both cases, the concept is we need a column that uniquely identifies each row.

upvoted 1 times

vrodriguesp 11 months, 1 week ago

Selected Answer: ABE

because I think the SupplierSystemId it's business key.

upvoted 1 times

SHENO000 11 months, 1 week ago

ABE is the correct Answer

upvoted 1 times

JosephVishal 11 months, 3 weeks ago

Since, the SupplierSystemId is an entity from Source ERP, it is business key. Hence, answer is

- 1.) Surrogate Primary Key
- 2.) Effective Start Date
- 3.) Effective End Date

upvoted 4 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

店舗: IT認証試験サービス
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

Correct 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

Box 1: Denormalize to a second normal form

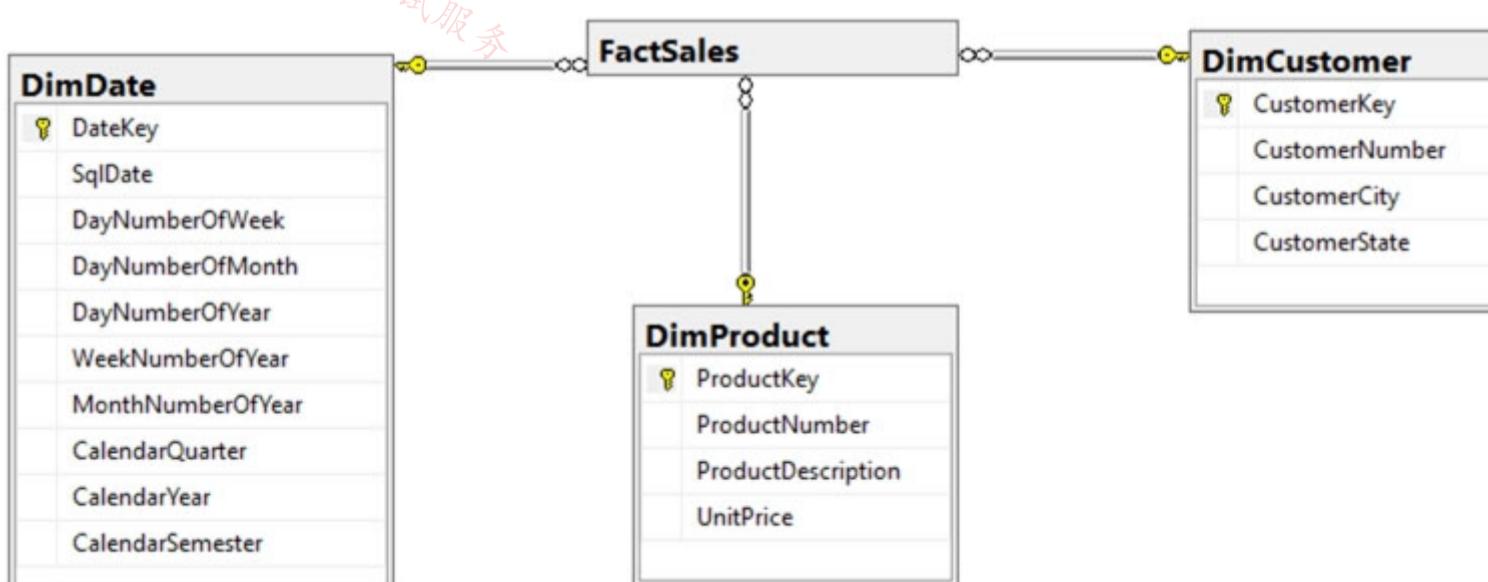
Denormalization is the process of transforming higher normal forms to lower normal forms via storing the join of higher normal form relations as a base relation.

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 1 year, 11 months ago

Answer correct.

upvoted 21 times

□ **ajhak** Highly Voted 7 months, 3 weeks ago

Always denormalize when moving from database to date warehouse.

upvoted 7 times

□ **Rob77** 7 months, 3 weeks ago

God no! Only dimensional models, read Inmon and Lindstedt approach.

upvoted 2 times

□ **kkk5566** Most Recent 4 months, 1 week ago

denormalizing and IDENTITY

upvoted 2 times

□ **SHENO000** 11 months, 1 week ago

Answer Is correct

upvoted 3 times

□ **Deepshikha1228** 1 year, 5 months ago

Answer is correct

upvoted 3 times

□ **JimZhang4123** 1 year, 7 months ago

'The solution must optimize read operations.' means denormalization

upvoted 5 times

□ **Mahesh_mm** 2 years ago

Answers are correct

upvoted 1 times

□ **PallaviPatel** 2 years ago

answer is correct

upvoted 4 times

□ **moreinva43** 2 years 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	▼
.range	▼
.sortBy	▼

("*")	▼
("StoreID", "Hour")	▼
("StoreID", "Year", "Month", "Day", "Hour")	▼

.mode("append")

.csv("/Purchases")	▼
.json("/Purchases")	▼
.parquet("/Purchases")	▼
.saveAsTable("/Purchases")	▼

Correct Answer:

Answer Area

df.write

.bucketBy	▼
.partitionBy	▼
.range	▼
.sortBy	▼

("*")	▼
("StoreID", "Hour")	▼
("StoreID", "Year", "Month", "Day", "Hour")	▼

.mode("append")

.csv("/Purchases")	▼
.json("/Purchases")	▼
.parquet("/Purchases")	▼
.saveAsTable("/Purchases")	▼

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** Highly Voted  2 years ago

Answers are correct

upvoted 22 times

□  **Aslam208** Highly Voted  2 years ago

correct

upvoted 8 times

□  **hodashiyam** Most Recent  3 months, 3 weeks ago

Answers are correct

upvoted 1 times

□  **kkk5566** 4 months, 1 week ago

correct

upvoted 1 times

□  **Rrk07** 1 year, 1 month ago

Why parquet option? Can anyone explain.

upvoted 3 times

□  **phydev** 2 months, 2 weeks ago

Because Parquet is always the answer.

upvoted 9 times

□  **DataSaM** 6 months, 1 week ago

I guess because of the requirement "reducing storage costs"

upvoted 3 times

□  **steveo123** 8 months, 1 week ago

The solution must minimize storage costs.

upvoted 3 times

□  **gabrielkuka** 1 year, 1 month ago

Can somebody explain why are we partitioning by StoreId, Year, Month, Day and Hour instead of just StoreID and Hour?

upvoted 5 times

□  **dduque10** 1 year, 1 month ago

if partitioned by storeid and hour only, the same hours from different days would go to the same partition, that would be inefficient

upvoted 28 times

□  **Keerthi24** 1 year, 4 months ago

Can someone explain why parquet and not saveAsTable option?

upvoted 3 times

□  **uira** 1 year, 1 month ago

Parquet is columnar, so faster to be read by Azure Synapse Analytics via CTEAs.

upvoted 6 times

□  **Deeksha1234** 1 year, 5 months ago

given answers are correct

upvoted 3 times

□  **hm358** 1 year, 7 months ago

Correct

upvoted 2 times

□  **sparkchu** 1 year, 9 months ago

ans should be saveAsTable. format is defined by format() method.

upvoted 4 times

□  **assU2** 1 year, 12 months ago

Can anyone explain why it's Partitioning and not Bucketing pls?

upvoted 6 times

□  **KashRaynardMorse** 1 year, 8 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/dataskipindex>

upvoted 6 times

✉  **bhanuprasad9331** 1 year, 11 months 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 9 times

✉  **assU2** 1 year, 12 months ago

Is it a question of correct syntax (numBuckets int the number of buckets to save) or is it smth else?

upvoted 2 times

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

Correct Answer: A

Each partition should have around 1 millions records. Dedicated SQL pools already have 60 partitions.

We have the formula: Records/(Partitions*60)= 1 million

$$\text{Partitions} = \text{Records}/(1 \text{ million} * 60)$$

$$\text{Partitions} = 2.4 \times 1,000,000,000 / (1,000,000 * 60) = 40$$

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

A (87%)	13%
---------	-----

曰 **Aslam208** Highly Voted 2 years ago

correct

upvoted 24 times

曰 **dom271219** Highly Voted 1 year, 4 months ago

Selected Answer: A

$$2,4\text{bn}/60=40\text{M}$$

upvoted 13 times

曰 **hassexat** Most Recent 4 months ago

Selected Answer: A

$$2,400,000,000 / 60,000,000 = 40$$

upvoted 1 times

曰 **kkk5566** 4 months, 1 week ago

Selected Answer: A

A is correct

upvoted 1 times

曰 **akhil5432** 5 months ago

Selected Answer: A

OPTION A

upvoted 1 times

曰 **AZLearn111** 11 months ago

No of automatic Distributions is 60. So each distribution will have $2.4 \text{ B} / 60 = 40\text{M}$. For a good performance each partition within a distribution (some time called buckets of data) should have 1M rows per bucket. So $40\text{M} / 1\text{M} = 40$ partitions.

upvoted 10 times

曰 **zekescokies** 8 months, 2 weeks 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 3 times

✉ **vrodriguesp** 11 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 partition (or just one partition) has 60Milion of records

I have used this logic, simple proportion:

1 partition : 60M = x : 2.4 B ==> 1 : 60 M = x : 2400 M ==> x = 2400 / 60 ==> x = 40 partitions
upvoted 2 times

✉ **vrodriguesp** 11 months ago

1 partition : 60M = x : 2.4 B
1 partition : 60 M = x : 2400 M
==> x = 2400 / 60
==> x = 40 partitions
upvoted 2 times

✉ **SHENO000** 11 months, 1 week ago

Selected Answer: A

Correct Answer

upvoted 1 times

✉ **NORLI** 1 year, 3 months ago

Very simple go with the smallest partition because too many partitions affect performance
upvoted 4 times

✉ **Deeksha1234** 1 year, 4 months ago

Selected Answer: A

correct

upvoted 1 times

✉ **hm358** 1 year, 7 months ago

Selected Answer: A

Optimal distribution is up to 60 instances

upvoted 1 times

✉ **sdokmak** 1 year, 7 months ago

Selected Answer: A

quick maths

upvoted 3 times

✉ **MS_Nikhil** 1 year, 8 months ago

Selected Answer: A

A is correct

upvoted 1 times

✉ **Egocentric** 1 year, 9 months ago

correct

upvoted 1 times

✉ **Twom** 1 year, 10 months ago

Selected Answer: A

Correct

upvoted 2 times

✉ **jskibick** 1 year, 10 months ago

Selected Answer: A

I am also confused.

So we have 2.400.000.000 rows that are already split in 60 nodes od SQL DW. That makes 40.000.000 per node.

Now is question how to order partitions to obtain efficiency for CCI.

Next, we know the partitions will be divided into CCI segments ~1mln per each. And here is my problem. Because CCI will autosplit data in partitions into 1mln row segments. We do not have to do it on our own in partitions. I would split data into monthly partitions i.e. #24 for 2 year, 2019 and 2020. The segments will autosplit partitions.

But there is not such answer.

I will have to go with A = 40

upvoted 6 times

✉ **Justin_beswick** 1 year, 11 months ago

Selected Answer: C

The Rule is Partitions= Records/(1 million * 60)

24,000,000,000/60,000,000 = 400

upvoted 5 times

✉ **panda_azzurro** 11 months, 2 weeks ago

2,400,000,000/60,000,000 = 40

upvoted 1 times

✉ **AlvaroEPMorais** 1 year, 11 months ago

The Rule is Partitions= Records/(1 million * 60)

2,400,000,000/60,000,000 = 40

upvoted 11 times

✉ **helpaws** 1 year, 10 months ago

it's 2.4 billion, not 24 billion

upvoted 14 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.

```
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
)
```

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:

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).

Correct Answer:

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 2 years, 8 months ago
product key is a surrogate key as it is an identity column
upvoted 201 times

111222333 2 years, 7 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 15 times

rikku33 2 years, 3 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 12 times

sagga Highly Voted 2 years, 8 months ago
Type2 because there are start and end columns and ProductKey is a surrogate key. ProductNumber seems a business key.
upvoted 46 times

DrC 2 years, 7 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 106 times

juanlu46 1 year, 3 months ago
Right, its Type 1. We have the updateddatetime and inserteddatetime, not historical
upvoted 10 times

SillyChili 1 month, 3 weeks ago
totally agree that this is type 1. the table has RowUpdatedDateTime, which seems to be the source system's audit column. SCD Type 2 should have date columns to capture its ExpiresOn (for previous record version) and EffectiveFrom (start of the record new version)
upvoted 2 times

borinot 1 year, 1 month 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

blazy001 Most Recent 1 month 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 5 times

 **hassexat** 4 months ago

Type 2 & Surrogate Key
upvoted 1 times

 **AvSUN** 4 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 1 times

 **kkk5566** 4 months, 1 week ago

t2 & is a surrogate key.
upvoted 1 times

 **subhratz** 4 months, 1 week ago

SCD: Type1
ProductKey is Surrogate key.
upvoted 4 times

 **kdp203** 4 months, 2 weeks ago

I agree that it is a surrogate key and not a business one.
upvoted 1 times

 **othman_ee** 6 months, 2 weeks ago

Type1 SCD and SurrogateKey
upvoted 13 times

 **auwia** 6 months, 3 weeks ago

The attribute: RowUpdatedDateTime is a clear indication of: Type 1 SCD.
Type 1 SCD involves updating the dimension attribute with the latest value whenever a change occurs. In this approach, the previous value is overwritten, and there is no historical tracking of changes. This means that only the current version of the dimension data is retained. Type 1 SCD is simple and suitable when historical data is not required or when the dimension changes are not significant.
upvoted 10 times

 **Debasish93** 8 months ago

It is type 2 SCD not because sale start date / sale end date which has no relation whatsoever but because of Insert date & update date. had it been SCD type 1, only one date column i.e. insert/update would have sufficed.
upvoted 2 times

 **rocky48** 8 months, 2 weeks ago

Type 2 SCD because there are start and end columns &
ProductKey is a surrogate key.
upvoted 1 times

 **Honour** 9 months, 2 weeks ago

Please, do not be confused. The ProductKey is a surrogate key, not a business key.
upvoted 3 times

 **esaade** 10 months ago

This table is a Type 1 slowly changing dimension because it overwrites the old data with new data when changes occur, without keeping a history of changes.

The ProductKey column is a surrogate key, which is a system-generated unique identifier for each row in the table. It is not related to any business meaning of the data.

upvoted 12 times

 **Abhishek_C86** 10 months, 3 weeks ago

I have a question for all....

I have an Employee location table that supposedly has columns like emp_id, emp_name, office_loc, start_dt, end_dt
Now if I consider that the employee joins the company and until his retirement, he gets transferred to 4 places so at the end of his career, that table would be having 4 rows.
There would be a time consistency in the start_dt and end_dt column for all 4 rows meaning there won't be any gap within his career that would not be included in this total time covered by the 4 rows

Now, in this example, the sale for a product can have a start_date and an end_date and that sale might happen say 3 times a year but most likely there will be gaps in the timeline as it's a "periodic sale" which does not happen throughout the year or throughout the lifecycle of that product

If for the second case, we consider it to be a SCD in-spite of the time inconsistency, then it's surely a SCD2 else it's SCD1

My question is ... do we consider tables that maintain a SCD type 2 structure but does not maintain a time consistency ... an SCD type2 ?
upvoted 1 times

□ **patvn** 11 months, 1 week ago

Type 1 & Surrogate Key:

Type 1: because there is no value start and end date. The sell start & end date in the picture is when the product actually is sold in the market.
Instead of that, there is modified date which indicates type 1.
surrogate key: even though type 1 doesn't require surrogate key, but this table has another column indicating the product source ID, so product key is most likely surrogate key

upvoted 9 times

□ **Mal2002** 4 months ago

if you are going with the surrogate key then it's most likely not type 1 SCD

upvoted 1 times

□ **SHENO000** 11 months, 1 week ago

It Should be Type 2 , Surrogate Key

upvoted 1 times

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.

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

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?

- A. replicated
- B. hash-distributed on PurchaseKey
- C. round-robin
- D. hash-distributed on DateKey

Correct 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

B (76%)

14%

10%

 AugustineUba  2 years, 5 months 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 54 times

 YipingRuan 2 years, 2 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** 1 year, 2 months ago

A distribution column should have high cardinality to ensure even distribution over nodes.
upvoted 2 times

□ **YipingRuan** 2 years, 2 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** Highly Voted 2 years, 5 months 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 18 times

□ **kamil_k** 1 year, 10 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

□ **AnandEMani** 2 years, 4 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 6 times

□ **waterbender19** 2 years, 5 months 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** 2 years ago

But the DateKey is used in the WHERE clause.
upvoted 2 times

□ **kamil_k** 1 year, 10 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

□ **jongert** Most Recent 2 weeks, 3 days ago

Selected Answer: B

Something not immediately clear to me was that distributing and partitioning are different, hence I was confused that one should not distribute over date columns.

Bottom line is, do not distribute over date columns but you can partition over them. In this question they specifically ask about distribution method. Query optimization for large tables directly points to hashing.

upvoted 2 times

□ **MarkJoh** 1 month, 1 week ago

You want to distribute by productKey and partition by date. Then all distributions will be looked at in parallel and then, within each distribution, only the desired partitions will be looked at. Thereby, the query is fully scaled out and the quickest it can be.
upvoted 3 times

□ **AlejandroU** 2 months, 2 weeks ago

B) the chosen distribution column should not be used in WHERE clauses; thus, we can discard DateKey (even though it is not a Date data type) to minimize data movement. The chosen distribution column must have many unique values; thus we potentially have 2 candidates: PurchaseKey or PurchaseOrderID; however, the chosen one should have no NULLS or only a few, making PurchaseKey the ideal in order to distribute evenly.
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>
upvoted 2 times

□ **Vanq69** 3 months, 1 week ago

Selected Answer: D

Is there any "official" answer to this?

- A. Replicated: Replicated tables have copies of the entire table on each distribution. While this option can eliminate data movement, it may not be the most efficient choice for very large tables with frequent updates.
- B. Hash-Distributed on PurchaseKey: Hash distribution on "PurchaseKey" may lead to data skew if "PurchaseKey" doesn't have a wide range of unique values. Additionally, it doesn't align with the primary filtering condition on "DateKey."
- C. Round-Robin: Round-robin distribution ensures even data distribution, but it doesn't take advantage of data locality for specific types of queries.

D. Hash-Distributed on DateKey: Distributing on "DateKey" aligns with your primary filtering condition, but it's a date column. This could lead to clustering by date, especially if many users filter on the same date.

None of the answers seem to fit. D could be the best guess but it's a date column.
upvoted 2 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: B

B is correct
upvoted 1 times

 **gozdek** 6 months, 2 weeks ago

Selected Answer: C

B is total nonsense if PurchaseKey has a unique value for every row it would end up distributing it evenly so same as round-robin. Distributing by date would slow down the query because in a situation presented in the question only 31 out of 60 distributions would be used. So in my opinion C is the correct answer.

upvoted 1 times

 **SHENO000** 11 months, 1 week ago

Selected Answer: B

B is the correct Answer
upvoted 2 times

 **DindaS** 11 months, 3 weeks ago

To me the answer should be D.
A query on the table that has a WHERE clause filtering on column A will perform partition elimination and scan one partition. That same query may run faster in scenario 2 as there are fewer rows to scan in a partition. A query that has a WHERE clause filtering on column B will scan all partitions.
The query may run faster in scenario 1 than in scenario 2 as there are fewer partitions to scan.
<https://learn.microsoft.com/en-us/sql/relational-databases/partitions/partitioned-tables-and-indexes?view=sql-server-ver16>

Wanted to hear from the experts here.

upvoted 1 times

 **shakes103** 11 months, 3 weeks ago

Selected Answer: B

B is the obvious answer. Hash is optimized for higher analytical performance while Round-robin is optimized for higher loading speed.
upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

B is right
upvoted 1 times

 **vlad888** 1 year, 6 months ago

Anyone who even one time run similar query in Synapse and look into execution plan understand that PurchaseKey doesn't help: there will be shuffle move dms operation! I suppose all these questions has mistake here. Because only column from GROUP BY clause will help. Or round_robin (although it will has almost the same cost as PurchaseKey if last one evenly distributed)

upvoted 1 times

 **Ramkrish39** 1 year, 9 months ago

Agree B is the right answer
upvoted 1 times

 **PallaviPatel** 1 year, 11 months ago

Selected Answer: C

I will go with round robin.
"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 2 times

 **yovi** 2 years ago

Anyone, when you finish an exam, do they give you the correct answers in the end?
upvoted 1 times

 **dev2dev** 1 year, 12 months ago

those finished exam will not know the answer. because answers are not reveled
upvoted 1 times

 **Mahesh_mm** 2 years ago

B is correct ans
upvoted 1 times

You are implementing a batch dataset in the Parquet format.

Data files will be produced by 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.

Correct 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

A (65%)	C (24%)	11%
---------	---------	-----

✉ **m2shines** Highly Voted 2 years ago

Answer should be A, because this talks about minimizing storage costs, not querying costs
upvoted 66 times

✉ **assU2** 1 year, 12 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 weeks, 3 days 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 1 times

✉ **Aslam208** Highly Voted 2 years ago

C is the correct answer, as an external table with a subset of columns with parquet files would be cost-effective.
upvoted 22 times

✉ **RehanRajput** 1 year, 7 months ago

This is not correct.
1. External tables 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

✉ **Aditya0891** 1 year, 6 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 different container and delete the source data from previous container after loading the filtered data to a different container in ADLS
upvoted 2 times

✉ **Aditya0891** 1 year, 6 months ago

check this <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas>. Answer C is correct
upvoted 4 times

✉ **Massy** 1 year, 8 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** 1 year, 1 month 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

 **Ram9198** Most Recent 4 months ago

Selected Answer: A

Snappy

upvoted 2 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: A

using compression

upvoted 2 times

 **kkk5566** 4 months, 2 weeks 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

 **andjurovicela** 5 months, 2 weeks ago

When presented with only the options of column pruning (variant of this is C) and compression (example of it would be snappy), ChatGPT chooses C.

upvoted 1 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: A

I would go by exclusion:

- A. Use Snappy compression for the files. --> nothing against this!
- B. Use OPENROWSET to query the Parquet files. --> doing this I just get a preview of the parquet files
- C. Create an external table that contains a subset of columns from the Parquet files. --> no body asked for a subset
- D. Store all data as string in the Parquet --> nobody asked that

upvoted 8 times

 **semauni** 5 months, 2 weeks ago

Storing data as a string would also make the file size bigger

upvoted 1 times

 **VittalManikonda** 6 months, 2 weeks ago

As per chat gpt , answer is A

upvoted 4 times

 **vctrhugo** 6 months, 2 weeks ago

Selected Answer: A

Snappy compression is a popular and efficient compression algorithm for Parquet files. It provides a good balance between compression ratio and query performance. By compressing the Parquet files using Snappy, you can significantly reduce the storage footprint, leading to lower storage costs.

C is not effective for minimizing storage costs: While creating an external table with a subset of columns can help reduce storage costs, it doesn't specifically address the Parquet format or compression. This option is more related to data modeling and selecting specific columns for query performance rather than minimizing storage costs.

upvoted 3 times

 **rocky48** 8 months, 2 weeks ago

Selected Answer: A

A. Use Snappy compression for the files.

The most effective way to minimize storage costs for the solution is to use compression. Parquet files can be compressed using a variety of codecs, including Snappy, which provides a good balance between compression ratio and query performance. By compressing the data, the file size is reduced, which results in lower storage costs.

Option B, using OPENROWSET to query the Parquet files, is a method to query the data, but it does not address storage costs.

Option C, creating an external table that contains a subset of columns from the Parquet files, is a useful optimization to reduce the amount of data scanned and therefore reduce query cost, but it does not address storage costs.

Option D, storing all data as strings in the Parquet files, is not a good approach because it would result in larger file sizes and potentially slower query performance due to the need to convert the data back to its original format.

Answer is A

upvoted 14 times

 **mamahani** 8 months, 4 weeks ago

Selected Answer: C

i belive the answer is correct; i dont think it should be A, see ms docs "(from the last link: However, when writing to a Parquet file, the service (i.e. data factory) chooses SNAPPY, which is the default for Parquet format. Currently, there is no option to override this behavior.)" this means data factory will already use snappy compression, and we cannot do anything about it, we cant change that; so how can this be a correct answer? if we are forced to use snappy, then we dont choose to apply snappy; i believe external table will be here the answer upvoted 2 times

□ **mamahani** 8 months, 4 weeks ago

also see the question 41;
upvoted 1 times

□ **Hisayuki** 9 months, 3 weeks ago

Selected Answer: C

Answer should be C
upvoted 1 times

□ **esaade** 10 months ago

Selected Answer: A
A. Use Snappy compression for the files.

The Snappy compression algorithm provides a good balance between compression ratio and decompression speed, making it an efficient choice for Parquet files. By compressing the files, you can significantly reduce the storage costs associated with storing large volumes of data. Additionally, Snappy compression is supported by both Azure Data Lake Storage Gen2 and Azure Synapse Analytics, making it an ideal solution for this scenario.

Option B is not a valid solution as OPENROWSET cannot directly query the Parquet files stored in Azure Data Lake Storage Gen2.

Option C may be a possible solution, but it requires additional configuration steps, such as creating external data sources, credentials, and schema mapping.

Option D is not a valid solution because storing all data as string in Parquet files will increase the storage requirements and may impact performance during data retrieval.

upvoted 5 times

□ **DAYENKAR** 11 months ago

Selected Answer: A
option A is correct because que is about storage not about querying
upvoted 1 times

□ **jhargett1** 11 months, 2 weeks ago

A. Use Snappy compression for the files.

Snappy is a high-performance data compression algorithm that is specifically designed for use with Parquet files. It is an open-source algorithm that is known to provide a good balance between compression ratio and processing speed. By using Snappy compression, you can reduce the overall storage costs for the data files, as the compressed files will take up less storage space.

B. Using OPENROWSET to query the Parquet files is a way to query external data from a SQL Server instance, but it does not minimize storage costs.

C. Creating an external table that contains a subset of columns from the Parquet files is a way to optimize the performance of queries, but it does not minimize storage costs.

D. Storing all data as string in the Parquet files will increase the storage size and hence it will increase the storage costs, not minimize it.
upvoted 7 times

□ **Lestrang** 11 months, 2 weeks ago

I do not have a certain solution but..

A. using snappy
it is default, so unless it was another format and changing to snappy. otherwise it is snappy by default, so using it again will not reduce costs because it is already snappy

creating an external table doesn't have to do with querying costs.

creating a standard table has its own storage costs attached.

but an external table serves as a virtual table and allows you to query the data in the external data source using standard SQL. The external table serves as a virtual table and allows you to query the data in the external data source using standard SQL.

So this is reducing storage compared to creating a standard table in the dataset.

I am more inclined to c.

upvoted 3 times

□ **Lestrang** 11 months, 2 weeks ago

When you query the external table, the SQL pool retrieves the data from the Data Lake Gen2 storage account and processes it in memory, it doesn't store a copy of the data in the SQL pool, so it doesn't require extra storage.

upvoted 1 times

□ **shakes103** 11 months, 3 weeks ago

Selected Answer: A

The answer is A. Consider the compression codec to use when writing to Parquet files. When reading from Parquet files, Data Factories automatically determine the compression codec based on the file metadata. Supported types are "none", "gzip", "snappy" (default), and "lzo".

Also you must consider the question is asking for storage cost and not operational (querying included) cost.

[https://learn.microsoft.com/en-us/azure/data-factory/format-parquet#:~:text=The%20compression%20codec%20to%20use%20when%20writing%20to%20Parquet%20files.%20When%20reading%20from%20Parquet%20files%2C%20Data%20Factories%20automatically%20determine%20the%20compression%20codec%20based%20on%20the%20file%20metadata.%0ASupported%20types%20are%20%22none%22%2C%20%22gzip%22%2C%20%22snappy%22%20\(default\)%2C%20and%20%22lzo%22.](https://learn.microsoft.com/en-us/azure/data-factory/format-parquet#:~:text=The%20compression%20codec%20to%20use%20when%20writing%20to%20Parquet%20files.%20When%20reading%20from%20Parquet%20files%2C%20Data%20Factories%20automatically%20determine%20the%20compression%20codec%20based%20on%20the%20file%20metadata.%0ASupported%20types%20are%20%22none%22%2C%20%22gzip%22%2C%20%22snappy%22%20(default)%2C%20and%20%22lzo%22.)

upvoted 1 times

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:

Actions

Create an external file format object

Create an external data source

Create a query that uses Create Table as Select

Create a table

Create an external table

Answer Area**Correct Answer:****Actions**

Create a query that uses Create Table as Select

Create a table

Answer Area

Create an external data source

Create an external file format object

Create an external table

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 2 years, 1 month ago

Looks correct answer

upvoted 34 times

vctrhugo Highly Voted 6 months, 2 weeks 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 13 times

kkk5566 Most Recent 4 months, 1 week ago

source, format ,external

upvoted 1 times

 **CrazyHorse** 1 year, 4 months ago

If I run CREATE EXTERNAL FILE FORMAT before CREATE EXTERNAL DATA SOURCE, will it change something?

upvoted 6 times

 **zekescookies** 8 months, 2 weeks 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** 10 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 2 times

 **aemilka** 9 months, 2 weeks 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** 1 year, 5 months ago

given answer is correct

upvoted 1 times

 **Dicer** 1 year, 5 months ago

why creating a query that creates a table is not correct?

upvoted 1 times

 **VeroDon** 1 year, 5 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** 1 year, 6 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** 1 year, 7 months ago

correct

upvoted 2 times

 **SandipSingha** 1 year, 8 months ago

correct

upvoted 1 times

 **lotuspetall** 1 year, 9 months ago

correct

upvoted 1 times

 **PallaviPatel** 1 year, 11 months ago

correct

upvoted 2 times

 **ANath** 2 years ago

Correct

upvoted 1 times

 **gf2tw** 2 years, 1 month ago

Correct

upvoted 1 times

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

Correct Answer: CE

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 2 years, 1 month ago

Correct Answer . Emp info as Dimension & trans table as fact
upvoted 21 times

 **ellala** Most Recent 3 months ago

Selected Answer: CE

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 1 times

 **hassexat** 4 months ago

Selected Answer: CE

Correct answers
upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: CE

CE is correct

upvoted 1 times

□ **SHENO000** 11 months, 1 week ago

Selected Answer: CE

Correct Answers

upvoted 3 times

□ **kl8585** 1 year, 1 month ago

Selected Answer: CE

Correct answer

upvoted 3 times

□ **Deeksha1234** 1 year, 5 months ago

correct answer - CE

upvoted 2 times

□ **Remedios79** 1 year, 6 months ago

correct

upvoted 2 times

□ **hm358** 1 year, 7 months ago

Selected Answer: CE

Correct

upvoted 3 times

□ **SandipSingha** 1 year, 8 months ago

correct

upvoted 1 times

□ **tg2707** 1 year, 8 months ago

why not fact table for employee and dim table for transactions

upvoted 1 times

□ **Aditya0891** 1 year, 7 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

□ **gabrysr1997** 1 year, 6 months ago

If we would know we wouldn't be here.

upvoted 12 times

□ **Egocentric** 1 year, 8 months ago

CE is correct

upvoted 1 times

□ **NewTuanAnh** 1 year, 9 months ago

Selected Answer: CE

CE is the correct answer

upvoted 2 times

□ **SebK** 1 year, 9 months ago

Selected Answer: CE

CE is correct

upvoted 2 times

□ **surya610** 1 year, 10 months ago

Selected Answer: CE

Dimension for employee and fact for transactions.

upvoted 1 times

□ **PallaviPatel** 1 year, 11 months ago

Selected Answer: CE

correct

upvoted 1 times

□ **gf2tw** 2 years, 1 month ago

Correct

upvoted 2 times

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

Correct 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 2 years, 1 month ago

Correct

upvoted 18 times

 **virendrapsingh** Highly Voted 1 year, 7 months ago

Kind of question that teacher leaves in paper for one free mark.

upvoted 6 times

 **hassexat** Most Recent 4 months ago

Selected Answer: C

Preserve history of changes... Type 2 is correct

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: C

SCD2 is correct

upvoted 1 times

 **akhil5432** 5 months ago

Selected Answer: C

Type 2

upvoted 1 times

 **GodfreyMbizo** 11 months, 2 weeks ago

Type 2 is correct

upvoted 3 times

 **LokeshJ** 1 year ago

Type 2- Maintains Row for each version on dimension data.

upvoted 3 times

 **dimbrici** 1 year, 1 month ago

Selected Answer: C

Correct
upvoted 3 times

 **kl8585** 1 year, 1 month ago

Selected Answer: C

Correct:
type 1 - does not keep history of previous values
type 3 - updates all rows, not required
upvoted 4 times

 **Deeksha1234** 1 year, 5 months ago

right answer given
upvoted 1 times

 **NikeJDI** 1 year, 6 months ago

C is right ~~answ~~
upvoted 1 times

 **Remedios79** 1 year, 6 months ago

correct
upvoted 1 times

 **SandipSingha** 1 year, 8 months ago

correct
upvoted 1 times

 **SandipSingha** 1 year, 8 months ago

correct
upvoted 1 times

 **AZ9997989798979789798979789797** 1 year, 8 months ago

Correct
upvoted 1 times

 **Onobhas01** 1 year, 9 months ago

Selected Answer: C
Correct!
upvoted 1 times

 **surya610** 1 year, 10 months ago

Selected Answer: C
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:

Actions

Create a database scoped credential that uses Azure Active Directory Application and a Service Principal Key

Create an external data source that uses the abfs location

Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages

Create an external file format and set the First_Row option

Answer Area**Correct Answer:****Actions**

Create a database scoped credential that uses Azure Active Directory Application and a Service Principal Key

Answer Area

Create an external data source that uses the abfs location



Create an external file format and set the First_Row option

Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages

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**  1 year, 3 months ago

- 1) create database scoped credentials
- 2) create external source
- 3) create file format
- 4) create external table (it not supports CTAS)

upvoted 28 times

□  **auwia** 6 months, 2 weeks 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 3 times

□  **OldSchool**  1 year, 1 month 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 16 times

□  **Rob77** 7 months, 3 weeks ago

No, CETAS is not used for loading Azure Synapse Analytics. It's used to export data from and not to!

upvoted 2 times

□  **hydmt07**  1 month 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 1 times

□  **ellala** 3 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** 3 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 1 times

□  **Ram9198** 4 months ago

Before building the loading pattern, you need to prepare the required database objects in Azure Synapse Analytics - Loading pattern is CETAS, so answer

DSC

DS

EFF

upvoted 1 times

□  **kkk5566** 4 months, 1 week ago

source ,format ,external

upvoted 2 times

□  **eladioyovera** 5 months, 2 weeks 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** 6 months, 2 weeks 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 4 times

□  **mamahani** 8 months, 3 weeks 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** 8 months, 3 weeks 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

□  **olegjdl** 11 months, 1 week 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 5 times

□  **Rakrah** 11 months, 3 weeks 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** 1 year 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 :

- 1) create database scoped credentials
- 2) create external source
- 3) create file format

upvoted 3 times

□  **rohanb1986** 1 year 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** 8 months, 3 weeks 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

□  **mamahani** 8 months, 3 weeks 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 1 times

□  **Bro111** 1 year, 1 month ago

What is azure active directory application? is it managed identity?

upvoted 2 times

□  **kl8585** 1 year, 1 month ago

I have many doubts on this question. Two points on which i focused:

- 1) The question states "Before building the loading pattern, you need to prepare the required database objects", so it doesn't seem to fit with CETAS statement because is asking about actions to do BEFORE implementing the loading pattern
- 2) With PolyBase you don't use CETAS statement, you should use CTAS with LOCATION option to create external table from Data Lake

Given these points, i will go with:

- Create database scoped credentials

- Create external data source
- Create file format

I will appreciate anyone confirming or confuting this answer. Thank you!
upvoted 7 times

✉ **mamahani** 8 months, 3 weeks ago

yes you do use CTAS with polybase , see ms docs "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

✉ **RBKasemodel** 1 year, 1 month ago

In the link below they make it very clear.
First, database scoped.
2nd External Data source,
3rd, File format.

Only after that you can create an external table in polybase

<https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects?view=sql-server-ver16>
upvoted 4 times

✉ **Billybob0604** 1 year, 1 month ago

The first step is only required only for Kerberos-secured Hadoop clusters.
upvoted 2 times

✉ **fionacanderson** 10 months, 2 weeks ago

you didnt scroll down the page far enough
upvoted 2 times

✉ **kl8585** 1 year, 1 month ago

thank you!
upvoted 1 times

✉ **Igor85** 1 year, 2 months ago

again, a question that is lacking precision. it suggests choosing three actions whereas all 4 makes sense, there is no indication that database scoped credentials are already in place
upvoted 4 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)
```

Correct 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
```

店铺：IT认证考试服务

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 2 years, 1 month ago

Correct

upvoted 22 times

OldSchool Highly Voted 1 year, 1 month ago

Correct answer, giveaway is "RANGE RIGHT"

upvoted 7 times

M_Anas_007 Most Recent 3 months, 1 week ago

Have a small doubt.

We are creating partition on what field?

Shouldn't it be the columnstore index.

upvoted 1 times

kkk5566 4 months, 1 week ago

Partition

upvoted 1 times

GodfreyMbizo 11 months, 2 weeks ago

correct answer given

upvoted 5 times

 **Deeksha1234** 1 year, 5 months ago

ans is correct

upvoted 3 times

 **NikeJDI** 1 year, 6 months 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** 1 year, 8 months ago

How are we ensuring "Minimizes the processing time to delete data that is older than 10 years"?

upvoted 2 times

 **Aditya0891** 1 year, 7 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 3 times

 **allagowf** 1 year, 3 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 2 times

 **wwdba** 1 year, 10 months ago

Correct

upvoted 2 times

 **PallaviPatel** 1 year, 11 months ago

correct

upvoted 2 times

 **saupats** 2 years ago

correct

upvoted 2 times

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.

```
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;
```

What do the query results include?

- A. Only CSV files in the tripdata_2020 subfolder.
- B. All files that have file names that begin with "tripdata_2020".
- C. All CSV files that have file names that contain "tripdata_2020".
- D. Only CSV that have file names that begin with "tripdata_2020".

Correct Answer: D

Community vote distribution

D (100%)

 **gf2tw** Highly Voted 2 years, 1 month ago

Correct

upvoted 18 times

 **akshaynag95** Highly Voted 11 months, 1 week ago

Selected Answer: D

D is the correct Answer

upvoted 5 times

 **hassexat** Most Recent 4 months ago

Selected Answer: D

Correct is D

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: D

D is correct

upvoted 1 times

 **henryphchan** 8 months, 1 week ago

Selected Answer: D

correct

upvoted 3 times

 **SHENO000** 11 months, 1 week ago

Selected Answer: D

D is the correct Answer

upvoted 3 times

 **panda_azzurro** 11 months, 2 weeks 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

 **Anton2020** 11 months, 1 week ago

Only file, there is no / after 2020 in the exercise.

upvoted 2 times

 **Stokstaartje420** 11 months, 2 weeks ago

Answer D is correct. Would be great if it was also correct grammatically.

upvoted 3 times

 **kl8585** 1 year, 1 month ago

Selected Answer: D

Correct

upvoted 1 times

 **GauravPurandare** 1 year, 5 months ago

Selected Answer: D

CSV that have file names that beginning with "tripdata_2020" .

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

D is correct

upvoted 1 times

 **objecto** 1 year, 7 months ago

Selected Answer: D

Correct

upvoted 2 times

 **Rrk07** 1 year, 7 months ago

D is correct

upvoted 1 times

 **Egocentric** 1 year, 8 months ago

on this one you need to pay attention to wording

upvoted 3 times

 **jskibick** 1 year, 9 months ago

Selected Answer: D

D all good

upvoted 1 times

 **sarapaisley** 1 year, 9 months ago

Selected Answer: D

D is correct

upvoted 1 times

 **SebK** 1 year, 9 months ago

Selected Answer: D

Correct

upvoted 1 times

DRAG DROP -

You use PySpark in Azure Databricks to parse the following JSON input.

```
{
  "persons": [
    {
      "name": "Keith",
      "age": 30,
      "dogs": ["Fido", "Fluffy"]
    },
    {
      "name": "Donna",
      "age": 46,
      "dogs": ["Spot"]
    }
  ]
}
```

You need to output the data in the following tabular format.

owner	age	dog
Keith	30	Fido
Keith	30	Fluffy
Donna	46	Spot

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 split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

Values	Answer Area
alias	dbutils.fs.put("/tmp/source.json", source_json, True) source_df = spark.read.option("multiline", "true").json("/tmp/source.json")
array_union	persons = source_df. <input type="text"/> Value <input type="text"/> Value ("persons").alias("persons")
createDataFrame	persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"),
explode	explode <input type="text"/> Value ("dog")) ("persons-dogs"). display(persons_dogs)
select	
translate	

Correct Answer:

Values	Answer Area
array_union	dbutils.fs.put("/tmp/source.json", source_json, True) source_df = spark.read.option("multiline", "true").json("/tmp/source.json")
createDataFrame	persons = source_df. select <input type="text"/> explode <input type="text"/> ("persons").alias("persons")
	persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"),
	explode <input type="text"/> alias <input type="text"/> ("dog")) ("persons-dogs"). display(persons_dogs)
translate	

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 1 year, 6 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 22 times

urassi Highly Voted 10 months, 1 week ago

ah "persons".alias("persons") what a fun and useful and nice alias
upvoted 12 times

kkk5566 Most Recent 4 months, 1 week ago

syntax is correct
upvoted 1 times

esaade 10 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 10 times

Deeksha1234 1 year, 5 months ago

Correct
upvoted 4 times

Dicer 1 year, 5 months ago

Correct, but last .alias("dog") is quite unnecessary because the column name is already 'dog'. I guess that is for safety measurement.
upvoted 5 times

Anton2020 10 months, 2 weeks ago

The column name in the json is dogs, not dog
upvoted 3 times

galacaw 1 year, 8 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

First week:

Archive
Cool
Hot

After one month:

Archive
Cool
Hot

After one year:

Archive
Cool
Hot

Answer Area

First week:

Archive
Cool
Hot

After one month:

Archive
Cool
Hot

After one year:

Archive
Cool
Hot

Correct Answer:

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.

	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

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 1 year, 5 months ago

right , Hot-cool-cool
upvoted 19 times

□ **Andy91** Highly Voted 1 year, 8 months ago

Correct answer!
Hot, Cool, Cool
upvoted 9 times

□ **hassexat** Most Recent 4 months ago

Hot
Cool
Cool
upvoted 2 times

□ **kkk5566** 4 months, 1 week ago

hot,cool,cool
upvoted 2 times

□ **rocky48** 8 months, 1 week ago

Hot, Cool, Cool
upvoted 3 times

□ **Fishy_Marcy** 1 year, 6 months 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** 1 year, 6 months 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 5 times

□  **nefarious_smalls** 1 year, 8 months ago

Why would it not be be Hot Cool and Archive

upvoted 1 times

□  **Guincimund** 1 year, 8 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 19 times

□  **SandipSingha** 1 year, 8 months ago

After one year, the data will be accessed infrequently but must be accessible within five minutes.

upvoted 5 times

□  **AnonymousJhb** 8 months, 1 week ago

hydration on the archive tier is hours, not minutes. hence its hot, cool, cool.

upvoted 2 times

□  **nefarious_smalls** 1 year, 8 months ago

I dont know

upvoted 1 times

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.

Correct 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%)	10%
---------	-----

 **galacaw** Highly Voted  1 year, 8 months ago

Should be D, it's about Apache Spark pool, not serverless SQL pool.
upvoted 34 times

 **ellala** Most Recent  3 months ago

Selected Answer: D

We have an "Azure Synapse Analytics Apache Spark pool" therefore, we use Spark. There is no information about a serverless SQL Pool
upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: D

Should be D
upvoted 2 times

 **vctrhugo** 6 months, 2 weeks ago

Selected Answer: D

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** 7 months ago

Selected Answer: D

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** 7 months, 3 weeks ago

Option D: Load the data by using PySpark

upvoted 1 times

✉ **henryphchan** 8 months, 1 week ago

Selected Answer: D

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** 8 months, 3 weeks 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 3 times

✉ **Tejashu** 1 month, 2 weeks 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

✉ **esaade** 10 months ago

Selected Answer: D

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** 10 months, 3 weeks ago

Selected Answer: D

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** 1 year, 1 month ago

Selected Answer: D

Should be D, it's about Apache Spark pool, not serverless SQL pool.

upvoted 2 times

✉ **smsme323** 1 year, 3 months ago

Selected Answer: D

Its a spark pool

upvoted 2 times

✉ **Deeksha1234** 1 year, 5 months ago

Both C and D looks correct

upvoted 2 times

✉ **Oli388** 1 year, 5 months ago

I agree with D, I already used PySpark for exactly that purpose. But what is the problem with C? Could it be that both answers are correct?
upvoted 1 times

✉ **Gg2** 1 year, 4 months ago

The problem is that "The solution MUST maintain the source data types."

upvoted 3 times

✉ **jiajiani** 3 months, 2 weeks ago

but openrowset can also work with different data types

upvoted 2 times

✉ **Dicer** 1 year, 5 months ago

D

First, there is the Apache Spark pool. And Azure Data Lake is based on Apache Spark Delta Lake. The most suitable answer is D.
upvoted 3 times

✉ **am85** 1 year, 6 months ago

Selected Answer: D

It should be D.

upvoted 3 times

✉ **jihenTR13** 1 year, 6 months ago

should be D, apache spark has an option inferschema that can be used to read the file and infer the schema from it automatically without the need to explicitly specify it

upvoted 1 times

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.

Correct 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 (71%)

C (29%)

✉  **hkay** Highly Voted 1 year, 7 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 8 months ago

D is accurate answer and this link show this info explicitly
<https://learn.microsoft.com/en-us/azure/databricks/clusters/cluster-config-best-practices>
upvoted 9 times

✉  **positivitypeople** Most Recent 2 weeks, 5 days ago

Got this question today on the exam. Selected Upgrade to Premium tier, was incorrect of course
upvoted 1 times

✉  **kkk5566** 4 months, 1 week ago

Selected Answer: D
D is correct
upvoted 1 times

✉  **Abdullah77** 4 months, 3 weeks ago

C is the answer
upvoted 1 times

✉  **auwia** 6 months, 3 weeks ago

Selected Answer: D
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** 7 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** 7 months ago

How can it be C), as the question clearly states 'minimize costs'

upvoted 1 times

 **janaki** 7 months, 3 weeks 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** 8 months ago

Selected Answer: D

D is correct answer

upvoted 1 times

 **rocky48** 8 months, 1 week ago

Selected Answer: D

Correct Answer: D

upvoted 1 times

 **henryphchan** 8 months, 1 week ago

Selected Answer: D

C and D are practical but the question asked to minimize the costs. Thus, the answer is D.

upvoted 2 times

 **jamesraju** 9 months ago

D is correct if anyone need paid pdf, then send me the 8885722344

upvoted 1 times

 **mehroosali** 9 months, 1 week ago

Selected Answer: D

D is correct

upvoted 2 times

 **halamgir15** 9 months, 4 weeks ago

Selected Answer: C

I think C makes more sense

upvoted 3 times

 **SHENO000** 11 months, 1 week ago

Selected Answer: D

D is the correct answer, as you need to lower the cost

upvoted 1 times

 **nicky87654** 12 months ago

Selected Answer: D

Correct answer is 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.)

Input Details ×

products

Container
 Create new Use existing

refdata

Path pattern 店铺: IT认证考试服务

Date format

Time format

Event serialization format * ①

Delimiter ①

Encoding ①

- ① If the chosen resource and the stream analytics job are located in different regions, you will be billed to move data between regions.

The storage account container view is shown in the Refdata exhibit. (Click the Refdata tab.)

refdata Container

« Renaming

Authentication method: Access key ([Switch to Azure AD User Account](#))
Location: refdata / 2020-03-20

Search blobs by prefix (case-sensitive)

Name
<input type="checkbox"/> [..]
<input type="checkbox"/> product.csv

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

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

Correct 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 1 year, 5 months ago

correct.

Reasons:

1. {date}/{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 18 times

□  **Rakrah** Highly Voted 11 months, 3 weeks 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 16 times

□  **kkk5566** Most Recent 4 months, 1 week ago

correct

upvoted 1 times

□  **Rakrah** 11 months, 3 weeks ago

The recommended way to refresh reference data is to:

Use {date}/{time} in the path pattern. Box 1 should be {date}/{time}/product.csv

upvoted 2 times

□  **Deeksha1234** 1 year, 5 months ago

given answer is correct

upvoted 2 times

□  **Franz58** 1 year, 5 months ago

1. {date}/product.csv

2. YYYY-MM-DD

upvoted 2 times

□  **temacc** 1 year, 5 months 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** 1 year, 6 months 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

□  **JayOR** 5 months, 1 week 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 1 times

□ **Revave2** 1 year, 6 months 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 4 times

□ **TriumphMC** 10 months, 4 weeks 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

□ **demirsamuel** 1 year, 7 months ago

answers are correct

upvoted 2 times

□ **inotbf83** 1 year, 8 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** 1 year, 8 months ago

The file is updated daily, i think `{date}/product.csv` is correct

upvoted 5 times

□ **Lotusss** 1 year, 8 months ago

Wrong! Path Pattern: {date}/{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

□ **KashRaynardMorse** 1 year, 8 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 7 times

HOTSPOT -

You have the following Azure Stream Analytics query.

```
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
```

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:

Answer Area

Statements	Yes	No
The query combines two streams of partitioned data.	<input type="radio"/>	<input type="radio"/>
The stream scheme key and count must match the output scheme.	<input type="radio"/>	<input type="radio"/>
Providing 60 streaming units will optimize the performance of the query.	<input type="radio"/>	<input type="radio"/>

Correct Answer:

Answer Area

Statements	Yes	No
The query combines two streams of partitioned data.	<input type="radio"/>	<input type="radio"/>
The stream scheme key and count must match the output scheme.	<input checked="" type="radio"/>	<input type="radio"/>
Providing 60 streaming units will optimize the performance of the query.	<input checked="" type="radio"/>	<input type="radio"/>

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 $6 \times 10 = 60$ 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 1 year, 7 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 33 times

oleg25 3 months, 1 week 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 understandng 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 Highly Voted 6 months, 2 weeks 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 8 times

Momoanwar Most Recent 1 month ago

Chatpt 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/>]](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/>]](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/>]](https://azure.microsoft.com/fr-fr/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/).

upvoted 3 times

ellala 3 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, $6 \times 10 = 60$.

upvoted 3 times

ellala 3 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 technically, it would be 21 SU V2.
Therefore, 60 SU is not correct.

Should be
FALSE
TRUE
FALSE
upvoted 2 times

✉ ahmadsayeed 2 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 3 months, 1 week 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 $22 \times 6 = 132$ SU V1.
upvoted 1 times

✉ mav2000 3 months, 1 week ago

Based on recommended Streaming units,
Step 1: 10 partitions
Step 2: 10 partitions

$(1 \times 10 + 1 \times 10) = 20$ 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 4 months, 2 weeks 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 6 months, 3 weeks 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 1 times

✉ auwia 6 months, 2 weeks 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 3 times

✉ **auwia** 6 months, 3 weeks 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** 4 months, 1 week ago

All answer is no.
upvoted 1 times

✉ **kkk5566** 4 months, 1 week ago

forgot it, no, yes, no
upvoted 1 times

✉ **UristMcFarmer** 1 year 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** 1 year 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** 11 months, 1 week 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** 1 year, 1 month ago

All 3 are YES
upvoted 3 times

✉ **OldSchool** 1 year, 1 month ago

At first I thought all 3 are Y but then; Streams are Input and UNION combines query results only so in that case first is No.
upvoted 1 times

✉ **dduque10** 1 year, 1 month ago

The explanation of the second Yes indicates that the first one is also Yes
upvoted 1 times

 **SomethingRight100** 1 year, 1 month ago

for the first box, I voted no. I think the correct expression may be something like "the query combine two partitioned data into one stream". I do not think there are two streams

upvoted 1 times

 **Bro111** 1 year, 1 month ago

do you have an example of combining two streams, please.

upvoted 3 times

 **ca_acc** 1 year, 2 months ago

The reasoning behind the first one could be: union combines two RESULTSETS, not two streams.

upvoted 1 times

 **Deeksha1234** 1 year, 4 months ago

all three yes

upvoted 4 times

 **dsp17** 1 year, 5 months ago

Should be ALL 3 YES.. Union will combine 2 stream output.

upvoted 6 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	applications
CREATE TABLE	
CREATE VIEW	

```

WITH (
  LOCATION = 'applications/',
  DATA_SOURCE = applications_ds,
  FILE_FORMAT = applications_file_format
)
AS
SELECT id, [address_housenumber] as addresshousenumber, [address_line1] as addressline1
FROM
  (BULK 'https://contosol1.dfs.core.windows.net/applications/year=*/*.parquet',
CROSS APPLY
OPENJSON
OPENROWSET
FORMAT='PARQUET') AS [r]
GO

```

Correct Answer:

Answer Area

CREATE EXTERNAL TABLE	applications
CREATE TABLE	
CREATE VIEW	

```

WITH (
  LOCATION = 'applications/',
  DATA_SOURCE = applications_ds,
  FILE_FORMAT = applications_file_format
)
AS
SELECT id, [address_housenumber] as addresshousenumber, [address_line1] as addressline1
FROM
  (BULK 'https://contosol1.dfs.core.windows.net/applications/year=*/*.parquet',
CROSS APPLY
OPENJSON
OPENROWSET
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 }
( <column_definition> [ ,...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 1 year, 5 months ago

Correct answer
upvoted 18 times

kkk5566 Most Recent 4 months, 1 week ago

correct
upvoted 1 times

cjb0 1 year, 4 months ago

Why is it External Table if it's a Serverless SQL pool?
upvoted 2 times

vctrhugo 6 months, 2 weeks ago

"You have data stored in Parquet files in an Azure Data Lake Storege Gen2 container."
upvoted 2 times

NORLI 1 year, 3 months ago

because serverless SQL pool does not have internal tables
upvoted 14 times

dkamat 1 year, 6 months ago

correct
upvoted 3 times

SandipSingha 1 year, 8 months ago

correct
upvoted 3 times

Feljoud 1 year, 8 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

```
店铺: IT认证考试服务
CREATE EXTERNAL DATA SOURCE source1
WITH
    ( LOCATION = 'https://account1..core.windows.net', .core.windows.net',
        PUSHDOWN = ON
        TYPE = BLOB_STORAGE
        TYPE = HADOOP
)
```

blob
dfs
table

Correct Answer:

Answer Area

```
CREATE EXTERNAL DATA SOURCE source1
WITH
    ( LOCATION = 'https://account1..core.windows.net', .core.windows.net',
        PUSHDOWN = ON
        TYPE = BLOB_STORAGE
        TYPE = HADOOP
)
```

blob
dfs
table

Box 1: blob -

The following example creates an external data source for Azure Data Lake Gen2

CREATE EXTERNAL DATA SOURCE YellowTaxi

```
WITH ( LOCATION = 'https://azuredatadatalakestorage.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 1 year, 8 months ago

1. dfs (for Azure Data Lake Storage Gen2)

upvoted 51 times

 **Vedjha** 11 months, 1 week ago

```
CREATE EXTERNAL DATA SOURCE mydatasource
WITH ( LOCATION = 'abfss://data@storageaccount.dfs.core.windows.net',
    CREDENTIAL = AzureStorageCredential,
    TYPE = HADOOP
)
```

upvoted 4 times

□ **Rob77** 7 months, 3 weeks 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 2 times

□ **jds0** 9 months, 2 weeks 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

□ **panda_azzurro** 11 months, 2 weeks ago

dfs is not valid

upvoted 1 times

□ **suvec** 9 months ago

dfs is valid

Data Lake Storage Gen2

abfs[s] <container>@<storage_account>.dfs.core.windows.net

http[s] <storage_account>.dfs.core.windows.net/<container>/subfolders

wasb[s] <container>@<storage_account>.blob.core.windows.net

upvoted 5 times

店铺：IT认证考试服务

□ **Kure87** Highly Voted 1 year, 1 month ago

1. blob. According 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 31 times

□ **vadiminski_a** 9 months, 3 weeks 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 6 times

□ **suvec** 9 months ago

@kure87 DFS is valid

Data Lake Storage Gen2

abfs[s] <container>@<storage_account>.dfs.core.windows.net

http[s] <storage_account>.dfs.core.windows.net/<container>/subfolders

wasb[s] <container>@<storage_account>.blob.core.windows.net

upvoted 5 times

□ **Sebastian1677** 1 year ago

please upvote this

upvoted 2 times

□ **Momoanwar** Most Recent 3 weeks 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** 3 weeks 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

店铺：IT认证考试服务

□ **Qordata** 3 months, 2 weeks ago

Answer is CORRECT: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#example-for-create-external-data-source>

CREATE EXTERNAL DATA SOURCE YellowTaxi

WITH (LOCATION = 'https://azuredataport.blob.core.windows.net/nyctlc/yellow/' ,

TYPE = HADOOP)

upvoted 1 times

□ **fahfouhi94** 3 months, 2 weeks 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 4 months, 1 week 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 4 months, 2 weeks ago

dfs should be the correct answer (ADLS Gen2)
upvoted 1 times

auwia 6 months, 3 weeks ago

Confirmed HADOOP and DFS:
External Data Source | Connector | Location path

```
Data Lake Storage Gen1 | adl | <storage_account>.azuredatalake.net
Data Lake Storage Gen2 | abfs[s] | <container>@<storage_account>.dfs.core.windows.net
Azure Blob Storage | wasbs | <container>@<storage_account>.blob.core.windows.net
Azure Blob Storage | https | <storage_account>.blob.core.windows.net/<container>/subfolders
Data Lake Storage Gen1 | http[s] | <storage_account>.azuredatalakestore.net/webhdfs/v1
Data Lake Storage Gen2 | http[s] | <storage_account>.dfs.core.windows.net/<container>/subfolders
Data Lake Storage Gen2 | wasb[s] | <container>@<storage_account>.blob.core.windows.net
```

upvoted 3 times

auwia 6 months, 3 weeks ago

```
CREATE EXTERNAL DATA SOURCE source1
WITH (
LOCATION = 'https://account1.dfs.core.windows.net',
TYPE = HADOOP
)
upvoted 1 times
```

aga444 7 months ago

```
CREATE EXTERNAL DATA SOURCE DataSourceName
WITH (
TYPE = HADOOP,
LOCATION = 'adl://Account1.dfs.core.windows.net/',
CREDENTIAL = SqlPoolCredential
);
upvoted 1 times
```

janaki 7 months, 3 weeks ago

```
CREATE EXTERNAL DATA SOURCE <datasource_name>
WITH (
TYPE = HADOOP,
LOCATION = 'adl://<account_name>.dfs.core.windows.net',
CREDENTIAL = <credential_name>
);
```

So answer is dfs and Type = Hadoop
upvoted 1 times

Reloadedvn 8 months ago

1. blob
2. TYPE=HADOOP
Source: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop>
The following example creates an external data source for Azure Data Lake Gen2 pointing to the publicly available New York data set:
CREATE EXTERNAL DATA SOURCE YellowTaxi
WITH (LOCATION = 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/',
TYPE = HADOOP)
upvoted 3 times

Fredward_95 8 months, 2 weeks ago

Hadoop is the only allowed type in dedicated SQL pools and for ADLS Gen2 with http[s] prefix it's definitely dfs as you can lookup here
<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 2 times

suvec 9 months ago

Data Lake Storage Gen2
abfs[s] <container>@<storage_account>.dfs.core.windows.net
http[s] <storage_account>.dfs.core.windows.net/<container>/subfolders
wasb[s] <container>@<storage_account>.blob.core.windows.net
upvoted 1 times

✉ **deutscher** 9 months ago

Answer is: Blob and Hadoop

I found this in Azure documentation:

```
CREATE EXTERNAL DATA SOURCE YellowTaxi
WITH ( LOCATION = 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/',
TYPE = HADOOP)
```

Link:

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop>

upvoted 2 times

✉ **[Removed]** 9 months, 2 weeks ago

correct answer

upvoted 1 times

✉ **Camarade_Emile** 10 months ago

1) dfs because of "Dedicated SQL Pool"

2) HADOOP

reference: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop>

upvoted 2 times

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

Correct 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 1 year, 6 months ago

If the answer has Parquet. Of course, you should choose that answer. :D
upvoted 16 times

✉  **ClassMistress** Highly Voted 1 year, 7 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 14 times

✉  **sdokmak** 1 year, 7 months ago

Good point, also better cost
upvoted 3 times

✉  **phydev** Most Recent 2 months, 1 week ago

Parquet is always the answer ;-)
upvoted 2 times

✉  **kkk5566** 4 months, 1 week ago

Selected Answer: B
is correct
upvoted 1 times

✉  **akhil5432** 5 months ago

Selected Answer: B
Parquet
upvoted 1 times

✉  **ajhak** 7 months, 3 weeks ago

Selected Answer: B
When in doubt, select Parquet.
upvoted 5 times

✉  **greenlever** 1 year, 3 months ago

Selected Answer: B
When reading from Parquet files, you can specify only the columns you want to read and skip the rest.
upvoted 4 times

 **Rahuar** 1 year, 4 months ago

CORRECT

upvoted 1 times

 **monibun** 1 year, 4 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 2 times

 **Ranjan6214** 1 week, 1 day ago

@monibun : great point. I agree with you . I am not sure if anyone wants to advise on your point .

upvoted 1 times

 **Deeksha1234** 1 year, 5 months ago

Parquet 

upvoted 2 times

 **shachar_ash** 1 year, 8 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)
    )
)

```

Correct Answer:

Values
CustomerKey
ROUND_ROBIN
REPLICATE
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 = HASH ([ProductKey])
    , PARTITION ( [ OrderDateKey ] RANGE RIGHT FOR VALUES
        (20170101,20180101,20190101,20200101,20210101)
    )
)

```

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** Highly Voted 1 year, 7 months ago

I think it is Hash because the question refer to a Fact table.

upvoted 19 times

□ **hereiamken** Highly Voted 10 months, 3 weeks ago

1. Hash -> Fact Table
 2. DateKey -> for Partition
- upvoted 13 times

□ **kkk5566** Most Recent 4 months, 1 week ago

the syntax is ok only for HASH &
Datekey

upvoted 2 times

□ **kumarsunny** 4 months, 1 week ago

Why not 'Product Key' for partition? can anyone explain me please.

upvoted 1 times

□ **MJamesP** 3 months, 2 weeks 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** 6 months, 2 weeks ago

if it is round robin, there is no key to specify, so hash
upvoted 1 times

□ **[Removed]** 9 months, 2 weeks ago

It must be HASH because of syntax.
upvoted 5 times

□ **SHENO000** 11 months, 1 week ago

The Answer is correct
upvoted 3 times

□ **astone42** 11 months, 2 weeks ago

The answer is correct.
upvoted 2 times

□ **Dindas** 11 months, 3 weeks ago

Should be Round Robin as the requirement is to have the data evenly. the second one should be on the date
upvoted 2 times

□ **auwia** 6 months, 3 weeks ago

You would use Round-Robin for staging table and not Fact table.
upvoted 3 times

□ **Ritik37** 9 months, 4 weeks ago

It should, but they have given attributes. So only hash supports attribute
upvoted 1 times

□ **allagowf** 1 year, 2 months ago

```
<distribution_option> ::=  
{  
  DISTRIBUTION = HASH ( distribution_column_name )  
  | DISTRIBUTION = ROUND_ROBIN  
  | DISTRIBUTION = REPLICATE  
}
```

+ fact table
it's for sure :: hash
upvoted 4 times

greenlever 1 year, 3 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 1 year, 4 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 4 times

Deeksha1234 1 year, 5 months ago

Answer is correct
upvoted 2 times

Dicer 1 year, 5 months ago

should be round robin because of distributed evenly.
upvoted 1 times

anks84 1 year, 4 months ago

syntax used is for HASH distribution.
upvoted 4 times

Franz58 1 year, 5 months ago

correct
upvoted 1 times

jebias 1 year, 8 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

Feljoud 1 year, 8 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 25 times

dduque10 1 year, 1 month ago

For small tables is recommended replicated, not round robin
upvoted 1 times

nefarious_smalls 1 year, 8 months ago

You are exactly righty
upvoted 1 times

sivva 1 year, 5 months ago

@Feljoud : Thanks for the clarification. Even I opted for Roundrobin, considering the keywords = "distributed evenly", but that's incorrect.
upvoted 2 times

Massy 1 year, 8 months ago

the syntax is ok only for HASH
upvoted 6 times

Muishkin 1 year, 8 months ago

yes i think so too
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

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.

Answer Area

Data over five years old:

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

Correct Answer:

Data over seven years old:

- 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.

	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

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

✉️ **sagur** Highly Voted 1 year, 8 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 31 times

✉️ **Feljoud** 1 year, 8 months ago

Would agree, but the question states: "a data archiving solution", so maybe to keep the data was implied with this?
 upvoted 16 times

✉️ **AnonymousJhb** 8 months ago

no. part of data management is deleted data that is no longer required. not keeping all data forever. you are allowed to delete data once it meets required guardrails. delete deprecated data > 7 years.
 upvoted 4 times

✉️ **Massy** 1 year, 8 months ago

I agree, should be deleted
 upvoted 5 times

✉️ **shakes103** 11 months, 3 weeks ago

Be careful with wording. Any answer given must be an "Archiving solution" & Delete the blob is not an archiving solution.
 upvoted 5 times

✉️ **AnonymousJhb** 8 months ago

of course it is.
 upvoted 4 times

✉️ **KashRaynardMorse** 1 year, 8 months ago

Deleting data older than 7 years is not an option available in the answer list. Be careful 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 24 times

✉️ **vctrhugo** 6 months, 2 weeks ago

God no... A single piece of data could be a BLOB (Binary Large OBJECT).
 upvoted 1 times

✉️ **Boompiee** 1 year, 8 months ago

I'm confused by your comment. It clearly does state an option to delete the blob after 7 years.
 upvoted 2 times

✉️ **Aditya0891** 1 year, 7 months ago

what he meant to say is all the data be it in hot, cool or archive 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

✉ **cgartiamarco1** 1 year, 3 months ago

Why is 'Delete the blob' not a valid option? Given that seven years data is not accessed, why don't we delete the blob?
upvoted 1 times

✉ **noobprogrammer** 1 year, 8 months ago

Makes sense to me
upvoted 1 times

✉ **PeteZaria** Highly Voted 1 year, 3 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 8 times

✉ **d046bc0** Most Recent 4 weeks, 1 day ago

correct. deleting files is not a type of archiving them (chatGPT)
upvoted 1 times

✉ **hassexat** 4 months ago

Cool & Archive

No deletion of data here because question require to perform a data archiving solution
upvoted 1 times

✉ **kkk5566** 4 months, 2 weeks ago

correct
upvoted 2 times

✉ **Saintu** 4 months, 3 weeks 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 sync 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** 6 months, 3 weeks ago

"Costs must be minimized while maintaining the required availability." ==> it means Move to archive storage"
upvoted 1 times

✉ **GodfreyMbizo** 11 months, 2 weeks ago

The Given answer is correct, the big tip here is archiving solution, deleting is not archiving solution
upvoted 5 times

✉ **sumanthss** 1 year, 2 months ago

Archive might be the right option as we need to retrieve before we access the data. so we cannot access without retrieval.
upvoted 1 times

✉ **cgartiamarco1** 1 year, 3 months ago

Why is 'Delete the blob' not a valid option? Given that seven years data is not accessed, why don't we delete the blob?
upvoted 1 times

✉ **Deeksha1234** 1 year, 5 months ago

Hot for 5 years and delete if older than 7 yrs since it'll not be accessed
upvoted 1 times

Data will be accessed rarely.. why waste money on high storage on hot? cool has low storage cost and high cost for accessing which suites better
upvoted 1 times

If data needs to be available even though we don't access it then we can put it in Archive layer
upvoted 1 times

I think shouldn't be deleted because we need data even though it is 7 years old, but we don't access it..
upvoted 1 times

✉️  **makkelijkzat** 8 months, 1 week ago

doesn't say that. It's not accessed, never ever. So delete seems the only logical option.

upvoted 1 times

✉️  **kishan_peter_pandey** 1 year, 5 months ago

Data should be archived as in question it's mentioned "while maintaining required availability"

upvoted 2 times

✉️  **namtn6** 1 year, 6 months ago

correct answer without doubt

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 |

Correct Answer:

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 |

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 1 year, 2 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 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 56 times

 **SM94** 1 month ago

its data center failure not region failure. ZRS is correct

upvoted 2 times

✉  **vctrhugo** 6 months, 2 weeks 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 8 times

✉  **SomethingRight100** 1 year, 1 month ago

Reading from the following, I do not think RAGZRS is ready only. I think it is read-only for the second region.

GZRS writes three copies of your data synchronously across multiple Azure Availability zones, similar to zone-redundant storage (ZRS), providing you continued read and write access even if a datacenter or availability zone is unavailable. In addition, GZRS asynchronously replicates your data to the secondary geo-pair region to protect against regional unavailability. RA-GZRS exposes a read endpoint on this secondary replica allowing you to read data in the event of primary region unavailability.

upvoted 15 times

✉  **aniagnesighile1** 1 year, 1 month ago

Again, based on the statement, 'Ensures that content remains available for writes if a primary data center fails', when the primary data center fails, you only have the secondary data center to work with. Now with RAGZRS, you only have the ability to read from the second region. You said it yourself 'I think it is read-only for the second region.'. Remember the primary data center is down, but the requirements states ENSURE THAT CONTENT REMAINS AVAILABLE FOR WRITES IF PRIMARY DATA CENTER FAILS. How are you going to write to the secondary datacenter.

upvoted 5 times

✉  **mamahani** 8 months, 3 weeks ago

aniagnesighile1 but you dont need the secondary region to write to in case of the fail of a data center; with ra-grs geo-zone redundant storage you still have 2 other zones with 2 other data centers that you can write to; so the requirement of being able to continue to write to is fullfilled; however with ra-grs you have also a secondary region, so in case of region fail you can failover to secondary region, and this gives you a higher resiliance than zrs only; so in my opinion the given answer is correct;

upvoted 6 times

✉  **AnonymousJhb** 8 months ago

the answer is wrong.
Microsoft recommends using ZRS in the primary region for Azure Data Lake Storage Gen2 Workloads.
With Microsoft failover.
upvoted 5 times

✉  **yogiazzaad** 11 months, 1 week ago

You are right.
upvoted 1 times

✉  **Sebastian1677** 1 year ago

Correct. ra-GZRS is the highest level that covered ZRS. When a DataCenter is downed, you still can write in another zone within the same region.
upvoted 6 times

✉  **suvec** 9 months ago

What about the highest level of data resiliency that cant be in ZRS answer is RA-GZRS
upvoted 5 times

✉  **chinomoreno** Highly Voted 1 year, 4 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 24 times

✉  **Gg2** 1 year, 4 months ago

RA-GZRS
Failover initiated by Microsoft.
upvoted 19 times

✉  **Metaalverf** 2 months, 2 weeks 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

✉  **BitacTeam** Most Recent 2 weeks, 5 days ago

according 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 answer is Customer managed

upvoted 1 times

 **ec255af** 1 month, 2 weeks ago

answer 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: initiated by customer. In Microsoft case they MAY failover only if greater disaster happens.

<https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance?toc=%2Fazur...%2Fstorage%2Fblobs%2Ftoc.json&bc=%2Fazur...%2Fstorage%2Fblobs%2Fbreadcrumb%2Ftoc.json#microsoft-managed-failover>

upvoted 1 times

 **Andrew_Chen** 2 months, 2 weeks ago

GZRS Because it mentioned that when a data center failed, not the entire region. When a data center failed, with ZRS, we can still read and write to the other two datacenters in the same region. So for the primary region, it must be ZRS. We still need a backup for the secondary region because it requires the highest degree of data resiliency.

Customer Managed Failover, No Doubt!!!

upvoted 1 times

 **ellala** 3 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

 **Chemmangat** 3 months, 3 weeks ago

Since it should be available for Writing in the secondary region, ZRS should be the one that we should opt for, since others only provide read access.

upvoted 1 times

 **kkk5566** 4 months, 2 weeks ago

<https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance#microsoft-managed-failover>
initiate by MS ,RAGZRS

upvoted 2 times

 **Saintu** 4 months, 3 weeks 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 sync 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 2 times

 **Ram9198** 5 months ago

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

With Microsoft failover - customer failover for hns is not yet supported

upvoted 2 times

 **Paulkuzzio** 6 months, 3 weeks ago

1. If your application requires resiliency, Microsoft recommends using geo-redundant storage. <https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance>

2. How an account failover works: <https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance>

upvoted 3 times

 **dp_learner** 7 months, 2 weeks ago

should be initiated by Microsoft.

==>

<https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance>

" Note

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.

In the event of a disaster that affects the primary region, Microsoft will manage the failover for accounts with a hierarchical namespace. For more information, see Microsoft-managed failover."

upvoted 3 times

 **janaki** 7 months, 3 weeks ago

I think failover is automatically initiated by azure automation job. customer do not have to direct control over initiating failover for GRS or RA-GRS storage

upvoted 1 times

 **rocky48** 8 months ago

It should be ZRS and Failover initiated by Microsoft.

upvoted 2 times

 **Rossana** 8 months, 3 weeks ago

ChatGPT: for the given requirements, it is recommended to use ZRS replication mechanism and an automatic failover initiated by Microsoft to provide the highest degree of data resiliency and ensure that the content remains available for writes if a primary data center fails.

upvoted 6 times

 **Honour** 8 months, 2 weeks ago

Lool.

Funny how I just used ChatGPT and the answer is different. I typed in all the questions and options and it recommended Read-Access Geo-Zone-Redundant Storage and Failover Initiated by Microsoft.

upvoted 1 times

 **frankanalysis** 9 months ago

A region contains multiple availability zones, and an availability zone contains multiple data centres. So if a data centre goes down, you still have other data centres in the same availability zone and same region to write to. RA-GZRS is the correct answer. If the question asked which solution is best when the primary REGION fails, then yes, ZRS would be the correct choice.

upvoted 2 times

 **Dhaval_Azure** 9 months, 3 weeks ago

Azure Storage supports account failover for geo-redundant storage accounts. With account failover, you can initiate the failover process for your storage account if the primary endpoint becomes unavailable. The failover updates the secondary endpoint to become the primary endpoint for your storage account. Once the failover is complete, clients can begin writing to the new primary endpoint.

<https://learn.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance?toc=%2Fazure%2Fstorage%2Fblobs%2Ftoc.json%20https%3A%2F%2Fdocs.microsoft.com%2Fen-us%2Fanswers%2Fquestions%2F32583%2Fazure-data-lake-gen2-disaster-recoverystorage-acco.html>

upvoted 1 times

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.

```
CREATE TABLE [DB0].[DimProduct] (
    [ProductKey] [int] IDENTITY(1,1) NOT NULL,
    [ProductSourceID] [int] NOT NULL,
    [ProductName] [nvarchar](100) NOT NULL,
    [Color] [nvarchar](15) NULL,
    [SellStartDate] [date] NOT NULL,
    [SellEndDate] [date] NULL,
    [RowInsertedDateTime] [datetime] NOT NULL,
    [RowUpdatedDateTime] [datetime] NOT NULL,
    [ETLAuditID] [int] NOT NULL
)
```

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.

[EffectiveEndDate] [datetime] NULL,

B.

[CurrentProductCategory] [nvarchar](100) NOT NULL,

C.

[ProductCategory] [nvarchar](100) NOT NULL,

D.

[EffectiveStartDate] [datetime] NOT NULL,

E.

[OriginalProductCategory] [nvarchar](100) NOT NULL,

Correct Answer: BE

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.

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

Reference:

<https://k21academy.com/microsoft-azure/azure-data-engineer-dp203-q-a-day-2-live-session-review/>

OldSchool Highly Voted 1 year, 1 month ago

B & E is correct answer

upvoted 14 times

erhard Highly Voted 1 year, 1 month ago

If BE is correct, then CE is also correct.

upvoted 7 times

 **hassexat** Most Recent 4 months ago

B & E is the correct answer

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

B & E is correct answer for SCD 3

upvoted 2 times

 **anks84** 1 year, 4 months ago

Answer is correct

upvoted 4 times

 **dom271219** 1 year, 4 months ago

Correct but SCD is always type 2. Type 3 is not SCD.

upvoted 1 times

 **gerrie1979** 1 year, 2 months ago

A SCD can be of type 0, 1, 2, 3 and so on, please read the documentation

upvoted 11 times

 **MuhammadK** 1 year, 4 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.

Name	Number of rows	Update frequency	Description
Common. Date	7,300	New rows inserted yearly	<ul style="list-style-type: none"> 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

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:

Table distribution types

- Hash
- Replicated
- Round-robin

Answer Area

Common.Data:

Marketing.Web.Sessions:

Staging. Web.Sessions:

Correct Answer:

Table distribution types

- Hash
- Replicated
- Round-robin

Answer Area

Common.Data:

Replicated

Marketing.Web.Sessions:

Hash

Staging. Web.Sessions:

Round-robin

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 1 year, 4 months ago

Replicated (Because its a Dimension table)
Hash (Fact table with High volume of data)
Round-Robin (Staging table)

upvoted 28 times

⊕  **kkk5566** Most Recent 4 months, 1 week ago

Replicated (Because its a Dimension table)
Hash (Fact table with High volume of data)
Round-Robin (Staging table)

upvoted 2 times

⊕  **spramanik_de** 6 months, 2 weeks 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

⊕  **wanchihh** 4 months, 3 weeks 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** 11 months, 1 week ago

Given answer is correct

upvoted 4 times

⊕  **DindaS** 11 months, 3 weeks ago

The dimension should be a replicated one. so that it will be available in all nodes for a better performance

fact table should be a HASH

staging table is always Round_ROBIN

upvoted 1 times

⊕  **nb1000** 1 year ago

correct

upvoted 1 times

⊕  **allagowf** 1 year, 3 months ago

the giving answer is correct, as the requirements
upvoted 4 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

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

Correct 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

Box 1: (CLUSTERED COLUMNSTORE INDEX)

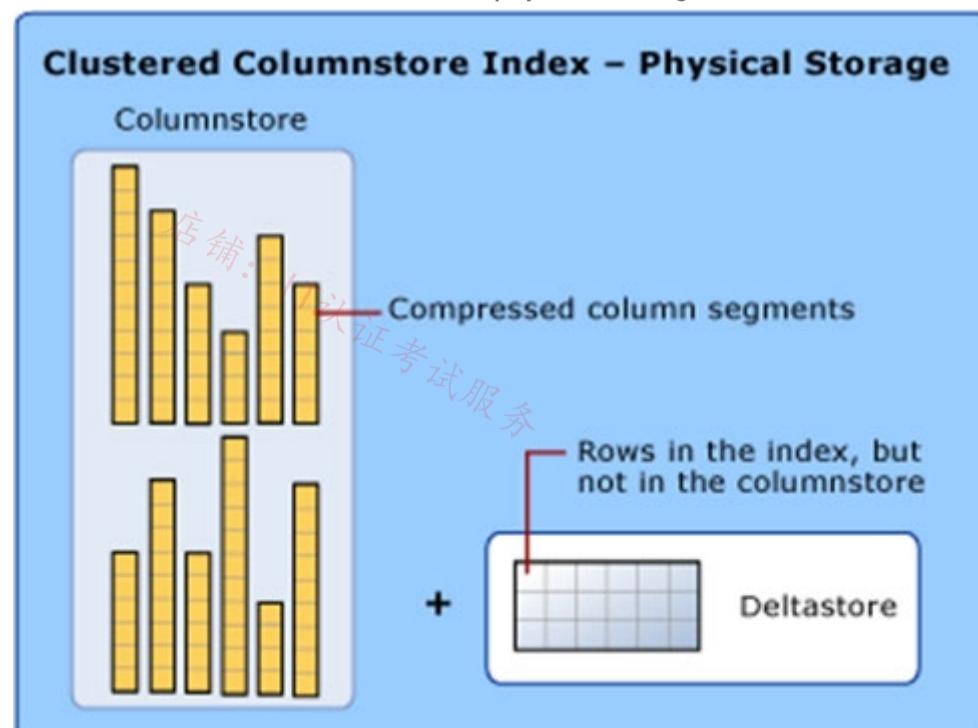
CLUSTERED COLUMNSTORE INDEX -

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 1 year, 2 months ago

you don't hash the date.. never..

upvoted 43 times

□ **Data_Analytics** 3 months ago

This sentence helped me soooo much - Hash the date NEVER.. thank you

upvoted 3 times

□ **Euanm28** 3 months ago

If you cant read don't bother commenting

upvoted 2 times

□ **Lestrang** Highly Voted 11 months, 2 weeks 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 9 times

□ **kkk5566** Most Recent 4 months, 1 week ago

Clustered columnstore

Hash(ProductKey)

upvoted 4 times

□ **smsme323** 1 year, 3 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 8 times

□ **Phund** 1 year, 4 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 7 times

□ **Lestrang** 1 year, 3 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** 1 year, 4 months ago

correct

upvoted 4 times

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

Correct 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

B (52%)	A (42%)	6%
---------	---------	----

✉  **vrodriguesp** Highly Voted  1 year 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 75 times

✉  **gggmaaster** 2 months, 1 week 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 1 times

✉  **yogiazzaad** 11 months, 1 week ago

I think you left out the fact that the table already has 1 billion records. This will change your calculations.

upvoted 6 times

✉  **CCCool77** 11 months, 1 week ago

You should consider the partition, not the table

upvoted 5 times

✉  **hiyoww** 5 months ago

agree, the question ask about partition, not the table distribution

upvoted 2 times

✉  **hiyoww** 5 months 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 5 months ago

sorry may be you are right
upvoted 1 times

Gcplearner8888 5 months ago

I have contributor access which is purchased for \$47.99 but no downloadable PDF with questions and explanations received yet.
upvoted 3 times

Paulkuzzio 3 months ago

Contact their Customer Care for answer with your complain and they will respond to you.
upvoted 1 times

vrodriguesp 11 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 1 times

anks84 Highly Voted 1 year, 4 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 10 times

sdg2844 Most Recent 5 days, 8 hours ago

Selected Answer: A

Did we ever come to a consensus on this? My thought would be once/month. You want to keep your records under 1 million, so the 60 million doesn't even come into it.

upvoted 1 times

d046bc0 4 weeks, 1 day ago

Once per month is correct
upvoted 1 times

Momoanwar 1 month ago

Chatpt say A

Considering the frequency of data loads and the typical access patterns, the best practice according to the Azure Synapse Analytics documentation would be:

A. **once per month**

This aligns with the monthly data load pattern and would facilitate efficient data management and query optimization.

upvoted 3 times

ellala 3 months ago

Selected Answer: B

If we partition per month, then each partition will have 30 million records. If each partition has 60 distributions, than this means each distribution has $30,000,000/60=500,000$ rows.

Now the key is in the Microsoft documentation sentence: " For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed." A MINIMUM OF 1 MILLION. If we partition by month, it will be half of the minimum. Therefore it cannot be a month. Ideally it would be 2 months, so that each partition had 1 million rows. However, that is not an option. Therefore correct option= PER YEAR

upvoted 4 times

Ram9198 4 months ago

Selected Answer: B

60 mill

upvoted 1 times

Ram9198 4 months ago

Selected Answer: A

Each partition 60million

upvoted 1 times

AvSUN 4 months ago

Once a month is correct, partition and distributions are not the same thing

upvoted 1 times

AvSUN 4 months ago

Can you delete this comment? I was wrong re read the docs

upvoted 1 times

kkk5566 4 months, 1 week ago

Selected Answer: A

A is correct
upvoted 2 times

 **akhil5432** 5 months ago

Can anyone help solving this question with proper reason
upvoted 1 times

 **Andrew_Chen** 5 months, 2 weeks ago

Selected Answer: B
Correct, only need to consider the newly added data here.
upvoted 1 times

 **Andrew_Chen** 5 months, 2 weeks ago

Selected Answer: B
Trick question, should consider that more than 1 million is better
upvoted 1 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: B
Provided answer is correct.
upvoted 2 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: A
The solution must optimize query performance and data loading. => Will receive about 30M records each month, means I could means single file arriving 1 time per month. Partitioning by month we have good performance on data loading and good performance in general, because probably somebody will check monthly data.
upvoted 1 times

 **janaki** 7 months, 3 weeks ago

per year will decrease the query performance as it will create larger partition size. per month will improve query performance , faster data loading and easier data management. So the answer is
Per month
As the question itself says that 30 million rows were added each month
upvoted 1 times

 **ajhak** 7 months, 2 weeks ago

Sir, you are incorrect.
upvoted 2 times

 **ajhak** 7 months, 3 weeks ago

Selected Answer: B
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>
upvoted 2 times

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

Correct 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

D (100%)

 labriji Highly Voted 12 months ago

Selected Answer: D

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 records. 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 8 times

 Naman1605 Highly Voted 1 year, 3 months ago

Selected Answer: D

correct D

upvoted 8 times

 **kkk5566** Most Recent ⓘ 4 months, 1 week ago

Selected Answer: D

To update or upsert records in a delta lake in Databricks use the "Merge" command.

upvoted 4 times

 **ajhak** 7 months, 2 weeks ago

Selected Answer: D

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** 1 year, 4 months ago

Correct

upvoted 3 times

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

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

Correct 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 `wide` (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%)

 **anks84** Highly Voted 1 year, 4 months ago

Correct as the requirement is Column based.

upvoted 7 times

 **ajhak** Highly Voted 7 months, 3 weeks ago

Selected Answer: D

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 6 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: D

D is correct

upvoted 1 times

 **Reloadedvn** 8 months ago

Selected Answer: D

Parquet format satisfies all reqs

upvoted 1 times

 **GodfreyMbizo** 11 months, 2 weeks ago

Parquet

upvoted 2 times

 **youngbug** 1 year ago

Selected Answer: D

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

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

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 ~~会~~ modify the files to ensure that each row is less than 1 MB.

Does this meet the goal?

A. Yes

B. No

Correct 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.

Load 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

A (75%)

B (25%)

□  **Tj87** Highly Voted 1 year, 4 months ago

I think we had this question in the previous pages and the correct answer was set as " compress the files"

upvoted 24 times

□  **semauni** 5 months, 2 weeks 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 2 times

□  **kim32** 8 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

□  **dom271219** 1 year, 4 months ago

Exactly
compress because a lot of row have more than 1MB length

upvoted 4 times

□  **Phund** Highly Voted 1 year, 4 months ago

Selected Answer: A

"ensure that each row is less than 1 MB" and the condition for polybase is <1M, whatever method you used

upvoted 15 times

□  **ExamDestroyer69** Most Recent 5 days, 8 hours ago

Selected Answer: A

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 2 times

✉ **hcq31818** 1 month, 3 weeks ago

Selected Answer: A

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,from%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** 4 months, 1 week ago

Selected Answer: A

A is correct
upvoted 1 times

✉ **auwia** 6 months, 2 weeks ago

Selected Answer: A

Yes, with less of 1Mb file we increase performance.
upvoted 2 times

✉ **e5019c6** 6 months, 2 weeks ago

i thought that polybase just query the tables and dont do any process of ETL or ELT.
upvoted 2 times

✉ **rocky48** 8 months ago

Selected Answer: A

"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

✉ **rocky48** 8 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 1 times

✉ **esaade** 10 months, 1 week ago

No, modifying the files to ensure that each row is less than 1 MB does not necessarily meet the goal of ensuring that the data copies quickly.

While it is true that large row sizes can impact data copy performance, simply reducing the row size to less than 1 MB may not be enough to optimize the data copy process. The performance of the data copy process can also be affected by factors such as network bandwidth, database design, and the method used to copy the data.

To ensure that the data copies quickly, you could consider other techniques such as compressing the data, using parallel data copy processes, and optimizing the database schema for efficient data loading.

Therefore, the correct answer is B. No.

upvoted 2 times

✉ **akk_1289** 11 months, 2 weeks ago

B. No

Modifying the files to ensure that each row is less than 1 MB may not be enough to ensure that the data copies quickly to Azure Synapse Analytics. Other factors such as network bandwidth, data compression, and parallel processing of data can also impact the speed of data transfer. To optimize data transfer, it may be necessary to implement data compression techniques, increase network bandwidth, or parallelize the data transfer process.

upvoted 2 times

✉ **panda_azzurro** 11 months, 2 weeks ago

Selected Answer: B

Question is very not clear.
upvoted 2 times

✉ **astone42** 11 months, 2 weeks ago

Selected Answer: A

It's A. The polybase has as condition <1M. I can't understand why people are confused.
It's extremely straightforward.

upvoted 2 times

 **sreekan2** 12 months ago

Selected Answer: B

Answer should be No
upvoted 1 times

 **Sima_al** 12 months ago

Selected Answer: B

Ensuring that each row is less than 1 MB will not necessarily ensure that the data copies quickly. While it is true that reducing the size of each row can help to improve the speed at which data is copied, there are other factors that can affect the performance of the data transfer. Some other things to consider would be the number of concurrent connections allowed, the network bandwidth and latency, and the file format used.

upvoted 2 times

 **VivekMadas** 1 year ago

<https://github.com/MicrosoftDocs/azure-docs/blob/main/articles/data-factory/v1/data-factory-azure-sql-data-warehouse-connector.md>
If you have source data with rows of size greater than 1 MB, you may want to split the source tables vertically into several small ones where the largest row size of each of them does not exceed the limit. The smaller tables can then be loaded using PolyBase and merged together in Azure Synapse Analytics.

Answer = Y (Modify the file to ensure that each row is less than 1 MB)

upvoted 3 times

 **temacc** 1 year, 1 month ago

1MB restriction is suitable for Copy command?

<https://learn.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse?tabs=data-factory#azure-sql-data-warehouse-as-sink>
upvoted 1 times

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 店铺: IT认证考试服务
- D. round-robin

Correct 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%)

□ **ank84** Highly Voted 1 year, 4 months ago

Given answer is correct !

Replicated because

- Dimension table
- Less than 2 GB (less than 1 GB in this case)

upvoted 9 times

□ **hcq31818** Most Recent 1 month, 3 weeks ago

Selected Answer: A

Replicated

upvoted 1 times

□ **kkk5566** 4 months, 1 week ago

Selected Answer: A

A is correct

upvoted 1 times

□ **Deeksha1234** 4 months, 4 weeks ago

Selected Answer: A

A is correct

upvoted 1 times

□ **akhil5432** 5 months ago

Selected Answer: A

replicated

upvoted 1 times

□ **mamahani** 8 months ago

Selected Answer: A

A is correct answer

upvoted 2 times

□ **GodfreyMbizo** 11 months, 2 weeks ago

answer is correct for Dimension tables and less than 2 GB microsoft recommends replicated tables

upvoted 2 times

□ **vigilante89** 1 year ago

Selected Answer: A

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** 1 year, 1 month ago

correct answer
upvoted 2 times

 **allagowf** 1 year, 3 months ago

Selected Answer: A
correct answer
upvoted 4 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

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

Correct 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 (94%) 6%

 **anks84** Highly Voted 1 year, 4 months ago

Selected Answer: C

Given answer is correct
upvoted 8 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: A

C is the answer.
upvoted 1 times

 **Deeksha1234** 4 months, 4 weeks ago

Selected Answer: C

C is correct
upvoted 2 times

 **akhil5432** 5 months ago

Selected Answer: C

Identity column
upvoted 2 times

 **nicky87654** 12 months ago

Selected Answer: C

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 4 times

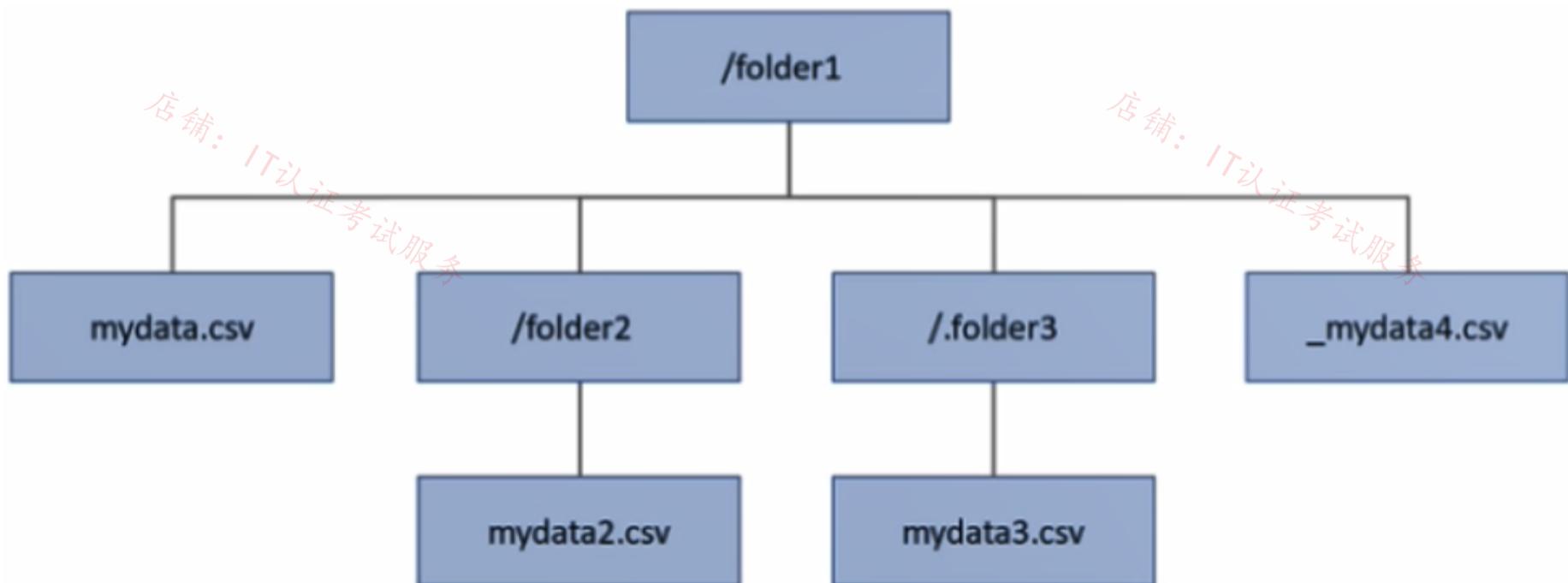
 **vigilante89** 1 year ago

Selected Answer: C

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 1 times

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.

```

CREATE EXTERNAL DATA SOURCE DataLake
WITH
(
    LOCATION      = 'https://mydatalake.dfs.core.windows.net/container1/folder1/**'
    , CREDENTIAL = DataLakeCred
);
  
```

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
When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.	<input type="radio"/>	<input type="radio"/>
When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	<input type="radio"/>	<input type="radio"/>
When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	<input type="radio"/>	<input type="radio"/>

Correct Answer:	Statements	Yes	No
	When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
	When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	<input type="checkbox"/>	<input checked="" type="checkbox"/>

shoottheduck Highly Voted 10 months, 3 weeks ago

I have just tested this on Synapse Serverless: ./Folder3 AND _mydata4.csv were ignored. Therefor; Yes, No, No
upvoted 65 times

MuruAzure 9 months, 1 week ago

its not ./Folder3 . it is ./Folder3 still ignored?
upvoted 5 times

□ **VikkiC** 7 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 15 times

□ **darkrajin** 3 months, 2 weeks 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

□ **AvSUN** 4 months ago

Thanks for the link

upvoted 1 times

□ **PGiagkoulas** Highly Voted 11 months, 4 weeks 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 excluded.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#arguments-create-external-table>

upvoted 38 times

□ **Paulkuzzio** 6 months, 3 weeks 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** 5 months, 2 weeks ago

But /** is specified at the end?

upvoted 4 times

□ **dakku987** 2 days ago

yes thats why it will return the all folders

upvoted 1 times

□ **matiandal** 7 months ago

ένα like για το nickname ;-)

upvoted 1 times

□ **blazy001** Most Recent 4 weeks, 1 day 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

□ **Lscrario** 1 month, 1 week 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** 3 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** 3 months, 3 weeks 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 3 times

□ **Ram9198** 4 months ago

Yes n n

upvoted 1 times

 **hassexat** 4 months ago

YES / NO / NO

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Yes, No, No

upvoted 1 times

 **UzairMir** 5 months, 3 weeks ago

The answer is Yes No No.

I found this resource

"COPY ignores hidden folders and doesn't return files that begin with an underline (_) or a period (.) unless explicitly specified in the path. This behavior is the same even when specifying a path with a wildcard."

<https://learn.microsoft.com/en-us/sql/t-sql/statements/copy-into-transact-sql?view=azure-sqldw-latest>.

upvoted 3 times

 **bakamon** 7 months, 3 weeks ago

Correct Answers :: 100000% True

Statement 1: Yes. The data from the mydata2.csv file will be returned when selecting all the rows in dbo.Table1 because the file is located in the folder structure specified by the external data source.

Statement 2: No. The data from the mydata3.csv file will not be returned when selecting all the rows in dbo.Table1 because folders that start with a dot are treated as hidden folders and are not included in wildcard searches.

Statement 3: No. The data from the _mydata4.csv file will not be returned when selecting all the rows in dbo.Table1 because files that start with an underscore are treated as hidden files and are not included in wildcard searches.

upvoted 14 times

 **rocky48** 8 months ago

1.Yes, 2.Yes, 3. No

upvoted 1 times

 **Victor_Kings** 8 months, 3 weeks ago

It's definitely Yes-No-No. According to Microsoft documentation, "It won't return mydata3.txt because it's a file in a hidden subfolder. And it won't return _hidden.txt because it's a hidden file.", and as we can see the folder is named "./hiddenfolder/", which means in this case the "./Folder3/" should be ignored too as it is hidden.

<https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=sql-server-ver16&tabs=dedicated>

upvoted 6 times

 **victorlie** 10 months, 2 weeks ago

Tricky question as Folder3 and mydata4.csv have (.) and (_) in the name. So they should be ignored. IMHO: Y N N

upvoted 10 times

 **hereiamken** 10 months, 3 weeks ago

Yes, Yes, No be correct

upvoted 1 times

 **bubby248** 11 months ago

All yes.

upvoted 3 times

 **UGOTCOOKIES** 11 months ago

Answer should be Yes Yes Yes.

From the learn documentation: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-folders-multiple-csv-files>
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.

upvoted 2 times

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

Correct Answer: B

Community vote distribution

B (100%)

✉ **nicky87654** Highly Voted 12 months ago

Selected Answer: B

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 15 times

✉ **AlviraTony** Highly Voted 4 months, 2 weeks 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 partition. But fourth one, requires you to use WHERE clause, so month is ideal for a partition

upvoted 12 times

✉ **ellala** 3 months ago

Thanks, this was helpful

upvoted 2 times

✉ **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: B

Should be B month.

upvoted 1 times

✉ **Deeksha1234** 4 months, 4 weeks ago

Selected Answer: B

B is right

upvoted 1 times

✉ **YikesYikes2023** 6 months, 1 week 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** 5 months, 2 weeks ago

Same. Every question refers to a different category, so why is month dominant over the others?

upvoted 4 times

 **Rossana** 8 months, 3 weeks 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** 8 months, 3 weeks ago

How can we partition by unique months unless we have the year too?

upvoted 2 times

 **hydmt07** 4 weeks, 1 day ago

If the Month column is the format YYYYMM, this won't be a problem.

upvoted 1 times

 **[Removed]** 11 months, 1 week ago

Selected Answer: B

Correct

upvoted 1 times

 **groozyn** 11 months, 2 weeks ago

Why B) month and not C) week?

upvoted 3 times

 **vrodriguesp** 11 months, 1 week ago

because it's doing aggregation (like the others answer A and D), instead partitions are powerful for where clause query

upvoted 4 times

 **[Removed]** 11 months, 1 week ago

Find all the orders for a given month.

Because of the above, monthly partitions are more efficient than weekly partitions.

If you have to read all the monthly data anyway, it is better to read one monthly partition than to read four to five weekly partitions.

upvoted 1 times

 **Dindas** 11 months, 3 weeks ago

B should be the correct answer here

upvoted 1 times

 **ZIMARAKI** 12 months ago

Selected Answer: B

Correct

upvoted 1 times

 **MrWood47** 12 months ago

Answer is correct

upvoted 1 times

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\FileDialog_YYYY_MM_DD.parquet
- B. \DataSource\SubjectArea\YYYY-WW\FileDialog_YYYY_MM_DD.parquet
- C. DataSource\SubjectArea\WW\YYYY\FileDialog_YYYY_MM_DD.parquet
- D. \YYYY\WW\DataSource\SubjectArea\FileDialog_YYYY_MM_DD.parquet
- E. WW\YYYY\SubjectArea\DataSource\FileDialog_YYYY_MM_DD.parquet

Correct Answer: A

Community vote distribution

A (86%)

14%

 nicky87654 Highly Voted 12 months ago

Selected Answer: A

A. \DataSource\SubjectArea\YYYY\WW\FileDialog_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 12 times

 ExamDestroyer69 Most Recent 3 weeks, 1 day ago

Selected Answer: B

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 4 months, 1 week 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 4 months, 1 week ago

for hot it , Option A can set permissions at the year level or at the week, should be A.

upvoted 2 times

 Deeksha1234 4 months, 4 weeks 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** 6 months, 2 weeks ago

Selected Answer: B

We need to query by week too, so better YYYY-WWW.

upvoted 1 times

□ **auwia** 6 months, 2 weeks 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** 5 months, 2 weeks ago

But why is it YYY-WW for you then instead of YYYY\WW?

upvoted 2 times

□ **VittalManikonda** 6 months, 3 weeks ago

Option B seems right answer as we can directly access the given yyyy-ww .

upvoted 2 times

□ **Rossana** 8 months, 3 weeks 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

□ **DP203Cert2023** 6 months, 4 weeks ago

I would not trust using Chat GPT for studying a certification...

upvoted 4 times

□ **Rob77** 7 months, 3 weeks ago

No, A it will give you additional separate column for week (WW).

upvoted 1 times

□ **auwia** 6 months, 2 weeks ago

"Most queries will include a filter on the current year OR week."

upvoted 1 times

□ **akshaynag95** 11 months, 1 week ago

Selected Answer: A

A is correct answer

upvoted 1 times

□ **vrodriguesp** 11 months, 1 week ago

Selected Answer: A

A is correct

upvoted 1 times

□ **DindaS** 11 months, 3 weeks ago

The answer A is correct

upvoted 1 times

□ **[Removed]** 12 months ago

Selected Answer: A

Given answer is correct

upvoted 1 times

□ **ZIMARAKI** 12 months ago

Selected Answer: A

Correct

upvoted 1 times

□ **MrWood47** 12 months ago

Selected Answer: A

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** 12 months ago

Selected Answer: A

My choice is A

upvoted 1 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

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

Correct Answer: D

Community vote distribution

D (79%) C (21%)

✉ **MrWood47** Highly Voted 12 months ago

Selected Answer: D

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 23 times

✉ **aemilka** Highly Voted 9 months, 2 weeks 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

✉ **OfficeSaracus** 8 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 5 times

✉ **aemilka** 8 months ago

Yes, 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 4 times

✉ **ukivanlamipi** Most Recent 2 months, 1 week ago

what is the different between "Alter Index Rebuild" or "DBCC DBREINDEX"

upvoted 1 times

✉ **kkk5566** 4 months, 1 week ago

Selected Answer: D

D. ALTER INDEX ALL on table1 REBUILD

upvoted 1 times

✉ **Deeksha1234** 4 months, 3 weeks ago

Selected Answer: D

D is correct

upvoted 1 times

✉ **vctrhugo** 6 months, 2 weeks ago

Selected Answer: D

ALTER INDEX REORGANIZE is used for rebuilding or reorganizing indexes, but it does not maximize columnstore compression.

upvoted 1 times

 **Matt2000** 4 months, 4 weeks 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** 7 months, 3 weeks ago

Reorganize is for row store indexes. The question here clearly mentions column store indexes. Correct answer is D

upvoted 2 times

 **MuhilMahil** 8 months ago

Selected Answer is C.
reorganizing only help in optimizing compression and performance.

upvoted 1 times

 **Vedjha** 11 months, 1 week 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

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.

```
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
```

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.

- 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

Correct Answer: BD

Community vote distribution

BE (72%)

BD (28%)

 aditya816 Highly Voted 9 months, 1 week ago

Selected Answer: BE

Surrogate is already there as customerkey column
upvoted 11 times

 laurasscastro 8 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 6 times

 **phydev** 2 months, 1 week 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 1 times

 **[Removed]**  8 months, 2 weeks 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 8 times

 **jiriz** 3 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 3 times

 **sdg2844**  5 days, 8 hours ago

Selected Answer: BE

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 1 times

 **jiriz** 3 months ago

Selected Answer: BE

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 2 times

 **hassexat** 4 months ago

Selected Answer: BE

B and E

upvoted 1 times

 **AvSUN** 4 months ago

B and D we need a unique row identifier

upvoted 2 times

 **kkk5566** 4 months, 2 weeks ago

B and D ,its a star schema on which has a fact table include a customerID property.

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

after think twice ,B&E

upvoted 1 times

 **Deeksha1234** 4 months, 3 weeks ago

B and D makes more sense, since inserted date is there already

upvoted 1 times

 **YikesYikes2023** 6 months, 1 week ago

Selected Answer: BE

If RowID was the surrogate, wouldn't it be an IDENTITY column? Therefore, it has to be B and E. Right? Please explain if this doesn't make sense make sense

upvoted 1 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: BE

<https://www.sqlshack.com/implementing-slowly-changing-dimensions-scds-in-data-warehouses/>

"For the SCD Type 2, we need to include three more attributes such as StartDate, EndDate and IsCurrent"
IsCurrentRow is already present! ... ;-)

CustomerKey (in reality is the RowID that many guys wants to add here),
effectiveEndDate will probably set to: 31.12.9999, (to justify the not null).

My final answer wil lbe : B and E.

upvoted 1 times

 **_ry_** 6 months, 2 weeks ago

what is the answer ?

upvoted 2 times

 **ArunMat** 7 months, 2 weeks ago

Selected Answer: BE

For SCD Type 2 we need record valid from and to date i.e effective date to identify latest row for that id.

upvoted 2 times

 **jlad26** 8 months, 1 week ago

I'm confused by the NOT NULL for the EffectiveEndDate. What value is this column going to hold for the row that holds the current information ?

upvoted 2 times

 **jlad26** 8 months, 1 week ago

OK seen elsewhere that typically would be e.g. Dec-31-9999

upvoted 1 times

 **AmrNegm** 9 months, 1 week ago

B and E

upvoted 2 times

 **wendyy** 9 months, 1 week ago

Should be BE

upvoted 2 times

 **SteveMcD** 9 months, 1 week ago

Selected Answer: BE

B and E. I don't think RowID is not needed, as there is already a surrogate key that exists with the CustomerKey column.

upvoted 2 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
```

```
)
```

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
)
```

Correct Answer:

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

)

FRANCIS_A_M Highly Voted 9 months, 1 week ago

Correct! 1. Hash(CustomerID) 2. Replicate
upvoted 14 times

zafnad 8 months ago

Could you please explain why 1. Hash([CustomerID]) is correct, and 2. Hash([OrderDate]) is incorrect.
upvoted 2 times

Spam_Account 6 months, 1 week ago

Don't hash on date, only partition on date
upvoted 10 times

vctrhugo 6 months, 2 weeks ago

Never distribute on Date.
upvoted 4 times

ajhak 7 months, 3 weeks 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 6 times

kkk5566 Most Recent 4 months, 1 week ago

1. Hash(CustomerID)
2. Replicate
upvoted 1 times

Deeksha1234 4 months, 3 weeks ago

given answer is correct

upvoted 1 times

✉  **examtopicsofyannick** 5 months, 1 week ago

Correct. Hash on Sales Table(Fact) and Replicate on Country table(Dimension)

upvoted 2 times

✉  **nmnm22** 9 months ago

correct

upvoted 4 times

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.

Correct Answer: A

Community vote distribution

A (100%)

✉️  **FRANCIS_A_M** Highly Voted 9 months, 1 week 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 25 times

✉️  **Rob77** 7 months, 3 weeks 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

✉️  **nmm22** 9 months ago

thanks a lot for the good explanation

upvoted 5 times

✉️  **vctrhugo** Highly Voted 6 months, 2 weeks ago

Selected Answer: A

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

✉️  **Ram9198** Most Recent 4 months ago

No support for storage acc key only ui, sas, sp, mi, apa

upvoted 1 times

✉️  **kkk5566** 4 months, 1 week ago

Selected Answer: A

is correct

upvoted 1 times

✉️  **Deeksha1234** 4 months, 3 weeks ago

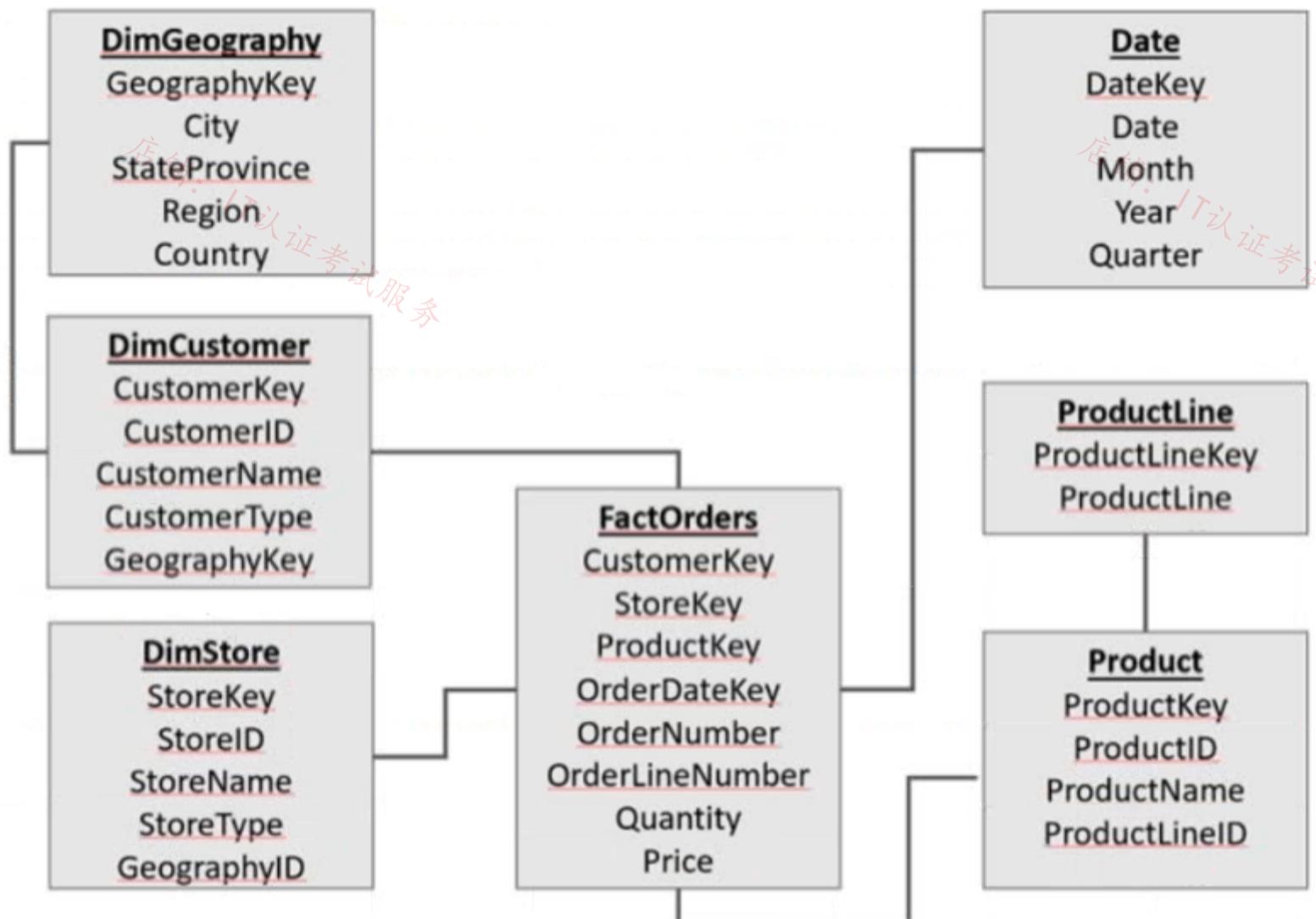
Selected Answer: A

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

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

Correct Answer:

Once the data model is converted into a star schema, there will be [answer choice] tables.

4
5
6
7

 **blenau** Highly Voted  9 months, 1 week 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 63 times

 **Dataminer** Highly Voted  8 months, 3 weeks ago

Agree with explanation. It will still be snowflake if Product and ProductLine is not combined

upvoted 16 times

 **sdg2844** Most Recent  5 days, 8 hours 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 1 times

 **surajpdh** 1 month, 2 weeks 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** 3 months ago

num of tables (dims + facts) == aka ==> 6

upvoted 4 times

 **ellala** 3 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 3 times

 **hassexat** 4 months ago

join DimGeography and DimCustomer

5 tables

upvoted 2 times

 **AvSUN** 4 months ago

shouldn't it be 5 tables?

upvoted 2 times

 **kkk5566** 4 months, 1 week ago

1 DimGeography and DimCustomer

2. 5 tables.

upvoted 2 times

 **ccesarrg** 4 months, 3 weeks 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** 4 months, 3 weeks ago

join DimGeography and DimCustomer and 5 tables

upvoted 1 times

 **Zak_Zakaria** 5 months, 3 weeks ago

ProductLine and Product also should be joined to switch to a star schema, if not we will be still on Snowflake Schema, so the remained tables should be 5, not 6.

upvoted 1 times

 **DataEngDP** 6 months ago

Customer is already joined with Geography (see the lines), the only thing needed is to combine it with Orders and ProductLine with Orders too, in order to convert this design to a star schema.

In this way we get 6 dimension tables plus the fact tables: Orders.

upvoted 1 times

 **Reloadedvn** 8 months ago

2. should be 5 tables

upvoted 6 times

 **rocky48** 8 months ago

DimGeography and DimCustomer and 5 tables.

upvoted 3 times

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

Correct Answer: C

Community vote distribution

C (100%)

✉  **Nikiboy**  9 months, 1 week 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 12 times

✉  **mr_examers** 8 months ago

Auto Loader does not support Azure Data Lake Storage Gen2

upvoted 1 times

✉  **vctrhugo** 6 months, 2 weeks 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 2 times

✉  **cloud_lady** 8 months ago

It does. Refer this link

<https://learn.microsoft.com/en-us/azure/databricks/ingestion/auto-loader/>

upvoted 1 times

✉  **ellala**  3 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** 4 months, 1 week ago

Selected Answer: C

Auto Loader

upvoted 1 times

 **Deeksha1234** 4 months, 3 weeks ago

Selected Answer: C

C is correct

upvoted 1 times

 **vctrhugo** 7 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** 8 months ago

Selected Answer: C

Auto Loader

upvoted 1 times

 **nicololmen** 8 months ago

D according to ChatGPT

upvoted 1 times

 **AHUI** 9 months, 1 week 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** 8 months, 4 weeks ago

Auto Loader is lower cost.

upvoted 1 times

You have an Azure subscription that contains the resources shown in the following table.

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

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?

- A. the WITH clause
- B. the ROWSET_OPTIONS bulk option
- C. the DATAFILETYPE bulk option
- D. the DATA_SOURCE parameter

Correct Answer: D

Community vote distribution

A (73%)

D (28%)

✉  **henryphchan**  8 months, 1 week ago

Selected Answer: A

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 10 times

✉  **19c1ee5**  8 months, 1 week ago

I think it's A. WITH CLAUSE

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-openrowset>

upvoted 7 times

✉  **jlad26** 8 months, 1 week 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 3 times

✉  **Momoanwar**  1 month ago

Selected Answer: A

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** 1 month, 2 weeks ago

Selected Answer: A

A - WITH Clause is the correct answer.

upvoted 1 times

✉  **Runaj** 2 months, 1 week ago

Selected Answer: D

D is right.

upvoted 1 times

✉  **jhargett1** 2 months, 1 week ago

Selected Answer: A

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 months, 2 weeks ago

Selected Answer: D

They ask for "in the function", not "in the query"

upvoted 3 times

 **ellala** 3 months ago

Selected Answer: A

Correct is A (explained in another comment)

upvoted 1 times

 **ellala** 3 months ago

It is not D because the files are PUBLICLY ACCESSIBLE in the storage account, therefore according to documentation:

"Any user can use OPENROWSET without DATA_SOURCE to read publicly available files on Azure storage." (<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-openrowset#security>)

Also DATA_SOURCE does not provide the necessary requirements, therefore WITH CLAUSE

Example:

```
OPENROWSET  
( { BULK 'unstructured_data_path' , [DATA_SOURCE = <data source name>, ]  
FORMAT= ['PARQUET' | 'DELTA'] }  
)  
[WITH ({'column_name' 'column_type'})]  
[AS] table_alias(column_alias,...n)  
upvoted 4 times
```

 **liwangbai123** 3 months, 1 week ago

Selected Answer: A

we can disable the header row in with clause

upvoted 1 times

 **gggqqqqq** 3 months, 1 week ago

with clause is not part of OPENROWSET function.

upvoted 2 times

 **HSZ** 4 months ago

Selected Answer: D

The correct answer is D. DATA_SOURCE parameter is the first parameter in OPENROWSET function.

upvoted 3 times

 **Metaalverf** 2 months, 2 weeks ago

Yes, they ask for "in the function", not "in the query"

upvoted 2 times

 **hassexat** 4 months ago

Selected Answer: A

WITH clause

upvoted 1 times

 **AvSUN** 4 months ago

I think D is the correct answer, it's mentioned that the TSV files have no headers

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: A

A is correct

upvoted 1 times

 **kkk5566** 4 months, 2 weeks ago

A is correct

upvoted 1 times

 **Deeksha1234** 4 months, 3 weeks ago

Selected Answer: A

answer is correct
upvoted 1 times

店铺: IT认证考试服务

店铺: IT认证考试服务

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

Correct Answer: A

Community vote distribution

D (81%)	Other
---------	-------

✉ **Ankit_Az** Highly Voted 7 months, 2 weeks 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 13 times

✉ **Vanq69** 3 months, 1 week ago

You mean the dedicated SQL pool has 60 distributions "by default"?

upvoted 2 times

✉ **Lscrario** 1 month ago

60 Million is correct

upvoted 1 times

✉ **58d2382** Most Recent 1 week, 4 days ago

Selected Answer: A

Question says "What is the minimum number of rows that Table1 should contain before you create (add/new/extr) 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 1 times

✉ **6d954df** 1 week, 6 days ago

60m, see <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition>

upvoted 1 times

✉ **blazy001** 4 weeks 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/extr) 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** 4 months ago

Selected Answer: D

60 million

upvoted 2 times

✉ **kkk5566** 4 months, 1 week ago

Selected Answer: D

is correct

upvoted 2 times

 **Deeksha1234** 4 months, 3 weeks ago

should be D
upvoted 3 times

 **akhil5432** 5 months ago

Selected Answer: C
WHY People mentioned option D..please explain how?
upvoted 2 times

 **akhil5432** 5 months ago

1 MILLION
upvoted 1 times

 **Lukis92** 6 months ago

Selected Answer: C
To achieve optimal data compression and query performance with clustered columnstore tables in Azure Synapse Analytics, it is recommended to have a minimum of 1 million rows per distribution and partition.
As Synapse Analytics automatically creates 60 distributions per table, to fulfill the 1 million rows per distribution recommendation, the table should ideally contain 60 million rows if no additional partitions are created.
However, the question is asking about the threshold for creating partitions, not necessarily a table of full 60 million rows. Therefore, you would want to ensure you have at least 1 million rows in each partition to maintain the optimal performance and compression. If the number of rows is less than 1 million, it's better to consider fewer partitions in order to increase the number of rows per partition.
upvoted 2 times

 **tankwayep** 4 months, 1 week ago

Question is about the minimum number of rows in the table not in a partition. And according to what you explained, which correct, the answer is D.
upvoted 4 times

 **janaki** 7 months, 3 weeks ago

answer shouls 60 million
The minimum number of rows that Table1 should contain before creating partitions in Azure Synapse Analytics dedicated SQL pool depends on various factors such as data size, query patterns, and performance requirements. However, a commonly recommended threshold is typically around 60 million rows before considering partitioning.
upvoted 3 times

 **maxstv** 7 months, 3 weeks ago

The minimum number of rows that Table1 should contain before you consider creating partitions in Azure Synapse Analytics dedicated SQL pool depends on multiple factors, such as the size of each row, the expected data growth rate, and the specific requirements of your workload.
However, considering the typical guidelines and best practices, a general rule of thumb is to consider creating partitions when the table size reaches around 60 million rows.

Therefore, the minimum number of rows that Table1 should contain before you create partitions is -

D. 60 million.

This is a commonly recommended threshold for optimizing data compression and query performance in large-scale data warehouses. However, it's important to note that this number can vary based on your specific scenario, so it's always advisable to conduct performance testing and consider the characteristics of your data and workload to determine the optimal time for partitioning your table.

upvoted 1 times

 **RamMovva** 7 months, 3 weeks ago

What is the minimum number of rows that Table1 should contain before you create partitions?

Answer : C
upvoted 1 times

 **ustefan11** 7 months, 3 weeks ago

Selected Answer: D
I've seen in the comments the explanation that this question has something to do with distribution and I don't think this is the case here. It's just that for a partition to have optimal compression, it has to be of at least 1 million rows, and since the idea of having a partition is to divide the data into smaller chunks, you need at least 2 partitions. Therefore, since there's no '2 mil' option, the only option left is '60M'.
upvoted 1 times

 **rocky48** 7 months, 4 weeks ago

Selected Answer: D
Clustered columnstore has the best compression with 1M rows. So it should be $1M * 60 = 60$ million rows
upvoted 2 times

 **mr_examers** 8 months ago

Selected Answer: D
For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Since Synapse Analytics divides each table into 60 distributions by default, the table should contain at least 1 million rows per distribution or 60 million rows in total before considering partitioning the table.
upvoted 4 times

 **jeroenmouse** 8 months, 1 week ago

Selected Answer: D

Hash-distributed tables work well for large fact tables in a star schema and along with that we need to use column store index for better compression and performance. By default it will have 60 distribution before partition and for better performance it is expected to have 1million rows per distribution.

upvoted 2 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

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.

Correct Answer: CD

Community vote distribution

CD (82%)	BC (18%)
----------	----------

 **Ankit_Az**  7 months, 2 weeks ago

Selected Answer: CD

CD is correct
upvoted 13 times

 **bakamon**  7 months, 3 weeks 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 11 times

 **bakamon** 7 months, 3 weeks ago

This will preserve the history of changes to the salesperson's last name while keeping the most current information in the table
upvoted 1 times

 **dakku987**  2 days ago

Selected Answer: CD

CD is correct as in scd2 we need startdate,enddate that is already present
what we need is "ISActive(flag)" and one more row that's all it takes to make scd2
upvoted 1 times

 **hassexat** 4 months ago

Selected Answer: CD

C & D are correct
upvoted 2 times

 **tankwayep** 4 months, 1 week ago

Selected Answer: CD

- Update one column: EndDate to the change date
 - Insert a new record with the new value of LastName, StartDate as the change date.
- upvoted 8 times

✉ **kkk5566** 4 months, 1 week ago

Selected Answer: CD

correct

upvoted 1 times

✉ **lfss** 4 months, 3 weeks ago

cd is correct

upvoted 1 times

✉ **Deeksha1234** 4 months, 3 weeks ago

Selected Answer: CD

answer should be CD, since activeRow flag is not present, we need to update only end date.

upvoted 6 times

✉ **Ram9198** 5 months ago

Selected Answer: CD

CD is correct

upvoted 1 times

✉ **Rob77** 7 months, 3 weeks 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

✉ **Vanq69** 3 months, 1 week 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 2 times

✉ **Rob77** 7 months, 3 weeks ago

* "original row"

upvoted 1 times

✉ **peches** 7 months, 2 weeks ago

but if you update the surname on the original row, don't you lose the previous value?

upvoted 6 times

✉ **laurasscastro** 8 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

✉ **ajhak** 7 months, 3 weeks 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** 8 months, 1 week ago

Selected Answer: CD

The answer is correct

upvoted 3 times

✉ **OfficeSaracus** 8 months, 1 week ago

Selected Answer: CD

Ans is correct

upvoted 3 times

✉ **jeroenmouse** 8 months, 1 week ago

Selected Answer: CD

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-scdfs-in-data-warehouses/>

upvoted 5 times

✉ **Yemeral** 8 months, 1 week ago

Selected Answer: CD

Correct. You need to insert a new row with the updated data and update the EndDate of the old row

upvoted 6 times

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.

Answer Area

For storage redundancy:

- Geo-zone-redundant storage (GZRS)
- Locally-redundant storage (LRS)
- Zone-redundant storage (ZRS)

For data deletion:

- A lifecycle management policy
- Soft delete
- Versioning

Answer Area

For storage redundancy:

- Geo-zone-redundant storage (GZRS)
- Locally-redundant storage (LRS)
- Zone-redundant storage (ZRS)

Correct Answer:

For data deletion:

- A lifecycle management policy
- Soft delete
- Versioning

 **bakamon** Highly Voted 7 months, 3 weeks 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 15 times

 **henryphchan** Highly Voted 8 months, 1 week 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 5 times

 **AvSUN** Most Recent 4 months ago

Correct

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

correct

upvoted 1 times

 **kkk5566** 4 months, 2 weeks ago

correct
upvoted 1 times

□ **Deeksha1234** 4 months, 3 weeks ago

correct answer
upvoted 2 times

□ **Rajan191083** 7 months, 3 weeks ago

Confusion is how the write access will be maintained in case primary zone failure?
upvoted 1 times

□ **Debasish93** 7 months, 3 weeks 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** 7 months, 4 weeks ago

Zone-redundant storage (ZRS) & Lifecycle Policy
upvoted 3 times

□ **makkelijkzat** 8 months, 1 week ago

correct
upvoted 2 times

□ **dksks** 8 months, 1 week 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

HR:

- Archive storage after one day and delete storage after 2,555 days.
- Archive storage after 2,555 days.
- Cool storage after one day.
- Cool storage after 180 days.
- Cool storage after 180 days and delete storage after 2,555 days.
- Delete after one day.
- Delete after 180 days.

Operations:

- Archive storage after one day and delete storage after 2,555 days.
- Archive storage after 2,555 days.
- Cool storage after one day.
- Cool storage after 180 days.
- Cool storage after 180 days and delete storage after 2,555 days.
- Delete after one day.
- Delete after 180 days.

Answer Area

HR:

- Archive storage after one day and delete storage after 2,555 days.
- Archive storage after 2,555 days.
- Cool storage after one day.
- Cool storage after 180 days.
- Cool storage after 180 days and delete storage after 2,555 days.
- Delete after one day.
- Delete after 180 days.

Correct Answer:

Operations:

- Archive storage after one day and delete storage after 2,555 days.
- Archive storage after 2,555 days.
- Cool storage after one day.
- Cool storage after 180 days.**
- ~~Cool storage after 180 days and delete storage after 2,555 days.~~
- Delete after one day.
- Delete after 180 days.

 **OfficeSaracus** Highly Voted 8 months, 1 week ago

The answer for HR depends on the meaning of "rarely" and the duration of "initial processing". If rarely is like once a year and initial processing is complete within 24 h the answer is correct. If rarely is like on a weekly basis, archiv might be the wrong way

upvoted 13 times

 **semauni** 5 months, 2 weeks ago

I agree, I also felt like I was missing information. In this case however, I'd say go for 'minimizing costs'. So the lowest cost option possible.

upvoted 3 times

 **dksks** Highly Voted 8 months, 1 week ago

correct

upvoted 12 times

 **AvSUN** Most Recent 4 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** 4 months, 1 week ago

correct

upvoted 1 times

 **Ankit_Az** 7 months, 2 weeks ago

Correct

upvoted 2 times

 **Rob77** 7 months, 3 weeks 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** 7 months, 4 weeks ago

Answer is correct

upvoted 3 times

 **henryphchan** 8 months, 1 week ago

the answer is correct

upvoted 4 times

 **shakes103** 8 months, 1 week ago

Answer is correct

upvoted 4 times

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:

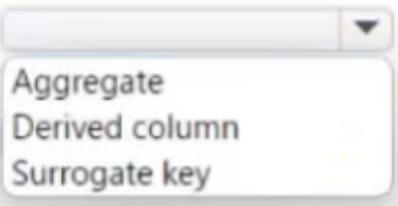
- Detect whether the data of a given customer has changed in the DimCustomer table.
- Perform an upsert to the DimCustomer table.

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

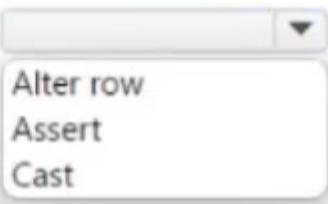
NOTE: Each correct selection is worth one point.

Answer Area

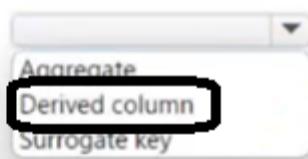
Detect whether the data of a given customer has changed in the DimCustomer table:



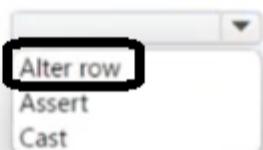
Perform an upsert to the DimCustomer table:

**Answer Area**

Detect whether the data of a given customer has changed in the DimCustomer table:

**Correct Answer:**

Perform an upsert to the DimCustomer table:



 **aemilka** Highly Voted 8 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 Upset 'Alter row' tranformation is used. The answer is definitely correct.

upvoted 17 times

 **kkk5566** Most Recent 4 months, 1 week ago

'Derived column' &'Alter row'

upvoted 1 times

 **Deeksha1234** 4 months, 3 weeks ago

answer is correct

upvoted 1 times

 **bakamon** 7 months, 3 weeks 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 4 times

✉ **rocky48** 7 months, 4 weeks ago

'Derived column' & 'Alter row'

upvoted 2 times

✉ **haythemsi** 8 months ago

surrogate key and assert

upvoted 1 times

✉ **nicololmen** 8 months ago

surrogate key and alter row according to chatgpt

upvoted 1 times

✉ **OfficeSaracus** 8 months, 1 week ago

It should be aggregate and alter row

As we talking Type 1 slowly changing dimension, we want to replace the current row with the updated one. This can be achieved by aggregate.

"A common use of the aggregate transformation is removing or identifying duplicate entries in source data. This process is known as deduplication. Based upon a set of group by keys, use a heuristic of your choosing to determine which duplicate row to keep. Common heuristics are first(), last(), max(), and min(). Use column patterns to apply the rule to every column except for the group by columns."

As in <https://learn.microsoft.com/en-us/azure/data-factory/data-flow-aggregate>

Alter row is correct:

<https://learn.microsoft.com/en-us/azure/data-factory/data-flow-alter-row#merges-and-upserts-with-azure-sql-database-and-azure-synapse>

upvoted 1 times

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

Values	Answer Area
BULK	SELECT TOP 100 *
DATA_SOURCE	FROM [] (
LOCATION	[] 'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',
OPENROWSET	FORMAT = 'CSV') AS rows

Correct Answer:	Answer Area
	SELECT TOP 100 *
	FROM OPENROWSET (
	BULK [] 'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',
	FORMAT = 'CSV') AS rows

 **rocky48** Highly Voted 7 months, 4 weeks 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/folder1/*.csv. The results are returned under the alias rows. Answer is correct.

upvoted 5 times

 **hassexat** Most Recent 4 months ago

OPENROWSET & Bulk

upvoted 4 times

 **kkk5566** 4 months, 1 week ago

openrowset..bulk
upvoted 2 times

 **Deeksha1234** 4 months, 3 weeks ago

ans is correct
upvoted 2 times

 **Ankit_Az** 7 months, 2 weeks ago

Correct
upvoted 2 times

 **henryphchan** 8 months, 1 week ago

The answer is correct
upvoted 3 times

 **OfficeSaracus** 8 months, 1 week ago

correct
upvoted 3 times

 **shakes103** 8 months, 1 week ago

Answer is correct
upvoted 3 times

店铺: IT认证考试服务

店铺: IT认证考试服务

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

Correct Answer: A

Community vote distribution

A (100%)

✉ **kam1122** 1 month ago

always pick parquet first
upvoted 2 times

✉ **kkk5566** 4 months, 1 week ago

Selected Answer: A

is correct
upvoted 1 times

✉ **Deeksha1234** 4 months, 3 weeks ago

Selected Answer: A

parquet. CSV , delta also possible but not an option here.
upvoted 2 times

✉ **akhil5432** 5 months ago

Selected Answer: A

parquet
upvoted 1 times

✉ **Ankit_Az** 7 months, 2 weeks ago

Selected Answer: A

Correct
upvoted 1 times

✉ **rocky48** 7 months, 4 weeks ago

Selected Answer: A

Parquet is the correct answer
upvoted 1 times

✉ **aemilka** 8 months ago

Selected Answer: A

<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 1 times

✉ **henryphchan** 8 months, 1 week ago

Selected Answer: A

Parquet is supported by serverless SQL pool
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-parquet-files>
upvoted 2 times

 **makkelijkzat** 8 months, 1 week ago

Selected Answer: A

correct

upvoted 1 times

 **makkelijkzat** 8 months, 1 week ago

correct

upvoted 1 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

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

Correct Answer: AD

Community vote distribution

AD (100%)

✉  **orionduo** Highly Voted 6 months, 3 weeks 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 9 times

✉  **vctrhugo** Highly Voted 7 months ago

Selected Answer: AD

Query acceleration supports CSV and JSON formatted data as input to each request.

upvoted 6 times

✉  **moize** Most Recent 3 days, 14 hours ago

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-query-acceleration>

upvoted 1 times

✉  **OldSchool** 3 months, 2 weeks ago

Selected Answer: AD

Query acceleration supports CSV and JSON formatted data as input to each request.

upvoted 1 times

✉  **kkk5566** 4 months, 1 week ago

Selected Answer: AD

CSV and JSON

upvoted 1 times

✉  **aga444** 7 months ago

Parquet and CSV

upvoted 1 times

✉  **IanKwok81** 7 months ago

Correct. <https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-query-acceleration>

upvoted 4 times

You have an Azure subscription that contains the resources shown in the following table.

Name	Type	Description
storage1	Azure Blob storage account	Contains publicly accessible JSON files
WS1	Azure Synapse Analytics workspace	Contains a serverless SQL pool

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?

- A. JSON
- B. DELTA
- C. PARQUET
- D. CSV

Correct Answer: A

Community vote distribution

D (97%)

✉ **phydev** Highly Voted 2 months, 1 week ago

Selected Answer: D

Was on my exam today (31.10.2023).
upvoted 8 times

✉ **kkk5566** Highly Voted 4 months, 1 week ago

Selected Answer: D

no json format, using CSV
upvoted 6 times

✉ **MJamesP** Most Recent 3 months, 2 weeks ago

Selected Answer: D

Please refer: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files#read-json-documents>
upvoted 2 times

✉ **susbhat** 4 months, 2 weeks ago

Selected Answer: D

Ignore my previous comment.
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files#read-json-files>
upvoted 4 times

✉ **susbhat** 4 months, 2 weeks ago

Selected Answer: A

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files#read-json-files>
upvoted 1 times

✉ **kdp203** 4 months, 2 weeks ago

CSV -> D is the correct answer
upvoted 2 times

✉ **lfss** 4 months, 3 weeks ago

D is the correct
upvoted 1 times

✉ **mmoayed** 4 months, 3 weeks ago

If most of all answered D, but the system says A, Who should I take then ?
upvoted 1 times

✉ **susbhat** 4 months, 2 weeks ago

There is no filetype as JSON. We have to use CSV and then query:
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files#read-json-files>
upvoted 4 times

□ **Deeksha1234** 4 months, 3 weeks ago

Selected Answer: D

D is correct

upvoted 1 times

□ **Andrew_Chen** 5 months, 2 weeks ago

Selected Answer: D

Exactly! OPENROWSET has no JSON files, so use FORMAT = 'csv' for the querying.

upvoted 1 times

□ **Zak_Zakaria** 5 months, 3 weeks 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 5 times

□ **auwia** 6 months, 2 weeks ago

Selected Answer: D

It should be D.

upvoted 1 times

□ **andjurovicela** 6 months, 4 weeks ago

Selected Answer: D

D is the correct answer, indeed: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files>

upvoted 4 times

□ **T4321** 7 months ago

Selected Answer: D

Correct answer is 'D'

upvoted 2 times

□ **vctrhugo** 7 months ago

Selected Answer: D

The easiest way to see to the content of your JSON file is to specify csv FORMAT.

upvoted 2 times

□ **abdallaissa** 7 months ago

Selected Answer: D

No Format for JSON to achieve this CSV is the right answer

upvoted 1 times

□ **abdallaissa** 7 months ago

To Read Json the format should be CSV

upvoted 1 times

HOTSPOT

You have an Azure subscription that contains the Azure Synapse Analytics workspaces shown in the following table.

Name	Primary storage account
workspace1	datalake1
workspace2	datalake2
workspace3	datalake1

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.

Answer Area

Statements	Yes	No
The shared catalog objects can be stored in Azure Database for MySQL.	<input type="radio"/>	<input type="radio"/>
For the Apache Hive Metastore of each workspace, you must configure a linked service that uses user-password authentication.	<input type="radio"/>	<input type="radio"/>
The users of workspace1 must be assigned the Storage Blob Contributor role for datalake1.	<input type="radio"/>	<input type="radio"/>

Answer Area

Statements	Yes	No
Correct Answer: The shared catalog objects can be stored in Azure Database for MySQL.	<input checked="" type="checkbox"/>	<input type="radio"/>
For the Apache Hive Metastore of each workspace, you must configure a linked service that uses user-password authentication.	<input checked="" type="checkbox"/>	<input type="radio"/>
The users of workspace1 must be assigned the Storage Blob Contributor role for datalake1.	<input type="radio"/>	<input checked="" type="checkbox"/>

 **auwia** Highly Voted  6 months, 2 weeks 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 11 times

 **aga444** Highly Voted  7 months ago

No-Yes-Yes

upvoted 11 times

□ **auwia** 6 months, 3 weeks 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** 6 months, 2 weeks ago

Sorry I was wrong.

upvoted 6 times

□ **ExamDestroyer69** [Most Recent] 2 weeks, 6 days ago

Confusing discussion section, no consensus

upvoted 2 times

□ **kkk5566** 4 months, 1 week ago

correct

upvoted 1 times

□ **ahmadsayeed** 2 months, 1 week ago

Shouldn't WS1 have blob data contributor access?

upvoted 1 times

□ **Deeksha1234** 4 months, 3 weeks ago

given answer is correct

<https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore>

upvoted 2 times

□ **Ram9198** 5 months ago

Yes , yes , no <https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore>

upvoted 1 times

□ **pavankr** 6 months, 1 week ago

Correct order should be Yes, No, Yes

upvoted 2 times

□ **DataSaM** 6 months, 1 week ago

<https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore>

upvoted 1 times

□ **Albeeliu** 6 months, 1 week ago

What are the correct answers???

upvoted 1 times

□ **Ram9198** 5 months ago

Yes, Yes , No you can check the document link

upvoted 2 times

□ **Paulkuzzio** 6 months, 3 weeks 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 5 times

□ **sridat** 7 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

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.

SCD Type Answer Area

Type 0	ProductName: <input type="text"/>
Type 1	Color: <input type="text"/>
Type 2	Size: <input type="text"/>
Type 3	

Answer Area

Correct Answer:	ProductName: <input checked="" type="text"/> Type 0
	Color: <input checked="" type="text"/> Type 1
	Size: <input checked="" type="text"/> Type 2

 **Ram0202** Highly Voted  7 months ago

Product name -type 0
color -type 2
size -type 3
upvoted 59 times

 **andjurovicela** Highly Voted  6 months, 4 weeks 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 21 times

 **hiyoww** 5 months, 3 weeks 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 14 times

 **Momoanwar** Most Recent  1 month 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 3 times

□ **Vanq69** 3 months, 1 week 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** 3 months, 1 week ago

I meant Name: 0
Color: 3
Size: 2
upvoted 4 times

□ **kkk5566** 4 months ago

SCD0-1-2
upvoted 1 times

□ **hassexat** 4 months ago

Product Name - Type 0
Color - Type 2
Size - Type 3
upvoted 2 times

□ **AvSUN** 4 months ago

The answers are 0, 3, 2
upvoted 2 times

□ **AvSUN** 4 months ago

ProductName - 0
ProductColor - 3
ProductSize - 2
upvoted 2 times

□ **Lucasmh** 1 month, 1 week 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

□ **AvSUN** 4 months ago

The requirements and blanks are out of order in the question
upvoted 2 times

□ **Deeksha1234** 4 months, 3 weeks ago

correct answer is - type 0, type 3,type 2
upvoted 4 times

□ **akhil5432** 5 months ago

type 0
type 3
type 2
upvoted 2 times

□ **ravigolu** 6 months, 1 week 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** 6 months, 3 weeks ago

Answer are corrects as reported in the solution: 0, 1, and 2.

upvoted 2 times

□ **auwia** 6 months, 3 weeks 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, means all the history)
- Retain all the current and previous values in ProductColor. => TYPE 3 (it includes a column for the previous value)

upvoted 3 times

□ **auwia** 6 months, 3 weeks 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

□ **auwia** 6 months, 3 weeks ago

- Retain only the current and the last values in ProductSize. => TYPE 1 (current and last, DOESN'T mean 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

□ **HimaC5991** 6 months, 3 weeks ago

my answer is 0,2,3

upvoted 4 times

□ **sridat** 7 months ago

ProductSize is Type 2 since it maintains current and last record. Type 1 can only have current value.

upvoted 2 times

□ **wendy** 7 months ago

Retain ONLY the current and the last values in ProductSize. type2 will include all changes. type 3 is correct.

upvoted 5 times

□ **mehroosali** 7 months ago

Correct answer is type 0, type 2, type 3

upvoted 2 times

□ **abdallaissa** 7 months ago

Correct!

upvoted 3 times

□ **IanKwok81** 7 months ago

Should be ProductName Type0, ProductColor Type2, ProductSize Type1

upvoted 2 times

□ **IanKwok81** 7 months ago

ProductSize Type3

upvoted 2 times

□ **abdallaissa** 7 months ago

Color need all the the values which mean you need all the rows for it, in type 1 any change happened a new row will be added

upvoted 1 times

□ **RoyP654** 7 months ago

ProductSize =/Type3? retain only the current and last value on the same row

upvoted 2 times

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

Table distribution:

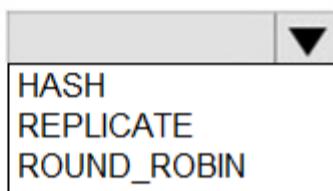


Table structure:

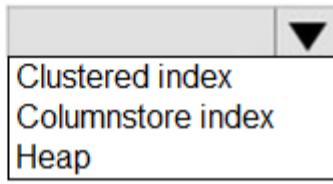
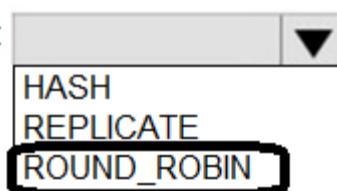
**Answer Area**

Table distribution:



Correct Answer:

Table structure:



orionduo Highly Voted 6 months, 3 weeks 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 16 times

kkk5566 Most Recent 4 months, 1 week ago

correct

upvoted 1 times

kkk5566 4 months, 1 week ago

For the staging data

upvoted 1 times

Deeksha1234 4 months, 3 weeks ago

ans is correct

upvoted 2 times

mehroosali 7 months ago

correct

upvoted 3 times

 **abdallaissa** 7 months ago

Correct

upvoted 2 times

 **IanKwok81** 7 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.

Correct Answer: ACD

Community vote distribution

ACD (100%)

 **pramod4lk** Highly Voted 5 months, 1 week ago

The answer is correct. We need to enable an analytical store in container1.

upvoted 5 times

 **AvSUN** Most Recent 4 months ago

Correct

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: ACD

correct

upvoted 1 times

 **gusztimm** 4 months, 2 weeks ago

Selected Answer: ACD

Correct

upvoted 1 times

 **Deeksha1234** 4 months, 3 weeks ago

The answer is correct

upvoted 2 times

HOTSPOT

You have an Azure subscription that contains the resources shown in the following table.

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

The storage1 account contains a container named container1. The container1 container contains the following files.

```
Webdata <root folder>
  Monthly <folder>
    _monthly.csv
    Monthly.csv
  .testdata.csv
  testdata.csv
```

In Pool1, you run the following script.

```
CREATE EXTERNAL DATA SOURCE Ds1
WITH
  ( LOCATION = 'abfss://container1@storage1.dfs.core.windows.net' ,
  CREDENTIAL = credential1,
  TYPE = HADOOP
) ;
```

In the Built-in serverless SQL pool, you run the following script.

```
CREATE EXTERNAL DATA SOURCE Ds2
WITH (
  LOCATION = 'https://storage1.blob.core.windows.net/container1/Webdata/',
  CREDENTIAL = credential2
);
```

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
An external table that uses Ds1 can read the _monthly.csv file.	<input type="radio"/>	<input checked="" type="radio"/>
An external table that uses Ds1 can read the Monthly.csv file.	<input checked="" type="radio"/>	<input type="radio"/>
An external table that uses Ds2 can read the .testdata.csv file.	<input type="radio"/>	<input checked="" type="radio"/>

Answer Area**Statements**

An external table that uses Ds1 can read the _monthly.csv file.

Yes**No****Correct Answer:**

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 5 months, 1 week ago

The answer is No, Yes, No

It will ignore "_" and "."

upvoted 25 times

pc1337xd Highly Voted 5 months, 1 week 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 7 times

hcq31818 Most Recent 1 month ago

No, Yes, No

upvoted 1 times

AvSUN 4 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 4 times

kkk5566 4 months, 1 week ago

NO

YES

NO

see previous quizzes.

upvoted 4 times

subhraz 4 months, 1 week ago

NO

YES

NO

upvoted 3 times

g2000 5 months, 1 week ago

The last one is No. File is prefixed 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

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.

Permissions

Execute	None
Read	Read and Execute
Read and Write	Write

Answer Area

container1/:

container1/folder1/:

Drag the permissions from the left pane to the correct folder in the right pane.

Answer Area**Correct Answer:**

container1/: Execute

container1/folder1/: Read and Execute

 **g2000** Highly Voted 5 months, 1 week ago

correct!

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control#levels-of-permission>

upvoted 16 times

 **TuxBingo** Most Recent 3 days, 11 hours ago

Correct

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

correct

upvoted 2 times

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

Correct Answer: A

Community vote distribution

A (81%) D (19%)

✉ **jonpert** 1 week, 6 days ago

Selected Answer: A

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** 3 weeks, 1 day 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 months ago

Selected Answer: D

Looks to me like Tumbling
upvoted 3 times

✉ **madhubanti123** 3 months ago

Selected Answer: A

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** 3 months, 1 week 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

✉ **ruggerofreddi** 3 months, 2 weeks ago

I would say tumbling with a max concurrency = 1. in this way if the first run doesn't stop the second one will not start
upvoted 4 times

✉ **Ram9198** 4 months ago

Selected Answer: A

Schedule
upvoted 2 times

✉ **hassexat** 4 months ago

Tumbling
upvoted 1 times

kkk5566 4 months, 1 week ago

Selected Answer: A

A is correct

upvoted 1 times

yassine70 4 months, 2 weeks 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 4 times

p_mks 4 months, 1 week ago

I don't see 'Retry' as a requirement, they mentioned only about the next pipeline execution. 'A' seems to be more appropriate.

upvoted 3 times

Biswada 4 months, 4 weeks ago

Tumbling

upvoted 3 times

Matt2000 4 months, 4 weeks ago

Sounds like a tumbling trigger with self-dependency: "if the trigger for pipeline1 stops, the next pipeline execution will occur at 2 AM, following a restart of the trigger", implicating that the trigger does not start if the previous run of the trigger is not yet completed.

By using a tumbling trigger with self-dependency, one can let a trigger only start if a previous run of the same trigger has completed. To achieve that maxConcurrency has to be set to '1'.

Ref: <https://learn.microsoft.com/en-us/azure/data-factory/tumbling-window-trigger-dependency>

upvoted 2 times

akhil5432 5 months ago

Selected Answer: A

Schedule

upvoted 3 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.

Aggregate settings Optimize Inspect Data preview

Output stream name * Learn more [\[\]](#)

Incoming stream *

[Group by](#) **Aggregates**

Grouped by: ProductID

+ Add [Clone](#) [Delete](#) [Open expression builder](#)

<input type="checkbox"/> Column	Expression
<input type="checkbox"/> Each column that matches <input type="text" value="name != 'ProductID'"/>	<input type="text" value="creates 1 column(s)"/> []
<input type="text" value="\$\$"/>	<input type="text" value="abc"/> []
	<input type="text" value="first(\$\$)"/> []
	ANY []

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

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.

Answer Area**Statements**

There will be six columns in the output of ProductRows.

Correct Answer:

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.

 **Ram9198** Highly Voted 5 months ago

Yes , no, yes - <https://learn.microsoft.com/en-us/azure/data-factory/data-flow-aggregate>
upvoted 17 times

 **jongert** Most Recent 1 week, 6 days 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** 1 month 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 1 times

 **Tincox** 1 month, 1 week ago

The syntax mentions `name!=` which means "not equal to" ProductID, so my answer would be: yes, yes, no
upvoted 1 times

 **Tincox** 3 weeks, 2 days 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 1 times

 **OldSchool** 3 months, 2 weeks 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 3 times

 **Vaq69** 3 months, 1 week 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 2 times

 **kkk5566** 4 months, 1 week ago

yes, no, yes
upvoted 4 times

 **kkk5566** 4 months, 2 weeks ago

Yes , no, yes
upvoted 3 times

 **mmoayed** 4 months, 3 weeks ago

I have notice that some answers might be wrong. What does this mean? who is confirming the correct answers ?
upvoted 4 times

 **mmoayed** 4 months, 3 weeks ago

yes, no, yes
upvoted 2 times

 **MSExpert** 4 months, 3 weeks ago

Yes No Yes
upvoted 2 times

 **DataEngDP** 4 months, 4 weeks ago

yes, no, yes
upvoted 3 times

 **ClydeZ** 5 months ago

6 columns from product + the aggregated measurement = 7 columsn in total, so answer is NO?
upvoted 1 times

 **Elxaxe** 4 months, 3 weeks ago

The irs no new aggregated measurement. You're grouping by ProductID, so that makes one column. All the other columns that are not 'ProductID' are grouped by choosing their first row. So you will obtain the same number of columns, like a `SELECT DISTINCT`.
upvoted 5 times

 **Heringer** 5 months ago

How can one know that the answer to the third question is either yes or no? In my understanding, you'd have to assume that there is no duplicates in the source table, i.e. there are no rows that share the same values in all columns except for productid upvoted 4 times

店铺: IT认证考试服务

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

Correct Answer: B

Community vote distribution

B (100%)

 **MSExpert** Highly Voted 5 months ago

Correct

upvoted 5 times

 **ndangalasi** Most Recent 2 months, 3 weeks ago

Selected Answer: B

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monitor-cache>
upvoted 1 times

 **Vanq69** 3 months, 1 week ago

B. Cache hit percentage should be correct since it only affects common used queries, which should be saved and loaded from cache.
upvoted 2 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: B

B is correct
upvoted 2 times

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

```
SELECT * FROM  
OPENROWSET(  
    [ ] N'https://myaccount.dfs.core.windows.net/mycontainer/mysubfolder/data.parquet', FORMAT = 'PARQUET')  
    BULK  
    DELTA  
    OPENQUERY  
    SINGLE_BLOB  
  
WITH [ ] as rows  
(Col1 int, Col2 varchar(20))  
FILEPATH(2)  
PARSER_VERSION = '2.0'  
SINGLE_BLOB
```

Answer Area

```
SELECT * FROM  
OPENROWSET(  
    [ ] N'https://myaccount.dfs.core.windows.net/mycontainer/mysubfolder/data.parquet', FORMAT = 'PARQUET')  
    BULK  
    DELTA  
    OPENQUERY  
    SINGLE_BLOB  
  
WITH [ ] as rows  
(Col1 int, Col2 varchar(20))  
FILEPATH(2)  
PARSER_VERSION = '2.0'  
SINGLE_BLOB
```

Correct Answer:

g2000 Highly Voted 5 months, 1 week ago

correct!

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-openrowset#data-source>

upvoted 12 times

kkk5566 Most Recent 4 months, 1 week ago

```
SELECT *  
FROM OPENROWSET(BULK 'http://<storage account>.dfs.core.windows.net/container/folder/*.parquet',  
FORMAT = 'PARQUET') AS [file]
```

upvoted 2 times

pramod4lk 5 months, 1 week ago

Correct, Serverless SQL pool uses BULK.

upvoted 4 times

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

Correct Answer: D

Community vote distribution

D (100%)

✉  **arihant_jain** 1 month, 2 weeks ago

Everything is mentioned here:

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop>

upvoted 2 times

✉  **ExamDestroyer69** 2 weeks, 6 days ago

Quote from link provided suggesting D is correct

"use the partitioned views instead of the external tables."

upvoted 1 times

✉  **metiii** 2 months ago

Selected Answer: D

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 2 times

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

Correct Answer: A

Community vote distribution

A (75%) D (25%)

 **Blablatest123** Highly Voted 2 months, 1 week ago

Partitioned by month and with 60 nodes, means it's 1M per combination
upvoted 8 times

 **d046bc0** Most Recent 4 weeks ago

Selected Answer: A

60 nodes so 1M per distribution and partition
upvoted 1 times

 **BitacTeam** 1 month, 1 week ago

Selected Answer: A

1 Mio per combination
upvoted 1 times

 **SimonQBDS** 1 month, 1 week ago

Selected Answer: A

same as Blablatest123 said
upvoted 1 times

 **mishoka23** 2 months, 1 week ago

Selected Answer: D

Duplicate Question
upvoted 1 times

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

Correct Answer: AE

Community vote distribution

AE (100%)

 **metiii** Highly Voted 2 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 5 times

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.

```
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
);
```

Which three columns should you add? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- 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

Correct Answer: CEF

Community vote distribution

CEF (100%)

 **dakku987** 2 weeks, 4 days ago

Selected Answer: CEF

CEF IS CORRECT

upvoted 4 times

 **d046bc0** 4 weeks ago

Selected Answer: CEF

correct

upvoted 3 times

 **Lucasmh** 1 month, 1 week 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 2 times

 **Kapello10** 4 weeks 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 3 times

 **metiii** 2 months ago

Selected Answer: CEF

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.

Answer Area

Partition the data:

- Partition by date with one partition per day.
- Partition by date with one partition per month.
- Partition by product.

Correct Answer:

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.

 **jongert** 4 days, 1 hour 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 2 times

You have an Azure subscription that contains an Azure Synapse Analytics serverless SQL pool.

You execute the following query.

```
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';
```

Where will the rows returned by the query be stored?

- 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

Correct Answer: A

 **jonpert** 4 days, 1 hour 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 1 times

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 店铺: IT认证考试服务
- D. facts and dimensions

Correct Answer: B

Community vote distribution

D (100%)

 **jongert** 4 days, 1 hour ago

Correct

<https://learn.microsoft.com/en-us/azure/synapse-analytics/database-designer/overview-database-templates>
upvoted 2 times

 **mrplmcc** 5 days, 18 hours ago

Selected Answer: D

Why? Shouldn't it be D?

upvoted 1 times

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.

Correct Answer: B

✉  **jongert** 4 days, 1 hour 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 1 times

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

Correct Answer: A

 **jongert** 4 days, 1 hour 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 2 times

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

Correct Answer: B

Community vote distribution

A (80%)

B (20%)

 **dakku987** 1 day, 22 hours 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 1 times

 **jongert** 4 days ago

Selected Answer: A

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 2 times

 **mrplmcc** 5 days, 18 hours ago

Selected Answer: A

Yes it should work

upvoted 2 times

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

Correct Answer: B

Community vote distribution

A (100%)

 **dakku987** 1 day, 22 hours ago

Selected Answer: A

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 1 times

 **mrplmcc** 5 days, 18 hours 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 3 times

 **jongert** 4 days ago

Agree, also the MS documentation for it.

<https://learn.microsoft.com/en-us/azure/data-factory/continuous-integration-delivery-manual-promotion>

upvoted 2 times

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

```
new_rows_df.write.  
    format("csv").  
    format("delta").  
    format("json").  
    format("parquet").  
        mode("append")  
        mode("error")  
        mode("ignore")  
        mode("overwrite")  
    .save(delta_table_path)
```

Answer Area

Correct Answer: new_rows_df.write.
 format("csv").
 format("delta").
 format("json").
 format("parquet").
 mode("append")
 mode("error")
 mode("ignore")
 mode("overwrite")
 .save(delta_table_path)

✉  **jongert** 4 days ago

Correct:

Add to a delta table => format is delta

Add the rows to existing table => append

upvoted 2 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.

Actions**Answer Area**

Create another container named Container1 and copy the contents of Container2 to Container1.

Set the TTL for Container1 to null.

Create a new container named Container2 and copy the contents of Container1 to Container2.



Set the TTL for Container1 to 0.

Delete Container1.

Delete Container2.

**Answer Area**

Create a new container named Container2 and copy the contents of Container1 to Container2.

Delete Container1.

Correct Answer:

Create another container named Container1 and copy the contents of Container2 to Container1.

Delete Container2.

Question #110

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.

Actions	Answer Area
Create a new table named Table1v2 by running CTAS.	
Rename Table1v2 as Table1.	
Rename Table1 as Table1_old.	▶
Run DBCC PDW_SHOWSPACEUSED.	◀
Drop Table1_old.	↑
Drop the indexes of Table1.	↓



Answer Area

Create a new table named Table1v2 by running CTAS.

Rename Table1 as Table1_old.

Rename Table1v2 as Table1.

Drop Table1_old.

Correct Answer:

店铺：IT认证考试服务

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

Correct Answer: B

Community vote distribution

B (100%)

 **dakku987** 1 day, 22 hours ago

Selected Answer: B

IT NEED GIT CONFIG TO SAVE THE CHANGES

upvoted 1 times

 **jongert** 4 days ago

Selected Answer: B

Correct, triggers have nothing to do with saving the pipeline logic.

upvoted 1 times

Topic 2 - Question Set 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:

Answer Area

Input type:

A dropdown menu with two options: "Stream" and "Reference".

Stream
Reference

Function:

A dropdown menu with three options: "Aggregate", "Geospatial", and "Windowing".

Aggregate
Geospatial
Windowing

Answer Area

Input type:

A dropdown menu with two options: "Stream" and "Reference". The "Stream" option is highlighted with a green background.

Stream
Reference

Correct Answer:

Function:

A dropdown menu with three options: "Aggregate", "Geospatial", and "Windowing". The "Geospatial" option is highlighted with a green background.

Aggregate
Geospatial
Windowing

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>

Correct solution!
upvoted 49 times

□ **MrWood47** Highly Voted 12 months 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 13 times

□ **kkk5566** Most Recent 4 months, 1 week ago

1-stream
2-geospatial
upvoted 1 times

□ **akhil5432** 5 months ago

1-stream
2-geospatial
upvoted 2 times

□ **hiyoww** 5 months, 2 weeks ago

this is microsoft link for reference:
<https://learn.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios>
upvoted 1 times

□ **dheeraj7654** 7 months, 2 weeks ago

Correct
upvoted 3 times

□ **Ankit_Az** 7 months, 2 weeks ago

Stream & Geospatial is Correct
upvoted 2 times

□ **rocky48** 7 months, 3 weeks ago

Stream & Geospatial
upvoted 2 times

□ **hanzocuk** 1 year 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 1 times

□ **vigilante89** 1 year ago

Input Type: Stream
Function: Geospatial
upvoted 1 times

□ **Amnoyana** 1 year, 1 month ago

Good application and need to learn more about it!
upvoted 4 times

□ **Deeksha1234** 1 year, 5 months ago

solution is correct
upvoted 4 times

□ **ClassMistress** 1 year, 7 months ago

Correct
upvoted 2 times

□ **NewTuanAnh** 1 year, 9 months ago

Correct!
upvoted 2 times

□ **PallaviPatel** 1 year, 11 months ago

Correct
upvoted 2 times

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.

店铺：IT认证考试服务

Correct Answer: DF

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 (46%) DF (44%) 6%

✉️  **manquak** Highly Voted 2 years, 4 months ago

Partition input and output.

REF: <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization>

upvoted 64 times

✉️  **kolakone** 2 years, 3 months ago

Agree. And partitioning Input and output with same number of partitions gives the best performance optimization..

upvoted 12 times

✉️  **Lio95** Highly Voted 2 years, 3 months ago

No event consumer was mentioned. Therefore, partitioning output is not relevant. Answer is correct

upvoted 14 times

✉️  **nicolas1999** 2 years, 1 month 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

✉️  **Boompiee** 1 year, 8 months ago

The stream analytics job is the consumer.

upvoted 1 times

✉️  **Khadija10** Most Recent 1 week, 3 days ago

Selected Answer: CF

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 1 times

✉️  **jongert** 1 week, 6 days ago

Selected Answer: CF

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 1 times

✉️  **d046bc0** 4 weeks ago

Selected Answer: CF

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

⊕ **Momoanwar** 1 month ago

Selected Answer: DF

Chatgpt say DF :

The question in the image relates to optimizing the performance of an Azure Stream Analytics job. The correct actions would typically involve scaling the Streaming Units (SUs) appropriately based on the throughput needs and implementing query parallelization. In this context:

- Scaling up the SU count (option D) would improve performance if the current SU allocation is insufficient.
- Implementing query parallelization by partitioning the data input (option F) could also optimize performance as it would allow the job to process multiple data partitions concurrently.

upvoted 1 times

⊕ **ellala** 3 months ago

I would say DF is correct. Despite C being a correct option to optimize performance, we have no information about the output. If the output is Power BI, it does not support partition. Therefore we cannot state output partition without more information. Therefore best option will be SU

upvoted 3 times

⊕ **fahfouhi94** 3 months, 1 week ago

Selected Answer: DF

the question is about actions should you perform, in case of power bi output , we cannot partition the stream analytics output.SO D & F

upvoted 2 times

⊕ **EliteAllen** 4 months ago

Selected Answer: CF

Implement query parallelization by partitioning the data input (Option F): Parallelizing the query by partitioning the data input allows the Stream Analytics job to process multiple data streams concurrently, which can significantly improve performance, especially when dealing with a large volume of data.

Implement query parallelization by partitioning the data output (Option C): Similar to partitioning the data input, partitioning the data output allows for parallel writing to the output sinks, which can also enhance performance.

upvoted 2 times

⊕ **kkk5566** 4 months, 1 week ago

avoiding embarrassingly parallel jobs, I would go C &F

upvoted 1 times

⊕ **dp_learner** 7 months, 2 weeks ago

"An embarrassingly parallel job is the most scalable scenario in Azure Stream Analytics. It connects one partition of the input to one instance of the query to one partition of the output. This parallelism has the following requirements:

- ...
4. The number of input partitions must equal the number of output partitions."

ref : <https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization>

upvoted 3 times

⊕ **bakamon** 7 months, 3 weeks ago

Selected Answer: CF

input and output

upvoted 3 times

⊕ **rocky48** 7 months, 3 weeks ago

Selected Answer: CF

C. Implement query parallelization by partitioning the data output.

F. Implement query parallelization by partitioning the data input.

upvoted 3 times

⊕ **dksks** 8 months, 1 week ago

Selected Answer: CD

C. Implement query parallelization by partitioning the data output.

D. Scale the SU count for the job up.

Explanation:

A higher SU count provides more processing power and can improve the performance of the Azure Stream Analytics job. Scaling up the job by increasing the SU count can reduce query latency and improve throughput.

Partitioning the data output allows for query parallelization, which can improve the performance of the job. By dividing the output into partitions, the job can process data simultaneously, reducing the time required to complete the job.

upvoted 2 times

⊕ **esaade** 10 months, 1 week ago

To optimize the performance of the Azure Stream Analytics job, you should perform the following two actions:

C. Implement query parallelization by partitioning the data output. Partitioning the data output helps to distribute query processing across multiple partitions, which can improve performance for queries that require a large amount of processing power.

D. Scale the SU count for the job up. Scaling up the number of Streaming Units (SU) will provide more processing power for the job, which can improve performance.

Therefore, the correct answers are C and D. Implement query parallelization by partitioning the data output, and scale the SU count for the job up. upvoted 3 times

 **akk_1289** 11 months, 1 week ago

C. Implement query parallelization by partitioning the data output.

D. Scale the SU count for the job up.

By partitioning the data output, the query processing can be split into smaller, parallel tasks which can lead to better performance. Scaling up the SU count for the job increases the processing power available for the job, which can also lead to improved performance.

Note: The specific optimizations required may vary based on the specific requirements and nature of the data analysis solution.

upvoted 1 times

 **VivekMadas** 1 year ago

Already 120 SU used (6 per node = 20 nodes) - Adding extra wont be any use.

Answer would be Partitioning Input & Output

upvoted 4 times

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

Correct 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 2 years, 1 month 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 30 times

✉ **medsimus** Highly Voted 2 years, 3 months ago

Correct

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger?tabs=data-factory>

upvoted 12 times

✉ **lisa710** Most Recent 2 weeks, 3 days ago

Selected Answer: C

c is correct

upvoted 1 times

✉ **kkk5566** 4 months, 1 week ago

Selected Answer: C

custom trigger use Event-grid

upvoted 1 times

✉ **kkk5566** 4 months, 2 weeks ago

Correct. C

upvoted 1 times

✉ **kkk5566** 4 months, 2 weeks ago

Correct

upvoted 1 times

✉ **Ankit_Az** 7 months, 2 weeks ago

Selected Answer: C

Correct

upvoted 2 times

✉ **esaade** 10 months, 1 week 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** 1 year ago

Selected Answer: C

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** 1 year, 1 month ago

Why it's not Microsoft.Automation?

upvoted 1 times

✉ **Amnoyana** 1 year, 1 month ago

Would that be part of the Power Platform instead?

upvoted 1 times

✉ **Deeksha1234** 1 year, 5 months ago

correct

upvoted 1 times

✉ **PallaviPatel** 1 year, 11 months ago

Selected Answer: C

Correct.

upvoted 2 times

✉ **romanzdk** 1 year, 11 months ago

But EventHub does not support ADLS, only Blob storage

upvoted 1 times

✉ **romanzdk** 1 year, 11 months ago

<https://docs.microsoft.com/en-us/azure/event-grid/overview>

upvoted 2 times

✉ **Swagat039** 1 year, 12 months ago

C. is correct.

You need storage event trigger (for this Microsoft.EventGrid service needs to be enabled).

upvoted 1 times

✉ **Vardhan_Brahmanapally** 2 years, 2 months ago

Why not eventhub?

upvoted 4 times

✉ **wijaz789** 2 years, 4 months ago

Absolutely correct

upvoted 4 times

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

Correct 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** Highly Voted 2 years, 3 months ago

Correct!

upvoted 16 times

✉ **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: B

Automatic/Jobs - best for jobs and automated batch processing.

upvoted 1 times

✉ **Ankit_Az** 7 months, 2 weeks ago

Given Answer and explanation is correct

upvoted 2 times

✉ **Lestrang** 11 months, 2 weeks 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** 1 year, 3 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** 1 year, 5 months ago

right answer

upvoted 1 times

✉ **HebaN** 1 year, 7 months ago

Selected Answer: B

Its correct answer

upvoted 2 times

✉ **necktru** 1 year, 8 months ago

Selected Answer: B

correct

upvoted 1 times

□ **PallaviPatel** 1 year, 11 months ago

Selected Answer: B

correct.

upvoted 1 times

□ **satyamkishoresingh** 2 years, 4 months ago

What is automated cluster ?

upvoted 2 times

□ **wijaz789** 2 years, 4 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 19 times

□ **youngbug** 1 year, 5 months ago

hi, may I ask where I can find these words?

upvoted 2 times

□ **Swagat039** 1 year, 12 months ago

Job cluster

upvoted 2 times

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

Correct Answer:

Answer Area

```
WITH LastInWindow AS
(
    SELECT
        COUNT | ▼ (Time) AS LastEventTime
        MAX
        MIN
        TOPONE
    FROM
        Input TIMESTAMP BY Time
    GROUP BY
        (minute, 10)
)
SELECT
    Input.License_plate,
    Input.Make,
    Input.Time
FROM
    Input TIMESTAMP BY Time
    INNER JOIN LastInWindow
    ON (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 2 years, 3 months ago

correct

upvoted 28 times

✉  **Jerrylolu** 2 years ago

Why not Hopping Window??

upvoted 1 times

✉  **auwia** 6 months, 2 weeks ago

it needs 3 parameters in input.

upvoted 3 times

✉  **Wijn4nd** 1 year, 12 months ago

Because a hopping window can overlap, and we need the data from 10 minute time frames that DON'T overlap

upvoted 9 times

✉  **GodfreyMbizo** Highly Voted 11 months, 2 weeks 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 14 times

✉  **ExamDestroyer69** 3 weeks ago

The referenced link shows the provided answer is correct

upvoted 1 times

✉  **goldy29** 6 months ago

You are great!

upvoted 1 times

✉  **sdg2844** Most Recent 1 week, 4 days 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** 3 weeks, 5 days ago

error in code:

wrong: DATEDIFF(minut, Input, LastInWindow)

correct: DATEDIFF(minut, Input, LastEventTime)

upvoted 1 times

✉  **gggqqqqq** 3 months, 1 week 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** 4 months ago

correct

upvoted 1 times

□ **mamahani** 8 months ago

max / tumblingwindow/datediff

upvoted 3 times

□ **AZLearn111** 11 months, 2 weeks 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 4 times

□ **alphilla** 3 weeks, 3 days 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

□ **meatpoof** 1 week, 2 days 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

□ **mamahani** 8 months, 3 weeks 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 1 times

□ **Bro111** 1 year, 1 month ago

Are "Input" and "LastInWindow" DateTime fields, to be compared with datediff???

upvoted 1 times

□ **Bro111** 1 year, 1 month ago

I have the answer here:

<https://learn.microsoft.com/en-us/stream-analytics-query/join-azure-stream-analytics>

upvoted 2 times

□ **anks84** 1 year, 4 months ago

CORRECT

upvoted 1 times

□ **Deeksha1234** 1 year, 4 months ago

correct

upvoted 2 times

□ **y203** 1 year, 5 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** 1 year, 5 months ago

Correct!

The keyword "each" in "last vehicle passed during each 10-minute window" pretty much makes it Clear!

upvoted 1 times

□ **Revave2** 1 year, 6 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 2 times

□ **sensaint** 1 year, 1 month 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** 1 year, 11 months ago

correct.

upvoted 1 times

□ **BusinessApps** 1 year, 11 months 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** 2 years ago

Answer is 100% correct.
upvoted 2 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

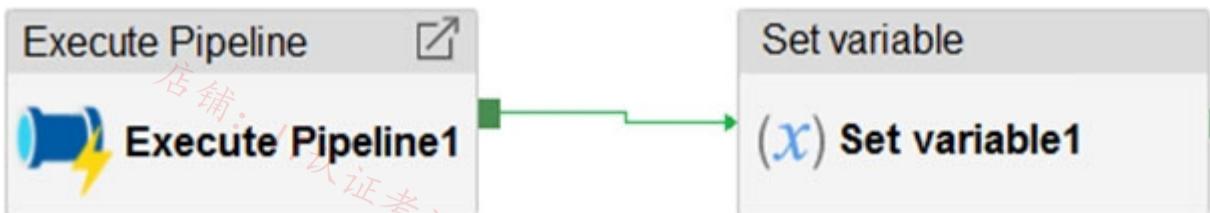
店铺: IT认证考试服务

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?

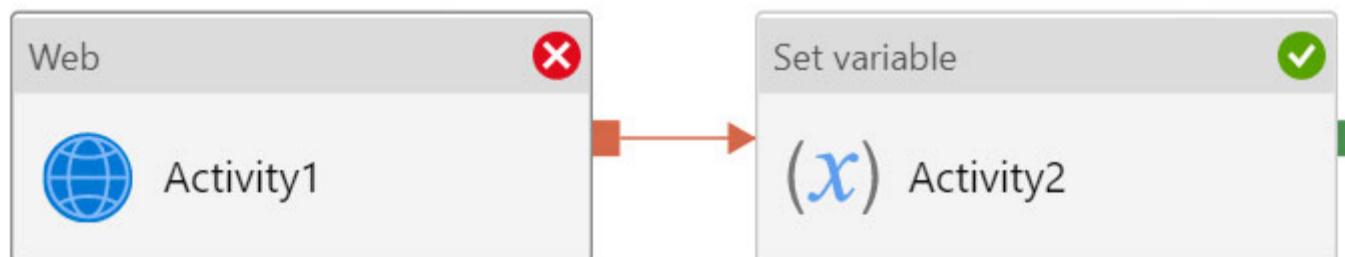
- A. Pipeline1 and Pipeline2 succeeded.
- B. Pipeline1 and Pipeline2 failed.
- C. Pipeline1 succeeded and Pipeline2 failed.
- D. Pipeline1 failed and Pipeline2 succeeded.

Correct 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 (89%)

11%

SaferSephysafersephy Highly Voted 2 years, 4 months ago

Correct answer is A. The trick is the fact that pipeline 1 only has a Failure dependency between the 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 44 times

Ram0202 Ram0202 7 months, 2 weeks ago

correct ,Execute pipeline explained here <https://youtu.be/Jkz1dtLrBE4>

upvoted 4 times

- **BK10** 1 year, 11 months ago
well explained! A is right
upvoted 1 times
- **echerish** Highly Voted 2 years, 4 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 27 times
- **kkk5566** Most Recent 4 months ago
Selected Answer: A
Correct answer is A.
upvoted 1 times
- **azaspirant** 4 months, 3 weeks ago
Selected Answer: B
How do we know which is the default behaviour? Its not mentioned anywhere whether there is a success dependency or failure dependency
upvoted 2 times
- **akhil5432** 5 months ago
Selected Answer: A
OPTION "A"
upvoted 1 times
- **joponlu** 7 months, 2 weeks ago
Selected Answer: A
Correct!!!
upvoted 2 times
- **mamahani** 8 months ago
Selected Answer: A
A is correct answer
upvoted 2 times
- **Billybob0604** 1 year, 1 month 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** 1 year, 3 months ago
Answer is correct.
This article helped: <https://www.sqlshack.com/dependencies-in-azure-data-factory/>
upvoted 1 times
- **Deeksha1234** 1 year, 5 months ago
Selected Answer: A
A is correct
upvoted 1 times
- **SebK** 1 year, 9 months ago
Selected Answer: A
Correct
upvoted 1 times
- **AngelJP** 1 year, 9 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 4 times
- **AvSUN** 4 months ago
Thanks
upvoted 1 times
- **PallaviPatel** 1 year, 11 months ago
Selected Answer: A
A correct. I agree with SaferSephy's comments below.
upvoted 3 times
- **dev2dev** 1 year, 11 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** 2 years ago

Selected Answer: A

Correct

upvoted 1 times

 **JSSA** 2 years ago

Correct answer is A

upvoted 1 times

 **rashjan** 2 years, 1 month ago

Selected Answer: A

correct

upvoted 1 times

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

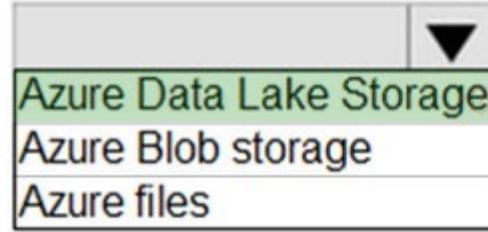
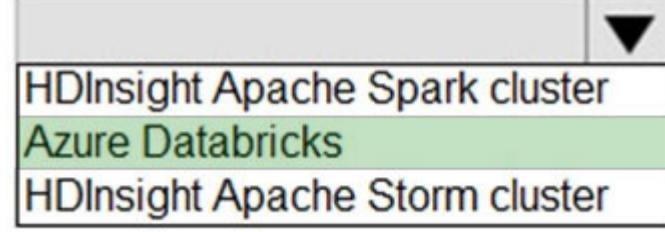
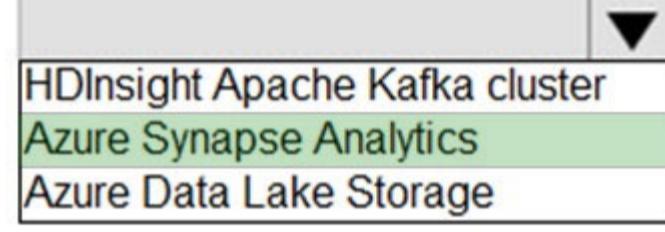
Prepare and Train

HDInsight Apache Spark cluster
Azure Databricks
HDInsight Apache Storm cluster

Model and Serve

HDInsight Apache Kafka cluster
Azure Synapse Analytics
Azure Data Lake Storage

Answer Area

Architecture requirement	Technology
Ingest	 <ul style="list-style-type: none">Logic AppsAzure Data FactoryAzure Automation
Store	 <ul style="list-style-type: none">Azure Data Lake StorageAzure Blob storageAzure files
Prepare and Train	 <ul style="list-style-type: none">HDInsight Apache Spark clusterAzure DatabricksHDInsight Apache Storm cluster
Model and Serve	 <ul style="list-style-type: none">HDInsight Apache Kafka clusterAzure Synapse AnalyticsAzure Data Lake Storage

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  2 years, 3 months ago

Is this an ad?

upvoted 63 times

 **phydev** 2 months, 1 week ago

You mean like all the Azure certification exams?

upvoted 1 times

□ **Bhabani83** 6 months, 1 week ago

sounds like one, replace databricks with Azure Fabric and then it will be for sure
upvoted 3 times

□ **Podavenna** Highly Voted 2 years, 3 months ago

Correct solution!
upvoted 47 times

□ **positivitypeople** Most Recent 2 weeks, 5 days ago

Got this question today on the exam
upvoted 2 times

□ **kkk5566** 4 months ago

Correct
upvoted 1 times

□ **anks84** 1 year, 4 months ago

Given answer is correct !
upvoted 5 times

□ **Deeksha1234** 1 year, 5 months ago

Correct
upvoted 2 times

□ **SebK** 1 year, 9 months ago

Correct
upvoted 1 times

□ **Massy** 1 year, 10 months ago

for the store, couldn't we use also Azure Blob Storage? It supports all the three requisites
upvoted 2 times

□ **kreative_feco** 1 day, 8 hours ago

ADLSG2 supports big data and large data storage as compared with blob
upvoted 1 times

□ **NewTuanAnh** 1 year, 9 months ago

Because ADLS Gen2 support Big Data Workload better
upvoted 4 times

□ **paras_gadhiya** 1 year, 10 months ago

Correct
upvoted 1 times

□ **PallaviPatel** 1 year, 11 months ago

Correct solution.
upvoted 1 times

□ **joeljohnrm** 2 years ago

Correct Solution
upvoted 1 times

□ **[Removed]** 2 years ago

for model~~and~~ server, HDI has all of this. Why DataBricks?
upvoted 1 times

□ **dara_44_6880** 1 year, 3 months ago

JAVA only supported in databricks
upvoted 3 times

□ **rockyc05** 1 year, 10 months ago

Support for SQL
upvoted 3 times

□ **rockyc05** 1 year, 10 months ago

Also seamless integration with AAD
upvoted 3 times

□ **corebit** 2 years ago

Would be best if people including answers that go against the popular responses provide some reference instead of blinding saying false
upvoted 3 times

□ **Akash0105** 2 years, 2 months ago

Answer is correct.

Azure Databricks supports java: <https://azure.microsoft.com/en-us/services/databricks/#overview>

upvoted 2 times

 **Pratikh** 2 years, 2 months ago

Databricks doesn't support Java so in the Prep and Train should be HDInsight Apache Spark Cluster

upvoted 4 times

 **KOSTA007** 2 years, 2 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 10 times

 **Aslam208** 2 years, 2 months ago

Databricks does not support Java, Prepare and Train should be Azure HDInsight Apache spark cluster

upvoted 1 times

 **Aslam208** 2 years 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 4 times

 **Samanda** 2 years, 2 months ago

false. kafka hd insight is the correct option in the last box

upvoted 1 times

DRAG DROP -

You have the following table named Employees.

first_name	last_name	hire_date	employee_type
Jane	Doe	2019-08-23	new
Ben	Smith	2017-12-15	Standard

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:

Values	Answer Area
CASE	<input type="text"/>
ELSE	<input type="text"/> WHEN hire_date >= '2019-01-01' THEN 'New'
OVER	<input type="text"/> 'standard'
PARTITION BY	END AS employee_type
ROW_NUMBER	FROM
	employees

Values	Answer Area
CASE	<input type="text"/> CASE
ELSE	<input type="text"/> WHEN hire_date >= '2019-01-01' THEN 'New'
OVER	<input type="text"/> ELSE 'standard'
PARTITION BY	END AS employee_type
ROW_NUMBER	FROM
	employees

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>

upvoted 41 times

- **[Removed]** Highly Voted 2 years, 4 months ago
The answer is correct. But, is this in the scope of this exam?
upvoted 9 times
- **mikutts** 2 years, 2 months ago
Got this question yesterday so yes.
upvoted 11 times
- **anto69** 1 year, 12 months ago
it seems
upvoted 1 times
- **parwa** 2 years, 4 months ago
make sense to me , data engineer should be able to write Queries
upvoted 10 times
- **kkk5566** Most Recent 4 months ago
Correct
upvoted 1 times
- **Ankit_Az** 7 months, 2 weeks ago
Correct
upvoted 1 times
- **GodfreyMbizo** 11 months, 2 weeks ago
they cant bring hard questions only for exam to balance
upvoted 1 times
- **DindaS** 11 months, 3 weeks ago
yes the answer is correct but not sure if they will be included the DP-203 exam
upvoted 1 times
- **Igor85** 1 year, 2 months ago
i wish all the questions in the exam were like that one :)
upvoted 5 times
- **Deeksha1234** 1 year, 5 months ago
correct
upvoted 1 times
- **NewTuanAnh** 1 year, 9 months ago
the answer is correct
CASE ...
WHEN ... THEN...
ELSE ...
upvoted 2 times
- **PallaviPatel** 1 year, 11 months ago
correct
upvoted 2 times

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.

```
{  
    "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"  
                    }  
                }  
            ]  
        }  
    }  
}
```

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:

Values**Answer Area**

```
select*
FROM
    (
        BULK 'https://contoso.blob.core.windows.net/contosodw',
        FORMAT= 'CSV',
        fieldterminator = '0x0b',
        fieldquote = '0x0b',
        rowterminator = '0x0b'
    )
    with (id varchar(50),
        contextdateeventTime 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 openrowset (contextcustomdimensions)
with ( ProfileType varchar(50) '$.customerInfo.ProfileType',
        RoomName varchar(50) '$.customerInfo.RoomName',
        CustomerName varchar(50) '$.customerInfo.CustomerName',
        UserName varchar(50) '$.customerInfo.UserName'
)
```

Correct Answer:

Values**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),
        contextdateeventTime 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 2 years, 4 months ago

Answer is correct

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files>

upvoted 45 times

□ **Lrng15** 2 years, 3 months ago

answer is correct as per this link

upvoted 2 times

□ **gf2tw** Highly Voted 2 years, 4 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 10 times

□ **vctrhugo** 6 months, 2 weeks 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 2 times

□ **anto69** 1 year, 12 months ago

agree with u

upvoted 1 times

□ **gsssd4scoder** 2 years, 2 months ago

agree with you, very misleading

upvoted 1 times

□ **dev2dev** 1 year, 11 months ago

Look at the WITH statement, the csv column can contain json data.

upvoted 1 times

□ **kkk5566** Most Recent 4 months ago

Answer is correct

upvoted 1 times

□ **mamahani** 8 months ago

openrowset / openjson

upvoted 1 times

□ **zorko10** 1 year, 2 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** 1 year, 5 months ago

correct

upvoted 1 times

□ **SebK** 1 year, 9 months ago

Correct

upvoted 1 times

□ **PallaviPatel** 1 year, 11 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.

Date	Temp
...	...
18-01-2021	3
19-01-2021	4
20-01-2021	2
21-01-2021	2
...	...

You need to produce the following table by using a Spark SQL query.

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

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:

Values 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 (
    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

```

CAST

COLLATE

CONVERT

FLATTEN

PIVOT

UNPIVOT

Correct Answer:

Values 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'
)
PIVOT (
    AVG ( CAST (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
```

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**  2 years, 3 months ago

correct answer, pivot and cast

upvoted 45 times

✉  **ggggyyyyy**  2 years, 3 months ago

correct. cast not convert

upvoted 5 times

✉  **MarkJoh**  1 month, 1 week 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

✉  **jongert** 1 week, 5 days 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 1 times

✉  **ellala** 3 months ago

Answer is correct

upvoted 1 times

✉  **kkk5566** 4 months ago

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

Correct 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 2 years, 4 months ago

Annotation

upvoted 22 times

✉  **anto69** 1 year, 12 months ago

Cause ADF pipelines are not first class resources

upvoted 3 times

✉  **AhmedDaffaie** Highly Voted 1 year, 9 months ago

What is the difference between resource tags and annotations?

upvoted 11 times

✉  **kkk5566** Most Recent 4 months ago

Selected Answer: D

D -Annotation

upvoted 1 times

✉  **akhil5432** 5 months ago

Selected Answer: D

OPTION -D

upvoted 1 times

✉  **Ankit_Az** 7 months, 2 weeks ago

Selected Answer: D

Correct

upvoted 1 times

✉  **joponlu** 7 months, 2 weeks ago

Selected Answer: D

D is correct!!

upvoted 1 times

✉  **esaade** 10 months, 1 week 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 2 times

✉  **DindaS** 11 months, 3 weeks ago

D -Annotation

upvoted 3 times

✉ **vigilante89** 1 year ago

Selected Answer: D

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 4 times

✉ **arunesh789** 1 year, 3 months ago

hjghfgh

upvoted 2 times

✉ **arunesh789** 1 year, 3 months ago

KIndly delete above comment. Answer is D.

upvoted 3 times

✉ **Deeksha1234** 1 year, 5 months ago

Selected Answer: D

correct

upvoted 1 times

✉ **paras_gadhiya** 1 year, 10 months ago

Correct!

upvoted 1 times

✉ **PallaviPatel** 1 year, 11 months ago

Selected Answer: D

correct

upvoted 1 times

✉ **huesazo** 1 year, 11 months ago

Selected Answer: D

Anotacion

upvoted 1 times

✉ **aarthy2** 2 years, 3 months ago

yes correct, annotation provides label functionality than show in pipeline monitoring.

upvoted 2 times

HOTSPOT -

The following code segment is used to create an Azure Databricks cluster.

```
{
  "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": []
}
```

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:

Answer Area

Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	<input type="radio"/>	<input type="radio"/>
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	<input type="radio"/>	<input type="radio"/>
The Databricks cluster supports the creation of a Delta Lake table.	<input type="radio"/>	<input type="radio"/>

Answer Area

Statements	Yes	No
Correct Answer: The Databricks cluster supports multiple concurrent users.	<input checked="" type="radio"/>	<input type="radio"/>
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	<input type="radio"/>	<input checked="" type="radio"/>
The Databricks cluster supports the creation of a Delta Lake table.	<input checked="" type="radio"/>	<input type="radio"/>

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 2 years, 2 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 50 times

□  **semauni** 5 months, 2 weeks ago

For 1, where do you see high concurrency?

upvoted 2 times

□  **semauni** 5 months, 2 weeks ago

I hadn't read the other comments yet, apparently it's in 'serverless' :)

upvoted 1 times

□  **Egocentric** 1 year, 8 months ago

agree on this one

upvoted 3 times

□  **Canary_2021** Highly Voted 2 years 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 34 times

□  **kkk5566** Most Recent 4 months ago

the answer is Yes, No, Yes.

upvoted 1 times

□  **hiyoww** 5 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** 8 months ago

yes / no / yes

upvoted 2 times

□  **Igor85** 1 year, 2 months ago

i guess this question won't be relevant anymore, since cluster creation UI has changed

upvoted 3 times

□  **WielK** 8 months, 3 weeks ago

They still use this question, I had this one on my exam this week

upvoted 3 times

□  **US007** 1 year, 5 months ago

1. should be 'No'. Its a standard cluster and it also has scala which is not supported on High Concurrency cluster.

upvoted 4 times

□  **Deeksha1234** 1 year, 5 months ago

Yes, No, Yes

upvoted 2 times

□ **PallaviPatel** 1 year, 11 months ago

Correct Answer. I agree with Canary_2021

upvoted 4 times

□ **edba** 2 years, 1 month 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** 2 years, 2 months ago

what is the answer lol

upvoted 3 times

□ **aarthy2** 2 years, 3 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** 2 years, 3 months ago

IMO NO, YES, YES

upvoted 1 times

□ **rav009** 2 years, 3 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** 2 years, 4 months ago

what is correct answer here please?

upvoted 2 times

□ **amma** 2 years, 4 months ago

Yes No No

upvoted 9 times

□ **Amyqwertyu** 2 years, 4 months ago

High Concurrency clusters are intended for use by multiple users. hence correct answer

upvoted 2 times

□ **Amalbenrebai** 2 years, 4 months ago

NO, NO, YES

upvoted 3 times

□ **petulda** 2 years, 4 months ago

The cluster enables autoscaling. Does this mean it minimize costs ?

upvoted 2 times

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

Correct Answer: C

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics>

Community vote distribution

B (71%) C (29%)

✉ **kolakone** Highly Voted 2 years, 3 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 85 times

✉ **anto69** 1 year, 11 months ago

But Python runs on Event Hubs why the other service does should support Python too?

upvoted 2 times

✉ **Aditya0891** 1 year, 6 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** 7 months ago

therefore i agree wih ASA

upvoted 1 times

✉ **RoyP654** 7 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

✉ **ExamDestroyer69** 3 weeks 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** Highly Voted 1 year, 11 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 14 times

✉ **RoyP654** 7 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

✉ **maxCarter** Most Recent 3 days, 14 hours ago

Azure Databricks

upvoted 1 times

✉ **HSZ** 4 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

✉ **mav2000** 3 weeks 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

✉ **kkk5566** 4 months ago

Selected Answer: B

Databricks supports Python.

upvoted 1 times

✉ **kkk5566** 4 months, 2 weeks ago

<https://learn.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql?view=sql-server-ver16>

upvoted 1 times

✉ **kkk5566** 4 months, 2 weeks ago

RIGHT :
t<20100101, 20100101<=t<20110101, 20110101<=t<20120101
20120101<=t
LEFT
t<=20100101, 20100101<=t<=20110101, 20110101<=t<=20120101,
t>20120101

upvoted 1 times

✉ **kkk5566** 4 months ago

post wrong quiz

upvoted 1 times

✉ **Abdullah77** 4 months, 3 weeks ago

C is the answer

upvoted 1 times

✉ **Zak_Zakaria** 5 months, 2 weeks ago

In this question, it's mentioned that Python will be used to produce the statistical solution (necessarily By Event Hubs), and the needed solution is the one which will follow and process the Statistical Solution (already processed by Python via Event Hubs). So in my opinion the answer is correct, I go for ASA too.

upvoted 1 times

✉ **andjurovicela** 5 months, 2 weeks ago

ChatGPT would choose Databricks as well :)

upvoted 1 times

✉ **phydev** 2 months, 1 week ago

Try again!

upvoted 2 times

✉ **yaberjorge** 6 months ago

Azure Stream Analytics

According Chatgpt

upvoted 2 times

✉ **auwia** 6 months, 2 weeks ago

Selected Answer: B

Databricks supports Python.

upvoted 2 times

✉ **vctrhugo** 7 months ago

Selected Answer: B

Azure Stream Analytics only supports programmability in SQL and JavaScript, while Apache Spark in Azure Databricks supports C#/F#, Java, Python, R, Scala.

<https://learn.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/stream-processing#general-capabilities>

upvoted 4 times

□ **TestingCRM** 7 months, 2 weeks ago

Azure Databricks. See <https://learn.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/stream-processing#general-capabilities>. Python is not supported in Azure Stream Analytics according to this documentation.

upvoted 5 times

□ **janaki** 7 months, 2 weeks ago

Selected Answer: C

Its stream analytucs because it's asking for real-time analytics in azure

upvoted 1 times

□ **auwia** 6 months, 3 weeks ago

Near real time

upvoted 1 times

□ **explorerhp** 7 months, 3 weeks ago

It should be C - Azure Stream Analytics. Don't reason other comments.

upvoted 1 times

□ **rocky48** 7 months, 3 weeks ago

Selected Answer: B

Azure Databricks

upvoted 2 times

□ **dksks** 8 months, 1 week ago

Selected Answer: B

Stream Analytics only supports a limited set of built-in functions and cannot execute custom Python code directly. Therefore, it may not be the best choice for this scenario where custom proprietary Python functions are required.

upvoted 2 times

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

	▼
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 FOR VALUES

RIGHT
LEFT

()

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, datecol>20100101 and datecol<=20110101, datecol>20110101 and datecol<=20120101, datecol>20120101

But if you use range RIGHT and 20100101,20110101,20120101

Partition will be, datecol<20100101, datecol>=20100101 and datecol<20110101, datecol>=20110101 and datecol<20120101, 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 2 years ago

Answer is correct.

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql?view=sql-server-ver15>

upvoted 25 times

victor90 Highly Voted 2 years 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 14 times

TestMitch 2 years ago

No! That's wrong! Number of partitions created = Number of partition boundaries specified + 1.

upvoted 25 times

onyerleft 2 years ago

Choosing box 2 with range right would create five partitions. The first partition would be <20090101. So the provided answer is correct

upvoted 5 times

kkk5566 Most Recent 4 months ago

RIGHT :

t<20100101, 20100101<=t<20110101, 20110101<=t<20120101

20120101<=t

LEFT

t<=20100101, 20100101<t<=20110101, 20110101<t<=20120101,

t>20120101

upvoted 2 times

Maddhy 1 year, 1 month ago

Answer is

- 1.right
2. Last option becoz there they mentioned 4 partitions (I'm sure that it is guaranteed)

upvoted 3 times

 **Maddhy** 1 year, 1 month ago

The reason we are using right that is here the values are not null

upvoted 1 times

 **rzeng** 1 year, 2 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** 1 year, 5 months ago

Answer is correct.. given 1st 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** 1 year, 5 months ago

Answer is correct

upvoted 1 times

 **dsp17** 1 year, 6 months ago

How to remember LEFT / RIGHT concept, it's confusing, pls help

upvoted 2 times

 **monibun** 1 year, 4 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** 1 year, 11 months ago

correct Answer.

upvoted 1 times

 **dev2dev** 1 year, 11 months ago

where does 2009 year stored? i think the 1st choice should be LEFT so th

upvoted 4 times

 **allagowf** 1 year, 3 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 and 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 partitions each partition contains all the orders placed during a given calendar year.

upvoted 3 times

 **VeroDon** 2 years ago

correct

upvoted 2 times

 **alexleonvalencia** 2 years, 1 month ago

Respuesta correcta. RIGTH, [3 VALORES].

upvoted 2 times

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.

```
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
)
```

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. [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,

Correct Answer: BE

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.

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

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 2 years, 4 months ago

Correct answer

upvoted 27 times

lukeonline Highly Voted 2 years 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 13 times

Dicer 1 year, 5 months ago

because if you look at Type 3 examples, usually there are "original" and "current".

upvoted 9 times

kkk5566 Most Recent 4 months ago

Selected Answer: BE

BE is Correct

upvoted 1 times

✉ **Ankit_Az** 7 months, 2 weeks ago

Selected Answer: BE

Correct

upvoted 1 times

✉ **esaade** 10 months, 1 week 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** 1 year ago

Selected Answer: BE

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** 1 year 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** 1 year, 1 month ago

This is the same question as Question #59Topic 1

upvoted 5 times

✉ **anks84** 1 year, 4 months ago

Selected Answer: BE

Given Answer is correct !!

upvoted 1 times

✉ **Deeksha1234** 1 year, 5 months ago

BE is correct

upvoted 1 times

✉ **Sriramiyer92** 1 year, 5 months ago

Right Answer!

upvoted 1 times

✉ **Fishy_Marcy** 1 year, 6 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 mentioning the Previous ProductCategory key wasn't there yet, but should have been)

upvoted 1 times

✉ **Egocentric** 1 year, 8 months ago

BE is correct

upvoted 1 times

✉ **SebK** 1 year, 9 months ago

Selected Answer: BE

Correct

upvoted 1 times

✉ **VeroDon** 2 years ago

correct

upvoted 1 times

 **dija123** 2 years, 1 month ago

Selected Answer: BE

B and E

upvoted 4 times

 **Leo0802** 2 years, 2 months ago

A&C are used for Type 2 SCD because when dimension changing you need to create a new row to store new effective date, D is just for Type 1 SCD update data, so the answer is B&E

upvoted 4 times

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

Correct 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

A (69%) B (31%)

 [Removed]  2 years, 4 months ago

The answer should be "Yes". Hopping window with hop size equals window size should be the same as Tumbling window.
upvoted 120 times

 DataEngineer7331 8 months, 3 weeks ago

A Tumbling Window would be correct. But as stated in the following, a hopping window can be the same as a tumbling window: "To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size."
<https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions#hopping-window>
upvoted 9 times

 YipingRuan 2 years, 2 months ago

ensure that each tweet is counted only once
upvoted 21 times

 dsp17 1 year, 5 months ago

Correct Ans: B. tumbling window
Read the question carefully - "The solution must ensure that each tweet is counted only once."
By definition, hopping window is not non-overlapping
upvoted 8 times

 strato 1 year, 5 months ago

a hopping window with equal window size to the hopping is effectively non-overlapping. If the size of the hop was smaller, sure.
upvoted 6 times

 medsimus  2 years, 2 months ago

Unlike tumbling windows, hopping windows model scheduled overlapping windows. A hopping window specification consists 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 17 times

 Amir_Abed  1 week, 1 day ago

Answer should be B
upvoted 1 times

 6d954df 2 weeks 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 weeks, 5 days ago

Got this question today on the exam
upvoted 2 times

 **Momoanwar** 1 month ago

Selected Answer: B

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 1 times

 **phydev** 2 months, 1 week ago

Selected Answer: A

Yes, the solution described using a hopping window with a hop size of 10 seconds and a window size of 10 seconds would meet the goal of counting tweets in each 10-second window.

In Azure Stream Analytics, a hopping window moves forward in time at regular intervals (the hop size) and collects data within the specified window size. In this case, with a window size of 10 seconds and a hop size of 10 seconds, you ensure that the system counts tweets

upvoted 1 times

 **jhargett1** 2 months, 1 week ago

Selected Answer: B

Solution: B. No

Using a hopping window with a hop size of 10 seconds and a window size of 10 seconds will not ensure that each tweet is counted only once. In a hopping window, data can be included in multiple windows as it "hops" forward in time. This means that a tweet could be counted in more than one 10-second window, depending on the specific timing of the tweets.

To ensure that each tweet is counted only once in 10-second windows, you should use a tumbling window with a size of 10 seconds. In a tumbling window, each event is assigned to a single, non-overlapping window, and it is only counted once. This is the appropriate window type to meet the goal.

upvoted 3 times

 **ellala** 3 months ago

Selected Answer: A

A hopping windows with same hop size as window size is the same as a tumbling window

upvoted 1 times

 **Chemmangat** 3 months, 3 weeks ago

i got this question in my Exam, and immediately after selecting Hopping Window as my answer, the next question was having the solution as Tumbling window, I am not sure if two 'Yes' are allowed, but both of them seems to be correct.

upvoted 1 times

 **kkk5566** 4 months ago

Selected Answer: A

should be A

upvoted 1 times

 **Kunal2699** 4 months, 3 weeks ago

The answer should be 'A'. Also will they update this question given twiter has changed to X

upvoted 2 times

 **akhil5432** 5 months ago

Selected Answer: B

NO is correct ans

upvoted 1 times

 **Ankit_Az** 7 months, 2 weeks ago

Selected Answer: A

Yes, a hopping window with hop size of window size is indirectly tumbling window and event will be counted once.

"A hopping window specification consists 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'."

<https://learn.microsoft.com/en-us/stream-analytics-query/hopping-window-azure-stream-analytics>

upvoted 1 times

 **mamahani** 8 months ago

Selected Answer: A

A is correct answer

upvoted 1 times

 **frankanalysis** 8 months, 4 weeks ago

Technically it is No because OffSetSize is not given, which would make all the difference in counting tweets more than once.

If the offset size is smaller than the size of the window, then it is possible for the same data point to be included in multiple windows, resulting in duplicate counts.

In the case of the proposed solution for counting tweets in 10-second windows using a hopping window with a hop size and window size of 10 seconds, each tweet will be counted only once as long as the offset size is set to the same value as the window size (10 seconds). However, if the offset size is set to a smaller value, such as 5 seconds, it is possible for the same tweet to be counted in two consecutive windows, resulting in duplicate counts.

For example, suppose that the offset size is set to 5 seconds, and a tweet is received at 00:00:08. This tweet would be included in the window that starts at 00:00:00 and ends at 00:00:10, as well as the window that starts at 00:00:05 and ends at 00:00:15. As a result, the tweet would be counted twice, once in each window.

upvoted 1 times

 **peches** 7 months, 1 week ago

I think you're referring to the `hopsize`, not the `offsetsize`. The `offsetsize` is optional and is used to pick up events that happened EXACTLY at the beginning or end of the window (by default they are inclusive in the end and exclusive in the beginning). The `hopsize` was stated to be 10s, so in your example the tweet at 08 would only be included in the 00-10 window, since the next window goes from 10-20.

<https://learn.microsoft.com/en-us/stream-analytics-query/hopping-window-azure-stream-analytics>

upvoted 1 times

 **esaade** 10 months, 1 week ago

A. Yes, this meets the goal. A hopping window moves continuously over the stream of events with each event assigned to one or more windows based on the hop and window size. In this case, a hopping window with a hop size of 10 seconds and a window size of 10 seconds would count each tweet only once within the designated window.

upvoted 1 times

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

Correct 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 2 years, 7 months ago

answer is correct
upvoted 20 times

 **allagowf** Highly Voted 1 year, 3 months ago

Selected Answer: B

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

 **positivitypeople** Most Recent 2 weeks, 5 days ago

Got this question today on the exam
upvoted 1 times

 **MioHaze** 3 months, 4 weeks ago

Selected Answer: B
correct
upvoted 1 times

 **kkk5566** 4 months ago

Selected Answer: B
B is correct
upvoted 1 times

 **Ankit_Az** 7 months, 2 weeks ago

Selected Answer: B
Correct
upvoted 1 times

 **DindaS** 11 months, 3 weeks ago

definitely No. as the twits will be counted more than once in this solution
upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

CORRECR
upvoted 1 times

 **Dicer** 1 year, 5 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** 1 year, 5 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** 1 year, 6 months ago

Selected Answer: B

B is the correct answer.

upvoted 1 times

 **Egocentric** 1 year, 8 months ago

hop size must be the same as window size... so B is correct

upvoted 1 times

 **Yohannesmulu** 1 year, 10 months ago

Selected Answer: B

B is the correct answer.

upvoted 2 times

 **scass** 1 year, 10 months ago

Selected Answer: B

Correct answer is B

upvoted 2 times

 **jh777** 1 year, 11 months ago

correct B

upvoted 1 times

 **PallaviPatel** 1 year, 11 months ago

Selected Answer: B

correct.

upvoted 2 times

 **Canary_2021** 2 years ago

Selected Answer: B

Answer is correct

upvoted 3 times

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

```

店铺: IT认证考试服务
SELECT
    [user],
    feature,
    DATEADD(
    DATEDIFF(
    DATEPART(
        second,
        (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
        ISFIRST
        LAST
        TOPONE
    Time) as duration
FROM input TIMESTAMP BY Time
WHERE
    Event = 'end'

```

Correct Answer:**Answer Area**

```

店铺: IT认证考试服务
SELECT
    [user],
    feature,
    DATEADD(
        DATEDIFF(
        DATEPART(
            second,
            (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
            ISFIRST
            LAST
            TOPONE
        Time) as duration
    FROM input TIMESTAMP BY Time
    WHERE
        Event = 'end'

```

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 2 years, 7 months ago

correct

upvoted 44 times

□ **ohana** Highly Voted 2 years, 2 months ago

Took the exam today. This question came out.

Ans: DateDiff, Last

upvoted 24 times

□ **romanzdk** 1 year, 11 months ago

how do you know?

upvoted 2 times

□ **assU2** 1 year, 11 months ago

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns#detect-the-duration-between-events>

upvoted 12 times

□ **kkk5566** Most Recent 4 months ago

Ans: DateDiff, Last

upvoted 1 times

□ **semauni** 5 months, 2 weeks ago

Why is the answer to the second blank LAST() instead of TOPONE(), when this is about the startdate?

upvoted 1 times

□ **Ram9198** 5 months ago

LAST gives the most recent event it contradicts the name.. ISFIRST gives the oldest event... TOP ONE is a aggregate function which requires ORDER BY mandatory more like a ranking function here no order by so eliminated

upvoted 2 times

□ **Deeksha1234** 1 year, 5 months ago

answer is correct

upvoted 2 times

□ **RajeshAzure** 1 year, 5 months ago

When Event = 'Start' should not be there in this question

upvoted 1 times

□ **PallaviPatel** 1 year, 11 months ago

correct answer.

upvoted 1 times

□ **hugoborda** 2 years, 3 months ago

The answer is correct

upvoted 3 times

□ **mdalorso** 2 years, 5 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

□ **AvithK** 2 years, 4 months ago

so is the answer DATEDIFF+LAST incorrect then?

upvoted 1 times

□ **Juan27** 2 years, 4 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** 2 years, 6 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

You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytic dedicated SQL pool. The CSV file contains three columns named username, comment, and date.

The data flow already contains the following:

- A source transformation.
- A Derived Column transformation to set the appropriate types of data.
- A sink transformation to land the data in the pool.

You need to ensure that the data flow meets the following requirements:

- All valid rows must be written to the destination table.
- Truncation errors in the comment column must be avoided proactively.
- Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. To the data flow, add a sink transformation to write the rows to a file in blob storage.
- B. To the data flow, add a Conditional Split transformation to separate the rows that will cause truncation errors.
- C. To the data flow, add a filter transformation to filter out rows that will cause truncation errors.
- D. Add a select transformation to select only the rows that will cause truncation errors.

Correct Answer: AB

B: Example:

1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream.

Any row that is larger than five will go into the BadRows stream.

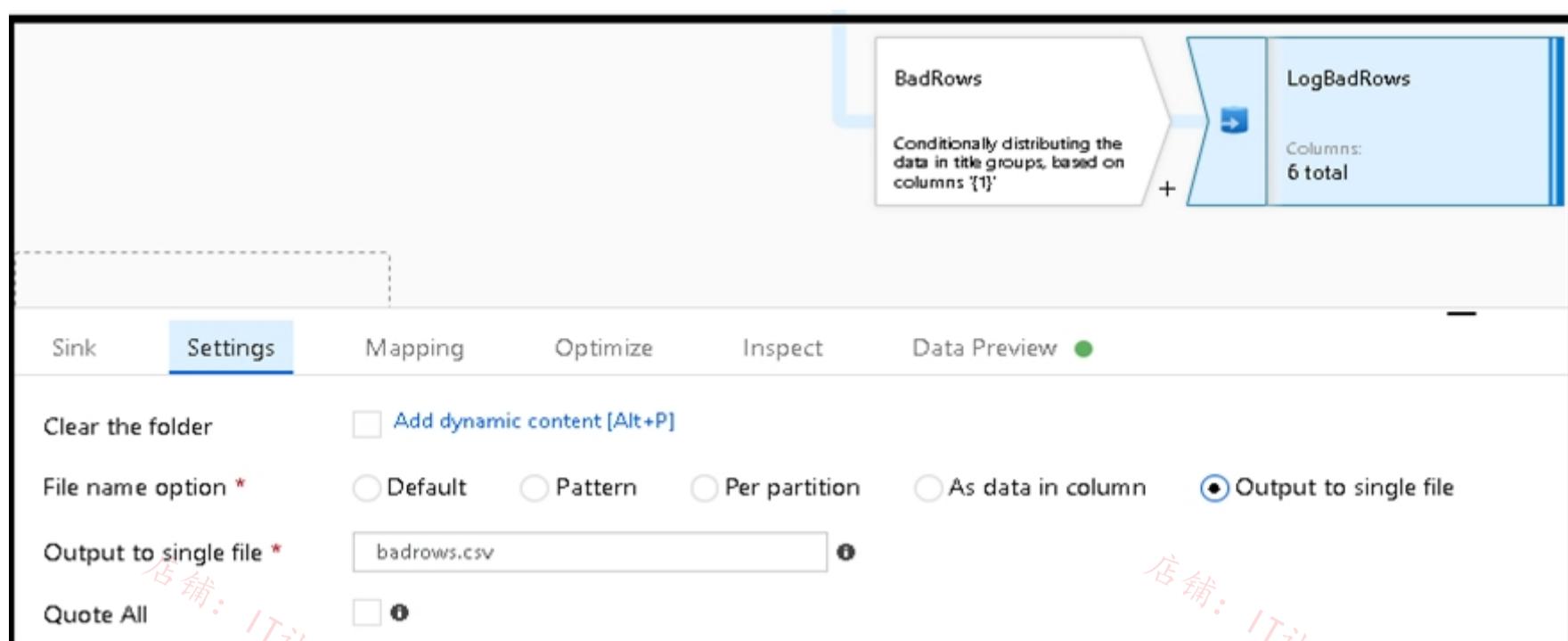
STREAM NAMES	CONDITION
GoodRows	length(title) <= 5
BadRows	Rows that do not meet any condition will use this output stream

2. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream.

Any row that is larger than five will go into the BadRows stream.

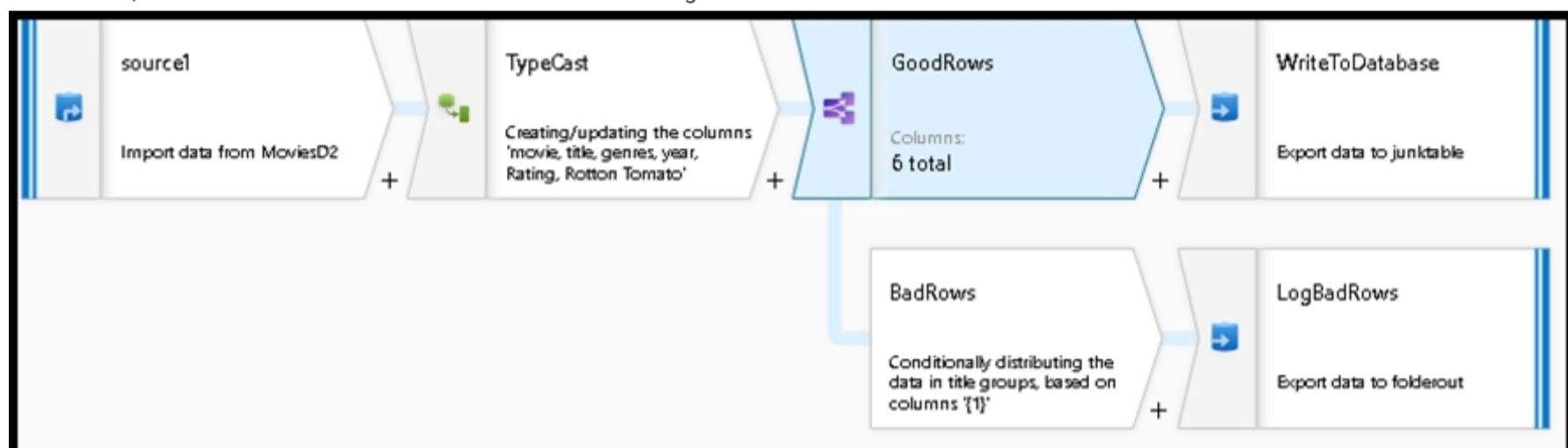
A:

3. Now we need to log the rows that failed. Add a sink transformation to the BadRows stream for logging. Here, we'll "auto-map" all of the fields so that we have logging of the complete transaction record. This is a text-delimited CSV file output to a single file in Blob Storage. We'll call the log file "badrows.csv".



4. The completed data flow is shown below. We are now able to split off error rows to avoid the SQL truncation errors and put those entries into a log file.

Meanwhile, successful rows can continue to write to our target database.



Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-data-flow-error-rows>

Community vote distribution

AB (100%)

Francesco1985 [Highly Voted] 2 years, 6 months ago

correct

upvoted 34 times

onyerleft [Highly Voted] 2 years ago

Selected Answer: AB

Conditional split with a sink transformation is the correct answer https://docs.microsoft.com/en-us/azure/data-factory/how-to-data-flow-error-rows?WT.mc_id=esi_studyguide_content_wwl#how-to-design-around-this-condition

upvoted 6 times

kkk5566 [Most Recent] 4 months ago

Selected Answer: AB

CORRECT

upvoted 1 times

anks84 1 year, 4 months ago

Selected Answer: AB

CORRECT

upvoted 3 times

Deeksha1234 1 year, 5 months ago

Selected Answer: AB

A&B are correct

upvoted 2 times

SebK 1 year, 9 months ago

Correct

upvoted 1 times

EmmettBrown 1 year, 11 months ago

Selected Answer: AB

Correct A and B
upvoted 2 times

 **PallaviPatel** 1 year, 11 months ago

Selected Answer: AB
Correct
upvoted 1 times

 **victor90** 2 years ago

I agree with the answer.
<https://www.microsoft.com/en-us/videoplayer/embed/RE4uOHj>
upvoted 1 times

 **rashjan** 2 years, 1 month ago

Selected Answer: AB
correct
upvoted 1 times

 **satyamkishoresingh** 2 years, 4 months ago

Why not C ?
upvoted 1 times

 **Marcus1612** 2 years, 4 months ago

You are right ! The job could be done by C only, BUT ! it has been asked for two(2) actions. No choice to add a SPLIT then another sink where to drop the bad records
upvoted 3 times

 **berserkzap** 2 years, 2 months ago

But you need to write to blob right ? How can only C help with that?
upvoted 3 times

 **AvithK** 2 years, 5 months ago

Bad rows go to 'folder out' and the good rows to the junk table? How come?
upvoted 1 times

 **data_guy** 1 month ago

The third point "A sink transformation to land the data in the pool" is for the good rows. You choose the conditional split to ensure that any bad rows are put somewhere else before they break the logic of the data flow.
upvoted 1 times

DRAG DROP -

You need to create an Azure Data Factory pipeline to process data for the following three departments at your company: Ecommerce, retail, and wholesale. The solution must ensure that data can also be processed for the entire company.

How should you complete the Data Factory data flow script? 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

```
all, ecommerce, retail, wholesale
dept=='ecommerce', dept=='retail',
dept=='wholesale'
dept=='ecommerce', dept==
'wholesale', dept=='retail'
disjoint: false
disjoint: true
ecommerce, retail, wholesale, all
```

Answer Area

```
CleanData
split(
    )
) ~> SplitByDept@()
```

Correct Answer:**Values**

```
all, ecommerce, retail, wholesale
dept=='ecommerce', dept=='retail',
dept=='wholesale'
dept=='ecommerce', dept==
'wholesale', dept=='retail'
disjoint: false
disjoint: true
ecommerce, retail, wholesale, all
```

Answer Area

```
CleanData
split(
    dept=='ecommerce', dept=='retail',
    dept=='wholesale'
        disjoint: false
) ~> SplitByDept@()
    ecommerce, retail, wholesale, all )
```

The conditional split transformation routes data rows to different streams based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language. The transformation evaluates expressions, and based on the results, directs the data row to the specified stream.

Box 1: dept=='ecommerce', dept=='retail', dept=='wholesale'

First we put the condition. The order must match the stream labeling we define in Box 3.

Syntax:

```
<incomingStream>
split(
<conditionalExpression1>
<conditionalExpression2>
...
disjoint: {true | false}
) ~> <splitTx>@(stream1, stream2,..., <defaultStream>)
```

Box 2: discount : false -

disjoint is false because the data goes to the first matching condition. All remaining rows matching the third condition go to output stream all.

Box 3: ecommerce, retail, wholesale, all

Label the streams -

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split>

I think "disjoint" should be True, so that data can be sent to all matching conditions. In this way the "all" output can get the data from every department, which ensures that "data can also be processed by the entire company".

upvoted 79 times

✉  **Stevyke** 2 years, 6 months ago

I concur with @Aleksx42 thought. Since we want to process for each dept (3 streams), then we must ensure we can still process for ALL depts at the same time (4th or default stream), hence DISJOINT:TRUE. Else, DISJOINT:FALSE.

upvoted 9 times

✉  **DataSaM** 6 months, 1 week ago

Disagree, all is like an else

upvoted 3 times

✉  **MrityunjayPrabhat** 1 year, 4 months ago

All is not defined in split so it has to be false. Refer

[](https://docs.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split#:~:text=CleanData%0A%20%20%20%20split(%0A%20%20%20%20%20%20%20%20year%20%3C%201960%2C%0A%09%20%20%20%20year%20%3E%201980%2C%0A%09%20%20%20%20%20disjoint%3A%20false%0A%20%20%20%20%20)%20~%3E%20SplitByYear%40(moviesBefore1960%2C%20moviesAfter1980%2C%20AllOtherMovies))

upvoted 6 times

✉  **kkk5566** 4 months, 2 weeks ago

disjoint is false because the data goes to the first matching condition rather than all matching conditions.

upvoted 1 times

✉  **rav009** 2 years, 3 months ago

agree, disjoint=true means the record will go through all the condition.

upvoted 4 times

✉  **mayank** Highly Voted 2 years, 7 months ago

As per the link provided in the explanation disjoint:false looks correct. I believe you must go through the link <https://docs.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split> and choose your answer for disjoint wisely. I will go with "False"

upvoted 43 times

✉  **auwia** 6 months, 3 weeks ago

From the link you've posted:

disjoint is false because the data goes to the first matching condition rather than all matching conditions.

So the correct answer is True, considering we have to "duplicate" records for the ALL category.

upvoted 2 times

✉  **dev2dev** 1 year, 11 months ago

you also need to read question to understand requirement. I will choose disjoint: true

upvoted 3 times

✉  **alphilla** Most Recent 3 weeks, 2 days ago

Guys Disjoint is True 110% and I will tell you why.

disjoint: false means that rows will be directed to the first branch whose condition is satisfied, and subsequent conditions are ignored. This might not fulfill the requirement because you want to process data for multiple departments, and with disjoint: false, a row would go to the first department branch it satisfies, ignoring the other departments.

Disjoint TRUE is more appropriate because it fulfills the requirement of processing data for individual departments (Ecommerce, retail, and wholesale) while also handling data for the entire company. Because all rows will match 2 conditions:

1st condition. They will have one of the three depts

2nd Condition. They will match the all condition

That's why it MUST BE TRUE.

upvoted 1 times

✉  **kkk5566** 4 months, 2 weeks ago

False is correct

upvoted 1 times

✉  **orionduo** 6 months, 3 weeks ago

I think the disjoint should be 'False'

By setting "disjoint true" for activities in a pipeline, you are essentially indicating that these activities are independent and can be executed concurrently. This can help improve the overall performance and efficiency of the pipeline by allowing for parallel execution of activities that do not have any interdependencies.

upvoted 1 times

✉  **bakamon** 7 months, 2 weeks ago

CleanData split(dept=='ecommerce', dept=='retail', dept=='wholesale') ~> SplitByDept@(disjoint: false)

This will split the data by department and allow for processing of data for the entire company as well as for individual departments.

upvoted 2 times

✉  **bakamon** 7 months, 2 weeks ago

The disjoint option in a split transformation determines whether the output streams are mutually exclusive or not. If disjoint is set to true, then each row of data can only be sent to one output stream. If disjoint is set to false, then a single row of data can be sent to multiple output streams.

In this case, setting disjoint to false allows for data to be processed for the entire company as well as for individual departments. This means that a single row of data can be sent to multiple output streams, allowing for processing at both the department and company level.

upvoted 2 times

✉ **markpumc** 9 months, 4 weeks ago

disjoin = true if you want all , if disjoint = false, nothing in ALL split

upvoted 3 times

✉ **DPMishra** 11 months, 1 week ago

Disjoint=False

upvoted 1 times

✉ **DindaS** 11 months, 3 weeks ago

disjoint=false

The below example is a conditional split transformation named SplitByYear that takes in incoming stream CleanData. This transformation has two split conditions year < 1960 and year > 1980. disjoint is false because the data goes to the first matching condition rather than all matching conditions. Every row matching the first condition goes to output stream moviesBefore1960. All remaining rows matching the second condition go to output stream moviesAfter1980. All other rows flow through the default stream AllOtherMovies.

from <https://learn.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split>

upvoted 4 times

✉ **nadahef** 1 year ago

Given answer correct

upvoted 2 times

✉ **Maddhy** 1 year, 1 month ago

The given answer is 100000% crct, don't confuse with others

upvoted 2 times

✉ **Aslam208** 1 year, 4 months ago

Given answer is 100% correct

upvoted 6 times

✉ **kiranSargar** 1 year, 7 months ago

Everyone is discussing about disjoint. But if disjoint is true then there is no ordering required of ecommerce, retail, wholesale, all .so we can fill 1st option with 2 or 3 and 3rd option with 1 or 6.

upvoted 2 times

✉ **nefarious_smalls** 1 year, 8 months ago

I think it should be disjoint is True based on microsofts example. it states that when disjoint is false each row will only go to the first matching condition. However in the example I believe each row will go to its matching department plus an aggregate stream that takes in every value regardless. Hence disjoint should be true

upvoted 1 times

✉ **Andushi** 1 year, 8 months ago

Definately Disjoint=True as per Microsoft doc

upvoted 2 times

✉ **kilowd** 1 year, 11 months ago

Answer: Disjoint=False

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split>

Example

The below example is a conditional split transformation named SplitByYear that takes in incoming stream CleanData. This transformation has two split conditions year < 1960 and year > 1980. disjoint is false because the data goes to the first matching condition. Every row matching the first condition goes to output stream moviesBefore1960. All remaining rows matching the second condition go to output stream moviesAfter1980. All other rows flow through the default stream AllOtherMovies.

upvoted 2 times

✉ **Onobhas01** 1 year, 9 months ago

The example has true as the data matches only one condition, it's either before 1960, after 1980 or Else... no two dataset matches more than one condition. But in the question they match more than one condition so disjoint has to be true.

upvoted 3 times

✉ **Onobhas01** 1 year, 9 months ago

sorry I meant to start off by saying "The example has false"

upvoted 1 times

✉ **kilowd** 1 year, 11 months ago

Disjoint = True

If true then split on all matching conditions, if false then only split on the first matching condition.

upvoted 3 times

 **PallaviPatel** 1 year, 11 months ago

I agree with yolap31172.

upvoted 1 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

DRAG DROP -

You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers. The file contains two attributes named FirstName and LastName.

You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the FirstName and LastName values.

You create the following components:

- A destination table in Azure Synapse
- An Azure Blob storage container
- A service principal

Which five actions should you perform in sequence next in is Databricks notebook? 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

- Mount the Data Lake Storage onto DBFS.
- Write the results to a table in Azure Synapse.
- Perform transformations on the file.
- Specify a temporary folder to stage the data.
- Write the results to Data Lake Storage.
- Read the file into a data frame.
- Drop the data frame.
- Perform transformations on the data frame.

Answer Area**Correct Answer:****Actions**

- Mount the Data Lake Storage onto DBFS.
- Write the results to a table in Azure Synapse.
- Perform transformations on the file.
- Specify a temporary folder to stage the data.
- Write the results to Data Lake Storage.
- Read the file into a data frame.
- Drop the data frame.
- Perform transformations on the data frame.

Answer Area

- Mount the Data Lake Storage onto DBFS.
- Read the file into a data frame.
- Perform transformations on the data frame.
- Specify a temporary folder to stage the data.
- Write the results to a table in Azure Synapse.

Step 1: Mount the Data Lake Storage onto DBFS

Begin with creating a file system in the Azure Data Lake Storage Gen2 account.

Step 2: Read the file into a data frame.

You can load the json files as a data frame in Azure Databricks.

Step 3: Perform transformations on the data frame.

Step 4: Specify a temporary folder to stage the data

Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse.

Step 5: Write the results to a table in Azure Synapse.

You upload the transformed data frame into Azure Synapse. You use the Azure Synapse connector for Azure Databricks to directly upload a dataframe as a table in a Azure Synapse.

Reference:

<https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-extract-load-sql-data-warehouse>

I think the correct order is:

- 1) mount onto DBFS
- 2) read into data frame
- 3) transform data frame
- 4) specify temporary folder
- 5) write to table in SQL data warehouse

About temporary folder, there is a note explain this:

<https://docs.microsoft.com/en-us/azure/databricks/scenarios/databricks-extract-load-sql-data-warehouse#load-data-into-azure-synapse>

Discussions about this question:

<https://www.examtopics.com/discussions/microsoft/view/11653-exam-dp-200-topic-2-question-30-discussion/>

upvoted 169 times

✉ **satyamkishoresingh** 2 years, 4 months ago

This order is absolutely correct.

upvoted 3 times

✉ **andylop04** 2 years, 6 months ago

Today I received this question in my exam. Only appeared the 5 options of this response. I only had to order, not choice. This solutions is the correct. Thanks sagga.

upvoted 38 times

✉ **Kingllo** 2 years, 1 month ago

Correct also received the only five options.

Also see:

<https://www.examtopics.com/discussions/microsoft/view/11653-exam-dp-200-topic-2-question-30-discussion/>

upvoted 2 times

✉ **snna4** 1 year, 11 months ago

OMG... the 5th step should be "Write the results to a table in Azure synapse". Who are those people "liked" this answer? Guys, just read the task.

upvoted 8 times

✉ **dev2dev** 1 year, 11 months ago

hehe, those who understand sql dw = azure synapse :D

upvoted 35 times

✉ **kkk5566** 4 months ago

it is the incorrect answer

upvoted 1 times

✉ **labasmuse** 2 years, 8 months ago

Hi sagga! Thank you. I do agree....

upvoted 2 times

✉ **InvisibleShadow** 2 years, 6 months ago

fix solution on site

upvoted 3 times

✉ **Miris** Highly Voted 2 years, 7 months ago

- 1) mount the data onto DBFS
- 2) Read the file into a data frame
- 3) Perform transformations on the file
- 4) Specify a temporary folder to stage the data
- 5) Write the results to a table in Azure synapse

upvoted 21 times

✉ **Tickxit** 1 year, 1 month ago

transformations on dataframe, not on the file.

upvoted 5 times

✉ **Momoanwar** Most Recent 1 month ago

Answer are correct, chatgpt say :

To accomplish the task in an Azure Databricks notebook, the logical sequence of actions would be:

1. **Mount the Data Lake Storage onto DBFS**: This allows access to the JSON file stored in Azure Data Lake Storage using the Databricks File System.
2. **Read the file into a data frame**: Use Spark to read the JSON file into a DataFrame for processing.
3. **Perform transformations on the data frame**: Apply transformations to concatenate the FirstName and LastName fields to create a new column.
4. **Specify a temporary folder to stage the data**: Before writing the data to Azure Synapse, it is a common practice to stage it in a temporary folder.
5. **Write the results to a table in Azure Synapse**: Finally, write the transformed DataFrame to the destination table in Azure Synapse Analytics.

These steps would ensure the JSON file data is properly transformed and loaded into Azure Synapse Analytics for further use.

upvoted 1 times

□ **EliteAllen** 1 month, 2 weeks ago

Just remember the initials first: M.R.P.S.W then go to the details.

upvoted 1 times

□ **bakamon** 7 months, 2 weeks ago

1. Mount the data lake storage onto DBFS.
2. Read the file into a data frame.
3. Perform transformations on the data frame.
4. Specify a temporary folder to stage the data.
5. Write the results to a table in Azure Synapse.

This will allow you to read the data from the JSON file into a data frame, perform the necessary transformations to concatenate the FirstName and LastName values, and then write the results to a table in Azure Synapse.

upvoted 3 times

□ **Deeksha1234** 1 year, 5 months ago

answer is correct, explained by the reference link in the given solution

upvoted 1 times

□ **carloalbe** 1 year, 8 months ago

I don not see the reason why "specify temporary folder" can not be both before or after the "read and transformation phase"

upvoted 3 times

□ **Davico93** 1 year, 6 months ago

I want to know the reason too!

upvoted 1 times

□ **Egocentric** 1 year, 8 months ago

given answer is correct, after reading and rereading stand with the given answer

upvoted 1 times

□ **Sandip4u** 2 years ago

I think the correct order is:

- 1) mount onto DBFS
- 2) read into data frame
- 3) transform data frame
- 4) specify temporary folder
- 5) write to table in SQL data warehouse

upvoted 3 times

□ **Canary_2021** 2 years ago

Here is my answer.

1) Create a service principal - Not sure why this step is not a choice in this question. I don't thing need to mount onto DBFS, but you do need to assign permission to allow databricks talk with Data Lake and read file.

2) Read the file into data frame

3) Perform transformations on the data frame

Data have been read into data from, so should transform data from data frame, not data file.

4) Specify temporary folder to stage the data

5) Write the results to a table in Azure Synapse

I reviewed this online document. No any place mentioned that the data frame needs to be dropped.

<https://docs.microsoft.com/en-us/azure/databricks/scenarios/databricks-extract-load-sql-data-warehouse>

upvoted 2 times

□ **Gina8008** 1 year, 10 months ago

you do not need to create a service principal, this is already exist

upvoted 1 times

□ **Sayour** 2 years ago

There Is A Contradiction Between Answers On The Drag & Drop And The Answers In The Steps Listing, And I ThinkThe Correct Ones Are That In The Listing And Not The Drag & Drop.

upvoted 1 times

□ **VJPR** 2 years ago

- 1) Mount the data onto DBFS
- 2) Read the file into a data frame
- 3) Perform transformations
- 4) Specify a temporary folder to stage the data
- 5) Write the results to a table in Azure synapse

upvoted 1 times

□ **[Removed]** 2 years, 4 months ago

The given answer is correct, after read the link provided carefully several times. There's already a service principal. With that, it's no need to mount. You do need to drop the dataframe as the last step.

upvoted 1 times

□ **GameLift** 2 years, 4 months ago

Service Principal has nothing to do with DataBricks.

upvoted 4 times

□ **nefarious_smalls** 1 year, 8 months ago

Actually you can assign a service principal to any data bricks account and use OAuth to connect with its tenant id app secret, and app id. You can then mount the data lake to databricks.

upvoted 1 times

□ **labasmuse** 2 years, 8 months ago

Correct solution:

Read the file into a data frame

Perform transformations on the file

Specify a temporary folder to stage the data

Write the results to a table in Azure synapse

Drop the data frame

upvoted 5 times

□ **ThiruthuvaRajan** 2 years, 7 months ago

you should not perform transformation on the file.

You need not to drop the dataframe.

sagga options are correct

upvoted 3 times

□ **Wisenut** 2 years, 7 months ago

I believe you perform transformation on the data frame and not on the file

upvoted 6 times

□ **hello2tomoki** 1 year, 9 months ago

Step 1: Read the file into a data frame.

You can load the json files as a data frame in Azure Databricks.

Step 2: Perform transformations on the data frame.

Step 3: Specify a temporary folder to stage the data

Specify a temporary folder to use while moving data between Azure Databricks and Azure SQL Data Warehouse.

Step 4: Write the results to a table in Azure synapse

Step 5: Drop the data frame - Clean up resources.

<https://www.examtopics.com/discussions/microsoft/view/11653-exam-dp-200-topic-2-question-30-discussion/>

upvoted 3 times

□ **kkk5566** 4 months ago

Correct

upvoted 1 times

HOTSPOT -

You build an Azure Data Factory pipeline to move data from an Azure Data Lake Storage Gen2 container to a database in an Azure Synapse Analytics dedicated SQL pool.

Data in the container is stored in the following folder structure.

/in/{YYYY}/{MM}/{DD}/{HH}/{mm}

The earliest folder is /in/2021/01/01/00/00. The latest folder is /in/2021/01/15/01/45.

You need to configure a pipeline trigger to meet the following requirements:

- Existing data must be loaded.
- Data must be loaded every 30 minutes.
- Late-arriving data of up to two minutes must be included in the load for the time at which the data should have arrived.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Type:

Event
On-demand
Schedule
Tumbling window

Additional properties:

Prefix: /in/, Event: Blob created
Recurrence: 30 minutes, Start time: 2021-01-01T00:00
Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes
Recurrence: 32 minutes, Start time: 2021-01-15T01:45

Correct Answer:**Answer Area**

Type:

Event
On-demand
Schedule
Tumbling window

Additional properties:

Prefix: /in/, Event: Blob created
Recurrence: 30 minutes, Start time: 2021-01-01T00:00
Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes
Recurrence: 32 minutes, Start time: 2021-01-15T01:45

Box 1: Tumbling window -

To be able to use the Delay parameter we select Tumbling window.

Box 2:

Recurrence: 30 minutes, not 32 minutes

Delay: 2 minutes.

The amount of time to delay the start of data processing for the window. The pipeline run is started after the expected execution time plus the amount of delay.

The delay defines how long the trigger waits past the due time before triggering a new run. The delay doesn't alter the window startTime.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger>

④  **Puneetgupta003** Highly Voted 2 years, 6 months ago

Answers are correct
upvoted 49 times

④  **positivitypeople** Most Recent 2 weeks, 5 days ago

Got this question today on the exam
upvoted 3 times

④  **kkk5566** 4 months ago

Answers are correct
upvoted 1 times

④  **Deeksha1234** 1 year, 4 months ago

correct
upvoted 4 times

④  **StudentFromAus** 1 year, 6 months ago

Answers are correct
upvoted 2 times

④  **parx** 1 year, 9 months ago

Correct. <https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers#trigger-type-comparison>
upvoted 4 times

④  **azurearmy** 2 years, 2 months ago

Why can't we use an event-based trigger here?
upvoted 2 times

because we also wanna do backfill with past data. Technically, the event-based trigger will also allow ADF to find all the old files from the source which ADF hasn't processed yet (and we could add a datetime filter when loading the data) but ADF is gonna choke on so many past events from experience. With tumbling windows, the trigger will kick off for each 30 minutes slices of the time span, emulating batch loads. be very careful when doing backfill with a tumbling window, by default, ADF will start 50 concurrent pipelines, it can be pricey, change the settings in advanced panel of the trigger creation form.

upvoted 10 times

④  **belha** 2 years, 6 months ago

not schedule ?
upvoted 2 times

④  **captainbee** 2 years, 6 months ago

As the solution says, you cannot use the Delay with Schedule.
upvoted 7 times

④  **escoins** 2 years, 6 months ago

why not schedule trigger?
upvoted 1 times

④  **Podavenna** 2 years, 4 months ago

Schedule trigger would not work because backfill is only possible with Tumbling window trigger. In this case, we need to use trigger for old data.
upvoted 7 times

HOTSPOT -

You are designing a near real-time dashboard solution that will visualize streaming data from remote sensors that connect to the internet. The streaming data must be aggregated to show the average value of each 10-second interval. The data will be discarded after being displayed in the dashboard.

The solution will use Azure Stream Analytics and must meet the following requirements:

- Minimize latency from an Azure Event hub to the dashboard.
- Minimize the required storage.
- Minimize development effort.

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

Azure Stream Analytics input type:

Azure Event Hub
Azure SQL Database
Azure Stream Analytics
Microsoft Power BI

Azure Stream Analytics output type:

Azure Event Hub
Azure SQL Database
Azure Stream Analytics
Microsoft Power BI

Aggregation query location:

Azure Event Hub
Azure SQL Database
Azure Stream Analytics
Microsoft Power BI

Answer Area

Azure Stream Analytics input type:

Azure Event Hub
Azure SQL Database
Azure Stream Analytics
Microsoft Power BI

Azure Stream Analytics output type:

Azure Event Hub
Azure SQL Database
Azure Stream Analytics
Microsoft Power BI

Aggregation query location:

Azure Event Hub
Azure SQL Database
Azure Stream Analytics
Microsoft Power BI

Correct Answer:

店铺：IT认证考试服务

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-power-bi-dashboard>

✉ **Sunnyb** Highly Voted 2 years, 7 months ago

Answer is correct

upvoted 66 times

✉ **captainbee** 2 years, 7 months ago

Agreed. So easy that even ExamTopics got it right.

upvoted 151 times

✉ **eskimolight** 5 months, 1 week ago

I badly needed a laugh after studying...LOL

upvoted 2 times

✉ **[Removed]** 2 years, 3 months ago

The best comment ever :)

upvoted 13 times

✉ **shaileshutd** 1 year, 1 month ago

super like for this comment

upvoted 4 times

✉ **positivitypeople** Most Recent 2 weeks, 5 days ago

Got this question today on the exam

upvoted 2 times

✉ **kkk5566** 4 months ago

correct

upvoted 1 times

✉ **nadahef** 1 year ago

We agree on this

upvoted 2 times

✉ **Deeksha1234** 1 year, 5 months ago

Correct solution

upvoted 2 times

✉ **Boompiee** 1 year, 8 months ago

Correct. Question so easy I wonder if it was really in the exam.

upvoted 4 times

✉ **paras_gadhiya** 1 year, 10 months ago

COrrECT

DRAG DROP -

You have an Azure Stream Analytics job that is a Stream Analytics project solution in Microsoft Visual Studio. The job accepts data generated by IoT devices in the JSON format.

You need to modify the job to accept data generated by the IoT devices in the Protobuf format.

Which three actions should you perform from Visual Studio on sequence? 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**

- Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.
- Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.
- Add .NET deserializer code for Protobuf to the custom deserializer project.
- Add .NET deserializer code for Protobuf to the Stream Analytics project.
- Add an Azure Stream Analytics Application project to the solution.

店铺：IT认证考试服务

Correct Answer:**Actions****Answer Area**

- Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.
- Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.
- Add .NET deserializer code for Protobuf to the custom deserializer project.
- Add .NET deserializer code for Protobuf to the Stream Analytics project.
- Add an Azure Stream Analytics Application project to the solution.

- Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.
- Add .NET deserializer code for Protobuf to the custom deserializer project.
- Add an Azure Stream Analytics Application project to the solution.

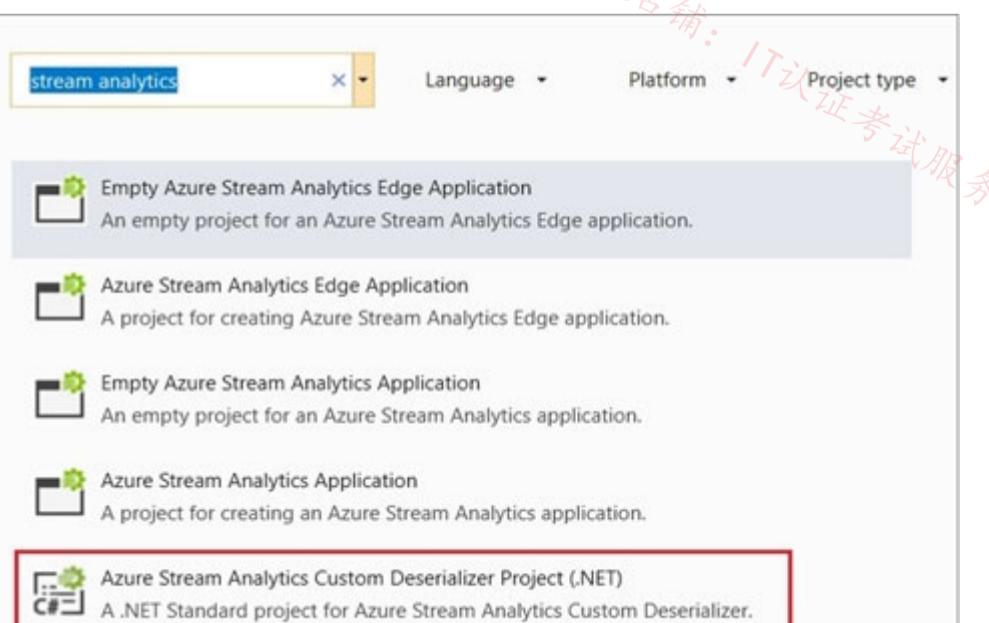
Step 1: Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.

Create a custom deserializer -

1. Open Visual Studio and select File > New > Project. Search for Stream Analytics and select Azure Stream Analytics Custom Deserializer Project (.NET). Give the project a name, like Protobuf Deserializer.

Create a new project**Recent project templates**

A list of your recently accessed templates will be displayed here.



2. In Solution Explorer, right-click your Protobuf Deserializer project and select Manage NuGet Packages from the menu. Then install the Microsoft.Azure.StreamAnalytics and Google.Protobuf NuGet packages.

3. Add the MessageBodyProto class and the MessageBodyDeserializer class to your project.

4. Build the Protobuf Deserializer project.

Step 2: Add .NET deserializer code for Protobuf to the custom deserializer project

Azure Stream Analytics has built-in support for three data formats: JSON, CSV, and Avro. With custom .NET deserializers, you can read data from other formats such as Protocol Buffer, Bond and other user defined formats for both cloud and edge jobs.

Step 3: Add an Azure Stream Analytics Application project to the solution

Add an Azure Stream Analytics project

1. In Solution Explorer, right-click the Protobuf Deserializer solution and select Add > New Project. Under Azure Stream Analytics > Stream Analytics, choose

Azure Stream Analytics Application. Name it ProtobufCloudDeserializer and select OK.

2. Right-click References under the ProtobufCloudDeserializer Azure Stream Analytics project. Under Projects, add Protobuf Deserializer. It should be automatically populated for you.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/custom-deserializer>

zarga Highly Voted 2 years, 6 months ago

The third one is wrong because the stream analytics application already exist in the project.

The goal is to modify the current stream analytics application in order to read protobuf data.

I think the right answer is the first one in the list (update input.json file and reference dll)

upvoted 63 times

Marcus1612 2 years, 4 months ago

Absolutely: <https://docs.microsoft.com/en-us/azure/stream-analytics/custom-deserializer>

upvoted 6 times

kamil_k 1 year, 10 months ago

this is a tricky question.. technically document here <https://docs.microsoft.com/en-us/azure/stream-analytics/custom-deserializer> describes it as follows:

1. Add custom deserializer project
2. Add the MessageBodyProto class and the MessageBodyDeserializer class to your project (it's actually merged with point 1)
3. add an Azure Stream Analytics project
4. Configure a Stream Analytics job (including changing input.json)

Technically, the question asks for 3 steps so we might to either skip the "add code" as a separate step, and include point 4 as the last step (i.e. 1,3,4 above) or we stop at 3 and then the answer is as listed.. As usual, the description of the question is confusing.

upvoted 6 times

HaliBrickclay Highly Voted 2 years, 2 months ago

1 Add an Azure Stream Analytics Customer Deserializer Project(.net) project to the Solution

2 Add .net deseriliazer Code to ProtoBuf to customer deserializer project

3. Change the event Serialization format to protobuf in the input.json File of the job and reference the DLL.

upvoted 56 times

Momoanwar Most Recent 1 month ago

Chatgpt :

To modify an Azure Stream Analytics job in Visual Studio to accept data in Protobuf format from IoT devices, you would typically need to:

1. **Add an Azure Stream Analytics Custom Deserializer Project (.NET project) to the solution**: This sets up a project that can include the custom deserialization logic.

2. **Add .NET deserializer code for Protobuf to the custom deserializer project**: Here, you would implement the Protobuf deserialization logic within the project you added in the previous step.

3. **Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL**: Finally, you need to update the job configuration to use the custom deserializer by changing the serialization format and pointing it to the compiled DLL from your custom deserializer project.

These actions will enable the Azure Stream Analytics job to deserialize and process data in Protobuf format instead of JSON.

upvoted 3 times

Lscranio 1 month ago

1- Add Azure Stream Analyticst Cuson Deserializer Project (.NET) project to the Solution;

2 - Add an Azure Stream Analytics Application project to the solution;

3 - Change... "Iput.json" is necessary your modification;

upvoted 1 times

kkk5566 4 months ago

1. Add Azure Stream Analytics Custom Deserializer Project (.NET) 2. Add Azure Stream Analytics Application 3. Configure a Stream Analytics job in Input.json

upvoted 1 times

janaki 7 months, 2 weeks ago

Correct answer is:

Add an Azure Stream Analytics Customer Deserializer Project(.net) project to the Solution

Add .net deserializer Code to ProtoBuf to the customer deserializer project

Change the event Serialization format to protobuf in the input.json File of the job and reference the DLL

upvoted 3 times

✉ **bakamon** 7 months, 2 weeks ago

1. Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.

2. Add .NET deserializer code for Protobuf to the custom deserializer project.

3. Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.

This will allow you to create a custom deserializer project and add .NET deserializer code for Protobuf to it. Then, you can change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL containing the custom deserializer code.

upvoted 2 times

✉ **Rossana** 8 months, 2 weeks ago

Chat GPT: B-C-A

upvoted 3 times

✉ **hiyoww** 9 months, 2 weeks ago

guessing seem not for DP203, anyone agree?

upvoted 5 times

✉ **esaade** 10 months, 1 week ago

To modify the Azure Stream Analytics job to accept data generated by the IoT devices in the Protobuf format, follow these steps in sequence:

Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.

Add .NET deserializer code for Protobuf to the custom deserializer project.

Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.

upvoted 1 times

✉ **zorko10** 1 year, 2 months ago

According to the documentation:

1- Create a custom deserializer for protocol buffer.

2- Add an Azure Stream Analytics project

3- Configure a Stream Analytics job (in here you specify reference the dll...) ==> Change the event Serialization format to protobuf in the input.json File of the job and reference the DLL.

<https://learn.microsoft.com/en-us/azure/stream-analytics/custom-deserializer>

upvoted 3 times

✉ **Lscrario** 1 month ago

I Agree;

1- Add Azure Stream Analyticst Cuson Deserializer Project (.NET) project to the Solution;

2 - Add an Azure Stream Analytics Application project to the solution;

3 - Change... "Input.json" is necessary your modification;

upvoted 1 times

✉ **nadahef** 1 year, 2 months ago

As stated in the documentation :

1- Create a custom deserializer project

2- Add an azure stream alaytics project

3- Configure a stream analytics job, (in this configuration, the dll is referenced) ==> update input.json file and reference dll

<https://learn.microsoft.com/en-us/azure/stream-analytics/custom-deserializer>

upvoted 1 times

✉ **ExamTopicsAshwin** 1 year, 3 months ago

I am still practicing and have not seen the answer yet. Have put my answer as 3-2-1. Will see the right answer in the end.

upvoted 1 times

✉ **Franz58** 1 year, 5 months ago

As described in <https://docs.microsoft.com/en-us/azure/stream-analytics/custom-deserializer>:

1. Add Azure Stream Analytics Custom Deserializer Project (.NET)

2. Add Azure Stream Analytics Application

3. Configure a Stream Analytics job in Input.json

upvoted 7 times

✉ **Paulkuzzio** 6 months, 2 weeks ago

Yes, I think your answer is correct following this link : <https://learn.microsoft.com/en-us/azure/stream-analytics/custom-deserializer>

upvoted 1 times

✉ **Johno1393** 2 years, 2 months ago

Has this question come up in the DP-203 exam?

upvoted 15 times

✉ **[Removed]** 2 years, 4 months ago

Third one should be the first action listed: Change file format in input.json

upvoted 4 times

 Gowthamr02 2 years, 7 months ago

Correct!

upvoted 1 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

You have an Azure Storage account and a data warehouse in Azure Synapse Analytics in the UK South region.

You need to copy blob data from the storage account to the data warehouse by using Azure Data Factory. The solution must meet the following requirements:

- ⇒ Ensure that the data remains in the UK South region at all times.
- ⇒ Minimize administrative effort.

Which type of integration runtime should you use?

- A. Azure integration runtime
- B. Azure-SSIS integration runtime
- C. Self-hosted integration runtime

Correct Answer: A

IR type	Public network	Private network
Azure	Data Flow Data movement Activity dispatch	
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

Incorrect Answers:

C: Self-hosted integration runtime is to be used On-premises.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime>

Community vote distribution

A (100%)

✉  **zarga** Highly Voted 2 years, 6 months ago

A is the right answer (don't use autoresolve region)
upvoted 43 times

✉  **SomethingRight100** 1 year, 1 month ago

Here I found in the docs <https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime>.
You can set the location region of an Azure IR, in which case the activity execution or dispatch will happen in the selected region

Self-hosted integration runtime can achieve the same goal with higher administrative effort

upvoted 1 times

✉  **kishorenayak** Highly Voted 2 years, 6 months ago

Should not this be option A??

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime>

"If you have strict data compliance requirements and need ensure that data do not leave a certain geography, you can explicitly create an Azure IR in a certain region and point the Linked Service to this IR using ConnectVia property. For example, if you want to copy data from Blob in UK South to Azure Synapse Analytics in UK South and want to ensure data do not leave UK, create an Azure IR in UK South and link both Linked Services to this IR."

upvoted 14 times

✉  **janaki** 7 months, 2 weeks ago

Yes, it is Azure integration runtime - option A

upvoted 2 times

✉  **Dicupillo** 2 years, 6 months ago

Yes it's option A

upvoted 2 times

 **kkk5566** Most Recent 4 months ago

Selected Answer: A

A is correct

upvoted 1 times

 **esaade** 10 months, 1 week ago

To ensure that the data remains in the UK South region and minimize administrative effort while copying blob data from the storage account to the data warehouse by using Azure Data Factory, you should use the Azure integration runtime.

upvoted 3 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: A

A is correct

upvoted 2 times

 **Fishy_Marcy** 1 year, 6 months ago

I think the first requirement isn't adding much to the equation, so it is primarily focussed on administration which is lowest with A
upvoted 2 times

 **StudentFromAus** 1 year, 6 months ago

Selected Answer: A

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime>

upvoted 2 times

 **DrTaz** 2 years ago

Selected Answer: A

Why would I want to move data to a local host then back to cloud? That sounds a bit unwise, eh?

upvoted 4 times

 **rashjan** 2 years, 1 month ago

Selected Answer: A

A is correct

upvoted 1 times

 **hsetin** 2 years, 4 months ago

A it is.

upvoted 2 times

 **saty_nl** 2 years, 6 months ago

Correct answer.

upvoted 2 times

 **damaldon** 2 years, 6 months ago

fully agree

upvoted 1 times

 **Sunnyb** 2 years, 7 months ago

A is correct

upvoted 2 times

HOTSPOT -

You have an Azure SQL database named Database1 and two Azure event hubs named HubA and HubB. The data consumed from each source is shown in the following table.

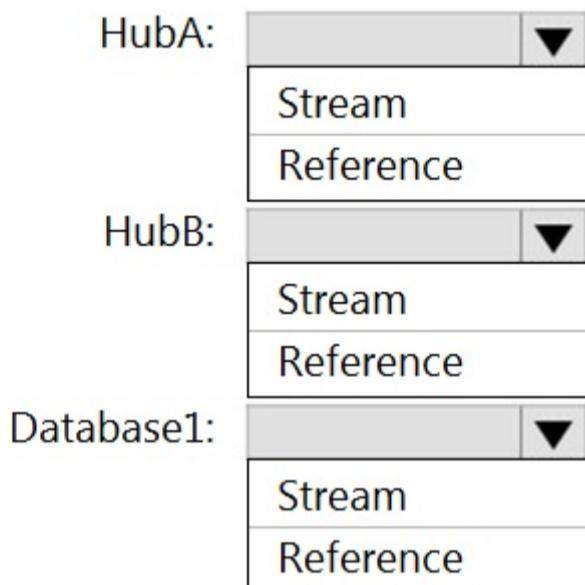
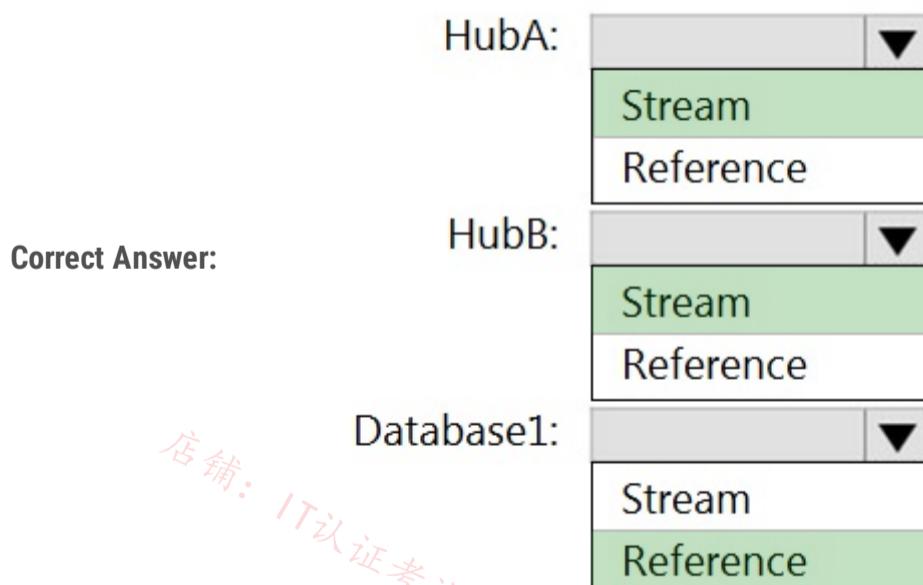
Source	Data
Database1	Driver's name Driver's license number
HubA	Ride route Ride distance Ride duration
HubB	Ride fare Ride payment

You need to implement Azure Stream Analytics to calculate the average fare per mile by driver.

How should you configure the Stream Analytics input for each source? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area**Answer Area**

HubA: Stream -

HubB: Stream -

Database1: Reference -

Reference data (also known as a lookup table) is a finite data set that is static or slowly changing in nature, used to perform a lookup or to augment your data streams. For example, in an IoT scenario, you could store metadata about sensors (which don't change often) in reference data and join it with real time IoT data streams. Azure Stream Analytics loads reference data in memory to achieve low latency stream processing

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data>

□  **Sunnyb**  2 years, 7 months ago

Answer is correct
upvoted 42 times

□  **Jaws1990**  1 year, 12 months ago

Crap question. With that data, how are you supposed to link the stream data with the reference data.
upvoted 14 times

□  **kkk5566**  4 months ago

correct
upvoted 1 times

□  **mafragias** 9 months, 1 week ago

Stream, STream, Reference
upvoted 3 times

□  **zorko10** 1 year, 2 months ago

Correct
Data stream input : is an unbounded sequence of events over time.
Reference Data input : Reference data is either completely static or changes slowly

<https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-add-inputs>
upvoted 4 times

□  **Deeksha1234** 1 year, 5 months ago

correct
upvoted 2 times

□  **wolf74** 1 year, 6 months ago

why hubA, that contains driver's name, should be processed as stream?? it's a dimension type and should be processed as reference in my opinion
upvoted 3 times

□  **dmitriypo** 1 year, 2 months ago

HubA doesn't contain driver's name. It is Database1 that contains driver's name.
upvoted 4 times

□  **gssd4scoder** 2 years, 2 months ago

Correct: <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-add-inputs>
upvoted 2 times

□  **GameLift** 2 years, 4 months ago

I doubt if these questions are really what they asking in the real exam.
upvoted 4 times

□  **Vanq69** 3 months ago

I mean they can't just ask hard questions where you need 5min to think about, no one would pass that. Also they want the certificate to be worth something but also have many people passing the exam and go for Azure instead of AWS, GCP.
upvoted 1 times

□  **[Removed]** 2 years, 3 months ago

it could be real, he is asking if you can realize the main difference between the real-time data vs the reference data and so you can choose the best service for each one
upvoted 4 times

□  **[Removed]** 2 years, 3 months ago

I have tried some pretty easy question like this one before in prev exams
upvoted 3 times

You have an Azure Stream Analytics job that receives clickstream data from an Azure event hub.

You need to define a query in the Stream Analytics job. The query must meet the following requirements:

- ⇒ Count the number of clicks within each 10-second window based on the country of a visitor.
- ⇒ Ensure that each click is NOT counted more than once.

How should you define the Query?

- A. SELECT Country, Avg(*) AS Average FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SlidingWindow(second, 10)
- B. SELECT Country, Count(*) AS Count FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, TumblingWindow(second, 10)
- C. SELECT Country, Avg(*) AS Average FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, HoppingWindow(second, 10, 2)
- D. SELECT Country, Count(*) AS Count FROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SessionWindow(second, 5, 10)

Correct Answer: B

Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

Example:

Incorrect Answers:

A: Sliding windows, unlike Tumbling or Hopping windows, output events only for points in time when the content of the window actually changes. In other words, when an event enters or exits the window. Every window has at least one event, like in the case of Hopping windows, events can belong to more than one sliding window.

C: Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap, so events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.

D: Session windows group events that arrive at similar times, filtering out periods of time where there is no data.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

Community vote distribution

B (100%)

 **saty_nl** Highly Voted 2 years, 6 months ago

Correct answer.

upvoted 25 times

 **GameLift** Highly Voted 2 years, 4 months ago

keyword : do not overlap

upvoted 5 times

 **sdokmak** 1 year, 7 months ago

Not really, the other Count option doesn't overlap. But it does skip.

upvoted 1 times

 **kkk5566** Most Recent 4 months ago

Selected Answer: B

correct

upvoted 1 times

 **joponlu** 7 months, 2 weeks ago

Selected Answer: B

Yes B is correct!!!

upvoted 2 times

 **cale** 9 months, 1 week ago

Selected Answer: B

B is correct

upvoted 3 times

 **igormmpinto** 1 year, 2 months ago

Selected Answer: B

Correct

upvoted 2 times

 **allagowf** 1 year, 3 months ago

A,C are excluded easily by the AVG function, the D also excluded by the session size that is less the window size and this will result in overlapped windows.

B is left alone Hahaha just want to describe a different method to answer exam questions.

upvoted 4 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: B

B is correct

upvoted 3 times

 **rashjan** 2 years, 1 month ago

Selected Answer: B

B is correct

upvoted 2 times

 **gssd4scoder** 2 years, 2 months ago

Correct: <https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics>

upvoted 2 times

 **damaldon** 2 years, 6 months ago

Correct, Tumbling Window is needed to use periodic time intervals

upvoted 2 times

 **Gowthamr02** 2 years, 7 months ago

Correct!

upvoted 2 times

HOTSPOT -

You are building an Azure Analytics query that will receive input data from Azure IoT Hub and write the results to Azure Blob storage.

You need to calculate the difference in the number of readings per sensor per hour.

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 sensorId,
       growth = reading -
               LAG (reading) OVER (PARTITION BY sensorId
               LAST (hour,1))
               LEAD (hour,1))
```

FROM input

Correct Answer:**Answer Area**

```
SELECT sensorId,
       growth = reading -
               LAG (reading) OVER (PARTITION BY sensorId
               LIMIT DURATION (hour,1))
               OFFSET
               WHEN
```

FROM input

Box 1: LAG -

The LAG analytic operator allows one to look up a previous event in an event stream, within certain constraints. It is very useful for computing the rate of growth of a variable, detecting when a variable crosses a threshold, or when a condition starts or stops being true.

Box 2: LIMIT DURATION -

Example: Compute the rate of growth, per sensor:

SELECT sensorId,

growth = reading -

LAG(reading) OVER (PARTITION BY sensorId LIMIT DURATION(hour, 1))

FROM input -

Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/lag-azure-stream-analytics>

 **ANath** Highly Voted 1 year, 11 months ago

The answer is correct

upvoted 17 times

 **UristMcFarmer** Highly Voted 12 months ago

The question and answer do not match. The question asks for "the difference in THE NUMBER OF READINGS per sensor per hour". The answer given is to compute the difference between the current sensor reading and the sensor reading from an hour ago.

upvoted 7 times

 **Momoanwar** Most Recent 1 month ago

Correct, cgatpgt:

To calculate the difference in the number of readings per sensor per hour using an Azure Stream Analytics query, you would use the LAG function to access the previous value and then calculate the difference. Here is how you would complete the query:

- Use `LAG` to get the previous reading.
- Use `LIMIT DURATION` to set the window of time for comparison, which in this case is per hour.

The completed query would look something like this:

```
```sql
SELECT sensorId,
 reading - LAG(reading) OVER (PARTITION BY sensorId LIMIT DURATION(hour, 1)) AS growth
FROM input
```
```

This query assumes `reading` is the column holding the sensor data and `sensorId` is the column to partition the data by each sensor. The `LAG` function gets the last reading for the same sensor from the previous hour, and then you subtract this value from the current reading to find the growth.

upvoted 1 times

□ **kkk5566** 4 months ago

should be correct

upvoted 1 times

□ **Rajan191083** 7 months, 3 weeks ago

LAG is the correct answer. Refer the below link. It mentions this example

<https://learn.microsoft.com/en-us/stream-analytics-query/lag-azure-stream-analytics>

upvoted 4 times

□ **Deeksha1234** 1 year, 5 months ago

right answer

upvoted 4 times

□ **Muishkin** 1 year, 8 months ago

what about LAST?

upvoted 1 times

□ **hbad** 1 year, 8 months ago

LAST is possible, however the code sample in this question does not include any WHEN syntax, so that will rule out LAST

upvoted 4 times

□ **UzairMir** 5 months, 3 weeks ago

When clause is optional

<https://learn.microsoft.com/en-us/stream-analytics-query/last-azure-stream-analytics>

Still cannot understand the difference between LAG and LAST :(

upvoted 1 times

□ **Ram9198** 5 months ago

LAST is the most recent event - literally opposite to the name.. that is why they take LAG which gives the previous event

upvoted 3 times

□ **onyerleft** 2 years ago

Answers as revealed are for computing the rate of growth per sensor. <https://docs.microsoft.com/en-us/stream-analytics-query/lag-azure-stream-analytics#examples>

upvoted 5 times

□ **bubububox** 2 years ago

yep, but the question here is unclear

upvoted 6 times

You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container. Which type of trigger should you use?

- A. on-demand
- B. tumbling window
- C. schedule
- D. event

Correct Answer: D.

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.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger>

Community vote distribution

D (100%)

 Miris Highly Voted 2 years, 7 months ago

correct

upvoted 37 times

 anto69 1 year, 11 months ago

hundered percent :)

upvoted 2 times

 positivitypeople Most Recent 2 weeks, 5 days ago

Got this question today on the exam

upvoted 1 times

 Liv3 3 months, 1 week ago

Selected Answer: D

Event-driven

upvoted 1 times

 kkk5566 4 months ago

Selected Answer: D

correct

upvoted 1 times

 akhil5432 5 months ago

Selected Answer: D

Option D EVENT

upvoted 1 times

 GodfreyMbizo 11 months, 2 weeks ago

Event grid

upvoted 2 times

 Deeksha1234 1 year, 5 months ago

correct

upvoted 1 times

 dsp17 1 year, 6 months ago

Selected Answer: D

Correct. One question related to Event-driven architecture (EDA) is must in exam.

upvoted 4 times

 sarapaisley 1 year, 9 months ago

Selected Answer: D

correct

upvoted 2 times

 **rashjan** 2 years, 1 month ago

Selected Answer: D

D is correct

upvoted 4 times

 **gssd4scoder** 2 years, 2 months ago

very complex... lol

upvoted 1 times

 **damaldon** 2 years, 6 months ago

Fully agree

upvoted 2 times

You have two Azure Data Factory instances named ADFdev and ADFprod. ADFdev connects to an Azure DevOps Git repository.

You publish changes from the main branch of the Git repository to ADFdev.

You need to deploy the artifacts from ADFdev to ADFprod.

What should you do first?

- A. From ADFdev, modify the Git configuration.
- B. From ADFdev, create a linked service.
- C. From Azure DevOps, create a release pipeline.
- D. From Azure DevOps, update the main branch.

Correct Answer: C

In Azure Data Factory, continuous integration and delivery (CI/CD) means moving Data Factory pipelines from one environment (development, test, production) to another.

Note: The following is a guide for setting up an Azure Pipelines release that automates the deployment of a data factory to multiple environments.

1. In Azure DevOps, open the project that's configured with your data factory.
2. On the left side of the page, select Pipelines, and then select Releases.
3. Select New pipeline, or, if you have existing pipelines, select New and then New release pipeline.
4. In the Stage name box, enter the name of your environment.
5. Select Add artifact, and then select the git repository configured with your development data factory. Select the publish branch of the repository for the Default branch. By default, this publish branch is adf_publish.
6. Select the Empty job template.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment>

Community vote distribution

C (100%)

 **damaldon** Highly Voted 2 years, 6 months ago

Correct!

upvoted 33 times

 **Gowthamr02** Highly Voted 2 years, 7 months ago

Answer in Correct!

upvoted 9 times

 **kkk5566** Most Recent 4 months ago

Selected Answer: C

C is correct

upvoted 1 times

 **allagowf** 1 year, 3 months ago

Selected Answer: C

correct

upvoted 4 times

 **Deeksha1234** 1 year, 5 months ago

correct

upvoted 2 times

 **Egocentric** 1 year, 8 months ago

answer is correct

upvoted 2 times

 **VeroDon** 2 years ago

Selected Answer: C

Correct

upvoted 2 times

 **rashjan** 2 years, 1 month ago

Selected Answer: C

pipeline is correct

upvoted 2 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

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 Cosmos DB
- B. Azure Blob storage
- C. Azure IoT Hub
- D. Azure Event Hubs

Correct Answer: B.

Stream Analytics supports Azure Blob storage and Azure SQL Database as the storage layer for Reference Data.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data>

Community vote distribution

B (100%)

 **trungngonpit** Highly Voted 2 years, 6 months ago

correct, blob storage or azure sql database

upvoted 33 times

 **Rob77** 7 months, 3 weeks ago

Correct <https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-add-inputs#reference-data-input>

upvoted 1 times

 **saty_nl** Highly Voted 2 years, 6 months ago

This is correct.

upvoted 6 times

 **kkk5566** Most Recent 4 months ago

Selected Answer: B

B is correct

upvoted 1 times

 **Deeksha1234** 1 year, 5 months ago

correct

upvoted 2 times

 **SebK** 1 year, 9 months ago

Selected Answer: B

Correct

upvoted 3 times

 **rashjan** 2 years, 1 month ago

Selected Answer: B

for reference data blob storage is correct (or also azure sql database)

upvoted 3 times

 **noranathalie** 2 years, 2 months ago

Why not connecting directly Event Hub? Answer D?

upvoted 1 times

 **Boompiee** 1 year, 8 months ago

Because the question is about reference data, not streaming data.

upvoted 1 times

 **berserksap** 2 years, 2 months ago

Since it is specified that Stream analytics is used , no need for event hub. It will just add cost. Correct me if I am wrong.

upvoted 3 times

You are designing an Azure Stream Analytics job to process incoming events from sensors in retail environments.

You need to process the events to produce a running average of shopper counts during the previous 15 minutes, calculated at five-minute intervals.

Which type of window should you use?

- A. snapshot
- B. tumbling
- C. hopping
- D. sliding

Correct Answer: C

Unlike tumbling windows, hopping windows model scheduled overlapping windows. A hopping window specification consists 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).

Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/hopping-window-azure-stream-analytics>

Community vote distribution

C (100%)

 **Guincimund** Highly Voted 1 year, 8 months ago

Selected Answer: C

Correct answer.

upvoted 16 times

 **BPW** Highly Voted 7 months, 3 weeks ago

Answer D

The running average is considered. So, it should be sliding window

upvoted 6 times

 **dakku987** Most Recent 1 week, 5 days ago

Selected Answer: C

For the scenario of producing a running average of shopper counts during the previous 15 minutes, calculated at five-minute intervals, you should use a:

C. Hopping Window
upvoted 1 times

 **kkk5566** 4 months ago

Selected Answer: C

correct

upvoted 1 times

 **auwia** 6 months, 3 weeks ago

Selected Answer: C

It is Hopping Window, see the picture and the green note:

<https://learn.microsoft.com/en-us/stream-analytics-query/hopping-window-azure-stream-analytics>

upvoted 3 times

 **janaki** 7 months, 2 weeks ago

Chat GPT says option D. sliding window

upvoted 3 times

 **bakamon** 7 months, 2 weeks ago

Selected Answer: C

A hopping window outputs events at a regular time interval (the hop size) and can be used to produce overlapping windows.

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

correct

upvoted 1 times

 **dsp17** 1 year, 6 months ago

Selected Answer: C

Correct. hoppingWindow(minute, 15, 5)

upvoted 2 times

 **Ajtk27** 1 year, 6 months ago

Selected Answer: C

Correct

upvoted 1 times

 **Saim8711** 1 year, 7 months ago

Correct Answer.

upvoted 2 times

 **Aurelkb** 1 year, 7 months ago

Correct

upvoted 3 times

HOTSPOT -

You are designing a monitoring solution for a fleet of 500 vehicles. Each vehicle has a GPS tracking device that sends data to an Azure event hub once per minute.

You have a CSV file in an Azure Data Lake Storage Gen2 container. The file maintains the expected geographical area in which each vehicle should be.

You need to ensure that when a GPS position is outside the expected area, a message is added to another event hub for processing within 30 seconds. The solution must minimize cost.

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

Service:

- An Azure Synapse Analytics Apache Spark pool
- An Azure Synapse Analytics serverless SQL pool
- Azure Data Factory
- Azure Stream Analytics

Window:

- Hopping
- No window
- Session
- Tumbling

Analysis type:

- Event pattern matching
- Lagged record comparison
- Point within polygon
- Polygon overlap

Answer Area

Service:

| |
|--|
| An Azure Synapse Analytics Apache Spark pool |
| An Azure Synapse Analytics serverless SQL pool |
| Azure Data Factory |
| Azure Stream Analytics |

Window:

| |
|-----------|
| Hopping |
| No window |
| Session |
| Tumbling |

Analysis type:

| |
|--------------------------|
| Event pattern matching |
| Lagged record comparison |
| Point within polygon |
| Polygon overlap |

Box 1: Azure Stream Analytics -

Box 2: Hopping -

Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.

Box 3: Point within polygon -

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

✉  Alekx42 Highly Voted 2 years, 7 months ago

You do not need a Window function. You just process the data and perform the geospatial check as it arrives. See the same example here:
<https://docs.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios>

upvoted 83 times

✉  Vanq69 3 months ago

The constraint is: "a message is added to another event hub for processing within 30 seconds".

The example you provided specifies that: "Devices can emit their ID and location every minute through a stream called DeviceStreamInput" so does DeviceStreamInput only emit every minute or can you adapt it? Else you would need a window.

upvoted 1 times

✉  isuru89 2 years ago

It works, But not the most const effective.

upvoted 8 times

✉  captainbee 2 years, 7 months ago

That's what I thought, there's no reporting over time periods. It's just a case of when this happens, ping it off.

upvoted 6 times

✉  JackArmitage Highly Voted 2 years, 6 months ago

1. Azure Stream Analytics
2. No Window
3. Point within Polygon

upvoted 77 times

✉  blazy001 Most Recent 3 weeks, 3 days ago

1. Azure Stream Analytics
2. Sliding (change on event) or No window for me
3. Point within Polygon

upvoted 1 times

✉  Momoanwar 1 month ago

Wrong, chatgpt :

- **Service**: Azure Stream Analytics – This service is ideal for processing real-time streaming data from IoT devices.
- **Window**: No window – As you're processing each GPS event as it arrives, you don't need to aggregate over a time window but process each event individually to check if it's within the expected area.
- **Analysis type**: Point within polygon – This analysis type is used for geospatial analytics, where you're checking if a point (the GPS position) is within a predefined polygon (the expected geographical area for the vehicle).

upvoted 2 times

□ **ThisAlreadyTaken** 4 months, 4 weeks ago

It is clear tumbling window guys.
requirements:-

- 1) new data created in every 1 min
- 2) if any out of range problem is there then need to send with in 30 sec.
so window size is 1 min, we will have 1 min of data., so we will not miss data.
- hop size will be 30 sec, so we can process the data in 30 sec and report the problem if any.

upvoted 2 times

□ **janaki** 7 months, 2 weeks ago

Correct answers are:

1. Azure Stream analytics
2. Tumbling window
3. Point within Polygon

upvoted 7 times

□ **akk_1289** 1 year ago

It sounds like you want to use Azure Stream Analytics for this task. Stream Analytics is a real-time analytics service that allows you to analyze and process high volumes of streaming data from various sources, such as Azure Event Hubs.

For the window, you should use a Tumbling window. A tumbling window is a fixed-sized, non-overlapping window of data. It is well-suited for this scenario because you want to process the data once per minute, and a tumbling window with a size of 1 minute would allow you to do this.

For the analysis type, you should use Point within polygon. This analysis type allows you to determine whether a GPS position falls within a specific geographical area. You can use the CSV file in the Azure Data Lake Storage Gen2 container to define the expected geographical areas for each vehicle.

upvoted 9 times

□ **youngbug** 1 year ago

Choosing no window is totally wrong. No window means you have to process a msg everytime a event happened. It's costly. But a session window can be triggered when there is a event and can combine all the events in maxduration size(here is 30s) to add a packet of events at one time. And other types of windows are not suitable for the situation here.

upvoted 8 times

□ **hoangcv** 1 year, 2 months ago

I think the answer is correct, should be hopping window with window size is 30 seconds and hope size is 1 minute.

upvoted 2 times

□ **semauni** 5 months, 1 week ago

Why exactly? That means that every point is going to fall inside two events, so you'll be pinged two times.

upvoted 2 times

□ **Deeksha1234** 1 year, 5 months ago

Answer is correct, agree to the point made by VyshakhUnnikrishnan

upvoted 3 times

□ **vishal10** 1 year, 5 months ago

when a GPS position is outside the expected area, a message is added to another event hub for processing within 30 seconds. - this has to be taken in to account

No Window

upvoted 2 times

□ **vishal10** 1 year, 5 months ago

1. Azure Stream Analytics
2. No Window
3. Point within Polygon

upvoted 2 times

□ **Aurelk** 1 year, 7 months ago

i thing it is correct

upvoted 2 times

□ **nefarious_smalls** 1 year, 8 months ago

I am not sure about tumbling window here. I think it would work but what would be the requirement to send an event out every thirty seconds with no data i.e no vehicle has traveled outside the geographic area. In my opinion, a no window could send data to another event hub as needed.

upvoted 3 times

□ **Khiem** 1 year, 9 months ago

It's Tumbling.

No Window: it should run 500 times per minute (500 messages are sent per minute) -> costly.

Tumbling Window: if we configure 15s -> it runs 4 times per minute -> much better.

Hopping Window: Some messages are processed twice (don't care allowing duplication or not) -> costly.
upvoted 12 times

✉ **adel182ff** 1 year, 11 months ago

Trumbling is the right answer because a message is added to another event hub for processing within 30 seconds -> Tumbling
upvoted 7 times

✉ **VyshakhUnnikrishnan** 1 year, 11 months ago

You need a window function to broadcast the messages every 30 seconds for the past one minute. It needs to geofence and update the second output with the location.

1. Azure Stream Analytics
2. Hopping Window
3. Point within Polygon

Ref: <https://docs.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios>

look for the text "The following query joins the device stream with the geofence reference data and calculates the number of requests per region on a time window of 15 minutes every minute."

This is a similar example

upvoted 11 times

✉ **adel182ff** 1 year, 11 months ago

Trumbling is the right answer because a message is added to another event hub for processing within 30 seconds -> Tumbling
upvoted 2 times

You are designing an Azure Databricks table. The table will ingest an average of 20 million streaming events per day.

You need to persist the events in the table for use in incremental load pipeline jobs in Azure Databricks. The solution must minimize storage costs and incremental load times.

What should you include in the solution?

- A. Partition by DateTime fields.
- B. Sink to Azure Queue storage.
- C. Include a watermark column.
- D. Use a JSON format for physical data storage.

Correct Answer: B

The Databricks ABS-AQS connector uses Azure Queue Storage (AQS) to provide an optimized file source that lets you find new files written to an Azure Blob storage (ABS) container without repeatedly listing all of the files. This provides two major advantages:

- ⇒ Lower latency: no need to list nested directory structures on ABS, which is slow and resource intensive.
- ⇒ Lower costs: no more costly LIST API requests made to ABS.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/spark/latest/structured-streaming/aqs>

Community vote distribution

B (81%)

A (19%)

 **bc5468521** Highly Voted 2 years, 7 months ago

The ABS-AQS source is deprecated. For new streams, we recommend using Auto Loader instead.

upvoted 27 times

 **manquak** Highly Voted 2 years, 4 months ago

Why not partition by date? What does the auto loader have to do with streaming jobs?

upvoted 14 times

 **dakku987** Most Recent 1 week, 5 days ago

Selected Answer: A

chat gpt

o design an efficient Azure Databricks table for ingesting an average of 20 million streaming events per day while minimizing storage costs and incremental load times, you should consider the following:

- A. Partition by DateTime fields.

Explanation:

Partitioning by DateTime fields is a common practice for time-series data in Azure

upvoted 1 times

 **kkk5566** 4 months ago

Selected Answer: B

should be B

upvoted 1 times

 **akhil5432** 5 months ago

Selected Answer: B

option B

upvoted 1 times

 **vctrhugo** 6 months, 2 weeks ago

Sinking to Azure Queue storage is not necessary for persisting the events in the Azure Databricks table. Azure Queue storage is typically used for decoupling and asynchronous messaging scenarios and may not directly contribute to minimizing storage costs or incremental load times for the Databricks table.

upvoted 1 times

 **auwia** 6 months, 3 weeks ago

Selected Answer: B

Probably it is B:

Partition by date&time is not the best, imagine events with each single partition because of (day, hour, minute, second) => the requirement is clear, minimize the space, etc..

You use Watermark when you need to reduce the amount of state data to improve latency during a long-running streaming operation.

JSON I would exclude because how it is formulated.

My answer is B, even if it's deprecated, it's clear that this question is an old one, but looking at the comments, we can still get in the exam.
upvoted 3 times

 **dkksk** 8 months, 1 week ago

Selected Answer: A

A. Partition by DateTime fields: Partitioning the table on frequently used columns such as DateTime fields can improve query performance and reduce incremental load times. Partitioning by DateTime can help to reduce the amount of data scanned during query execution and facilitate incremental loading.

upvoted 2 times

 **hiyoww** 9 months, 2 weeks ago

is the question outdated?

upvoted 2 times

 **haidebelognime** 11 months ago

Selected Answer: A

Im sure it is A. Partition by DateTime!!

upvoted 1 times

 **kckalahasti** 1 year, 1 month ago

<https://learn.microsoft.com/en-us/azure/databricks/structured-streaming/aqs>

upvoted 2 times

 **Igor85** 1 year, 1 month ago

question is deprecated, AutoLoader is the way to do the incremental loads

upvoted 4 times

 **RajashekharC** 1 year, 4 months ago

As per requirement: "You need to persist the events in the table for use in incremental load pipeline jobs in Azure Databricks.

Wha I understood from this is, dataset which will stored would be used by Databricks and load type is incremental. Considering this, I see "watermark column" makes more sense.

upvoted 4 times

 **Deeksha1234** 1 year, 5 months ago

B is correct

upvoted 2 times

 **Aurelkb** 1 year, 7 months ago

Selected Answer: B

Corret

upvoted 2 times

 **kamil_k** 1 year, 10 months ago

Do we want to persist this data in a table or in a message queue? From what the question asks it has to be a table. Why would we use queue storage for this task?

upvoted 4 times

 **Canary_2021** 2 years ago

Selected Answer: B

A. Partition by DateTime field

Each partition will generate a file. Loading latency may reduce, but file storage cost will increase because generate more folders and files for different partition. Is it right???

B. Sink to Azure Queue Storage.

Read this document. Spark table files are stored in DBFS. Mount Azure Blob storage containers to Databricks File System (DBFS). If The Databricks ABS-AQS provides these two benefits, sounds like it is a correct answer.

<https://docs.microsoft.com/en-us/azure/databricks/spark/latest/structured-streaming/aqs>

C. Include a watermark column: For sure it is not correct

Watermarks define how long your aggregate should wait around for data delay.

D. User a Json format for physical data storage. - ???

Don't find any documents to compare physical data storage of JSON, CSV, and Parquet.

upvoted 10 times

HOTSPOT -

You have a self-hosted integration runtime in Azure Data Factory.

The current status of the integration runtime has the following configurations:

- Status: Running
- Type: Self-Hosted
- Version: 4.4.7292.1
- Running / Registered Node(s): 1/1
- High Availability Enabled: False
- Linked Count: 0
- Queue Length: 0
- Average Queue Duration: 0.00s

The integration runtime has the following node details:

- Name: X-M
- Status: Running
- Version: 4.4.7292.1
- Available Memory: 7697MB
- CPU Utilization: 6%
- Network (In/Out): 1.21KBps/0.83KBps
- Concurrent Jobs (Running/Limit): 2/14
- Role: Dispatcher/Worker
- Credential Status: In Sync

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

If the X-M node becomes unavailable, all executed pipelines will:

| |
|---------------------------------------|
| fail until the node comes back online |
| switch to another integration runtime |
| exceed the CPU limit |

The number of concurrent jobs and the CPU usage indicate that the Concurrent Jobs (Running/Limit) value should be:

| |
|------------|
| raised |
| lowered |
| left as is |

Correct Answer:

Answer Area

If the X-M node becomes unavailable, all executed pipelines will:

| |
|---------------------------------------|
| fail until the node comes back online |
| switch to another integration runtime |
| exceed the CPU limit |

The number of concurrent jobs and the CPU usage indicate that the Concurrent Jobs (Running/Limit) value should be:

| |
|------------|
| raised |
| lowered |
| left as is |

Box 1: fail until the node comes back online

We see: High Availability Enabled: False

Note: Higher availability of the self-hosted integration runtime so that it's no longer the single point of failure in your big data solution or cloud data integration with Data Factory.

Box 2: lowered -

We see:

Concurrent Jobs (Running/Limit): 2/14

CPU Utilization: 6%

Note: When the processor and available RAM aren't well utilized, but the execution of concurrent jobs reaches a node's limits, scale up by increasing the number of concurrent jobs that a node can run

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime>

✉️ **Sunnyb** Highly Voted 2 years, 7 months ago

$1/14 = 0.07$
 $6\% = 0.06$

should be lowered.
upvoted 32 times

✉️ **romanzdk** 1 year, 11 months ago

$0.06/2 = 0.03$
 $0.03 * 14 = 0.42$ = maximally 42% of cpu for all jobs
isn't this better?
upvoted 9 times

✉️ **semauni** 5 months, 1 week ago

Why is this the calculation you make? I see 6% utilization, so 94% to go, so the amount can be raised.
upvoted 2 times

✉️ **shachar_ash** 1 year, 7 months ago

The question mentions 2/14 which is 0.14, therefore it can be increased.
upvoted 4 times

✉️ **MirandaL** Highly Voted 2 years, 6 months ago

"We recommend that you increase the concurrent jobs limit only when you see low resource usage with the default values on each node."
<https://docs.microsoft.com/en-us/azure/data-factory/monitor-integration-runtime>
upvoted 22 times

✉️ **Momoanwar** Most Recent 1 month ago

Chatgpt :
Given that high availability is not enabled for the self-hosted integration runtime, the correct answer for the first statement is:
- **Fail until the node comes back online.**

For the second statement regarding the number of concurrent jobs, considering that the CPU utilization is quite low at 6%, and there is a significant difference between the number of running jobs (2) and the limit (14), the correct answer should be:

- **Left as is.**

There is no indication from the given data that the concurrent jobs limit needs to be adjusted, as the system is currently underutilized.
upvoted 2 times

✉ **phydev** 2 months, 1 week ago

ChatGPT says: *it should be raised* because there's currently a very low CPU utilization (only 6%) and two concurrent jobs running out of a limit of 14. The fact that the CPU utilization is quite low suggests that your integration runtime has available processing capacity.
upvoted 2 times

✉ **Vanq69** 3 months ago

I don't know why there is this one standard bs answer with lowered everywhere.
So only 2 concurrent jobs are running out of 14 possible and CPU usage is at 6%, does it not make sense to raise the concurrent jobs to be even 14/14 and still have only 42% CPU usage. Or is this question aiming at something else?
upvoted 1 times

✉ **Ram9198** 4 months ago

Left as it is
upvoted 2 times

✉ **kkk5566** 4 months ago

be lowered
upvoted 1 times

✉ **Omkarrokee** 6 months, 2 weeks ago

Based on the information provided, the CPU Utilization is 6% and the Concurrent Jobs (Running/Limit) is 2/14. This indicates that the integration runtime is utilizing only 6% of the available CPU capacity and currently running 2 out of a maximum limit of 14 concurrent jobs.

Given this information, the appropriate answer choice for the completion statement would be:

Concurrent Job should be scaled up

Since the current CPU utilization is relatively low at 6% and there is still capacity available for running additional jobs, scaling up the concurrent job limit would allow for more jobs to run simultaneously and make better use of the available resources.

upvoted 4 times

✉ **auwia** 6 months, 3 weeks ago

We are talking about max number of job running in parallel!
If you have available resource of course it is recommended to raise up the current limit to afford future load.
Also Microsoft recommend that:
<https://learn.microsoft.com/en-us/azure/data-factory/monitor-integration-runtime>
We recommend that you increase the concurrent jobs limit only when you see low resource usage with the default values on each node. I think this is the case, also the question doesn't tell you it's mandatory, what should! So I think we should follow recommendation and raise up the limit.
upvoted 2 times

✉ **pavankr** 7 months, 3 weeks ago

when the explanation is "scale up by increasing the number" then why the answer is "Lowered"???

upvoted 2 times

✉ **norbitek** 12 months ago

I would leave it as it is.
See:
<https://learn.microsoft.com/en-us/azure/data-factory/monitor-integration-runtime>

"The default value of the concurrent jobs limit is set based on the machine size. The factors used to calculate this value depend on the amount of RAM and the number of CPU cores of the machine. So the more cores and the more memory, the higher the default limit of concurrent jobs."

You scale out by increasing the number of nodes. When you increase the number of nodes, the concurrent jobs limit is the sum of the concurrent job limit values of all the available nodes. For example, if one node lets you run a maximum of twelve concurrent jobs, then adding three more similar nodes lets you run a maximum of 48 concurrent jobs (that is, 4×12). We recommend that you increase the concurrent jobs limit only when you see low resource usage with the default values on each node."

upvoted 3 times

✉ **martcerv** 1 year ago

the concurrent jobs limit is the sum of the concurrent job limit values of all the available nodes
upvoted 1 times

✉ **OldSchool** 1 year, 1 month ago

Here is my thinking of this. High availability is False, so no scaling. The 2nd Q is what should be done with the number of concurrent jobs, not scaling up CPU. Since there are only 2 running jobs of possible 14 and CPU utilization is only 6% the number of concurrent jobs should be increased. If left as is we are overspending, if decreased we are still overspending even more since CPU utilization will be lowered too.
upvoted 3 times

✉ **OldSchool** 1 year, 1 month ago

When the processor and available RAM aren't well utilized, but the execution of concurrent jobs reaches a node's limits, scale up by increasing the number of concurrent jobs that a node can run.

<https://learn.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime?tabs=data-factory#scale-up>
upvoted 1 times

✉ **kamil_k** 1 year, 10 months ago

according to this article [https://docs.microsoft.com/en-us/azure/data-factory/monitor-integration-runtime#:~:text=When%20you%20increase%20the%20number,is%2C%204%20x%2012\)](https://docs.microsoft.com/en-us/azure/data-factory/monitor-integration-runtime#:~:text=When%20you%20increase%20the%20number,is%2C%204%20x%2012)).
it is not advisable to touch the default calculated limits unless we encounter issues..

upvoted 1 times

✉ **VyshakhUnnikrishnan** 1 year, 11 months ago

The CPU is only used 6% with 2 parallel jobs running. This gives the opportunity for the cluster to scale up the number of concurrent jobs. The number of parallel/concurrent jobs should hence be increase

upvoted 4 times

✉ **kamil_k** 1 year, 10 months ago

This question is poorly written, it doesn't paint the whole picture. You would need to monitor resource utilisation over a prolonged period of time e.g. 24 hours to see what happens at peak times when you can have all 14 spots taken. Each job can take different amount of compute power. For instance you may find that at some times one job consumes 50% CPU.

upvoted 1 times

✉ **edba** 2 years ago

It sounds many people were confused regarding 2nd question. after check the link below, I think it means to lower or raise limit concurrent jobs. Apparently 14 is too high as usage is 2 only, so it should be "lowered". ref:<https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime?tabs=data-factory#scale-considerations>

upvoted 17 times

✉ **Billybob0604** 1 year, 1 month ago

Yes, this is the right explanation. the question is about lowering the limit not the running jobs.

upvoted 4 times

✉ **Dusica** 1 year, 1 month ago

agree with this; practically paying more capacity then they need

upvoted 1 times

✉ **sdokmak** 1 year, 7 months ago

maybe they worded the question poorly deliberately to test you on what you can scale up or down, which would be the limit not the number of jobs.

upvoted 3 times

✉ **onyerleft** 2 years ago

- 1) Fail until the node comes back online
- 2) left as is

For all those who are extrapolating CPU based on the 2 jobs that are running - you have no idea what the other 12 concurrent jobs could look like. You could have one additional job that maxes the CPU. You could have 12 easy jobs that bring it up to 10% utilization. Since we don't know, leave things as they are until one of the values becomes a bottleneck.

upvoted 6 times

✉ **Davico93** 1 year, 6 months ago

You're right, a professional would say that, but I think Ms is making it easier. I would say also left as is, but that they have to monitor

upvoted 1 times

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier.

You need to configure workspace1 to support autoscaling all-purpose clusters. The solution must meet the following requirements:

- Automatically scale down workers when the cluster is underutilized for three minutes.
- Minimize the time it takes to scale to the maximum number of workers.
- Minimize costs.

What should you do first?

- A. Enable container services for workspace1.
- B. Upgrade workspace1 to the Premium pricing tier.
- C. Set Cluster Mode to High Concurrency.
- D. Create a cluster policy in workspace1.

Correct Answer: B

For clusters running Databricks Runtime 6.4 and above, optimized autoscaling is used by all-purpose clusters in the Premium plan

Optimized autoscaling:

Scales up from min to max in 2 steps.

Can scale down even if the cluster is not idle by looking at shuffle file state.

Scales down based on a percentage of current nodes.

On job clusters, scales down if the cluster is underutilized over the last 40 seconds.

On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds.

The spark.databricks.aggressiveWindowDownS Spark configuration property specifies in seconds how often a cluster makes down-scaling decisions. Increasing the value causes a cluster to scale down more slowly. The maximum value is 600.

Note: Standard autoscaling -

Starts with adding 8 nodes. Thereafter, scales up exponentially, but can take many steps to reach the max. You can customize the first step by setting the spark.databricks.autoscaling.standardFirstStepUp Spark configuration property.

Scales down only when the cluster is completely idle and it has been underutilized for the last 10 minutes.

Scales down exponentially, starting with 1 node.

Reference:

<https://docs.databricks.com/clusters/configure.html>

Community vote distribution

| | |
|---------|----|
| B (93%) | 7% |
|---------|----|

□  ROBERSONWM Highly Voted 2 years, 4 months ago

B is the correct answer.

Automated (job) clusters always use optimized autoscaling. The type of autoscaling performed on all-purpose clusters depends on the workspace configuration.

Standard ~~optimized~~ autoscaling is used by all-purpose clusters in workspaces in the Standard pricing tier. Optimized autoscaling is used by all-purpose clusters in the Azure Databricks Premium Plan.

<https://docs.databricks.com/clusters/cluster-config-best-practices.html>

upvoted 17 times

□  elcholo Highly Voted 2 years, 4 months ago

QUE PASA CHOLO!

upvoted 14 times

□  mulumbi 8 months, 3 weeks ago

delete this please, highly offensive especially to us in rajput region

upvoted 1 times

□  dakku987 1 week, 5 days ago

how it is offensive to rajput?? it from peru!

upvoted 1 times

□  kkk5566 Most Recent 4 months ago

Selected Answer: B

B is correct

upvoted 1 times

akhil5432 5 months ago

Selected Answer: B

Option B

upvoted 1 times

auwia 6 months, 3 weeks ago

Selected Answer: D

I've finally found a valid answer:

<https://learn.microsoft.com/en-us/azure/databricks/administration-guide/clusters/policies>

upvoted 1 times

auwia 6 months, 2 weeks ago

False, Cluster policies require the Premium plan. :) So B is the correct answer.

upvoted 3 times

Rossana 8 months, 2 weeks ago

B doesn't minimize the costs. To support autoscaling all-purpose clusters in Azure Databricks, you need to create a cluster policy that specifies the auto-scaling settings. The cluster policy allows you to specify when to add or remove workers based on the workload on the cluster.

For this scenario, the cluster policy should be configured to automatically scale down workers when the cluster is underutilized for three minutes. This will help to minimize costs by reducing the number of idle workers. The policy should also be configured to scale to the maximum number of workers quickly to minimize the time it takes to process workloads.

Enabling container services for workspace1 (option A) is not necessary for autoscaling all-purpose clusters. Upgrading workspace1 to the Premium pricing tier (option B) may not be necessary and may not be cost-effective depending on your specific requirements. Setting Cluster Mode to High Concurrency (option C) is not related to autoscaling all-purpose clusters.

upvoted 5 times

JG1984 6 months, 3 weeks ago

Cluster policies are available only in the Premium pricing tier of Azure Data bricks, and not in the Standard pricing tier.

upvoted 3 times

Deeksha1234 1 year, 5 months ago

B is correct

upvoted 1 times

Aurelkb 1 year, 7 months ago

Selected Answer: B

correct

upvoted 2 times

Egocentric 1 year, 8 months ago

B is the correct answer

upvoted 1 times

Jaws1990 1 year, 12 months ago

Not sure if this is a valid question anymore. This link shows that the standard pricing tier supports optimised autoscaling.

<https://databricks.com/product/azure-pricing>

upvoted 6 times

allagowf 1 year, 3 months ago

the autoscaling is under the premium plan not the standard one and this is clear in the link you shared.

upvoted 1 times

Igor85 1 year, 1 month ago

no difference anymore between Standard and Premium, indeed

upvoted 1 times

cosarac 1 year, 1 month ago

as Jaws1990 says it is available on both on the link. I have green for both types

upvoted 1 times

lukeonline 2 years ago

Selected Answer: B

We definitely need "Optimized Autoscaling" (not Standard Autoscaling) which is only part of Premium Plan.

Reason: We need to scale down after 3 min underutilization and Standard Autoscaling only allows scaling down after at least 10 minutes.

Standard autoscaling: "Scales down only when the cluster is completely idle and it has been underutilized for the last 10 minutes."

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

upvoted 7 times

 **trietnv** 2 years ago

Selected Answer: B

They need to use Optimized autoscaling for adapting requirements.

- Optimized autoscaling is used by all-purpose clusters in the Azure Databricks Premium Plan.
- On job clusters, scales down if the cluster is underutilized over the last 40 seconds.
- On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds.

reference:

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

upvoted 3 times

 **Canary_2021** 2 years ago

1. Both standard and premium pricing tier support Autopilot Cluster. Autopilot supports Autoscaling and Terminate after X minutes of inactivity.
2. 'Cluster Policies' is only supported by premium pricing tier. Control cost by limiting per cluster maximum cost.

3. standard pricing tier is cheaper than premium pricing tier.

Base on these 3 items, I don't figure out why it has to upgrade to Premium pricing tier.

upvoted 1 times

 **Canary_2021** 2 years ago

A .Enable Databricks Container Service only when you need to use customer containers, so it is not a correct answer.

I vote C to be the correct Answer.

upvoted 1 times

 **Larrave** 2 years, 1 month ago

Answer B is correct. One has to check on the documentation. There are two autoscaling solutions:
standard autoscaling (Standard Tier) and optimized autoscaling (Premium Tier).

Since there is a requirement of downscaling after three minutes of underutilization, only optimized autoscaling can offer such a solution.

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure#optimized-autoscaling>

On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds.

upvoted 6 times

 **certstowinirl** 2 years, 2 months ago

Why is the answer not D? Autoscaling is available in the Standard pricing tier. Since "costs" is also a factor in this question, why upgrade to premium?

upvoted 5 times

 **brendy** 2 years, 4 months ago

Is this correct?

upvoted 2 times

 **Sudheer_K** 2 years, 3 months ago

Not sure, what about the cost factor and premium doesn't minimize cost.

upvoted 1 times

 **husseyn** 2 years, 7 months ago

Concurrent Jobs should be raised - There is less CPU utilization

upvoted 3 times

 **husseyn** 2 years, 7 months ago

Please ignore this, it was meant for the question before

upvoted 10 times

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 tumbling window, and you set the window size to 10 seconds.

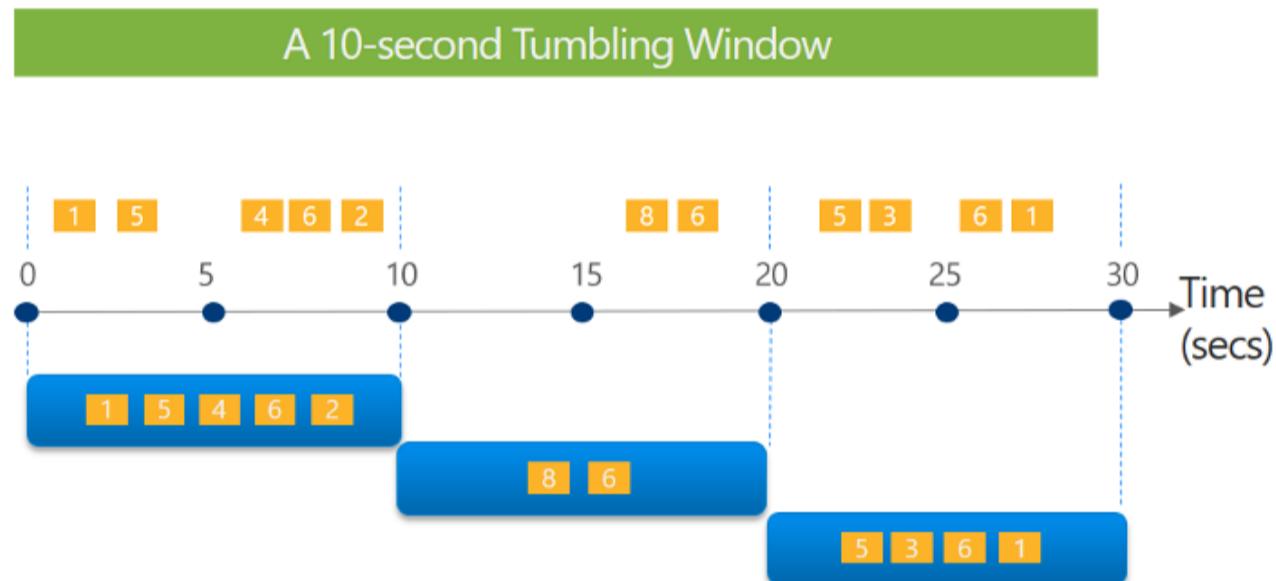
Does this meet the goal?

- A. Yes
B. No

Correct Answer: A

Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.

Tell me the count of tweets per time zone every 10 seconds



```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics>

Community vote distribution

A (100%)

Prabagar Highly Voted 2 years, 7 months ago

correct answer

upvoted 37 times

positivitypeople Most Recent 2 weeks, 5 days ago

Got this question today on the exam

upvoted 1 times

kkk5566 4 months ago

correct

<https://learn.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics>

upvoted 1 times

Deeksha1234 1 year, 5 months ago

correct
upvoted 1 times

□ **practicewizards** 1 year, 6 months ago

this question appears at topic 2 question 18 and it said the correct answer was hopping window with 10" window... so, what's the right correct answer?

upvoted 2 times

□ **kmrrch** 1 year, 3 months ago

Both are correct. A Hopping window with hop-size = window-size is identical to a Tumbling window.

upvoted 5 times

□ **sarapaisley** 1 year, 9 months ago

Selected Answer: A

correct

upvoted 2 times

□ **agar** 1 year, 10 months ago

correct "D cholo

upvoted 1 times

□ **anto69** 1 year, 11 months ago

quite trivial, yes - correct answer: <https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics#:~:text=Tumbling%20windows%20are%20a%20series,into%2010%2Dsecond%20tumbling%20windows>.

upvoted 1 times

□ **Teraflow** 2 years ago

A is correct

upvoted 1 times

□ **lukeonline** 2 years ago

Selected Answer: A

correct

upvoted 1 times

□ **Canary_2021** 2 years ago

Selected Answer: A

A is Correct Answer

upvoted 1 times

□ **rashjan** 2 years, 1 month ago

Selected Answer: A

correct

upvoted 1 times

□ **paoloscott** 2 years, 1 month ago

Correct answer !

upvoted 1 times

□ **AnandEMani** 2 years, 3 months ago

correct

upvoted 1 times

□ **hugoborda** 2 years, 3 months ago

Answer is correct

upvoted 1 times

□ **damaldon** 2 years, 6 months ago

Fully agree

upvoted 2 times

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 session window that uses a timeout size of 10 seconds.

Does this meet the goal?

- A. Yes
B. No

Correct 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 2 years, 7 months ago

answer correct
upvoted 22 times

 **MoDar** Highly Voted 2 years, 4 months ago

False as we need to count tweets in EACH 10 sec. Session windows can have gaps if there is no event happening during the window size
upvoted 12 times

 **robdale** Most Recent 4 months ago

A session window is designed to group events together that occur within a certain time frame, but it doesn't guarantee that each event will be counted only once.
upvoted 2 times

 **kkk5566** 4 months ago

Selected Answer: B
NO is the answer
upvoted 1 times

 **mamahani** 8 months ago

Selected Answer: B
B is correct answer
upvoted 1 times

 **MScapris** 1 year ago

Selected Answer: B
NO is correct
upvoted 1 times

 **Deeksha1234** 1 year, 5 months ago

B is correct
upvoted 3 times

 **sarapaisley** 1 year, 9 months ago

Selected Answer: B
session window wouldn't count periods with no tweets
upvoted 4 times

 **Sandip4u** 2 years ago

This should be yes as the max duration of window is 10 secs and timeout size is also 10 sec . So this means irrespective of any events comes or not timeout is happening or not window size will be remain as 10 sec.
upvoted 1 times

 **Teraflow** 2 years ago

B - it has to be tumbling window
upvoted 1 times

 **rashjan** 2 years, 1 month ago

Selected Answer: B
correct: no
upvoted 1 times

 **dragos_dragos62000** 2 years, 6 months ago

I think you can use a session window with 10 sec timeout... is like tumbling window with 10 second window size.
upvoted 3 times

 **RyuHayabusa** 2 years, 5 months ago

The important thing to remember in a session window is the maximum duration. So theoretically a 10 second timeout can still result in a window of 20 minutes for example (if every 9 seconds a new event comes in and the window never "closes"). If the maximum duration would be 10 seconds, I would agree. But as the question is worded right now, the answer is NO.

<https://docs.microsoft.com/en-us/stream-analytics-query/session-window-azure-stream-analytics>
upvoted 15 times

 **TedoG** 2 years, 5 months ago

I Disagree. The session could be extended if the maximum duration is set longer than the timeout.
upvoted 4 times

 **EddyRoboto** 2 years, 5 months ago

Agree, cause it doesn't overlap any event, just group them in a given time that we can define;
upvoted 1 times

You use Azure Stream Analytics to receive data from Azure Event Hubs and to output the data to an Azure Blob Storage account.

You need to output the count of records received from the last five minutes every minute.

Which windowing function should you use?

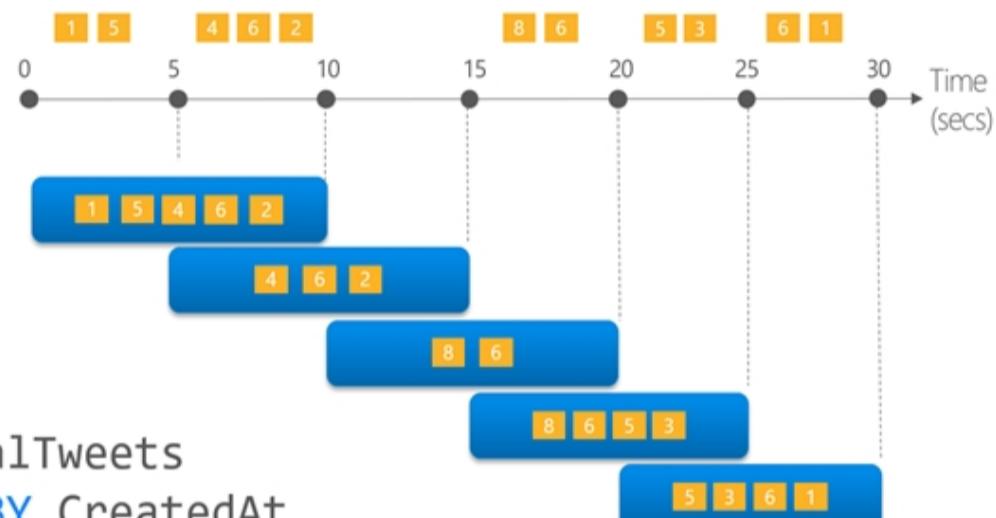
- A. Session
- B. Tumbling
- C. Sliding
- D. Hopping

Correct Answer: D

Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.

Every 5 seconds give me the count of Tweets over the last 10 seconds

A 10-second Hopping Window with a 5-second "Hop"



```
SELECT Topic, COUNT(*) AS TotalTweets
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY Topic, HoppingWindow(second, 10 , 5)
```

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

Community vote distribution

D (100%)

alexleonvalencia Highly Voted 2 years, 1 month ago

Respuesta Correcta: Hopping
upvoted 15 times

DataEngDP Most Recent 3 months, 3 weeks ago

Selected Answer: D

Updated comment: Hopping window since the we are 'hopping' every minute to count the number of records received in the last 5 minutes, 3 arguments: hopping(minute, 5, 1)
upvoted 1 times

DataEngDP 3 months, 3 weeks ago

Selected Answer: D

Hopping window since the we are 'hopping' every minute to count the number of records received in the last 5 minutes, 2 arguments:
hopping(minute, 1)
upvoted 1 times

kkk5566 4 months ago

Selected Answer: D

Hopping
upvoted 1 times

akhil5432 5 months ago

Selected Answer: D

D is correct answer

upvoted 2 times

Rossana 8 months, 2 weeks ago

A hopping window would not be the best option for this scenario because it does not allow you to set a sliding interval that is less than the window size.

In a hopping window, the window size is fixed, and the window "hops" forward by a specified number of intervals. For example, if you set a hopping window size of five minutes and a hop size of one minute, then the first window would include data from the first five minutes, the second window would include data from the second through sixth minutes, the third window would include data from the third through seventh minutes, and so on.

In this scenario, if you set the hopping window size to five minutes, you would only output the count of records every five minutes, which does not meet the requirement of outputting the count of records every minute. Therefore, a sliding window would be a better choice as it allows you to output data at smaller sliding intervals, which is required in this scenario.

upvoted 3 times

Deeksha1234 1 year, 5 months ago

Selected Answer: D

Hopping is right

upvoted 4 times

StudentFromAus 1 year, 6 months ago

Why shouldn't it be sliding?

upvoted 1 times

C1995 1 year, 8 months ago

Why is sliding not correct?

upvoted 1 times

Davico93 1 year, 6 months ago

I want to know too

upvoted 1 times

[Removed] 1 year, 2 months ago

I was thinking sliding as well but a sliding window wouldn't have advanced or returned a result if there was no data e.g. the count was zero.

Hopping advances when there is no input.

upvoted 2 times

SebK 1 year, 9 months ago

Selected Answer: D

Correct

upvoted 3 times

wwdba 1 year, 10 months ago

Hopping is correct!

upvoted 1 times

metallicjade 1 year, 11 months ago

Selected Answer: D

hopping window

upvoted 1 times

Teraflow 2 years ago

Hopping window is correct

upvoted 4 times

HOTSPOT -

You configure version control for an Azure Data Factory instance as shown in the following exhibit.

Git repository

Git repository information associated with your data factory. [CI/CD best practices](#)

Setting **Disconnect**

| | |
|----------------------|------------------|
| Repository type | Azure DevOps Git |
| Azure DevOps Account | CONTOSO |
| Project name | Data |
| Repository name | dwh_batchetl |
| Collaboration branch | main |
| Publish branch | adf_publish |
| Root folder | / |

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:

Answer Area

Azure Resource Manager (ARM) templates for the pipeline assets are stored in [answer choice]

| | |
|---------------------------|---|
| | ▼ |
| / | |
| adf_publish | |
| main | |
| Parameterization template | |

A Data Factory Azure Resource Manager (ARM) template named contososales can be found in [answer choice]

| | |
|--|---|
| | ▼ |
| / | |
| /contososales | |
| /dwh_batchetl/adf_publish/contososales | |
| /main | |

Correct Answer:

Answer Area

Azure Resource Manager (ARM) templates for the pipeline assets are stored in [answer choice]

| | |
|---------------------------|---|
| / | ▼ |
| adf_publish | |
| main | |
| Parameterization template | |

A Data Factory Azure Resource Manager (ARM) template named contososales can be found in [answer choice]

| | |
|--|---|
| / | ▼ |
| contososales | |
| /dwh_batchetl/adf_publish/contososales | |
| /main | |

Box 1: adf_publish -

The Publish branch is the branch in your repository where publishing related ARM templates are stored and updated. By default, it's adf_publish.

Box 2: / dwh_batchetl/adf_publish/contososales

Note: RepositoryName (here dwh_batchetl): Your Azure Repos code repository name. Azure Repos projects contain Git repositories to manage your source code as your project grows. You can create a new repository or use an existing repository that's already in your project.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/source-control>

✉  **Aurelk8** Highly Voted 1 year, 6 months ago

Correct Answer i test it in devops

upvoted 25 times

✉  **VyshakhUnnikrishnan** Highly Voted 1 year, 11 months ago

The assets are in the main branch which is the collaboration branch.

The template is repository/adf_publish/datafactoryname

upvoted 11 times

✉  **Massy** 1 year, 9 months ago

could you please paste the source of that? I can't find it

upvoted 2 times

✉  **kkk5566** Most Recent 4 months ago

correct

upvoted 1 times

✉  **mamahani** 8 months ago

Adf_publish
/Dwh_batchetl/adf_publish/contososales

upvoted 3 times

✉  **Deeksha1234** 1 year, 5 months ago

correct

upvoted 2 times

✉  **SameerL** 1 year, 6 months ago

Can someone please refer to below link which includes video as well to answer this question?

>>> <https://docs.microsoft.com/en-us/azure/data-factory/source-control>

upvoted 1 times

✉  **Amsterliese** 1 year, 9 months ago

I'm not sure, but I think we don't see the complete picture, so we cannot answer the second question for sure. See: <https://docs.microsoft.com/en-us/azure/data-factory/source-control#github-settings>

upvoted 2 times

✉  **BK10** 1 year, 11 months ago

Can someone confirm the correct answer? Is it:

1. adf_publish

2. /.

Please let me know

upvoted 3 times

✉  **ANath** 1 year, 11 months ago

The answers are correct.

upvoted 2 times

- **Rohan21** 1 year, 12 months ago
Second answer should be contososales
upvoted 2 times
- **varmal** 2 years ago
I dont think we ever refer to locations as REPO/BRANCH/PATH in devops. For me it is / as we assume the branch could be any and still location would be /
upvoted 6 times
- **romanzdk** 1 year, 11 months ago
I would say so as well
upvoted 3 times
- **Davico93** 1 year, 6 months ago
You are thinking in an agnostic way and this is Azure DevOps
upvoted 2 times
- **anto69** 1 year, 11 months ago
Yeah, totally agree. Nobody uses this notation
upvoted 4 times
- **VeroDon** 2 years ago
correct
upvoted 2 times
- **Skeinofi** 2 years ago
Correct
upvoted 2 times

HOTSPOT -

You are designing an Azure Stream Analytics solution that receives instant messaging data from an Azure Event Hub.

You need to ensure that the output from the Stream Analytics job counts the number of messages per time zone every 15 seconds.

How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Select ~~店铺: IT认证考试服务~~ TimeZone, count (*) AS MessageCount

FROM MessageStream

| | |
|--------------------------|--------------------|
| <input type="checkbox"/> | ▼ |
| <input type="checkbox"/> | LAST |
| <input type="checkbox"/> | OVER |
| <input type="checkbox"/> | SYSTEM.TIMESTAMP() |
| <input type="checkbox"/> | TIMESTAMP BY |

~~店铺: IT认证考试服务~~ CreatedAt

GROUP BY TimeZone,

| | |
|--------------------------|----------------|
| <input type="checkbox"/> | ▼ |
| <input type="checkbox"/> | HOPPINGWINDOW |
| <input type="checkbox"/> | SESSIONWINDOW |
| <input type="checkbox"/> | SLIDINGWINDOW |
| <input type="checkbox"/> | TUMBLINGWINDOW |

(second, 15)

Correct Answer:

Answer Area

Select TimeZone, count (*) AS MessageCount

FROM MessageStream

| | |
|-------------------------------------|--------------------|
| <input type="checkbox"/> | ▼ |
| <input type="checkbox"/> | LAST |
| <input type="checkbox"/> | OVER |
| <input type="checkbox"/> | SYSTEM.TIMESTAMP() |
| <input checked="" type="checkbox"/> | TIMESTAMP BY |

CreatedAt

GROUP BY TimeZone,

| | |
|-------------------------------------|----------------|
| <input type="checkbox"/> | ▼ |
| <input type="checkbox"/> | HOPPINGWINDOW |
| <input type="checkbox"/> | SESSIONWINDOW |
| <input type="checkbox"/> | SLIDINGWINDOW |
| <input checked="" type="checkbox"/> | TUMBLINGWINDOW |

(second, 15)

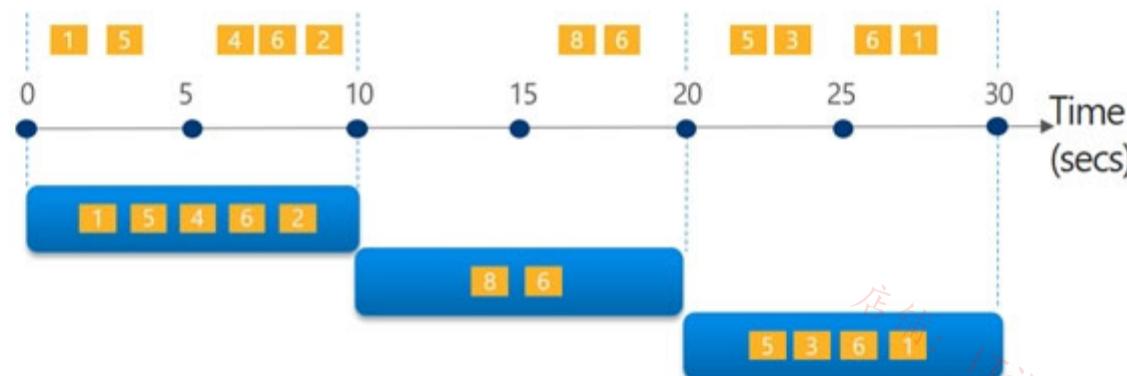
Box 1: timestamp by -

Box 2: TUMBLINGWINDOW -

Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

Tell me the count of Tweets per time zone every 10 seconds

A 10-second Tumbling Window



```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

□ **ANath** Highly Voted 1 year, 11 months ago

The answers are correct
upvoted 27 times

□ **seba020** Highly Voted 9 months, 2 weeks ago

reason for "TIMESTAMPBY":
<https://learn.microsoft.com/en-us/stream-analytics-query/timestamp-by-azure-stream-analytics>
upvoted 5 times

□ **semauni** 5 months, 1 week ago

But a TIMESTAMP BY would need an OVER clause if I see it correctly, and there is none here.
upvoted 2 times

□ **Preksha_** 1 month, 2 weeks ago

As per the syntax, OVER is optional.
<https://learn.microsoft.com/en-us/stream-analytics-query/timestamp-by-azure-stream-analytics>
upvoted 1 times

□ **kkk5566** Most Recent 4 months ago

correct
upvoted 1 times

□ **Dhaval_Azure** 10 months ago

which one is correct? many question left confusing depending on discussion.
and I am not trusting the answer as many answers are wrong.
upvoted 1 times

□ **bch9994** 4 months, 3 weeks ago

you need to use the answers from ET and discussions to find the best answer yourself just like everyone on here.
upvoted 1 times

□ **Deeksha1234** 1 year, 5 months ago

correct answer
upvoted 2 times

□ **May99** 1 year, 12 months ago

I think it's system.timestamp()
upvoted 3 times

□ **sdokmak** 1 year, 7 months ago

From examples, I can only see system.timestamp() used after SELECT, not FROM.
upvoted 1 times

□ **jv2120** 2 years ago

It only says about window size, not sure why tumbling window not hopping.

upvoted 2 times

✉️ **Andreas_K** 2 years ago

Syntax would not be correct since hopping window expects three parameters.

<https://docs.microsoft.com/en-us/stream-analytics-query/hopping-window-azure-stream-analytics>

Tumbling window is the correct answer

upvoted 21 times

✉️ **TestMitch** 2 years ago

Correcto

upvoted 2 times

HOTSPOT -

You have an Azure Data Factory instance named ADF1 and two Azure Synapse Analytics workspaces named WS1 and WS2.

ADF1 contains the following pipelines:

- ☞ P1: Uses a copy activity to copy data from a nonpartitioned table in a dedicated SQL pool of WS1 to an Azure Data Lake Storage Gen2 account
- ☞ P2: Uses a copy activity to copy data from text-delimited files in an Azure Data Lake Storage Gen2 account to a nonpartitioned table in a dedicated SQL pool of WS2

You need to configure P1 and P2 to maximize parallelism and performance.

Which dataset settings should you configure for the copy activity if each pipeline? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

P1:

| | |
|--|---|
| | ▼ |
| Set the Copy method to Bulk insert | |
| Set the Copy method to PolyBase | |
| Set the Isolation level to Repeatable read | |
| Set the Partition option to Dynamic range | |

P2:

| | |
|--|---|
| | ▼ |
| Set the Copy method to Bulk insert | |
| Set the Copy method to PolyBase | |
| Set the Isolation level to Repeatable read | |
| Set the Partition option to Dynamic range | |

Answer Area

P1:

| | |
|--|---|
| | ▼ |
| Set the Copy method to Bulk insert | |
| Set the Copy method to PolyBase | |
| Set the Isolation level to Repeatable read | |
| Set the Partition option to Dynamic range | |

Correct Answer:

店铺: IT认证考试服务

| | |
|--|--------------|
| | 正确: IT认证考试服务 |
| Set the Copy method to Bulk insert | |
| Set the Copy method to PolyBase | |
| Set the Isolation level to Repeatable read | |
| Set the Partition option to Dynamic range | |

Box 1: Set the Copy method to PolyBase

While SQL pool supports many loading methods including non-Polybase options such as BCP and SQL BulkCopy API, the fastest and most scalable way to load data is through PolyBase. PolyBase is a technology that accesses external data stored in Azure Blob storage or Azure Data Lake Store via the T-SQL language.

Box 2: Set the Copy method to Bulk insert

Polybase not possible for text files. Have to use Bulk insert.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/load-data-overview>

marcin1212 Highly Voted 2 years ago

how to use PolyBase when copy data from Synapse to file ? I don't have idea.
Moreover PolyBase option is available only when the target is Synapse

it should be

P1: Set the partition option to "Dynamic range "

P2: PolyBase

regarding to P1

<https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse?tabs=data-factory#parallel-copy-from-synapse-analytics>

Scenario: "Full load from large table, without physical partitions." ->

Suggested settings: Partition options: Dynamic range partition.

upvoted 87 times

Matt2000 5 months ago

It should be:

P1: PolyBase

P2: PolyBase

"PolyBase is the best choice when you are loading or exporting large volumes of data, or you need faster performance."
Ref: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool>

Regarding "dynamic range partitions":

" As repartitioning data takes time, Use [sic] current partitioning is recommended in most scenarios." -> dynamic partitioning is NOT selected
Ref: <https://learn.microsoft.com/en-us/azure/data-factory/concepts-data-flow-performance>

upvoted 1 times

Canary_2021 Highly Voted 2 years ago

P1: Copy data from SQL to Data Lake.

• Bulk insert and PolyBase are not a choice in Sink tab if target is Data Lake. So they are not correct.

• Isolation level can be setup if SQL database is the source. Repeatable Read means that locks are placed on all data that is used in a query. Don't think it maximizes parallelism and performance.

• Set the Partition option to Dynamic range

Can be setup if source is SQL in copy activity. And it maximizes parallelism and performance. So I select this option.

P2: Copy data from Data Lake to SQL. It is for sure to select PolyBase.

upvoted 41 times

6d954df Most Recent 2 weeks ago

For the Azure Data Factory pipelines P1 and P2, the dataset settings for the copy activity should be configured as follows:

P1:

b. Set the Copy method to PolyBase

PolyBase is a technology that accesses data outside of the database via the T-SQL language. It's designed to leverage parallelism, which can lead to significant performance improvements when copying large amounts of data12.

P2:

a. Set the Copy method to Bulk insert

Bulk insert is a process that can be used to import large amounts of data into a SQL Server table. It's a highly efficient way to push data into a table, especially when dealing with text-delimited files12.

Please note that the actual performance may vary depending on the specific requirements and the structure of your data12.

Learn more

1

learn.microsoft.com

2

learn.microsoft.com

3

social.msdn.microsoft.com

upvoted 1 times

positivitypeople 2 weeks, 5 days ago

Got this question today on the exam

upvoted 1 times

d046bc0 3 weeks, 5 days ago

P1: Dynamic range according to

<https://techcommunity.microsoft.com/t5/fasttrack-for-azure/leverage-copy-data-parallelism-with-dynamic-partitions-in-adf/ba-p/3692133>

upvoted 1 times

Momoanwar 1 month ago

Correct, chatgpt

For P1, where data is copied from a non-partitioned table in a SQL pool to Azure Data Lake Storage Gen2:

- **Set the Copy method to PolyBase**: This is because PolyBase is designed to efficiently transfer large amounts of data to and from SQL-based data stores into Azure Data Lake Storage.

For P2, which copies data from text-delimited files in Azure Data Lake Storage Gen2 to a non-partitioned table in a SQL pool:

- **Set the Copy method to Bulk insert**: Bulk insert is an efficient way to load data from files into SQL tables, especially when dealing with non-partitioned tables where PolyBase might not be applicable or the most optimal choice.

upvoted 2 times

fahfouhi94 3 months, 1 week ago

P1 : set the partition option to dynamic range (see here : <https://techcommunity.microsoft.com/t5/fasttrack-for-azure/leverage-copy-data-parallelism-with-dynamic-partitions-in-adf/ba-p/3692133>)

P2: Polybase give the best performance

PolyBase loads data from UTF-8 and UTF-16 encoded delimited text files. PolyBase also loads from the Hadoop file formats RC File, ORC, and Parquet. PolyBase can also load data from Gzip and Snappy compressed files. PolyBase currently does not support extended ASCII, fixed-width format, and nested formats such as WinZip, JSON, and XML.

upvoted 1 times

kkk5566 4 months ago

P1) Set the partition option to dynamic range p2) set the copy method to PolyBase

upvoted 1 times

JezWalters 5 months, 1 week ago

There's a really interesting video regarding PolyBase/COPY INTO here:

<https://microsoft.github.io/PartnerResources/skilling/modern-analytics-academy/vignettes/polybase-vs-copy>

This video indicates that PolyBase can actually be used to pull/push data from/to Azure Data Lake Storage to/from Azure Synapse Analytics Dedicated SQL Pool tables (via CTAS & CETAS statements).

upvoted 4 times

faabbasi 6 months ago

P1 dynamic range, link is pretty clear: <https://learn.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse?tabs=data-factory#parallel-copy-from-azure-synapse-analytics>

upvoted 1 times

Rossana 8 months, 2 weeks ago

for P1, you should set the copy method to Polybase, and for P2, you should set the copy method to Bulk.

The reason is that Polybase is better suited for copying data between Azure Synapse Analytics and Azure Data Lake Storage Gen2, and can achieve better performance than Bulk copy in this scenario. On the other hand, Bulk copy is the fastest method for copying data from text-delimited files in Azure Data Lake Storage Gen2 to Azure Synapse Analytics.

Setting the partition option to Dynamic range for both pipelines can help to maximize parallelism and performance by allowing the copy activity to split the data into multiple partitions based on the data range.

upvoted 2 times

 **vrodriguesp** 11 months ago

I tried to create a copy activity in adf and these were results:

P1) Synapse to ADLS --> Source Partition option: None/Dynamic range

Sink Copy behavior: Add dynamic content/None/Flatten hierarchy/Merge files/Preserve hierarchy

P2) ADLS to Synapse --> Source Copy method: NA

Sink Copy method: Copy command/PolyBase/Bulk insert/Upsert

So I think correct answers should be:

P1) Set the partition option to dynamic range

p2) set the copy method to PolyBase

upvoted 9 times

 **DAYENKAR** 11 months, 3 weeks ago

Both answer are polybase

upvoted 1 times

 **XiltroX** 1 year, 1 month ago

I think you can put both as PolyBase. PolyBase is much faster and supports text delimited files as well now.

<https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse?tabs=data-factory#use-polybase-to-load-data-into-azure-synapse-analytics>

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

Agree with marcin1212

it should be

P1: Set the partition option to "Dynamic range "

P2: PolyBase

upvoted 4 times

 **NamitSehgal** 1 year, 6 months ago

P2 should be Polybase

<https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse?tabs=data-factory#use-polybase-to-load-data-into-azure-synapse-analytics>

P1

set the partition option to "Dynamic range "

upvoted 2 times

 **Towin** 1 year, 8 months ago

Both are PolyBase

<https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-guide?view=sql-server-ver15>

"Azure Synapse Analytics can Read/Write Azure Storage"

<https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse?tabs=data-factory#parallel-copy-from-synapse-analytics>

"As a sink, load data by using COPY statement or PolyBase or bulk insert. We recommend COPY statement or PolyBase for better copy performance"

upvoted 1 times

HOTSPOT -

You have an Azure Storage account that generates 200,000 new files daily. The file names have a format of {YYYY}/{MM}/{DD}/{HH}/{CustomerID}.csv.

You need to design an Azure Data Factory solution that will load new data from the storage account to an Azure Data Lake once hourly. The solution must minimize load times and costs.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Load methodology:

| | |
|--------------------------------------|---|
| | ▼ |
| Full Load | |
| Incremental Load | |
| Load individual files as they arrive | |

Trigger:

| | |
|-----------------|---|
| | ▼ |
| Fixed schedule | |
| New file | |
| Tumbling window | |

Answer Area

Load methodology:

| | |
|--------------------------------------|---|
| | ▼ |
| Full Load | |
| Incremental Load | |
| Load individual files as they arrive | |

Correct Answer:

Trigger:

| | |
|-----------------|---|
| | ▼ |
| Fixed schedule | |
| New file | |
| Tumbling window | |

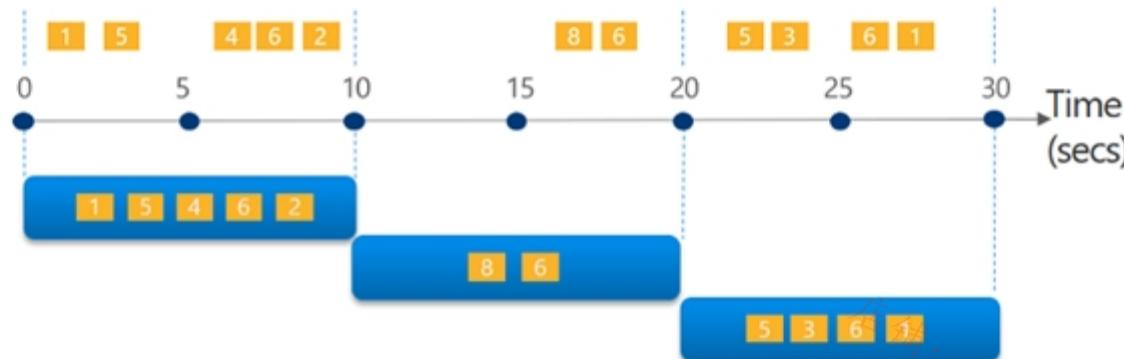
Box 1: Incremental load -

Box 2: Tumbling window -

Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.

Tell me the count of tweets per time zone every 10 seconds

A 10-second Tumbling Window



`SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)`

Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics>

⊕ **onyerleft** Highly Voted 2 years ago

- 1) Incremental Load
- 2) Tumbling Window

Seems like you could go with either Schedule trigger or Tumbling Window here. I would use the latter option, and pass the windowStart system variable to the pipeline as a parameter, allowing me to more easily navigate to the proper directory in the storage account.

upvoted 31 times

⊕ **xcsakubara** Highly Voted 1 year, 10 months ago

Since, we are loading NEW data and not going back in time, it should be Schedule as we are scheduling it for every 1 hour in the future. It would've been Tumbling if we scheduled it for every 1 hour in the past.

upvoted 15 times

⊕ **phydev** 2 months, 1 week ago

Besides, a scheduled trigger is a better option for this specific scenario than a tumbling window due to precision, efficiency and cost savings.

upvoted 1 times

⊕ **positivitypeople** Most Recent 2 weeks, 5 days ago

Got this question today on the exam

upvoted 2 times

⊕ **Momoanwar** 1 month ago

Correct, chatgpt :

For the scenario described, to load data from an Azure Storage account to an Azure Data Lake hourly and to minimize load times and costs, you would configure the Azure Data Factory solution as follows:

- **Load methodology**: Incremental Load - Because you are loading new data every hour, and the goal is to minimize the load times and costs, you would incrementally load only the new data that has arrived since the last load.

- **Trigger**: Tumbling window - This trigger is suitable for fixed-duration, repeating intervals in Azure Data Factory, which fits the requirement of loading data hourly.

Using a tumbling window trigger ensures that each window of time is processed once and only once, and by doing an incremental load, you are only processing the new data that has appeared since the last hour, rather than reprocessing all existing data.

upvoted 2 times

⊕ **Andrew_Chen** 3 months ago

I think one thing very important here is that Tumbling window manages state between runs, that means that it will not be counted twice.

upvoted 1 times

⊕ **auwia** 6 months, 3 weeks ago

From Azure Data Factory Studio when you create a new trigger, you can choose TYPE in ('Schedule', 'Tumbling window', 'Storage events', 'Custom events').

We should exclude "Fixed Schedule" because of 'fixed'! :)

So my final answer will be Incremental Load and Tumbling Window.

upvoted 6 times

 **vedantnj** 8 months, 1 week ago

Hi there

upvoted 4 times

 **Rossana** 8 months, 2 weeks ago

To minimize load times and costs for loading new data from the storage account to an Azure Data Lake once hourly, you should configure the solution to use incremental load and a trigger based on new files arriving.

Load methodology: With 200,000 new files generated daily, a full load every hour could be time-consuming and expensive. Incremental load is a better option in this scenario because it only loads new or changed data since the last successful execution of the pipeline, which can significantly reduce load times and costs.

Trigger: A trigger based on new files arriving is the most efficient option because it only runs the pipeline when new files are detected in the storage account. This avoids unnecessary pipeline executions and reduces costs. A fixed schedule trigger runs the pipeline at fixed intervals, regardless of whether there is new data to process or not. A tumbling window trigger runs the pipeline at specified intervals, but still processes all data within the window, regardless of whether there is new data or not. Therefore, a new file trigger is the best option in this scenario.

upvoted 5 times

 **martcerv** 1 year ago

A schedule for an activity creates a series of tumbling windows with in the pipeline start and end times

I think is "Fixed schedule" because "Tumbling windows" are more related to streams analytics questions according to MS doc.

<https://learn.microsoft.com/en-us/azure/data-factory/v1/data-factory-scheduling-and-execution>

upvoted 6 times

 **Deeksha1234** 1 year, 5 months ago

- 1) Incremental Load
- 2) Tumbling Window

upvoted 4 times

 **jskibick** 1 year, 7 months ago

With Scheduled trigger executions can overlaps if the process does not finish within 1 hour, Tumbling window is better, with concurrency setting it can allow only one ongoing execution.

upvoted 12 times

 **Massy** 1 year, 9 months ago

both Tumbling Window and Schedule trigger will reach the goal. Which one is more cost effective?

upvoted 1 times

 **Boompiee** 1 year, 8 months ago

I think because every hour you're only processing the past hour's data. With a tumbling window you can define which messages to process, whereas with a schedule trigger you'd have to implement that filter separately.

upvoted 2 times

 **xcsakubara** 1 year, 10 months ago

why not schedule trigger?

upvoted 1 times

 **sparkchu** 1 year, 9 months ago

for backfill purpose? just guessing.

upvoted 1 times

 **jv2120** 2 years ago

incremental, fixed schedule every hour.

upvoted 6 times

 **jv2120** 2 years ago

correct answer..tumbling window

upvoted 1 times

 **Ayan3B** 2 years ago

As a input we are receiving csv files so why not trigger mechanism to the pipeline when file arrived.

upvoted 2 times

 **ItHYMeRlsh** 2 years ago

The question says, "load new data from the storage account to the Azure Data Lake once hourly." This already indicates a tumbling window to run every hour.

On top of that, if you executed this as an event every time a file arrived, you'd have 200,000 ADF pipeline executions per day - one per file. If you ran the pipeline once per hour per day, you'd have just 24.

1,000 ADF runs is \$1. In this situation, 1 day is 24 runs when executed on a tumbling window. That's 2.4 cents. If we ran 200,000 pipelines, that'd be \$200/day. This excludes other costs.

<https://azure.microsoft.com/en-us/pricing/details/data-factory/data-pipeline/>

upvoted 35 times

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 plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- A workload for data engineers who will use Python and SQL.
- A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a Standard cluster for the data engineers, and a High Concurrency cluster for the jobs.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer: B

We need a High Concurrency cluster for the data engineers and the jobs.

Note: Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.

A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

Reference:

<https://docs.azuredatabricks.net/clusters/configure.html>

Community vote distribution

| | |
|---------|----|
| B (96%) | 4% |
|---------|----|

✉  lukeonline  2 years ago

Selected Answer: B

B is correct but the explanation is wrong.

- A workload for data engineers who will use Python and SQL. --> high concurrency
- A workload for jobs that will run notebooks that use Python, Scala, and SQL. --> standard
- A workload that data scientists will use to perform ad hoc analysis in Scala and R. --> standard because high concurrency does not support Scala

<https://stackoverflow.com/questions/65869399/high-concurrency-clusters-in-databricks>

upvoted 37 times

✉  kamil_k 1 year, 10 months ago

or rather Scala does not support concurrent instances (but yes, it implies HC cluster will not support Scala)

upvoted 2 times

✉  d046bc0  3 weeks, 5 days ago

Standard is enough for all workloads. High concurrency (due to Scala) possible only for data engineers

upvoted 2 times

✉  Momoanwar 1 month ago

Correct, chatgpt :

For the given scenario, where data engineers must share a cluster, data scientists need their own clusters with auto-termination, and a managed job cluster is required for running notebooks, the solution provided may not fully meet the goal. Here's why:

- Data engineers should share a cluster, so creating a single Standard cluster for all data engineers would meet this requirement.
- For data scientists, the solution suggests a Standard cluster for each, but it should specify that these clusters have auto-termination settings

configured to minimize costs.

- The High Concurrency cluster is suitable for running jobs because it allows multiple users to share the cluster and run jobs concurrently. However, it should be managed as per the enterprise team's standards.

The provided solution does not fully adhere to these standards, especially regarding the auto-termination requirement for data scientists' clusters. Thus, the answer would be:

B. No, the solution does not meet the goal.

upvoted 1 times

□ **kkk5566** 4 months ago

Selected Answer: B

high concurrency does not support Scala

upvoted 2 times

□ **akhil5432** 5 months ago

Selected Answer: B

Correct option is B-NO

upvoted 1 times

□ **Rossana** 8 months, 2 weeks ago

A)Yes

The use of a shared Standard cluster for data engineers, a High Concurrency cluster for jobs, and individual Standard clusters for each data scientist that auto-terminates after 120 minutes of inactivity aligns with the specified standards and is a valid approach for creating a tiered Databricks workspace.

upvoted 1 times

□ **kckalahasti** 1 year, 1 month ago

<https://docs.databricks.com/clusters/configure.html>

upvoted 1 times

□ **Igor85** 1 year, 1 month ago

high concurrency cluster is already a legacy cluster mode. question is not relevant anymore

upvoted 2 times

□ **greenlever** 1 year, 2 months ago

Selected Answer: A

Standard mode can be shared by multiple users and terminate automatically, on the other hand High do not terminate automatically and Scala workload is not supported.

upvoted 2 times

□ **Babu99** 1 year, 3 months ago

NO IS CORRECT ANSWER

upvoted 1 times

□ **Deeksha1234** 1 year, 5 months ago

correct , answer B, agree with lukeonline

upvoted 1 times

□ **mkthoma3** 1 year, 6 months ago

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

upvoted 1 times

□ **Hanse** 1 year, 10 months ago

As per Link: <https://docs.azuredatabricks.net/clusters/configure.html>

Standard and Single Node clusters terminate automatically after 120 minutes by default. --> Data Scientists

High Concurrency clusters do not terminate automatically by default.

A Standard cluster is recommended for a single user. --> Standard for Data Scientists & High Concurrency for Data Engineers

Standard clusters can run workloads developed in any language: Python, SQL, R, and Scala.

High Concurrency clusters can run workloads developed in SQL, Python, and R. The performance and security of High Concurrency clusters is provided by running user code in separate processes, which is not possible in Scala. --> Jobs needs Standard

upvoted 3 times

□ **bad_atitude** 2 years ago

B is correct

upvoted 2 times

□ **alexleonvalencia** 2 years, 1 month ago

Selected Answer: B

Respuesta correcta; Standar para Cientificos y jobs. Alta concurrencia para ingenieros de datos.

upvoted 3 times

□ **Sanand** 2 years ago

Agree! - Correct answer; Standard for Scientists and jobs. High concurrency for data engineers.

upvoted 3 times

You have the following Azure Data Factory pipelines:

- Ingest Data from System1
- Ingest Data from System2
- Populate Dimensions
- Populate Facts

Ingest Data from System1 and Ingest Data from System2 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest

Data from System2. Populate Facts must execute after Populate Dimensions pipeline. All the pipelines must execute every eight hours.

What should you do to schedule the pipelines for execution?

- 店铺：
A. Add an event trigger to all four pipelines.
B. Add a schedule trigger to all four pipelines.
C. Create a parent pipeline that contains the four pipelines and use a schedule trigger.
D. Create a parent pipeline that contains the four pipelines and use an event trigger.

Correct Answer: C

Schedule trigger: A trigger that invokes a pipeline on a wall-clock schedule.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers>

Community vote distribution

C (100%)

 **onyerleft** Highly Voted 2 years ago

Selected Answer: C

C is correct, but with poor wording. Should be 'parent pipeline' with a schedule trigger.

The parent pipeline has 4 execute pipeline activities. Ingest 1 and Ingest 2 have no dependencies. Dimension pipeline has two dependencies from 'on completion' outputs of both Ingest 1 and Ingest 2 pipelines. Fact pipeline has one 'on completion' dependency on the Dimension pipeline. Absolutely nothing to do with a tumbling window trigger

upvoted 61 times

 **Vaq69** 3 months ago

Also looked up "patient pipeline" and was confused xD

upvoted 1 times

 **lukeonline** 2 years ago

Lol, I searched in the internet for the "patient pipeline".... should have read the comments first :)

upvoted 20 times

 **Remedios79** 1 year, 6 months ago

Thank you. I was wondering about "patient" and related it on my poor english!

upvoted 2 times

 **dsp17** 1 year, 6 months ago

Big thanks onyerleft :)

upvoted 2 times

 **phydev** Most Recent 2 months, 1 week ago

Selected Answer: C

Was on my exam today (31.10.2023).

upvoted 2 times

 **kkk5566** 4 months ago

Selected Answer: C

c is correct

upvoted 1 times

 **steveo123** 6 months, 3 weeks ago

C is correct

upvoted 1 times

 **vigilante89** 1 year ago

Selected Answer: C

Its not patient pipeline, it should be parent pipeline. Since there are 3 types of triggers in ADF:

- 1) Schedule Trigger - trigger a pipeline at a fixed hour/minute of the day.
- 2) Tumbling Window Trigger - trigger a pipeline which usually works for real time data
- 3) Event-based Trigger - trigger a pipeline incase of an event i.e. new file coming to blob/adls etc.

Since the 4 pipelines must be triggered every 8 hrs, then it should be schedule trigger.

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

right C

upvoted 1 times

 **DrTaz** 2 years ago

what the ~~h~~ is a patient pipeline?

upvoted 2 times

 **anto69** 1 year, 11 months ago

lol I think they mean "parent"

upvoted 1 times

 **jv2120** 2 years ago

It should be tumbling window since 2 dependent pipelines on run state. from given option only schedule event fits but its not correct.

upvoted 3 times

 **AzureJobsTillRetire** 1 year, 1 month ago

If those pipelines finish quickly, schedule trigger should be fine. If there is possibility that those pipelines may run for close to or more than 8 hours, definitely tumbling window should be used instead

upvoted 1 times

 **VJPR** 2 years ago

Shouldn't the answer be A/D?

upvoted 1 times

 **TashaP** 1 year, 4 months ago

So the question tries to trick you, they don't want to ask about individual pipeline configurations where you need to account for dependencies, they literally want to know how you will schedule the pipelines for execution. The additional information is there to confuse you and make you overthink, focus on the question. In this case, it is C.

upvoted 1 times

 **dpBBC** 2 years ago

I think it should be Tumbling window

upvoted 2 times

 **corebit** 2 years ago

The question or answers do not mention Tumbling Window. What is the basis for the response? Any more context?

upvoted 2 times

DRAG DROP -

You are responsible for providing access to an Azure Data Lake Storage Gen2 account.

Your user account has contributor access to the storage account, and you have the application ID and access key.

You plan to use PolyBase to load data into an enterprise data warehouse in Azure Synapse Analytics.

You need to configure PolyBase to connect the data warehouse to storage account.

Which three components should you create in sequence? To answer, move the appropriate components from the list of components to the answer area and arrange them in the correct order.

Select and Place:

Components

- a database scoped credential
- an asymmetric key
- an external data source
- a database encryption key
- an external file format

Answer Area

Correct Answer:

Components

- a database encryption key
- an external file format

Answer Area

an asymmetric key

a database scoped credential

an external data source

Step 1: an asymmetric key -

A master key should be created only once in a database. The Database Master Key is a symmetric key used to protect the private keys of certificates and asymmetric keys in the database.

Step 2: a database scoped credential

Create a Database Scoped Credential. A Database Scoped Credential is a record that contains the authentication information required to connect an external resource. The master key needs to be created first before creating the database scoped credential.

Step 3: an external data source -

Create an External Data Source. External data sources are used to establish connectivity for data loading using Polybase.

Reference:

<https://www.sqlservercentral.com/articles/access-external-data-from-azure-synapse-analytics-using-polybase>

✉ **alexonvalencia** Highly Voted 2 years, 1 month ago

- 1.- A database scoped credential
 - 2.- an External data sorce
 - 3.- a external file format
- upvoted 185 times

✉ **DiscussoR** 9 months, 2 weeks ago

Agree:

<https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects?view=sql-server-ver16#create-external-tables-for-hadoop>

upvoted 2 times

✉ **Franz58** 1 year, 5 months ago

you need to connect to the DW, not to a specific file. Therefore :

- 1- Create a Database Encryption Key
- 2 - Create a Database Scoped Credential
- 3 - Create an External Data Source

upvoted 18 times

✉ **auwia** 6 months, 3 weeks ago

It should be: "Create a Master Key" not 'DataBase Encryption Key' if you are trying to findout a predecessor step before "create a Database Scoped Credential".

upvoted 2 times

✉ **DiscussoR** 9 months, 2 weeks ago

File format is not related to a specific file

upvoted 1 times

✉ **Bilal2** 1 year ago

agreed.

<https://www.sqlshack.com/sql-server-polybase-external-tables-with-azure-blob-storage/>

upvoted 1 times

✉ **engrbrain** Highly Voted 2 years ago

According to the documentation, the first thing you are to create is
CREATE MASTER KEY ENCRYPTION BY PASSWORD = 'S0me!Info';

I don't think this means an asymmetric key. It is simply a database encryption key. So I think the answer is

- 1- Create a Database Encryption Key
- 2 - Create a Database Scoped Credential
- 3 - Create an External Data Source

upvoted 53 times

✉ **vanrell** 1 year, 9 months ago

Does the text not say you already have an access key? Should the correct answer not be

- 1.- A database scoped credential
- 2.- an External data source
- 3.- a external file format

as alex mentions?

upvoted 4 times

✉ **sdokmak** 1 year, 7 months ago

access key is for storage account so you still need a master/asymmetric key for the database.

upvoted 4 times

✉ **sdokmak** 1 year, 7 months ago

*sorry, not asymmetric

upvoted 3 times

✉ **kamil_k** 1 year, 9 months ago

Btw yes even in the description it says that the master key is a symmetric key, not an asymmetric one. It

upvoted 2 times

✉ **kamil_k** 1 year, 10 months ago

also, the question only mentions storage account in general not a file or folder, so I believe we don't need to go as far as creating file format anyway

upvoted 3 times

✉ **jongert** Most Recent 1 week, 4 days ago

Given answer appears correct. This section specifies we need the master key (which is an asymmetric key), then create the scoped credential. The scoped credential can finally be used to create an external data source.

<https://learn.microsoft.com/en-us/sql/t-sql/statements/create-database-scoped-credential-transact-sql?view=sql-server-ver16#c-creating-a-database-scoped-credential-for-polybase-connectivity-to-azure-data-lake-store>

upvoted 1 times

✉ **6d954df** 2 weeks ago

check this out, the answers provided are correct <https://learn.microsoft.com/en-us/sql/analytic-platform-system/polybase-configure-azure-blob-storage?view=aps-pdw-2016-au7>

upvoted 1 times

✉ **Momoanwar** 1 month ago

Chatgpt :

To configure PolyBase to connect an Azure Synapse Analytics data warehouse to an Azure Data Lake Storage Gen2 account, you need to create:

1. **A database scoped credential**: This stores the necessary authentication to access the data lake, such as the storage account access key.

2. **An external data source**: This defines the location of the data in the storage account and uses the scoped credential for authentication.
3. **An external file format**: This specifies the format of the data files (e.g., CSV, Parquet) in the data lake so that PolyBase knows how to parse the data.

These components should be created in the sequence listed above to ensure that PolyBase has the information it needs to authenticate, locate, and read the data from the storage account.

upvoted 3 times

✉ **abhijectbgmcanada** 1 month, 3 weeks ago

Seems like answers are differing. Do we get partial points if answer is partially correct ?

upvoted 2 times

✉ **ellala** 3 months ago

The database master key is a SYMMETRIC KEY (<https://learn.microsoft.com/en-us/sql/t-sql/statements/create-master-key-transact-sql?view=sql-server-ver16#remarks>) therefore the answer should be:

1- Create a Database Encryption Key (CREATE MASTER KEY ENCRYPTION BY PASSWORD)

2 - Create a Database Scoped Credential (CREATE DATABASE SCOPED CREDENTIAL)

3 - Create an External Data Source (CREATE EXTERNAL DATA SOURCE)

(<https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects?view=sql-server-ver16#1-create-database-scoped-credential>)

upvoted 1 times

✉ **oturbo** 3 months, 1 week ago

CREATE MASTER KEY;

-- SECRET: Provide your Azure storage account key.

CREATE DATABASE SCOPED CREDENTIAL AzureStorageCredential

WITH

IDENTITY = 'user',

SECRET = '<azure_storage_account_key>'

;

CREATE EXTERNAL DATA SOURCE AzureStorage

WITH (

TYPE = HADOOP,

LOCATION = 'wasbs://<blob_container_name>@<azure_storage_account_name>.blob.core.windows.net',

CREDENTIAL = AzureStorageCredential

);

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-load-from-azure-blob-storage-with-polybase>

upvoted 1 times

✉ **kkk5566** 4 months ago

1.- A database scoped credential 2.- an External data sorce 3.- a external file format

upvoted 1 times

✉ **Ram9198** 5 months ago

1.A database scoped credential

2.an External data source

3 a external file format

Create master key is an symmetric key <https://learn.microsoft.com/en-us/sql/t-sql/statements/create-master-key-transact-sql?view=sql-server-ver16>

DEK comes under the concept of azure SQL TDE and no way related to this question

Hence proved

upvoted 2 times

✉ **mcwest002** 6 months ago

Create external tables for Azure Data Lake Store

From <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>

1. Create database scoped credential

2. Create external data source to reference Azure Data Lake Store (ADLS)

3. Create external file format

upvoted 2 times

✉ **rocky48** 7 months, 2 weeks ago

1.- A database scoped credential

2.- an External data sorce

3.- a external file format

upvoted 1 times

✉ **Rossana** 8 months, 2 weeks ago

Create an external data source (C) that specifies the location of the data in the storage account.

Create an external file format (E) that describes the format of the data in the external data source.

Create a database scoped credential (A) that contains the credentials needed to access the storage account.
Note that asymmetric keys and database encryption keys are not required for configuring PolyBase with Azure Data Lake Storage Gen2.
upvoted 1 times

 **DipikaChavan** 8 months, 4 weeks ago

- 1.A database scoped credential
- 2.an External data source
- 3 a external file format

upvoted 2 times

 **DiscussoR** 9 months, 2 weeks ago

The final answer is:
Master key (to encrypt credentials)
Scoped credential (to provide credentials for storage account)
External data source (to point to a specific storage account)

Source: <https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects?view=sql-server-ver16#create-external-tables-for-hadoop>

upvoted 2 times

 **esaade** 10 months ago

To configure PolyBase to connect the data warehouse to the storage account, you should create the following components in sequence:

An asymmetric key in the data warehouse database.
A database scoped credential using the application ID and access key.
An external data source that references the database scoped credential and specifies the storage account details.

upvoted 1 times

 **esaade** 10 months, 1 week ago

To configure PolyBase to connect the data warehouse to the storage account, you should create the following components in sequence:

An asymmetric key (to secure the database scoped credential).
A database scoped credential (to provide authentication to the storage account).
An external data source (to define the connection to the storage account).

upvoted 1 times

You are monitoring an Azure Stream Analytics job by using metrics in Azure.

You discover that during the last 12 hours, the average watermark delay is consistently greater than the configured late arrival tolerance.

What is a possible cause of this behavior?

- A. Events whose application timestamp is earlier than their arrival time by more than five minutes arrive as inputs.
- B. There are errors in the input data.
- C. The late arrival policy causes events to be dropped.
- D. The job lacks the resources to process the volume of incoming data.

Correct Answer: D

Watermark Delay indicates the delay of the streaming data processing job.

There are a number of resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to:

1. Not enough processing resources in Stream Analytics to handle the volume of input events. To scale up resources, see Understand and adjust Streaming Units.
2. Not enough throughput within the input event brokers, so they are throttled. For possible solutions, see Automatically scale up Azure Event Hubs throughput units.
3. Output sinks are not provisioned with enough capacity, so they are throttled. The possible solutions vary widely based on the flavor of output service being used.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling>

Community vote distribution

D (100%)

 **Skeinofi** Highly Voted 2 years ago

Selected Answer: D

The link provided is the source of truth
upvoted 14 times

 **Teraflow** Highly Voted 2 years ago

Selected Answer: D

D is correct
upvoted 7 times

 **positivitypeople** Most Recent 2 weeks, 5 days ago

Got this question today on the exam
upvoted 1 times

 **kkk5566** 4 months ago

Selected Answer: D

D is correct
upvoted 1 times

 **akhil5432** 5 months ago

Selected Answer: D

option D
upvoted 1 times

 **Deeksha1234** 1 year, 5 months ago

correct
upvoted 2 times

 **austin06112000** 1 year, 9 months ago

D is correct.
upvoted 3 times

 **DrTaz** 2 years ago

Selected Answer: D

Correct. D is the one that makes most sense.
upvoted 3 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

HOTSPOT -

You are building an Azure Stream Analytics job to retrieve game data.

You need to ensure that the job returns the highest scoring record for each five-minute time interval of each game.

How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

SELECT

| | |
|--|---|
| Collect(Score) | ▼ |
| CollectTop(1) OVER(ORDER BY Score Desc) | ▼ |
| Game, MAX(Score) | ▼ |
| TopOne() OVER(PARTITION BY Game ORDER BY Score Desc) | ▼ |

as HighestScore

FROM input TIMESTAMP BY CreatedAt

GROUP BY

| | |
|---|---|
| Game | ▼ |
| Hopping(minute,5) | ▼ |
| Tumbling(minute,5) | ▼ |
| Windows(TumblingWindow(minute,5),Hopping(minute,5)) | ▼ |

Correct Answer:**Answer Area**

SELECT

| | |
|--|---|
| Collect(Score) | ▼ |
| CollectTop(1) OVER(ORDER BY Score Desc) | ▼ |
| Game, MAX(Score) | ▼ |
| TopOne() OVER(PARTITION BY Game ORDER BY Score Desc) | ▼ |

as HighestScore

FROM input TIMESTAMP BY CreatedAt

GROUP BY

| | |
|---|---|
| Game | ▼ |
| Hopping(minute,5) | ▼ |
| Tumbling(minute,5) | ▼ |
| Windows(TumblingWindow(minute,5),Hopping(minute,5)) | ▼ |

Box 1: TopOne OVER(PARTITION BY Game ORDER BY Score Desc)

TopOne returns the top-rank record, where rank defines the ranking position of the event in the window according to the specified ordering.

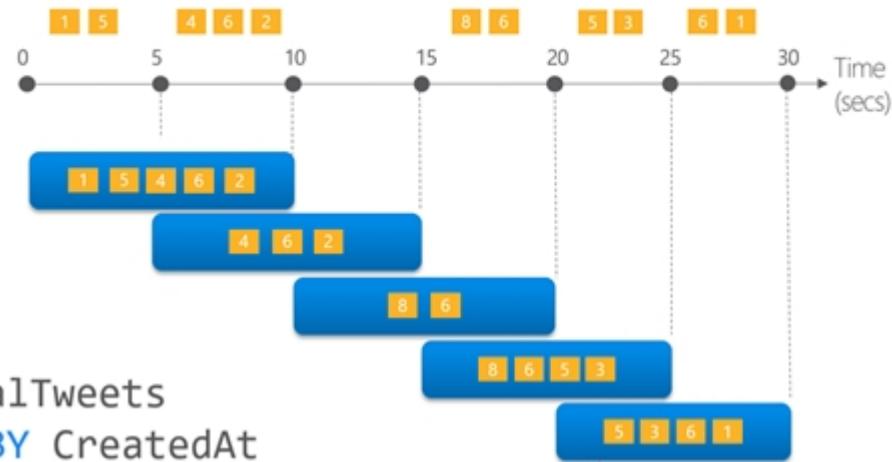
Ordering/ranking is based on event columns and can be specified in ORDER BY clause.

Box 2: Hopping(minute,5)

Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.

Every 5 seconds give me the count of Tweets over the last 10 seconds

A 10-second Hopping Window with a 5-second "Hop"



`SELECT Topic, COUNT(*) AS TotalTweets
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY Topic, HoppingWindow(second, 10, 5)`

Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/topone-azure-stream-analytics> [https://docs.microsoft.com/en-us/azure/stream-analytics/window-functions](https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions)

✉ **alexleonvalencia** Highly Voted 2 years, 1 month ago

TopOne() / Tumbling
upvoted 126 times

✉ **MarcelIT** 11 months, 2 weeks ago

The Select built with the TopOne() option would return one row for each game. Still, it would not tell you the game (SELECT TOP_ONE... FROM). On the other hand, the GAME,MAX() option clearly informs the Game.
upvoted 10 times

✉ **_lene_** 9 months ago

The question was "the highest scoring record of each game", so that's what we need - one row for each game
upvoted 2 times

✉ **cr727** 11 months, 2 weeks ago

I think its TopOne() as "TopOne() OVER(partition by Game order by Score Desc)", it orders by descending of Score and by partition, and top one of each of them.
upvoted 2 times

✉ **gf2tw** Highly Voted 2 years, 1 month ago

Syntax for Hopping window requires 3 arguments, seems this should be Tumbling Window which fulfils the exact same requirements.
upvoted 42 times

✉ **anto69** 1 year, 11 months ago

Yeah sure
upvoted 2 times

✉ **6d954df** Most Recent 1 week, 6 days ago

SELECT
TopOne() OVER(PARTITION BY Game ORDER BY Score Desc) as HighestScore
FROM input TIMESTAMP BY CreatedAt
GROUP BY
Game, TumblingWindow(minute,5)
Here's what this query does:

TopOne() OVER(PARTITION BY Game ORDER BY Score Desc) selects the highest scoring record in each partition of games.
TIMESTAMP BY CreatedAt specifies the timestamp of each event.

GROUP BY Game, TumblingWindow(minute,5) groups the output by game and in five-minute intervals. The TumblingWindow function creates a series of fixed-sized, non-overlapping and contiguous time intervals.

Please note that this is a general guidance and the actual query might need to be adjusted based on the specific requirements and data schema of your game data.

upvoted 1 times

✉ **Momoanwar** 1 month ago

Wrong, chatgpt :

For the Azure Stream Analytics job that needs to return the highest scoring record for each five-minute interval of each game, the query should use the following options:

1. **SELECT**: `TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)` as HighestScore

- This function returns the highest score for each partitioned group (each game in this case), ordered by score in descending order, ensuring that the highest score is selected.

2. **GROUP BY**: `TumblingWindow(minute, 5)`

- This window function groups the events into non-overlapping, continuous five-minute intervals, which is what's required to get the highest score in each five-minute time slice.

This configuration will ensure that you get the highest score for each game every five minutes.

upvoted 1 times

✉ **MarkJoh** 1 month, 1 week ago

There really isn't a solution based on the options given.

First off, Hopping() and Tumbling() don't exist, it's HoppingWindow and TumblingWindow
So, the group by only has

Game

Windows(TumblingWindow(minute, 5, Hopping(minute, 5))

But, that last one isn't valid either in multiple ways.

So, only Group by Game is valid.

So, you can try it this way...

```
SELECT TOPOne() OVER (PARTITION BY Game ORDER BY Score DESC)
```

```
FROM input TIMESTAMP BY CreatedDt
```

```
GROUP BY game
```

But, that does nothing with 5 minute window.

So, there is NO solution based on the data given.

I would write it this way

```
SELECT game, MAX(Score) as MaxScore
```

```
FROM intput TIMESTAMP by CreatedAt
```

```
Group by Game, TumblingWindow(minute, 5)
```

But, those aren't options provided.

upvoted 2 times

✉ **kkk5566** 4 months ago

Game(max) and tumbling window is the correct answer

upvoted 5 times

✉ **pavankr** 6 months, 1 week ago

"Hopping(minute,5)" - the syntax itself is wrong. Not sure who is preparing these answers???

upvoted 5 times

✉ **akk_1289** 1 year ago

minute time interval of each game, you can use the TumblingWindow function to define a five-minute tumbling window over the data, and then use the MAX function to select the highest scoring record within each window.

upvoted 7 times

✉ **mroova** 11 months ago

Totally agree:

- Game, max() - you need to have [Game] column to know, which game the max score refers to,
- tumbling window - requires 2 arguments, hopping window could be used, but requires 3 arguments

upvoted 9 times

✉ **auwia** 6 months, 3 weeks ago

Then you have to put "game" attribute in the group by as well, but there is only 1 option and it's without the window! So ripone / tumbling!

upvoted 3 times

✉ **auwia** 6 months, 3 weeks ago

Topone

upvoted 2 times

✉ **Achu24** 1 year ago

Game(max) and tumbling window is the correct answer

upvoted 7 times

✉ **mesloth** 12 months ago

Correct. Here top score is being asked, instead of Rank

upvoted 3 times

✉ **JosephVishal** 1 year ago

Tumbling window seems to be correct, in question there is no fixed time interval specified.

upvoted 1 times

✉ **THAYTRUONG** 1 year, 1 month ago

TopOne() / Tumbling is correct answer

upvoted 2 times

✉ **bakstorage00001** 1 year, 3 months ago

This is clearly a fu**-up, it's a tumbling Window. For sure! I wonder what would happen in the exam if you select Tumbling...

upvoted 3 times

✉ **allagowf** 1 year, 3 months ago

true, if tumbling is not the correct answer in the exam then we fu**-up really fu**-up hahaha

upvoted 3 times

✉ **Deeksha1234** 1 year, 5 months ago

It should be TopOne() and Tumbling
upvoted 2 times

✉ **jainparag1** 1 year, 5 months ago

It should be Tumbling window.
upvoted 4 times

✉ **agar** 1 year, 10 months ago

it is Tumbling "D
upvoted 1 times

✉ **chxzqw** 1 year, 11 months ago

pls why not game, max(score) ?
upvoted 7 times

IT认证考试服务

✉ **svik** 1 year, 11 months ago

If max(score) is used then we have to have Game in the group by clause

upvoted 12 times

IT认证考试服务

✉ **assU2** 1 year, 11 months ago

pls can someone explain in details why not game, max(score) ?

upvoted 3 times

IT认证考试服务

✉ **adfgasd** 1 year, 11 months ago

If you use max(score) in first box and game in second, you would not have a max(score) every 5 minutes. If you choose max(score) in first box and tumbling in second, query would return an error, because it misses the group by game clause.

upvoted 11 times

IT认证考试服务

✉ **stunner85_** 1 year, 11 months ago

When you don't group them by Game, when you run Game, Max(Score) it will get the max score out of all games.

upvoted 3 times

✉ **Teraflow** 2 years ago

It should be tumbling window

upvoted 3 times

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 Data Lake Storage account that contains a staging zone.

You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script.

Does this meet the goal?

A. Yes

B. No

Correct Answer: A

If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity with your own data processing logic and use the activity in the pipeline.

Note: You can use data transformation activities in Azure Data Factory and Synapse pipelines to transform and process your raw data into predictions and insights at scale.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/transform-data>

Community vote distribution

| | |
|---------|----|
| B (94%) | 6% |
|---------|----|

✉️  **NaiCob**  2 years ago

Correct answer: No - you cannot execute the R script using a stored procedure activity
upvoted 52 times

✉️  **auwia** 6 months, 3 weeks ago

I agree, I've found an articol where it's saying that you can run R script from a Custom .net activity, or better if you have BYOC HDInsight cluster that already has R Installed on it.
upvoted 1 times

✉️  **Daemon69**  1 year, 11 months ago

I select A because you can use R script in sp_execute_external_script
upvoted 13 times

✉️  **Rossana** 8 months, 2 weeks ago

The answer is NO for other reasons than the SP.
Concerning the SP: To execute an R script within a stored procedure in Synapse Analytics, you can use the sp_execute_external_script system stored procedure. This procedure can be used to execute R scripts, as well as scripts written in other languages such as Python.
upvoted 2 times

✉️  **sparkchu** 1 year, 9 months ago

i admire your thought, but context looks wanna discriminate the inavailability of R in SP not like that in Databricks.
upvoted 4 times

✉️  **ExamDestroyer69**  5 days, 11 hours ago

Selected Answer: B

VARIATIONS OF THIS QUESTION

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script. **NO**

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes an Azure Databricks notebook, and then inserts the data into the data warehouse. **YES**

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data flow, and then inserts the data into the data warehouse. **NO**

Solution: You schedule an Azure Databricks job that executes an R notebook, and then inserts the data into the data warehouse. **YES**
upvoted 1 times

 **TheReg81** 3 months, 1 week ago

While the most voted is 'No', the green bar indicates 'Yes'. My first inclination was 'No' as well, because it's not the easiest or most logical way to do it. Then again, there's many roads to Rome. Does it meet the goal, yes.

upvoted 2 times

 **kkk5566** 4 months ago

Selected Answer: B

B should be correct

upvoted 1 times

 **FRANCIS_A_M** 9 months ago

Correct Answer B:

The solution proposed does not meet the goal because it suggests executing the R script using a stored procedure in the data warehouse. Azure Synapse Analytics does not support executing R scripts directly within stored procedures. Instead, you should use Azure Data Factory to orchestrate the process, using an Azure Machine Learning activity to execute the R script for data transformation before loading the transformed data into Azure Synapse Analytics.

upvoted 3 times

 **Kamekung** 10 months, 1 week ago

Btw.. Is it worth to pay for accessing the rest of pages? Since the actual value is community discussion. And beyond this point, it's supposed to be less people.

upvoted 2 times

FRANCIS_A_M 9 months, 2 weeks ago

I have paid for further access and would say it is worth it. The community discussion continues

upvoted 1 times

 **bubby248** 11 months ago

Cant we fix answers correctly in the portal, instead of relying on votes

upvoted 2 times

 **mckovin** 11 months, 1 week ago

Correct

upvoted 1 times

 **millusmiley** 11 months, 3 weeks ago

Next page is asking for contributor access, anyone have credentials or how we can skip the payment

upvoted 2 times

CNBOOST2 11 months, 2 weeks ago

I think this is not possible we have to pay :(

upvoted 3 times

 **Dusica** 12 months ago

Selected Answer: B

there is a staging zone in Azure Data Lake Storage. The very fact that A suggest copying into DWH staging zone makes it invalid so any other discussion is unnecessary. It is B

upvoted 2 times

 **akk_1289** 1 year ago

his solution does not meet the goal of the daily process you have described. While using an Azure Data Factory schedule trigger to execute a pipeline is a good approach for scheduling the process to run on a daily basis, the pipeline you have described does not include any steps to transform the data using an R script.

To meet the goal of the daily process, you will need to include a step in the pipeline to execute the R script that transforms the data. One way to do this would be to use an Azure Data Factory activity, such as an Execute R Script activity, to run the R script on the data as it is being copied from the staging zone to the staging table in the data warehouse. You can then use a stored procedure or another Data Factory activity, such as an SQL activity, to insert the transformed data into the final destination table in the data warehouse.

upvoted 1 times

 **Tj87** 1 year, 5 months ago

Synapse doesn't support R at the moment

<https://docs.microsoft.com/en-us/answers/questions/222624/is-azure-synapse-analytics-supporting-r-language.html>

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

should be B

upvoted 3 times

 **nilubabu** 1 year, 6 months ago

As per problem, Azure Data Lake Storage account that contains a staging zone. From staging zone, transform the data and insert into Azure Synapse Analytics.

But the solution providing as copy data to a staging table in data warehouse.

As per problem, staging will be in Azure Data Lake Storage account, not in data warehouse.

Answer is 'B'

upvoted 5 times

 rafaelptu 1 year, 9 months ago

Selected Answer: R

Sim, o script vai ser executado e carregado posteriormente a execução pode ser chamada pela sp_exec_external_script
upvoted 1 times

 Philipp 1 year, 11 months ago

Selected Answer: B

Should be NO
upvoted 1 times

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 plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- A workload for data engineers who will use Python and SQL.
- A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a High Concurrency cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.

Does this meet the goal?

A. Yes

B. No

Correct Answer: B

Need a High Concurrency cluster for the jobs.

Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.

A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

Reference:

<https://docs.azuredatabricks.net/clusters/configure.html>

Community vote distribution

B (100%)

 **djinche** Highly Voted 2 years, 4 months ago

data scientist need scala so standard, jobs need scala so standard, so B but for different reasons
upvoted 45 times

 **Gina8008** 1 year, 11 months ago

engineer has to share the cluster so high -concurrency is correct. the answer should be A
upvoted 2 times

 **Aditya0891** 1 year, 7 months ago

gina8008 you are missing a point here that data scientists uses scala as per question and scala is not supported in high concurrency cluster.
So the answer is no
upvoted 7 times

 **111222333** Highly Voted 2 years, 7 months ago

Correct is A
upvoted 16 times

 **dfdsfdfsfsd** 2 years, 7 months ago

Agree. Jobs cannot use a high-concurrency cluster because it does not support Scala.
upvoted 5 times

 **Aditya0891** 1 year, 7 months ago

and what about the data scientists requirement? Read the question properly and don't mislead people looking for answers. Scala is not supported in high concurrency and data scientists are using scala as per question so answer in No
upvoted 6 times

 **Chemmangat** Most Recent 3 months, 3 weeks ago

Selected Answer: B

Answer : B

"High Concurrency clusters can run workloads developed in SQL, Python, and R."

<https://learn.microsoft.com/en-us/azure/databricks/archive/compute/configure>

upvoted 1 times

 **kkk5566** 4 months ago

Selected Answer: B

B should be correct

upvoted 1 times

 **akhil5432** 5 months ago

Selected Answer: B

NO is correct answer

upvoted 2 times

 **Hanse** 1 year, 10 months ago

As per Link: <https://docs.azure.databricks.net/clusters/configure.html>

Standard and Single Node clusters terminate automatically after 120 minutes by default. --> Data Scientists

High Concurrency clusters do not terminate automatically by default.

A Standard cluster is recommended for a single user. --> Standard for Data Scientists & High Concurrency for Data Engineers

Standard clusters can run workloads developed in any language: Python, SQL, R, and Scala.

High Concurrency clusters can run workloads developed in SQL, Python, and R. The performance and security of High Concurrency clusters is provided by running user code in separate processes, which is not possible in Scala. --> Jobs needs Scala, hence: Standard

upvoted 2 times

 **avijitd** 2 years ago

Selected Answer: B

NO - as High concurrency not support Scala

upvoted 6 times

 **rashjan** 2 years, 1 month ago

Selected Answer: B

correct: no

upvoted 5 times

 **arjunbhai** 2 years, 1 month ago

Like djincheg said, Data scientists need scala so B.

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

upvoted 2 times

 **Julius7000** 2 years, 3 months ago

-Data Engineers: Correct, they are working together, they need High-Concurreny cluster

-Jobs: Correct, Standad Cluster since it supports SCALA

HOWEVER:

- Data Scientists need cluster who terminates after 120 minutes automatically: THAT MEANS ONLY STANDARD AND SINGLE NODE CLUSTERS CAN SUPPORT THAT.

Since this is the holistic question, the answer is NO.

upvoted 14 times

 **Julius7000** 2 years, 3 months ago

All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity.

That means they need standard cluster, not high-concurreny cluster. STANDARD cluster terminates automatically after 120 minutes:

"Standard and Single Node clusters terminate automatically after 120 minutes by default."

IMO the answer is NO, since all 3 solutions have to be correct.

upvoted 2 times

 **michals** 2 years, 4 months ago

It's correct that standard cluster is for job workload, but they assigned high concurrency cluster for data scientist, who want to use scala too, so it's false

upvoted 4 times

 **damaldon** 2 years, 6 months ago

Answer: A

-Data scientist should have their own cluster and should terminate after 120 mins - STANDARD

-Cluster for Jobs should support scala - STANDARD

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

upvoted 2 times

 **kimalto452** 2 years, 3 months ago

Solution: You create a High Concurrency cluster for each data scientist
Does this meet the goal?
A. Yes

Answer: A
-Data scientist should have their own cluster and should terminate after 120 mins - STANDARD

GENIUSSSSSSSSS
upvoted 1 times

 **Sunnyb** 2 years, 7 months ago

A is the right answer because Standard cluster supports scala
upvoted 2 times

 **Wisenut** 2 years, 7 months ago

I too agree on the comment by 111222333. As per the requirement " A workload for jobs that will run notebooks that use Python, Scala, and SOL".
Scala is ~~only~~ supported by Standard
upvoted 6 times

You are designing an Azure Databricks cluster that runs user-defined local processes.

You need to recommend a cluster configuration that meets the following requirements:

- ⇒ Minimize query latency.
- ⇒ Maximize the number of users that can run queries on the cluster at the same time.
- ⇒ Reduce overall costs without compromising other requirements.

Which cluster type should you recommend?

- A. Standard with Auto Termination
- B. High Concurrency with Autoscaling
- C. High Concurrency with Auto Termination
- D. Standard with Autoscaling

Correct Answer: B

A High Concurrency cluster is a managed cloud resource. The key benefits of High Concurrency clusters are that they provide fine-grained sharing for maximum resource utilization and minimum query latencies.

Databricks chooses the appropriate number of workers required to run your job. This is referred to as autoscaling. Autoscaling makes it easier to achieve high cluster utilization, because you don't need to provision the cluster to match a workload.

Incorrect Answers:

C: The cluster configuration includes an auto terminate setting whose default value depends on cluster mode:

Standard and Single Node clusters terminate automatically after 120 minutes by default.

High Concurrency clusters do not terminate automatically by default.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

Community vote distribution

B (71%)

C (29%)

Canary_2021 [Highly Voted] 2 years ago

Selected Answer: B

B is correct answer.

High concurrency cluster cannot terminated, so C is wrong.

Standard cluster cannot shared by multiple tasks, so A and D are wrong.

upvoted 17 times

HaBroNounen 2 years ago

"High Concurrency clusters do not terminate automatically by default."

but u can change that default so your argument about C is incorrect..

Link: <https://docs.microsoft.com/en-us/azure/databricks/clusters/configure#cluster-mode>

upvoted 16 times

ANath [Highly Voted] 2 years ago

Auto terminate for high concurrency cluster is possible. But due to the 2nd point 'Maximize the number of users that can run queries on the cluster at the same time', I will go with option B. High Concurrency and Auto Scaling

upvoted 5 times

kamil_k 1 year, 9 months ago

Also to minimise query latency you don't want to have to wait for a cluster to spin up after it terminates

upvoted 3 times

kkk5566 [Most Recent] 4 months ago

Selected Answer: B

B is correct

upvoted 1 times

UzairMir 5 months, 4 weeks ago

Here is my take on this. Autoscaling addresses first requirement i-e latency and Auto terminate addresses third requirement i-e cost. Now we have to implement both of them manually, meaning none of them is by default. But in the question it says reduce cost without compromising other requirements. So if we go for auto-termination that means we are not autoscaling, compromising the first requirement i-e the latency requirement. therefore the given option is correct i-e B: High Concurrency with autoscaling

upvoted 3 times

Reloadedvn 8 months ago

Selected Answer: B

Agree to B
upvoted 1 times

 **_lene_** 9 months ago

Selected Answer: C

The cluster does auto-scaling by default. Auto-termination should be set up manually
upvoted 5 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: B

B is correct
upvoted 1 times

 **jz10** 1 year, 8 months ago

Selected Answer: B

Just because auto termination is eligible for high concurrency clusters, doesn't mean we have to use it.
A key requirement is to "minimize query latency", which makes autoscaling more favorable.

Ref: "Workloads can run faster compared to a constant-sized under-provisioned cluster."
<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure#cluster-size-and-autoscaling>
upvoted 4 times

 **Amsterliese** 1 year, 9 months ago

High Concurrency clusters can be configured with auto termination (I just checked). BUT: The questions says: reduce costs WITHOUT compromising the other requirements. So I would still go for autoscaling, since there is no answer option that offers both (autoscaling and auto termination)

upvoted 4 times

 **sunithagsk** 1 year, 9 months ago

Answer should be B as per below
The key benefits of High Concurrency clusters are that they provide fine-grained sharing for maximum resource utilization and minimum query latencies. Autoscaling clusters can reduce overall costs compared to a statically-sized cluster.
upvoted 4 times

 **alex1491** 1 year, 10 months ago

Selected Answer: C

i try it and it's possible to create a cluster with auto termination.
upvoted 3 times

 **AngelJP** 1 year, 10 months ago

Selected Answer: C

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure#cluster-mode>
- High Concurrency clusters do not terminate automatically by default.
- A Standard cluster is recommended for a single user.
upvoted 2 times

 **danielmt** 1 year, 10 months ago

I would say C. High Concurrency with Auto Termination.
Although the default is no auto terminate we can still overwrite that setting.
upvoted 2 times

 **BK10** 1 year, 11 months ago

B is correct answer.
High concurrency cluster cannot AUTO terminated
upvoted 2 times

 **venkatibm** 2 years ago

it's correct
upvoted 1 times

HOTSPOT -

You are building an Azure Data Factory solution to process data received from Azure Event Hubs, and then ingested into an Azure Data Lake Storage Gen2 container.

The data will be ingested every five minutes from devices into JSON files. The files have the following naming pattern.

/{{deviceType}}/in/{{YYYY}}/{{MM}}/{{DD}}/{{HH}}/{{deviceID}}_{{YYYY}}{{MM}}{{DD}}{{HH}}{{mm}}.json

You need to prepare the data for batch data processing so that there is one dataset per hour per deviceType. The solution must minimize read times.

How should you configure the sink for 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

Parameter:

- @pipeline(),TriggerTime
- @pipeline(),TriggerType
- @trigger().outputs.windowStartTime
- @trigger().startTime

Naming pattern:

- /{deviceID}/out/{{YYYY}}/{{MM}}/{{DD}}/{{HH}}.json
- /{{YYYY}}/{{MM}}/{{DD}}/{{deviceType}}.json
- /{{YYYY}}/{{MM}}/{{DD}}/{{HH}}.json
- /{{YYYY}}/{{MM}}/{{DD}}/{{HH}}_{{deviceType}}.json

Copy behavior:

- Add dynamic content
- Flatten hierarchy
- Merge files

Answer Area

Parameter:

@pipeline(),TriggerTime
@pipeline(),TriggerType
@trigger().outputs.windowStartTime
@trigger().startTime

Naming pattern:

Correct Answer:

{deviceID}/out/{YYYY}/{MM}/{DD}/{HH}.json
/{YYYY}/{MM}/{DD}/{deviceType}.json
/{YYYY}/{MM}/{DD}/{HH}.json
/{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json

Copy behavior:

Add dynamic content
Flatten hierarchy
Merge files

Box 1: @trigger().startTime -

startTime: A date-time value. For basic schedules, the value of the startTime property applies to the first occurrence. For complex schedules, the trigger starts no sooner than the specified startTime value.

Box 2: /{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json

One dataset per hour per deviceType.

Box 3: Flatten hierarchy -

- FlattenHierarchy: All files from the source folder are in the first level of the target folder. The target files have autogenerated names.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers> <https://docs.microsoft.com/en-us/azure/data-factory/connector-file-system>

ItHYMeRish Highly Voted 2 years ago

The correct copy behavior is merge - not flatten hierarchy.

The question starts with a folder structure as the following:

{deviceType}/in/{YYYY}/{MM}/{DD}/{HH}/{deviceID}_{YYYY}{MM}{DD}{HH}{mm}.json

It indicates there are multiple device ID JSON files per deviceType. Those need to be merged to get the target naming pattern - "one file per device type per hour."

The target naming pattern is the following:
/{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json

The correct copy behavior is "Merge" because there are multiple files in the source folder that are merged into a single folder per device type per hour.

upvoted 97 times

Bro111 1 year, 1 month ago

Why not /{deviceType}/out/{YYYY}/{MM}/{DD}/{HH}.json ?

upvoted 3 times

sensaint 1 year, 1 month ago

It is not an option. It says /{deviceID}/out/{YYYY}/{MM}/{DD}/{HH}.json

upvoted 6 times

onyerleft Highly Voted 2 years ago

1) @trigger().outputs.windowStartTime - this output is from a tumbling window trigger, and is required to identify the correct directory at the /{HH}/ level. Using windowStartTime will give the hour with complete data. The @trigger().startTime is for a schedule trigger, which corresponds to the hour for which data has not arrived yet.

2) /{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json is the naming pattern to achieve an hourly dataset for each device type.

3) Multiple files for each device type will exist on the source side, since the naming pattern starts with {deviceID}... so the files must be merged in the sink to create a single file per device type.

upvoted 83 times

✉ **Davico93** 1 year, 6 months ago

but, the solution must minimize read times, I think is @trigger().startTime

upvoted 2 times

✉ **blazy002** Most Recent 2 weeks, 5 days ago

The files must be MERGED > each hour,
so on @trigger().outputs.windowStartTime => start time of the window

The author made 2/3 errors on this Q, grrr :)

@trigger().outputs.windowStartTime:

Gives the start time of the current window n

@trigger().StartTime:

Gives the start time of each trigger within that window n

upvoted 1 times

✉ **phydev** 2 months, 1 week ago

Was on my exam today (31.10.2023).

upvoted 4 times

✉ **Chemmangat** 3 months, 3 weeks ago

It's @trigger().outputs.windowStartTime

Ref : <https://learn.microsoft.com/en-us/azure/data-factory/control-flow-system-variables>

Under Tumbling Window Trigger

upvoted 1 times

✉ **kkk5566** 4 months ago

1) @trigger().outputs.windowStartTime

/{{YYYY}}/{{MM}}/{{DD}}/{{HH}}_{deviceType}.json

Merge

upvoted 1 times

✉ **pavankr** 6 months, 1 week ago

what exactly you want to "FLATTEN"??? You need to Merge files. period.

upvoted 2 times

✉ **rocky48** 7 months, 2 weeks ago

1) @trigger().outputs.windowStartTime

2) /{{YYYY}}/{{MM}}/{{DD}}/{{HH}}_{deviceType}.json

3) Merge

upvoted 9 times

✉ **rzenq** 1 year, 2 months ago

1. windowstarttime

2. yyyy/mm/dd/hh_devicetype.json

3. Merge

upvoted 6 times

✉ **Deeksha1234** 1 year, 5 months ago

1) @trigger().outputs.windowStartTime

2) /{{YYYY}}/{{MM}}/{{DD}}/{{HH}}_{deviceType}.json

3) Merge

agree with onyer

upvoted 3 times

✉ **Rafafouille76** 1 year, 10 months ago

Of course it is a merge, can't believe the official provided answers are so wrong ... Who wrote that

upvoted 9 times

✉ **kamil_k** 1 year, 9 months ago

I know it's almost as bad as Microsoft documentation about Azure.. That's why we see so much confusion over so many questions

upvoted 3 times

✉ **Jaws1990** 1 year, 11 months ago

Would you have to delay the tumbling processing by 60minutes to pick up data that hasn't arrived for that hour yet?

upvoted 1 times

✉ **Canary_2021** 2 years ago

The batch job runs in Data Factory should use Tumbling window trigger, so system variable trigger().outputs.windowStartTime should be passed in as the parameter.

upvoted 3 times

DRAG DROP -

You are designing an Azure Data Lake Storage Gen2 structure for telemetry data from 25 million devices distributed across seven key geographical regions. Each minute, the devices will send a JSON payload of metrics to Azure Event Hubs.

You need to recommend a folder structure for the data. The solution must meet the following requirements:

⇒ Data engineers from each region must be able to build their own pipelines for the data of their respective region only.

⇒ The data must be processed at least once every 15 minutes for inclusion in Azure Synapse Analytics serverless SQL pools.

How should you recommend completing the structure? 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 | Answer Area |
|----------------------------|-------------|
| {deviceID} | Value |
| {mm}/{HH}/{DD}/{MM}/{YYYY} | Value |
| {regionID}/{deviceID} | Value |
| {regionID}/raw | Value |
| {YYYY}/{MM}/{DD}/{HH} | Value |
| {YYYY}/{MM}/{DD}/{HH}/{mm} | Value |
| raw/{deviceID} | Value |
| raw/{regionID} | Value |

Correct Answer:

| Values | Answer Area |
|----------------------------|-------------|
| {deviceID} | Value |
| {mm}/{HH}/{DD}/{MM}/{YYYY} | Value |
| {regionID}/{deviceID} | Value |
| {regionID}/raw | Value |
| {YYYY}/{MM}/{DD}/{HH} | Value |
| {YYYY}/{MM}/{DD}/{HH}/{mm} | Value |
| raw/{deviceID} | Value |
| raw/{regionID} | Value |

Box 1: {raw/regionID}

Box 2: {YYYY}/{MM}/{DD}/{HH}/{mm}

Box 3: {deviceID}

Reference:

<https://github.com/paolosalvatori/StreamAnalyticsAzureDataLakeStore/blob/master/README.md>

By  **ItHYMeRish** Highly Voted  2 years ago

The correct answer is

{raw/regionID}/{YYYY}/{MM}/{DD}/{HH}/{mm}/{deviceID}.json

{raw/regionID} is the first level because raw is the container name for the raw data. RegionID follows it for ease of managing security.

{YYYY}/{MM}/{DD}/{HH}/{mm}/{deviceID}.json instead of {deviceID}/{YYYY}/{MM}/{DD}/{HH}/{mm}.json. The primary reason is that you want your namespace structure to have as few folders as high up and narrow those down as you get deeper into your structure.

For example, if you have 1 year worth of data and 25 million devices, using {YYYY}/{MM}/{DD}/{HH}/{mm}/ results in 2.1 million folders (1 year * 12 months * 30 days [estimate] * 24 hours * 60 minutes). If you start your folder structure with {deviceID}, you end up with 25 million folders - one for each device - before you even get to including the date in the hierarchy.

upvoted 191 times

By  **ML_Novice** 1 year, 4 months ago

ItHYMeRish you're a genius man

upvoted 4 times

By  **auwia** 6 months, 3 weeks ago

In IoT workloads, there can be a great deal of data being ingested 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}/

For example, landing telemetry for an airplane engine within the UK might look like the following structure:

UK/Planes/BA1293/Engine1/2017/08/11/12/

In this example, by putting the date at the end of the directory structure, you can use ACLs to more easily secure regions and subject matters to specific users and groups. If you put the date structure at the beginning, it would be much more difficult to secure these regions and subject matters. For example, if you wanted to provide access only to UK data or certain planes, you'd need to apply a separate permission for numerous directories under every hour directory. This structure would also exponentially increase the number of directories as time went on.

upvoted 2 times

✉ **auwia** 6 months, 3 weeks ago

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices#monitor-telemetry>

upvoted 3 times

✉ **nmm22** 9 months ago

thats such a cool explanation, i aspire to have the same critical thinking skills u have

upvoted 3 times

✉ **DataEX** 11 months ago

The correct structure answer will have 561.600 folders per year.

upvoted 1 times

✉ **gf2tw** **Highly Voted** 2 years, 1 month ago

raw/RegionId should be in the first box as raw is the name of your container. Furthermore, putting RegionId as one of the first foldernames allows easy partitioning and simpler RBAC for the Data Engineers.

upvoted 14 times

✉ **SAli12** 2 years ago

Yes I agree, raw/regionId --> timestamp --> deviceId.json

upvoted 5 times

✉ **auwia** **Most Recent** 6 months, 3 weeks ago

I'll follow best practice from Microsoft:

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices#monitor-telemetry>

So: /raw/regionid/deviceid/YYYY/MM/DD/HH

(without minutes).

upvoted 2 times

✉ **rocky48** 7 months, 2 weeks ago

The correct answer is

{raw/regionID}/{YYYY}/{MM}/{DD}/{HH}/{mm}/{deviceID}.json

upvoted 1 times

✉ **georgich87** 1 year, 9 months ago

I think that link will help us to find the correct answer:

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices>

The given example for a directory structure is: *{Region}/{SubjectMatter(s)}/{yyyy}/{mm}/{dd}/{hh}/*

upvoted 3 times

✉ **wwdba** 1 year, 9 months ago

{raw/regionID}/{YYYY}/{MM}/{DD}/{HH}/{mm}/{deviceID}.json

upvoted 2 times

✉ **staniopolis** 1 year, 10 months ago

IMHO {YYYY}/{MM}/{DD}/{HH}/{regionID/raw}/{deviceID}.json (given answer) is correct. Please pay attention that there is no minutes {mm} course it is not supported by Time format

<https://docs.microsoft.com/en-us/azure/stream-analytics/blob-storage-azure-data-lake-gen2-output>

upvoted 3 times

✉ **staniopolis** 1 year, 10 months ago

{raw/regionID}/{YYYY}/{MM}/{DD}/{HH}/{deviceID}.json

Time Format [optional]: if the time token is used in the prefix path, specify the time format in which your files are organized. Currently the only supported value is

HH.

upvoted 3 times

✉ **Canary_2021** 1 year, 11 months ago

Question 54: the correct answer of box 2 is {YYYY}/{MM}/{DD}/{HH}_{deviceType}.json

One dataset per hour per deviceType.

So looks like regionid and deviceid should be put after {YYYY}/{MM}/{DD}/{HH}/{mm} .

{YYYY}/{MM}/{DD}/{HH}/{mm}/{raw/regionID}/{deviceID}.json

upvoted 1 times

✉ **Canary_2021** 1 year, 11 months ago

Still feel {raw/RegionID} / {YYYY/MM/DD/mm} /{DeviceID} is correct. Just have some questions after compare answers of question 54.

upvoted 2 times

✉ **engrbrain** 2 years ago

The Question says : Each minute, the devices will send a JSON payload. That means the data is demarcated by region and by minutes.
{raw/RegionID} / {YYYY/MM/DD/mm} /{DeviceID}

upvoted 2 times

✉ **SabaJamal2010AtGmail** 2 years ago

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

upvoted 2 times

✉ **PA7** 2 years ago

raw/regionid -> DeviceId -> YYYY/MM/dd/HH-mm

upvoted 4 times

✉ **auwia** 6 months, 3 weeks ago

without minute info.

upvoted 2 times

✉ **mr_corte** 2 years ago

{raw/regionID}/{deviceID}/{YYYY}/{MM}/{DD}/{HH}{mm} imo.

upvoted 4 times

✉ **auwia** 6 months, 3 weeks ago

without minute in my opinion

upvoted 2 times

✉ **tsmk** 6 months ago

IMO, with {mm}.

Otherwise, every HH dir will have 25mil (device) * 60 (freq. of incoming files)

upvoted 1 times

HOTSPOT -

You are implementing an Azure Stream Analytics solution to process event data from devices.

The devices output events when there is a fault and emit a repeat of the event every five seconds until the fault is resolved. The devices output a heartbeat event every five seconds after a previous event if there are no faults present.

A sample of the events is shown in the following table.

| DeviceID | EventType | EventTime |
|--------------------------------------|------------------------|-----------------------|
| 78cc5ht9-w357-684r-w4fr-kr16h6p9874e | HeartBeat | 2020-12-01T19:00.000Z |
| 78cc5ht9-w357-684r-w4fr-kr16h6p9874e | HeartBeat | 2020-12-01T19:05.000Z |
| 78cc5ht9-w357-684r-w4fr-kr16h6p9874e | TemperatureSensorFault | 2020-12-01T19:07.000Z |

You need to calculate the uptime between the faults.

How should you complete the Stream Analytics SQL query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

SELECT

DeviceID,

MIN(EventTime) as StartTime,

MAX(EventTime) as EndTime,

DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds

FROM input TIMESTAMP BY EventTime

WHERE EventType='HeartBeat'

WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType

WHERE IsFirst(second,5) = 1

GROUP BY

DeviceID

,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID)

,TumblingWindow(second,5)

HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5

Answer Area

```
SELECT  
    DeviceID,  
    MIN(EventTime) as StartTime,  
    MAX(EventTime) as EndTime,  
    DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds
```

Correct Answer: FROM input TIMESTAMP BY EventTime

```
WHERE EventType='HeartBeat'  
WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType  
WHERE IsFirst(second,5) = 1
```

GROUP BY

DeviceID

```
,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID)  
,TumblingWindow(second,5)  
HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5
```

Box 1: WHERE EventType='HeartBeat'

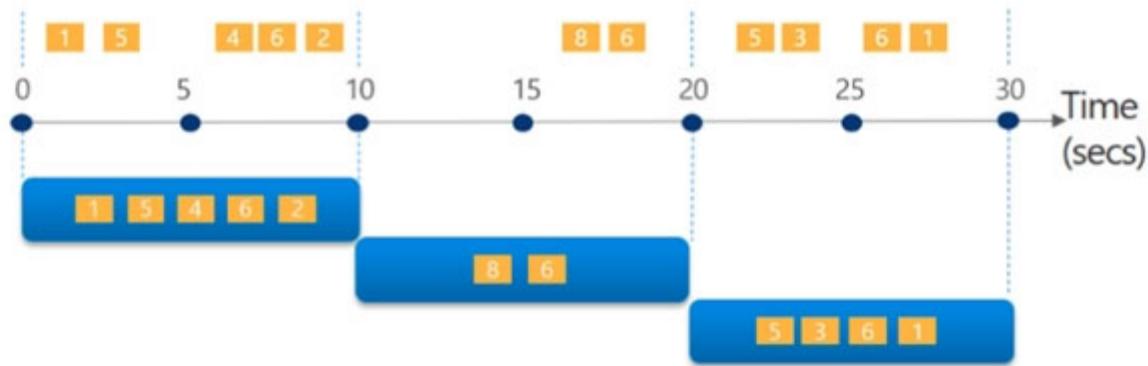
Box 2: ,TumblingWindow(Second, 5)

Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.

Tell me the count of tweets per time zone every 10 seconds

A 10-second Tumbling Window



```
SELECT TimeZone, COUNT(*) AS Count  
FROM TwitterStream TIMESTAMP BY CreatedAt  
GROUP BY TimeZone, TumblingWindow(second,10)
```

Incorrect Answers:

,SessionWindow.. : Session windows group events that arrive at similar times, filtering out periods of time where there is no data.

Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/session-window-azure-stream-analytics> <https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics>

Fer079 Highly Voted 2 years ago

I think the right answers should be WHERE EventType='HeartBeat' and Session window. If we want to calculate the uptime between the faults, we must use session window for each device, we know that will be receiving events for each 5 seconds if there is no error, so when an error occurs (or if we reach the maximum size of the window) then a new event will not be received within the next 5 seconds and the window will close, calculating the uptime. However if We use Tumbling window, it's not possible to calculate the uptime beyond 5 seconds

upvoted 80 times

✉ **yogiazaad** 11 months ago

This link is relevant here <https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns#session-windows>

upvoted 1 times

✉ **ovokpus** 1 year, 10 months ago

I concur!

upvoted 1 times

✉ **onyerleft** 2 years ago

Yes this sounds right

upvoted 2 times

✉ **Davico93** 1 year, 6 months ago

what happen if the event continues and the 50,000 second finishes? you cannot count that as a fault event

upvoted 1 times

✉ **Davico93** 1 year, 6 months ago

Sorry, you are right @Fer079!

upvoted 2 times

✉ **Canary_2021** [Highly Voted] 2 years ago

My answer is:

Question 1: B. Use LAG function as a filter to only filter out the events that switch from 'HeartBeat' to fault or switch from fault to 'HeartBeat'.

Question 2: C. No matter if there is a fault, device always sends message every 5min. Calculate the uptime between the faults don't need any window here. Any duration > 5s should between fault line and heartbeat line should be part of items that need to count into to calculate duration.

upvoted 22 times

✉ **Fer079** 1 year, 11 months ago

You cannot use the LAG function here because the "partition by" by deviceId is not included here, so the change between the status could be between different devices. This LAG function is evaluated before the "group by" clause of the query.

If you see the Microsoft documentation:

<https://docs.microsoft.com/en-us/stream-analytics-query/lag-azure-stream-analytics>

It says clearly that "LAG isn't affected by predicates in the WHERE clause, join conditions in the JOIN clause, or grouping expressions in the GROUP BY clause of the current query because it's evaluated before those clauses."

upvoted 8 times

✉ **mamahani** 8 months, 2 weeks ago

you do not need partition by with LAG function; it's an optional parameter; however in this scenario this is not the reason why we should not be using this function; with LAG we will receive in the query result only the "transition" events i.e. the device works correctly (eventtype='heartbeat') and then there is fault ('fault')-> we would receive only the record with "faul" (as it's different than previous line event i.e. heartbeat; by this one record we will not know how long the device was operational correctly, because we don't have these records anymore; we need to have 'startting' record for correctly operating device with heartbeat event , and this for every single "re-start" after the fault; LAG function would be good to calculate e.g. the increasing heartbeat by comparing the heartbeat of previous records with current one; but not in this user case;

upvoted 1 times

✉ **ubaldo1002** 1 year, 9 months ago

LAG does not require the PARTITION BY this is optional..

upvoted 2 times

✉ **kkk5566** [Most Recent] 4 months ago

WHERE EventType='HeartBeat' and Session window.

upvoted 2 times

✉ **rocky48** 7 months, 2 weeks ago

1. Where EventType = 'HeartBeat'
2. SessionWindow

upvoted 4 times

✉ **SinSS** 8 months, 1 week ago

1. Where EventType = 'HeartBeat'
2. SessionWindow

upvoted 1 times

✉ **dom271219** 1 year, 4 months ago

The where clause must be EventType != 'HeartBeat' otherwise you're not counting the uptime between the fault

upvoted 1 times

✉ **dom271219** 1 year, 4 months ago

Sorry ignore it

upvoted 1 times

✉ **Deeksha1234** 1 year, 5 months ago

Agree with Fer079 , EventType='HeartBeat' and Session window is correct

upvoted 1 times

✉  **uzairahm** 1 year, 6 months ago

WHERE EventType='HeartBeat' is definitely correct as you would need to filter out other events to calculate the uptime.
If you look at the example in link <https://docs.microsoft.com/en-us/stream-analytics-query/session-window-azure-stream-analytics> it would be crystal clear that sessionwindow is the right answer and @iooj (a lot of thanks) has already tested it
upvoted 1 times

✉  **sparkchu** 1 year, 9 months ago

tricky but instructive question.
upvoted 2 times

✉  **iooj** 1 year, 11 months ago

I created a Stream Analytics job and tested all combinations and here is my answer.
With a tumbling window, you will never be able to accumulate the correct interval.
The session is suitable here, but if the session closes earlier (by timeout) than the event occurs, then it will also fail to accumulate. So please note that in the timeout should be 6, not 5. A working version: EventType='HeartBeat' and SessionWindow(second, 6, 50000). But...
P.S. In the data example on the screenshot, the difference is generally indicated in minutes, in this situation, none of the answers will work, you will need to change seconds to minutes.

upvoted 13 times

店铺: IT认证考试服务

✉  **MadEgg** 1 year, 7 months ago

Thanks for testing it, but I think your conclusion is wrong.
We should calculate the difference (without any limitation). If you use SessionWindow with a timeout of 6 you limit this functionality. You get the right answer for the data in the table but what happens if you have a failure after >6 seconds?
I think Canary_2021 is right -> B, C

P.S. :-D didn't recognize it... but would say that this is a typo in the table.

upvoted 1 times

✉  **romanzdk** 1 year, 11 months ago

B and A?

upvoted 1 times

✉  **engrbrain** 2 years ago

The answer is BC. Every T_SQL Group by Query that needs to calculate max based on certain criteria should use the HAVING function to group that criteria

upvoted 5 times

✉  **MFR** 2 years ago

For me the session window suits for the given scenario. Also no device ID has been considered in the given answer, which is essential for calculating the uptime period per device

upvoted 4 times

店铺: IT认证考试服务

店铺: IT认证考试服务

You are creating a new notebook in Azure Databricks that will support R as the primary language but will also support Scala and SQL. Which switch should you use to switch between languages?

- A. %<language>
- B. @<Language >
- C. \\[<language >]
- D. \\(<language >)

Correct Answer: A

To change the language in Databricks' cells to either Scala, SQL, Python or R, prefix the cell with '%', followed by the language.

%python //or r, scala, sql

Reference:

<https://www.theta.co.nz/news-blogs/tech-blog/enhancing-digital-twins-part-3-predictive-maintenance-with-azure-databricks>

Community vote distribution

A (100%)

 **bad_atitude** Highly Voted 2 years ago

I wish you a DP203 as easy as this question folks
upvoted 49 times

 **CodingOwl** 1 year, 2 months ago

Username suits youe wis! :D
upvoted 2 times

 **anto69** 1 year, 11 months ago

We all hope man
upvoted 4 times

 **Deeksha1234** Most Recent 1 year, 5 months ago

Selected Answer: A

A is correct
upvoted 4 times

 **romanzdk** 1 year, 11 months ago

Selected Answer: A

Correct
upvoted 3 times

 **leandrors** 2 years ago

Selected Answer: A

Correct
upvoted 3 times

 **Will_KaiZuo** 2 years ago

Selected Answer: A

Correct
upvoted 4 times

You have an Azure Data Factory pipeline that performs an incremental load of source data to an Azure Data Lake Storage Gen2 account.

Data to be loaded is identified by a column named LastUpdatedDate in the source table.

You plan to execute the pipeline every four hours.

You need to ensure that the pipeline execution meets the following requirements:

- Automatically retries the execution when the pipeline run fails due to concurrency or throttling limits.
- Supports backfilling existing data in the table.

Which type of trigger should you use?

- A. event
- B. on-demand
- C. schedule
- D. tumbling window

Correct Answer: D

In case of pipeline failures, tumbling window trigger can retry the execution of the referenced pipeline automatically, using the same input parameters, without the user intervention. This can be specified using the property "retryPolicy" in the trigger definition.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger>

Community vote distribution

D (100%)

 **Canary_2021** Highly Voted  2 years ago

Selected Answer: D

D is correct answer.

<https://www.sqlshack.com/how-to-schedule-azure-data-factory-pipeline-executions-using-triggers/>

Azure Data Factory pipeline executions using Triggers:

- Schedule Trigger: The schedule trigger is used to execute the Azure Data Factory pipelines on a wall-clock schedule.
- Tumbling Window Trigger: Can be used to process history data. Also can define Delay, Max concurrency, retry policy etc.
- Event-Based Triggers : The event-based trigger executes the pipelines in response to a blob-related event, such as creating or deleting a blob file, in an Azure Blob Storage

upvoted 25 times

 **steveo123** Most Recent  6 months, 3 weeks ago

Selected Answer: D

D is correct.

upvoted 1 times

 **Ankit_Az** 7 months, 1 week ago

Selected Answer: D

Correct. As soon as you see backfill, its tumbling.

upvoted 1 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: D

D is correct

upvoted 1 times

 **Sriramiyer92** 1 year, 5 months ago

(D)

Tumbling window trigger

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers#tumbling-window-trigger>

Retry capability:

Is 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).

upvoted 3 times

 **dev2dev** 1 year, 11 months ago

Selected Answer: D

D is correct. Tumbling Window has more advance options for setting retry and concurrency policies which schedule doesn't have.

upvoted 2 times

You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account. The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/. You need to design a daily Azure Data Factory data load to minimize the data transfer between the two accounts. Which two configurations should you include in the design? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point

- A. Specify a file naming pattern for the destination.
- B. Delete the files in the destination before loading the data.
- C. Filter by the last modified date of the source files.
- D. Delete the source files after they are copied.

Correct Answer: AC

Copy only the daily files by using filtering.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage>

Community vote distribution

| | | |
|----------|----------|----|
| AC (76%) | AD (15%) | 9% |
|----------|----------|----|

 **Philipp** Highly Voted 1 year, 11 months ago

Selected Answer: AC

AC is correct, there is no point about deletion in source and might be the case that the data should stay in source too.

upvoted 14 times

 **necktru** Highly Voted 1 year, 8 months ago

Selected Answer: AC

I think the C option has impact in data transfer, B are incorrect, D is irrelevant for the question, and A is a complement of the task

upvoted 7 times

 **kkk5566** Most Recent 4 months ago

Selected Answer: AC

AC is correct

upvoted 1 times

 **Spinozabubble** 8 months, 1 week ago

A. Specify a file naming pattern for the destination:

By specifying a file naming pattern for the destination files in the Azure Data Lake Storage Gen2 account, you can ensure that the files are organized and stored in a structured manner. This can help with data management and subsequent processing.

C. Filter by the last modified date of the source files:

By filtering the source files based on the last modified date, you can select only the files that have been modified on the current day. This reduces the amount of data transferred and improves the efficiency of the data load process.

upvoted 5 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: AC

should be AC

upvoted 4 times

 **Boumisasound** 1 year, 10 months ago

I will go for AC

Why not D? Cause they are not mentionned some cost optimisation

upvoted 3 times

 **boopathi** 1 year, 10 months ago

AD are correct ?

upvoted 1 times

 **Istiaque** 1 year, 11 months ago

The requirement is to minimize the data transfer.

If we delete the files in source then there is no need to filter for daily load. So answer C,D is incorrect. Beside, there is no requirement to for minimizing the cost.

To my point of view, AC is correct because, even though filter by the modified date will take long time for lot of files, it won't impact the transfer.

upvoted 4 times

 **dev2dev** 1 year, 11 months ago

Selected Answer: AD

Normally we move the files after being processed, so it has to be D.

upvoted 5 times

 **yo1233** 1 year, 11 months ago

is A,D correct

upvoted 2 times

 **rainbowyu** 2 years ago

Should it be A &D as the requirement is to minimize the process time. Will option C take longer compared to D?

upvoted 2 times

 **djblue** 1 year, 10 months ago

Minimizing the process time is not part of the question. "Minimizing the data transfer", whatever that is - either time or amount.

upvoted 4 times

 **Canary_2021** 2 years ago

Selected Answer: CD

Either C or D can realize daily incremental load. Not sure why need to setup both of them.

upvoted 3 times

 **edba** 2 years ago

should it be C, D?

upvoted 2 times

 **Dusica** 12 months ago

YOU CAN'T GO WITHOUT A

upvoted 2 times

You plan to build a structured streaming solution in Azure Databricks. The solution will count new events in five-minute intervals and report only events that arrive during the interval. The output will be sent to a Delta Lake table.

Which output mode should you use?

- A. update
- B. complete
- C. append

Correct Answer: C

Append Mode: Only new rows appended in the result table since the last trigger are written to external storage. This is applicable only for the queries where existing rows in the Result Table are not expected to change.

Incorrect Answers:

B: Complete Mode: The entire updated result table is written to external storage. It is up to the storage connector to decide how to handle the writing of the entire table.

A: Update Mode: Only the rows that were updated in the result table since the last trigger are written to external storage. This is different from Complete Mode in that Update Mode outputs only the rows that have changed since the last trigger. If the query doesn't contain aggregations, it is equivalent to Append mode.

Reference:

<https://docs.databricks.com/getting-started/spark/streaming.html>

Community vote distribution

C (100%)

 **ANath** Highly Voted 1 year, 11 months ago

Correct Answer.

upvoted 8 times

 **kkk5566** Most Recent 4 months ago

Selected Answer: C

Append

upvoted 1 times

 **steveo123** 6 months, 3 weeks ago

Selected Answer: C

C is correct

upvoted 1 times

 **Jiaa** 1 year ago

C is correct

upvoted 2 times

 **Daniko** 1 year, 2 months ago

Selected Answer: C

C is correct

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

C is correct

upvoted 1 times

 **Remedios79** 1 year, 6 months ago

Append is correct

upvoted 1 times

 **bad_atitude** 2 years ago

agree with append

upvoted 4 times

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 Synapse Analytics dedicated SQL pool that contains a table named Table1.

You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of

Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: In an Azure Synapse Analytics pipeline, you use a data flow that contains a Derived Column transformation.

Does this meet the goal?

A. Yes

B. No

Correct Answer: A

Use the derived column transformation to generate new columns in your data flow or to modify existing fields.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column>

Community vote distribution

A (80%)

B (20%)

corebit [Highly Voted] 2 years ago

Selected Answer: A

"Data flows are available both in Azure Data Factory and Azure Synapse Pipelines"

"Use the derived column transformation to generate new columns in your data flow or to modify existing fields."

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column>

upvoted 22 times

Canary_2021 [Highly Voted] 1 year, 11 months ago

Selected Answer: B

Derived Column cannot get DateTime (created or lastmodified datetime) of the files.

Get Metadata activity can retrieve the DateTime of the files.

so answer should be B.

upvoted 6 times

Canary_2021 1 year, 11 months ago

If it is a real time process and pipeline is triggered to load data to table1 when file drop to container immediately, the created datetime of the file is similar as the pipeline process datetime. In this way Derived Column works.

The question is not clear.

upvoted 8 times

Jerrie86 11 months, 3 weeks ago

Can we just use the current datetime when the data is loaded. It doesn't say that we need to get data from the files. Just datetime which is kind of confusing. I will say, use derived column

upvoted 3 times

kkk5566 [Most Recent] 4 months, 2 weeks ago

A. Use this transformation to add any new columns to existing data.

upvoted 1 times

Deeksha1234 1 year, 5 months ago

correct

upvoted 2 times

Anandtr 1 year, 5 months ago

Selected Answer: A

Correct

upvoted 2 times

 **mkthoma3** 1 year, 6 months ago

What is the DateTime measuring? The DML transaction time or a file property?

If the measurement gives respect to the DML transaction time, you can use this: <https://docs.microsoft.com/en-us/azure/data-factory/data-flow-expressions-usage#currentTimestamp>

upvoted 1 times

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 Synapse Analytics dedicated SQL pool that contains a table named Table1.

You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of

Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: You use a dedicated SQL pool to create an external table that has an additional DateTime column.

Does this meet the goal?

A. Yes

B. No

Correct Answer: B

Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column>

Community vote distribution

B (100%)

Canary_2021 [Highly Voted] 2 years ago

Selected Answer: B

Answer should be B.

An external table is based on a source flat file structure. It seems to make no sense to add additional date time columns to such a table.

upvoted 18 times

AlejandroU [Most Recent] 1 month, 2 weeks ago

Answer B. The answer is incomplete because 2 additional steps were missing. After the 1st step which is creating the external table in the dedicated SQL pool with the additional DateTime column, the 2nd step is to load data using for example PolyBase to load data from the files in container1 into the external table of your dedicated SQL pool. 3rd step is to transform and insert once the data is in the dedicated SQL pool, and then insert the transformed data into your actual Table1, including the additional DateTime column.

upvoted 1 times

Deeksha1234 1 year, 5 months ago

Selected Answer: B

B is correct

upvoted 1 times

youngbug 1 year, 5 months ago

From the words in the Solution part, it seems to use PolyBase to read external tables. PolyBase can't change the schemas of external tables(files). You can only transform the data after loading data in the staging directory. And then load the data into tables.

upvoted 3 times

sdokmak 1 year, 7 months ago

Selected Answer: B

serverless works for data lake
dedicated doesn't

upvoted 2 times

GDJ2022 1 year, 11 months ago

Its clearly mentioned "You plan to insert data from the files in container1 into Table1". External tables dont get the data inserted into themselves, but instead refer outside data.

upvoted 4 times

edba 2 years ago

If using dedicated SQL pool, after creating an external table, need a further CTAS for adding derived columns.

upvoted 3 times

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 Synapse Analytics dedicated SQL pool that contains a table named Table1.

You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of

Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: You use an Azure Synapse Analytics serverless SQL pool to create an external table that has an additional DateTime column.

Does this meet the goal?

A. Yes

B. No

Correct Answer: B

Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column>

Community vote distribution

B (57%)

A (43%)

✉ **rainbowyu** Highly Voted 1 year, 11 months ago

You can't use serverless pool to create table in dedicate pool

upvoted 22 times

✉ **Knoushore1** Highly Voted 1 year, 2 months ago

Selected Answer: B

Table1 is in dedicated sql pool

upvoted 7 times

✉ **kkk5566** Most Recent 4 months ago

Selected Answer: B

should be no

upvoted 2 times

✉ **kkk5566** 4 months ago

correct to yes

upvoted 1 times

✉ **Ram9198** 5 months ago

Selected Answer: B

You can't use serverless pool to create table in dedicate pool

upvoted 1 times

✉ **AliakseiM** 5 months ago

Selected Answer: B

B since Table1 is in dedicated pool

upvoted 2 times

✉ **g2000** 5 months, 1 week ago

Selected Answer: A

You can use external tables to read external data using dedicated SQL pool or serverless SQL pool.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop>

upvoted 2 times

✉ **OldSchool** 1 year, 1 month ago

Selected Answer: A

Q:"You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1. You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1. You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1." Park for a while Table1 and dedicated SQL pool, that is where the transformation will happen AFTER loading from container1 to Table1. Here is about loading data to ADLSG2 continer1 and adding a column which can be done with serverless SQL as an external table.
upvoted 3 times

⊕ **berend1** 1 year, 2 months ago

if table 1 would be serverless, yes, now no
upvoted 1 times

⊕ **emna2022** 1 year, 3 months ago

The job is to insert data from the files in container1 into Table1 (in the dedicated sql pool) and transform the data after that and we need to add a new additional column.

External table are just references to the data, only metadata is really stored in the sql pool.
Hence anything including external table will be not a solution.

If you follow the different proposed solutions from previous questions, the most efficient solution is to use derived column transformation.
upvoted 3 times

⊕ **Deeksha1234** 1 year, 5 months ago

Selected Answer: A
yes, with serverless pool we can add a new column while creating an external table
upvoted 1 times

⊕ **youngbug** 1 year, 5 months ago

The aim of the solution is to load data from Data Lake's files to dedicated SQL pool's tables. There are three ways: DF's Copy Activity, PolyBase and Bulk insert. It's not serverless SQL pool's business...

upvoted 1 times

⊕ **StudentFromAus** 1 year, 6 months ago

The answer should be yes as we can create an additional column using CETAS in a serverless SQL pool though it is not a complete solution but a step closer to the required result.

upvoted 1 times

⊕ **sdokmak** 1 year, 7 months ago

Serverless pool works for data lake
Dedicated doesn't
upvoted 1 times

⊕ **nefarious_smalls** 1 year, 8 months ago

Apparently when dealing with dedicated sql pools you can only create an external table by importing the data from source using ctas. However, when using serverless using cetas will actually export a new file to your data source as well as create an external table. With that being said I think the answer is A.

upvoted 3 times

⊕ **Andushi** 1 year, 8 months ago

Selected Answer: A
Answer should be Yes
<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas#examples>
upvoted 3 times

⊕ **Billybob0604** 1 year, 1 month ago

it doesn't say in the link you can add a column using external table, so no.
upvoted 3 times

⊕ **ranjsi01** 1 year, 9 months ago

answer is Yes

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas>
upvoted 1 times

⊕ **g2000** 1 year, 8 months ago

Table1 is not an externa table
upvoted 1 times

⊕ **edba** 2 years ago

correct to me.
upvoted 4 times

⊕ **edba** 2 years ago

after further looking into it, I think the answer should be YES. pls refer to <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas#examples>
upvoted 6 times

✉ **Aditya0891** 1 year, 7 months ago

edba can you please suggest where in the link is it mentioned that you can use extra columns ?

upvoted 1 times

✉ **Aditya0891** 1 year, 7 months ago

Ignore my comments, I got your point thanks :)

upvoted 1 times

✉ **alex623** 1 year, 11 months ago

I think it's possible modify the files using cetas, but you have to create very much cetas to modify the files, so I think thw answer is no

upvoted 1 times

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 Synapse Analytics dedicated SQL pool that contains a table named Table1.

You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of

Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: In an Azure Synapse Analytics pipeline, you use a Get Metadata activity that retrieves the DateTime of the files.

Does this meet the goal?

A. Yes

B. No

Correct Answer: B

Instead use a serverless SQL pool to create an external table with the extra column.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables>

Community vote distribution

B (70%)

A (30%)

juanlu46 Highly Voted 1 year, 8 months ago

Selected Answer: B

Is part of a possible solution, but it isn't sufficient to meet the goal, yo need to pass the "Get metadata"'s output as a parameter to the ingest process, processing each file inside a "for" loop, for example.

<https://docs.microsoft.com/en-us/azure/data-factory/control-flow-get-metadata-activity>

upvoted 17 times

oldpony Highly Voted 1 year, 7 months ago

Selected Answer: A

<https://docs.microsoft.com/en-us/azure/data-factory/control-flow-get-metadata-activity>
points that Get Metadata activity can retrieve the corresponding Metadata type of: Created datetime of the file or folder.
upvoted 9 times

kkk5566 Most Recent 4 months ago

Selected Answer: B

nonono

upvoted 1 times

Ram9198 5 months ago

Selected Answer: B

Does not meet the goal

upvoted 1 times

vctrhugo 6 months, 4 weeks ago

According to ChatGPT:

B. No

The proposed solution of using a Get Metadata activity in an Azure Synapse Analytics pipeline will retrieve the DateTime of the files, but it does not address the requirement of storing the DateTime as an additional column in Table1.

upvoted 3 times

Qordata 3 months, 2 weeks ago

chatGpt is not to be trusted at all.

upvoted 2 times

esaade 10 months, 1 week ago

No, using a Get Metadata activity in an Azure Synapse Analytics pipeline to retrieve the DateTime of the files does not meet the goal of storing the DateTime as an additional column in Table1. The Get Metadata activity retrieves metadata about the files, such as file size, file name, or last modified date, but it does not provide the file content needed to extract the DateTime value and store it as an additional column in Table1. To achieve the goal, you need to use a data flow in the pipeline that loads the data from container1, extracts the DateTime value, and transforms the data by adding the DateTime column to Table1.

upvoted 5 times

 **OldSchool** 1 year, 1 month ago

If DateTime is part of data in files in container1 than answer is A, but if it is not part of data in files but only Meta data of files then B. Wording in question is really strange but I think it is A because it says "data from files in container1"

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

Its confusing, if we need to insert the dateTime of insertion then answer should be No, but if we need to insert the datetime of file modified then answer should be yes.

To me looks like the question is about 1st case so the answer should be No

upvoted 2 times

 **Dusica** 12 months ago

AGREED

upvoted 1 times

 **Strix** 1 year, 5 months ago

Selected Answer: B

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables>

upvoted 2 times

 **Davico93** 1 year, 6 months ago

I'm confusing more every time I read the solution, I don't know if it says that you have to do it in two steps, that changes everything

upvoted 1 times

 **MvanG** 1 year, 6 months ago

It seems rather odd that in the same two previous questions "Use the derived column transformation to generate new columns in your data flow or to modify existing fields." was the answer. This is very confusing.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column>

upvoted 3 times

 **g2000** 1 year, 8 months ago

Get Metadata seems possible

<https://www.mssqltips.com/sqlservertip/6246/azure-data-factory-get-metadata-example/>

upvoted 2 times

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 Data Lake Storage account that contains a staging zone.

You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes an Azure Databricks notebook, and then inserts the data into the data warehouse.

Does this meet the goal?

A. Yes

B. No

Correct Answer: B

If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity, not an Azure Databricks notebook, with your own data processing logic and use the activity in the pipeline. You can create a custom activity to run R scripts on your HDInsight cluster with R installed.

Reference:

<https://docs.microsoft.com/en-US/azure/data-factory/transform-data>

Community vote distribution

A (95%) 5%

juanlu46 Highly Voted 1 year, 8 months ago

Selected Answer: A

I think is A. Yes.

You can execute R code in a notebook, and then call it from Data Factory.

You can check it at "Databricks Notebook activity" header:

<https://docs.microsoft.com/en-US/azure/data-factory/transform-data>

And also:

<https://docs.microsoft.com/en-us/azure/databricks/spark/latest/sparkr/introduction>

upvoted 21 times

Deeksha1234 Highly Voted 1 year, 5 months ago

Selected Answer: A

answer should be A

upvoted 5 times

ExamDestroyer69 Most Recent 5 days, 11 hours ago

Selected Answer: A

VARIATIONS OF THIS QUESTION

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script. **NO**

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes an Azure Databricks notebook, and then inserts the data into the data warehouse. **YES**

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data flow, and then inserts the data into the data warehouse. **NO**

Solution: You schedule an Azure Databricks job that executes an R notebook, and then inserts the data into the data warehouse. **YES**
upvoted 2 times

kkk5566 4 months ago

Selected Answer: A

yes can do it

upvoted 1 times

vctrhugo 6 months, 4 weeks ago

Selected Answer: A

A. Yes

The proposed solution meets the goal of designing a daily process to ingest incremental data from the staging zone, transform the data using an R script, and insert the transformed data into a data warehouse in Azure Synapse Analytics. The solution involves using an Azure Data Factory (ADF) schedule trigger to execute a pipeline that executes an Azure Databricks notebook and then inserts the data into the data warehouse.

upvoted 3 times

 **esaade** 10 months, 1 week ago

Yes, this solution meets the goal of ingesting incremental data from the staging zone, transforming the data by executing an R script, and inserting the transformed data into a data warehouse in Azure Synapse Analytics. By using an Azure Data Factory schedule trigger, you can schedule the pipeline to run on a daily basis. The pipeline can execute an Azure Databricks notebook, which can perform the transformation using R scripts, and then insert the transformed data into the data warehouse.

upvoted 4 times

 **vrodriguesp** 1 year ago

Selected Answer: A
yes, you can execute R script in notebook and call it via adf

upvoted 4 times

 **urielramoss** 1 year, 1 month ago

Selected Answer: A
the answer is YES. I already used this solution in a previous project.

upvoted 4 times

 **rzeng** 1 year, 2 months ago

should be YES
upvoted 2 times

 **dom271219** 1 year, 4 months ago

Selected Answer: A
We do sth like it in my company
upvoted 4 times

 **Sriramiyer92** 1 year, 5 months ago

Selected Answer: A
A.
R Language is supported in ADB.
ADB notebooks, can be called from ADF pipeline(Use Notebook Activity) to link to the ADB notebook
upvoted 3 times

 **Davico93** 1 year, 6 months ago

I don't know guys, it's kind of tricky, in 2 next questions, it says "inser the TRANSFORMED data" and here it says jus "DATA".... what do you think?
upvoted 2 times

 **evega** 1 year, 7 months ago

Selected Answer: U
Para mi es la respuesta A. En un pipeline de ADF puede tener una actividad de notebook para databricks, el cual permitirá ejecutar el notebook una vez al día a través de un trigger.
upvoted 3 times

 **OCHT** 1 year, 8 months ago

Selected Answer: A
R in notebook and call via Data Factory
upvoted 4 times

 **MS_Nikhil** 1 year, 8 months ago

Selected Answer: A
You can execute R code in a notebook.
upvoted 4 times

 **hbad** 1 year, 8 months ago

The correct answer should be No, based on the how it is worded and the following logic:

In Azure Data Factory a Databricks Activity can be used to execute a Databricks notebook. However, it cannot pass the data along to the next activity (dbutils.notebook.exit("returnValue") only passes a string). Given that the way this is worded it says " execute a pipeline that executes an Azure Databricks notebook, and then inserts the data " the "then" implies a next step which wont work as cant pass the data along. If the transformation and insert both happened in the notebook only then it would work.

<https://docs.microsoft.com/en-US/azure/data-factory/transform-data-databricks-notebook>

upvoted 2 times

 **nefarious_smalls** 1 year, 8 months ago

Yea but you do not have to pass the data along in ADF. You can insert it into Synapse from the notebook.
upvoted 2 times

 **hbad** 1 year, 7 months ago

precisely my point, either both things (R and Insert) should be in the one workbook OR you need two workbooks. The wording indicates 2 steps rather than all in one book: "notebook" being the first step and "then" indicating another step.

upvoted 2 times

 **Igor85** 1 year, 1 month ago

i don't see any problem to run R and write Synapse dedicated SQL pool in the same notebook

<https://learn.microsoft.com/en-us/azure/databricks/external-data/synapse-analytics>

upvoted 2 times

 **romega2** 1 year, 8 months ago

Selected Answer: A

I agree that Yes

upvoted 2 times

 **gauravgogs** 1 year, 8 months ago

I think it should be Yes. i.e. A

R Script is well supported by databricks notepad

upvoted 3 times

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 Data Lake Storage account that contains a staging zone.

You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data flow, and then inserts the data into the data warehouse.

Does this meet the goal?

A. Yes

B. No

Correct Answer: B

If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity, not a mapping flow,⁵ with your own data processing logic and use the activity in the pipeline. You can create a custom activity to run R scripts on your HDInsight cluster with R installed.

Reference:

<https://docs.microsoft.com/en-US/azure/data-factory/transform-data>

Community vote distribution

B (100%)

 juanlu46 Highly Voted 1 year, 8 months ago

Selected Answer: B

Is correct.

Mapping Dataflows can't execute R code that is a requirement, so not meet the goal.

upvoted 8 times

 ExamDestroyer69 Most Recent 5 days, 11 hours ago

Selected Answer: B

VARIATIONS OF THIS QUESTION

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script. **NO**

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes an Azure Databricks notebook, and then inserts the data into the data warehouse. **YES**

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data flow, and then inserts the data into the data warehouse. **NO**

Solution: You schedule an Azure Databricks job that executes an R notebook, and then inserts the data into the data warehouse. **YES**
upvoted 1 times

 kkk5566 4 months ago

Selected Answer: B

no is correct

upvoted 1 times

 dom271219 1 year, 4 months ago

Selected Answer: B

There is no R in ADF dataflow

upvoted 4 times

 Deeksha1234 1 year, 5 months ago

Selected Answer: B

B is right

upvoted 2 times

 Remedios79 1 year, 6 months ago

The answer is no.

upvoted 3 times

店铺：IT认证考试服务

店铺：IT认证考试服务

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 Data Lake Storage account that contains a staging zone.

You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.

Solution: You schedule an Azure Databricks job that executes an R notebook, and then inserts the data into the data warehouse.

Does this meet the goal?

A. Yes

B. No

Correct Answer: B

Must use an Azure Data Factory, not an Azure Databricks job.

Reference:

<https://docs.microsoft.com/en-US/azure/data-factory/transform-data>

Community vote distribution

A (100%)

juanlu46 Highly Voted 1 year, 8 months ago

Selected Answer: A

The correct answer is "A. Yes"

You can execute R code in a notebook, and then call it from Data Factory.

You can check it at "Databricks Notebook activity" header:

<https://docs.microsoft.com/en-US/azure/data-factory/transform-data>

And also:

<https://docs.microsoft.com/en-us/azure/databricks/spark/latest/sparkr/overview>

upvoted 13 times

juanlu46 1 year, 8 months ago

I'm Sorry, in the statement there isn't mention to "Data factory", but you can use a Databrick's job also, therefore the solution meet the goal.
<https://docs.microsoft.com/en-us/azure/databricks/jobs#--run-a-job>

upvoted 12 times

bp_a_user 8 months, 2 weeks ago

...but where is the ingest done?

upvoted 1 times

ExamDestroyer69 Most Recent 5 days, 11 hours ago

Selected Answer: A

VARIATIONS OF THIS QUESTION

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script. **NO**

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes an Azure Databricks notebook, and then inserts the data into the data warehouse. **YES**

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data flow, and then inserts the data into the data warehouse. **NO**

Solution: You schedule an Azure Databricks job that executes an R notebook, and then inserts the data into the data warehouse. **YES**

kkk5566 4 months ago

Selected Answer: A

Yes, this solution would meet the goal.

upvoted 1 times

esaade 10 months, 1 week ago

Yes, this solution would meet the goal. An Azure Databricks job can be scheduled to run on a regular basis, such as daily, and can execute an R notebook that reads data from Azure Data Lake Storage, transforms the data using R code, and then writes the transformed data to the data warehouse in Azure Synapse Analytics.

upvoted 3 times

 **vrodriguesp** 1 year ago

Selected Answer: A

should be yes, you can schedule notebook directly from databricks

upvoted 3 times

 **lemonpotato** 1 year ago

Selected Answer: A

Has to be Yes

upvoted 1 times

 **XiltroX** 1 year, 1 month ago

The Answer is A. You can only execute R notebook in Databricks and not in Data Factory. The key word here is Databricks.

upvoted 1 times

 **greenlever** 1 year, 2 months ago

Selected Answer: A

1. extract data from Azure Data Lake Storage Gen2 into Azure Databricks,

2. run transformations on the data in Azure Databricks,

3. load the transformed data into Azure Synapse Analytics.

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: A

yes, its possible

upvoted 1 times

 **demirsamuel** 1 year, 7 months ago

Selected Answer: A

I go for A as well

upvoted 2 times

 **observador081** 1 year, 7 months ago

You have an Azure subscription that includes the following resources:

VNet1, a virtual network

Subnet1, a subnet in VNet1

WebApp1, a web app application service

NSG1, a network security group

You create an application security group named ASG1.

Which resource can use ASG1?

Selecione somente uma resposta.

VNet1

Subnet1

WebApp1

NSG1

upvoted 2 times

 **allagowf** 1 year, 2 months ago

the answer is : VNet1

upvoted 1 times

 **cuongthh** 1 year, 7 months ago

Selected Answer: A

I go for A.

upvoted 2 times

 **HoangTr** 1 year, 7 months ago

I go for A.

Databrick should have an option to trigger the job on selected schedule, it doesn't need data factory to trigger.

upvoted 2 times

 **KHawk** 1 year, 8 months ago

I would go for No. You can create a Spark Submit Job to run R Code but as shown in the second link, Databricks Utilities is not supported which would be necessary in my opinion to connect to Data Lake

<https://docs.microsoft.com/en-us/azure/databricks/jobs>

What do you think ?

<https://docs.microsoft.com/en-us/azure/databricks/dev-tools/api/latest/examples#spark-submit-api-example-r>

upvoted 2 times

✉️ **Davico93** 1 year, 6 months ago

you made me doubt about it
upvoted 1 times

✉️ **Andushi** 1 year, 8 months ago

Selected Answer: A

The solution meet the goal
upvoted 2 times

Question #66

Topic 2

You plan to create an Azure Data Factory pipeline that will include a mapping data flow.

You have JSON data containing objects that have nested arrays.

You need to transform the JSON-formatted data into a tabular dataset. The dataset must have one row for each item in the arrays.

Which transformation method should you use in the mapping data flow?

- A. new branch
- B. unpivot
- C. alter row
- D. flatten

Correct Answer: D

Use the flatten transformation to take array values inside hierarchical structures such as JSON and unroll them into individual rows. This process is known as denormalization.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-flatten>

Community vote distribution

D (100%)

✉️ **gauravgogs** **Highly Voted** 1 year, 8 months ago

Correct

upvoted 7 times

✉️ **juanlu46** **Highly Voted** 1 year, 8 months ago

Selected Answer: D

Is correct

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-flatten>

upvoted 6 times

✉️ **kkk5566** **Most Recent** 4 months ago

Selected Answer: D

Correct

upvoted 1 times

✉️ **Deeksha1234** 1 year, 5 months ago

Selected Answer: D

D is correct

upvoted 3 times

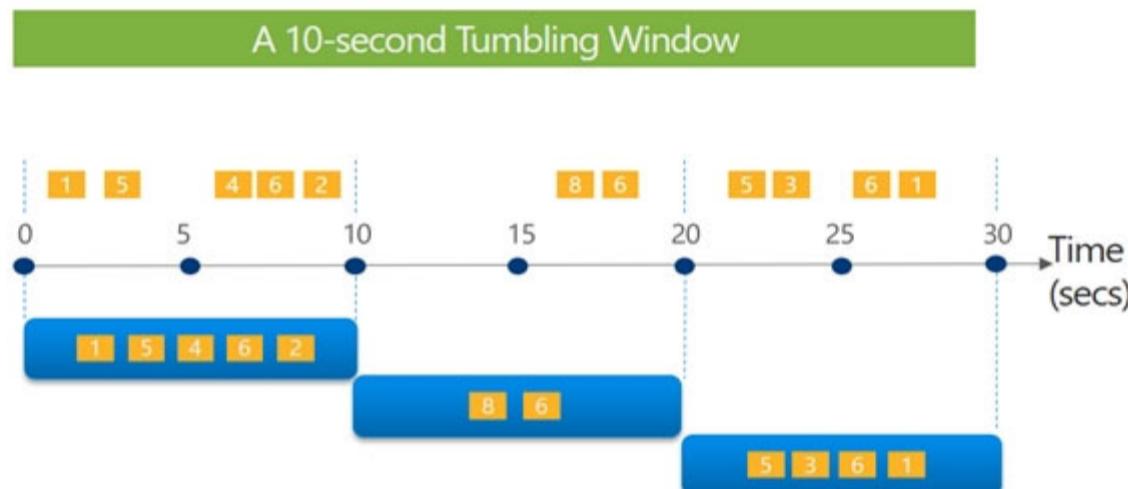
You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account. You need to output the count of tweets during the last five minutes every five minutes. Each tweet must only be counted once. Which windowing function should you use?

- A. a five-minute Sliding window
- B. a five-minute Session window
- C. a five-minute Hopping window that has a one-minute hop
- D. a five-minute Tumbling window

Correct Answer: D

Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

Tell me the count of Tweets per time zone every 10 seconds



```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

Community vote distribution

D (100%)

Remedios79 Highly Voted 1 year, 6 months ago

corrett. It would be corret also a hopping window with hop and size both to 5 seconds
upvoted 6 times

Kavya_sri Most Recent 1 month ago

correct
upvoted 1 times

kkk5566 4 months ago

Selected Answer: D
repeated
upvoted 1 times

Kezzah 1 year, 4 months ago

Selected Answer: D
correct
upvoted 3 times

Deeksha1234 1 year, 5 months ago

Selected Answer: D
correct

upvoted 2 times

 **nefarious_smalls** 1 year, 7 months ago

correct

upvoted 2 times

 **juanlu46** 1 year, 8 months ago

Selected Answer: D

Is correct

upvoted 3 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

You are planning a streaming data solution that will use Azure Databricks. The solution will stream sales transaction data from an online store. The solution has the following specifications:

The output data will contain items purchased, quantity, line total sales amount, and line total tax amount.

- Line total sales amount and line total tax amount will be aggregated in Databricks.

- Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.

You need to recommend an output mode for the dataset that will be processed by using Structured Streaming. The solution must minimize duplicate data.

What should you recommend?

- A. Update
- B. Complete
- C. Append

Correct Answer: A

By default, streams run in append mode, which adds new records to the table.

Incorrect Answers:

B: Complete mode: replace the entire table with every batch.

Reference:

<https://docs.databricks.com/delta/delta-streaming.html>

Community vote distribution

C (62%)

A (38%)

✉  **necktru** Highly Voted 1 year, 8 months ago

Selected Answer: A

I think Update is correct, because " new rows will be added to adjust a sale" , that means that in the course of a day you must update de daily import with the new sales, the group by process generates new amounts, keep in mind that when it say "sales transactions will never be updated" its about the online store, not the aggregated rows.

upvoted 17 times

✉  **vctrhugo** 6 months, 2 weeks ago

Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.

upvoted 6 times

✉  **[Removed]** Highly Voted 11 months, 3 weeks ago

Selected Answer: C

Using chatgpt : Append

upvoted 13 times

✉  **jsav1** Most Recent 1 week, 1 day ago

Selected Answer: C

Append: you are only adding new rows and existing rows do not need to be updated

upvoted 1 times

✉  **dakku987** 1 week, 5 days ago

Selected Answer: C

when you see new rows will be added to APPEND is always the answers

upvoted 1 times

✉  **d046bc0** 3 weeks, 4 days ago

Selected Answer: A

(ChatGPT) The Append output mode is used when new rows are added to the result table. This mode is suitable for scenarios where the output table is a summary of the input data, and the input data is not updated

upvoted 1 times

✉  **dawoodiee** 3 months, 2 weeks ago

Sales transactions will NEVER be updated.

Append.

upvoted 1 times

 **Ram9198** 4 months ago

Selected Answer: A

Update

upvoted 1 times

 **EliteAllen** 4 months ago

Selected Answer: C

C. Append

This mode is used when you are always adding new records to the output data. Given that sales transactions will never be updated and new rows will be added to adjust a sale, this mode seems to be the most suitable. It will also help in minimizing duplicate data since it only adds new records and does not modify existing ones.

upvoted 1 times

 **kkk5566** 4 months ago

Selected Answer: A

A is correct 确

upvoted 1 times

 **kkk5566** 4 months ago

ignore it

upvoted 1 times

 **Tightbot** 4 months, 3 weeks ago

Selected Answer: C

Append is the right choice . Update is for modifications and append is to add new rows

upvoted 1 times

 **Ram9198** 5 months ago

Selected Answer: A

When say adjusted, it means update because we need to reaggregate to get the latest total amount after adjustment, also there is a hint to minimise duplicates.. Append will ignore the updated state records only emit new records

upvoted 1 times

 **wanchihh** 3 months, 4 weeks ago

The question specifically stated "Sales transactions will never be updated".

upvoted 1 times

 **pavankr** 6 months, 1 week ago

The requirement says "not to update", your answer says "update"???????

upvoted 2 times

 **vctrhugo** 6 months, 4 weeks ago

Selected Answer: C

C. Append

For the given scenario, where sales transactions are never updated but new rows are added to adjust a sale, the recommended output mode for the dataset processed by using Structured Streaming in Azure Databricks is "Append".

The "Append" output mode ensures that only new rows are added to the output data as they arrive in the streaming data source. It appends the new rows to the existing result without modifying or updating previously processed data. This mode is suitable when you want to continuously append new records to the output data without duplicating or modifying existing data.

In this case, as new rows are added to adjust a sale, the "Append" mode will capture these new rows and include them in the output data, allowing you to aggregate the line total sales amount and line total tax amount in Databricks while minimizing duplicate data.

upvoted 4 times

 **MarkJoh** 1 month ago

This is definitely the correct answer. The statement "Sales transactions will never be updated. Instead, new rows will be added to adjust a sale." is not about the actions that "you do", it's about what the incoming rows will look like. There will never be an update of a row, if an update is needed, a new row will come in as an adjustment. So, it's kind of a trick question. The answer is Append.

upvoted 1 times

 **Ankit_Az** 7 months, 1 week ago

I feel Append is correct here

upvoted 3 times

 **janaki** 7 months, 2 weeks ago

It's Append as the 3rd instruction says

Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.

So it's not UPDATE but an APPEND

upvoted 4 times

 **esaade** 10 months, 1 week ago

Selected Answer: C

I would recommend using the "Append" output mode for the dataset processed by using Structured Streaming in this scenario.

The "Append" output mode is appropriate when the output dataset is a set of new records and does not include any updates or deletions. It will only append new rows to the output dataset, which means there will be no duplicate data created as a result of the streaming data solution. Since the solution will never update existing rows, but rather add new rows, the "Append" mode is the best choice to meet the requirements.

upvoted 7 times

 **Rakrah** 11 months ago

Very Correct Answer is "APPEND" MODE - Because Sales transaction never be updated using Update Mode, would not provide any benefits, rather "Append" mode will be add new row to the output dataset and correctly aggregate the line total sales amount and line total tax amount without any duplicates. So Append mode 200% meet the requirement.

upvoted 5 times

You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1.

You need to determine the size of the transaction log file for each distribution of DW1.

What should you do?

- A. On DW1, execute a query against the sys.database_files dynamic management view.
- B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.
- C. Execute a query against the logs of DW1 by using the Get-AzOperationalInsightsSearchResult PowerShell cmdlet.
- D. On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.

Correct Answer: A

For information about the current log file size, its maximum size, and the autogrow option for the file, you can also use the size, max_size, and growth columns for that log file in sys.database_files.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/logs/manage-the-size-of-the-transaction-log-file>

Community vote distribution

| | |
|---------|---------|
| D (78%) | A (22%) |
|---------|---------|

✉ **Saransundar** Highly Voted 1 year, 7 months ago

The question asks for transaction log size on each distribution. The correct answer is D: Link below: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor>

-- Transaction log size

```
SELECT
instance_name as distribution_db,
cntr_value*1.0/1048576 as log_file_size_used_GB,
pdw_node_id
FROM sys.dm_pdw_nodes_os_performance_counters
WHERE
instance_name like 'Distribution_%'
AND counter_name = 'Log File(s) Used Size (KB)'
```

upvoted 18 times

✉ **Davico93** 1 year, 6 months ago

but you don't need it from master, just DW1

upvoted 4 times

✉ **Saim8711** Highly Voted 1 year, 6 months ago

Selected Answer: D

D is totally correct. Link has this very clearly mentioned

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor>

upvoted 9 times

✉ **Sachmett** Most Recent 3 weeks, 6 days ago

Selected Answer: A

Table sys.dm_pdw_nodes_os_performance_counters contains information about current size of file log each distribution.
You can use sys.database_files to determine size of file log of DW1 (each distribution the same).

upvoted 1 times

✉ **kkk5566** 4 months ago

Selected Answer: D

Should be D

upvoted 1 times

✉ **Ram9198** 5 months ago

Selected Answer: D

A is wrong it applies for SQL server and non distributed non MPP database.. question clearly says per distribution and synapse

upvoted 1 times

✉ **pavankr** 6 months, 1 week ago

the question is about distribution, so D should be answer.

upvoted 1 times

✉ **vctrhugo** 6 months, 2 weeks ago

```
-- Transaction log size
SELECT
instance_name as distribution_db,
cntr_value*1.0/1048576 as log_file_size_used_GB,
pdw_node_id
FROM sys.dm_pdw_nodes_os_performance_counters
WHERE
instance_name like 'Distribution_%'
AND counter_name = 'Log File(s) Used Size (KB)'
```

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor#monitor-transaction-log-size>

upvoted 2 times

✉ **vctrhugo** 6 months, 2 weeks ago

This query returns the transaction log size on each distribution.

upvoted 1 times

✉ **auwia** 6 months, 3 weeks ago

Selected Answer: D

Probably A and D are correct, but I would choose D, because it's clearly described as the question:

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor#monitor-transaction-log-size>

upvoted 2 times

✉ **vctrhugo** 6 months, 4 weeks ago

Selected Answer: D

Monitor transaction log size

The following query returns the transaction log size on each distribution. If one of the log files is reaching 160 GB, you should consider scaling up your instance or limiting your transaction size.

```
-- Transaction log size
SELECT
instance_name as distribution_db,
cntr_value*1.0/1048576 as log_file_size_used_GB,
pdw_node_id
FROM sys.dm_pdw_nodes_os_performance_counters
WHERE
instance_name like 'Distribution_%'
AND counter_name = 'Log File(s) Used Size (KB)'
```

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor#monitor-transaction-log-size>

upvoted 1 times

✉ **TestingCRM** 7 months, 2 weeks ago

D. See this article <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor#monitor-transaction-log-size>

upvoted 1 times

✉ **agold96** 11 months, 4 weeks ago

Selected Answer: A

According to the documentation:

"For information about the current log file size, its maximum size, and the autogrow option for the file, you can also use the size, max_size, and growth columns for that log file in sys.database_files."

A seems enough, I am not sure it gives the results for each distribution but it seems so.

upvoted 2 times

✉ **cokey** 1 year, 1 month ago

Selected Answer: D

i think "D"

upvoted 1 times

✉ **allagowf** 1 year, 2 months ago

Selected Answer: D

Answer is On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.

The following query returns the transaction log size on each distribution. If one of the log files is reaching 160 GB, you should consider scaling up your instance or limiting your transaction size.

```
-- Transaction log size
SELECT
instance_name as distribution_db, cntr_value*1.0/1048576 as log_file_size_used_GB, pdw_node_id
FROM sys.dm_pdw_nodes_os_performance_counters
WHERE
instance_name like 'Distribution_%'
AND counter_name = 'Log File(s) Used Size (KB)'
```

References:

<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-manage-monitor>

upvoted 2 times

✉ **ads5891** 1 year, 5 months ago

Selected Answer: D

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor#monitor-transaction-log-size>

upvoted 2 times

✉ **youngbug** 1 year, 5 months ago

DW is a distributed system, and you can run view queries on any node. So it doesn't matter on the master database.

upvoted 1 times

✉ **zxc01** 1 year, 5 months ago

I cannot find any correct answer if the question is correct. Someone said D, but how can you run it in master database? you should execute it in DW1. you will get error message if you run it in master database "Invalid object name 'sys.dm_pdw_nodes_os_performance_counters'." it is correct if change ~~answer~~ D to execute on DW1.

upvoted 3 times

✉ **Saim8711** 1 year, 6 months ago

D is totally correct. Link has this very clearly mentioned

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor>

The following query returns the transaction log size on each distribution. If one of the log files is reaching 160 GB, you should consider scaling up your instance or limiting your transaction size.

upvoted 3 times

✉ **MS2710** 1 year ago

A is also close, but D wil give exact answer (log size for each distribution). Not sure if same can be achieved using A.

upvoted 1 times

You are designing an anomaly detection solution for streaming data from an Azure IoT hub. The solution must meet the following requirements:

- Send the output to Azure Synapse.
- Identify spikes and dips in time series data.
- Minimize development and configuration effort.

Which should you include in the solution?

- A. Azure Databricks
- B. Azure Stream Analytics
- C. Azure SQL Database

Correct Answer: B

You can identify anomalies by routing data via IoT Hub to a built-in ML model in Azure Stream Analytics.

Reference:

<https://docs.microsoft.com/en-us/learn/modules/data-anomaly-detection-using-azure-iot-hub/>

Community vote distribution

B (100%)

□  **SAYAK7** Highly Voted 1 year, 7 months ago

Selected Answer: B

Obviously B, IoT is event hub of stream data so we need stream analytics for sure.
upvoted 11 times

□  **kkk5566** Most Recent 4 months ago

Selected Answer: B

B is correct
upvoted 1 times

□  **Ankit_Az** 7 months, 1 week ago

Selected Answer: B

Correct
upvoted 1 times

□  **Coderhbti** 8 months, 4 weeks ago

Selected Answer: B

B is correct
upvoted 2 times

□  **Deeksha1234** 1 year, 5 months ago

B is correct
upvoted 1 times

□  **Remedios79** 1 year, 6 months ago

i agree
upvoted 2 times

A company uses Azure Stream Analytics to monitor devices.

The company plans to double the number of devices that are monitored.

You need to monitor a Stream Analytics job to ensure that there are enough processing resources to handle the additional load.

Which metric should you monitor?

- A. Early Input Events
- B. Late Input Events
- C. Watermark delay
- D. Input Deserialization Errors

Correct Answer: C

There are a number of resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to:

- ⇒ Not enough processing resources in Stream Analytics to handle the volume of input events.
- ⇒ Not enough throughput within the input event brokers, so they are throttled.
- ⇒ Output sinks are not provisioned with enough capacity, so they are throttled. The possible solutions vary widely based on the flavor of output service being used.

Incorrect Answers:

A: Deserialization issues are caused when the input stream of your Stream Analytics job contains malformed messages.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling>

Community vote distribution

C (100%)

✉ **nicky87654** Highly Voted 11 months, 4 weeks ago

C-->it measures the amount of delay in the processing of the input events. If the watermark delay increases, it could indicate that the Stream Analytics job is not able to keep up with the incoming data and may not have enough processing resources to handle the additional load.
upvoted 6 times

✉ **kkk5566** Most Recent 4 months ago

Selected Answer: C

sure it

upvoted 1 times

✉ **MScapris** 1 year ago

Selected Answer: C

correct !

upvoted 2 times

✉ **Deeksha1234** 1 year, 5 months ago

correct

upvoted 2 times

✉ **dsp17** 1 year, 6 months ago

Watermark delay - correct

upvoted 2 times

✉ **sagur** 1 year, 7 months ago

Selected Answer: C

seems ok

upvoted 4 times

HOTSPOT -

You are designing an enterprise data warehouse in Azure Synapse Analytics that will store website traffic analytics in a star schema.

You plan to have a fact table for website visits. The table will be approximately 5 GB.

You need to recommend which distribution type and index type to use for the table. The solution must provide the fastest query performance.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

店铺: IT认证考试服务
Distribution:

- Hash
- Round robin
- Replicated

Index:

- Clustered columnstore
- Clustered
- Nonclustered

Answer Area

Distribution:

- Hash
- Round robin
- Replicated

Correct Answer:

Index:

- Clustered columnstore
- Clustered
- Nonclustered

Box 1: Hash -

Consider using a hash-distributed table when:

The table size on disk is more than 2 GB.

The table has frequent insert, update, and delete operations.

Box 2: Clustered columnstore -

Clustered columnstore tables offer both the highest level of data compression and the best overall query performance.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-index>

 **Deeksha1234** Highly Voted  1 year, 5 months ago

Hash and clustered columnstore..right

upvoted 10 times

 **Ankit_Az** Most Recent  7 months, 1 week ago

Correct

hash and CCI

upvoted 4 times

 **objecto** 1 year, 6 months ago

I'm not sure about the hash distribution since we don't have enough information on what columns we get. In any case I would choose Round Robin to just have a even distribution.

upvoted 1 times

 **vctrhugo** 6 months, 4 weeks ago

Anything above 2GB you should go with Hash. Round Robin/Heap is for staging tables.

upvoted 4 times

 **henryphchan** 8 months ago

Round Robin is used for Staging Table

upvoted 1 times

 **Revave2** 1 year, 6 months ago

I would go with Hash as the table is >2gb and is a fact table...

upvoted 9 times

You have an Azure Stream Analytics job.

You need to ensure that the job has enough streaming units provisioned.

You configure monitoring of the SU % Utilization metric.

Which two additional metrics should you monitor? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Backlogged Input Events
- B. Watermark Delay
- C. Function Events
- D. Out of order Events
- E. Late Input Events

Correct Answer: AB

To react to increased workloads and increase streaming units, consider setting an alert of 80% on the SU Utilization metric. Also, you can use watermark delay and backlogged events metrics to see if there is an impact.

Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job, by increasing the SUs.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-monitoring>

Community vote distribution

AB (100%)

✉  **demirsamuel** Highly Voted 1 year, 7 months ago

Selected Answer: AB

A and B are correct
upvoted 12 times

✉  **kkk5566** Most Recent 4 months ago

Selected Answer: AB

A and B are correct
upvoted 1 times

✉  **Ankit_Az** 7 months, 1 week ago

Selected Answer: AB

Correct
upvoted 2 times

✉  **Coderhbt** 8 months, 4 weeks ago

Selected Answer: AB

Correct Answers
upvoted 2 times

✉  **Deeksha1234** 1 year, 5 months ago

Selected Answer: AB

correct answer
upvoted 2 times

You have an activity in an Azure Data Factory pipeline. The activity calls a stored procedure in a data warehouse in Azure Synapse Analytics and runs daily.

You need to verify the duration of the activity when it ran last.

What should you use?

- A. activity runs in Azure Monitor
- B. Activity log in Azure Synapse Analytics
- C. the sys.dm_pdw_wait_stats data management view in Azure Synapse Analytics
- D. an Azure Resource Manager template

Correct Answer: A

Monitor activity runs. To get a detailed view of the individual activity runs of a specific pipeline run, click on the pipeline name.

Example:

The screenshot shows the 'Pipeline runs' page in the Azure Data Factory portal. The top navigation bar includes tabs for 'Triggered' (which is selected), 'Debug', and other options like 'Rerun', 'Cancel', 'Refresh', and 'Edit columns'. Below the navigation is a search bar labeled 'Search by run ID or name' and a time filter set to 'Pacific Time (US & C... : Last 7 days)'. The main area displays a table titled 'Showing 1 - 21 items' with columns for 'Pipeline name', 'Run start', and 'Run end'. The first row, which has a red box around its 'Pipeline name' cell containing 'S3ToDataLakeCopy', is highlighted. The other rows show runs for different pipelines: 'DatabricksJarPipeline' and two more 'S3ToDataLakeCopy' runs. All runs occurred on November 5, 2020, between 6:00:18 AM and 6:03:15 AM.

| Pipeline name | Run start ↑↓ | Run end |
|-----------------------|---------------------|---------------------|
| S3ToDataLakeCopy | 11/5/20, 6:00:18 AM | 11/5/20, 6:03:15 AM |
| DatabricksJarPipeline | 11/4/20, 6:04:11 PM | 11/4/20, 6:10:15 PM |
| S3ToDataLakeCopy | 11/4/20, 6:00:18 PM | 11/4/20, 6:03:15 PM |
| S3ToDataLakeCopy | 11/4/20, 6:00:19 AM | 11/4/20, 6:04:15 AM |

The list view shows activity runs that correspond to each pipeline run. Hover over the specific activity run to get run-specific information such as the JSON input,

JSON output, and detailed activity-specific monitoring experiences.

SalesAnalyticsMLPipeline

Activity runs

Pipeline run ID a600eabe-19fb-4d0b-bd8d-d20b21223923

All status ▾

Showing 1 - 5 of 5 items

| Activity name | Activity type | Run start ↑↓ | Duration | Status |
|---------------------|---------------|----------------------|----------|--|
| Location_HTTP | Copy | 11/5/20, 12:12:44 PM | 00:00:15 | ✓ Succeeded |
| Clickstream_S3 | Copy | 11/5/20, 12:12:44 PM | 00:00:27 | ✗ Failed |
| Customer_Salesforce | Copy | 11/5/20, 12:12:44 PM | 00:00:10 | ✓ Succeeded |
| POS_SQL | Copy | 11/5/20, 12:12:44 PM | 00:00:36 | ✗ Failed |
| Products_SAP | Copy | 11/5/20, 12:12:44 PM | 00:00:08 | ✓ Succeeded |

You can check the Duration.

Incorrect Answers:

C: sys.dm_pdw_wait_stats holds information related to the SQL Server OS state related to instances running on the different nodes.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-visualy>

Community vote distribution

A (100%)

Deeksha1234 Highly Voted 1 year, 5 months ago

Selected Answer: A

Monitor, in ADF we have monitor to check all activity runs
upvoted 9 times

kkk5566 Most Recent 4 months ago

Selected Answer: A

Monitor, in ADF we have monitor to check all activity runs
upvoted 1 times

TechMgr 8 months, 1 week ago

Selected Answer: A

A is correct
upvoted 1 times

MScapris 1 year ago

Selected Answer: A

is correct answer
upvoted 4 times

nicky87654 1 year ago

Selected Answer: A

A IS CORRECT ANSWER

upvoted 2 times

✉ **Franz58** 1 year, 5 months ago

I'd go with A using this:

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor>

upvoted 3 times

✉ **RanjitManuel** 1 year, 6 months ago

Azure monitor is different from the Monitor option shown in the screenshot.

upvoted 4 times

✉ **demirsamuel** 1 year, 7 months ago

answer is correct. screenshot shows azure data factory pipeline run

upvoted 2 times

✉ **demirsamuel** 1 year, 7 months ago

* under monitor section in ADF

upvoted 3 times

✉ **g2000** 1 year, 8 months ago

based upon the screen shot, isn't that part of the azure synapse analytics (one of the icons from the left)?

upvoted 2 times

✉ **sdokmak** 1 year, 7 months ago

Looks like Data Factory to me. If Data Factory was there I would have picked it.

upvoted 1 times

✉ **upliftinghut** 1 year, 7 months ago

answer is correct, monitoring is under Monitor and Dashboard

upvoted 3 times

You have an Azure Data Factory pipeline that is triggered hourly.

The pipeline has had 100% success for the past seven days.

The pipeline execution fails, and two retries that occur 15 minutes apart also fail. The third failure returns the following error.

ErrorCode=UserErrorFileNotFoundException, Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException, Message=ADLS Gen2 operation failed for: Operation returned an invalid status code 'NotFound'. Account: 'contosoproductsouth'. Filesystem: wwi. Path: 'BIKES/CARBON/year=2021/month=01/day=10/hour=06'. ErrorCode: 'PathNotFound'. Message: 'The specified path does not exist.'. RequestId: '6d269b78-901f-001b-4924-e7a7bc000000'. TimeStamp: 'Sun, 10 Jan 2021 07:45:05'

What is a possible cause of the error?

- A. The parameter used to generate year=2021/month=01/day=10/hour=06 was incorrect.
- B. From 06:00 to 07:00 on January 10, 2021, there was no data in wwi/BIKES/CARBON.
- C. From 06:00 to 07:00 on January 10, 2021, the file format of data in wwi/BIKES/CARBON was incorrect.
- D. The pipeline was triggered too early.

Correct Answer: A

A file is missing.

Community vote distribution

| | |
|---------|----|
| B (94%) | 6% |
|---------|----|

✉  **KashRaynardMorse** Highly Voted 1 year, 8 months ago

Selected Answer: B

The error message says a missing file, which matches with answer B: missing data from 06:00. The process had re-tried three times, 15 mins apart, which explains that the error was generated 07:45.

upvoted 27 times

✉  **jongert** 1 week, 4 days ago

To elaborate, the pipeline is triggered hourly, any processing that is done has to be applied to the data that was received in the last hour. In other words, at 07:00 the data received between 06:00 and 07:00 is processed. Accounting for 3*15m results in 07:45.

More elegantly however, the timestamps worked fine for a week which indicates that path creation is not a problem. Answer is B as you say.
upvoted 2 times

✉  **Billybob0604** 1 year, 1 month ago

i don't agree. the path is not created correctly and therefore the file is 'missing'. It is in the error message too.

upvoted 2 times

✉  **MS2710** 1 year ago

Answer A. The path in error message shows hour=06 whereas the hour of retry run is 07.

upvoted 3 times

✉  **Remedios79** 1 year, 6 months ago

Thank you for the detail

upvoted 1 times

✉  **sugias** Highly Voted 10 months ago

For 7 days, this job was succeeding.

So path rule seems to be right.

upvoted 5 times

✉  **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: B

B is correct

upvoted 1 times

✉  **Altgeeky** 4 months, 2 weeks ago

A. The parameter used to generate year=2021/month=01/day=10/hour=06 was incorrect.

The error message indicates that the specified path 'BIKES/CARBON/year=2021/month=01/day=10/hour=06' does not exist. This suggests that the parameters used to generate the path might be incorrect, resulting in an invalid or non-existent path. Double-checking the parameter values used to construct the path would be the most likely reason for this error.

upvoted 1 times

✉  **Tightbot** 4 months, 3 weeks ago

Selected Answer: B

This option matches with the parameter 6th hour of Jan 10th . Option A might or might not be true but we have no data about what hour of data to be retrieved. Whereas, Option B clearly points to the time mentioned in the question and Path not found could be due to no data present upvoted 1 times

 esaade 10 months, 1 week ago

Selected Answer: B

The error message states that the specified path does not exist. Therefore, a possible cause of the error could be that the data for the specified path, which is wwi/BIKES/CARBON/year=2021/month=01/day=10/hour=06, does not exist in the storage account. This could be due to missing data or incorrect path or container name. Option B is the most likely cause of the error as it suggests that there was no data in the specified path during the given time frame.

upvoted 3 times

 raphasc 1 year ago

Selected Answer: A

Provided answer is correct : PATH NOT FOUND . The right path must be BIKES/CARBON/2021/01/06/Filename.*
Filesystem: wwi. Path: 'BIKES/CARBON/year=2021/month=01/day=10/hour=06'. ErrorCode: 'PathNotFound'. Message: 'The specified path does not exist.'

upvoted 2 times

 Venub28 12 months ago

question says that it ran fine earlier. Parameter must have been set correctly. Answer is B

upvoted 3 times

 dmitriypo 1 year, 2 months ago

Selected Answer: B

No file

upvoted 2 times

 Rohanh 1 year, 3 months ago

Selected Answer: B

B is correct

upvoted 3 times

 dom271219 1 year, 4 months ago

Selected Answer: B

B of course

upvoted 3 times

 CloudixExamTopics 1 year, 4 months ago

The question states the pipeline runs hourly and in the timestamp of the error we can see that the time is 7:45 for the third run. So the initial run was at 7:00, but the folder it was looking at is hour=06, which is wrong, it should be hour = 07. So I agree with the option A
upvoted 5 times

 pangas2567 1 year, 4 months ago

With run starting at 7.00 pointing to the hour=07 folder, you wouldn't have anything to work with. One hour delay needed here.

upvoted 9 times

 Deeksha1234 1 year, 5 months ago

Selected Answer: B

B is correct

upvoted 3 times

 ROLLINGROCKS 1 year, 5 months ago

The issue with B is that it says that there is no data in BIKES/CARBON which is false, because it has been loading for a week. There might not be data in a subdirectory of BIKES/CARBON but there is data in BIKES/CARBON for sure, making B false in my opinion.
upvoted 5 times

 Sriramiyer92 1 year, 5 months ago

B

Reason : Operation returned an invalid status code 'NotFound'. & 'Message: 'The specified path does not exist.''

upvoted 1 times

 virendrapsingh 1 year, 7 months ago

Selected Answer: B

Answer should be B.

upvoted 2 times

 demirsamuel 1 year, 7 months ago

Selected Answer: B

100% B. The error shows at 7:45. So 45 min after 7. o'clock. and that's equal to 3 times 15 min interval. Additionally the stacktrace shows that no filepath exists.

upvoted 4 times

 **RamboRinky** 1 year, 7 months ago

Selected Answer: A

Since the parameter that generates the path reference did not generate properly, we cannot look into the proper folder to check if the file is really missing. No telling if the file is missing if you do not look at the proper place where the file is supposed to be. year = 2021 should be 2021/....

upvoted 1 times

 **sdokmak** 1 year, 7 months ago

it was correct for 7 days so there's no way that's true.

upvoted 6 times

Question #76

Topic 2

You have an Azure Synapse Analytics job that uses Scala.

You need to view the status of the job.

What should you do?

- A. From Synapse Studio, select the workspace. From Monitor, select SQL requests.
- B. From Azure Monitor, run a Kusto query against the AzureDiagnostics table.
- C. From Synapse Studio, select the workspace. From Monitor, select Apache Sparks applications.
- D. From Azure Monitor, run a Kusto query against the SparkLoggingEvent_CL table.

Correct Answer: C

Use Synapse Studio to monitor your Apache Spark applications. To monitor running Apache Spark application Open Monitor, then select Apache Spark applications. To view the details about the Apache Spark applications that are running, select the submitting Apache Spark application and view the details. If the

Apache Spark application is still running, you can monitor the progress.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/monitoring/apache-spark-applications>

Community vote distribution

C (100%)

 **nefarious_smalls** Highly Voted  1 year, 7 months ago

I think this is one is correct.

upvoted 9 times

 **kkk5566** Most Recent  4 months, 1 week ago

Selected Answer: C

Correct

upvoted 1 times

 **Ankit_Az** 7 months, 1 week ago

Selected Answer: C

Correct

upvoted 1 times

 **akk_1289** 11 months ago

The correct answer is C. From Synapse Studio, select the workspace. From Monitor, select Apache Spark applications.

upvoted 2 times

 **pangas2567** 1 year, 4 months ago

Selected Answer: C

Correct

<https://docs.microsoft.com/en-us/azure/synapse-analytics/monitoring/how-to-monitor-spark-applications>

upvoted 4 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: C

C is correct

upvoted 2 times

DRAG DROP -

You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers. The file contains two attributes named FirstName and LastName.

You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the FirstName and LastName values.

You create the following components:

- A destination table in Azure Synapse
- An Azure Blob storage container
- A service principal

In which order should you perform the actions? 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

- Mount the Data Lake Storage onto DBFS.
- Write the results to a table in Azure Synapse.
- Specify a temporary folder to stage the data.
- Read the file into a data frame.
- Perform transformations on the data frame.

Answer Area

Correct Answer:

Actions

- Mount the Data Lake Storage onto DBFS.
- Read the file into a data frame.
- Perform transformations on the data frame.
- Specify a temporary folder to stage the data.
- Write the results to a table in Azure Synapse.

Answer Area

- Mount the Data Lake Storage onto DBFS.
- Read the file into a data frame.
- Perform transformations on the data frame.
- Specify a temporary folder to stage the data.
- Write the results to a table in Azure Synapse.

Step 1: Mount the Data Lake Storage onto DBFS

Begin with creating a file system in the Azure Data Lake Storage Gen2 account.

Step 2: Read the file into a data frame.

You can load the json files as a data frame in Azure Databricks.

Step 3: Perform transformations on the data frame.

Step 4: Specify a temporary folder to stage the data

Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse.

Step 5: Write the results to a table in Azure Synapse.

You upload the transformed data frame into Azure Synapse. You use the Azure Synapse connector for Azure Databricks to directly upload a dataframe as a table in a Azure Synapse.

Reference:

<https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-extract-load-sql-data-warehouse>

 **Feljoud** Highly Voted 1 year, 8 months ago

Similar to another question in this dump. Seems correct!

upvoted 18 times

 **kkk5566** Most Recent 4 months, 1 week ago

correct

upvoted 1 times

 **rzeng** 1 year, 2 months ago

correct

upvoted 3 times

✉ **dom271219** 1 year, 4 months ago

"Specify a temporary folder to stage the data" must be before creating the DF : I am wrong ?

upvoted 1 times

✉ **Karl_Cen** 1 year ago

As mentioned earlier, the Azure Synapse connector uses Azure Blob storage as temporary storage to upload data between Azure Databricks and Azure Synapse

so it means only before you loading data into ADLS, you need this temporary folder.

<https://learn.microsoft.com/en-us/azure/databricks/scenarios/databricks-extract-load-sql-data-warehouse>

upvoted 2 times

✉ **Deeksha1234** 1 year, 5 months ago

correct

upvoted 2 times

✉ **nefarious_smalls** 1 year, 7 months ago

correct

upvoted 1 times

✉ **demirsamuel** 1 year, 7 months ago

answer is correct. Similar to a duplicated question in this question catalog.

upvoted 3 times

You have an Azure data factory named ADF1.

You currently publish all pipeline authoring changes directly to ADF1.

You need to implement version control for the changes made to pipeline artifacts. The solution must ensure that you can apply version control to the resources currently defined in the UX Authoring canvas for ADF1.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. From the UX Authoring canvas, select Set up code repository.
- B. Create a Git repository.
- C. Create a GitHub action.
- D. Create an Azure Data Factory trigger.
- E. From the UX Authoring canvas, select Publish.
- F. From the UX Authoring canvas, run Publish All.

Correct Answer: BF

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/source-control>

Community vote distribution

| | | |
|----------|----------|-----|
| AB (74%) | AF (16%) | 11% |
|----------|----------|-----|

✉ dom271219 Highly Voted 1 year, 4 months ago

Selected Answer: AB

They are asking to "implement version control".

B Create Git repo

A From the UX Set up code repository

upvoted 17 times

✉ esaade Highly Voted 10 months, 1 week ago

Selected Answer: AB

To implement version control for changes made to pipeline artifacts in ADF1 while ensuring that version control can be applied to the resources currently defined in the UX Authoring canvas, you should perform the following two actions:

A. From the UX Authoring canvas, select Set up code repository: This will allow you to configure ADF1 to integrate with a version control system such as Git, which will enable you to track changes made to pipeline artifacts over time.

B. Create a Git repository: This will provide the version control system needed to track changes made to pipeline artifacts in ADF1.

Therefore, options A and B are the correct answers.

C, D, E, and F are not relevant to implementing version control for changes made to pipeline artifacts in ADF1.

upvoted 6 times

✉ kkk5566 Most Recent 4 months, 1 week ago

Selected Answer: AB

Correct

upvoted 1 times

✉ Ankit_Az 7 months, 1 week ago

Selected Answer: AB

Correct

upvoted 1 times

✉ Debasish93 8 months ago

I think the answer should be AF as "Set up code repository" gives us the option of creating new repository if not already created so option B is redundant. More over we should not individually publish existing artifacts rather should go for "Publish All".

upvoted 3 times

✉ azure_user11 8 months ago

Selected Answer: AB

"ensuring that version control can be applied to the resources currently defined in the UX Authoring canvas"

When creating a Git repository this option is ticked by default, so all available resources at the time of the creation are imported into Git, no need to publish which is what the last answers are trying to imply.

Import existing resources to repository Specifies whether to import existing data factory resources from the UX authoring canvas into a GitHub repository. Select the box to import your data factory resources into the associated Git repository in JSON format. This action exports each resource individually (that is, the linked services and datasets are exported into separate JSONs). When this box isn't selected, the existing resources aren't imported. Selected (default)

upvoted 1 times

akk_1289 11 months ago

The correct answers are B. Create a Git repository and A. From the UX Authoring canvas, select Set up code repository.

upvoted 1 times

[Removed] 12 months ago

Selected Answer: AB

option A is correct because it allows you to set up a code repository to store and manage the changes made to pipeline artifacts in ADF1. Option B is correct because it allows you to create a Git repository, which is a version control system that stores the history of changes made to the pipeline artifacts. This allows you to easily roll back to a previous version or compare changes made over time.

upvoted 2 times

OldSchool 1 year, 1 month ago

Selected Answer: AF

Since there is no mention of GitHub or DevOps the solution that works for both is A & F

upvoted 2 times

dmitriypo 1 year, 2 months ago

Selected Answer: AE

I did a setup of the version control for my test ADF instance in the following way:

A. From the UX Authoring canvas, select Set up code repository.

Here I configured a connection to the Azure DevOps organization, chose a project, and created a new repo.

E. From the UX Authoring canvas, select Publish.

upvoted 4 times

Titokyo 1 year, 2 months ago

A & B: The documentation attached to this question states the first step is to set up a code repository from the UX and this question is around setting up version control, not saving your changes which is what F suggests

upvoted 3 times

coolin 1 year, 3 months ago

F A is the correct order. Save changes then set up code repos

upvoted 3 times

anks84 1 year, 4 months ago

Selected Answer: AF

Should be AF as we want to achieve the version control for the code changes.

upvoted 4 times

federc 1 year, 4 months ago

why not A and B ? I would set up the code repository after creating the git repo

upvoted 3 times

yuorrik 10 months, 2 weeks ago

in my opinion, A & B resulted the same - repo created

upvoted 1 times

DRAG DROP -

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1. Branch1 contains an Azure Synapse pipeline named pipeline1.

In workspace1, you complete testing of pipeline1.

You need to schedule pipeline1 to run daily at 6 AM.

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.

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:

Actions**Answer Area**

Create a new branch in Repo1.



Merge the changes from branch1 into main.

Associate the schedule trigger with pipeline1.

Switch to Synapse live mode.

Create a schedule trigger.

Publish the contents of main.

Correct Answer:**Actions****Answer Area**

Create a new branch in Repo1.

Create a schedule trigger.

Associate the schedule trigger with pipeline1.

Merge the changes from branch1 into main.

Switch to Synapse live mode.

Publish the contents of main.

MJSnail Highly Voted 11 months, 1 week ago

If it's hard to remember, memorize it as CAMP.

upvoted 39 times

azure_user11 Highly Voted 8 months ago

Correct. I've worked with this many times. It's the right order.

upvoted 7 times

kkk5566 Most Recent 4 months, 1 week ago

C->A->M->P

upvoted 2 times

✉ **Karl_Cen** 1 year ago

you should associate the trigger before merge the code into main, because schedule also is part of code. all code store in main, do not change it directly, that is the purpose of version control.

upvoted 3 times

✉ **vrodriguesp** 1 year ago

why not this?

- 1.merge th echanges from branch1 into main
- 2.publish the contents of main
- 3.create a schedule trigger
- 4.associate the schedule trigger with pipeline1

upvoted 4 times

✉ **mroova** 11 months ago

This order is also possible, but not recommended. As the trigger would not be visible in the repo, which can be misleading to anyone that is reviewing or auditing the solution.

upvoted 3 times

✉ **vrodriguesp** 1 year ago

sorry, I noticed that the question claims:

NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

upvoted 3 times

✉ **dmitriypo** 1 year, 2 months ago

Correct

upvoted 2 times

✉ **TiredDad** 1 year, 2 months ago

Should it not be merge, then publish, then create a schedule trigger, finally associate the schedule trigger with pipeline1?

upvoted 3 times

✉ **TiredDad** 1 year, 2 months ago

Pls ignore, I agree with the suggested answer

upvoted 2 times

✉ **rzeng** 1 year, 2 months ago

correct

upvoted 1 times

✉ **anks84** 1 year, 4 months ago

Looks correct

upvoted 2 times

HOTSPOT -

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers.

You need to create an external data source in Pool1 that will be used to read .orc files in storage1.

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

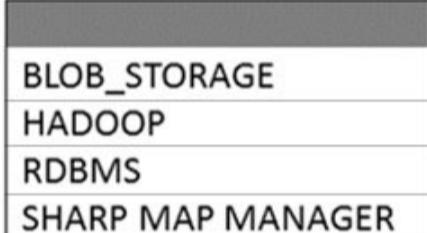
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore

WITH

(**Location1** '  ://data@newyorktaxidataset.dfs.core.windows.net' ,
 abfs
 abfss
 wasb
 wasbs

credential = ADLS_credential ,

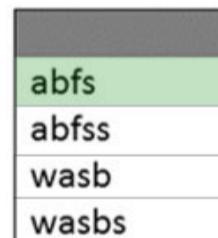
TYPE -

) ; 
 BLOB_STORAGE
 HADOOP
 RDBMS
 SHARP MAP MANAGER

Correct Answer:**Answer Area**

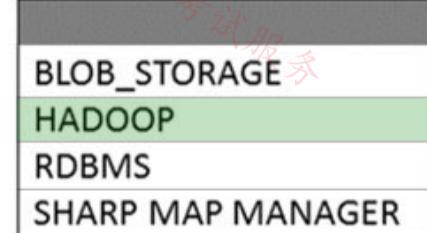
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore

WITH

(**Location1** '  ://data@newyorktaxidataset.dfs.core.windows.net' ,
 abfs
 abfss
 wasb
 wasbs

credential = ADLS_credential ,

TYPE -

) ; 
 BLOB_STORAGE
 HADOOP
 RDBMS
 SHARP MAP MANAGER

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw-latest&preserve-view=true&tabs=dedicated>

  **Hema_V**  1 year, 4 months ago

Answer: abfss and Hadoop

Hint: Storage1 requires secure transfers --> The default option is to use enable secure SSL connections when provisioning Azure Data Lake Storage Gen2. When this is enabled, you must use abfss when a secure TLS/SSL connection is selected.

Reference: <https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw-latest&preserve-view=true&tabs=dedicated>

upvoted 33 times

✉ **vigilante89** Highly Voted 1 year ago

abfss and Hadoop

upvoted 7 times

✉ **kkk5566** Most Recent 4 months, 1 week ago

Answer: abfss and Hadoop

upvoted 1 times

✉ **ZIMARAKI** 12 months ago

abfss & hadoop

upvoted 5 times

✉ **greenlever** 1 year, 2 months ago

abfss

Hadoop

abfss endpoint when your account has secure transfer enabled

upvoted 6 times

✉ **sesank** 1 year, 3 months ago

abfss

Hadoop

upvoted 4 times

✉ **anks84** 1 year, 4 months ago

-abfss

-hadoop

upvoted 6 times

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named SQLPool1.

SQLPool1 is currently paused.

You need to restore the current state of SQLPool1 to a new SQL pool.

What should you do first?

- A. Create a workspace.
- B. Create a user-defined restore point.
- C. Resume SQLPool1.
- D. Create a new SQL pool.

Correct Answer: B

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-restore-active-paused-dw>

Community vote distribution

C (84%) B (16%)

 **yogiazaad** Highly Voted 11 months, 2 weeks ago

Selected Answer: C

You won't be able to create restore point when the SQL pool is paused. The correct answer is Result SQL Pool. See below from Microsoft documentation.

User-defined restore points can also be created through Azure portal.

Sign in to your Azure portal account.

Navigate to the dedicated SQL pool (formerly SQL DW) that you want to create a restore point for.

Select Overview from the left pane, select + New Restore Point. If the New Restore Point button isn't enabled, make sure that the dedicated SQL pool (formerly SQL DW) isn't paused.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-restore-points>
upvoted 13 times

 **esaade** Highly Voted 10 months, 1 week ago

Selected Answer: C

Before restoring the state of SQLPool1 to a new SQL pool, you should resume SQLPool1. Therefore, the correct answer is:

C. Resume SQLPool1.

upvoted 6 times

 **Lucasmh** Most Recent 4 weeks, 1 day ago

Selected Answer: B

Correcto

upvoted 1 times

 **msb** 3 months, 2 weeks ago

Selected Answer: C

Correct answer is C.

Not restore points are created when pool is paused. And there default retention is 7 days.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/backup-and-restore#automatic-restore-points>
upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: C

is correct

upvoted 1 times

 **auwia** 6 months, 3 weeks ago

Selected Answer: C

Resume.

upvoted 2 times

 **Ast999** 10 months, 1 week ago

Selected Answer: C

you cannot create user-defined restore points when the Azure Synapse Analytics dedicated SQL pool is currently paused. In order to create a user-defined restore point, the SQL pool must be running.

upvoted 3 times

 **RV123** 1 year ago

Selected Answer: B

Correct

upvoted 2 times

 **dom271219** 1 year, 4 months ago

Selected Answer: B

Agreed

upvoted 2 times

 **kumarrahul1107** 1 year, 4 months ago

Correct

upvoted 1 times

You are designing an Azure Synapse Analytics workspace.

You need to recommend a solution to provide double encryption of all the data at rest.

Which two components should you include in the recommendation? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. an X.509 certificate
- B. an RSA key
- C. an Azure virtual network that has a network security group (NSG)
- D. an Azure Policy initiative
- E. an Azure key vault that has purge protection enabled

Correct Answer: BE

Synapse workspaces encryption uses existing keys or new keys generated in Azure Key Vault. A single key is used to encrypt all the data in a workspace.

Synapse workspaces support RSA 2048 and 3072 byte-sized keys, and RSA-HSM keys.

The Key Vault itself needs to have purge protection enabled.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption>

Community vote distribution

BE (94%) 6%

 **allagowf** Highly Voted 1 year, 2 months ago

Selected Answer: BE

Answer is correct : BE

upvoted 6 times

 **d046bc0** Most Recent 3 weeks, 3 days ago

Selected Answer: BE

BE is correct

<https://learn.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption#workspace-encryption-configuration>

upvoted 1 times

 **DataEngDP** 3 months, 3 weeks ago

Selected Answer: BE

BE is correct to provide double encryption at REST. RSA key and Azure key vault.

upvoted 1 times

 **kkk5566** 4 months ago

Selected Answer: BE

BE is correct

upvoted 1 times

 **Pradeep2675** 5 months, 2 weeks ago

Selected Answer: BE

Answer: BE

Explanation:

Synapse workspaces encryption uses existing keys or new keys generated in Azure Key Vault. A single key is used to encrypt all the data in a workspace.

Synapse workspaces support RSA 2048 and 3072 byte-sized keys, and RSA-HSM keys.

The Key Vault itself needs to have purge protection enabled.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption>

upvoted 1 times

 **vctrhugo** 6 months, 2 weeks ago

Selected Answer: AE

A. Including an X.509 certificate in the solution can be used to provide encryption at rest for the data in Azure Synapse Analytics. X.509 certificates are widely used for securing data and communications.

E. An Azure Key Vault with purge protection enabled can be utilized to securely store and manage encryption keys. By storing the encryption keys in Azure Key Vault, you can ensure that the keys are well protected and access to them is tightly controlled.

upvoted 1 times

 **auwia** 6 months, 3 weeks ago

Selected Answer: BE

Correct.

upvoted 1 times

 **dmitriypo** 1 year, 2 months ago

Selected Answer: BE

Agree with the answer

upvoted 4 times

 **amitshinde14** 1 year, 3 months ago

Correct ans.

upvoted 3 times

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

You have an Azure Synapse Analytics serverless SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named storage1. The AllowBlobPublicAccess property is disabled for storage1.

You need to create an external data source that can be used by Azure Active Directory (Azure AD) users to access storage from Pool1.

What should you create first?

- A. an external resource pool
- B. an external library
- C. database scoped credentials
- D. a remote service binding

Correct Answer: C

Security -

User must have SELECT permission on an external table to read the data. External tables access underlying Azure storage using the database scoped credential defined in data source.

Note: A database scoped credential is a record that contains the authentication information that is required to connect to a resource outside SQL Server. Most credentials include a Windows user and password.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables> <https://docs.microsoft.com/en-us/sql/t-sql/statements/create-database-scoped-credential-transact-sql>

Community vote distribution

C (100%)

 **anks84** Highly Voted 1 year, 4 months ago

Selected Answer: C

Correct Answer !

upvoted 5 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: C

correct

upvoted 2 times

 **Ankit_Az** 7 months, 1 week ago

Selected Answer: C

Correct

upvoted 2 times

 **GodfreyMbizo** 11 months, 2 weeks ago

correct answer

upvoted 2 times

 **GodfreyMbizo** 11 months, 2 weeks ago

database scoped credentials first

upvoted 3 times

You have an Azure Data Factory pipeline named Pipeline1. Pipeline1 contains a copy activity that sends data to an Azure Data Lake Storage Gen2 account.

Pipeline1 is executed by a schedule trigger.

You change the copy activity sink to a new storage account and merge the changes into the collaboration branch.

After Pipeline1 executes, you discover that data is NOT copied to the new storage account.

You need to ensure that the data is copied to the new storage account.

What should you do?

- A. Publish from the collaboration branch.
- B. Create a pull request.
- C. Modify the schedule trigger.
- D. Configure the change feed of the new storage account.

Correct Answer: A

CI/CD lifecycle -

1. A development data factory is created and configured with Azure Repos Git. All developers should have permission to author Data Factory resources like pipelines and datasets.
2. A developer creates a feature branch to make a change. They debug their pipeline runs with their most recent changes
3. After a developer is satisfied with their changes, they create a pull request from their feature branch to the main or collaboration branch to get their changes reviewed by peers.
4. After a pull request is approved and changes are merged in the main branch, the changes get published to the development factory.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-delivery>

Community vote distribution

A (100%)

✉  **kkk5566** 4 months, 1 week ago

Selected Answer: A

correct

upvoted 2 times

✉  **Iudaka** 6 months, 3 weeks ago

Selected Answer: A

Correct answer.

upvoted 2 times

✉  **kim32** 8 months ago

I selected B, pull request

upvoted 1 times

✉  **auwia** 6 months, 2 weeks ago

I guessed the same, but in adf publish step is manually and not automatic, so when the question says code is merged, you can automatically assume that pull request was done and also approved by reviewers. Therefore the correct answer for me it's A.

upvoted 2 times

✉  **Xinyuehong** 1 year, 2 months ago

I had heard "publish to", never heard of "publish from". So confused.

upvoted 2 times

✉  **Igor85** 1 year, 1 month ago

probably it was meant to publish from collaboration branch to adf_publish branch

upvoted 2 times

✉  **debarun** 1 year, 4 months ago

Why not B ?

upvoted 1 times

✉  **DataEX** 1 year, 4 months ago

Because the pull request is already implicit in the statement as it is said to be merged into the collaborating branch:
"You change the copy activity sink to a new storage account and MERGE the CHANGES INTO the COLLABORATION BRANCH."

upvoted 11 times

店铺: IT认证考试服务

店铺: IT认证考试服务

You have an Azure Data Factory pipeline named pipeline1 that is invoked by a tumbling window trigger named Trigger1. Trigger1 has a recurrence of 60 minutes.

You need to ensure that pipeline1 will execute only if the previous execution completes successfully.

How should you configure the self-dependency for Trigger1?

- A. offset: "-00:01:00" size: "00:01:00"
- B. offset: "01:00:00" size: "-01:00:00"
- C. offset: "01:00:00" size: "01:00:00"
- D. offset: ^{大小: 1小时} "-01:00:00" size: "01:00:00"

Correct Answer: D

Tumbling window self-dependency properties

In scenarios where the trigger shouldn't proceed to the next window until the preceding window is successfully completed, build a self-dependency. A self-dependency trigger that's dependent on the success of earlier runs of itself within the preceding hour will have the properties indicated in the following code.

Example code:

```
"name": "DemoSelfDependency",
"properties": {
  "runtimeState": "Started",
  "pipeline": {
    "pipelineReference": {
      "referenceName": "Demo",
      "type": "PipelineReference"
    }
  },
  "type": "TumblingWindowTrigger",
  "typeProperties": {
    "frequency": "Hour",
    "interval": 1,
    "startTime": "2018-10-04T00:00:00Z",
    "delay": "00:01:00",
    "maxConcurrency": 50,
    "retryPolicy": {
      "intervalInSeconds": 30
    },
    "dependsOn": [
      {
        "type": "SelfDependencyTumblingWindowTriggerReference",
        "size": "01:00:00",
        "offset": "-01:00:00"
      }
    ]
  }
}
```

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/tumbling-window-trigger-dependency>

Community vote distribution

D (100%)

 **Azurre**  10 months ago

Correct Answer: D

Offset of "-01:00:00" indicates to start the next trigger instance only after the previous trigger instance completes, and size of "01:00:00" indicates to wait for 1 hour after the previous trigger instance completes before starting the next one.

upvoted 5 times

 **Okea** Highly Voted  1 year ago

Answer: D

offset

Offset of the dependency trigger. Provide a value in time span format and both negative and positive offsets are allowed. This property is mandatory if the trigger is depending on itself and in all other cases it is optional. Self-dependency should always be a negative offset. If no value specified, the window is the same as the trigger itself.

size

Size of the dependency tumbling window. Provide a positive timespan value. This property is optional.

<https://learn.microsoft.com/en-us/azure/data-factory/tumbling-window-trigger-dependency>

upvoted 5 times

 **kkk5566** Most Recent  4 months, 1 week ago

Selected Answer: D

D is correct

upvoted 1 times

 **vctrhugo** 6 months, 4 weeks ago

Selected Answer: D

Offset: "-01:00:00"

Size: "01:00:00"

This configuration ensures that the trigger waits for the completion of the previous execution before starting a new one. The offset of "-01:00:00" indicates that the trigger should start one hour before the current time, and the size of "01:00:00" indicates that the trigger should have a duration of one hour.

Therefore, the correct option is:

D. offset: "-01:00:00" size: "01:00:00"

upvoted 4 times

 **dom271219** 1 year, 4 months ago

Selected Answer: D

```
"dependsOn": [
{
  "type": "SelfDependencyTumblingWindowTriggerReference",
  "size": "01:00:00",
  "offset": "-01:00:00"
}]
```

upvoted 5 times

HOTSPOT -

You have an Azure Synapse Analytics pipeline named Pipeline1 that contains a data flow activity named Dataflow1.

Pipeline1 retrieves files from an Azure Data Lake Storage Gen 2 account named storage1.

Dataflow1 uses the AutoResolveIntegrationRuntime integration runtime configured with a core count of 128.

You need to optimize the number of cores used by Dataflow1 to accommodate the size of the files in storage1.

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

Hot Area:

Answer Area

店铺：IT认证考试服务

To Pipeline1, add:

| |
|--------------------------|
| A custom activity |
| A Get Metadata activity |
| An If Condition activity |

For Dataflow1, set the core count by using:

| |
|-----------------|
| Dynamic content |
| Parameters |
| User properties |

Correct Answer:**Answer Area**

To Pipeline1, add:

| |
|--------------------------------|
| A custom activity |
| A Get Metadata activity |
| An If Condition activity |

For Dataflow1, set the core count by using:

| |
|-----------------|
| Dynamic content |
| Parameters |
| User properties |

Box 1: A Get Metadata activity -

Dynamically size data flow compute at runtime

The Core Count and Compute Type properties can be set dynamically to adjust to the size of your incoming source data at runtime. Use pipeline activities like

Lookup or Get Metadata in order to find the size of the source dataset data. Then, use Add Dynamic Content in the Data Flow activity properties.

Box 2: Dynamic content -

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/control-flow-execute-data-flow-activity>

 **dom271219** Highly Voted  1 year, 4 months ago

Correct:

Use pipeline activities like Lookup or Get Metadata in order to find the size of the source dataset data. Then, use Add Dynamic Content in the Data Flow activity properties. You can choose small, medium, or large compute sizes. Optionally, pick "Custom" and configure the compute types and number of cores manually.

upvoted 12 times

 **dmitriypo** Highly Voted  1 year, 2 months ago

Looks correct. Checked in the doc.

upvoted 6 times

 **kkk5566** Most Recent  4 months, 1 week ago

Get Metadata & Dynamic Content

upvoted 1 times

 **anks84** 1 year, 4 months ago

Looks Correct !!

upvoted 3 times

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 plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- A workload for data engineers who will use Python and SQL.
- A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.

Does this meet the goal?

- A. Yes
- B. No

Correct Answer: B

We would need a High Concurrency cluster for the jobs.

Note:

Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.

A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

Reference:

<https://docs.azuredatabricks.net/clusters/configure.html>

Community vote distribution

| | |
|---------|---------|
| A (83%) | B (17%) |
|---------|---------|

 **Amalbenrebai** Highly Voted  2 years, 4 months ago

- data engineers: high concurrency cluster
 - jobs: Standard cluster
 - data scientists: Standard cluster
- upvoted 91 times

 **gogosgh** 8 months, 1 week ago

The issue is the jobs are going to be ran by multiple users i.e. engineers and scientists? So it needs to be a high concurrency cluster?
upvoted 1 times

 **auwia** 6 months, 2 weeks ago

If you enable high concurrency then all scale scripts don't work, so scientists will stop to work). Standard cluster is scalable, will support all jobs and users! ;-)
upvoted 2 times

 **supriyako** 1 year, 3 months ago

Correct. Because jobs could be for Scala notebook, which is supported by Standard cluster mode
upvoted 2 times

 **Egocentric** 1 year, 8 months ago

agreed
upvoted 1 times

 **Julius7000** 2 years, 3 months ago

Tell me one thing: is this answer 9jobs) based on the text:
"A Single Node cluster has no workers and runs Spark jobs on the driver node.

In contrast, a Standard cluster requires at least one Spark worker node in addition to the driver node to execute Spark jobs."?
I dont understand the connection between worker noodes and the requirements given in the question about jobs workspace.
upvoted 1 times

✉ **Aditya0891** 1 year, 7 months ago

single node cluster and standard cluster are different. In single node cluster you only have 1 node which act as driver and worker node while in standard cluster you can have separate driver and worker node and for jobs you can use standard or high concurrency cluster as well. So the requirements are satisfied here

upvoted 1 times

✉ **gangstfear** Highly Voted 2 years, 4 months ago

The answer must be A!

upvoted 34 times

✉ **dakku987** Most Recent 1 week, 2 days ago

Selected Answer: B

we need HC cluster for data engineer,data scientist,jobs

upvoted 1 times

✉ **kkk5566** 4 months, 1 week ago

Selected Answer: A

correct

- data engineers: high concurrency cluster
- jobs: Standard cluster
- data scientists: Standard cluster

upvoted 1 times

✉ **auwia** 6 months, 2 weeks ago

Selected Answer: A

High concurrence doesn't support scala.

upvoted 2 times

✉ **auwia** 6 months, 2 weeks ago

Selected Answer: A

True, correct.

upvoted 1 times

✉ **Ast999** 10 months, 1 week ago

Selected Answer: A

SCALA = STANDARD

upvoted 4 times

✉ **allagowf** 1 year, 2 months ago

Selected Answer: A

data scientists and Job --> Scala --> Standard cluster .

upvoted 3 times

✉ **greenlever** 1 year, 2 months ago

Selected Answer: A

Correct

upvoted 1 times

✉ **anks84** 1 year, 4 months ago

Selected Answer: A

We would need a Standard cluster for the jobs to support Scala. High-concurrecny cluster does not support Scala.
Hence, the Answer is A !

upvoted 1 times

✉ **Deeksha1234** 1 year, 5 months ago

Selected Answer: B

the answer should be No

upvoted 1 times

sorry 'A' should be correct

upvoted 3 times

✉ **sethuramansp** 1 year, 6 months ago

The answer should be "NO" as per the given statement "The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster." since the Job cluster is standard it will not allow data scientists and engineers to collectively deploy their Notebooks in standard cluster as it requires High Concurrency Cluster

upvoted 2 times

□ **wanchihh** 3 months, 4 weeks ago

High Concurrency Cluster does not support Scala.

upvoted 1 times

□ **Eyepatch993** 1 year, 9 months ago

Selected Answer: B

Standard clusters do not have fault tolerance. Both the data scientist and data engineers will be using the job cluster for processing their notebooks, so if a standard cluster is chosen and a fault occurs in the notebook of any one user, there is a chance that other notebooks might also fail. Due to this a high concurrency cluster is recommended for running jobs.

upvoted 4 times

□ **Aditya0891** 1 year, 7 months ago

Read the question properly. it states that each data scientist will have a standard cluster and a separate standard cluster for running jobs. So there is no question of fault due to other users. The answer is A

upvoted 2 times

□ **Boompiee** 1 year, 8 months ago

It may not be a best practice, but the question asked is: does the solution meet the stated requirements, and it does..

upvoted 1 times

□ **Hanse** 1 year, 10 months ago

As per Link: <https://docs.azuredatabricks.net/clusters/configure.html>

Standard and Single Node clusters terminate automatically after 120 minutes by default. --> Data Scientists

High Concurrency clusters do not terminate automatically by default.

A Standard cluster is recommended for a single user. --> Standard for Data Scientists & High Concurrency for Data Engineers

Standard clusters can run workloads developed in any language: Python, SQL, R, and Scala.

High Concurrency clusters can run workloads developed in SQL, Python, and R. The performance and security of High Concurrency clusters is provided by running user code in separate processes, which is not possible in Scala. --> Jobs needs Standard

upvoted 6 times

□ **ovokpus** 1 year, 10 months ago

Selected Answer: A

Yes it seems to be!

upvoted 2 times

□ **PallaviPatel** 1 year, 11 months ago

Selected Answer: A

correct

upvoted 2 times

□ **kilowd** 1 year, 11 months ago

Selected Answer: A

Data Engineers - High Concurrency cluster as it provides for sharing . Also caters for SQL,Python and R.

Data Scientist - Standard Clusters which automatically terminates after 120 minutes and caters for Scala,SQL,Python and R.

JOBS- Standard Cluster

upvoted 2 times

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 plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- A workload for data engineers who will use Python and SQL.
- A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a High Concurrency cluster for the jobs.

Does this meet the goal?

A. Yes

B. No

Correct Answer: A

We need a High Concurrency cluster for the data engineers and the jobs.

Note: Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.

A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

Reference:

<https://docs.azuredatabricks.net/clusters/configure.html>

Community vote distribution

B (100%)

 **dfdsfdsfsd** Highly Voted 2 years, 7 months ago

High-concurrency clusters do not support Scala. So the answer is still 'No' but the reasoning is wrong.

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

upvoted 46 times

 **Preben** 2 years, 7 months ago

I agree that High concurrency does not support Scala. But they specified using a Standard cluster for the jobs, which does support Scala. Why is the answer 'No'?

upvoted 3 times

 **eng1** 2 years, 6 months ago

Because the High Concurrency cluster for each data scientist is not correct, it should be standard for a single user!

upvoted 6 times

 **FRAN_CO_HO** Highly Voted 2 years, 6 months ago

Answer should be NO, which

Data scientist: STANDARD as need to run scala

Jobs: STANDARD as need to run scala

Data Engineers: High-concurrency clusters as better resource sharing

upvoted 14 times

 **Tactable** Most Recent 2 months, 2 weeks ago

High concurrency doesn't support Scala

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: B

Answer is No.

upvoted 1 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: B

Correct answer: false (no)

upvoted 1 times

 **Pais** 1 year, 1 month ago

Selected Answer: B

High-concurrency cluster does not support Scala.

upvoted 1 times

 **OldSchool** 1 year, 1 month ago

Selected Answer: B

Jobs require Scala so the answer is B) No.

upvoted 1 times

 **greenlever** 1 year, 2 months ago

Selected Answer: B

Cluster for Jobs should support scala - STANDARD

upvoted 1 times

 **anks84** 1 year, 4 months ago

We would need a Standard cluster for the jobs to support Scala. High-concurrency cluster does not support Scala.

Hence, Answer is NO

upvoted 1 times

 **Hema_V** 1 year, 4 months ago

Selected Answer: B

High Concurrency clusters can run workloads developed in SQL, Python, and R. The performance and security of High Concurrency clusters is provided by running user code in separate processes, which is not possible in Scala.

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

upvoted 1 times

 **Deeksha1234** 1 year, 4 months ago

No is correct

upvoted 1 times

 **ClassMistress** 1 year, 7 months ago

Selected Answer: B

High Concurrency clusters is provided by running user code in separate processes, which is not possible in Scala.

upvoted 1 times

 **narendra399** 1 year, 9 months ago

1 and 2 are same questions but answers are different why?

upvoted 2 times

 **Hanse** 1 year, 10 months ago

As per Link: <https://docs.azuredatabricks.net/clusters/configure.html>

Standard and Single Node clusters terminate automatically after 120 minutes by default. --> Data Scientists

High Concurrency clusters do not terminate automatically by default.

A Standard cluster is recommended for a single user. --> Standard for Data Scientists & High Concurrency for Data Engineers

Standard clusters can run workloads developed in any language: Python, SQL, R, and Scala.

High Concurrency clusters can run workloads developed in SQL, Python, and R. The performance and security of High Concurrency clusters is provided by running user code in separate processes, which is not possible in Scala. --> Jobs needs Standard

upvoted 2 times

 **lukeonline** 2 years ago

Selected Answer: B

high concurrency does not support scala

upvoted 2 times

 **rashjan** 2 years, 1 month ago

Selected Answer: B

wrong: no

upvoted 1 times

 **FredNo** 2 years, 1 month ago

Selected Answer: B

Answer is no because high concurrency does not support scala
upvoted 5 times

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

You are designing a folder structure for the files in an Azure Data Lake Storage Gen2 account. The account has one container that contains three years of data.

You need to recommend a folder structure that meets the following requirements:

- Supports partition elimination for queries by Azure Synapse Analytics serverless SQL pools
- Supports fast data retrieval for data from the current month
- Simplifies data security management by department

Which folder structure should you recommend?

- A. \Department\DataSource\YYYY\MM\DataFile_YYYYMMDD.parquet
- B. \DataSource\Department\YYYYMM\DataFile_YYYYMMDD.parquet
- C. \DD\MM\YYYY\Department\DataSource\DataFile_DDMMYY.parquet
- D. \YYYY\MM\DD\Department\DataSource\DataFile_YYYYMMDD.parquet

Correct Answer: A

Department top level in the hierarchy to simplify security management.

Month (MM) at the leaf/bottom level to support fast data retrieval for data from the current month.

Community vote distribution

| | |
|---------|-----|
| A (88%) | 12% |
|---------|-----|

 **anks84** Highly Voted 1 year, 4 months ago

Selected Answer: A

Answer is Correct !

upvoted 9 times

 **dmitriypo** Highly Voted 1 year, 2 months ago

Selected Answer: A

Of course A

upvoted 5 times

 **kkk5566** Most Recent 4 months ago

Selected Answer: A

A is correct

upvoted 1 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: B

The raw zone may be organised by source system, then entity. Here is an example folder structure, optimal for folder security:
\Raw\DataSource\Entity\YYYY\MM\DD\File.extension

Typically each source system will be granted write permissions at the DataSource folder level with default ACLs (see section on ACLs below) specified. This will ensure permissions are inherited as new daily folders and files are created.

Whilst many use time based partitioning there are a number of options which may provide more efficient access paths. ->this justify the YYYYMM to get easily the current month. Correct answer for me is B.

upvoted 2 times

 **MarkJoh** 1 month ago

Except is contradicts the requirement "Simplifies data security management by department". You need minimum (#DataSource) * (#Department) ACLs with option B

Whereas with option A, you need just minimum #Department ACLs.

Also, Option A allows you to have additional datasource specific ACLs within a particular department, if that is so desired.

upvoted 1 times

 **rzeng** 1 year, 2 months ago

A is right

upvoted 4 times

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 receives new data once every 24 hours.

You have the following function.

```
create function dbo.udfFtoC(F decimal)
return decimal
as
begin
    return (F - 32) * 5.0 / 9
end
```

You have the following query.

```
select avg_date, sensorid, avg_f, dbo.udfFtoC(avg_temperature) as avg_c from SensorTemps
where avg_date = @parameter
```

The query is executed once every 15 minutes and the @parameter value is set to the current date.

You need to minimize the time it takes for the query to return results.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Create an index on the avg_f column.
- B. Convert the avg_c column into a calculated column.
- C. Create an index on the sensorid column.
- D. Enable result set caching.
- E. Change the table distribution to replicate.

Correct Answer: BD

D: 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.

Incorrect:

Not A, not C: No joins so index not helpful.

Not E: What is a replicated table?

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. If the data is static and does not change, you can replicate larger tables.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-caching>

Community vote distribution

BD (71%)

DE (23%)

3%

 esaade Highly Voted 10 months ago

Selected Answer: BD

- B. Convert the avg_c column into a calculated column.
- D. Enable result set caching.

Explanation:

A calculated column is a column that uses an expression to calculate its value based on other columns in the same table. In this case, the udfFtoC function can be used to calculate the avg_c value based on the avg_temperature column, eliminating the need to call the UDF in the SELECT statement.

Enabling result set caching can improve query performance by caching the result set of the query, so subsequent queries that use the same parameters can be retrieved from the cache instead of executing the query again.

Creating an index on the avg_f column or the sensorid column is not useful because there are no join or filter conditions on these columns in the WHERE clause. Changing the table distribution to replicate is also not necessary because it does not affect the query performance in this scenario

upvoted 11 times

□  **dakku987** Most Recent 1 week, 2 days ago

Selected Answer: AC

chatgpt

To minimize the time it takes for the query to return results, you should consider the following actions:

A. Create an index on the avg_f column:

Creating an index on the avg_f column can improve the query performance, especially if there are frequent searches or filtering based on this column.

C. Create an index on the sensorid column:

If the sensorid column is frequently used in filtering or joins, creating an index on this column can improve the query performance.

upvoted 1 times

□  **OldSchool** 3 months, 2 weeks ago

Selected Answer: DE

First the wording of the question is ridiculous. "Query is executed once every 15 minutes".

So what is it, "Once" or "every 15 minutes"?

Either way, they are asking what to do to speed up the query.

D Setting result caching

E Replicated distribution

upvoted 1 times

□  **kkk5566** 4 months, 1 week ago

Selected Answer: DE

correct

upvoted 1 times

□  **kkk5566** 4 months, 1 week ago

Selected Answer: BD

correct

upvoted 1 times

□  **kkk5566** 4 months, 1 week ago

D & E should be correct

upvoted 1 times

□  **Matt2000** 5 months ago

Calculated columns exist in Power BI, not dedicated SQL pools. Computed columns are not supported in dedicated SQL pools.

Ref: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overview>

upvoted 2 times

□  **dumbled** 8 months, 2 weeks ago

Selected Answer: BD

correct

upvoted 2 times

□  **Lestrang** 11 months, 3 weeks ago

Selected Answer: AB

With that point by erhard being made (caching does work with queries using UDF), the most commonly voted D is wrong, so B and what now? Replicated cannot be right because it received date everyday and has aggregations so not a dim table and we have no clue about its size. by elimination that leaves us A and C

Indexing is less useful with no joins but it does improve some performance being on where clause target. so I'd go with A and B.

upvoted 1 times

□  **Lestrang** 11 months, 3 weeks ago

Creating an index on the avg_f column will improve the performance of the query, as it will allow the query to find the relevant data more quickly. Converting the avg_c column into a calculated column will allow the query to return the temperature in Celsius without the need to perform the calculation at runtime, which will also improve the performance of the query.

upvoted 1 times

□  **Lestrang** 11 months, 2 weeks ago

After re-considering, I am unsure whether the indexing would help. That would only leave Replication as the viable option even though it is not viable design but the request is to minimize query time and that is what it will do, so I guess final answer is BE

upvoted 1 times

□  **Karforcerts** 1 year, 1 month ago

Selected Answer: BD

need to first change UDF to a calculated column and then enable result set caching. agreed with the answer

upvoted 4 times

□  **erhard** 1 year, 1 month ago

Queries using user defined functions are not cached.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-caching>

upvoted 4 times

kl8585 1 year, 1 month ago

Selected Answer: DE

A,C not right since index don't help if join are not involved.

D for sure help query performance.

I don't get why B:

"A computed column is a virtual column whose value is calculated from other values in the table. By default, the expression's outputted value is not physically stored. Instead, SQL Server runs the expression when the column is queried and returns the value as part of the result set ... In many cases, non-persistent computed columns put too much burden on the processor, resulting in SLOWER QUERIES and unresponsive applications"

Since the only requirements is faster execution times for queries, i don't think calculated columns will improve performance.

Si second option for me would be D (replicate). Although it will cause more effort writing, because updates should be written to every partition, optimized writes aren't a requirement in the question.

upvoted 3 times

rzeng 1 year, 2 months ago

pool ingest data once per 24 hrs, while query happens every 15mins, caching result can definitely avoid the some duplicate calculation, I'll go with BD.

upvoted 1 times

Xinyuehong 1 year, 2 months ago

Selected Answer: DE

I think should be DE.

since "the query is executed once every 15 minutes and the @parameter value is set to the current date", and the it receives new data once every 24 hours, it means the query result isn't change in one day even you run it every 15 mins. The data is static within a day. Replication could help the performance.

upvoted 2 times

anks84 1 year, 4 months ago

Selected Answer: BD

Answer is Correct !

upvoted 4 times

You need to design a solution that will process streaming data from an Azure Event Hub and output the data to Azure Data Lake Storage. The solution must ensure that analysts can interactively query the streaming data.

What should you use?

- A. Azure Stream Analytics and Azure Synapse notebooks
- B. Structured Streaming in Azure Databricks
- C. event triggers in Azure Data Factory
- D. Azure Queue storage and read-access geo-redundant storage (RA-GRS)

Correct Answer: B

Community vote distribution

B (61%)

A (39%)

 **esaade** Highly Voted 9 months, 4 weeks ago

Selected Answer: B

B. Structured Streaming in Azure Databricks is the best option for this scenario as it allows for processing of streaming data and outputting it to Azure Data Lake Storage, while also providing the ability for analysts to interactively query the data using Databricks notebooks.

Azure Stream Analytics and Azure Synapse notebooks (option A) can also process streaming data and output to Data Lake Storage, but they may not provide the same level of interactivity for analysts.

Event triggers in Azure Data Factory (option C) can help automate data movement between Event Hubs and Data Lake Storage, but they do not provide the necessary functionality for processing and querying streaming data.

Azure Queue Storage and read-access geo-redundant storage (RA-GRS) (option D) are not relevant for this scenario as they do not provide capabilities for processing and querying streaming data.

upvoted 14 times

 **Sirstyle** Most Recent 2 weeks, 2 days ago

selected answer : B

to visualize data in stream analytics you use SQL query in the Azure portal inside synapse , not a notebook , therefore the answer is B

upvoted 1 times

 **Andrew_Chen** 2 months, 3 weeks ago

Selected Answer: A

A. Azure Stream Analytics and Azure Synapse notebooks:

Azure Stream Analytics can be used to process the streaming data from Azure Event Hub and output the data to Azure Data Lake Storage. Azure Synapse notebooks provide interactive querying capabilities, and they can be integrated with the Azure Data Lake Storage to enable analysts to run their analytics on the stored data.

B. Structured Streaming in Azure Databricks:

Structured Streaming in Azure Databricks indeed supports streaming data and can write outputs to Azure Data Lake Storage. However, the question emphasizes "interactively querying" the streaming data, and while Databricks notebooks allow for interactive queries, Azure Synapse notebooks are better integrated with Microsoft's suite of data tools for broader analytics purposes.

upvoted 4 times

 **ExamDestroyer69** 3 weeks ago

@Andrew_Chen Azure Databricks with Structured Streaming is preferred over Azure Stream Analytics and Azure Synapse Notebooks for real-time streaming data processing from Azure Event Hubs due to its native support for continuous processing, live querying, and seamless integration with Azure Data Lake Storage.

upvoted 2 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: A

should be correct

upvoted 1 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: A

What streaming sources and sinks does Azure Databricks support?

Databricks recommends using Auto Loader to ingest supported file types from cloud object storage into Delta Lake. For ETL pipelines, Databricks recommends using Delta Live Tables (which uses Delta tables and Structured Streaming). You can also configure incremental ETL workloads by streaming to and from Delta Lake tables.

In addition to Delta Lake and Auto Loader, Structured Streaming can connect to messaging services such as Apache Kafka.

<https://learn.microsoft.com/en-us/azure/databricks/structured-streaming/>

I don't see data lake in the list, so probably the answer is A.

upvoted 1 times

✉ **vadiminski_a** 9 months, 1 week ago

Selected Answer: B

I am in favour of B because of this piece of information I have encountered:

<https://www.databricks.com/spark/getting-started-with-apache-spark/streaming>

upvoted 3 times

✉ **vadiminski_a** 9 months, 1 week ago

On the other hand, there is this: <https://learn.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics>

So I believe both to be valid, Azure Stream Analytics seems to be more straightforward

upvoted 1 times

✉ **Kate0204** 10 months, 1 week ago

Selected Answer: A

An Azure Stream Analytics job consists of an input, query, and an output.

upvoted 1 times

✉ **Karl_Cen** 11 months, 2 weeks ago

"The solution must ensure that analysts can interactively query the streaming data"

Streaming analysis can't query streaming data interactively

upvoted 2 times

✉ **Lestrang** 11 months, 3 weeks ago

Selected Answer: A

B. Structured Streaming in Azure Databricks is incorrect because while it allows you to process streaming data using Spark's structured streaming API, it is not designed to directly output the data to Azure Data Lake Storage. Instead, it typically outputs the data to storage systems like HDFS, S3, or Cosmos DB. Additionally, Databricks is a separate service that does not integrate with Azure Synapse for interactive querying. While it's possible to use Databricks to read the data from Data Lake Storage and use Spark to process the data and then write it back to Data Lake Storage, it will not be as efficient as using Azure Stream Analytics for this use case as it is specifically designed for streaming data processing and also has built-in connectors to various data storage and analytics services like Data Lake Storage

upvoted 2 times

✉ **Lestrang** 11 months, 2 weeks ago

Although this might be true, after some pondering, the given solution A. Azure Stream Analytics and Azure Synapse notebooks requires a Synapse workspace which is not implied.

So I guess it would be databricks.

upvoted 1 times

✉ **Mal2002** 7 months ago

It's implied. Solutions said Azure Stream Analytics and Azure Synapse Notebook, Azure Synapse notebook cannot be created without Azure Synapse Workspace.

upvoted 2 times

✉ **alexnicolita** 11 months, 4 weeks ago

Selected Answer: A

Why not Azure Stream Analytics and Azure Synapse Analytics?

upvoted 2 times

You are creating an Apache Spark job in Azure Databricks that will ingest JSON-formatted data.

You need to convert a nested JSON string into a DataFrame that will contain multiple rows.

Which Spark SQL function should you use?

- A. explode
- B. filter
- C. coalesce
- D. extract

Correct Answer: A

Community vote distribution

A (100%)

✉ **Rajcse03** Highly Voted 11 months, 4 weeks ago

Selected Answer: A

<https://learn.microsoft.com/en-us/azure/databricks/kb-scala/flatten-nested-columns-dynamically>
upvoted 6 times

✉ **warre** Most Recent 1 day, 18 hours ago

Selected Answer: A

The explode function in Spark SQL is used to transform an array or a map column into multiple rows, essentially "exploding" the nested structure
upvoted 1 times

✉ **hassexat** 4 months ago

Selected Answer: A

A is correct answer!
upvoted 1 times

✉ **kkk5566** 4 months, 1 week ago

Selected Answer: A

A is correct
upvoted 1 times

✉ **[Removed]** 12 months ago

Selected Answer: A

correct
upvoted 3 times

DRAG DROP

You have an Azure subscription that contains an Azure Databricks workspace. The workspace contains a notebook named Notebook1.

In Notebook1, you create an Apache Spark DataFrame named df_sales that contains the following columns:

- Customer
- SalesPerson
- Region
- Amount

You need to identify the three top performing salespersons by amount for a region named HQ.

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.

| Values | Answer Area |
|---|--|
| <code>agg(col('SalesPerson'))</code> | <code>df_sales.filter(col('Region')=='HQ').</code> <input type="text"/> |
| <code>filter(col('SalesPerson'))</code> | <code>.agg(sum('Amount').alias('TotalAmount')).</code> <input type="text"/> <code>.limit(3)</code> |
| <code>groupBy(col('SalesPerson'))</code> | |
| <code>groupBy(col('TotalAmount'))</code> | |
| <code>orderBy(col('TotalAmount'))</code> | |
| <code>orderBy(desc('TotalAmount'))</code> | |

Answer Area

Correct Answer: `df_sales.filter(col('Region')=='HQ').` `groupBy(col('SalesPerson'))`
`.agg(sum('Amount').alias('TotalAmount')).` `orderBy(desc('TotalAmount'))` `limit(3)`

 esaade Highly Voted 10 months ago

```
df_sales.filter(col("Region") == "HQ")
.groupBy(col('SalesPerson'))
.agg(sum('Amount').alias('TotalAmount'))
.orderBy(desc('TotalAmount'))
.limit(3)
```

upvoted 13 times

 kkk5566 Highly Voted 4 months, 1 week ago

```
.groupBy(col('SalesPerson')) and .orderBy(desc('TotalAmount'))
upvoted 5 times
```

 aurorafang Most Recent 11 months, 2 weeks ago

for the sequence, group by usually put before the order by operations
upvoted 4 times

 [Removed] 12 months ago

correct

upvoted 3 times

You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container.

Which type of trigger should you use?

- A. on-demand
- B. tumbling window
- C. schedule
- D. storage event

Correct Answer: D

Community vote distribution

D (100%)

 **FRANCIS_A_M** Highly Voted 9 months, 1 week ago

Selected Answer: D

Correct, D

upvoted 9 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: D

correct

upvoted 2 times

 **haythemsi** 8 months ago

Correct, D

upvoted 2 times

 **AHUI** 9 months, 1 week ago

ans is correct

upvoted 1 times

DRAG DROP

You have a project in Azure DevOps that contains a repository named Repo1. Repo1 contains a branch named main.

You create a new Azure Synapse workspace named Workspace1.

You need to create data processing pipelines in Workspace1. The solution must meet the following requirements:

- Pipeline artifacts must be stored in Repo1
- Source control must be provided for pipeline artifacts.
- All development must be performed in a feature branch.

Which four actions should you perform in sequence in Synapse Studio? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

| Actions | Answer Area |
|---|-------------|
| Create pipeline artifacts and save them in the main branch. | |
| Set the main branch as the collaboration branch. | |
| Create a pull request to merge the contents of the main branch into the new branch. | |
| Create pipeline artifacts and save them in the new branch. | |
| Create a new branch. | |
| Configure a code repository and select Repo1. | |

Correct Answer:

| Answer Area |
|---|
| Configure a code repository and select Repo1. |
| Create a new branch. |
| Create pipeline artifacts and save them in the new branch. |
| Create a pull request to merge the contents of the main branch into the new branch. |

✉️  **SinSS** Highly Voted 7 months, 3 weeks ago

Configure a code repo and select Repo1
Set the main branch as the collaboration branch
Create a new brach
Create pipeline artifacts and save them in the new branch
upvoted 18 times

✉️  **macinpune9** 1 month ago

This is wrong, given answer by exam topics is right
upvoted 3 times

✉️  **j4g092t** 4 months ago

Isn't the main branch by default the collaboration branch?
upvoted 1 times

✉️  **mhi** Highly Voted 7 months, 4 weeks ago

Shouldn't you merge the new branch into the main branch?
upvoted 12 times

✉️  **peches** 7 months ago

Agree, you create a feature branch from the collaboration branch, work on it, and after you finished you merge back to the collaboration branch (by default is main). Source: <https://learn.microsoft.com/en-us/azure/synapse-analytics/cicd/source-control#version-control>
upvoted 3 times

✉️  **kkk5566** Most Recent 4 months, 1 week ago

correct
upvoted 1 times

✉️  **kkk5566** 4 months, 1 week ago

Configure a code repo and select Repo1
Set the main branch as the collaboration branch
Create a new brach
and PR
upvoted 2 times

 **kkk5566** 4 months ago

forgot it ,the given answer is right.
upvoted 3 times

 **AlviraTony** 4 months, 2 weeks ago

Given solution is correct
upvoted 3 times

 **Matt2000** 5 months ago

"Configure a code repository and select Repo2" is not required as you already have a repo Repo1 with main as branch.
upvoted 1 times

 **FRANCIS_A_M** 9 months, 1 week ago

Correct
upvoted 6 times

You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1.

You need to use an Azure Synapse pipeline to copy data from the selected tables in DB1 to the files in storage1. The solution must meet the following requirements:

- The Copy activity in the pipeline must be parameterized to use the data in File1.txt to identify the source and destination of the copy.
- Copy activities must occur in parallel as often as possible.

Which two pipeline activities should you include in the pipeline? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Get Metadata
- B. Lookup
- C. ForEach
- D. If Condition

Correct Answer: AC

Community vote distribution

BC (100%)

✉  **FRANCIS_A_M** Highly Voted 9 months, 1 week ago

Selected Answer: BC

It's BC. Use the LookUp Activity to read the .txt file. ForEach to Loop though making sure Sequential is off (which off by default) for parallelization
upvoted 9 times

✉  **ExamDestroyer69** 3 weeks ago

This is correct, Get Metadata doesn't directly retrieve the content of the file itself. If you want to obtain the content of a .txt file like File1.txt, specifically the table names contained within it, you'd typically use Lookup.

Correct Answer: BC

upvoted 1 times

✉  **aemilka** Highly Voted 8 months, 4 weeks ago

Selected Answer: BC

Lookup activity reads and returns the content of a configuration file or table. It also returns the result of executing a query or stored procedure. The output can be a singleton value or an array of attributes, which can be consumed in a subsequent copy, transformation, or control flow activities like ForEach activity.

<https://learn.microsoft.com/en-us/azure/data-factory/control-flow-lookup-activity>

upvoted 6 times

✉  **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: BC

correct

upvoted 2 times

✉  **examtopicsofyannick** 5 months ago

Selected Answer: BC

BC - Lookup and ForEach.

Lookup - reads .txt file

ForEach - iteration through contents read in Lookup activity for COPY activity

upvoted 2 times

✉  **auwia** 6 months, 2 weeks ago

Selected Answer: BC

Correct answers.

upvoted 2 times

✉  **vctrhugo** 6 months, 3 weeks ago

Selected Answer: BC

B. Lookup: The Lookup activity can be used to read the contents of File1.txt from the storage account. It will retrieve the names of selected tables in DB1 as parameter values for the Copy activity.

C. ForEach: The ForEach activity can be used to iterate over the retrieved table names from File1.txt. Inside the loop, you can configure the Copy activity with the source and destination information based on the current table name.

upvoted 2 times

 **shakes103** 9 months ago

Selected Answer: BC

Answer is B and C.

upvoted 4 times

 **Sibaprasad** 9 months, 1 week ago

'Get Metadata' cannot read the content of the file. Its Lookup and ForEach.

Refer to link : <https://learn.microsoft.com/en-us/azure/data-factory/control-flow-get-metadata-activity> and <https://learn.microsoft.com/en-us/azure/data-factory/control-flow-lookup-activity>

upvoted 3 times

You have an Azure data factory that connects to a Microsoft Purview account. The data factory is registered in Microsoft Purview.

You update a Data Factory pipeline.

You need to ensure that the updated lineage is available in Microsoft Purview.

What should you do first?

- A. Disconnect the Microsoft Purview account from the data factory.
- B. Execute the pipeline.
- C. Execute an Azure DevOps build pipeline.
- D. Locate the related asset in the Microsoft Purview portal.

Correct Answer: B

Community vote distribution

B (100%)

✉  **Sibaprasad**  9 months, 1 week ago

B. Execute the Pipeline is correct answer.

Refer link : <https://learn.microsoft.com/en-us/azure/data-factory/tutorial-push-lineage-to-purview> and <https://learn.microsoft.com/en-us/azure/data-factory/connect-data-factory-to-azure-purview>

upvoted 11 times

✉  **aemilka** 8 months ago

Correct.

"The lineage data will automatically be captured during the activities execution."

upvoted 2 times

✉  **kkk5566**  4 months, 1 week ago

Selected Answer: B

executing the pipeline

upvoted 1 times

✉  **vctrhugo** 6 months, 3 weeks ago

Selected Answer: B

By executing the pipeline, the Data Factory will generate the lineage information and propagate it to the connected Microsoft Purview account. This will update the lineage in Purview and reflect any changes made in the pipeline.

upvoted 2 times

✉  **Gopinath123** 8 months, 2 weeks ago

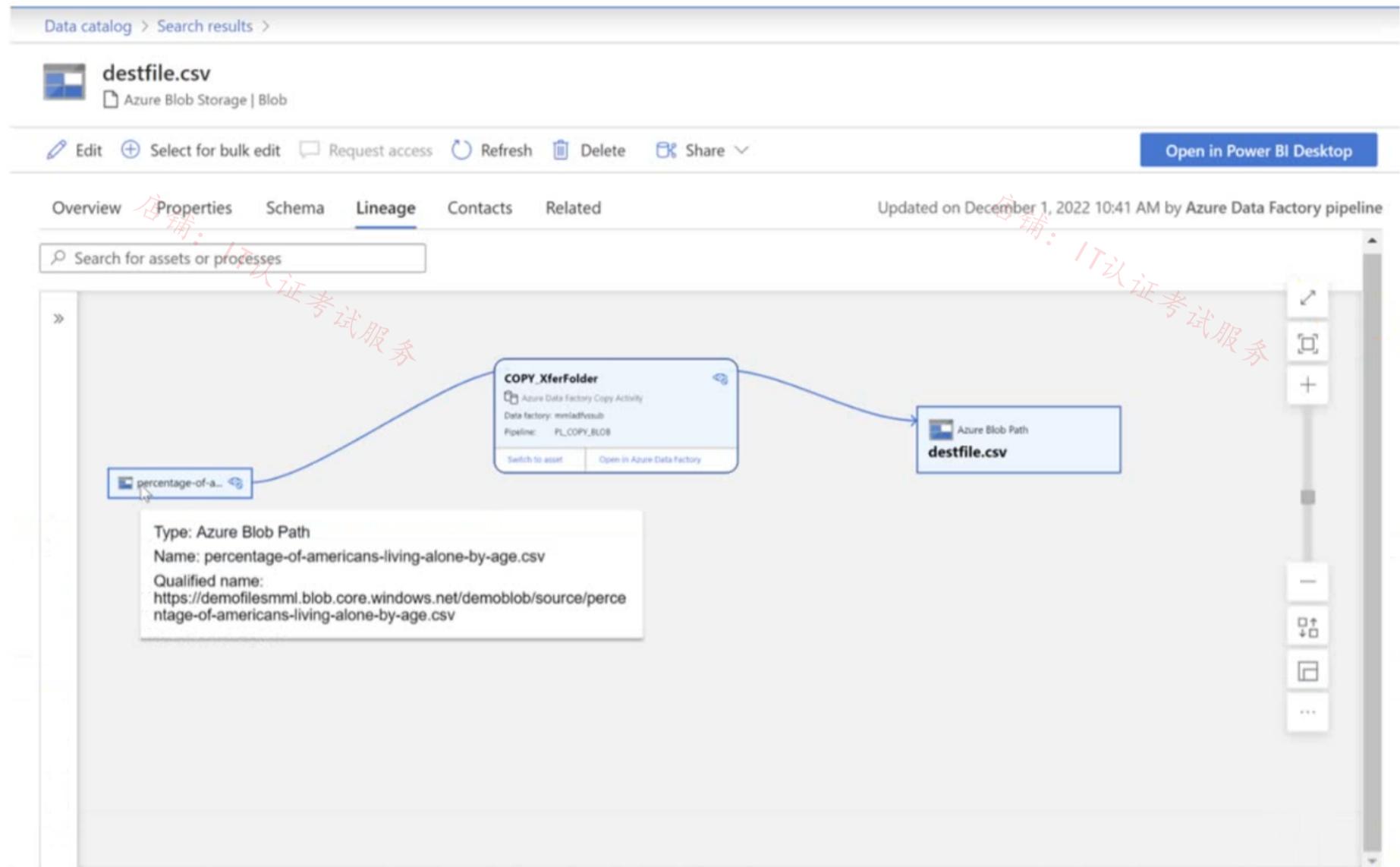
Selected Answer: B

<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-push-lineage-to-purview>

upvoted 2 times

You have a Microsoft Purview account.

The Lineage view of a CSV file is shown in the following exhibit.



How is the data for the lineage populated?

- A. manually
- B. by scanning data stores
- C. by executing a Data Factory pipeline

Correct Answer: C

Community vote distribution

C (100%)

✉️ **shakes103** Highly Voted 8 months, 3 weeks ago

Selected Answer: C

Answer is C

Find reason here: <https://learn.microsoft.com/en-us/azure/data-factory/tutorial-push-lineage-to-purview#run-pipeline-and-push-lineage-data-to-microsoft-purview>

upvoted 5 times

✉️ **shakes103** 8 months, 3 weeks ago

The answer is also displayed on the top right corner of the image displayed.

upvoted 14 times

✉️ **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: C

C is correct

upvoted 1 times

You have an Azure subscription that contains a Microsoft Purview account named MP1, an Azure data factory named DF1, and a storage account named storage1. MP1 is configured to scan storage1. DF1 is connected to MP1 and contains a dataset named DS1. DS1 references a file in storage1.

In DF1, you plan to create a pipeline that will process data from DS1.

You need to review the schema and lineage information in MP1 for the data referenced by DS1.

Which two features can you use to locate the information? Each correct answer presents a complete solution.

NOTE: Each correct answer is worth one point.

- A. the search bar in the Microsoft Purview governance portal
- B. the Storage browser of storage1 in the Azure portal
- C. the search bar in the Azure portal
- D. the search bar in Azure Data Factory Studio

Correct Answer: AB

Community vote distribution

AD (92%) 8%

 **Sibaprasad** Highly Voted 9 months, 1 week ago

From ChatGPT :

- A. the search bar in the Microsoft Purview governance portal
- D. the search bar in Azure Data Factory Studio

To review the schema and lineage information in MP1 for the data referenced by DS1, you can use the following two features:

The search bar in the Microsoft Purview governance portal: You can search for the file in storage1 that is referenced by DS1 in the search bar of the Purview governance portal. Once you locate the file, you can view the schema and lineage information for it.

The search bar in Azure Data Factory Studio: You can search for the dataset DS1 in the Azure Data Factory Studio search bar. Once you locate the dataset, you can view the schema and lineage information for the data it references in storage1, which can also be viewed in Purview.

upvoted 15 times

 **chryckie** Highly Voted 8 months, 3 weeks ago

Selected Answer: AD

You need lineage info. Lineage is in Purview. Also, the lineage is all based off what the Data Factory pipeline is doing. I'd say A and D.

<https://learn.microsoft.com/en-us/azure/purview/how-to-search-catalog#searching-microsoft-purview-in-connected-services>
upvoted 7 times

 **ExamDestroyer69** Most Recent 3 weeks ago

Selected Answer: AB

I don't fully understand this discussion, I have been led to believe that Azure Data Factory Studio is primarily used to search for components within the data factory itself, such as datasets, pipelines, activities, linked services, etc. It is useful for finding and managing resources within the data factory BUT not for exploring external metadata or lineage information stored in Purview.

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: AD

is correct

upvoted 1 times

 **peches** 7 months, 1 week ago

Selected Answer: AD

If the Data Factory resource is connected to a Purview account there will be a column in the monitoring view of the Pipeline with the lineage status. <https://learn.microsoft.com/en-us/azure/data-factory/tutorial-push-lineage-to-purview#step-3-monitor-lineage-reporting-status>
upvoted 3 times

HOTSPOT

You have an Azure Blob storage account that contains a folder. The folder contains 120,000 files. Each file contains 62 columns.

Each day, 1,500 new files are added to the folder.

You plan to incrementally load five data columns from each new file into an Azure Synapse Analytics workspace.

You need to minimize how long it takes to perform the incremental loads.

What should you use to store the files and in which format? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area**Storage:**

- Multiple blob storage accounts
- Multiple containers in the blob storage account
- Timeslice partitioning in the folders

Format:

- Apache Parquet
- CSV
- JSON

Answer Area**Storage:**

- Multiple blob storage accounts
- Multiple containers in the blob storage account
- Timeslice partitioning in the folders

Correct Answer:**Format:**

- Apache Parquet
- CSV
- JSON

 **ababatunde_hs** Highly Voted 9 months, 1 week ago

Time partitioning is correct as the fastest way to load only new files, but requires that the timeslice information be part of the file or folder name (<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-overview>)

However, Parquet is the correct file format since it's a columnar format
upvoted 42 times

 **kkk5566** Highly Voted 4 months, 1 week ago

Time partitioning and parquet
upvoted 8 times

 **vctrhugo** Most Recent 6 months, 3 weeks ago

You need to minimize how long it takes to perform the incremental loads. With Parquet, which is a columnar format, it is way faster to select a few columns than csv.
upvoted 2 times

 **vegeta379** 7 months, 2 weeks ago

we can do incremental load just with deltatable for a parquet file which supported by datarbricks or synapse spark and here he didn't give details so I think it will be CSV

upvoted 1 times

 **pavankr** 7 months, 2 weeks ago

I think the requirement is to select specific columns, hence CSV?

upvoted 1 times

 **verisdev** 7 months, 4 weeks ago

it supposed to be Parquet instead of CSV

upvoted 5 times

DRAG DROP

You are batch loading a table in an Azure Synapse Analytics dedicated SQL pool.

You need to load data from a staging table to the target table. The solution must ensure that if an error occurs while loading the data to the target table, all the inserts in that batch are undone.

How should you complete the Transact-SQL 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 split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

| Values | Answer Area |
|--|---|
| <code>BEGIN DISTRIBUTED TRANSACTION</code> | |
| <code>BEGIN TRAN</code> | <code>BEGIN TRY</code> |
| <code>COMMIT TRAN</code> | <code>INSERT INTO dbo.Table1 (col1, col2, col3)</code> |
| <code>ROLLBACK TRAN</code> | <code>SELECT col1, col2, col3 FROM stage.Table1;</code> |
| <code>SET RESULT_SET_CACHING ON</code> | <code>END TRY</code> |
| | <code>BEGIN CATCH</code> |
| | <code>IF @@TRANCOUNT > 0</code> |
| | <code>BEGIN</code> |
| | <code>ROLLBACK TRAN</code> |
| | <code>END</code> |
| | <code>END CATCH;</code> |
| | <code>IF @@TRANCOUNT > 0</code> |
| | <code>BEGIN</code> |
| | <code>COMMIT TRAN;</code> |
| | <code>END</code> |

| Answer Area | |
|---|--|
| <code>BEGIN DISTRIBUTED TRANSACTION</code> | |
| <code>BEGIN TRY</code> | |
| <code>INSERT INTO dbo.Table1 (col1, col2, col3)</code> | |
| <code>SELECT col1, col2, col3 FROM stage.Table1;</code> | |
| <code>END TRY</code> | |
| <code>BEGIN CATCH</code> | |
| <code>IF @@TRANCOUNT > 0</code> | |
| <code>BEGIN</code> | |
| <code>ROLLBACK TRAN</code> | |
| <code>END</code> | |
| <code>END CATCH;</code> | |
| <code>IF @@TRANCOUNT > 0</code> | |
| <code>BEGIN</code> | |
| <code>COMMIT TRAN;</code> | |
| <code>END</code> | |

Correct Answer:

Given answer is wrong. It should be BEGIN TRAN as SQL pool in Azure Synapse Analytics does not support distributed transaction.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-develop-transactions>

"Limitations

SQL pool does have a few other restrictions that relate to transactions.

They are as follows:

No distributed transactions
No nested transactions permitted
No save points allowed
No named transactions
No marked transactions
No support for DDL such as CREATE TABLE inside a user-defined transaction
"

Distributed Transactions are only allowed in SQL Server and Azure SQL Managed Instance:

<https://learn.microsoft.com/de-de/sql/t-sql/language-elements/begin-distributed-transaction-transact-sql?view=sql-server-ver16>
upvoted 21 times

 **janaki** Highly Voted 7 months, 1 week ago

Its BEGIN TRAN
then ROLLBACK TRAN
upvoted 10 times

 **Ram9198** Most Recent 1 week, 6 days ago

It should be BEGIN TRAN
and ROLLBACK TRAN
upvoted 1 times

 **kkk5566** 4 months, 1 week ago

BEGIN TRAN
ROLLBACK TRAN
upvoted 1 times

HOTSPOT

You have two Azure SQL databases named DB1 and DB2.

DB1 contains a table named Table1. Table1 contains a timestamp column named LastModifiedOn. LastModifiedOn contains the timestamp of the most recent update for each individual row.

DB2 contains a table named Watermark. Watermark contains a single timestamp column named WatermarkValue.

You plan to create an Azure Data Factory pipeline that will incrementally upload into Azure Blob Storage all the rows in Table1 for which the LastModifiedOn column contains a timestamp newer than the most recent value of the WatermarkValue column in Watermark.

You need to identify which activities to include in the pipeline. The solution must meet the following requirements:

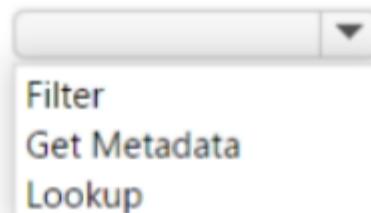
- Minimize the effort to author the pipeline.
- Ensure that the number of data integration units allocated to the upload operation can be controlled.

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

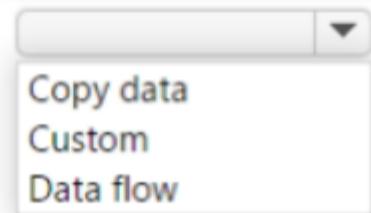
NOTE: Each correct answer is worth one point.

Answer Area

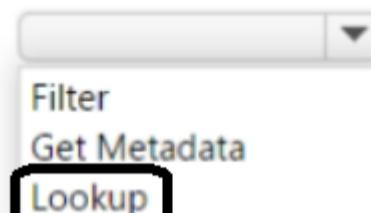
To retrieve the watermark value, use:



To perform the upload, use:

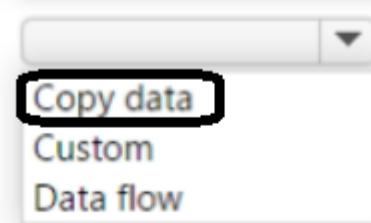
**Answer Area**

To retrieve the watermark value, use:



Correct Answer:

To perform the upload, use:



 **DarKru** Highly Voted 7 months, 1 week ago

Correct. The example is here
<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-portal>
upvoted 13 times

 **OfficeSaracus** Highly Voted 8 months, 1 week ago

Seems correct to me
upvoted 11 times

 **kkk5566** Most Recent 4 months, 1 week ago

lookup & copy activity
upvoted 1 times

 **haythemsi** 8 months ago

Filter not lookup, because we have to "Minimize the effort to author the pipeline" and we have only the LastModifiedOn column as information, we are not sure for lookup.

upvoted 3 times

 **auwia** 6 months, 2 weeks ago

The Filter activity in Azure Data Factory is used to filter an array of objects from a previous activity's output (typically from a Lookup activity). It cannot directly query a database or compare a value from a database (watermark in this case) against data in another database.

upvoted 4 times

HOTSPOT

You have an Azure Synapse serverless SQL pool.

You need to read JSON documents from a file by using the OPENROWSET function.

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 *  
FROM OPENROWSET  
(  
    BULK  
    'https://sourcedatalake.blob.core.windows.net/public/docs.json',  
    FORMAT =  
        'CSV'  
        'DELTA'  
        'JSON'  
        'PARQUET'  
    FIELDTERMINATOR = '0x0b',  
    FIELDQUOTE =  
        '0x09'  
        '0x0a'  
        '0x0b'  
        '0x0c'  
    ROWTERMINATOR = '0x0b'  
)  
WITH (jsondoc nvarchar(max) AS JsonDocuments
```

Answer Area

```
SELECT *
FROM OPENROWSET
(
    BULK
    'https://sourcedatalake.blob.core.windows.net/public/docs.json',
    FORMAT = 'CSV'
    FIELDTERMINATOR = '0x0b',
    FIELDQUOTE = '0x0b'
    ROWTERMINATOR = '0x0b'
)
WITH (jsondoc nvarchar(max) AS JsonDocuments)
```

Correct Answer:

店铺：IT认证考试服务

 **Yemeral** Highly Voted 8 months, 1 week ago

Correct. It's weird but best way to open a json is as a csv and with 0x0b for fieldterminator and fieldquote.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files>

upvoted 30 times

 **kkk5566** Most Recent 4 months, 1 week ago

Correct

upvoted 4 times

You use Azure Data Factory to create data pipelines.

You are evaluating whether to integrate Data Factory and GitHub for source and version control.

What are two advantages of the integration? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. additional triggers
- B. lower pipeline execution times
- C. the ability to save without publishing
- D. the ability to save pipelines that have validation issues

Correct Answer: CD

Community vote distribution

CD (83%) BC (17%)

✉ **akk_1289** Highly Voted 8 months, 1 week ago

- C. the ability to save without publishing
- D. the ability to save pipelines that have validation issues

When you integrate Data Factory and GitHub, you can save your pipelines to a GitHub repository without publishing them to Azure. This allows you to work on your pipelines in a development environment and then publish them to Azure when you are ready.

You can also save pipelines that have validation issues. This is because GitHub does not validate your pipelines when you save them. This allows you to work on your pipelines and fix the validation issues before you publish them to Azure.

upvoted 18 times

✉ **abhijit1990** 3 months, 2 weeks ago

absolutely right
upvoted 2 times

✉ **henryphchan** 8 months ago

agree with you
upvoted 3 times

✉ **kkk5566** 4 months, 1 week ago

agree with you
upvoted 1 times

✉ **vctrhugo** Most Recent 6 months, 3 weeks ago

Selected Answer: CD

C. The ability to save without publishing: Integrating Data Factory with GitHub allows you to save changes to your pipelines without immediately publishing them. This provides flexibility in terms of saving work-in-progress or experimental changes without impacting the production pipelines.

D. The ability to save pipelines that have validation issues: With GitHub integration, you can save pipelines that have validation issues. This is useful when you want to save your work-in-progress changes or modifications to a pipeline, even if it doesn't currently pass validation. You can continue to work on resolving the validation issues without losing your progress.

upvoted 2 times

✉ **aemilka** 8 months ago

Correct
upvoted 3 times

✉ **haythemsi** 8 months ago

Selected Answer: CD

Correct
upvoted 3 times

✉ **Aninina** 8 months, 1 week ago

Selected Answer: BC

I think B and C
upvoted 1 times

DRAG DROP

You have an Azure Synapse Analytics workspace named Workspace1.

You perform the following changes:

- Implement source control for Workspace1.
- Create a branch named Feature based on the collaboration branch.
- Switch to the Feature branch.
- Modify Workspace1.

You need to publish the changes to Azure Synapse.

From which branch should you perform each change? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point

| Branches | Answer Area |
|--|---|
| <input type="checkbox"/> Collaboration | Create a pull request: <input type="text"/> |
| <input type="checkbox"/> Publish | Publish the changes: <input type="text"/> |
| <input type="checkbox"/> Feature | |

Correct Answer:

| Answer Area |
|------------------------------------|
| Create a pull request: Feature |
| Publish the changes: Collaboration |

 **henryphchan** (Highly Voted) 8 months ago

Correct! It's a easy one.

upvoted 9 times

 **kkk5566** (Most Recent) 4 months, 1 week ago

Correct

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Seems correct.

upvoted 3 times

 **shakes103** 8 months, 1 week ago

Answer is correct

upvoted 3 times

You have two Azure Blob Storage accounts named account1 and account2.

You plan to create an Azure Data Factory pipeline that will use scheduled intervals to replicate newly created or modified blobs from account1 to account2.

You need to recommend a solution to implement the pipeline. The solution must meet the following requirements:

- Ensure that the pipeline only copies blobs that were created or modified since the most recent replication event.
- Minimize the effort to create the pipeline.

What should you recommend?

- A. Run the Copy Data tool and select Metadata-driven copy task.
- B. Create a pipeline that contains a Data Flow activity.
- C. Create a pipeline that contains a flowlet.
- D. Run the Copy Data tool and select Built-in copy task.

Correct Answer: A

Community vote distribution

D (100%)

 **Sabbath** Highly Voted 7 months ago

Selected Answer: D

Just use Built-in copy task, according to: <https://learn.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-lastmodified-copy-data-tool>

upvoted 12 times

 **vctrhugo** Highly Voted 6 months, 3 weeks ago

Selected Answer: D

"[...] use the Copy Data tool to create a pipeline that incrementally copies new and changed files only, from Azure Blob storage to Azure Blob storage. It uses LastModifiedDate to determine which files to copy."

<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-lastmodified-copy-data-tool>

upvoted 6 times

 **MarkJoh** Most Recent 1 month ago

A is correct because of the requirement "Minimize the effort to create the pipeline."

upvoted 3 times

 **Lewiasskick** 2 days, 11 hours ago

metadriven is for copy huge amounts of objects (for example, thousands of tables) or load data from large variety of sources

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: D

D is correct

upvoted 1 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: D

Create a data factory.

Use the Built-in Copy Data tool to create a pipeline.

Monitor the pipeline and activity runs.

upvoted 2 times

You have an Azure Data Factory pipeline named pipeline1 that contains a data flow activity named activity1.

You need to run pipeline1.

Which runtime will be used to run activity1?

- A. Azure Integration runtime
- B. Self-hosted integration runtime
- C. SSIS integration runtime

Correct Answer: A

Community vote distribution

A (100%)

 **kkk5566** 4 months, 1 week ago

Selected Answer: A

correct

upvoted 4 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: A

Probably the correct answer.

upvoted 2 times

HOTSPOT

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 contains a dedicated SQL pool named SQLPool1 and an Apache Spark pool named sparkpool1. Sparkpool1 contains a DataFrame named pyspark_df.

You need to write the contents of pyspark_df to a table in SQLPool1 by using a PySpark notebook.

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

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")  
  
%%local  
%%spark  
%%sql  
  
val scala_df = spark.sqlContext.sql ("select * from pysparkdftemptable")  
scala_df.write.  
    jdbc  
    saveAsTable  
    synapsesql
```

Answer Area

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")  
  
%%local  
%%spark  
%%sql  
  
val scala_df = spark.sqlContext.sql ("select * from pysparkdftemptable")  
scala_df.write.  
    jdbc  
    saveAsTable  
    synapsesql
```

 **Azure_2023**  7 months ago

Correct

<https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/synapse-spark-sql-pool-import-export?tabs=scala%2Cscala1%2Cscala2%2Cscala3%2Cscala4%2Cscala5>

upvoted 8 times

 **kkk5566**  4 months, 1 week ago

%%spark

&&

df.write.synapsesql

upvoted 2 times

 **alegiordx** 6 months ago

Correct, also according to this link

https://microsoftlearning.github.io/DP-203-Data-Engineer/Instructions/Labs/LAB_04_data_warehouse_using_apache_spark.html

%%spark

// Make sure the name of the dedicated SQL pool (SQLPool01 below) matches the name of your SQL pool.

```
val df = spark.sqlContext.sql("select * from top_purchases")
df.write.synapsesql("SQLPool01.wwi.TopPurchases", Constants.INTERNAL)
```

upvoted 2 times

You have an Azure data factory named ADF1 and an Azure Synapse Analytics workspace that contains a pipeline named SynPipeline1. SynPipeline1 includes a Notebook activity.

You create a pipeline in ADF1 named ADPPipeline1.

You need to invoke SynPipeline1 from ADPPipeline1.

Which type of activity should you use?

- A. Web
- B. Spark
- C. Custom
- D. Notebook

Correct Answer: A

Community vote distribution

A (100%)

 **ludaka** Highly Voted 6 months, 3 weeks ago

Selected Answer: A

Web Activity
<https://learn.microsoft.com/en-us/azure/data-factory/solution-template-synapse-notebook>
upvoted 7 times

 **auwia** Highly Voted 6 months, 2 weeks ago

Selected Answer: A

To invoke a Synapse pipeline from a Data Factory pipeline, you should use a Web activity.
upvoted 5 times

 **jsav1** Most Recent 6 days, 11 hours ago

Selected Answer: A

Correct
upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: A

is correct
upvoted 1 times

 **Mani_V** 6 months, 2 weeks ago

its a notebook activity
<https://learn.microsoft.com/en-us/azure/data-factory/solution-template-synapse-notebook>
upvoted 1 times

 **vctrhugo** 6 months, 2 weeks ago

Selected Answer: A

Web calls a Synapse pipeline with a notebook activity.
upvoted 2 times

 **sandpl203** 7 months ago

Selected Answer: A

Web Activity as per this article.

<https://fnuson.medium.com/invoke-synapse-notebook-spark-job-by-azure-data-factory-adf-fc19cef89bdd>
upvoted 4 times

HOTSPOT

You have an Azure data factory that contains the linked service shown in the following exhibit.

Edit linked service Azure SQL Database [Learn more](#)

i To avoid publishing immediately to Data Factory, please use Azure Key Vault to retrieve secrets securely. Learn more [here](#)

Name *

AzureSqlDatabase1

Description

Connect via integration runtime * ①

AutoResolveIntegrationRuntime

[Connection string](#)[Azure Key Vault](#)

Account selection method ①

 From Azure subscription Enter manually

Fully qualified domain name *

ssio2022.database.windows.net

Database name *

Contoso

Authentication type *

SQL authentication

User name *

SQLAdmin

[Password](#)[Azure Key Vault](#)

Password *

.....

Always encrypted ①

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 answer is worth one point.

Answer Area

When working in a feature branch, changes to the linked service will be published to the live service

upon publishing the changes
upon saving the changes
when the changes are merged into the collaboration branch

A Copy activity that uses the linked service as the source will perform the Copy activity

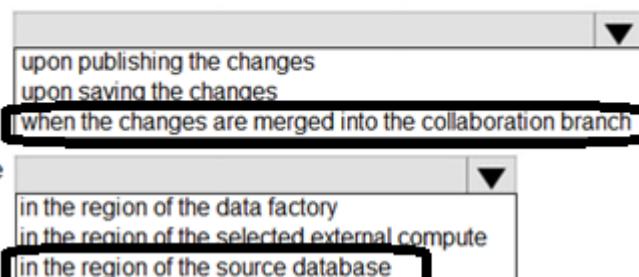
in the region of the data factory
in the region of the selected external compute
in the region of the source database

Answer Area

When working in a feature branch, changes to the linked service will be published to the live service

Correct Answer:

A Copy activity that uses the linked service as the source will perform the Copy activity



peches Highly Voted 7 months ago

According to Microsoft, AutoResolveIntegrationRuntime will attempt to use the sink location to get an IR in the same region (or the closest available) to execute the Copy activity, not the source location. I would go with the region of data factory, since that is the default option when the sink's location is not detectable. Source: <https://learn.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime#azure-ir-location>

upvoted 13 times

matiandal 2 months, 1 week ago

"the sink's location is not detectable" is any wording in the Q that confirms ?

If not, no confirmation about the undetectable sink source, the correct answer is the selected from ET ("in the region of the source database").

r: <https://learn.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime#azure-ir-location> - 1st bullet

upvoted 1 times

Galvanir Highly Voted 5 months, 2 weeks ago

the first one should be "upon publishing changes to the service". See <https://learn.microsoft.com/en-us/azure/data-factory/source-control>

upvoted 5 times

msb Most Recent 3 months, 2 weeks ago

For Linked Services, changes are published immediately unless you use key vault, which means basically upon saving.

"Changes to Linked Services are published immediately, unless you use Key Vault. This can mean a branch change to a Linked Service could impact other branch tests. The reason for this is credential protection. While "Live-Mode" retrieves definitions from the back-end, "Github" or "Devops" mode construct the definitions from the repository. Putting credentials into repository code is a very bad idea. This is why, without Key Vault, credentials are stored and encrypted in Data Factory back end."

<https://learn.microsoft.com/en-us/answers/questions/568057/advanced-feature-branch-development>

upvoted 2 times

EliteAllen 4 months ago

1. upon publishing changes to the service
2. in the region of data factory

upvoted 3 times

kkk5566 4 months, 1 week ago

upon publishing changes to the service
upvoted 1 times

abradabra200 6 months ago

Shouldn't we choose the 'upon saving the changes' option in the first dropdown?

Link: <https://learn.microsoft.com/en-us/azure/data-factory/source-control#stale-publish-branch>

upvoted 4 times

andjurovicela 6 months, 3 weeks ago

I did not manage to find a clear answer to this one, but based on cross-reading a few articles, I think "in the region of data factory" should be the correct answer, and this article explains it a bit better than the others I found: <https://asankap.wordpress.com/2021/10/26/why-you-shouldnt-use-auto-resolve-integration-runtime-in-azure-data-factory-or-synapse/>

upvoted 3 times

JG1984 6 months, 3 weeks ago

When using the AutoResolveIntegrationRuntime with a Copy activity in Azure Data Factory that uses a linked service as the source, the copy operation will be performed in the region of the source data store.

The AutoResolveIntegrationRuntime is a system-assigned integration runtime that automatically routes data movement and activity dispatch to the optimal region based on the location of the source and sink data stores. When using a linked service as the source, the service will attempt to detect the location of the source data store and use an Integration Runtime in the same region to perform the copy operation.

upvoted 1 times

vctrhugo 6 months, 3 weeks ago

For copy activity, a best effort is made to automatically detect your sink data store's location, then use the IR in either the same region, if available, or the closest one in the same geography, otherwise; if the sink data store's region is not detectable, the IR in the instance's region is used instead.

upvoted 1 times

mehroosali 7 months ago

correct

upvoted 1 times

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

HOTSPOT

In Azure Data Factory, you have a schedule trigger that is scheduled in Pacific Time.

Pacific Time observes daylight saving time.

The trigger has the following JSON file.

```
{  
    "name": "Trigger 1",  
    "properties": {  
        "annotations": [],  
        "runtimeState": "Started",  
        "pipelines": [],  
        "type": "ScheduleTrigger",  
        "typeProperties": {  
            "recurrence": {  
                "frequency": "Week",  
                "interval": 1,  
                "startTime": "2022-08-05T04:00:00",  
                "timeZone": "Pacific Standard Time",  
                "schedule": {  
                    "minutes": [  
                        0  
                    ],  
                    "hours": [  
                        3,  
                        21  
                    ],  
                    "weekDays": [  
                        "Sunday",  
                        "Saturday"  
                    ]  
                }  
            }  
        }  
    }  
}
```

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

NOTE: Each correct selection is worth one point.

Answer Area

The trigger will execute [answer choice] on Sunday, March 3, 2024.

| |
|------------|
| ▼ |
| one time |
| two times |
| zero times |

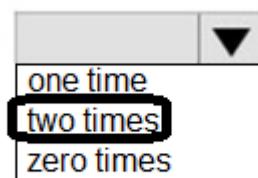
The trigger [answer choice] daylight saving time.

| |
|--------------------------------|
| ▼ |
| is unaffected by |
| will automatically adjust for |
| will require an adjustment for |

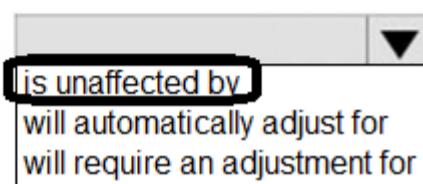
Answer Area

The trigger will execute [answer choice] on Sunday, March 3, 2024.

Correct Answer:



The trigger [answer choice] daylight saving time.



✉ **ludaka** Highly Voted 6 months, 3 weeks ago

1. two times
2. will automatically adjust

"For time zones that observe daylight saving, trigger time will auto-adjust for the twice a year change, if the recurrence is set to Days or above. To opt out of the daylight saving change, please select a time zone that does not observe daylight saving, for instance UTC."

<https://learn.microsoft.com/en-us/azure/data-factory/how-to-create-schedule-trigger?tabs=data-factory#azure-data-factory-and-synapse-portal-experience>

upvoted 17 times

✉ **mrplmcc** 1 month, 1 week ago

Why is it two times instead of one time, if it will automatically adjust?

upvoted 5 times

✉ **JezWalters** Most Recent 5 months, 1 week ago

There's a catch here, as daylight savings actually starts on the SECOND Sunday of March, and March 3 2024 is before this date.

upvoted 2 times

✉ **vctrhugo** 6 months, 3 weeks ago

"[...] we are also adding support for Daylight Saving auto-adjustment: for time zones that observe Daylight Saving, auto change schedule trigger time twice a year (e.g. 8AM daily trigger will fire at 8AM, whether it's PST or PDT)"

<https://techcommunity.microsoft.com/t5/azure-data-factory-blog/time-zone-and-daylight-saving-support-for-schedule-trigger/ba-p/1840199>

upvoted 3 times

✉ **iVath** 6 months, 3 weeks ago

2nd answer should be : will require an adjustment for

ref to : <https://learn.microsoft.com/en-us/azure/data-factory/how-to-create-schedule-trigger?tabs=data-factory>

The timeZone element specifies the time zone that the trigger is created in. This setting affects both startTime and endTime.

upvoted 2 times

✉ **peches** 7 months ago

Agree, as of 2020 ADF supports auto-adjustment for Daylight Saving in Schedule Triggers for time zones that aren't UTC. Since here we are using Pacific time, answer seems correct. Source: <https://techcommunity.microsoft.com/t5/azure-data-factory-blog/time-zone-and-daylight-saving-support-for-schedule-trigger/ba-p/1840199>

upvoted 4 times

✉ **wendyy** 7 months ago

Azure Data Factory only supports time zones UTC. I think it should require the adjustment.

upvoted 1 times

✉ **vctrhugo** 6 months, 3 weeks ago

Incorrect. As of 2020 you can create schedule triggers in your local time zone, without the need to convert timestamps to Coordinated Universal Time (UTC) first.

upvoted 2 times

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a pipeline that will execute a stored procedure in the dedicated SQL pool and use the returned result set as the input for a downstream activity. The solution must minimize development effort.

Which type of activity should you use in the pipeline?

- A. U-SQL
- B. Stored Procedure
- C. Script
- D. Notebook

Correct Answer: B

Community vote distribution

C (73%) B (27%)

✉  **ludaka** Highly Voted 6 months, 3 weeks ago

Selected Answer: C

For me the correct answer is C.

The store procedure activity doesn't return any data.

In the description of the script activity is written that it can be used for : "Run stored procedures. If the SQL statement invokes a stored procedure that returns results from a temporary table, use the WITH RESULT SETS option to define metadata for the result set."

<https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-script>

upvoted 8 times

✉  **andjurovicela** 6 months, 3 weeks ago

I also think this one is correct. One of the things script activity can do is "...Save the rowset returned from a query as activity output for downstream consumption." which is pretty much what is needed here. This is not viable with 'execute SP' activity as it doesn't cannot return any data.

upvoted 2 times

✉  **auwia** Highly Voted 6 months, 2 weeks ago

Selected Answer: C

<https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-script>

The script may contain either a single SQL statement or multiple SQL statements that run sequentially. You can use the Script task for the following purposes:

Truncate a table in preparation for inserting data.

Create, alter, and drop database objects such as tables and views.

Re-create fact and dimension tables before loading data into them.

Run stored procedures. If the SQL statement invokes a stored procedure that returns results from a temporary table, use the WITH RESULT SETS option to define metadata for the result set.

Save the rowset returned from a query as activity output for downstream consumption.

upvoted 6 times

✉  **jongert** Most Recent 1 week ago

Selected Answer: C

The key is 'Output Result set support' which the stored procedure activity does not have. Therefore we have to use a script which supports running stored procedures.

<https://techcommunity.microsoft.com/t5/azure-data-factory-blog/execute-sql-statements-using-the-new-script-activity-in-azure/ba-p/3239969>

upvoted 1 times

✉  **kkk556** 4 months, 1 week ago

Selected Answer: C

C is corret

upvoted 1 times

✉  **CoinUmbrella** 6 months, 1 week ago

Selected Answer: B

B. Chat GPT says the given answer is correct. Stored Procedure is specifically designed to execute stored procedures within Azure Synapse Analytics and is the most suitable option for the scenario, minimizing development effort.

upvoted 1 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: B

The "Script" activity in Azure Data Factory is primarily used to run HDInsight scripts such as Hive, Pig, MapReduce, and Spark. These are typically used for big data processing tasks.

upvoted 1 times

 **auwia** 6 months, 2 weeks ago

False, finally I've found the link, it's C:

<https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-script>

upvoted 4 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: B

"In Azure Synapse Analytics, you can use the SQL pool Stored Procedure Activity to invoke a stored procedure in a dedicated SQL pool."

<https://learn.microsoft.com/en-us/azure/synapse-analytics/data-integration/sql-pool-stored-procedure-activity>

upvoted 2 times

 **sandpl203** 7 months ago

Selected Answer: B

<https://learn.microsoft.com/en-us/azure/synapse-analytics/data-integration/sql-pool-stored-procedure-activity>

upvoted 2 times

You have an Azure SQL database named DB1 and an Azure Data Factory data pipeline named pipeline1.

From Data Factory, you configure a linked service to DB1.

In DB1, you create a stored procedure named SP1. SP1 returns a single row of data that has four columns.

You need to add an activity to pipeline1 to execute SP1. The solution must ensure that the values in the columns are stored as pipeline variables.

Which two types of activities can you use to execute SP1? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Script
- B. Copy
- C. Lookup
- D. Stored Procedure

Correct Answer: AD

Community vote distribution

AC (73%)

AD (27%)

□ **Ram9198** Highly Voted 6 months, 1 week ago

Selected Answer: AC

<https://learn.microsoft.com/en-us/answers/questions/925742/how-to-process-output-from-stored-procedure-in-azu>

SP Activity does not capture result.. use lookup instead of script
upvoted 6 times

□ **andie123** Highly Voted 5 months, 2 weeks ago

why not CD?

A Lookup activity can be used to execute a query or stored procedure against a data source and retrieve a single row of data. The returned values can then be stored as pipeline variables and used in subsequent activities.

A Stored Procedure activity can be used to directly execute a stored procedure against a data source. The returned values can be captured as output parameters and stored as pipeline variables for use in subsequent activities.

upvoted 5 times

□ **Ram9198** Most Recent 4 months ago

Selected Answer: AC

Script lookup
upvoted 1 times

□ **Mal2002** 4 months, 1 week ago

Here is an example of how we can use the Script activity to execute SP1:

```
Script activity (name: "ExecuteSP1Script")
{
    ScriptSource = "<![CDATA[
        var results = SqlCommand('EXEC SP1', connection);
        var myVar = results[0];
    ]]>"
```

In this example, the ScriptSource property specifies the script that is used to execute SP1. The script first executes the SQL statement EXEC SP1. The script then stores the results of SP1 in the variable myVar.

Correct Answers are: A & D

upvoted 1 times

□ **kkk5566** 4 months, 1 week ago

Selected Answer: AC

is correct

upvoted 1 times

□ **Ram9198** 6 months, 1 week ago

sorry, Answer - AC

upvoted 1 times

□ **Ram9198** 6 months, 1 week ago

Answer is wrong - <https://learn.microsoft.com/en-us/answers/questions/925742/how-to-process-output-from-stored-procedure-in-azu>

Answer is CD. SP act cannot emit result..

upvoted 2 times

□ **Mani_V** 6 months, 2 weeks ago

CD is the rite answer

upvoted 1 times

□ **vctrhugo** 6 months, 3 weeks ago

Selected Answer: AD

<https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-stored-procedure>

<https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-script>

upvoted 3 times

□ **WayOps** 6 months, 4 weeks ago

- C. Lookup
- D. Stored Procedure

Explanation:

C. Lookup: The Lookup activity is used to retrieve a dataset from a data source and the output can be used in subsequent activities. It is often used to fetch a small amount of data to be used as parameters in other activities. In this case, it can be used to execute the stored procedure and capture the result into pipeline variables.

D. Stored Procedure: The Stored Procedure activity is used specifically to execute stored procedures. You can capture the output of the stored procedure and assign it to pipeline variables. This activity is designed specifically for executing stored procedures, making it a direct option for this requirement.

A. Script: There is no "Script" activity in Azure Data Factory.

B. Copy: The Copy activity is primarily used for copying data from a source to a destination and is not suitable for executing a stored procedure and capturing its output into pipeline variables.

upvoted 3 times

□ **vctrhugo** 6 months, 3 weeks ago

There is Script activity in ADF.

"The script may contain either a single SQL statement or multiple SQL statements that run sequentially. You can use the Script task for the following purposes:

[...]

Run stored procedures. [...]"

<https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-script>

upvoted 4 times

□ **mehroosali** 6 months, 4 weeks ago

I think the correct answer is C and D.

upvoted 1 times

□ **vctrhugo** 6 months, 3 weeks ago

You use ~~lookup~~ to consume, not to get.

upvoted 1 times

You have an Azure data factory named ADF1.

You currently publish all pipeline authoring changes directly to ADF1.

You need to implement version control for the changes made to pipeline artifacts. The solution must ensure that you can apply version control to the resources currently defined in the Azure Data Factory Studio for ADF1.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each ~~correct~~ selection is worth one point.

- A. From the Azure Data Factory Studio, run Publish All.
- B. Create an Azure Data Factory trigger.
- C. Create a Git repository.
- D. Create a GitHub action.
- E. From the Azure Data Factory Studio, select Set up code repository.
- F. From the Azure Data Factory Studio, select Publish.

Correct Answer: CE

Community vote distribution

CE (100%)

✉  **Paulkuzzio** Highly Voted 4 months, 2 weeks ago

Answer is highly correct. I know this one for sure.
upvoted 8 times

✉  **vernellen** Most Recent 1 week ago

Selected Answer: CE

Version control = always github
upvoted 1 times

You have an Azure data factory named ADF1 that contains a pipeline named Pipeline1.

Pipeline1 must execute every 30 minutes with a 15-minute offset.

You need to create a trigger for Pipeline1. The trigger must meet the following requirements:

- Backfill data from the beginning of the day to the current time.
- If Pipeline1 fails, ensure that the pipeline can re-execute within the same 30-minute period.
- Ensure that only one concurrent pipeline execution can occur.
- Minimize development and configuration effort.

Which type of trigger should you create?

- A. schedule
- B. event-based
- C. manual
- D. tumbling window

Correct Answer: D

Community vote distribution

D (67%)

C (33%)

 **jasmd2** 1 week, 3 days ago

Selected Answer: D

Tumbling window because of the backfill
upvoted 1 times

 **Lucasmh** 4 weeks ago

Selected Answer: C

data is expected to arrive at regular intervals, and you want to trigger a pipeline with a fixed window size.

For the specific requirements mentioned, especially the need to execute every 30 minutes with a 15-minute offset and backfill data from the beginning of the day to the current time, a schedule trigger is more suitable. The tumbling window trigger is generally used for scenarios where you want to process data in fixed windows based on its arrival time.

upvoted 1 times

 **msb** 3 months, 2 weeks ago

Tumbling window is correct.

Backfill scenario: <https://learn.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger?tabs=data-factory%2Cazure-powershell#execution-order-of-windows-in-a-backfill-scenario>

offset, concurrency, ... :

<https://learn.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger?tabs=data-factory%2Cazure-powershell#tumbling-window-trigger-type-properties>

upvoted 3 times

 **DataEngDP** 4 months ago

schedule because it occurs every 30 minutes

upvoted 2 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: D

seem to correct

upvoted 1 times

 **Paulkuzzio** 4 months, 2 weeks ago

It seems correct but the 15mins offset is throwing me off. Somebody please explain. Thanks

upvoted 4 times

 **Lewiasskick** 1 day, 13 hours ago

offset refers to the delay of start of the trigger : A timespan value that must be negative in a self-dependency. If no value specified, the window is the same as the trigger itself.

upvoted 1 times

 **g2000** 2 months, 2 weeks ago

Ensure that only one concurrent pipeline execution can occur <--- this suggests the tumbling window
upvoted 2 times

Question #116

Topic 2

You have an Azure Data Lake Storage Gen2 account named account1 and an Azure event hub named Hub1. Data is written to account1 by using Event Hubs Capture.

You plan to query account by using an Apache Spark pool in Azure Synapse Analytics.

You need to create a notebook and ingest the data from account1. The solution must meet the following requirements:

- Retrieve multiple rows of records in their entirety.
- Minimize query execution time.
- Minimize data processing.

Which data format should you use?

- A. Parquet -
- B. Avro
- C. ORC
- D. JSON

Correct Answer: A

 **jongert** 1 week ago

Answer is B, here is why:

Avro and Parquet are both binary compressed file formats which makes them preferred over CSV which stores the data as strings (much larger files).

Now the difference between Parquet and Avro is the format, as Parquet is column based while Avro provides row based store. Since the requirement is to retrieve rows in their entirety, it is better to use Avro. Scenarios where we only retrieve a subset of columns for analysis would favour the use of Parquet.

upvoted 1 times

 **matiandal** 2 months, 1 week ago

Parquet showed either similar or better results on every test [than Avro]. The query-performance differences on the larger datasets in Parquet's favor are partly due to the compression results; when querying the wide dataset, Spark had to read 3.5x less data for Parquet than Avro. Avro did not perform well when processing the entire dataset, as suspected."

R: <https://blog.cloudera.com/benchmarking-apache-parquet-the-allstate-experience/>

upvoted 3 times

You have an Azure Blob Storage account named blob1 and an Azure Data Factory pipeline named pipeline1.

You need to ensure that pipeline1 runs when a file is deleted from a container in blob1. The solution must minimize development effort.

Which type of trigger should you use?

- A. schedule
- B. storage event
- C. tumbling window
- D. custom event

Correct Answer: B

Community vote distribution

B (100%)

 **vernille** 1 month, 4 weeks ago

Selected Answer: B

"A Storage Event trigger in Azure Data Factory is designed to initiate a pipeline in response to an event happening in Azure Blob Storage, such as the deletion of a file"

So B.

upvoted 3 times

HOTSPOT

You have Azure Data Factory configured with Azure Repos Git integration. The collaboration branch and the publish branch are set to the default values.

You have a pipeline named pipeline1.

You build a new version of pipeline1 in a branch named feature1.

From the Data Factory Studio, you select Publish.

The source code of which branch will be built, and which branch will contain the output of the Azure Resource Manager (ARM) template? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Source code:

A dropdown menu containing three items: "adf_publish", "feature1", and "main".

ARM template output:

A dropdown menu containing three items: "adf_publish", "feature1", and "main".

Answer Area

Source code:

A dropdown menu containing three items: "adf_publish", "feature1", and "main". The "main" option is highlighted with a black rectangle.

Correct Answer:

ARM template output:

A dropdown menu containing three items: "adf publish", "feature1", and "main". The "adf publish" option is highlighted with a black rectangle.

by **jonpert** 1 week ago

Correct, publish branch contains only ADF related code in JSON format. All the source code can be found in the collaboration branch which is by default the main branch.

<https://learn.microsoft.com/en-us/azure/data-factory/source-control>
upvoted 1 times

by **y154707** 2 months ago

Seems correct, as there was not a PR of the changes made in the feature branch. Thus, if you "Publish" the built will be based on what is in the collab branch, not in the feature branch.

upvoted 2 times

DRAG DROP

You have an Azure subscription that contains an Azure data factory.

You are editing an Azure Data Factory activity JSON.

The script needs to copy a file from Azure Blob Storage to multiple destinations. The solution must ensure that the source and destination files have consistent folder paths.

How should you complete the script? 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.

Values**Answer Area**

| | | |
|-------------------|--------------------------|--------------------------|
| FlattenHierarchy | { | "name": "Pipeline1", |
| ForEach | "properties": { | "activities": [|
| MergeFiles | "activities": [| { |
| PreserveHierarchy | "name": "Activity1", | "activity": "Activity1", |
| Switch | "type": [| "activity": "Activity1", |
| Until | "activity": "Activity1", | "activity": "Activity1", |

```

    "activities": [
        {
            "name": "Activity1",
            "type": [ "activity": "Activity1" ],
            "typeProperties": {
                "isSequential": "true",
                "items": [
                    {
                        "value": "@pipeline
() .parameters.mySinkDatasetFolderPath",
                        "type": "Expression"
                    }
                ],
                "activities": [
                    {
                        "name": "MyCopyActivity",
                        "type": "Copy",
                        "typeProperties": {
                            "source": {
                                "type": "BlobSource",
                                "recursive": "false" },
                            "sink": {
                                "type": "BlobSink",
                                "CopyBehavior": [ ]
                            }
                        }
                    }
                ]
            }
        }
    ]
}
```

Answer Area**Correct Answer:**

```

{
    "name": "Pipeline1",
    "properties": {
        "activities": [
            {
                "name": "Activity1",
                "type": "ForEach",
                "typeProperties": {
                    "isSequential": "true",
                    "items": [
                        {
                            "value": "@pipeline
() .parameters.mySinkDatasetFolderPath",
                            "type": "Expression"
                        }
                    ],
                    "activities": [
                        {
                            "name": "MyCopyActivity",
                            "type": "Copy",
                            "typeProperties": {
                                "source": {
                                    "type": "BlobSource",
                                    "recursive": "false" },
                                "sink": {
                                    "type": "BlobSink",
                                    "CopyBehavior": "PreserveHierarchy"
                                }
                            }
                        }
                    ]
                }
            }
        ]
    }
}
```

□  **Lewiasskick** 1 day, 13 hours ago

Correct, refer to <https://learn.microsoft.com/en-us/azure/data-factory/control-flow-for-each-activity>

upvoted 2 times

Question #120

Topic 2

You are building a data flow in Azure Data Factory that upserts data into a table in an Azure Synapse Analytics dedicated SQL pool.

You need to add a transformation to the data flow. The transformation must specify logic indicating when a row from the input data must be upserted into the sink.

Which type of transformation should you add to the data flow?

- A. join
- B. alter row
- C. surrogate key
- D. select

Correct Answer: B

□  **warre** 1 day, 14 hours ago

correct, chatGPT:

For upsert operations (insert or update), you typically need to determine whether a record already exists in the destination table based on some condition. In Azure Data Factory's data flow, you would use the "Alter Row" transformation for this purpose.

upvoted 1 times

You have an on-premises database named db1 and a self-hosted integration runtime.

You have an Azure subscription that contains an Azure Data Lake Storage account named dl1.

You need to develop four data pipeline projects that will use Microsoft Power Query to copy data from db1 to dl1. The solution must meet the following requirements:

- All pipelines must use the self-hosted integration runtime.
- Each project must be stored in a separate Git repository.
- Development effort must be minimized.

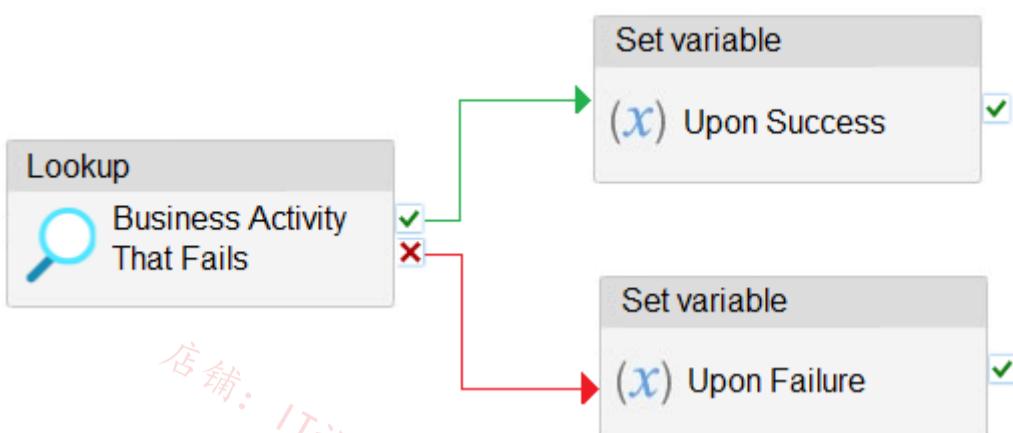
What should you use?

- A. Azure Synapse Analytics
- B. Azure Logic Apps.
- C. Azure Data Factory
- D. Microsoft Power BI

Correct Answer: C

 warre 1 day, 14 hours ago
correct, chat gpt confirmed
upvoted 1 times

You have the Azure Synapse Analytics pipeline shown in the following exhibit.



You need to add a set variable activity to the pipeline to ensure that after the pipeline's completion, the status of the pipeline is always successful.

What should you configure for the set variable activity?

- A. a skipped dependency on the Upon Failure activity
- B. a skipped dependency on the Upon Success activity
- C. a success dependency on the Business Activity That Fails activity
- D. a failure dependency on the Upon Failure activity

Correct Answer: A

Community vote distribution

B (100%)

⊕ **jongert** 3 days, 19 hours ago

Selected Answer: B

Should be B, creating an 'Do If Skip Else' statement by adding a skipped dependency to the success path means that the pipeline is always succeeded.

upvoted 1 times

You have an on-premises Linux server that contains a database named DB1.

You have an Azure subscription that contains an Azure data factory named ADF1 and an Azure Data Lake Storage account named ADLS1.

You need to create a pipeline in ADF1 that will copy data from DB1 to ADLS1.

Which type of integration runtime should you use to read the data from DB1?

- A. self-hosted integration runtime
- B. Azure integration runtime
- C. Azure-SQL Server Integration Services (SSIS)

Correct Answer: A

 **JJFortunato** 18 hours, 57 minutes ago

Correct

upvoted 1 times

DRAG DROP

You have an Azure Data Lake Storage account named account1.

You use an Azure Synapse Analytics serverless SQL pool to access sales data stored in account1.

You need to create a bar chart that displays sales by product. The solution must minimize development effort.

In which order should you perform the actions? To answer, move all actions from the list of actions to the answer area and arrange them in the correct order.

Actions**Answer Area**

Add a `SELECT` statement that will return the sales by product data.

Switch to the Chart view.

Modify the Chart settings.

Create a SQL script by using Synapse Studio.

Execute the script.

**Answer Area**

Create a SQL script by using Synapse Studio.

Add a `SELECT` statement that will return the sales by product data.

Correct Answer:

Execute the script.

Switch to the Chart view.

Modify the Chart settings.

DRAG DROP

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a copy of the data warehouse and make the copy available for 28 days. The solution must minimize costs.

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.

| Actions | Answer Area |
|---|---|
| Create a new user-defined restore point. | |
| Restore the latest automatic restore point to a new data warehouse. | |
| Pause the restored data warehouse.
 |  |
| Restore the copy from the latest automatic restore point to the current data warehouse. | |
| Restore the copy from the new user-defined restore point to a new data warehouse. | |

| Answer Area |
|---|
| Create a new user-defined restore point. |
| Correct Answer: Restore the copy from the new user-defined restore point to a new data warehouse. |
| Pause the restored data warehouse. |

HOTSPOT

You have an Azure Synapse Analytics workspace that contains an Apache Spark pool named Pool1.

You need to read data from a CSV file and write the data to a Delta table by using Pool1.

How should you complete the PySpark code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

```
from delta.tables import *
from pyspark.sql.functions import *
df = spark.read.load
('abfss://container@mydatalake.dfs.core.windows.net/stage/
products.csv', format = 'csv', header = True)
delta_table_path = "/delta/products-delta"
df. ▼ .save(delta_table_path)

cache()
inputFiles()
write.format("delta")
write.parquet

deltaTable = ▼ (spark, delta_table_path)
deltaTable.alias
deltaTable.convertToDelta
deltaTable.forPath
deltaTable.update
```

Answer Area

```
from delta.tables import *
from pyspark.sql.functions import *
df = spark.read.load
('abfss://container@mydatalake.dfs.core.windows.net/stage/
products.csv', format = 'csv', header = True)
delta_table_path = "/delta/products-delta"
df. ▼ .save(delta_table_path)

cache()
inputFiles()
write.format("delta")
write.parquet

deltaTable = ▼ (spark, delta_table_path)
deltaTable.alias
deltaTable.convertToDelta
deltaTable.forPath
deltaTable.update
```

Correct Answer:

HOTSPOT

You have an Azure Data Lake Storage account that contains one CSV file per hour for January 1, 2020, through January 31, 2023. The files are partitioned by using the following folder structure.

```
csv/system1/{year}/{month}/{filename}.csv
```

You need to query the files by using an Azure Synapse Analytics serverless SQL pool. The solution must return the row count of each file created during the last three months of 2022.

How should you complete the query? To answer, select the appropriate options in the answer area.

Answer Area

```

SELECT
    r.filepath() AS filepath
    ,COUNT_BIG(*) AS [rows]
FROM OPENROWSET(
    BULK
        'csv/system1/2022',
        'csv/system1/2022/',
        'csv/system1/2022/*/*.csv',
    DATA_SOURCE = 'MyDataLake',
    FORMAT = 'CSV',
    PARSER_VERSION = '2.0',
    FIRSTROW = 2 )
WITH (vendor_id INT) AS [r]
WHERE
    r.filepath()
    r.filepath(1)
    r.filepath(2)
GROUP BY

```

Answer Area

```

SELECT
    r.filepath() AS filepath
    ,COUNT_BIG(*) AS [rows]
FROM OPENROWSET(
    BULK
        'csv/system1/2022',
        'csv/system1/2022/',
        'csv/system1/2022/*/*.csv',
    DATA_SOURCE = 'MyDataLake',
    FORMAT = 'CSV',
    PARSER_VERSION = '2.0',
    FIRSTROW = 2 )
WITH (vendor_id INT) AS [r]
WHERE
    r.filepath()
    r.filepath(1)
    r.filepath(2)
GROUP BY

```

Correct Answer:

```

DATA_SOURCE = 'MyDataLake',
FORMAT = 'CSV',
PARSER_VERSION = '2.0',
FIRSTROW = 2 )
WITH (vendor_id INT) AS [r]
WHERE

```

Lewiasskick 1 day, 12 hours ago

correct: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/query-specific-files>

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains an external table named Sales. Sales contains sales data. Each row in Sales contain data on a single sale, including the name of the salesperson.

You need to implement row-level security (RLS). The solution must ensure that the salespeople can access only their respective sales.

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

NOTE: Each correct selection is worth one point.

Answer Area

Create:

- A materialized view in Pool1
- A security policy for Sales
- Database scoped credentials in Pool1

Restrict row access by using:

- A masking rule
- A table-valued function
- The CONTAINS predicate

Answer Area

Correct Answer:

Create:

- A materialized view in Pool1
- A security policy for Sales**
- Database scoped credentials in Pool1

Restrict row access by using:

- A masking rule
- A table-valued function**
- The CONTAINS predicate

You have an Azure Data Factory pipeline named P1.

You need to schedule P1 to run at 10:15 AM, 12:15 PM, 2:15 PM, and 4:15 PM every day.

Which frequency and interval should you configure for the scheduled trigger?

A. Frequency: Month -

Interval: 1

B. Frequency: Day -

Interval: 1

C. Frequency: Minute -

Interval: 60

D. Frequency: Hour -

Interval: 2

Correct Answer: D

Topic 3 - Question Set 3

DRAG DROP -

You have an Azure Active Directory (Azure AD) tenant that contains a security group named Group1. You have an Azure Synapse Analytics dedicated SQL pool named dw1 that contains a schema named schema1.

You need to grant Group1 read-only permissions to all the tables and views in schema1. The solution must use the principle of least privilege. 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:

| Actions | Answer Area |
|--|-------------|
| Create a database role named Role1 and grant Role1 SELECT permissions to schema1. | |
| Create a database role named Role1 and grant Role1 SELECT permissions to dw1. | |
| Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1. | |
| Create a database user in dw1 that represents Group1 and uses the FROM EXTERNAL PROVIDER clause. | |
| Assign Role1 to the Group1 database user. | |

Correct Answer:

| Actions | Answer Area |
|--|--|
| Create a database role named Role1 and grant Role1 SELECT permissions to schema1. | Create a database user in dw1 that represents Group1 and uses the FROM EXTERNAL PROVIDER clause. |
| Create a database role named Role1 and grant Role1 SELECT permissions to dw1. | Create a database role named Role1 and grant Role1 SELECT permissions to schema1. |
| Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1. | Assign Role1 to the Group1 database user. |
| Create a database user in dw1 that represents Group1 and uses the FROM EXTERNAL PROVIDER clause. | |
| Assign Role1 to the Group1 database user. | |

Step 1: Create a database user named dw1 that represents Group1 and use the FROM EXTERNAL PROVIDER clause.

Step 2: Create a database role named Role1 and grant Role1 SELECT permissions to schema1.

Step 3: Assign Role1 to the Group1 database user.

Reference:

<https://docs.microsoft.com/en-us/azure/data-share/how-to-share-from-sql>

Rob77 Highly Voted 2 years, 7 months ago

1. create user from external provider for Group1
2. create Role1 with select on schema1
3. add user to the Role1

upvoted 91 times

SameerL 1 year, 6 months ago

- Step 1: Create a database user named dw1 that represents Group1 and use the FROM EXTERNAL PROVIDER clause.
- Step 2: Create a database role named Role1 and grant Role1 SELECT permissions to schema1.
- Step 3: Assign Role1 to the Group1 database user.

upvoted 8 times

AlexLo 1 year, 11 months ago

Sorry, but "add user to the Role1" is not part of the answers. Or, which option is that?

upvoted 4 times

thomas02 1 year, 10 months ago

Assign Role1 to the Group1 database user
upvoted 4 times

□ **PallaviPatel** 1 year, 11 months ago

add user to the role1 option isn't available in the given choices not sure why this answer is suggested then? what is the need for creating external provider for Group1 can you explain?

upvoted 2 times

□ **Lotusss** 1 year, 8 months ago

UDEMY says this as well. So correct

upvoted 1 times

□ **patricka95** Highly Voted 2 years, 5 months ago

The suggested answer is wrong. As others have identified, the correct steps are;

1. create user <> from external provider
2. create role <> with select permission on schema
3. add user to role

upvoted 10 times

□ **lukeonline** 2 years ago

Can somebody explain why we have to create the user first and not the role?

upvoted 3 times

□ **SQLDev0000** 1 year, 10 months ago

There is a note in the question that says "More than one order of answer choices is correct". Create role and create user can be interchanged.

upvoted 4 times

□ **vanrell** 1 year, 9 months ago

They do mention that: : More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Creating user or role first does not matter. As long as you assign the role to the user in the end.

upvoted 3 times

□ **sachabess79** 2 years, 3 months ago

Agreed 100%

upvoted 3 times

□ **Aditya0891** 1 year, 6 months ago

Answer is not wrong. Read the question properly

upvoted 1 times

□ **kkk5566** Most Recent 4 months, 1 week ago

1. create database user in dw1 that represent Group1 and uses From External Provider clause
2. create database role named Role1 with grant Role1 select permission on dw1
3. add Role1 to Group1 database user

upvoted 2 times

□ **_Lukas_** 5 months, 3 weeks ago

```
CREATE USER security_group_lk FROM EXTERNAL PROVIDER;
CREATE ROLE security_group_role;
GRANT SELECT ON SCHEMA::app TO security_group_role;
ALTER ROLE security_group_role ADD MEMBER security_group_lk;
```

upvoted 1 times

□ **kornat** 9 months, 1 week ago

correct
upvoted 1 times

□ **Deeksha1234** 1 year, 5 months ago

given answer is correct
upvoted 2 times

□ **SabaJamal2010AtGmail** 2 years ago

1. create database user in dw1 that represent Group1 and uses From External Provider clause
2. create database role named Role1 with grant Role1 select permission on dw1
3. add Role1 to Group1 database user

upvoted 6 times

□ **ADHDBA** 1 year, 9 months ago

it should be least privileged so select on schema is correct not on dw1
upvoted 1 times

□ **eng1** 2 years, 6 months ago

It should be D-E-A
upvoted 1 times

✉ eng1 2 years, 6 months ago

Please ignore my previous answer, it should be

D: Create a database user in dw1 that represents Group1 and uses FROM EXTERNAL PROVIDE clause

A: Create a database role named Role1 and grant Role1 SELECT permissions to schema1

E: Assign Rol1 to the Group1 database user

upvoted 18 times

✉ eng1 2 years, 6 months ago

It should be C-A-E

upvoted 1 times

✉ SG1705 2 years, 7 months ago

Is the answer correct ??

upvoted 1 times

✉ Marcello83 2 years, 6 months ago

No, in my opinion it is D, A, E. If you give a reader role to the group, the users will have the possibility to query all the tables, not only the selected schema.

upvoted 6 times

✉ Davico93 1 year, 6 months ago

But, the answer shown in solution es D,A,E....

upvoted 1 times

HOTSPOT -

You have an Azure subscription that contains a logical Microsoft SQL server named Server1. Server1 hosts an Azure Synapse Analytics SQL dedicated pool named Pool1.

You need to recommend a Transparent Data Encryption (TDE) solution for Server1. The solution must meet the following requirements:

☞ Track the usage of encryption keys.

Maintain the access of client apps to Pool1 in the event of an Azure datacenter outage that affects the availability of the encryption keys.

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

To track encryption key usage:

- Always Encrypted
- TDE with customer-managed keys
- TDE with platform-managed keys

To maintain client app access in the event of a datacenter outage:

- Create and configure Azure key vaults in two Azure regions.
- Enable Advanced Data Security on Server1.
- Implement the client apps by using a Microsoft .NET Framework data provider.

Correct Answer:

Answer Area

To track encryption key usage:

- Always Encrypted
- TDE with customer-managed keys
- TDE with platform-managed keys

To maintain client app access in the event of a datacenter outage:

- Create and configure Azure key vaults in two Azure regions.
- Enable Advanced Data Security on Server1.
- Implement the client apps by using a Microsoft .NET Framework data provider.

Box 1: TDE with customer-managed keys

Customer-managed keys are stored in the Azure Key Vault. You can monitor how and when your key vaults are accessed, and by whom. You can do this by enabling logging for Azure Key Vault, which saves information in an Azure storage account that you provide.

Box 2: Create and configure Azure key vaults in two Azure regions

The contents of your key vault are replicated within the region and to a secondary region at least 150 miles away, but within the same geography to maintain high durability of your keys and secrets.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption> <https://docs.microsoft.com/en-us/azure/key-vault/general/logging>

Francesco1985 Highly Voted 2 years, 6 months ago

Guys the answers are correct: <https://docs.microsoft.com/en-us/azure/azure-sql/database/transparent-data-encryption-byok-overview>
upvoted 59 times

Slena 2 years, 3 months ago

Agreed. "Link each server with two key vaults that reside in different regions and hold the same key material, to ensure high availability of encrypted databases. Mark only the key from the key vault in the same region as a TDE protector. System will automatically switch to the key vault in the remote region if there is an outage affecting the key vault in the same region."

<https://docs.microsoft.com/en-us/azure/azure-sql/database/transparent-data-encryption-byok-overview>

upvoted 5 times

bhavesh_wadhwani Highly Voted 2 years, 3 months ago

First answer is correct.
2nd box answer should be " Implement the client apps by using .NET framework data provider" as key vault is by default replicated in two or more regions for HA.
upvoted 7 times

bhavesh_wadhwani 2 years, 3 months ago

Link from Microsoft docs : <https://docs.microsoft.com/en-us/azure/key-vault/general/disaster-recovery-guidance#:~:text=The%20contents%20of%20your%20key%20vault%20are%20replicated%20within%20the%20region%20and%20to%20a%20secondary%20region%20at%20least%20150%20miles%20away%2C%20but%20within%20the%20same%20geography%20to%20maintain%20high%20durability%20of%20your%20keys%20and%20secrets>

upvoted 1 times

kkk5566 Most Recent 4 months, 1 week ago

correct

upvoted 1 times

pavankr 6 months, 1 week ago

why "two" azure regions? the requirement never mentioned how many regions?

upvoted 2 times

Deeksha1234 1 year, 5 months ago

correct answer

upvoted 1 times

SabaJamal2010AtGmail 2 years ago

Both answers Correct 1) Transparent Data Encryption with customer-managed key 2) key vault in 2 regions

upvoted 2 times

Skeinofi 2 years ago

Correct.

Recommendations when configuring customer-managed TDE: Recommendations when configuring AKV:

- Enable auditing and reporting on all encryption keys: Key vault provides logs that are easy to inject into other security information and event management tools. Operations Management Suite Log Analytics is one example of a service that is already integrated.

- Link each server with two key vaults that reside in different regions and hold the same key material, to ensure high availability of encrypted databases. Mark the key from one of the key vaults as the TDE protector. System will automatically switch to the key vault in the second region with the same key material, if there's an outage affecting the key vault in the first region.

upvoted 1 times

kimalto452 2 years, 3 months ago

Transparent Data Encryption with customer-managed key

<https://docs.microsoft.com/en-us/azure/azure-sql/database/transparent-data-encryption-byok-overview>

upvoted 1 times

terajuana 2 years, 7 months ago

TDE doesn't use client managed keys

answer therefore is

1) always encrypted

2) key vault in 2 regions

upvoted 1 times

Alekx42 2 years, 7 months ago

Moreover, always encrypted is NOT TDE option. The question asks to enable TDE.

upvoted 3 times

Reel 1 year, 1 month ago

you need to create key vault separately on two regions and then linked it together

"Even in cases when there's no configured geo-redundancy for server, it's highly recommended to configure the server to use two different

key vaults in two different regions with the same key material."

<https://learn.microsoft.com/en-us/azure/azure-sql/database/transparent-data-encryption-byok-overview?view=azuresql#high-availability-with-customer-managed-tde>

upvoted 1 times

✉ **Alekx42** 2 years, 7 months ago

TDE can be configured with Customer Managed keys:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/transparent-data-encryption-tde-overview?tabs=azure-portal#customer-managed-transparent-data-encryption---bring-your-own-key>

Key vault is configured in multiple regions by microsoft itself. I also double-checked by creating a key vault and there are no geo-redundancy options. Also see here:

<https://docs.microsoft.com/en-us/azure/key-vault/general/disaster-recovery-guidance>

upvoted 5 times

✉ **Alekx42** 2 years, 7 months ago

The first answer is correct. You need to enable TDE with customer keys in order to track the key usage in Azure key vault.

The second answer seems wrong, as pointed out by Rob77. AKV does have replication to 2 additional regions by default. So I guess that it makes more sense to use a Microsoft .NET framework data provider <https://docs.microsoft.com/en-us/dotnet/framework/data/adonet/data-providers>

upvoted 2 times

✉ **terajuana** 2 years, 7 months ago

TDE doesn't operate with customer keys but always encrypted does

upvoted 1 times

✉ **Rob77** 2 years, 7 months ago

second answer does not seem to be correct - AKV is already replicated within the region locally (and also 2 pair regions). Therefore if the datacentre fails (or even whole region) the traffic will be redirected. <https://docs.microsoft.com/en-us/azure/key-vault/general/disaster-recovery-guidance>

upvoted 3 times

✉ **corebit** 2 years ago

"The contents of your key vault are replicated within the region and to a secondary region at least 150 miles away, but within the same geography to maintain high durability of your keys and secrets."

<https://docs.microsoft.com/en-us/azure/key-vault/general/disaster-recovery-guidance>

upvoted 1 times

You plan to create an Azure Synapse Analytics dedicated SQL pool.

You need to minimize the time it takes to identify queries that return confidential information as defined by the company's data privacy regulations and the users who executed the queries.

Which two components should you include in the solution? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. sensitivity-classification labels applied to columns that contain confidential information
- B. resource tags for databases that contain confidential information
- C. audit logs sent to a Log Analytics workspace
- D. dynamic data masking for columns that contain confidential information

Correct Answer: AC

A: You can classify columns manually, as an alternative or in addition to the recommendation-based classification:

| Schema | Table | Column |
|---------|------------------|------------------------|
| SalesLT | Customer | FirstName |
| SalesLT | Customer | LastName |
| SalesLT | Customer | EmailAddress |
| SalesLT | Customer | Phone |
| SalesLT | Customer | PasswordHash |
| SalesLT | ErrorLog | UserName |
| dbo | Address | AddressLine1 |
| SalesLT | Address | AddressLine2 |
| SalesLT | Address | City |
| SalesLT | Address | PostalCode |
| SalesLT | CustomerAddress | AddressType |
| SalesLT | SalesOrderHeader | AccountNumber |
| SalesLT | SalesOrderHeader | CreditCardApprovalCode |
| SalesLT | SalesOrderHeader | TaxAmt |

1. Select Add classification in the top menu of the pane.

2. In the context window that opens, select the schema, table, and column that you want to classify, and the information type and sensitivity label.

3. Select Add classification at the bottom of the context window.

C: An important aspect of the information-protection paradigm is the ability to monitor access to sensitive data. Azure SQL Auditing has been enhanced to include a new field in the audit log called data_sensitivity_information. This field logs the sensitivity classifications (labels) of the data that was returned by a query. Here's an example:

| d | client_ip | application_name | duration_milliseconds | response_rows | affected_rows | connection_id | data_sensitivity_information |
|---|-----------|--|-----------------------|---------------|---------------|------------------|-----------------------------------|
| | 7.125 | Microsoft SQL Server Management Studio - Query | 1 | 847 | 847 | C244A066-2271... | Confidential - GDPR |
| | 7.125 | Microsoft SQL Server Management Studio - Query | 2 | 32 | 32 | C244A066-2271... | Confidential |
| | 7.125 | Microsoft SQL Server Management Studio - Query | 41 | 32 | 32 | A7088FD4-759E... | Confidential, Confidential - GDPR |

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview>

damaldon Highly Voted 2 years, 6 months ago

Correct!

upvoted 27 times

 **saty_nl** Highly Voted 2 years, 6 months ago

Answer is correct. Dynamic data masking will limit the exposure of sensitive data.

upvoted 11 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: AC

audit and sensitivity-classification labels

upvoted 2 times

 **anks84** 1 year, 4 months ago

Selected Answer: AC

Given Answers are correct !

upvoted 4 times

 **Deeksha1234** 1 year, 5 months ago

correct

upvoted 3 times

 **Remedios79** 1 year, 6 months ago

Also for me is correct

upvoted 2 times

 **sparkchu** 1 year, 9 months ago

log auditing & tracing is important for data governance, therefore necessary for any data solution.

upvoted 2 times

 **rashjan** 2 years, 1 month ago

Selected Answer: AC

correct

upvoted 2 times

 **rashjan** 2 years, 1 month ago

Correct: "The solution needs to identify the users who executed queries, not to hide confidential information." thanks @DirectX from this discussion: <https://www.examtopics.com/discussions/microsoft/view/51257-exam-dp-201-topic-3-question-32-discussion/>

upvoted 7 times

 **dduque10** 2 years, 3 months ago

Is it really C correct?

upvoted 1 times

 **rashjan** 2 years, 1 month ago

Yes, the logs are used to identify the user who executed the query.

upvoted 3 times

 **Dizzystar** 2 years, 2 months ago

wondering the same thing.

upvoted 1 times

You are designing an enterprise data warehouse in Azure Synapse Analytics that will contain a table named Customers. Customers will contain credit card information.

You need to recommend a solution to provide salespeople with the ability to view all the entries in Customers. The solution must prevent all the salespeople from viewing or inferring the credit card information.

What should you include in the recommendation?

- A. data masking
- B. Always Encrypted
- C. column-level security
- D. row-level security

Correct Answer: C

Column-level security simplifies the design and coding of security in your application, allowing you to restrict column access to protect sensitive data.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security>

✉ **Alekx42** Highly Voted 2 years, 7 months ago

C is the right answer. Check the discussion here:

<https://www.examtopics.com/discussions/microsoft/view/18788-exam-dp-201-topic-3-question-12-discussion/>
upvoted 32 times

✉ **mikerss** 2 years, 6 months ago

the key word is 'infer'. as listed in the below documentation, data masking is not used to protect against malicious intent to infer the underlying data. I would therefore choose C

upvoted 11 times

✉ **Marcus1612** 2 years, 3 months ago

I agree with mikerss, the key word is 'infer'. Data masking is a kind of column-level security but it is only partial. A malicious person could infer the credit card number. The good answer is C

upvoted 4 times

✉ **Deeksha1234** 1 year, 5 months ago

I agree with the logic provided

upvoted 3 times

✉ **anto69** 1 year, 11 months ago

yeah, from ms docs: "ensuring that specific users can access only certain columns of a table pertinent to their department"
upvoted 3 times

✉ **Tracy_Anderson** 2 years, 5 months ago

The link below show how you can infer a column that is data masked. It is also referenced in the 201 topic, <https://docs.microsoft.com/nl-nl/sql/relational-databases/security/dynamic-data-masking?view=sql-server-ver15>
upvoted 2 times

✉ **FredNo** Highly Voted 2 years, 1 month ago

Selected Answer: C

Data masking does not protect against inferring with the data
upvoted 10 times

✉ **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: C

Data masking does not protect against inferring with the data
upvoted 1 times

✉ **kkk5566** 4 months ago

go to A
upvoted 1 times

✉ **vctrhugo** 6 months, 3 weeks ago

Infer data means:

```
SELECT ID, Name, Salary FROM Employees
WHERE Salary > 99999 and Salary < 100001;
```

```
+-----+-----+-----+
|Id |Name |Salary|
+-----+-----+-----+
|62543 |Jane Doe |0 |
|91245 |John Smith |0 |
+-----+-----+-----+
upvoted 1 times
```

□ **SinSS** 7 months, 3 weeks ago

Only with DDM, you can guess with trying some queries

upvoted 1 times

□ **Okea** 11 months, 2 weeks ago

C is the answer

As an example, consider a database principal that has sufficient privileges to run ad-hoc queries on the database, and tries to 'guess' the underlying data and ultimately infer the actual values. Assume that we have a mask defined on the [Employee].[Salary] column, and this user connects directly to the database and starts guessing values, eventually inferring the [Salary] value of a set of Employees:
<https://learn.microsoft.com/en-us/sql/relational-databases/security/dynamic-data-masking?view=sql-server-ver16#security-note-bypassing-masking-using-inference-or-brute-force-techniques>

upvoted 1 times

□ **anks84** 1 year, 4 months ago

Selected Answer: C

Column level security is the correct answer !!

upvoted 2 times

□ **Deeksha1234** 1 year, 5 months ago

correct

upvoted 1 times

□ **orm33** 1 year, 8 months ago

There is nothing that says that you must use the credit card masking rule, you can use another one. This way, the sales persons has access to all entries but cannot infer the credit card. The answer is A

upvoted 1 times

□ **Aditya0891** 1 year, 7 months ago

data masking will only help in not viewing the credit card information however it won't help in inferring the column so column level security is required. In this way you can view all the rows(entries) without using the credit card column

upvoted 1 times

□ **juanlu46** 1 year, 8 months ago

Selected Answer: C

Column-level security prevent get "credit card" column, you not be able to infer the credit card information contrary to "masking".

upvoted 1 times

□ **GDJ2022** 1 year, 11 months ago

There are 2 parts to it:

1. provide salespeople with the ability to **view** all the entries in Customers.
2. should not be able to infer.

DDM is the only solution if you have to comply with both requirements

upvoted 2 times

□ **dev2dev** 1 year, 11 months ago

Selected Answer: C

C is correct. The requirement is to put restriction on viewing or inferring. In other words, don't allow to access the column. My previous choice A was wrong.

upvoted 1 times

□ **dev2dev** 1 year, 11 months ago

Selected Answer: A

You get 'The SELECT permission was denied on the column...' error if you use column level security. You need to allow to query the column with protection which is achieved using data masking. So A is correct

upvoted 1 times

□ **Sabajamal2010AtGmail** 2 years ago

to provide salespeople with the ability to view all the entries in Customers. (Column level security prevents that) The solution must prevent all the salespeople from viewing or inferring the credit card information. (Data masking helps infer information even when you can view the column)

upvoted 1 times

□ **vj84** 2 years ago

Data Masking is the correct Answer, it is not necessarily he need to use credit card masking. we can even use Default or Random and avoid users from inferring the data.

Hence A is the Right Answer.

upvoted 2 times

 **aasarii** 2 years ago

Selected Answer: C

upvoted 1 times

 **Amalbenrebai** 2 years, 4 months ago

Dynamic Data Masking should not be used as an isolated measure to fully secure sensitive data from users running ad-hoc queries on the database.

It is appropriate for preventing accidental sensitive data exposure, but will not protect against malicious intent to infer the underlying data.
==> as we would that salespeople can't infer the data so we will use CLS

upvoted 2 times

You develop data engineering solutions for a company.

A project requires the deployment of data to Azure Data Lake Storage.

You need to implement role-based access control (RBAC) so that project members can manage the Azure Data Lake Storage resources.

Which three actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Create security groups in Azure Active Directory (Azure AD) and add project members.
- B. Configure end-user authentication for the Azure Data Lake Storage account.
- C. Assign Azure AD security groups to Azure Data Lake Storage.
- D. Configure Service-to-service authentication for the Azure Data Lake Storage account.
- E. Configure access control lists (ACL) for the Azure Data Lake Storage account.

Correct Answer: ACE

AC: Create security groups in Azure Active Directory. Assign users or security groups to Data Lake Storage Gen1 accounts.

E: Assign users or security groups as ACLs to the Data Lake Storage Gen1 file system

Reference:

<https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-secure-data>

 **rashjan** Highly Voted 2 years, 1 month ago

Selected Answer: ACE

correct

upvoted 17 times

 **Nathan_W** Highly Voted 2 years, 3 months ago

nice question!

upvoted 7 times

 **jsav1** Most Recent 10 hours, 36 minutes ago

Selected Answer: ACE

correct

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: ACE

ACE is correct method.

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: ACE

Create security groups in Azure Active Directory (Azure AD) and add project members: Start by creating the necessary security groups in Azure AD and adding the project members to these groups. This step allows you to organize users and manage their permissions collectively.

Configure access control lists (ACL) for the Azure Data Lake Storage account: Next, configure the access control lists (ACL) for the Azure Data Lake Storage account. ACLs provide granular control over permissions at the individual file or folder level within the storage. By setting up ACLs, you can define specific access rights for different data assets.

Assign Azure AD security groups to Azure Data Lake Storage: Once the security groups and ACLs are set up, assign the Azure AD security groups to the Azure Data Lake Storage account. This step associates the security groups with the storage resources and enables you to grant permissions based on group membership rather than individually managing permissions for each user.

upvoted 4 times

 **francocalvo** 8 months ago

Isn't this a ACL model instead of RBAC?

upvoted 2 times

 **Jerrie86** 11 months, 3 weeks ago

1.Create security group.
2. Assign the Group/users to data lake.
3. Assign ACL (access control on the data which is stored inside the lake)

upvoted 1 times

 **anks84** 1 year, 4 months ago

Selected Answer: ACE

CORRECT !!

upvoted 3 times

 **ML_Novice** 1 year, 4 months ago

E-> A->C

is the order right ?

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: ACE

correct

upvoted 2 times

 **juanlu46** 1 year, 8 months ago

Selected Answer: ACE

Is correct!

upvoted 1 times

 **nss8500** 1 year, 11 months ago

Selected Answer: ACE

correct

upvoted 1 times

 **Podavenna** 2 years, 3 months ago

Correct answer!

upvoted 4 times

You have an Azure Data Factory version 2 (V2) resource named Df1. Df1 contains a linked service.

You have an Azure Key vault named vault1 that contains an encryption key named key1.

You need to encrypt Df1 by using key1.

What should you do first?

- A. Add a private endpoint connection to vault1.
- B. Enable Azure role-based access control on vault1.
- C. Remove the linked service from Df1.
- D. Create a self-hosted integration runtime.

Correct Answer: C

Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources.

Incorrect Answers:

D: A self-hosted integration runtime copies data between an on-premises store and cloud storage.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key> <https://docs.microsoft.com/en-us/azure/data-factory/concepts-linked-services> <https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime>

✉ **gnulf69** Highly Voted 2 years, 4 months ago

I believe this is correct, based on the question: What should you do FIRST?

A DF needs to be empty to be encrypted: <https://docs.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key#post-factory-creation-in-data-factory-ui>

So FIRST we need to empty the DF - then we can move on.

upvoted 39 times

✉ **hanzocuk** 1 year ago

B!!!

Enable Azure RBAC permissions on Key Vault:

<https://learn.microsoft.com/en-us/azure/key-vault/general/rbac-guide?tabs=azure-cli>

upvoted 1 times

✉ **auwia** Highly Voted 6 months, 2 weeks ago

Selected Answer: C

Correct answer:

A customer-managed key can only be configured on an empty data Factory. The data factory can't contain any resources such as linked services, pipelines and data flows.

<https://learn.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key#post-factory-creation-in-data-factory-ui>

upvoted 5 times

✉ **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: C

is the first step.

upvoted 1 times

✉ **Ram9198** 6 months, 1 week ago

Selected Answer: C

Correct

upvoted 1 times

✉ **vctrhugo** 6 months, 3 weeks ago

A customer-managed key can only be configured on an empty data Factory. The data factory can't contain any resources such as linked services, pipelines and data flows.

<https://learn.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key#post-factory-creation-in-data-factory-ui>

upvoted 1 times

✉ **rzeng** 1 year, 2 months ago

so you need to encrypt the df, you need to remove the bonded service first , answer is correct

upvoted 1 times

✉ **RajashekharC** 1 year, 4 months ago

Its C:

Your ADF should be empty during encryption process using a KEY
upvoted 3 times

Deeksha1234 1 year, 5 months ago

Selected Answer: C

correct answer
upvoted 3 times

juanlu46 1 year, 8 months ago

Selected Answer: C

You don't need to enable "RBAC", access policies is a default and more simple way to assign permissions, so B option is not necessary, but it is a requirement to delete the linked services to configure customer-managed key. So the correct answer is C - Delete linked services first.

<https://docs.microsoft.com/en-us/azure/key-vault/general/assign-access-policy?tabs=azure-portal>
<https://docs.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key#enable-customer-managed-keys>
upvoted 1 times

ploer 1 year, 11 months ago

Selected Answer: C

Correct. "A customer-managed key can only be configured on an empty data Factory. The data factory can't contain any resources such as linked services, pipelines and data flows."
upvoted 1 times

MFR 2 years ago

A customer-managed key can only be configured on an empty data Factory. The data factory can't contain any resources such as linked services, pipelines and data flows. It is recommended to enable customer-managed key right after factory creation.

Note: Azure Data Factory encrypts data at rest, including entity definitions and any data cached while runs are in progress. By default, data is encrypted with a randomly generated Microsoft-managed key that is uniquely assigned to your data factory.

Reference: <https://docs.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key>
upvoted 3 times

Canary_2021 2 years ago

Selected Answer: B

B should be the correct answer.
<https://docs.microsoft.com/en-us/azure/key-vault/general/rbac-guide?tabs=azure-cli>
upvoted 1 times

x089797 2 years, 1 month ago

Should it be D?
<https://docs.microsoft.com/en-us/powershell/module/az.datafactory/new-azdatafactoryv2linkedserviceencryptedcredential?view=azps-7.0.0>
upvoted 1 times

eoicp 2 years, 2 months ago

I think it's B. I recently changed a linked service pwf to key vault. I didn't delete the service and just added the managed Identity access to the vault with all the desired rules.
upvoted 2 times

Satschi 2 years, 4 months ago

Isn't B Correct ?
upvoted 2 times

You are designing an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that you can audit access to Personally Identifiable Information (PII).

What should you include in the solution?

- A. column-level security
- B. dynamic data masking
- C. row-level security (RLS)
- D. sensitivity classifications

Correct Answer: D

Data Discovery & Classification is built into Azure SQL Database, Azure SQL Managed Instance, and Azure Synapse Analytics. It provides basic capabilities for discovering, classifying, labeling, and reporting the sensitive data in your databases.

Your most sensitive data might include business, financial, healthcare, or personal information. Discovering and classifying this data can play a pivotal role in your organization's information-protection approach. It can serve as infrastructure for:

- ⇒ Helping to meet standards for data privacy and requirements for regulatory compliance.
- ⇒ Various security scenarios, such as monitoring (auditing) access to sensitive data.
- ⇒ Controlling access to and hardening the security of databases that contain highly sensitive data.

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview>

✉  **Podavenna** Highly Voted 2 years, 3 months ago

Correct answer!

upvoted 27 times

✉  **Deeksha1234** Highly Voted 1 year, 5 months ago

Selected Answer: D

An important aspect of the classification is the ability to monitor access to sensitive data. Azure SQL Auditing has been enhanced to include a new field in the audit log called `data_sensitivity_information`. This field logs the sensitivity classifications (labels) of the data that was returned by a query.

Ref - <https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview?view=azuresql>
upvoted 6 times

✉  **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: D

correct

upvoted 1 times

✉  **vctrhugo** 6 months, 3 weeks ago

Selected Answer: D

By implementing Data Discovery & Classification we can Audit access to sensitive data.

<https://learn.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview?view=azuresql#audit-sensitive-data>
upvoted 1 times

✉  **kornat** 9 months, 1 week ago

correct

upvoted 1 times

✉  **TimboobmiT** 1 year, 2 months ago

Why not dynamic data masking?

upvoted 2 times

✉  **vctrhugo** 6 months, 3 weeks ago

Data mask you hide the data, but can't audit who read it.

upvoted 2 times

✉  **juanlu46** 1 year, 8 months ago

Selected Answer: D

Is correct!

<https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview#audit-sensitive-data>
upvoted 3 times

 **AIcubeHead** 1 year, 9 months ago

Correct!

upvoted 1 times

Question #8

Topic 3

HOTSPOT -

You have an Azure subscription that contains an Azure Data Lake Storage account. The storage account contains a data lake named DataLake1.

You plan to use an Azure data factory to ingest data from a folder in DataLake1, transform the data, and land the data in another folder.

You need to ensure that the data factory can read and write data from any folder in the DataLake1 container. The solution must meet the following requirements:

- Minimize the risk of unauthorized user access.
- Use the principle of least privilege.
- Minimize maintenance effort.

How should you configure access to the storage account for the data factory? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Use

Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra
 a shared access signature (SAS)
 a shared key

to authenticate by using

a managed identity.
 a stored access policy.
 an Authorization header.

Correct Answer:

Answer Area

Use

Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra
 a shared access signature (SAS)
 a shared key

to authenticate by using

a managed identity.
 a stored access policy.
 an Authorization header.

 **kkk5566** Highly Voted  4 months, 1 week ago

correcr

upvoted 5 times

 **abhijit1990** Most Recent  3 months ago

should be B SAS

upvoted 2 times

 **MarkJoh** 1 month ago

No, because of the requirement "Minimize maintenance effort"

upvoted 3 times

 **MSExpert** 4 months, 4 weeks ago

Correct

upvoted 2 times

HOTSPOT -

You are designing an Azure Synapse Analytics dedicated SQL pool.

Groups will have access to sensitive data in the pool as shown in the following table.

| Name | Enhanced access |
|------------|--------------------------------------|
| Executives | No access to sensitive data |
| Analysts | Access to in-region sensitive data |
| Engineers | Access to all numeric sensitive data |

You have policies for the sensitive data. The policies vary by region as shown in the following table.

| Region | Data considered sensitive |
|---------|---|
| RegionA | Financial, Personally Identifiable Information (PII) |
| RegionB | Financial, Personally Identifiable Information (PII), medical |
| RegionC | Financial, medical |

You have a table of patients for each region. The tables contain the following potentially sensitive columns.

| Name | Sensitive data | Description |
|--------------|----------------|---|
| CardOnFile | Financial | Debit/credit card number for charges |
| Height | Medical | Patient's height in cm |
| ContactEmail | PII | Email address for secure communications |

You are designing dynamic data masking to maintain compliance.

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:

Answer Area

- | Statements | Yes | No |
|---|-----------------------|-----------------------|
| Analysts in RegionA require dynamic data masking rules for [Patients_RegionA]. | <input type="radio"/> | <input type="radio"/> |
| Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height] | <input type="radio"/> | <input type="radio"/> |
| Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height] | <input type="radio"/> | <input type="radio"/> |

Answer Area

- | Statements | Yes | No |
|--|----------------------------------|----------------------------------|
| Correct Answer: Analysts in RegionA require dynamic data masking rules for [Patients_RegionA]. | <input checked="" type="radio"/> | <input type="radio"/> |
| Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height] | <input type="radio"/> | <input checked="" type="radio"/> |
| Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height] | <input checked="" type="radio"/> | <input type="radio"/> |

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview>

 **[Removed]** Highly Voted  2 years, 4 months ago

The Answer should be No, No, No. Analysts have access to in-region sensitive data, so the first one should be No. Engineers have access to all numeric sensitive data, Height is patient's height in CM, so the second and third one should also No.

upvoted 130 times

 **Amalbenrebai** 2 years, 4 months ago

I agree: NO NO NO

upvoted 14 times

范69 2 months, 4 weeks ago

The wording of the question is bad but the GIVEN ANSWER is actually RIGHT.

The question is about masking data for certain roles that have access to that data.

1. YES because Analysts in Region A have access to Region A sensitive data and in region A "financial and PII" are considered sensitive so these fields should be masked for the analyst (can leave out height since it's not sensitive).

2. NO, because in Region C [Height] is not considered sensitive.

3. YES, because in Region B [Height] which is medical is considered sensitive so it should be masked for the Engineer because he has access to the numeric [Height] data.

Note: it would be different when you had to consider [Patients_Region] and [Height] separately.

upvoted 2 times

auwia 6 months, 2 weeks ago

The first 2 are false because the Patient table is not present in Region A, NO Medical for region A.

Probably the third one is Yes because of comma between table name and height, so probably it means give access to other columns too.

upvoted 2 times

Seansmyke 1 year, 10 months ago

Its no, yes, yes

Engineers only have access to numeric data. the contact email is considered sensitive in the regions and is not numeric

upvoted 10 times

janaki 7 months, 2 weeks ago

but the questions is asked about region B and region C.

upvoted 1 times

ADHDBA 1 year, 9 months ago

but they clearly specify only height, no mention of email and height is numeric so steeee is correct

upvoted 4 times

g2000 1 year, 8 months ago

there's a comma between height and patients_regionA. i would assume they are two distinct items.. namely height in any region and patients_region_A. in region a, PII is considered sensitive which is something engineers have no access

upvoted 2 times

Aditya0891 1 year, 7 months ago

g2000 read the question carefully. It's clearly mentioned you have table for patients in each region. So patients_regionA means table in region A and then height is the column which is being referred to in 2nd second question and similarly for 3rd as well. SO the answer is No, No, No

upvoted 5 times

HaBroNounen Highly Voted 2 years, 3 months ago

the solution is correct: Yes, no, yes. Just because somebody has access, doesn't mean that they don't need any dynamic masking. It just means that they have access and a policy is required. If they had no access, then obviously no data masking is required.

Statement 1: Analysts in Region A have access to (all) the following sensitive data in region A: CardOnFile, Height and ContactEmail. Since financial (CardOnFile) and PII (ContactEmail) are considered sensitive data you need dynamic data masking: so Yes.

Statement 2 & 3: Engineers have access to all numeric sensitive data (which means in every region). So they have access to height. Height is medical and therefore only sensitive in Region B according to the second table, but not in Region A. So Statement 2 is "No" and Statement 3 is "Yes"

upvoted 81 times

Julius7000 2 years, 3 months ago

I think You are correct

upvoted 6 times

noranathalie 2 years, 2 months ago

I would go for this answer as well.. otherwise the double question 2 and 3 would be useless.

upvoted 2 times

YLiу 2 years, 2 months ago

But for statement 1, [height] is not considered sensitive data for Region A, so it should not require data mask on [height]. -> A is NO

Also I am confused about whether we should apply the policy of sensitive data based on the region of data or the region of the requester (eg engineer from region C requesting data of region A)?

upvoted 2 times

kkk5566 Most Recent 4 months, 1 week ago

in oder, y,n,y

upvoted 1 times

auwia 6 months, 2 weeks ago

First: NO, because there is no medical data in the region A. Second and Third, NO, because data engineers can see numeric data in all regions (height is number).

upvoted 2 times

janaki 7 months, 1 week ago

Answer should be NO, NO, NO. Analyst have access to in-region sensitive data, Engineers have access to all numeric sensitive data.

upvoted 1 times

✉ **g2000** 2 months, 4 weeks ago

last one is yes... in region b, financial, pii and medical are sensitive data. but engineers have access to all numeric sensitive data. pii is sensitive data.

upvoted 1 times

✉ **chryckie** 8 months, 2 weeks ago

Q1: Yes, these users need to see past any default masking.

Analysts have access to in-region sensitive data. So, since they're in RegionA looking at RegionA data, the default masking should be dynamically removed for them.

Q2: No, these users should see data with default masking.

You have to assume that Enhanced Access only apply to users when they are in their own region. Since the Engineers are outside of the region, they are treated as regular users, with default masking. Perhaps there's some documentation in Azure that says you can't enhance access for users outside of a given region, but I'm not aware of any. Personally, I feel the wording of the Enhanced Access makes me assume it's "region agnostic". However, the given answer (of No) seems to imply otherwise.

Q3: Yes, these users need to see past SOME default masking.

There's a lot to consider, but I assume because the Engineers need to see numeric data, and both Financial and Medical data is numeric, they need to SOME data unmasked.

upvoted 1 times

✉ **chryckie** 8 months, 2 weeks ago

This is a poorly worded question, in my opinion. I eventually came to accept the given answer of Yes, No, Yes. However, my gut would have had me say No (no masking), Yes (mask e-mail), Yes (mask e-mail).

These were the questions I had when trying to sort through this one.

1. Is Enhanced Access truly defined as only applicable should the user be in the same region as the data? (I didn't want to.)

2. Should we only be considering the Height field for Q2, Q3? (Hard to say, with that comma....)

3. If we're meant to consider the full table, then (a) is it a "Yes" if ANY data needs to be unmasked, or (b) is it only a "Yes" if ALL data needs to be unmasked? (I'd assume A.)

4. Does the region of the Engineer matter at all? (I doubt it.)

Not fun to sort through before committing to an answer. (I spent way too long typing this up too.)

upvoted 2 times

✉ **chryckie** 8 months, 2 weeks ago

Answer: Yes, No, Yes.

This is a poorly worded question, in my opinion. I eventually came to accept the given answer of Yes, No, Yes. However, my gut would have had me say No (no masking), Yes (mask e-mail), Yes (mask e-mail).

I initially assumed that "Yes" meant the user should have the data masked/treated for them. Based on the given answers (of Yes, No, Yes) it seems like it's the opposite

upvoted 1 times

✉ **chryckie** 8 months, 2 weeks ago

Answer: Yes, No, Yes.

This is a poorly worded question, in my opinion. I eventually came to accept the given answer of Yes, No, Yes. However, my gut would have had me say No (no masking), Yes (mask e-mail), Yes (mask e-mail).

upvoted 1 times

✉ **chryckie** 8 months, 2 weeks ago

Sorry for the spam. The site was throwing an error when I would try to submit my full comment....

upvoted 1 times

✉ **Dhaval_Azure** 9 months, 3 weeks ago

after reading discussion very confused. What could be the answer.

upvoted 6 times

✉ **rcpaudel** 7 months, 3 weeks ago

Correct answer is YES, NO & YES, look at the explanation from esaade underneath. The fact that the data should be unmasked for certain group, these are masked by some rules. After masking, some are unmasked for required group- this holds for Q1 & Q3. Q2 does not have height on it and hence no rule is needed.

upvoted 1 times

✉ **esaade** 10 months ago

Analysts in RegionA require dynamic data masking rules for [Patients RegionA].

Yes. Since analysts in RegionA have access to in-region sensitive data, which includes PII, dynamic data masking rules should be implemented for the [Patients RegionA] table to mask the [ContactEmail] column which contains PII.

Engineers in RegionC require a dynamic data masking rule for [Patients RegionA], [Height].

No. Engineers in RegionC have access to all numeric sensitive data, but [Height] is not considered sensitive data in RegionC, only in RegionB. Therefore, there is no need to implement a dynamic data masking rule for [Height] in RegionC.

Engineers in RegionB require a dynamic data masking rule for [Patients RegionB], [Height].

Yes. Engineers in RegionB have access to sensitive data, including medical data, which includes the [Height] column in the [Patients RegionB] table. Therefore, dynamic data masking should be implemented for the [Height] column in the [Patients RegionB] table.
upvoted 5 times

□ **Billybob0604** 1 year, 1 month ago

This answer is clearly NO, NO, NO
upvoted 1 times

□ **XiltroX** 1 year, 1 month ago

The answer is No for all questions. Engineers have full access to all data so no need for data masking. Analysts have access to in region data already.
upvoted 1 times

□ **dmitriypo** 1 year, 2 months ago

I would go for Yes, Yes, Yes.
Engineers have access to medical info (Height) in regions B and C, thus Height needs to be masked.
upvoted 1 times

□ **rzeng** 1 year, 2 months ago

agree with No, No, No
upvoted 1 times

□ **debarun** 1 year, 3 months ago

I think its No, No, yes

Patients data is not sensitive in region A so need of masking

Patients data is not sensitive in region A so engineers C who has access to all numeric sensitive data has access to it but since it is not sensitive so no need of masking

patients data and height (medical data) is sensitive on region B and engineers and engineers have access to it, so it surely needs masking.
upvoted 8 times

□ **dom271219** 1 year, 4 months ago

No no no
Obviously
upvoted 1 times

□ **Deeksha1234** 1 year, 5 months ago

Correct, Agree with Habronounen
upvoted 2 times

DRAG DROP -

You have an Azure Synapse Analytics SQL pool named Pool1 on a logical Microsoft SQL server named Server1.

You need to implement Transparent Data Encryption (TDE) on Pool1 by using a custom key named key1.

Which five 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.

Select and Place:

Actions**Answer Area**

Enable TDE on Pool1.

店铺：
IT认证考试服务

Assign a managed identity to Server1.

店铺：
IT认证考试服务

Configure key1 as the TDE protector for Server1.



Add key1 to the Azure key vault.



Create an Azure key vault and grant the managed identity permissions to the key vault.

店铺：IT认证考试服务

Correct Answer:**Actions****Answer Area**

Assign a managed identity to Server1.

Create an Azure key vault and grant the managed identity permissions to the key vault.



Add key1 to the Azure key vault.



Configure key1 as the TDE protector for Server1.

Enable TDE on Pool1.

Step 1: Assign a managed identity to Server1

You will need an existing Managed Instance as a prerequisite.

Step 2: Create an Azure key vault and grant the managed identity permissions to the vault

Create Resource and setup Azure Key Vault.

Step 3: Add key1 to the Azure key vault

The recommended way is to import an existing key from a .pfx file or get an existing key from the vault. Alternatively, generate a new key directly in Azure Key Vault.

Vault.

Step 4: Configure key1 as the TDE protector for Server1

Provide TDE Protector key -

Step 5: Enable TDE on Pool1 -

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/managed-instance/scripts/transparent-data-encryption-byok-powershell>

Sudheer_K Highly Voted 2 years, 3 months ago

Answer is right!

upvoted 18 times

Liz42 Highly Voted 2 years, 2 months ago

Shouldn't the last two be switched? Enable TDE then configure the key?

upvoted 7 times

□ **anto69** 1 year, 11 months ago

I also think so, but not sure

upvoted 1 times

□ **noranathalie** 2 years, 2 months ago

I think the correct answer is the one provided.

Please see the link below:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/transparent-data-encryption-byok-configure?tabs=azure-powershell>

upvoted 10 times

□ **shoottheduck** 10 months, 2 weeks ago

Checked this link and it supports the answer given

upvoted 2 times

□ **kkk5566** Most Recent 4 months, 1 week ago

correct

upvoted 1 times

□ **hanzocuk** 1 year ago

1. Get a KV
2. Add key to KV
3. Assign MI to server
4. Enable TDE
5. Config TDE

upvoted 3 times

□ **Okea** 11 months, 2 weeks ago

4 and 5 should be swapped

upvoted 2 times

□ **rzeng** 1 year, 2 months ago

correct

upvoted 1 times

□ **Deeksha1234** 1 year, 4 months ago

given ans is correct

upvoted 1 times

□ **dev2dev** 1 year, 11 months ago

options looks correct but. i am bit lost. I dont see tde settings for the logical server it creates by default while creating synapse analytics ws. and there is no option to create synapse analytics pool when I create logic server and then try to create database.

upvoted 1 times

You have a data warehouse in Azure Synapse Analytics.

You need to ensure that the data in the data warehouse is encrypted at rest.

What should you enable?

- A. Advanced Data Security for this database
- B. Transparent Data Encryption (TDE)
- C. Secure transfer required
- D. Dynamic Data Masking

Correct Answer: B

Azure SQL Database currently supports encryption at rest for Microsoft-managed service side and client-side encryption scenarios.

⇒ Support for server encryption is currently provided through the SQL feature called Transparent Data Encryption.

⇒ Client-side encryption of Azure SQL Database data is supported through the Always Encrypted feature.

Reference:

<https://docs.microsoft.com/en-us/azure/security/fundamentals/encryption-atrest>

✉  **Podavenna** Highly Voted  2 years, 3 months ago

Correct!

upvoted 22 times

✉  **juanlu46** Highly Voted  1 year, 8 months ago

Selected Answer: B

Correct!

upvoted 5 times

✉  **kkk5566** Most Recent  4 months ago

Selected Answer: B

B is correct

upvoted 1 times

✉  **anks84** 1 year, 4 months ago

Selected Answer: B

Correct !

upvoted 3 times

✉  **Deeksha1234** 1 year, 5 months ago

Selected Answer: B

correct

upvoted 3 times

You are designing a streaming data solution that will ingest variable volumes of data.

You need to ensure that you can change the partition count after creation.

Which service should you use to ingest the data?

- A. Azure Event Hubs Dedicated
- B. Azure Stream Analytics
- C. Azure Data Factory
- D. Azure Synapse Analytics

Correct Answer: A

You can't change the partition count for an event hub after its creation except for the event hub in a dedicated cluster.

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features>

 **Canary_2021** Highly Voted 2 years ago

Selected Answer: A

A is the correct Answer.

You can specify the number of partitions at the time of creating an event hub. In some scenarios, you may need to add partitions after the event hub has been created. This article describes how to dynamically add partitions to an existing event hub.

Dynamic additions of partitions is available only in premium and dedicated tiers of Event Hubs.

<https://docs.microsoft.com/en-us/azure/event-hubs/dynamically-add-partitions>

upvoted 13 times

 **mshakir** Highly Voted 2 years, 3 months ago

Answer is Correct according to given link

upvoted 12 times

 **kkk5566** Most Recent 4 months, 1 week ago

Answer is Correct

upvoted 1 times

 **esaade** 10 months ago

Selected Answer: A

A. Azure Event Hubs Dedicated would be the best choice to ingest the variable volumes of data and change the partition count after creation.

Azure Event Hubs Dedicated is a highly scalable and fully managed event hub service that can ingest millions of events per second. It allows you to create and manage partitions, and you can dynamically increase or decrease the number of partitions to accommodate changes in data volume or throughput requirements.

Azure Stream Analytics, Azure Data Factory, and Azure Synapse Analytics are not specifically designed to manage the partition count after creation. Although they can be used to ingest streaming data, they may not provide the flexibility to change the partition count dynamically.

upvoted 3 times

 **shoottheduck** 10 months, 2 weeks ago

Selected Answer: A

Correct

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: A

A is correct

upvoted 3 times

 **Dothy** 1 year, 8 months ago

As A is correct

upvoted 3 times

 **BerendJan** 2 years, 3 months ago

From the provided link: "We recommend that you choose at least as many partitions as you expect that are required during the peak load of your application for that particular event hub. You can't change the partition count for an event hub after its creation except for the event hub in a dedicated cluster. The partition count for an event hub in a dedicated Event Hubs cluster can be increased after the event hub has been created, but the distribution of streams across partitions will change when it's done as the mapping of partition keys to partitions changes, so you should try hard to avoid such changes if the relative order of events matters in your application."

upvoted 5 times

□  **dikkieknor** 2 years, 2 months ago

I think you're focusing on the wrong part. It says that the partition count can be increased in a dedicated event hubs cluster. And this question is about event hubs dedicated (cluster?), so I think event hubs is the correct answer.

upvoted 3 times

Question #13

Topic 3

You are designing a date dimension table in an Azure Synapse Analytics dedicated SQL pool. The date dimension table will be used by all the fact tables.

Which distribution type should you recommend to minimize data movement during queries?

- A. HASH
- B. REPLICATE
- C. ROUND_ROBIN

Correct Answer: B

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.

Incorrect Answers:

A: A hash distributed table is designed to achieve high performance for queries on large tables.

C: 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/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overview>

□  **allagowf** Highly Voted 1 year, 2 months ago

Selected Answer: B

correct B

upvoted 7 times

□  **vctrhugo** Highly Voted 6 months, 3 weeks ago

Selected Answer: B

HASH = Fact/2+Gb table

REPLICATE = Dimensionn

ROUND_ROBIN = Staging

upvoted 5 times

□  **hassexat** Most Recent 4 months ago

Selected Answer: B

Replicate

upvoted 1 times

□  **kkk5566** 4 months, 1 week ago

Selected Answer: B

correct

upvoted 1 times

□  **anks84** 1 year, 4 months ago

Selected Answer: B

REPLICATE

upvoted 5 times

HOTSPOT -

You develop a dataset named DBTBL1 by using Azure Databricks.

DBTBL1 contains the following columns:

- SensorTypeID
- GeographyRegionID
- Year
- Month
- Day
- Hour
- Minute
- Temperature
- WindSpeed
- Other

You need to store the data to support daily incremental load pipelines that vary for each GeographyRegionID. 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 | (<code>"*</code>) |
| .format | (<code>"GeographyRegionID"</code>) |
| .partitionBy | (<code>"GeographyRegionID", "Year", "Month", "Day"</code>) |
| .sortBy | (<code>"Year", "Month", "Day", "GeographyRegionID"</code>) |

```
.mode("append")
```

| |
|--|
| .csv(<code>"/DBTBL1"</code>) |
| .json(<code>"/DBTBL1"</code>) |
| .parquet(<code>"/DBTBL1"</code>) |
| .saveAsTable(<code>"/DBTBL1"</code>) |

Answer Area

```
df.write
```

Correct Answer:

| | |
|--------------------|--|
| .bucketBy | (<code>"*</code>) |
| .format | (<code>"GeographyRegionID"</code>) |
| partitionBy | (<code>"GeographyRegionID", "Year", "Month", "Day"</code>) |
| .sortBy | (<code>"Year", "Month", "Day", "GeographyRegionID"</code>) |

```
.mode("append")
```

| |
|--|
| .csv(<code>"/DBTBL1"</code>) |
| .json(<code>"/DBTBL1"</code>) |
| .parquet(<code>"/DBTBL1"</code>) |
| saveAsTable(<code>"/DBTBL1"</code>) |

Box 1: .partitionBy -

Incorrect Answers:

- .format:

Method: format():

Arguments: "parquet", "csv", "txt", "json", "jdbc", "orc", "avro", etc.

↪ .bucketBy:

Method: bucketBy()

Arguments: (numBuckets, col, col..., coln)

The number of buckets and names of columns to bucket by. Uses Hive's bucketing scheme on a filesystem.

Box 2: ("Year", "Month", "Day", "GeographyRegionID")

Specify the columns on which to do the partition. Use the date columns followed by the GeographyRegionID column.

Box 3: .saveAsTable("/DBTBL1")

Method: saveAsTable()

Argument: "table_name"

The table to save to.

Reference:

<https://www.oreilly.com/library/view/learning-spark-2nd/9781492050032/ch04.html> <https://docs.microsoft.com/en-us/azure/databricks/delta/delta-batch>

□  **PallaviPatel** Highly Voted  2 years ago

1. Partition by
2. GeographyRegionID, Year, Month, Day as the pipelines are per region this seems right choice
3. Parquet

upvoted 88 times

□  **uzairahm** 1 year, 6 months ago

regarding point 2 Solution needs to support daily incremental load so having Year, Month, Day first would be more useful
upvoted 5 times

□  **petilda** Highly Voted  2 years, 4 months ago

I suggest storing the data in parquet
upvoted 52 times

□  **xmety** Most Recent  1 week ago

1. Partition by
2. Year, Month, Day, GeographyRegionID (it said to minimize storage cost, not query performance. if GeographyRegionID goes first, each regionID will have repeated folders for different dates)
3. Parquet

upvoted 1 times

□  **pperf** 3 months, 1 week ago

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-best-practices>
upvoted 1 times

□  **VikkiC** 5 months, 4 weeks ago

This question is similar to #36Topic 1, if you reference that question, the answer should be 1. Partitioned By, 2. GeographyRegionID, Year, Month, Day, 3. Parquet
upvoted 3 times

□  **JosephVishal** 1 year ago

For 3.) if parquet with partitions, then it should "overwrite" mode instead of "append". Since, it is "append" mode, I think saveAsTable is more appropriate.

upvoted 1 times

□  **Deeksha1234** 1 year, 5 months ago

Agree with Pallavi
1. Partition by
2. GeographyRegionID, Year, Month, Day
3. Parquet
upvoted 7 times

□  **OldSchool** 1 year, 1 month ago

Agree on 1) & 3) but for 2) it should be year/month/day/GeographyRegionId and for each day we would generate several GeographyRegionId.parquet files
upvoted 2 times

Disregard my comment on 2). Provided answer is the correct one.

upvoted 2 times

□  **dsp17** 1 year, 6 months ago

Parquet is must (offer higher compression rates)- "The solution must minimize storage costs."
upvoted 3 times

⊕ **Aurelkb** 1 year, 6 months ago

it is the same question on Topic 1 Question 36.

Then

1. Partition by
2. GeographyRegionID, Year, Month, Day
3. Parquet

upvoted 9 times

⊕ **Backy** 1 year, 8 months ago

// the correct answer is

```
df.write.partitionBy("GeographyRegionID").mode("append").parquet("/DBTBL1")
```

// or

```
df.write.partitionBy("GeographyRegionID","Year","Month","Day").mode("append").parquet("/DBTBL1")
```

// Question says "minimize storage costs" so I would select the first one

upvoted 4 times

⊕ **Davico93** 1 year, 6 months ago

Agree, but if you choose the first one, you won't have the daily data

upvoted 1 times

⊕ **allagowf** 1 year, 2 months ago

no mentionning for daily data in the question

upvoted 2 times

⊕ **Spinozabubble** 7 months, 4 weeks ago

daily incremental load pipelines

upvoted 2 times

⊕ **Amsterliese** 1 year, 9 months ago

I was wondering if the incremental load is supported for parquet, but since "append" mode is used, this should be alright. The question asks to minimize costs, so I go for parquet (not saveAsTable).

partitionBy

GeopgraphyRegionID, Year, Month, Day (pipelines per region; daily load)

parquet

upvoted 2 times

⊕ **dev2dev** 1 year, 11 months ago

its recommend to use partitions first before Y/M/D so that they can be managed easily such as assigning security, or processing by business unit such as zone/country/area etc., GeographyRegionId/Year/Month/Day and Paraquet are answers

upvoted 9 times

⊕ **bad_atitude** 2 years ago

Mes chers amis:

- 1.Sortby
- 2.GeographyRegionId, Year, Month, Day
- 3.Parquet

upvoted 7 times

⊕ **jv2120** 2 years ago

only reason for using .parquet is option seems to be dataset path not table else saveable is right.

upvoted 2 times

⊕ **Aslam208** 2 years, 2 months ago

I agreed with @hryniwka, saveAsTable takes db name and table name not path.

upvoted 4 times

⊕ **hryniwka** 2 years, 2 months ago

saveAsTable is wrong as in saveAsTable we specify name for the table and here is a path, so I would suggest that correct answer is parquet

upvoted 4 times

⊕ **sparkchu** 1 year, 9 months ago

u got the right answer with wrong reasoning, saveAsTable() can also take file path when a unmanaged table is created in such case. Like rav009 said, the correct answer for this not to choose saveAsTable() is because of the more disk space required for Delta format.

upvoted 1 times

⊕ **rav009** 2 years, 3 months ago

saveAsTable will use the delta format to save the dataset.

delta format is based on parquet with versions

so delta will cost more on storage

Box 3 should be parquet

upvoted 6 times

You are designing a security model for an Azure Synapse Analytics dedicated SQL pool that will support multiple companies.

You need to ensure that users from each company can view only the data of their respective company.

Which two objects should you include in the solution? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. a security policy
- B. a custom role-based access control (RBAC) role
- C. a predicate function
- D. a column encryption key
- E. asymmetric keys

Correct Answer: AB

A: Row-Level Security (RLS) enables you to use group membership or execution context to control access to rows in a database table.

Implement RLS by using the CREATE SECURITY POLICY Transact-SQL statement.

B: Azure Synapse provides a comprehensive and fine-grained access control system, that integrates:

Azure roles for resource management and access to data in storage,

▪

▫ Synapse roles for managing live access to code and execution,

▫ SQL roles for data plane access to data in SQL pools.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security> <https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-access-control-overview>

✉ lukeonline Highly Voted 2 years ago

Selected Answer: AB

A and B

upvoted 18 times

✉ alexleonvalencia Highly Voted 2 years ago

Selected Answer: AC

Respuesta A/C

upvoted 15 times

✉ VJPR 2 years ago

why not RBAC?

upvoted 6 times

✉ sensaint 1 year ago

Assuming RBAC is already in place, predicate function for row-level security would be next step. However, it's not clearly stated in question which makes it confusing.

upvoted 1 times

✉ dakku987 1 week ago

see you can not add even row level security bcz you are saying some company will have access to some of its rows even that is not allowed

AB

upvoted 1 times

✉ zizonesol 10 months ago

That's why I went with AB instead because it wasn't mentioned. Therefore, we should assume that the system does not already have the RBAC already in place.

upvoted 5 times

✉ Momoanwar Most Recent 2 weeks, 2 days ago

Selected Answer: AB

Chatgpt:

If only two responses must be selected from the given options, based on the question asked, the two most relevant objects to ensure that users can view only the data of their respective company would be:

A. **A security policy**: This would define the rules and conditions for data access based on company affiliation.

B. **A custom role-based access control (RBAC) role**: This would allow for the assignment of specific access rights depending on the user's company.

Even though a predicate function could be used as part of a security policy implementation, it is typically a component of such a policy, rather than a standalone object. Options D and E are related to encryption and are not directly used to control data views based on the user's company.

Therefore, the two most appropriate answers, according to the question, would be A and B.

upvoted 1 times

 **MarkJoh** 1 month ago

Selected Answer: AC

Answer is A & C. Although as many have indicated, the steps are

- Create the users or groups you want to isolate access.
- Create the inline table-valued function that will filter the results based on the predicate defined.
- Create a security policy for the table, assigning the function created above

The first step may look like "objects"/option B but option B says "A custom role-based access control (RBAC) role".

In reality, you would want to create a domain table with companyId and RoleName and create one Role per companyId. (Or maybe a set of roles per companyId depending on what the requirements are). Then the predicate function would use the meta data driven companyIdRoleName table.

upvoted 2 times

 **Shanuramasubbu** 1 month, 3 weeks ago

Based on this MS doc, A&C is the right answer

<https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=azure-sqldw-latest&preserve-view=true>

upvoted 2 times

 **y154707** 2 months ago

Question says: "Which two objects should you include in the solution?". It seems that answers A, B and C should be part of the solution, so any combination of the 3 should be ok in terms of a valid answer. If the question would ask for "the sequence of the first 2 steps required to achieve the goal" then the answer would be B => C => A.

upvoted 1 times

 **pperf** 3 months, 1 week ago

It's A & C

upvoted 1 times

 **EliteAllen** 4 months ago

Selected Answer: BC

B: To define roles that have specific permissions to access certain data (company-specific).

C: To implement a function that filters the data a user can access, based on their company.

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: AC

correct

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: AB

a and B

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

sorry a and C

upvoted 1 times

 **pavankr** 6 months, 1 week ago

RBAC is for to use "internal" company. So 100% wrong.

upvoted 2 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: AC

<https://learn.microsoft.com/en-gb/training/modules/implement-compliance-controls-sensitive-data/5-implement-row-level-security>

Create the users or groups you want to isolate access.

Create the inline table-valued function that will filter the results based on the predicate defined.

Create a security policy for the table, assigning the function created above.

upvoted 1 times

 **klayytech** 6 months, 4 weeks ago

Option C, a predicate function, is not wrong. It can be a helpful tool for fine-grained control over data access. However, it is not strictly necessary for the solution described in the question. A security policy and a custom RBAC role can be used to achieve the desired outcome without a predicate function.

Here is an example of how you could use a security policy and a custom RBAC role to control access to data in a Synapse Analytics dedicated SQL pool:

Create a security policy that defines who can access data, what data they can access, and how they can access it.

Create a custom RBAC role that gives users specific permissions to data in the pool.

Assign the custom RBAC role to users based on their company name.

With this configuration in place, users will only be able to access data that they are authorized to access.

The use of a predicate function can be helpful in some scenarios, but it is not always necessary. In the case of the question, a security policy and a custom RBAC role are sufficient to achieve the desired outcome.

upvoted 3 times

✉️ **mamahani** 8 months, 2 weeks ago

I think A/C as per examples in docs:

<https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16#CodeExamples>
i dont think its RBAC; according to documentation Synapse RBAC is used to manage who can:

Publish code artifacts and list or access published code artifacts,
Execute code on Apaches Spark pools and Integration runtimes,
Access linked (data) services protected by credentials
Monitor or cancel job execution, review job output, and execution logs."

I do not see the direct link with limiting retrieved data here;

<https://learn.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-synapse-rbac>

upvoted 2 times

✉️ **esaade** 10 months ago

Selected Answer: BC

Answer: B and C

To ensure that users from each company can view only the data of their respective company in an Azure Synapse Analytics dedicated SQL pool, you can use custom role-based access control (RBAC) roles to define specific permissions for each company, and use predicate functions to apply row-level security (RLS) to restrict access based on company membership. By doing this, you can limit the scope of access to the appropriate company data.

A security policy is a mechanism for implementing automatic security controls to enforce compliance requirements, which may not be directly related to company-specific data access.

A column encryption key is used for encrypting sensitive data, but it does not necessarily restrict access based on company membership.

Asymmetric keys are used for secure communication and authentication, but they do not directly relate to company-specific data access control.
upvoted 2 times

✉️ **jz10** 9 months, 2 weeks ago

ChatGPT isn't always reliable

upvoted 4 times

✉️ **janaki** 7 months, 2 weeks ago

@jz10 you're correct. After ChatGPT answers any of your certification exam questions, you then type -- sure? ChatGPT will change its answer...so 'Yes' ChatGPT is not reliable.

upvoted 2 times

✉️ **EliteAllen** 4 months ago

GPT-4 Chose BC, I asked are you sure his response:

Yes, I am confident with the choices.

To restrict data access at a granular level, where users from different companies can only view data related to their respective companies, implementing row-level security is a well-accepted strategy. Here's a bit more detail on why the selected options are pertinent:

B. Custom Role-Based Access Control (RBAC) Role: Setting up custom roles can allow you to grant specific permissions for accessing certain data. By setting up roles based on companies, you can ensure that users are only able to access the data related to their company.

C. Predicate Function: In the context of row-level security, a predicate function is used to filter the rows of the table that are visible to the user. You can define a function that filters rows based on the company attribute, ensuring that users can only access data from their own company.

upvoted 1 times

✉️ **AHUI** 10 months, 1 week ago

A, C

<https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16#CodeExamples>

upvoted 2 times

✉️ **haidebelognime** 10 months, 2 weeks ago

Selected Answer: B

The answer is B

upvoted 1 times

You have a SQL pool in Azure Synapse that contains a table named dbo.Customers. The table contains a column name Email. You need to prevent nonadministrative users from seeing the full email addresses in the Email column. The users must see values in a format of aXXX@XXXX.com instead. What should you do?

- A. From Microsoft SQL Server Management Studio, set an email mask on the Email column.
- B. From the Azure portal, set a mask on the Email column.
- C. From Microsoft SQL Server Management Studio, grant the SELECT permission to the users for all the columns in the dbo.Customers table except Email.
- D. From the Azure portal, set a sensitivity classification of Confidential for the Email column.

Correct Answer: A

The Email masking method, which exposes the first letter and replaces the domain with XXX.com using a constant string prefix in the form of an email address. aXX@XXXX.com

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview>

✉ **edba** Highly Voted 2 years ago

I think it's a terrible question, both A(using T-SQL) and B (via GUI) can do the job.
upvoted 19 times

✉ **rzeng** Highly Voted 1 year, 2 months ago

Selected Answer: A

Go with A, reason for not B, if email column is string type ,default masking will make it as xxxxxxxx, so here I go with email mask on email column.
<https://learn.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview?view=azuresql>
upvoted 12 times

✉ **andjurovicela** 5 months, 1 week ago

but you don't have to opt for the default and the add masking rule step from the link shows the exact same format as in the task. Therefore I would go with B to avoid overthinking :D
upvoted 2 times

✉ **Momoanwar** Most Recent 2 weeks, 2 days ago

Selected Answer: A

Its A ! B can work but not by default.
Cgatgpt :

Based solely on the information provided in the prompt and considering that any unspecified option would use a default value, the appropriate response would be:

A. From Microsoft SQL Server Management Studio, set an email mask on the Email column.

This is because option A specifically mentions setting an email mask, which is the type of masking required by the scenario. The other options do not mention configuring a custom masking format for email addresses.

upvoted 1 times

✉ **pperf** 3 months, 1 week ago

Both A & B are correct capable of achieving the same. But let's go for A.
upvoted 1 times

✉ **kkk5566** 4 months, 1 week ago

Selected Answer: B

A or B
upvoted 1 times

✉ **kkk5566** 4 months ago

agree with @auwia ,B
upvoted 1 times

✉ **Ram9198** 4 months, 4 weeks ago

Selected Answer: A

default masking will make it as xxxxxxxx,
upvoted 1 times

 **Ram9198** 6 months, 1 week ago

Selected Answer: A

B says just mask and not email mask
upvoted 3 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: B

The link provided in the solution is correctly pointing to the solution: Dynamic Data Masking, that is done from the Azure Portal, so the correct answer is B! :)
upvoted 3 times

 **Shanmahi** 1 year, 1 month ago

Selected Answer: B

email masking option via ssms
upvoted 1 times

 **OldSchool** 1 year, 1 month ago

Selected Answer: B

Vote for B because of "You set up a dynamic data masking policy in the Azure portal by selecting the Dynamic Data Masking blade under Security in your SQL Database configuration pane."
Source: <https://learn.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview?view=azuresql#:~:text=You%20set%20up%20a%20dynamic%20data%20masking%20policy%20in%20the%20Azure%20portal%20by%20selecting%20the%20Dynamic%20Data%20Masking%20blade%20under%20Security%20in%20your%20SQL%20Database%20configuration%20pane.>
upvoted 1 times

 **amitshinde14** 1 year, 3 months ago

B correct
upvoted 1 times

 **Deeksha1234** 1 year, 5 months ago

both A and B are correct
upvoted 1 times

 **ROLLINGROCKS** 1 year, 5 months ago

Selected Answer: A

Occams razor with this one
upvoted 2 times

 **Glen711** 1 year, 5 months ago

Selected Answer: A

There are lots of comments here saying that the question does not ask for the default masking format. I'd be interested in hearing from people who saw this question on the exam. Because the way I read this question - it IS asking for the default format. There's just a line break in the question. The text says "in a format of a XXX@XXXX.com" it's just that someone with less command of English put a space between the "a" and the "XXX" so the space got turned into a line break.

So I think that if the question is actually the default format, then "A".
upvoted 3 times

 **StudentFromAus** 1 year, 6 months ago

The answer should be B as it's not the default email mask format.
upvoted 1 times

 **Navthing** 1 year, 6 months ago

Selected Answer: A

Both A & B are correct But I will prefer A.
upvoted 2 times

 **NamitSehgal** 1 year, 6 months ago

Sorry A is my proffered way, I can not edit my earlier comment.
upvoted 2 times

You have an Azure Data Lake Storage Gen2 account named adls2 that is protected by a virtual network.

You are designing a SQL pool in Azure Synapse that will use adls2 as a source.

What should you use to authenticate to adls2?

- A. an Azure Active Directory (Azure AD) user
- B. a shared key
- C. a shared access signature (SAS)
- D. a managed identity

Correct Answer: D

Managed Identity authentication is required when your storage account is attached to a VNet.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-examples>

⊕ **ploer** Highly Voted 1 year, 11 months ago

Selected Answer: D

D is the way we do it in our company. So it works at least.

upvoted 9 times

⊕ **PallaviPatel** Highly Voted 2 years ago

the answer and explanation given is correct.

upvoted 8 times

⊕ **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: D

Vnet = managed identity

upvoted 2 times

⊕ **vctrhugo** 6 months, 3 weeks ago

Selected Answer: D

VNet = managed identity

upvoted 4 times

⊕ **janaki** 7 months, 2 weeks ago

Vnet = managed identity

upvoted 3 times

⊕ **yogiazaad** 12 months ago

Answer is correct.

The blow link has more details.

"Analytic capabilities such as Dedicated SQL pool and Serverless SQL pool use multi-tenant infrastructure that is not deployed into the managed virtual network. In order for traffic from these capabilities to access the secured storage account, you must configure access to your storage account based on the workspace's system-assigned managed identity by following the steps below."

<https://learn.microsoft.com/en-us/azure/synapse-analytics/security/connect-to-a-secure-storage-account#grant-your-azure-synapse-workspace-access-to-your-secure-storage-account-as-a-trusted-azure-service>

upvoted 2 times

⊕ **Deeksha1234** 1 year, 5 months ago

Selected Answer: D

yes, correct

upvoted 3 times

⊕ **EmmettBrown** 1 year, 9 months ago

Selected Answer: D

Managed identity is correct

upvoted 3 times

⊕ **anto69** 1 year, 11 months ago

I too I think is correct, anyway for sure it's possible

upvoted 4 times

⊕ **bad_atitude** 2 years ago

I believe so

upvoted 4 times

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

HOTSPOT -

You have an Azure Synapse Analytics SQL pool named Pool1. In Azure Active Directory (Azure AD), you have a security group named Group1.

You need to control the access of Group1 to specific columns and rows in a table in Pool1.

Which Transact-SQL commands 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

To control access to the columns:

| |
|-------------------------------|
| CREATE CRYPTOGRAPHIC PROVIDER |
| CREATE PARTITION FUNCTION |
| CREATE SECURITY POLICY |
| GRANT |

To control access to the rows:

| |
|-------------------------------|
| CREATE CRYPTOGRAPHIC PROVIDER |
| CREATE PARTITION FUNCTION |
| CREATE SECURITY POLICY |
| GRANT |

Correct Answer:

Answer Area

To control access to the columns:

| |
|-------------------------------|
| CREATE CRYPTOGRAPHIC PROVIDER |
| CREATE PARTITION FUNCTION |
| CREATE SECURITY POLICY |
| GRANT |

To control access to the rows:

| |
|-------------------------------|
| CREATE CRYPTOGRAPHIC PROVIDER |
| CREATE PARTITION FUNCTION |
| CREATE SECURITY POLICY |
| GRANT |

Box 1: GRANT -

You can implement column-level security with the GRANT T-SQL statement. With this mechanism, both SQL and Azure Active Directory (Azure AD) authentication are supported.

Box 2: CREATE SECURITY POLICY -

Implement RLS by using the CREATE SECURITY POLICY Transact-SQL statement, and predicates created as inline table-valued functions.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security> <https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security>

□  **RajBathani**  2 years ago

Correct Answer

upvoted 21 times

□  **HaBroNounen**  2 years ago

Answer is correct.

for Row LLevel Security: <https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver15>

For Column Level Security: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security>

upvoted 12 times

□  **janaki** 7 months, 2 weeks ago

You are correct! :-)

upvoted 2 times

□  **kkk5566**  4 months, 1 week ago

Row = CREATE SECURITY POLICY

Column = GRANT

upvoted 3 times

□  **vctrhugo** 6 months, 3 weeks ago

Row = CREATE SECURITY POLICY

Column = GRANT

upvoted 3 times

□  **vrodriguesp** 11 months ago

correct, as documentation claims:

to control access to the columns)-->Implement RLS by using the CREATE SECURITY POLICY Transact-SQL statement, and predicates created as inline table-valued functions.

to control access to the rows) -->You can implement column-level security with the GRANT T-SQL statement. With this mechanism, both SQL and Azure Active Directory (Azure AD) authentication are supported.

upvoted 3 times

□  **vctrhugo** 6 months, 2 weeks ago

It should be swaped.

Row = CREATE SECURITY POLICY

Column = GRANT

upvoted 3 times

□  **Deeksha1234** 1 year, 5 months ago

correct

upvoted 2 times

□  **Remedios79** 1 year, 6 months ago

Correct!

upvoted 1 times

□  **juanlu46** 1 year, 8 months ago

Totally correct!

upvoted 1 times

HOTSPOT -

You need to implement an Azure Databricks cluster that automatically connects to Azure Data Lake Storage Gen2 by using Azure Active Directory (Azure AD) integration.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Tier:

| | |
|----------|--|
| Premium | |
| Standard | |

Advanced option to enable:

| | |
|--|--|
| Azure Data Lake Storage Credential Passthrough | |
| Table Access Control | |

Correct Answer:

Answer Area

Tier:

| | |
|----------|--|
| Premium | |
| Standard | |

Advanced option to enable:

| | |
|--|--|
| Azure Data Lake Storage Credential Passthrough | |
| Table Access Control | |

Box 1: Premium -

Credential passthrough requires an Azure Databricks Premium Plan

Box 2: Azure Data Lake Storage credential passthrough

You can access Azure Data Lake Storage using Azure Active Directory credential passthrough.

When you enable your cluster for Azure Data Lake Storage credential passthrough, commands that you run on that cluster can read and write data in Azure Data

Lake Storage without requiring you to configure service principal credentials for access to storage.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough>

ANath Highly Voted 1 year, 11 months ago

Correct

upvoted 14 times

HaBroNounen Highly Voted 2 years ago

Provided answer is correct

<https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough>

upvoted 7 times

kkk5566 Most Recent 4 months ago

correct

upvoted 1 times

Matt2000 5 months ago

The question seems outdated. Credential passthrough is a legacy data governance model. Use unity catalogue instead.
Ref: <https://learn.microsoft.com/en-us/azure/databricks/data-governance/credential-passthrough/adls-passthrough>
upvoted 1 times

✉ **vctrhugo** 6 months, 2 weeks ago

Azure Active Directory credential passthrough requires a Premium plan.

<https://learn.microsoft.com/en-us/azure/databricks/data-governance/credential-passthrough/adls-passthrough#--requirements>
upvoted 2 times

✉ **anks84** 1 year, 4 months ago

Given Answer is correct

upvoted 3 times

✉ **Deeksha1234** 1 year, 5 months ago

correct

upvoted 4 times

✉ **edba** 2 years ago

I think answer is correct!

upvoted 3 times

You are designing an Azure Synapse solution that will provide a query interface for the data stored in an Azure Storage account. The storage account is only accessible from a virtual network.

You need to recommend an authentication mechanism to ensure that the solution can access the source data.

What should you recommend?

- A. a managed identity
- B. anonymous public read access
- C. a shared key

Correct Answer: A

Managed Identity authentication is required when your storage account is attached to a VNet.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-examples>

 **PallaviPatel** Highly Voted 2 years ago

correct

upvoted 11 times

 **Jerrie86** Highly Voted 11 months, 2 weeks ago

Whenever you see Vnet , answer is usually managed Identity

upvoted 5 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: A

correct

upvoted 1 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: A

Managed Idendity = VNET

upvoted 2 times

 **orionduo** 6 months, 2 weeks ago

Selected Answer: A

correct

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: A

A. A managed identity: By assigning a managed identity to the Azure Synapse solution, you can enable it to authenticate and access the Azure Storage account securely. The managed identity acts as a service principal and provides a way to authenticate to Azure services without the need for explicit credentials. By granting the managed identity appropriate permissions on the Azure Storage account, the solution can access the data while ensuring security and avoiding the need for storing and managing explicit credentials.

B. Anonymous public read access is not recommended in this scenario as it would expose the data publicly without any authentication, which can lead to unauthorized access.

C. A shared key is not recommended in this scenario as it involves managing and distributing the storage account's access keys, which can be cumbersome, less secure, and not ideal for scenarios where the storage account is only accessible from a virtual network.

upvoted 2 times

 **anks84** 1 year, 4 months ago

Correct, Managed Identity authentication is required when your storage account is attached to a VNet.

upvoted 4 times

 **Deeksha1234** 1 year, 5 months ago

correct

upvoted 1 times

 **ravi2931** 1 year, 9 months ago

Correct

upvoted 1 times

 **alex1491** 1 year, 9 months ago

the key here is virtual network. Correct!

upvoted 1 times

 **ANath** 1 year, 11 months ago

Correct

upvoted 3 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

You are developing an application that uses Azure Data Lake Storage Gen2.

You need to recommend a solution to grant permissions to a specific application for a limited time period.

What should you include in the recommendation?

- A. role assignments
- B. shared access signatures (SAS)
- C. Azure Active Directory (Azure AD) identities
- D. account keys

Correct Answer: B

A shared access signature (SAS) provides secure delegated access to resources in your storage account. With a SAS, you have granular control over how a client can access your data. For example:

What resources the client may access.

What permissions they have to those resources.

How long the SAS is valid.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/storage-sas-overview>

 **bad_atitude** Highly Voted  2 years ago

Agree with the answer => B

upvoted 18 times

 **kkk5566** Most Recent  4 months, 1 week ago

Selected Answer: B

correct

upvoted 1 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: B

Correct.

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: B

You are developing an application that uses Azure Data Lake Storage Gen2.

You need to recommend a solution to grant permissions to a specific application for a limited time period.

What should you include in the recommendation?

- A. role assignments
- B. shared access signatures (SAS)
- C. Azure Active Directory (Azure AD) identities
- D. account keys

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

correct

upvoted 2 times

 **Remedios79** 1 year, 6 months ago

the key here is "limited time period", so SAS.

upvoted 3 times

 **juanlu46** 1 year, 8 months ago

Selected Answer: B

Correct!

upvoted 4 times

HOTSPOT -

You use Azure Data Lake Storage Gen2 to store data that data scientists and data engineers will query by using Azure Databricks interactive notebooks. Users will have access only to the Data Lake Storage folders that relate to the projects on which they work.

You need to recommend which authentication methods to use for Databricks and Data Lake Storage to provide the users with the appropriate access. The solution must minimize administrative effort and development effort.

Which authentication method should you recommend for each Azure service? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Databricks:

| | |
|---|--------------------------|
| Azure Active Directory credential passthrough | <input type="checkbox"/> |
| Azure Key Vault secrets | <input type="checkbox"/> |
| Personal access tokens | <input type="checkbox"/> |

Data Lake Storage:

| | |
|---|--------------------------|
| Azure Active Directory credential passthrough | <input type="checkbox"/> |
| Shared access keys | <input type="checkbox"/> |
| Shared access signatures | <input type="checkbox"/> |

Answer Area

Databricks:

| | |
|---|--------------------------|
| Azure Active Directory credential passthrough | <input type="checkbox"/> |
| Azure Key Vault secrets | <input type="checkbox"/> |
| Personal access tokens | <input type="checkbox"/> |

Correct Answer:

Data Lake Storage:

| | |
|---|--------------------------|
| Azure Active Directory credential passthrough | <input type="checkbox"/> |
| Shared access keys | <input type="checkbox"/> |
| Shared access signatures | <input type="checkbox"/> |

Box 1: Personal access tokens -

You can use storage shared access signatures (SAS) to access an Azure Data Lake Storage Gen2 storage account directly. With SAS, you can restrict access to a storage account using temporary tokens with fine-grained access control.

You can add multiple storage accounts and configure respective SAS token providers in the same Spark session.

Box 2: Azure Active Directory credential passthrough

You can authenticate automatically to Azure Data Lake Storage Gen1 (ADLS Gen1) and Azure Data Lake Storage Gen2 (ADLS Gen2) from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that you use to log into Azure Databricks. When you enable your cluster for Azure Data Lake

Storage credential passthrough, commands that you run on that cluster can read and write data in Azure Data Lake Storage without requiring you to configure service principal credentials for access to storage.

After configuring Azure Data Lake Storage credential passthrough and creating storage containers, you can access data directly in Azure Data Lake Storage

Gen1 using an adl:// path and Azure Data Lake Storage Gen2 using an abfss:// path:

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/data/data-sources/azure/adls-gen2/azure-datalake-gen2-sas-access>

<https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough>

 **ItHYMeRish**  2 years ago

Accessing the ADLS via Databricks should be using Azure Active Directory with Passthrough. Accessing the files in ADLS should be SAS, based on the options provided.

The explanation provided for this question is incorrect.

upvoted 58 times

 **Billybob0604** 1 year ago

This is it. Correct

upvoted 2 times

 **edba** 1 year, 11 months ago

To be more clear, for box it shall be user delegation SAS which is secured with ADD credentials.

upvoted 2 times

 **vivekazure**  1 year, 12 months ago

1. Accessing the Databricks should be using Personal Tokens
2. Accessing the ADLS should be using Shared Access Signatures. (Because of controlled access to project folders they work).
upvoted 14 times

 **kkk5566**  4 months, 1 week ago

box1 Azure Active Directory with Passthrough
box2 SAS
upvoted 3 times

 **Ram9198** 6 months, 1 week ago

Box 1 - Pass through Databricks
Box 2 - SAS - DL Gen 2
upvoted 3 times

 **auwia** 6 months, 2 weeks ago

Databricks: Azure Active Directory credential passthrough or personal access tokens.

Data Lake Storage: Azure Active Directory credential passthrough.

Please note that while shared access keys and shared access signatures are valid authentication methods for Data Lake Storage, they do not meet the requirement of minimizing administrative effort and providing granular access control based on projects in this scenario.

upvoted 1 times

 **gogosgh** 8 months, 1 week ago

I think the answers given are correct. The question is which authentication to use "for" Databricks and Gen2. So we look at authenticating for (or "into") either of them. The question then becomes which authentication can you use to access databricks and then through that which authentication can you use to authenticate for gen2?

upvoted 1 times

 **JG1984** 6 months, 3 weeks ago

Personal Access Tokens are an alternative authentication method for Azure Databricks that can be used to authenticate to the Databricks REST API and to access Databricks resources. While PATs can provide a high level of security, they require more administrative effort to manage and maintain than Azure Active Directory Credential Passthrough.

upvoted 1 times

 **OldSchool** 1 year, 1 month ago

As we need to access Databricks via ADLS use Azure Databricks access tokens or AAD tokens as explained here: <https://learn.microsoft.com/en-us/azure/databricks/dev-tools/api/latest/aad/>

Data Lake Storage with Passthrough

upvoted 1 times

 **Pais** 1 year, 1 month ago

Both should be Azure Active Directory with Passthrough

1. Shared Key and SAS authorization grants access to a user (or application) without requiring them to have an identity in Azure Active Directory (Azure AD). With these two forms of authentication, Azure RBAC and ACLs have no effect.

ACLs let you grant "fine-grained" access, such as write access to a specific directory or file.

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model>

Azure AD provides superior security and ease of use over Shared Key for authorizing requests to Blob storage. For more information, see Authorize access to data in Azure Storage.

<https://learn.microsoft.com/en-us/azure/storage/blobs/security-recommendations>

2. Azure AD Passthrough will ensure a user can only access the data that they have previously been granted access to via Azure AD in ADLS Gen2.

<https://www.databricks.com/blog/2019/10/24/simplify-data-lake-access-with-azure-ad-credential-passthrough.html>

upvoted 6 times

 **KR8055** 1 year, 2 months ago

Databricks- Azure Active Directory with Passthrough

<https://learn.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough>

Data Lake Storage - SAS

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model>

upvoted 4 times

 **sunil_smile** 1 year, 3 months ago

the question is about how to authenticate the ADLS gen2 dataset both in Databricks and ADLSGen2... Its not about how you authenticate the Databricks.

- 1) Credential Pass through
 - 2) SAS
- upvoted 5 times

✉  **vrodriguesp** 11 months, 2 weeks ago

I agree with you, plus looking at the definitions here:

-) SAS = A shared access signature provides secure delegated access to resources in your storage account. With a SAS, you have granular control over how a client can access your data
-) Azure Active Directory with Passthrough = Credential passthrough allows you to authenticate automatically to Azure Data Lake Storage from Azure Databricks clusters using the identity that you use to log in to Azure Databricks.
-) Shared Access Key = Access keys give you full rights to everything in your storage account

The more explicit question will be:

Which authentication method should you recommend for each Azure service to provide the users with the appropriate access?

- 1) how to authenticate the ADLS gen2 dataset using databricks? ---> Credential Pass through
 - 2) how to authenticate the ADLS gen2 dataset using Data Lake Storage? ---> SAS
- upvoted 1 times

✉  **vrodriguesp** 11 months, 2 weeks ago

Sorry but I missed completely one definition:

-) personal acces token = Personal Access Tokens (PATs) can be used to authenticate to the Databricks REST API, allowing for programmatic access to your Databricks workspace

So by using a PAT, you can automate data movements between Databricks and Data Lake Storage Gen 2 and control user permission to appropriate access

Correct answer should be:

- 1) how to authenticate the ADLS gen2 dataset using databricks? ---> personal acces token
 - 2) how to authenticate the ADLS gen2 dataset using Data Lake Storage? ---> SAS
- upvoted 1 times

✉  **Deeksha1234** 1 year, 5 months ago

Given answer seems correct, agree with HaBroNounen's explanation

upvoted 1 times

✉  **vishal10** 1 year, 5 months ago

Azure Data Lake Storage Gen2 also supports Shared Key and SAS methods for authentication.

To authenticate to and access Databricks REST APIs, you can use Azure Databricks personal access tokens or Azure Active Directory (Azure AD) tokens

upvoted 2 times

✉  **luis1220** 1 year, 5 months ago

It is not mentioning REST API, so it is not personal tokens. I think a normal user will log in databricks using the Active directory. Also, databricks will use Active directory passthrough to use ADLS gen2. Of course, ACLs will be needed to restrict to the folder level which is compatible to the answer.

upvoted 1 times

✉  **HaBroNounen** 2 years ago

Access Databricks with personal access tokens:

<https://docs.databricks.com/dev-tools/api/latest/authentication.html>

Access ADLS from Databricks with Credential Passthrough:

<https://databricks.com/de/blog/2019/10/24/simplify-data-lake-access-with-azure-ad-credential-passthrough.html>

upvoted 10 times

✉  **elahi** 1 week, 6 days ago

Azure Databricks is different from Databricks:

based on this link: <https://learn.microsoft.com/en-us/azure/databricks/security/auth-authz/>

so the answer is Azure Active Directory

upvoted 1 times

✉  **Canary_2021** 2 years ago

Question 1: I select B 'Azure Key Vault secrets'

A: credential passthrough let you access ADLS Gen1 and Gen 2 using same login as Databricks.

B: Key Vault secrets can create a shared login to Databricks. In this way, you don't need to create diff login for diff user any more.

C. Personal access tocks is special for Databricks rest API call. For this question, data scientists and data engineers will query by using Azure Databricks interactive notebooks. So I don't select C.

Question 2: I select A 'Azure Active Directory credential passthrough'. The answer is correct.

upvoted 1 times

✉  **Canary_2021** 2 years ago

Correct my answer.

Question 1: A

Access ADLS Gen2 from Databricks by running query interactively from notebooks.

Question 2: C 'Shared access signatures'

Users also need directly access to the Data Lake Storage for specific folders.

upvoted 2 times

 **tony4fit** 2 years ago

The answer is correct. personal token is the default authentication method for databricks. <https://docs.microsoft.com/en-us/azure/databricks/dev-tools/api/latest/authentication>

upvoted 3 times

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Contacts. Contacts contains a column named Phone. You need to ensure that users in a specific role only see the last four digits of a phone number when querying the Phone column. What should you include in the solution?

- A. table partitions
- B. a default value
- C. row-level security (RLS)
- D. column encryption
- E. dynamic data masking

Correct Answer: E

Dynamic data masking helps prevent unauthorized access to sensitive data by enabling customers to designate how much of the sensitive data to reveal with minimal impact on the application layer. It's a policy-based security feature that hides the sensitive data in the result set of a query over designated database fields, while the data in the database is not changed.

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview>

 **PallaviPatel** Highly Voted 2 years ago

correct

upvoted 13 times

 **ANath** Highly Voted 1 year, 11 months ago

Correct

upvoted 5 times

 **74gjd_37** Most Recent 3 months, 4 weeks ago

Selected Answer: C

The correct answer is row level security ("to allow specific roles")

See <https://learn.microsoft.com/en-us/azure/data-explorer/kusto/management/rowlevelsecuritypolicy>

More use cases

A call center support person may identify callers by several digits of their social security number. This number shouldn't be fully exposed to the support person. An RLS policy can be applied on the table to mask all but the last four digits of the social security number in the result set of any query.

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: E

correct

upvoted 1 times

 **borinot** 1 year, 1 month ago

And Topic 3 question 24 is column-level encryption?

upvoted 2 times

 **vrodriguesp** 11 months ago

I think the key is "when querying the Phone column". Column encryption encrypts individual columns of database on db level, instead Dynamic data masking does not store masked data, only display it.

upvoted 3 times

 **Deeksha1234** 1 year, 5 months ago

correct!

upvoted 2 times

 **juanlu46** 1 year, 8 months ago

Selected Answer: E

Correct!

upvoted 4 times

 **wwdba** 1 year, 10 months ago

Correct

upvoted 4 times

店铺：IT认证考试服务

店铺：IT认证考试服务

You are designing database for an Azure Synapse Analytics dedicated SQL pool to support workloads for detecting ecommerce transaction fraud.

Data will be combined from multiple ecommerce sites and can include sensitive financial information such as credit card numbers.

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

Users must be able to identify potentially fraudulent transactions.

-

▫ Users must be able to use credit cards as a potential feature in models.

▫ Users must NOT be able to access the actual credit card numbers.

What should you include in the recommendation?

- A. Transparent Data Encryption (TDE)
- B. row-level security (RLS)
- C. column-level encryption
- D. Azure Active Directory (Azure AD) pass-through authentication

Correct Answer: C

Use Always Encrypted to secure the required columns. You can configure Always Encrypted for individual database columns containing your sensitive data.

Always Encrypted is a feature designed to protect sensitive data, such as credit card numbers or national identification numbers (for example, U.S. social security numbers), stored in Azure SQL Database or SQL Server databases.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/always-encrypted-database-engine>

✉  juanlu46 Highly Voted 1 year, 8 months ago

Selected Answer: C

By discard, is C, you can create a symmetric key to encrypt a data, for example one column, and then use this data as feature of the model <https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/encrypt-a-column-of-data?view=sql-server-ver15>

The other options that not meet the requirements:

- TDE encrypt data, but decrypt when you query <https://docs.microsoft.com/en-us/azure/azure-sql/database/transparent-data-encryption-tde-overview?tabs=azure-portal>
- RLS is for row restriction, not meet the requirement
- Azure AD pass-through is for authentication

upvoted 15 times

✉  kkk5566 Most Recent 4 months, 1 week ago

Selected Answer: C

C. Column-level encryption

upvoted 1 times

✉  _Lukas_ 5 months, 3 weeks ago

C. Column-level encryption

Explanation:

The given requirement is to enable users to utilize credit card data for model features but to not have access to the actual credit card numbers. Column-level encryption serves this purpose best as it allows for specific columns (in this case, the credit card number column) to be encrypted, while still enabling operations on the data.

A. Transparent Data Encryption (TDE): This encrypts the physical files of the database, but not specific columns. It doesn't fit the requirement here.

B. Row-level security (RLS): This restricts data access at the row level based on certain filters. It doesn't offer column-specific security, and thus isn't the best choice here.

D. Azure Active Directory (Azure AD) pass-through authentication: This is an authentication method, not an encryption method. It would not be applicable for protecting specific data within the database.

upvoted 3 times

✉  yogiazaad 12 months ago

Looks like the column level encryption is still in preview.

<https://azure.microsoft.com/en-us/updates/columnlevel-encryption-for-azure-synapse-analytics/>

upvoted 2 times

✉  yogiazaad 12 months ago

IS column level encryption supported on Dedicated SQL Pools? The question is relate to Dedicated Pool?

upvoted 1 times

 Deeksha1234 1 year, 5 months ago

correct

upvoted 1 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

You have an Azure subscription linked to an Azure Active Directory (Azure AD) tenant that contains a service principal named ServicePrincipal1. The subscription contains an Azure Data Lake Storage account named adls1. Adls1 contains a folder named Folder2 that has a URI of <https://adls1.dfs.core.windows.net/> container1/Folder1/Folder2/. ServicePrincipal1 has the access control list (ACL) permissions shown in the following table.

| Resource | Permission |
|------------|------------------|
| container1 | Access – Execute |
| Folder1 | Access – Execute |
| Folder2 | Access – Read |

You need to ensure that ServicePrincipal1 can perform the following actions:

- Traverse child items that are created in Folder2.
- Read files that are created in Folder2.

The solution must use the principle of least privilege.

Which two permissions should you grant to ServicePrincipal1 for Folder2? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Access "Read"
- B. Access "Write"
- C. Access "Execute"
- D. Default "Read"
- E. Default "Write"
- F. Default "Execute"

Correct Answer: DF

Execute (X) permission is required to traverse the child items of a folder.

There are two kinds of access control lists (ACLs), Access ACLs and Default ACLs.

Access ACLs: These control access to an object. Files and folders both have Access ACLs.

Default ACLs: A "template" of ACLs associated with a folder that determine the Access ACLs for any child items that are created under that folder. Files do not have Default ACLs.

Reference:

<https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control>

 **kl8585** Highly Voted 1 year, 1 month ago

Selected Answer: CD

Phrased different, the question for me says: if you create "Folder3" inside Folder2, you should be able to read files created in Folder3.

This means that you for sure need Executive and Read permissions to Folder2 (Executive to traverse child folder, read to read the files).

Now, starting from the least privilege, suppose you give "Access" permission both for read and execute. In this case, you can't read files created in Folder3. This is a requirement ("child items that are created in Folder2"), so you need Default Read access.

You don't need Default Execute, otherwise you would have access to a Folder created in Folder3 (say Folder 4) and this is not required so for the least privilege you must give Access Execute and not Default Execute.

upvoted 16 times

 **yogiazaad** 12 months ago

Requirement 1 says Traverse child items that are created in Folder2. Means that you need to be able to traverse the subFolders under Folder2.
So Default:Execute is a required permission.

upvoted 2 times

 **bokLuci** Highly Voted 1 year, 2 months ago

Selected Answer: CD

C - You need to traverse the Folder2 only and no potential children folders - Principle of least privilege.

D- You need to pass on the READ access to the files in Folder2. Default ACLs are not passed to files but we are not setting the permission on a file level, we are setting it on Folder2.

upvoted 9 times

 **MarkJoh** Most Recent 6 days, 11 hours ago

Selected Answer: AF

I'm going with AF and here is why.
The requirement "Traverse child items that are created in Folder2" -> This requires default execute so that if any child folders under folder2 get created, the user can list those folders and files.
Now, because of principle of least privilege, it does NOT say that if a file is created under a subfolder (like folder2/folder2/file1.json) that they need access to it.
So, it should be Access Read on folder2 so that the users only get read access to the files in folder2 and not in /folder2/folder3/*.json, for instance.
upvoted 1 times

 **jasmd2** 1 week, 3 days ago

Selected Answer: DF

Default Execute and Default Read as you don't know in advance the files/folder to be created, and you need to access to all of them.
upvoted 1 times

 **kkk5566** 4 months, 1 week ago

"Default - Read" and "Default - Execute"
upvoted 2 times

 **Ram9198** 4 months, 4 weeks ago

Selected Answer: CD

Traverse require access execute, file reads need default read
upvoted 1 times

 **Ram9198** 6 months, 1 week ago

Selected Answer: DF

Default Execute is mandatory to traverse child items through cascade.. Default Read by process of elimination
upvoted 3 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: AF

⇒ Traverse child items that are created in Folder2. => DEFAULT EXECUTE
⇒ Read files that are created in Folder2. => ACCESS READ (that was already given).
upvoted 2 times

 **esaade** 10 months ago

Selected Answer: DF

Based on the permissions table provided, the ServicePrincipal1 has "Access - Execute" permission on container1, "Access - Execute" permission on Folder1, and "Access - Read" permission on Folder2. To allow ServicePrincipal1 to traverse child items that are created in Folder2 and read files created in Folder2, you should grant the "Default - Read" and "Default - Execute" permissions on Folder2. The "Default - Read" permission allows ServicePrincipal1 to read files created in Folder2, and the "Default - Execute" permission allows ServicePrincipal1 to traverse child items that are created in Folder2.

Therefore, the correct answer is:

- D. Default - Read
 - F. Default - Execute
- upvoted 6 times

 **yogiazzaad** 12 months ago

Traverse child items that are created in Folder2.
This needs Default:Execute Because user needs to traverse any child Items(Sub Folders) created under under Folder2.
Read files that are created in Folder2.
Since the The Access:read ACL is already set on Folder2.Any files that are created under Folder2 can be access by User. But to see (or list) the items/files under Folder2 we need Access:Execute .
SO the answer is Access: Execute and Default: Execute
upvoted 4 times

 **AzureJobsTillRetire** 1 year, 1 month ago

Selected Answer: DF

Default Read and Execute are required. The reason is as below.

In the POSIX-style model that's used by Data Lake Storage Gen2, permissions for an item are stored on the item itself. In other words, permissions for an item cannot be inherited from the parent items if the permissions are set after the child item has already been created. Permissions are only inherited if default permissions have been set on the parent items before the child items have been created.

Reference: <https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

upvoted 6 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: DF

so the answer is correct
upvoted 4 times

 **Deeksha1234** 1 year, 5 months ago

I think the given answer is correct. Since we should be able to traverse and read the child items from the folder 2 .

From one of the DP 203 Microsoft lab exercise -

Access ACLs control access to an object. Files and directories both have access ACLs.

Default ACLs are templates of ACLs associated with a directory that determine the access ACLs for any child items that are created under that directory. Files do not have default ACLs.

upvoted 4 times

 **Davico93** 1 year, 6 months ago

Selected Answer: AF

Default is not related to files so, if we want to read files, we need ACCESS - READ

upvoted 3 times

 **Aditya0891** 1 year, 7 months ago

Please make a note how the sentence is framed "Traverse child items that are created in Folder2". Access ACL doesn't propagate the permissions to child items but default ACL does. So it is obvious that new files or folders can be created in Folder2 and that requires default ACL. So according to me default execute and default read on folder2 should be the correct answer

upvoted 1 times

 **Aditya0891** 1 year, 7 months ago

Please ignore this. It's not correct. Examtopics should provide a delete option here.

upvoted 4 times

 **sdokmak** 1 year, 7 months ago

Following principle of least privilege, isn't Access Execute and Default Read enough? You only need to traverse the files in Folder2, not the folders within Folder2 (even though there aren't any)

upvoted 4 times

 **virendrapsingh** 1 year, 7 months ago

Agreed with your comment on least privilege as it is mentioned specifically in the question.

Choices A & F should be the answer.

upvoted 5 times

 **Aditya0891** 1 year, 7 months ago

sdokmak not sure but it's not mentioned that there are only files inside folder2 and in the next line it specifically mentioned that to read files inside folder 2. I think the answers are correct as per requirement. Please correct me if I'm wrong

upvoted 1 times

 **MadEgg** 1 year, 7 months ago

Selected Answer: DF

Correct

upvoted 1 times

HOTSPOT -

You have an Azure subscription that is linked to a hybrid Azure Active Directory (Azure AD) tenant. The subscription contains an Azure Synapse Analytics SQL pool named Pool1.

You need to recommend an authentication solution for Pool1. The solution must support multi-factor authentication (MFA) and database-level authentication.

Which authentication solution or solutions 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

MFA:

| | |
|-------------------------------------|---|
| Azure AD authentication | ▼ |
| Microsoft SQL Server authentication | ▼ |
| Passwordless authentication | ▼ |
| Windows authentication | ▼ |

Database-level authentication:

| | |
|-----------------------------|---|
| Application roles | ▼ |
| Contained database users | ▼ |
| Database roles | ▼ |
| Microsoft SQL Server logins | ▼ |

Answer Area

MFA:

| | |
|-------------------------------------|---|
| Azure AD authentication | ▼ |
| Microsoft SQL Server authentication | ▼ |
| Passwordless authentication | ▼ |
| Windows authentication | ▼ |

Correct Answer:

Database-level authentication:

| | |
|-----------------------------|---|
| Application roles | ▼ |
| Contained database users | ▼ |
| Database roles | ▼ |
| Microsoft SQL Server logins | ▼ |

Box 1: Azure AD authentication -

Azure AD authentication has the option to include MFA.

Box 2: Contained database users -

Azure AD authentication uses contained database users to authenticate identities at the database level.

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/authentication-mfa-ssms-overview> <https://docs.microsoft.com/en-us/azure/azure-sql/database/authentication-aad-overview>

  **Skeinofi**  2 years ago

Correct

upvoted 22 times

  **Amsterliese**  1 year, 9 months ago

"SQL Database and Azure Synapse Analytics support Azure Active Directory identities as contained database users"
<https://docs.microsoft.com/en-us/sql/relational-databases/security/contained-database-users-making-your-database-portable?view=sql-server-ver15#contained-database-user-model>

upvoted 7 times

 **kkk5566** Most Recent 4 months, 1 week ago

Correct

upvoted 1 times

 **JG1984** 6 months, 3 weeks ago

Azure Synapse Analytics supports two types of database-level authentication:

Azure Active Directory (Azure AD) authentication: This uses your Azure AD identity to authenticate to Synapse SQL. This is the recommended authentication method, as it provides a single sign-on experience and allows you to manage permissions using Azure AD groups.

SQL Server authentication: This uses a traditional SQL Server username and password to authenticate to Synapse SQL. This authentication method is less secure than Azure AD authentication, but it may be necessary if you are using legacy applications that do not support Azure AD.

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

answer is correct

upvoted 4 times

 **dev2dev** 1 year, 11 months ago

B is wrong. Contained users not supported by synapse analytics. D is correct ('MS SQL Server logins')

upvoted 4 times

 **PallaviPatel** 1 year, 11 months ago

<https://docs.microsoft.com/en-us/azure/azure-sql/database/authentication-aad-overview> this document says contained users are supported by synapse analytics, so this is correct answer.

upvoted 16 times

 **dev2dev** 1 year, 11 months ago

correct

upvoted 3 times

DRAG DROP -

You have an Azure data factory.

You need to ensure that pipeline-run data is retained for 120 days. The solution must ensure that you can query the data by using the Kusto query language.

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.

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:

| Actions | Answer Area |
|---|--------------------|
| Select the PipelineRuns category. | |
| Create a Log Analytics workspace that has Data Retention set to 120 days. | |
| Stream to an Azure event hub. | |
| Create an Azure Storage account that has a lifecycle policy. | |
| From the Azure portal, add a diagnostic setting. | |
| Send the data to a Log Analytics workspace. | |
| Select the TriggerRuns category. | |

| Actions | Answer Area |
|---|---|
| Select the PipelineRuns category. | Create an Azure Storage account that has a lifecycle policy. |
| Create a Log Analytics workspace that has Data Retention set to 120 days. | Create a Log Analytics workspace that has Data Retention set to 120 days. |
| Stream to an Azure event hub. | From the Azure portal, add a diagnostic setting. |
| Create an Azure Storage account that has a lifecycle policy. | Send the data to a Log Analytics workspace. |
| From the Azure portal, add a diagnostic setting. | |
| Send the data to a Log Analytics workspace. | |
| Select the TriggerRuns category. | |

Step 1: Create an Azure Storage account that has a lifecycle policy

To automate common data management tasks, Microsoft created a solution based on Azure Data Factory. The service, Data Lifecycle

Management, makes frequently accessed data available and archives or purges other data according to retention policies. Teams across the company use the service to reduce storage costs, improve app performance, and comply with data retention policies.

Step 2: Create a Log Analytics workspace that has Data Retention set to 120 days.

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time. With Monitor, you can route diagnostic logs for analysis to multiple different targets, such as a Storage Account: Save your diagnostic logs to a storage account for auditing or manual inspection. You can use the diagnostic settings to specify the retention time in days.

Step 3: From Azure Portal, add a diagnostic setting.

Step 4: Send the data to a log Analytics workspace,

Event Hub: A pipeline that transfers events from services to Azure Data Explorer.

Keeping Azure Data Factory metrics and pipeline-run data.

Configure diagnostic settings and workspace.

Create or add diagnostic settings for your data factory.

1. In the portal, go to Monitor. Select Settings > Diagnostic settings.

2. Select the data factory for which you want to set a diagnostic setting.

3. If no settings exist on the selected data factory, you're prompted to create a setting. Select Turn on diagnostics.

4. Give your setting a name, select Send to Log Analytics, and then select a workspace from Log Analytics Workspace.

5. Select Save.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor>

✉️  **Sunnyb** Highly Voted 2 years, 7 months ago

Step 1: Create a Log Analytics workspace that has Data Retention set to 120 days.

Step 2: From Azure Portal, add a diagnostic setting.

Step 3: Select the PipelineRuns Category

Step 4: Send the data to a Log Analytics workspace.

upvoted 174 times

✉️  **matiandal** 2 months ago

why not the ET Answer?

-- > the ET answer misses the step "Add diagnostic setting-> PipelineRuns option

4 a Confirmation of Sunnyb's answer see the also the following link (with screenshots)

--> <https://davidalzamendi.com/long-running-azure-data-factory-pipelines/>

upvoted 1 times

✉️  **kkk5566** 4 months, 1 week ago

correct

upvoted 2 times

✉️  **datapc** 1 year, 2 months ago

<https://learn.microsoft.com/en-us/azure/azure-monitor/essentials/tutorial-resource-logs?source=recommendations>

Above order is mentioned here.

upvoted 8 times

✉️  **RajashekharC** 1 year, 4 months ago

This is correct order, I have tried this on Azure portal.

upvoted 4 times

✉️  **hercilian_effort** Highly Voted 2 years, 6 months ago

step 1. From Azure Portal, add a diagnostic setting.

step 2. Send data to a Log analytics workspace.

step 3. Create a Log Analytics workspace that has Data Retention set to 120 days.

step 4. Select the PipelineRuns Category.

The video in the below link walks you through the process step by step, start watching at 2min 30sec mark

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor#keeping-azure-data-factory-metrics-and-pipeline-run-data>

upvoted 39 times

✉️  **matiandal** 2 months ago

WRONG.

In order to add the "PipelineRuns" setting you have to follow the steps below:

STEPS: ADF -> "under monitoring group" select Diagnostic Settings -> Add diagnostic Setting -> Select PipelineRuns

For confirmation diy or see the provided link -)

R: <https://davidalzamendi.com/long-running-azure-data-factory-pipelines/>

Cheers !

upvoted 1 times

- **KashRaynardMorse** 1 year, 8 months ago
Read the text surrounding the video; it is for Azure Monitoring which provides only base-level services; of only 45 days. So the video is incorrect, for the question asked.
upvoted 4 times
- **klaytech** 6 months, 1 week ago
45 for Monitoring not Azure Log Analytics
upvoted 1 times
- **Igor85** 1 year, 1 month ago
steps 2 & 3 must be swapped. you can't send data to log analytics workspace that isn't created yet
upvoted 2 times
- **Armandoo** 2 years, 5 months ago
This is the correct answer
upvoted 1 times
- **Sriramiyer92** [Most Recent] 1 year, 5 months ago
Can see multiple answers that are correct in the discussion!
Also note the question states :"More than one order of answer choices is correct"
upvoted 2 times
- **NamitSehgal** 1 year, 6 months ago
Output is either SA, LA or Eventhub
Retention is configured during setting up the diag on any Azure resource , so take out option 1 which says configure SA retention.
Just stick to LA solution and include all the points related to it.
upvoted 1 times
- **[Removed]** 2 years, 4 months ago
I am not very familiar with this topic, but follow the link below, we can know With Monitor, you can route diagnostic logs for analysis to multiple different targets: Storage account, Event Hub and Log Analytics. It also needs to query the data by use Kusto query language, so we can know we should use Log Analytics for this scenario. With this in mind, we can exclude anything related with storage account and Event Hub. Then the question talks about Pipeline runs log, so we can also exclude the Trigger run log one. Then there are 4 options left there as listed in the solution raised by @Sunnyb.
<https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor#keeping-azure-data-factory-metrics-and-pipeline-run-data>
upvoted 12 times
- **Amalbenrebai** 2 years, 4 months ago
in this case we will not use a storage Account to save the diagnostic logs to a storage account, but we will send them to Log Analytics:
1: Create a Log Analytics workspace that has Data Retention set to 120 days.
2: From Azure Portal, add a diagnostic setting.
3: Select the PipelineRuns Category
4: Send the data to a Log Analytics workspace
upvoted 9 times
- **mss1** 2 years, 5 months ago
If you create diagnostics from the Datafactory you will notice that you can only set the retentiondays when you select a storage account for the PipelineRuns. So you need a storage account first. You do not have an option in the selection to create a diagnostic from the datafactory and thus the option "select the pipelineruns" is not an option. I agree with the current selection.
upvoted 2 times
- **mss1** 2 years, 5 months ago
To complete my answer. I also agree with "Sunnyb". There are more solutions to this question.
upvoted 2 times
- **Marcus1612** 2 years, 3 months ago
When you create diagnostic, you have to select "Log Analytics" as destination target. Log Analytics Workspace has its own Data Retention Properties under General/Usage and Estimated Cost/Data Retention. So the good answer is:Step 1: Create a Log Analytics workspace that has Data Retention set to 120 days.
Step 2: From Azure Portal, add a diagnostic setting.
Step 3: Select the PipelineRuns Category
Step 4: Send the data to a Log Analytics workspace.
upvoted 1 times
- **mr1c** 2 years, 6 months ago
According to the linked article, it's: first Storage Account, then Event Hub, and finally Log Analytics.
So I would say:
1- Create an Azure Storage Account with a lifecycle policy
2- Stream to an Azure Event Hub
3- Create a Log Analytics workspace that has a Data Retention set to 120 days
4- Send the data to a Log Analytics Workspace
Source: <https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor#keeping-azure-data-factory-metrics-and-pipeline-run-data>
upvoted 4 times
- **det_wizard** 2 years, 7 months ago
Take off the storage account and After add diagnostic setting it would be select pipelineruns then send to log analytics

upvoted 2 times

✉ **teofz** 2 years, 7 months ago

regarding the storage account, what is it for?!

upvoted 1 times

✉ **sagga** 2 years, 7 months ago

I don't know if you need to, see this discussion: <https://www.examtopics.com/discussions/microsoft/view/49811-exam-dp-200-topic-3-question-19-discussion/>

upvoted 2 times

✉ **Amsterliese** 1 year, 9 months ago

In this case, not needed (imo). MS advises to store log data in a storage account (if needed) since Data Factory only retains it for 45 days. However, in this case you don't have to store it longer than 2 years and you want to use Kusto, so Log Analytics makes more sense.

upvoted 1 times

You have an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that data in the pool is encrypted at rest. The solution must NOT require modifying applications that query the data.

What should you do?

- A. Enable encryption at rest for the Azure Data Lake Storage Gen2 account.
- B. Enable Transparent Data Encryption (TDE) for the pool.
- C. Use a customer-managed key to enable double encryption for the Azure Synapse workspace.
- D. Create an Azure key vault in the Azure subscription grant access to the pool.

Correct Answer: B

Transparent Data Encryption (TDE) helps protect against the threat of malicious activity by encrypting and decrypting your data at rest.

When you encrypt your database, associated backups and transaction log files are encrypted without requiring any changes to your applications. TDE encrypts the storage of an entire database by using a symmetric key called the database encryption key.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-overview-manage-security>

 **damaldon** Highly Voted 2 years, 6 months ago

Correct!

upvoted 39 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: B

correct

upvoted 1 times

 **vctrhugo** 6 months, 2 weeks ago

Selected Answer: B

Transparent Data Encryption (TDE) is a feature provided by Azure SQL Database and Azure Synapse Analytics that encrypts the database files at rest. It performs real-time I/O encryption and decryption of the database files, ensuring that the data is encrypted on disk. TDE operates transparently and does not require any changes to the application code or queries.

By enabling TDE for the dedicated SQL pool in Azure Synapse Analytics, you can achieve encryption at rest for the data stored in the pool without impacting the applications that access the data.

upvoted 1 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: B

B is correct

upvoted 2 times

 **youlitai003** 1 year, 8 months ago

B is right, however using CMK configed at workspace level to achieve double encryption is also right.
<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption>

upvoted 1 times

 **bigw** 1 year, 1 month ago

you can only enable double encryption when you are creating a new workspace.

upvoted 2 times

 **djblue** 1 year, 10 months ago

Selected Answer: B

TDE is used for encrypting data at rest.

upvoted 3 times

DRAG DROP -

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named storage1. Storage1 contains a container named container1.

Container1 contains a directory named directory1. Directory1 contains a file named file1.

You have an Azure Active Directory (Azure AD) user named User1 that is assigned the Storage Blob Data Reader role for storage1.

You need to ensure that User1 can append data to file1. The solution must use the principle of least privilege.

Which permissions should you grant? To answer, drag the appropriate permissions to the correct resources. 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.

Select and Place:

| Permissions | Answer Area |
|--------------------|---|
| Read | container1: <input type="button" value="Permission"/> |
| Write | directory1: <input type="button" value="Permission"/> |
| Execute | file1: <input type="button" value="Permission"/> |

| Permissions | Answer Area |
|--------------------|--|
| Read | container1: <input type="button" value="Execute"/> |
| Write | directory1: <input type="button" value="Execute"/> |
| Execute | file1: <input type="button" value="Write"/> |

Box 1: Execute -

If you are granting permissions by using only ACLs (no Azure RBAC), then to grant a security principal read or write access to a file, you'll need to give the security principal Execute permissions to the root folder of the container, and to each folder in the hierarchy of folders that lead to the file.

Box 2: Execute -

On Directory: Execute (X): Required to traverse the child items of a directory

Box 3: Write -

On file: Write (W): Can write or append to a file.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

✉  **anks84**  1 year, 4 months ago

-Execute
-Execute
-Write
upvoted 12 times

✉  **Matt2000** 5 months ago

Supported by the following two references:

without additional permissions: <https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

with additional permissions such as storage blob data reader: <https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model#permissions-table-combining-azure-rbac-abac-and-acls>

upvoted 1 times

□ **AlviraTony** 4 months, 1 week ago

In the above link, the use case is given for appending to Data.txt file, then the answers would be

- Execute
- Execute
- Read and Write

upvoted 2 times

□ **dom271219** Highly Voted 1 year, 4 months ago

Correct : Execute to traverse the folders and Write to append the file

upvoted 6 times

□ **Ram9198** Most Recent 4 months ago

X X RW need both rw for append

upvoted 3 times

□ **hassexat** 4 months ago

Execute

Execute

Write

The provided answer is correct!

upvoted 1 times

□ **kkk5566** 4 months, 1 week ago

correct

upvoted 1 times

□ **bakamon** 7 months, 2 weeks ago

container1 : Read access [by default because User1 that is assigned the Storage Blob Data Reader role for storage1]

directory1: Execute [since requirement is only to append file1 so traverse (execute) permission will be enough for it]

file1 : Write [because execute cannot append the file in Azure Data Lake Storage Gen2]

only write permission can append a file.

upvoted 1 times

□ **OldSchool** 1 year, 1 month ago

Can't remember if the wording on actual exam was the same or very similar but instead of Append was Delete and the Q was like this:
You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named storage1. Storage1 contains a container named container1.

Container1 contains a directory named directory1. Directory1 contains a file named file1.

You have an Azure Active Directory (Azure AD) user named User1 that is assigned the Storage Blob Data Reader role for storage1.

You need to ensure that User1 can delete file1. The solution must use the principle of least privilege.

Permission:

--WX

---X

Answer Area and my answers:

container1 ---X

directory1 ---X

file1 --WX

upvoted 5 times

□ **mamahani** 8 months, 2 weeks ago

i dont think you gave correct answers;

see this doc: <https://learn.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control#common-scenarios-related-to-permissions>

to delete a file you dont need any permissions on the file itself; only on the folder where it resides (read + execute)

upvoted 1 times

□ **Matt2000** 5 months ago

mamahani is correct. See the following references:

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model#permissions-table-combining-azure-rbac-abac-and-acls>

upvoted 1 times

□ **renukahouse** 5 months, 3 weeks ago

if you give write access to entire folder , the user can delete/modify other folders , whihc is not correct

upvoted 1 times

□ **vctrhugo** 6 months, 2 weeks ago

The solution must use the principle of least privilege. You shouldn't do -WX on folder, only on file.

upvoted 2 times

HOTSPOT -

You have an Azure subscription that contains an Azure Databricks workspace named databricks1 and an Azure Synapse Analytics workspace named synapse1.

The synapse1 workspace contains an Apache Spark pool named pool1.

You need to share an Apache Hive catalog of pool1 with databricks1.

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

NOTE: Each correct selection is worth one point.

Hot Area:

From synapse1, create a linked service to:

| |
|------------------------------|
| Azure Cosmos DB |
| Azure Data Lake Storage Gen2 |
| Azure SQL Database |

Configure pool1 to use the linked service as:

| |
|----------------------------------|
| An Azure Purview account |
| A Hive metastore |
| A managed Hive metastore service |

Correct Answer:

From synapse1, create a linked service to:

| |
|------------------------------|
| Azure Cosmos DB |
| Azure Data Lake Storage Gen2 |
| Azure SQL Database |

Configure pool1 to use the linked service as:

| |
|----------------------------------|
| An Azure Purview account |
| A Hive metastore |
| A managed Hive metastore service |

Box 1: Azure SQL Database -

Use external Hive Metastore for Synapse Spark Pool

Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog.

Set up linked service to Hive Metastore

Follow below steps to set up a linked service to the external Hive Metastore in Synapse workspace.

1. Open Synapse Studio, go to Manage > Linked services at left, click New to create a new linked service.
2. Set up Hive Metastore linked service
3. Choose Azure SQL Database or Azure Database for MySQL based on your database type, click Continue.
4. Provide Name of the linked service. Record the name of the linked service, this info will be used to configure Spark shortly.
5. You can either select Azure SQL Database/Azure Database for MySQL for the external Hive Metastore from Azure subscription list, or enter the info manually.
6. Provide User name and Password to set up the connection.
7. Test connection to verify the username and password.
8. Click Create to create the linked service.

Box 2: A Hive Metastore -

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore>

 **federic** Highly Voted  1 year, 4 months ago

I would say:

1. sql - this is correct
2. managed hive metastore

upvoted 9 times

✉  **federc** 1 year, 4 months ago

scrath that, given answers are correct. sql + hive metastore

upvoted 16 times

✉  **matiandal** Most Recent 2 months ago

b1-sql

b2. managed hive metastore

Why b2 a managed ?

A Hive metastore is a central repository that stores metadata about the data stored in a Hive warehouse. A managed Hive metastore is a type of Hive metastore that is fully managed by Azure Databricks. It provides the following benefits over a self-managed Hive metastore:

It is automatically created and configured when you create a Databricks workspace.

It is automatically backed up and restored by Databricks.

It is automatically scaled and optimized by Databricks.

It is compatible with all Databricks features, such as Delta Lake, SQL Analytics, and Unity Catalog.

A managed Hive metastore is recommended for most use cases, unless you have specific requirements that need a self-managed Hive metastore, such as:

You want to use an external metastore service, such as AWS Glue or Azure SQL Database.

You want to share the same metastore across multiple Databricks workspaces or other applications.

upvoted 1 times

✉  **matiandal** 2 months ago

b2. hive metastore, not a managed ! (sorry)

Why b2 have to be " A Hive metastore" and not a managed one

upvoted 1 times

✉  **kkk5566** 4 months, 1 week ago

1. sql db

2. hive metastore

upvoted 2 times

✉  **andjurovicela** 6 months ago

1 - definitely correct per documentation TestingCRM provided.

2 - I think the devil's in the detail here :/ documentation says "Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog".

The word managed may sway you towards the answer managed hive metasotre SERVICE but the docs don't mention "service" at all, which is why I would go with Hive metastore

upvoted 2 times

✉  **TestingCRM** 7 months, 1 week ago

1. sql - this is correct

2. managed hive metastore

See <https://learn.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore>

upvoted 1 times

HOTSPOT -

You have an Azure subscription.

You need to deploy an Azure Data Lake Storage Gen2 Premium account. The solution must meet the following requirements:

- * Blobs that are older than 365 days must be deleted.
- * Administrative effort must be minimized.
- * Costs must be minimized.

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

NOTE: Each correct selection is worth one point.

Hot Area:

To minimize costs:

Locally-redundant storage (LRS)
The Archive access tier
The Cool access tier
Zone-redundant storage (ZRS)

To delete blobs:

Azure Automation runbooks
Azure Storage lifecycle management
Soft delete

Correct Answer:

To minimize costs:

Locally-redundant storage (LRS)
The Archive access tier
The Cool access tier
Zone-redundant storage (ZRS)

To delete blobs:

Azure Automation runbooks
Azure Storage lifecycle management
Soft delete

Box 1: The Archive access tier -

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 Storage lifecycle management

With the lifecycle management policy, you can:

- * Delete current versions of a blob, previous versions of a blob, or blob snapshots at the end of their lifecycles.

Transition blobs from cool to hot immediately when they're accessed, to optimize for performance.

Transition current versions of a blob, previous versions of a blob, or blob snapshots to a cooler storage tier if these objects haven't been accessed or modified for a period of time, to optimize for cost. In this scenario, the lifecycle management policy can move objects from hot to cool, from hot to archive, or from cool to archive.

Etc.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview> <https://docs.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-overview>

□  **goxxx** Highly Voted 1 year, 3 months ago

If u choose premium storage account, there is no possibility to choose tiers (hot, cool, archive), its always hot, sa LRS and lifecycle storage mgmt
upvoted 40 times

□  **mamahani** 8 months, 2 weeks ago

its not the same as hot; see this microsoft article: <https://azure.microsoft.com/nl-nl/blog/azure-premium-block-blob-storage-is-now-generally-available/>
"Premium Blob Storage is a new performance tier in Azure Blob Storage for block blobs and append blobs, complimenting the existing Hot, Cool, and Archive access tiers."

upvoted 1 times

□  **allagowf** 1 year, 2 months ago

Agree no mention for tiering in the question so LRS is the best option to minimize the cost
upvoted 4 times

□  **dom271219** Highly Voted 1 year, 4 months ago

The statement doesn't mention requirement for a tiering storage archive nor cool nor hot before deletion.
Then I think it is LRS and lifecycle storage mgmt
upvoted 15 times

□  **Momoanwar** Most Recent 2 weeks, 2 days ago

Chatgpt :
To minimize costs, select **The Archive access tier** since it is optimized for data that is rarely accessed and offers the lowest storage cost. For the deletion of blobs older than 365 days, you would use **Azure Storage lifecycle management** to automate the deletion process, reducing administrative effort.
upvoted 1 times

□  **AlfredPennyworth** 3 weeks, 3 days ago

For your Azure Data Lake Storage Gen2 Premium account, considering the requirements:

To minimize costs: Locally-redundant storage (LRS). This is cost-effective and provides high durability within a single region.

To delete blobs older than 365 days: Azure Automation runbooks. Since Azure Storage lifecycle management isn't applicable to Premium tier, automation runbooks can be used to programmatically delete older blobs, minimizing administrative effort.
upvoted 1 times

□  **hassexat** 4 months ago

LRS & Lifecycle
upvoted 2 times

□  **kkk5566** 4 months, 1 week ago

LRS and LCM
upvoted 1 times

□  **Ram9198** 4 months, 4 weeks ago

As per the response from the Microsoft <https://github.com/MicrosoftDocs/azure-docs/issues/100695> tiering is not supported for premium but delete through LCM is supported.. but still not clearly mentioned in this document <https://learn.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-overview>

Answer LRS and LCM
upvoted 1 times

□  **pavankr** 6 months, 1 week ago

Why you want to "Archive"???
upvoted 1 times

□  **vctrhugo** 6 months, 2 weeks ago

LRS and data lifecycle. Even tho you can't switch data from tier-to-tier, you can still apply a rule to delete the BLOB once it reaches 365 days.
upvoted 2 times

□  **BPW** 7 months, 3 weeks ago

According to
<https://learn.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview?tabs=azure-portal>
"Data stored in a premium block blob storage account cannot be tiered to hot, cool, cold or archive by using Set Blob Tier or using Azure Blob

Storage lifecycle management."
So answers are LRS and Soft delete
<https://learn.microsoft.com/en-us/azure/storage/blobs/soft-delete-blob-overview>
<https://learn.microsoft.com/en-us/azure/storage/blobs/soft-delete-container-enable?tabs=azure-portal>
upvoted 2 times

□ **JG1984** 6 months, 3 weeks ago

Azure Blob Storage Lifecycle Management allows you to create rules to automatically delete blobs based on their age, reducing administrative effort and minimizing costs. This makes it a better option for meeting the requirements specified in your scenario.
Soft delete is an option for protecting against accidental deletion of blobs, but it is not the best option for automatically deleting blobs that are older than 365 days. Soft delete works by retaining deleted blobs for a specified period of time, allowing you to recover them if needed. However, it does not automatically delete blobs based on their age.

upvoted 1 times

□ **mamahani** 8 months, 2 weeks ago

according to microsoft: "Premium Blob Storage is a new performance tier in Azure Blob Storage for block blobs and append blobs, complimenting the existing Hot, Cool, and Archive access tiers."
<https://azure.microsoft.com/nl-nl/blog/azure-premium-block-blob-storage-is-now-generally-available/>
so the only two other options left are LRS and ZRS; LRS is cheaper; so it must be this one;

upvoted 1 times

□ **mamahani** 8 months, 2 weeks ago

also in the documentation all the three tiers are greyed out for premium
<https://learn.microsoft.com/en-us/azure/storage/blobs/storage-feature-support-in-storage-accounts#premium-block-blob-accounts>
so you cannot possibly choose this as an answer;

upvoted 1 times

□ **youngbug** 1 year ago

I strongly doubt they didn't offer the whole question. The question is not clear.

upvoted 1 times

□ **AzureJobsTillRetire** 1 year, 1 month ago

Box1: Locally-redundant storage (LRS)
In the question, it specifically states that "You need to deploy an Azure Data Lake Storage Gen2 Premium account", and Azure Data Lake Storage Gen2 premium tier is neither an Archive access tier nor a Cool Access tier, and so those two options are out. Locally-redundant storage (LRS) is less expensive than Zone-redundant storage (ZRS), so we choose LRS.
<https://learn.microsoft.com/en-us/azure/storage/blobs/premium-tier-for-data-lake-storage>

Box2: Azure Storage Lifecycle management

Well explained in the answer already.

upvoted 7 times

HOTSPOT -

You are designing an application that will use an Azure Data Lake Storage Gen 2 account to store petabytes of license plate photos from toll booths. The account will use zone-redundant storage (ZRS).

You identify the following usage patterns:

- * The data will be accessed several times a day during the first 30 days after the data is created. The data must meet an availability SLA of 99.9%.

- * After 90 days, the data will be accessed infrequently but must be available within 30 seconds.

- * After 365 days, the data will be accessed infrequently but must be available within five minutes.

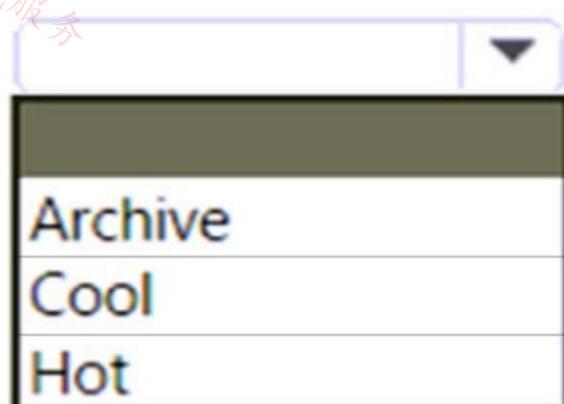
You need to recommend a data retention solution. The solution must minimize costs.

Which access tier should you recommend for each time frame? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

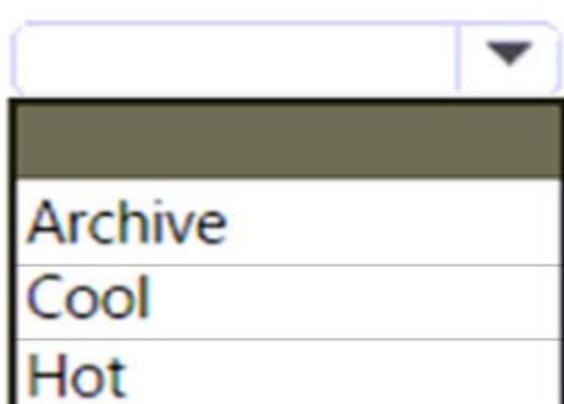
First 30 days:



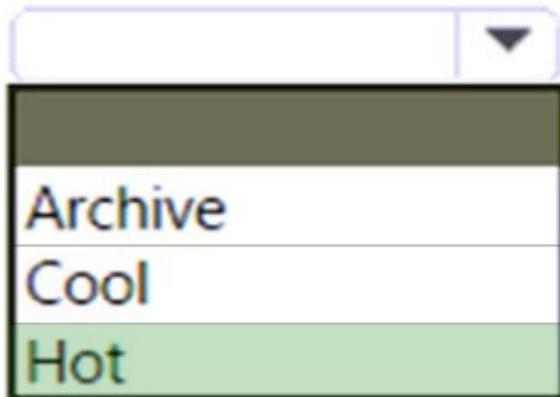
After 90 days:



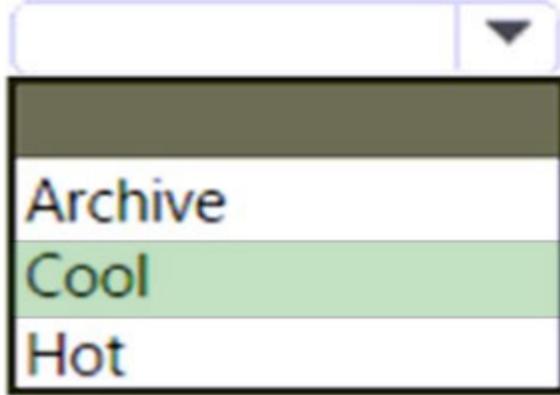
After 365 days:



First 30 days:

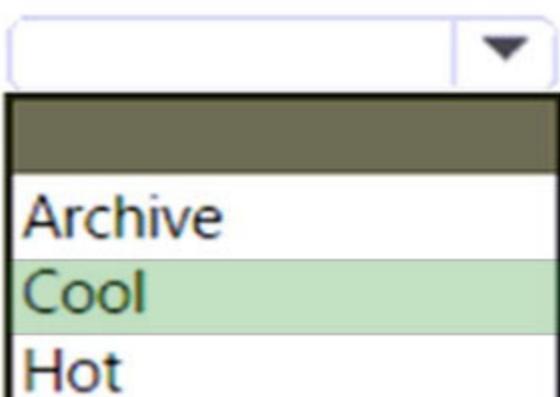


After 90 days:



Correct Answer:

After 365 days:



Box 1: Hot -

The data will be accessed several times a day during the first 30 days after the data is created. The data must meet an availability SLA of 99.9%.

Box 2: Cool -

After 90 days, the data will be accessed infrequently but must be available within 30 seconds.

Data in the Cool tier should be stored for a minimum of 30 days.

When your data is stored in an online access tier (either Hot or Cool), users can access it immediately. The Hot tier is the best choice for data that is in active use, while the Cool tier is ideal for data that is accessed less frequently, but that still must be available for reading and writing.

Box 3: Cool -

After 365 days, the data will be accessed infrequently but must be available within five minutes.

Incorrect:

Not Archive:

While a blob is in the Archive access tier, it's considered to be offline and can't be read or modified. In order to read or modify data in an archived blob, you must first rehydrate the blob to an online tier, either the Hot or Cool tier.

Rehydration priority -

When you rehydrate a blob, you can set the priority for the rehydration operation via the optional `x-ms-rehydrate-priority` header on a Set Blob Tier or Copy Blob operation. Rehydration priority options include:

Standard priority: The rehydration request will be processed in the order it was received and may take up to 15 hours.

High priority: The rehydration request will be prioritized over standard priority requests and may complete in less than one hour for objects under 10 GB in size.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview> <https://docs.microsoft.com/en-us/azure/storage/blobs/archive-rehydrate-overview>

OdogwuSaina Highly Voted 11 months, 3 weeks ago

Hot, Cool, Cool is correct.

Ref: <https://learn.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview>

upvoted 9 times

AlfredPennyworth Most Recent 3 weeks, 3 days ago

First 30 Days: Use the Hot tier for frequent access and meeting the 99.9% availability SLA.

After 90 Days: Shift to the Cool tier, suitable for infrequent access with availability within 30 seconds.

After 365 Days: Transition to the Archive tier for rare access and longer retrieval time.

upvoted 1 times

hassexat 4 months ago

Hot

Cool

Cool

upvoted 1 times

kkk5566 4 months, 1 week ago

Hot, Cool, Cool is correct.

upvoted 1 times

Sima_al 11 months, 3 weeks ago

1. Hot - because of the 99.9% availability.
2. Hot - because Cool tier needs several minutes to give back an answer (but 30 sec. is asked for).
3. Cool - because the answer is needed within 5 minutes. Thats what cool tier does.

upvoted 2 times

kkk5566 4 months, 1 week ago

<https://learn.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview#summary-of-access-tier-options>

upvoted 1 times

shoottheduck 10 months, 2 weeks ago

Cool has a response time of Milliseconds. So Hot, Cool, Cool

upvoted 7 times

gabrys1997 1 year, 3 months ago

I think that 'cool' tier is just enough, it provides availability on 99.9%

upvoted 2 times

hanzocuk 1 year ago

Keep this in mind --> "The data will be accessed several times a day during the first 30 days". Cool tier is more expensive to read from.
hot, cool, cool looks correct.

upvoted 4 times

Marcohcm 1 year, 3 months ago

Cool Tier provides 99.9% availability only on RA-GRS. For ZRS, it should be 99% .

<https://learn.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview#summary-of-access-tier-options>

upvoted 4 times

Strix 1 year, 4 months ago

Correct!

upvoted 2 times

DRAG DROP

You have an Azure Data Lake Storage Gen 2 account named storage1.

You need to recommend a solution for accessing the content in storage1. The solution must meet the following requirements:

- List and read permissions must be granted at the storage account level.
- Additional permissions can be applied to individual objects in storage1.
- Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.

What should you use? To answer, drag the appropriate components to the correct requirements. Each component 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.

| Components | Answer Area |
|--|---|
| Access control lists (ACLs) | To grant permissions at the storage account level: <input type="text"/> |
| Role-based access control (RBAC) roles | To grant permissions at the object level: <input type="text"/> |
| Shared access signatures (SAS) | |
| Shared account keys | |

| Answer Area |
|--|
| Correct Answer:
To grant permissions at the storage account level: <input checked="" type="checkbox"/> Role-based access control (RBAC) roles
To grant permissions at the object level: <input checked="" type="checkbox"/> Access control lists (ACLs) |

 **SannPro** Highly Voted 11 months, 2 weeks ago

Correct

upvoted 8 times

 **aemilka** Highly Voted 8 months, 3 weeks ago

Correct.

Azure Data Lake Storage Gen2 implements an access control model that supports both Azure role-based access control (Azure RBAC) and POSIX-like access control lists (ACLs).

Azure RBAC scope are storage accounts and containers.

ACL scope are directories and files.

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

upvoted 6 times

 **EliteAllen** Most Recent 4 months ago

1. Role-based access control (RBAC) rules
2. Access control lists (ACLs)

upvoted 2 times

 **kkk5566** 4 months, 1 week ago

correct

upvoted 2 times

 **Venub28** 12 months ago

Given answer is correct

upvoted 4 times

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains a table named Sales.

Sales has row-level security (RLS) applied. RLS uses the following predicate filter.

```
CREATE FUNCTION Security.fn_securitypredicate(@SalesRep AS sysname)
    RETURNS TABLE
    WITH SCHEMABINDING
AS
    RETURN SELECT 1 AS fn_securitypredicate_result
    WHERE @SalesRep = USER_NAME() OR USER_NAME() = 'Manager';
```

A user named SalesUser1 is assigned the db_datareader role for Pool1.

Which rows in the Sales table are returned when SalesUser1 queries the table?

- A. only the rows for which the value in the User_Name column is SalesUser1
- B. all the rows
- C. only the rows for which the value in the SalesRep column is Manager
- D. only the rows for which the value in the SalesRep column is SalesUser1

Correct Answer: D

✉ **mamahani** Highly Voted 8 months, 2 weeks ago

Selected Answer: D

here is the same example directly from microsoft docs:

|<https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16#Typical>
its definitely D

upvoted 5 times

✉ **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: D

D is correct , see link

<https://learn.microsoft.com/en-us/training/modules/secure-data-warehouse-azure-synapse-analytics/6-exercise-manage-authorization-through-column-row-level-security>

upvoted 1 times

✉ **kkk5566** 4 months, 1 week ago

<https://learn.microsoft.com/en-us/training/modules/secure-data-warehouse-azure-synapse-analytics/6-exercise-manage-authorization-through-column-row-level-security>

upvoted 1 times

✉ **AHUI** 9 months, 1 week ago

Ans is C.

The function returns 1 when a row in the SalesRep column is the same as the user executing the query (@SalesRep = USER_NAME()) or if the user executing the query is the Manager user (USER_NAME() = 'Manager').

ref: <https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16>

upvoted 2 times

✉ **zekescokies** 8 months, 4 weeks ago

It's D. It clearly states that the user querying the table is SalesUser1. I feel they should have mentioned it being a manager if it's C.

upvoted 7 times

✉ **shakes103** 9 months ago

I have looked it up too. Answer is C

upvoted 2 times

✉ **Mal2002** 5 months ago

If you really looked up then what did you understand from this?

```
EXECUTE AS USER = 'SalesRep1';
SELECT * FROM Sales.Orders;
```

```
REVERT;
```

```
EXECUTE AS USER = 'SalesRep2';
SELECT * FROM Sales.Orders;
REVERT;
```

```
EXECUTE AS USER = 'Manager';
SELECT * FROM Sales.Orders;
REVERT;
```

The manager should see all six rows. The Sales1 and Sales2 users should only see their own sales.

<https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16#Typical>

It's clearly D
upvoted 1 times

 **aemilka** 8 months, 3 weeks ago

In the "Scenario for users who authenticate to the database" there is the same code snippet and it's clearly stated that after applying security policy adding the function as a filter predicate "the manager should see all rows. The Sales1 and Sales2 users should only see their own sales."

So the answer is D.
upvoted 6 times

HOTSPOT

You have an Azure Data Lake Storage Gen2 account named account1 that contains the resources shown in the following table.

| Name | Type | Description |
|------------|-----------|---------------------------|
| container1 | Container | A container |
| Directory1 | Directory | A directory in container1 |
| File1 | File | A file in Directory1 |

You need to ~~configure~~ access control lists (ACLs) to allow a user named User1 to delete File1. User1 is NOT assigned any role-based access control (RBAC) roles for account1. The solution must use the principle of least privilege.

Which type of ACL should you ~~configure~~ for each resource? To answer select the appropriate options in the answer area.

Answer Area

container1:

| | |
|-----------------|---|
| | ▼ |
| --- permissions | |
| -WX permissions | |
| --X permissions | |

Directory1:

| | |
|-----------------|---|
| | ▼ |
| --- permissions | |
| -WX permissions | |
| --X permissions | |

File1:

| | |
|-----------------|---|
| | ▼ |
| --- permissions | |
| -WX permissions | |
| --X permissions | |

Answer Area

container1:

| | |
|------------------------|---|
| | ▼ |
| --- permissions | |
| -WX permissions | |
| --X permissions | |

Correct Answer:

Directory1:

| | |
|------------------------|---|
| | ▼ |
| --- permissions | |
| -WX permissions | |
| --X permissions | |

File1:

| | |
|------------------------|---|
| | ▼ |
| --- permissions | |
| -WX permissions | |
| --X permissions | |

 **BPW**  8 months, 3 weeks ago

Answer is

--x/ -wx/ ---

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

upvoted 33 times

 Ahmad_Abukhater  9 months ago

last box file1 should be --- (Frist option)

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

upvoted 7 times

 vctrhugo 6 months, 3 weeks ago

"So long as the previous two conditions are true."

upvoted 1 times

 DataEngineer7331 8 months, 3 weeks ago

According to this your Link, the Directory should have "-WX" and the File "---

upvoted 1 times

 matiandal  2 months ago

Correct Answer: --X, -WX, ---

IN general : X until the last folder, the last folder needs WX, and on the file needs nothing(---)

R:

Common scenarios related to permissions

<https://learn.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control#common-scenarios-related-to-permissions>

upvoted 2 times

 SenMia 3 weeks, 3 days ago

please clarify why should Directory1 be --WX? why write access for a directory? shouldn't it be just --X?

upvoted 1 times

 kkk5566 4 months, 1 week ago

--x/ --x/ ---

upvoted 2 times

 kkk5566 4 months ago

<https://learn.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control#common-scenarios-related-to-permissions>

upvoted 1 times

 wanchihh 3 months, 2 weeks ago

--x/ -wx/ ---

upvoted 2 times

 Ram9198 6 months, 1 week ago

--x/ -wx/ ---

upvoted 4 times

 vctrhugo 6 months, 3 weeks ago

The solution must use the principle of least privilege!!!

You shouldn't grant -WX to the entire Directory1. Instead, do -x / --w

upvoted 2 times

 mamahani 8 months, 2 weeks ago

you do not need any permissions on a file itself to delete it; you only need permissions on the folder where the file resides;

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control#common-scenarios-related-to-acl-permissions>
so answer -x / -wx / ---

upvoted 4 times

 andjurovicela 6 months ago

I agree with everything except the write&execute permission for directory. According to the "famous" link on ACLs the directory permissions should be only execute for deleting actions.

upvoted 1 times

 wanchihh 3 months, 2 weeks ago

Not according to this link

<https://learn.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control#common-scenarios-related-to-permissions>

The directory on which the file resides have to be -wx in order to delete the file.

upvoted 1 times

 wanchihh 3 months, 2 weeks ago

The link should be this instead:

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control#common-scenarios-related-to-acl-permissions>

upvoted 1 times

 DataEngDP 4 months, 1 week ago

You need write and execute in order to create child items in a directory. And for deleting you don't need permissions so ---.

upvoted 2 times

You have an Azure subscription that is linked to a tenant in Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra. The tenant that contains a security group named Group1. The subscription contains an Azure Data Lake Storage account named myaccount1. The myaccount1 account contains two containers named container1 and container2.

You need to grant Group1 read access to container1. The solution must use the principle of least privilege.

Which role should you assign to Group1?

- A. Storage Table Data Reader for myaccount1
- B. Storage Blob Data Reader for container1
- C. Storage Blob Data Reader for myaccount1
- D. Storage Table Data Reader for container1

Correct Answer: B

✉  **lamthealpha** Highly Voted 8 months, 2 weeks ago

The appropriate role to assign to Group1 to grant read access to container1 with the principle of least privilege is option B, Storage Blob Data Reader for container1.

Option A, Storage Table Data Reader for myaccount1, is incorrect because it grants read access to all tables in the storage account, not just container1.

Option C, Storage Blob Data Reader for myaccount1, is incorrect because it grants read access to all containers in the storage account, not just container1.

Option D, Storage Table Data Reader for container1, is incorrect because it grants read access to tables in the specified container only, not blobs in container1.

Therefore, option B, Storage Blob Data Reader for container1, is the most appropriate role to assign Group1 to grant read access to container1 with the principle of least privilege.

upvoted 10 times

✉  **kkk5566** Most Recent 4 months ago

Selected Answer: B

correct

upvoted 1 times

✉  **xymtyk** 8 months, 2 weeks ago

Selected Answer: B

Correct.

upvoted 3 times

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named dbo.Users.

You need to prevent a group of users from reading user email addresses from dbo.Users.

What should you use?

- A. column-level security
- B. row-level security (RLS)
- C. Transparent Data Encryption (TDE)
- D. dynamic data masking

Correct Answer: A

✉ **iamthealpha** Highly Voted 8 months, 2 weeks ago

The appropriate feature to use to prevent a group of users from reading user email addresses from dbo.Users in an Azure Synapse Analytics dedicated SQL pool is option A, column-level security.

Option B, row-level security (RLS), is used to filter rows in a table based on the user executing a query, but it cannot prevent certain columns from being read by a group of users.

Option C, Transparent Data Encryption (TDE), encrypts data at rest and does not prevent a group of users from reading specific columns in a table.

Option D, dynamic data masking, is used to mask sensitive data in query results, but it does not prevent a group of users from reading the actual values in a column.

Therefore, option A, column-level security, is the most appropriate feature to use to prevent a group of users from reading user email addresses from dbo.Users in an Azure Synapse Analytics dedicated SQL pool. Column-level security can be used to deny read access to specific columns in a table based on a user or group's permissions.

upvoted 12 times

✉ **shakes103** Highly Voted 9 months ago

Selected Answer: A

A is correct. Column-level security simplifies the design and coding of security in your application, allowing you to restrict column access to protect sensitive data. For example, ensuring that specific users can access only certain columns of a table pertinent to their department.

upvoted 5 times

✉ **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: A

correct

upvoted 1 times

✉ **RoyP654** 7 months, 2 weeks ago

I guess i missed reading about it, but how do you implement column-level security? If via view, folks still have access to the underlying table. Let me know.

upvoted 1 times

✉ **RoyP654** 7 months ago

sorry, pls ignore my comment here

upvoted 1 times

✉ **halamgir15** 9 months ago

I think it should be D:
dynamic data masking

upvoted 4 times

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool that hosts a database named DB1.

You need to ensure that DB1 meets the following security requirements:

- When credit card numbers show in applications, only the last four digits must be visible.
- Tax numbers must be visible only to specific users.

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

NOTE: Each correct selection is worth one point.

Answer Area

Credit card numbers:

Column-level security
Dynamic Data Masking
Row-level security (RLS)

Tax numbers:

Column-level security
Row-level security (RLS)
Transparent Database Encryption (TDE)

Answer Area

Credit card numbers:

Column-level security
Dynamic Data Masking
Row-level security (RLS)

Correct Answer:

Tax numbers:

Column-level security
Row-level security (RLS)
Transparent Database Encryption (TDE)

 **haythemsi** Highly Voted 8 months ago

Correct

upvoted 8 times

 **kim32** Highly Voted 8 months ago

It should be Row Level security not column since limited for some users

upvoted 7 times

 **francocalvo** 8 months ago

I think the answer is correct. Imagine a team where all have access to the table, but just one person needs access to the tax numbers, you can use column-level to disable access for all the other people except the one that needs it

upvoted 16 times

 **Matt2000** 5 months ago

yes. that is what row-level security is designed for.

upvoted 1 times

 **kkk5566** Most Recent 4 months, 1 week ago

masking ,row Level

upvoted 1 times

✉ **kkk5566** 4 months, 1 week ago

should be Column-level security

upvoted 4 times

✉ **eternalamit5** 4 months, 3 weeks ago

It should be Column-level security as it ensures those specific users can access only certain columns of a table.

Whereas, RLS can help you to create a group membership or execution context in order to control not just columns in a database table, but actually, the rows.

upvoted 3 times

Question #39

Topic 3

You have an Azure subscription that contains a storage account named storage1 and an Azure Synapse Analytics dedicated SQL pool. The storage1 account contains a CSV file that requires an account key for access.

You plan to read the contents of the CSV file by using an external table.

You need to create an external data source for the external table.

What should you create first?

- A. a database role
- B. a database scoped credential
- C. a database view
- D. an external file format

Correct Answer: B

✉ **cloud_lady** Highly Voted 8 months ago

Given answer is correct.

Refer this link - <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables>

upvoted 12 times

✉ **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: B

correct

upvoted 3 times

✉ **MSExpert** 4 months, 4 weeks ago

Correct

upvoted 2 times

You have a tenant in Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra. The tenant contains a group named Group1.

You have an Azure subscription that contains the resources shown in the following table.

| Name | Type | Description |
|-------------|-----------------------------------|---|
| ws1 | Azure Synapse Analytics workspace | None |
| storage1 | Azure Storage account | Contains CSV files |
| credential1 | Database-scoped credential | Stored in the Azure Synapse Analytics serverless SQL pool in ws1 and used to authenticate to storage1 |

You need to ensure that members of Group1 can read CSV files from storage1 by using the OPENROWSET function. The solution must meet the following requirements:

- The members of Group1 must use credential1 to access storage1.
- The principle of least privilege must be followed.

Which permission should you grant to Group1?

- A. EXECUTE
- B. CONTROL
- C. REFERENCES
- D. SELECT

Correct Answer: A

 **Sachmett** Highly Voted  3 weeks, 3 days ago

Selected Answer: C

"Caller must have REFERENCES permission on credential to use it to authenticate to storage."
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-openrowset>
 upvoted 8 times

 **vernillen** Highly Voted  1 month, 3 weeks ago

Selected Answer: D

When you're using the OPENROWSET function to read data from the storage account, you're actually performing a read operation, not an execute operation. The credential is used implicitly by Azure Synapse to authenticate the session with the storage account and does not require the EXECUTE permission for the user or group accessing it. Instead, you grant permissions that are appropriate for data access. In this case, the SELECT permission is the correct one to use because it allows the members of Group1 to read or select the data.
 upvoted 5 times

 **vernillen** Most Recent  4 days, 23 hours ago

Selected Answer: C

Database users who access external storage must have permission to use credentials. To use the credential, a user must have the REFERENCES permission on a specific credential.

To grant the REFERENCES permission on a server-level credential for a login, use the following T-SQL query in the master database:

```
GRANT REFERENCES ON CREDENTIAL::[server-level_credential] TO [login_name];
```

To grant a REFERENCES permission on a database-scoped credential for a database user, use the following T-SQL query in the user database:

```
GRANT REFERENCES ON DATABASE SCOPED CREDENTIAL::[database-scoped_credential] TO [user_name];
```

upvoted 1 times

 **vernillen** 4 days, 21 hours ago

To be fair, I know I answered with both answers, but the context isn't at all that clear. So depending on the context it could be D or C really.
 upvoted 1 times

 **jasmd2** 1 week, 3 days ago

Selected Answer: C

Caller must have REFERENCES permission on credential to use it to authenticate to storage.
 upvoted 1 times

BitacTeam 3 weeks, 4 days ago

I am new this topic, however to read from a file you should grant Execute on the above level. so Storage1 should grant Execute on file level should be read.

upvoted 1 times

kam1122 1 month, 1 week ago

Selected Answer: D

D , SELECT

upvoted 1 times

Question #41

Topic 3

You have an Azure subscription that contains an Azure Data Lake Storage account named dl1 and an Azure Analytics Synapse workspace named workspace1.

You need to query the data in dl1 by using an Apache Spark pool named Pool1 in workspace1. The solution must ensure that the data is accessible Pool1.

Which two actions achieve the goal? Each correct answer presents a complete solution.

NOTE: Each correct answer is worth one point.

- A. Implement Azure Synapse Link.
- B. Load the data to the primary storage account of workspace1.
- C. From workspace1, create a linked service for the dl1.
- D. From Microsoft Purview, register dl1 as a data source.

Correct Answer: CD

jongert 3 days, 18 hours ago

Selected Answer: BC

Would say Purview only registers the data source to track lineage, should not have anything to do with access. Synapse Link is not concerned with data lake.

Seemingly, data lake storage has to be configured as primary storage as per:

<https://learn.microsoft.com/en-us/troubleshoot/azure/synapse-analytics/spark/spark-jobexec-storage-access#common-issues-and-solutions>

upvoted 2 times

Topic 4 - Question Set 4

You implement an enterprise data warehouse in Azure Synapse Analytics.

You have a large fact table that is 10 terabytes (TB) in size.

Incoming queries use the primary key SaleKey column to retrieve data as displayed in the following table:

| SaleKey | CityKey | CustomerKey | StockItemKey | InvoiceDateKey | Quantity | UnitPrice | TotalExcludingTax |
|---------|---------|-------------|--------------|----------------|----------|-----------|-------------------|
| 49309 | 90858 | 70 | 69 | 10/22/13 | 8 | 16 | 128 |
| 49313 | 55710 | 126 | 69 | 10/22/13 | 2 | 16 | 32 |
| 49343 | 44710 | 234 | 68 | 10/22/13 | 10 | 16 | 160 |
| 49352 | 66109 | 163 | 70 | 10/22/13 | 4 | 16 | 64 |
| 49448 | 65312 | 230 | 70 | 10/22/13 | 8 | 16 | 128 |
| 49646 | 85877 | 271 | 70 | 10/24/13 | 1 | 16 | 16 |
| 49798 | 41238 | 288 | 69 | 10/24/13 | 1 | 16 | 16 |

You need to distribute the large fact table across multiple nodes to optimize performance of the table.

Which technology should you use?

- A. hash distributed table with clustered index
- B. hash distributed table with clustered Columnstore index
- C. round robin distributed table with clustered index
- D. round robin distributed table with clustered Columnstore index
- E. heap table with distribution replicate

Correct Answer: B

Hash-distributed tables improve query performance on large fact tables.

Columnstore indexes can achieve up to 100x better performance on analytics and data warehousing workloads and up to 10x better data compression than traditional rowstore indexes.

Incorrect Answers:

C, D: Round-robin tables are useful for improving loading speed.

Reference:

<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute> <https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-query-performance>

✉  **rjile** Highly Voted 2 years, 6 months ago

correct B

upvoted 39 times

✉  **aortega** 2 years, 3 months ago

For Example:

```
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
, DISTRIBUTION = HASH([ProductKey])
)
;
```

upvoted 6 times

✉  **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: B

is correct

upvoted 1 times

✉  **temmytak** 9 months, 1 week ago

Selected Answer: B

Correct B

upvoted 2 times

 **Shanmahi** 1 year, 1 month ago

Selected Answer: B

Hash on SaleKey distribution column using Columnstore clustered index; Why? (1) petabyte scale data (2) incoming query on SaleKey therefore, SaleKey will be used in WHERE condition and clustered columnstore index will be efficient.

upvoted 2 times

 **dmitriypo** 1 year, 2 months ago

Selected Answer: B

B is correct

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: B

yes, B is correct

upvoted 2 times

 **Remedios79** 1 year, 6 months ago

correct

upvoted 2 times

 **LiLy91** 1 year, 11 months ago

Clustered indexes may outperform clustered columnstore tables when a single row needs to be quickly retrieved. For queries where a single or very few row lookup is required to perform with extreme speed, consider a clustered index or nonclustered secondary index. The disadvantage to using a clustered index is that only queries that benefit are the ones that use a highly selective filter on the clustered index column. To improve filter on other columns, a nonclustered index can be added to other columns. However, each index that is added to a table adds both space and processing time to loads.

upvoted 1 times

 **jv2120** 2 years ago

Clustered columnstore indexes are the most efficient way you can store your data in Azure SQL Data Warehouse. Storing your data in tables that have a clustered columnstore index are the fastest way to query your data. It will give you the greatest data compression and lower your storage costs.

Hash-distributed tables work well for large fact tables in a star schema. They can have very large numbers of rows and still achieve high performance.

Consider using a hash-distributed table when:

The table size on disk is more than 2 GB.

The table has frequent insert, update, and delete operations.

ANS B

upvoted 1 times

 **SujithaVulchi** 2 years, 3 months ago

A heap is a table without a clustered index. One or more nonclustered indexes can be created on tables stored as a heap. Data is stored in the heap without specifying an order. Usually data is initially stored in the order in which the rows are inserted into the table, but the Database Engine can move data around in the heap to store the rows efficiently; so the data order cannot be predicted. To guarantee the order of rows returned from a heap, you must use the ORDER BY clause. To specify a permanent logical order for storing the rows, create a clustered index on the table, so that the table is not a heap.

Correct answer: Non clustered

upvoted 2 times

 **Avinash75** 2 years, 6 months ago

Incoming queries use the primary key SaleKey column to retrieve data as displayed in the following table ..doesn't this mean Salekey will be used in where clause , which makes Salekey not suitable for hashkey distribution .

Choosing a distribution column that helps minimize data movement is one of the most important strategies for optimizing performance of your dedicated SQL pool:

- Is not used in WHERE clauses. This could narrow the query to not run on all the distributions.

with no obvious choice i feel it should be round robin with column clustered index i.e D

upvoted 1 times

 **[Removed]** 2 years, 2 months ago

Consider using a hash-distributed table when:

The table size on disk is more than 2 GB

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

upvoted 1 times

 **Aditya0891** 1 year, 6 months ago

when you don't have any good candidate for hashkey you can also go for composite key. And here the size of the table is huge and using round robin you will never obtain good performance

upvoted 1 times

 **erssiws** 2 years, 6 months ago

I understand that hash distribution mainly for improving the joins and group-by to reduce the data shuffling. In this case, there is no join or group-by mentioned. I think round-robin would be a better option.

upvoted 1 times

 **Yatoom** 2 years, 7 months ago

If the answer is hash distributed, then what would be the key? If there is no obvious joining key, round-robin should be chosen (<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute#round-robin-distributed>)
upvoted 1 times

 **Preben** 2 years, 7 months ago

It says it uses the SaleKey.
Round-robin is generally not effective at these large scale tables. The 10 tb was a very important hint here.
upvoted 15 times

You have an Azure Synapse Analytics dedicated SQL pool that contains a large fact table. The table contains 50 columns and 5 billion rows and is a heap.

Most queries against the table aggregate values from approximately 100 million rows and return only two columns.

You discover that the queries against the fact table are very slow.

Which type of index should you add to provide the fastest query times?

- A. nonclustered columnstore
- B. clustered columnstore
- C. nonclustered
- D. clustered

Correct Answer: B

Clustered columnstore indexes are one of the most efficient ways you can store your data in dedicated SQL pool.

Columnstore tables won't benefit a query unless the table has more than 60 million rows.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool>

 **damaldon** Highly Voted 2 years, 6 months ago

correct!

upvoted 30 times

 **Miris** Highly Voted 2 years, 7 months ago

correct

upvoted 14 times

 **AlfredPennyworth** Most Recent 3 weeks, 3 days ago

Clustered index

* Tables with up to 100 million rows

* Large tables (more than 100 million rows) with only 1-2 columns heavily used

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: D

If your table size is less than the recommended 60 million rows for clustered columnstore indexing, consider using heap or clustered index tables.

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

upvoted 1 times

 **Ram9198** 4 months, 3 weeks ago

Selected Answer: C

It asks which index to add answer c

upvoted 1 times

 **Ram9198** 4 months, 4 weeks ago

Selected Answer: A

A heap is a table without a clustered index. One or more nonclustered indexes can be created on tables stored as a heap.
Question says already it's a heap table and asks what to add? So answer is A

upvoted 2 times

 **matiandal** 5 months, 3 weeks ago

why not a NCCI - why not A?

Nonclustered columnstore index on a disk-based heap or B-tree index Use for:

1) An OLTP workload that has some analytics queries. You can drop B-tree indexes created for analytics and replace them with one nonclustered columnstore index.

2) Many traditional OLTP workloads that perform Extract Transform and Load (ETL) operations to move data to a separate data warehouse. You can eliminate ETL and a separate data warehouse by creating a nonclustered columnstore index on some of the OLTP tables. NCCI is an additional index that requires 10% more storage on average.

R: <https://learn.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-design-guidance?view=sql-server-ver16#choose-the-best-columnstore-index-for-your-needs>

Enjoy !

upvoted 1 times

□ **Matt2000** 5 months ago

it is a currently a heap. thus clustered columnstore makes most sense.

upvoted 1 times

□ **Ram9198** 6 months, 1 week ago

Selected Answer: B

Only 2 columns returned

upvoted 1 times

□ **auwia** 6 months, 2 weeks ago

Selected Answer: B

B of course, there are a few scenarios where clustered columnstore may not be a good option:

Columnstore tables do not support varchar(max), nvarchar(max), and varbinary(max). Consider heap or clustered index instead.

Columnstore tables may be less efficient for transient data. Consider heap and perhaps even temporary tables.

Small tables with less than 60 million rows. Consider heap tables.

upvoted 1 times

□ **vctrhugo** 6 months, 3 weeks ago

Selected Answer: B

B. clustered columnstore index.

Given the large fact table with 50 columns and 5 billion rows, and the fact that most queries aggregate values from approximately 100 million rows and return only two columns, a clustered columnstore index would be the most suitable choice. Clustered columnstore indexes are designed for large-scale data warehousing scenarios and provide excellent compression and query performance for analytical workloads.

A clustered columnstore index stores the data in columnar format, enabling efficient data compression and batch-based query execution. It allows for significant query performance improvements, especially for aggregations and large-scale data retrieval.

upvoted 2 times

□ **mamahani** 8 months, 2 weeks ago

im really baffled by all the answers here; noone is even considering clustered index, which is what microsoft is recommending for this particular user case scenario;

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet#index-your-table>

for a table up to 100 mln records and using heavily 1-2 columns and performing queries with lots of joins and aggregations (group by clause) microsoft recommends clustered index; why is this recommendation not applicable here? could someone explain?

upvoted 4 times

□ **mamahani** 8 months, 1 week ago

ignore pls; instead of reading watch out if....i read just if, must have been tired?; so clustered index is NOT good when group by operations; its good if you need to retrieve 1 single row or few rows (but aggregate is not just few rows -> its many many rows aggregating to 1 row, which is not the same); by this i believe its indeed clustered columnstore index so the given answer is correct

upvoted 5 times

□ **AHUI** 9 months, 1 week ago

Selected Answer: B

correct

upvoted 3 times

□ **Rakrah** 11 months, 1 week ago

Answer is correct (B) clustered columnstore - This index reordered the physical table data with columnar format which is stored with index and compressed. All the query will fetch from index columnstored data and it is designed specially Data warehouse complex query and aggregated data too.

upvoted 3 times

□ **OldSchool** 1 year, 1 month ago

Selected Answer: B

It's B

"Do not use a heap when ranges of data are frequently queried from the table. A clustered index on the range column will avoid sorting the entire heap."

<https://learn.microsoft.com/en-us/sql/relational-databases/indexes/heaps-tables-without-clustered-indexes?toc=%2Fazure%2Fsynapse-analytics%2Fsql-data-warehouse%2Ftoc.json&bc=%2Fazure%2Fsynapse-analytics%2Fsql-data-warehouse%2Fbreadcrumb%2Ftoc.json&view=sql-server-ver15&preserve-view=true#when-not-to-use-a-heap>

upvoted 1 times

□ **stunner85_** 1 year, 3 months ago

Selected Answer: C

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-index>
upvoted 2 times

 **dom271219** 1 year, 4 months ago

Selected Answer: B
"return only two columns" => don't be confused. It's 2 col and not 2 rows => then Clustered columnstore
upvoted 6 times

 **Ast999** 1 year, 4 months ago

Selected Answer: C
<https://docs.microsoft.com/en-us/sql/relational-databases/indexes/heaps-tables-without-clustered-indexes?view=sql-server-ver16>
upvoted 1 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

You create an Azure Databricks cluster and specify an additional library to install.

When you attempt to load the library to a notebook, the library is not found.

You need to identify the cause of the issue.

What should you review?

- A. notebook logs
- B. cluster event logs
- C. global init scripts logs
- D. workspace logs

Correct Answer: C

Cluster-scoped Init Scripts: Init scripts are shell scripts that run during the startup of each cluster node before the Spark driver or worker JVM starts. Databricks customers use init scripts for various purposes such as installing custom libraries, launching background processes, or applying enterprise security policies.

Logs for Cluster-scoped init scripts are now more consistent with Cluster Log Delivery and can be found in the same root folder as driver and executor logs for the cluster.

Reference:

<https://databricks.com/blog/2018/08/30/introducing-cluster-scoped-init-scripts.html>

 **Dizzystar** Highly Voted 2 years, 2 months ago

I should say Cluster Event logs:

Azure Databricks provides three kinds of logging of cluster-related activity:

Cluster event logs, which capture cluster lifecycle events, like creation, termination, configuration edits, and so on.

Apache Spark driver and worker logs, which you can use for debugging.

Cluster init-script logs, valuable for debugging init scripts.

<https://docs.microsoft.com/en-us/azure/databricks/clusters/clusters-manage#event-log>

upvoted 31 times

 **dragos_dragos62000** Highly Voted 2 years, 6 months ago

Correct

upvoted 11 times

 **jongert** Most Recent 6 days, 21 hours ago

Selected Answer: B

The correct answer is B

Cluster event logs capture two init script events: INIT_SCRIPTS_STARTED and INIT_SCRIPTS_FINISHED, indicating which scripts are scheduled for execution and which have completed successfully. INIT_SCRIPTS_FINISHED also captures execution duration.

<https://docs.databricks.com/en/init-scripts/logs.html>

upvoted 1 times

 **Momoanwar** 2 weeks, 1 day ago

Selected Answer: B

ChatGpt :

if the library was to be installed through:

- Standard Databricks library installation methods: Check the cluster event logs (B).

- A global init script: Check the global init scripts logs (C).

Without additional context or explicit mention of an init script being used, option B is typically the more standard choice for initial troubleshooting.

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Legacy global init scripts and cluster-named init scripts are deprecated and cannot be used in new workspaces starting February 21, 2023. On September 1st, 2023, Azure Databricks will disable legacy global init scripts for all workspaces.

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: C

should be C

upvoted 1 times

 **Ram9198** 6 months, 1 week ago

Selected Answer: B

Cluster event logs
upvoted 2 times

 **vctrhugo** 6 months, 2 weeks ago

Selected Answer: B

Cluster event logs in Azure Databricks provide detailed information about the cluster's lifecycle events, including the installation and initialization of libraries. By reviewing the cluster event logs, you can examine the events related to library installation and determine if any errors or issues occurred during the process.

upvoted 3 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: C

Cluster event logs do not log init script events for each cluster node; only one node is selected to represent them all.

<https://learn.microsoft.com/en-us/azure/databricks/clusters/init-scripts>
upvoted 1 times

 **vctrhugo** 6 months, 2 weeks ago

Installation and initialization of libraries is not part of init scripts.
upvoted 2 times

 **bch9994** 5 months ago

That's incorrect. It is a part of init scripts.

Some examples of tasks performed by init scripts include:

Set system properties and environment variables used by the JVM.

Modify Spark configuration parameters.

Modify the JVM system classpath in special cases.

Install packages and libraries not included in Databricks Runtime. To install Python packages, use the Azure Databricks pip binary located at /databricks/python/bin/pip to ensure that Python packages install into the Azure Databricks Python virtual environment rather than the system Python environment. For example, /databricks/python/bin/pip install <package-name>.

[https://learn.microsoft.com/en-us/azure/databricks/init-scripts/](https://learn.microsoft.com/en-us/azure/databricks/init-scripts)

upvoted 1 times

 **aemilka** 8 months, 3 weeks ago

Selected Answer: C

Additional libraries are installed in global init scripts, so correct answer is C.

Some examples of tasks performed by init scripts include:

- Install packages and libraries not included in Databricks Runtime. To install Python packages, use the Azure Databricks pip binary located at /databricks/python/bin/pip to ensure that Python packages install into the Azure Databricks Python virtual environment rather than the system Python environment. For example, /databricks/python/bin/pip install <package-name>.
- Modify the JVM system classpath in special cases.
- Set system properties and environment variables used by the JVM.
- Modify Spark configuration parameters.

ref: <https://learn.microsoft.com/en-us/azure/databricks/clusters/init-scripts>

upvoted 2 times

 **vctrhugo** 6 months, 2 weeks ago

There are two primary ways to install a library on a cluster:

- Install a workspace library that has been already been uploaded to the workspace.
- Install a library for use with a specific cluster only.

upvoted 1 times

 **kornat** 9 months, 1 week ago

Selected Answer: C

correct

upvoted 2 times

 **esaade** 10 months ago

Selected Answer: B

the best option in this scenario would be to review the cluster event logs to identify the cause of the issue where an additional library is not found in the Azure Databricks cluster.

upvoted 3 times

 **lafita** 11 months ago

Answer C.

A global init script runs on every cluster created in your workspace. Global init scripts are useful when you want to enforce organization-wide library configurations or security screens. Only admins can create global init scripts. You can create them using either the UI or REST API.

upvoted 2 times

 **youngbug** 11 months, 2 weeks ago

Selected Answer: C

cluster event logs only record start and finish event, so C is right, init script logs record the details of running.

upvoted 2 times

 **gerrie1979** 1 year, 1 month ago

Selected Answer: B

<https://learn.microsoft.com/en-us/azure/databricks/clusters/init-scripts>:

Init script start and finish events are captured in cluster event logs. Details are captured in cluster logs. Global init script create, edit, and delete events are also captured in account-level diagnostic logs.

Cluster event logs capture two init script events: INIT_SCRIPTS_STARTED and INIT_SCRIPTS_FINISHED, indicating which scripts are scheduled for execution and which have completed successfully. INIT_SCRIPTS_FINISHED also captures execution duration.

Global init scripts are indicated in the log event details by the key "global" and cluster-scoped init scripts are indicated by the key "cluster".

upvoted 2 times

 **dmitriypo** 1 year, 2 months ago

Selected Answer: C

Agree with the given answer - C

Database customers use init scripts for various purposes such as installing custom libraries, launching background processes, or applying enterprise security policies.

Reference:

<https://www.databricks.com/blog/2018/08/30/introducing-cluster-scoped-init-scripts.html>

upvoted 2 times

 **Raghul08** 1 year, 11 months ago

My Answer is B

upvoted 1 times

You have an Azure data factory.

You need to examine the pipeline failures from the last 60 days.

What should you use?

- A. the Activity log blade for the Data Factory resource
- B. the Monitor & Manage app in Data Factory
- C. the Resource health blade for the Data Factory resource
- D. Azure Monitor

Correct Answer: D

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor>

 **erssiws** Highly Voted 2 years, 6 months ago

Activity logs show only activities, e.g., trigger the pipeline, stop the pipeline, ...

Resource health check shows only the healthiness of the resource.

The monitor app indeed contains the pipeline run failure information. But it keep the data only for 45 days.
upvoted 30 times

 **snna4** 2 years ago

"Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time."

upvoted 10 times

 **damaldon** Highly Voted 2 years, 6 months ago

Correct!

upvoted 7 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: D

correct

upvoted 1 times

 **dmitriypo** 1 year, 2 months ago

Selected Answer: D

Agree with D

upvoted 3 times

 **Deeksha1234** 1 year, 5 months ago

correct

upvoted 3 times

 **KrishIC** 2 years ago

Selected Answer: D

CORRECT

upvoted 4 times

 **FredNo** 2 years, 1 month ago

Selected Answer: D

Correct

upvoted 4 times

 **Jayant68** 2 years, 1 month ago

Correct..

upvoted 3 times

You are monitoring an Azure Stream Analytics job.
The Backlogged Input Events count has been 20 for the last hour.
You need to reduce the Backlogged Input Events count.
What should you do?

- A. Drop late arriving events from the job.
- B. Add an Azure Storage account to the job.
- C. Increase the streaming units for the job.
- D. Stop the job.

Correct Answer: C

General symptoms of the job hitting system resource limits include:

If the backlog event metric keeps increasing, it's an indicator that the system resource is constrained (either because of output sink throttling, or high CPU).

Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job: adjust Streaming Units.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-scale-jobs> <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-monitoring>

 **MinionVII** Highly Voted 2 years, 6 months ago

Correct.

"Backlogged Input Events Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job."

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-monitoring>

upvoted 18 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: C

correct

upvoted 1 times

 **Deeksha1234** 1 year, 5 months ago

correct

upvoted 2 times

 **juanlu46** 1 year, 8 months ago

Selected Answer: C

Correct!

upvoted 4 times

You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination. You need to ensure that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs. What should you do?

- A. Pin the cluster.
- B. Create an Azure runbook that starts the cluster every 90 days.
- C. Terminate the cluster manually when processing completes.
- D. Clone the cluster after it is terminated.

Correct Answer: A

Azure Databricks retains cluster configuration information for up to 70 all-purpose clusters terminated in the last 30 days and up to 30 job clusters recently terminated by the job scheduler. To keep an all-purpose cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.

Reference:

<https://docs.microsoft.com/en-us/azure/databricks/clusters/>

 **FredNo** Highly Voted 2 years, 1 month ago

Selected Answer: A

Correct

upvoted 11 times

 **markpumc** Highly Voted 10 months ago

To ensure that the cluster configuration is retained indefinitely after the cluster is terminated while minimizing costs, you should pin the cluster.

Pinning a cluster in Azure Databricks prevents it from being terminated by the auto-termination feature. This means that the cluster configuration and installed libraries will be retained even if the cluster is not being used. This is the most efficient and cost-effective way to ensure that the cluster configuration is retained indefinitely after the cluster is terminated.

Creating an Azure runbook to start the cluster every 90 days would require additional resources and would not be a cost-effective solution. Terminating the cluster manually when processing completes would not retain the cluster configuration. Cloning the cluster after it is terminated would create a new cluster with the same configuration, but this would also result in additional costs. Should be A

upvoted 9 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: A

is correct

upvoted 1 times

 **akk_1289** 6 months, 1 week ago

got this question for my exam

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

correct

upvoted 3 times

 **Podavenna** 2 years, 3 months ago

Correct answer!

upvoted 6 times

You have an Azure data solution that contains an enterprise data warehouse in Azure Synapse Analytics named DW1.

Several users execute ad hoc queries to DW1 concurrently.

You regularly perform automated data loads to DW1.

You need to ensure that the automated data loads have enough memory available to complete quickly and successfully when the adhoc queries run.

What should you do?

- A. Hash distribute the large fact tables in DW1 before performing the automated data loads.
- B. Assign a smaller resource class to the automated data load queries.
- C. Assign a larger resource class to the automated data load queries.
- D. Create sampled statistics for every column in each table of DW1.

Correct Answer: C

The performance capacity of a query is determined by the user's resource class. Resource classes are pre-determined resource limits in Synapse SQL pool that govern compute resources and concurrency for query execution.

Resource classes can help you configure resources for your queries by setting limits on the number of queries that run concurrently and on the compute- resources assigned to each query. There's a trade-off between memory and concurrency.

Smaller resource classes reduce the maximum memory per query, but increase concurrency.

Larger resource classes increase the maximum memory per query, but reduce concurrency.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/resource-classes-for-workload-management>

 **Podavenna** Highly Voted 2 years, 3 months ago

Correct answer!

upvoted 21 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: C

ic correct

upvoted 1 times

 **vctrhugo** 6 months, 2 weeks ago

Selected Answer: C

Assigning a larger resource class to the automated data load queries prioritizes their resource allocation, allowing them to complete without being heavily impacted by the concurrent ad hoc queries. This helps avoid contention and ensures that the data loads can utilize the necessary resources to complete successfully.

upvoted 2 times

 **AHUI** 9 months, 1 week ago

Selected Answer: C

agreed

upvoted 3 times

 **Deeksha1234** 1 year, 5 months ago

correct

upvoted 3 times

 **juanlu46** 1 year, 8 months ago

Selected Answer: C

Is correct!

upvoted 3 times

 **aortega** 2 years, 3 months ago

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/resource-classes-for-workload-management>

upvoted 2 times

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1. You need to identify the extent of the data skew in Table1. What should you do in Synapse Studio?

- A. Connect to the built-in pool and run DBCC PDW_SHOWSPACEUSED.
- B. Connect to the built-in pool and run DBCC CHECKALLOC.
- C. Connect to Pool1 and query sys.dm_pdw_node_status.
- D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_stats.

Correct Answer: D

Microsoft recommends use of sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

✉ **wuespe** Highly Voted 2 years, 3 months ago

The right answer is D, I tested it in Synapse and it's the only one that actually runs without an error
upvoted 29 times

✉ **wijaz789** Highly Voted 2 years, 4 months ago

-- Find data skew for a distributed table
DBCC PDW_SHOWSPACEUSED('dbo.FactInternetSales');

upvoted 18 times

✉ **ItHYMeRlsh** 2 years ago

This will only work if you connect to the dedicated pool. The answer you've chosen says you are connecting to the built-in (serverless) pool.
upvoted 10 times

✉ **d046bc0** Most Recent 3 weeks, 3 days ago

Selected Answer: D

dm_pdw_nodes_db_partition_stats because we need to verify it on Pool1 (not built-in pool!)
upvoted 1 times

✉ **kkk5566** 4 months, 1 week ago

Selected Answer: D

dm_pdw_nodes_db_partition_stats
upvoted 1 times

✉ **kkk5566** 4 months ago

You can use DBCC PDW_SHOWSPACEUSED to find the skew, however only on dedicated pools.
upvoted 2 times

✉ **vctrhugo** 6 months, 3 weeks ago

Use sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data.
upvoted 1 times

✉ **vctrhugo** 6 months, 2 weeks ago

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet#distributed-or-replicated-tables>
upvoted 1 times

✉ **aemilka** 8 months, 3 weeks ago

Selected Answer: D

Correct answer is D.

A quick way to check for data skew is to use DBCC PDW_SHOWSPACEUSED, but DBCC PDW_SHOWSPACEUSED is not supported by serverless SQL pool in Azure Synapse Analytics. So A option can't be performed.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>

The only correct option here is to check sys.dm_pdw_nodes_db_partition_stats using dedicated SQL pool.
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

upvoted 2 times

✉ **JG1984** 6 months, 3 weeks ago

DBCC PDW_SHOWSPACEUSED is a command that can be used to show space usage information for a Database in an Azure Synapse Analytics dedicated SQL pool. However, it is not the best option for identifying data skew in a specific table.

upvoted 2 times

□ **Okea** 11 months, 1 week ago

A quick way to check for data skew is to use DBCC PDW_SHOWSPACEUSED.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>

upvoted 2 times

□ **Lestrang** 11 months, 2 weeks ago

This has been explained by others, but not clear enough to get it. I certainly had to look around and ponder for a bit. So, to give a more lucid explanation for why this is D and why the later question is DBCC PDW_SHOWSPACEUSED , it comes down to the small differences.

You can use DBCC PDW_SHOWSPACEUSED to find the skew, however only on dedicated pools. Well if you are like me, you would be shouting WELL THE QUESTION SAID DEDICATED POOL DUH. But if you read it carefully, it says connect to the "built-in pool" AKA serverless pool and run DBCC PDW_SHOWSPACEUSED.

Well, we ain't in a serverless pool are we? so that leaves D as the solution.

in the other question the given answers are so

- A. Connect to Pool1 and run DBCC PDW_SHOWSPACEUSED.
- B. Connect to the built-in pool and run DBCC PDW_SHOWSPACEUSED.
- C. Connect to Pool1 and run DBCC CHECKALLOC.
- D. Connect to the built-in pool and query sys.dm_pdw_sys_info.

Here we see that db_partition_stats is in a built in, which is a no go, so obviously we use PDW_SHOWSPACEUSED.

Hopefully this help any airheaded kindred spirits.

upvoted 8 times

□ **youngbug** 11 months, 3 weeks ago

A is a quicker way, but you can run DBCC in a serverless SQL pool, the built-in pool.

upvoted 2 times

□ **steve7** 12 months ago

Right answer is A. DBCC PDW_SHOWSPACEUSED. google it

upvoted 1 times

□ **Deeksha1234** 1 year, 5 months ago

Selected Answer: D

D is correct

upvoted 4 times

□ **Franz58** 1 year, 5 months ago

I think that first we need to connect to Pool 1, this excludes the first two options (and especially DBCC PDW_SHOWSPACEUSED). In the other two options, after connecting to Pool1, we execute query sys.dm_pdw_nodes_db_partition_stats.

upvoted 1 times

□ **StudentFromAus** 1 year, 6 months ago

Selected Answer: D

For dedicated SQL Pool this is the correct answer.

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

upvoted 2 times

□ **Andushi** 1 year, 8 months ago

Selected Answer: D

Use sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data.

ref: <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

upvoted 4 times

□ **FelixI** 1 year, 8 months ago

Selected Answer: A

DBCC PDW_SHOWSPACEUSED

upvoted 1 times

□ **AlCubeHead** 1 year, 9 months ago

Selected Answer: A

Firstly, this is for DEDICATED SQL Pool.

Here is what both likely outputs give you:

sys.dm_pdw_nodes_db_partition_stats:

object_id, partition_id, in_row_data_page_count, in_row_used_page_count

These columns are not useful in identifying skew

However, if you're using PDW_SHOWSPACEUSED:

ROWS, RESERVED_SPACE, DATA_SPACE, INDEX_SPACE, UNUSED_SPACE

These columns are definitely useful in identifying skew as you can calculate the Space allocation per row and look at any unused space

upvoted 1 times

□ **ladywhiteadder** 1 year, 9 months ago

Selected Answer: D

A does not work as in this answer we connect to the build in pool NOT the dedicated pool. This leaves D as valid option
upvoted 3 times

 **AIcubeHead** 1 year, 9 months ago

The question specifies dedicated Pool NOT Built-in Pool, so it is A
upvoted 1 times

 **Amsterliese** 1 year, 9 months ago

Please read the answer options carefully. In options A + B, you connect to the serverless SQL pool, in options C + D, you connect to the dedicated SQL pool.
upvoted 3 times

HOTSPOT -

You need to collect application metrics, streaming query events, and application log messages for an Azure Databrick cluster.

Which type of library and workspace should you implement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Library: 店铺:

| |
|---|
| Azure Databricks Monitoring Library |
| Microsoft Azure Management Monitoring Library |
| PyTorch |
| TensorFlow |

店铺: IT认证考试服务

Workspace:

| |
|------------------------|
| Azure Databricks |
| Azure Log Analytics |
| Azure Machine Learning |

Answer Area

Correct Answer:

Library:

| |
|---|
| Azure Databricks Monitoring Library |
| Microsoft Azure Management Monitoring Library |
| PyTorch |
| TensorFlow |

Workspace:

| |
|------------------------|
| Azure Databricks |
| Azure Log Analytics |
| Azure Machine Learning |

You can send application logs and metrics from Azure Databricks to a Log Analytics workspace. It uses the Azure Databricks Monitoring Library, which is available on GitHub.

Reference:

<https://docs.microsoft.com/en-us/azure/architecture/databricks-monitoring/application-logs>

店铺:
leandrors Highly Voted 2 years, 2 months ago

Correct!

upvoted 11 times

kkk5566 Most Recent 4 months ago

correct

upvoted 1 times

Igor85 1 year, 1 month ago

the solution works for databricks runtime 10.x only, though.
newer version isn't supported yet

upvoted 3 times

dmitriypo 1 year, 2 months ago

The given answer is correct
upvoted 2 times

Deeksha1234 1 year, 5 months ago

Correct

upvoted 3 times

 **wwdba** 1 year, 10 months ago

Correct!

upvoted 2 times

 **Start** 2 years, 3 months ago

Answer is correct

<https://docs.microsoft.com/en-us/azure/architecture/databricks-monitoring/application-logs>

upvoted 4 times

 **MFO_FM** 2 years, 3 months ago

is it correct

upvoted 1 times

You have a SQL pool in Azure Synapse.
You discover that some queries fail or take a long time to complete.
You need to monitor for transactions that have rolled back.
Which dynamic management view should you query?

- A. sys.dm_pdw_request_steps
- B. sys.dm_pdw_nodes_tran_database_transactions
- C. sys.dm_pdw_waits
- D. sys.dm_pdw_exec_sessions

Correct Answer: B

You can use Dynamic Management Views (DMVs) to monitor your workload including investigating query execution in SQL pool.

If your queries are failing or taking a long time to proceed, you can check and monitor if you have any transactions rolling back.

Example:

-- Monitor rollback

SELECT -

```
SUM(CASE WHEN t.database_transaction_next_undo_lsn IS NOT NULL THEN 1 ELSE 0 END), t.pdw_node_id, nod.[type]  
FROM sys.dm_pdw_nodes_tran_database_transactions t  
JOIN sys.dm_pdw_nodes nod ON t.pdw_node_id = nod.pdw_node_id  
GROUP BY t.pdw_node_id, nod.[type]
```

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor#monitor-transaction-log-rollback>

 **Podavenna** Highly Voted  2 years, 3 months ago

Correct Answer!
upvoted 15 times

 **kkk5566** Most Recent  4 months, 1 week ago

Selected Answer: B
correct
upvoted 1 times

 **Jerrie86** 11 months, 3 weeks ago

Rollback works with transactions. answer B
upvoted 3 times

 **nicky87654** 11 months, 4 weeks ago

Selected Answer: B
Correct Answer! B. sys.dm_pdw_nodes_tran_database_transactions
upvoted 3 times

 **dmitriypo** 1 year, 2 months ago

Selected Answer: B
The given answer is correct
upvoted 3 times

 **allagowf** 1 year, 2 months ago

Selected Answer: B
B. sys.dm_pdw_nodes_tran_database_transactions
upvoted 4 times

 **Deeksha1234** 1 year, 5 months ago

correct
upvoted 2 times

 **ladywhiteadder** 1 year, 9 months ago

Selected Answer: B

see <https://docs.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sys-dm-tran-database-transactions-transact-sql?view=sql-server-ver15>
upvoted 2 times

Question #11

Topic 4

You are monitoring an Azure Stream Analytics job.

You discover that the Backlogged Input Events metric is increasing slowly and is consistently non-zero.

You need to ensure that the job can handle all the events.

What should you do?

- A. Change the compatibility level of the Stream Analytics job.
- B. Increase the number of streaming units (SUs).
- C. Remove any named consumer groups from the connection and use \$default.
- D. Create an additional output stream for the existing input stream.

Correct Answer: B

Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job. You should increase the Streaming Units.

Note: 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.

Reference:

<https://docs.microsoft.com/bs-cyril-ba/azure/stream-analytics/stream-analytics-monitoring>

 **Lrng15** Highly Voted 2 years, 3 months ago

duplicate question. correct answer B

upvoted 15 times

 **Jerrie86** Highly Voted 11 months, 3 weeks ago

Selected Answer: B

Money is the answer to all problems. Answer B. increase SU units.

upvoted 11 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: B

answer B

upvoted 1 times

 **yogiazzaad** 11 months, 3 weeks ago

<https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-job-metrics>

This link is useful

upvoted 1 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: B

correct!

upvoted 1 times

 **snna4** 2 years ago

It's just a similar question. Proposed answers are different.

upvoted 1 times

 **Sudheer_K** 2 years, 3 months ago

Repeated

upvoted 4 times

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

| Table | Comment |
|-----------------------|---|
| EventDate | One million records are added to the table each day |
| EventTypeID | The table contains 10 million records for each event type. |
| WarehouseID | The table contains 100 million records for each warehouse. |
| ProductCategoryTypeID | The table contains 25 million records for each product category type. |

You identify the following usage patterns:

- Analysts will most commonly analyze transactions for a warehouse.
- Queries will summarize by product category type, date, and/or inventory event type.

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

On which column should you partition the table?

- A. EventTypeID
 B. ProductCategoryTypeID
 C. EventDate
 D. WarehouseID

Correct Answer: D

The number of records for each warehouse is big enough for a good partitioning.

Note: Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.

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 distributed databases.

 **Lio95** Highly Voted 2 years, 3 months ago

It is recommended to have at least 1 million rows per partition and distribution. Since there are 60 distributions, the number of rows for each partition must exceed 60 millions. Answer is correct

upvoted 26 times

 **yassine70** 2 years, 3 months ago

I fully Agree! Answer is correct

Link below :<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition>

"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 distributed databases.

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."

upvoted 7 times

 **LiamRT** 2 years, 1 month ago

Partitioning by EventDate does not mean a partition for each day. Partitioning by quarter years would be effective.

upvoted 1 times

 **Canary_2021** Highly Voted 2 years ago

Selected Answer: D

D is the correct answer.

Analysts will most commonly analyze transactions for a warehouse. This means that warehouseID is always in where clause. Partition field should be in where clause to improve query performance.

upvoted 17 times

 **Canary_2021** 2 years ago

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

upvoted 1 times

 **Matt2000** 4 months, 4 weeks ago

However the cheat sheet says: "In 99 percent of cases, the partition key should be based on date"

upvoted 1 times

✉ **dakku987** 1 week ago

but only when there is more than 1 billion records "You might partition your table when you have a large fact table (greater than 1 billion rows). In 99 percent of cases, the partition key should be based on date."

upvoted 1 times

✉ **kkk5566** [Most Recent] 4 months, 1 week ago

Selected Answer: D

is correct

upvoted 2 times

✉ **Karl_Cen** 11 months, 2 weeks ago

Selected Answer: C

I don't think the answer is right, the answer should be C, EventDate .

The total row number in this inventory updates table is determined before it's created. And here the question is asking us to chose the partition column, not distribution column.

upvoted 2 times

✉ **dmitriypo** 1 year, 2 months ago

Selected Answer: C

I would go for a date column since positions are most often created for a date column

upvoted 2 times

✉ **dmitriypo** 1 year, 2 months ago

Forget it. I agree with the provided answer D

upvoted 2 times

✉ **dom271219** 1 year, 4 months ago

Selected Answer: D

Tables ? These are the columns, aren't they ?

upvoted 1 times

✉ **Deeksha1234** 1 year, 5 months ago

D is right

upvoted 2 times

✉ **nefarious_smalls** 1 year, 7 months ago

Selected Answer: C

I will go C. We are querying about warehouses. Therefore I think the distribution column would have to be warehouse. If not then we would most likely have to do a shuffle to aggregate all the transactions for the same warehouse which would be spread out amongst the 60 distributions.

upvoted 1 times

✉ **Aditya0891** 1 year, 6 months ago

It's about partition not distribution. Read the question carefully first

upvoted 2 times

✉ **Dizzystar** 2 years, 2 months ago

I agree on date column. "In most cases, table partitions are created on a date column." <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition>

upvoted 1 times

✉ **ploer** 1 year, 11 months ago

But only in most cases. In most cases old data is not needed so date column often shows up in the where clause. This is why partitioning often makes sense on date columns. In this case the "Analysts will most commonly analyze transactions for a warehouse", so WarehouseID will be in the where clause and therefore we should partition on this column.

upvoted 2 times

✉ **sreejani** 2 years, 3 months ago

Aren't partition supposed to be done on columns of group by?. So here it's product type on which analysts summarise.so partition should be on productype

upvoted 2 times

✉ **Samanda** 2 years, 2 months ago

are you thinking of hash distributions instead of partitions?

upvoted 5 times

✉ **rikku33** 2 years, 3 months ago

For effective partitions its good to have one million rows per partitions for an ideal optimized scenario. This is also mentioned in the Microsoft documentation. C

upvoted 2 times

✉ **Samanda** 2 years, 2 months ago

You don't have to put each warehouse into it's own partition though so the sizing argument doesn't make sense....Answer is D as you will benefit from partition elimination when you use the warehouseID in the where clause

upvoted 2 times

✉ **sachabess79** 2 years, 3 months ago

WHERE is applied on the WarehouseID, so D

upvoted 6 times

✉ **YipingRuan** 2 years, 2 months ago

Nope, don't use WHERE

upvoted 2 times

✉ **mbl** 2 years, 2 months ago

it does : "Analysts will most commonly analyze transactions for a warehouse"

upvoted 3 times

✉ **AppleVan** 2 years, 3 months ago

I think it faster to go by date (C)....Otherwise, the query time will be extremely long since it has wrangled here and there...

upvoted 2 times

✉ **Amalbenrebai** 2 years, 3 months ago

can someone confirm this ?

upvoted 1 times

✉ **Samanda** 2 years, 2 months ago

It's 100% D

upvoted 3 times

✉ **rav009** 2 years, 3 months ago

I will go C

upvoted 3 times

You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date.

You need to ensure that the design provides the fastest query times of the records when querying for arbitrary date ranges and aggregating by fiscal calendar attributes.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Create a date dimension table that has a DateTime key.
- B. Use built-in SQL functions to extract date attributes.
- C. Create a date dimension table that has an integer key in the format of YYYYMMDD.
- D. In the fact table, use integer columns for the date fields.
- E. Use DateTime columns for the date fields.

Correct Answer: BD

 **echerish** Highly Voted 2 years, 4 months ago

Should be C and D

upvoted 63 times

 **anto69** 1 year, 11 months ago

Yup, that makes sense

upvoted 4 times

 **GervasioMontaNelas** Highly Voted 2 years, 4 months ago

100% CD

upvoted 14 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: CD

C&D. You can see some examples from modules of MS' dp-203 training.

upvoted 1 times

 **auwia** 6 months, 2 weeks ago

Selected Answer: CD

Because of fixed date ranges used to query.

upvoted 2 times

 **BPW** 7 months, 3 weeks ago

Should be A and E

upvoted 2 times

 **esaade** 9 months, 4 weeks ago

Selected Answer: AB

A. Create a date dimension table that has a DateTime key. A date dimension table that has a DateTime key can provide fast query times when querying for arbitrary date ranges and aggregating by fiscal calendar attributes. The DateTime key allows for easy sorting and filtering of dates, and can be used to join with the fact table on the order date, order due date, and order ship date fields.

B. Use built-in SQL functions to extract date attributes. Using built-in SQL functions to extract date attributes (such as year, quarter, month, week, day) from the DateTime key in the date dimension table can help with aggregating data by fiscal calendar attributes. This can improve query performance by reducing the amount of data that needs to be scanned and aggregated.

Therefore, the correct actions to perform are A and B.

upvoted 2 times

 **gogosgh** 8 months, 1 week ago

we are not querying against time. the fact table has only dates

upvoted 1 times

 **XiltroX** 1 year, 1 month ago

For sure its CD

upvoted 2 times

 **Xinyuehong** 1 year, 2 months ago

Selected Answer: CD

CD with no doubt.
upvoted 2 times

- Deeksha1234** 1 year, 5 months ago
correct - C&D agree with StudentFromAus M
upvoted 4 times

- StudentFromAus** 1 year, 6 months ago

Selected Answer: CD

The question has many clues, it states fiscal calendar year and then star schema which hints we need proper fact and dim tables and appropriate date keys to link these.
upvoted 5 times

- Davico93** 1 year, 6 months ago

Selected Answer: CD

basic knowledge for fact and dim tables
upvoted 3 times

- AIcubeHead** 1 year, 9 months ago

Selected Answer: CD

Who gives these answers?? It's so obviously C and D. You want a Date Dim with an Integer key and the fact table also with that integer key
upvoted 7 times

- wwdba** 1 year, 9 months ago

Should be CD!
upvoted 2 times

- Boumisasound** 1 year, 10 months ago

Selected Answer: CD

I'm agree for CD
upvoted 2 times

- ovokpus** 1 year, 10 months ago

Selected Answer: AE

this makes the most sense
upvoted 2 times

- kanak01** 1 year, 11 months ago

Selected Answer: CD

C & D should be correct
upvoted 2 times

- bahamutedean** 1 year, 11 months ago

should be CD
upvoted 2 times

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices.

The company must be able to monitor the devices in real-time.

You need to design the solution.

What should you recommend?

- A. Azure Analysis Services using Azure Portal
- B. Azure Analysis Services using Azure PowerShell
- C. Azure Stream Analytics cloud job using Azure Portal
- D. Azure Data Factory instance using Microsoft Visual Studio

Correct Answer: C

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. Your Stream Analytics job would ingest these events from Event Hubs and run real-time analytics queries against the streams.

Create a Stream Analytics job:

In the Azure portal, select + Create a resource from the left navigation menu. Then, select Stream Analytics job from Analytics.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-get-started-with-azure-stream-analytics-to-process-data-from-iot-devices>

✉ **Raghul108** Highly Voted 1 year, 11 months ago

Repeated question
upvoted 10 times

✉ **PallaviPatel** Highly Voted 1 year, 11 months ago

Selected Answer: C
C is correct
upvoted 8 times

✉ **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: C
C is correct
upvoted 1 times

✉ **ZIMARAKI** 12 months ago

Selected Answer: C
C is correct
upvoted 2 times

✉ **Deeksha1234** 1 year, 5 months ago

correct
upvoted 1 times

✉ **Sabajamal2010AtGmail** 2 years ago

C is correct
upvoted 4 times

You have a SQL pool in Azure Synapse.

A user reports that queries against the pool take longer than expected to complete. You determine that the issue relates to queried columnstore segments.

You need to add monitoring to the underlying storage to help diagnose the issue.

Which two metrics should you monitor? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Snapshot Storage Size
- B. Cache used percentage
- C. DWU Limit
- D. Cache hit percentage

Correct Answer: BD

D: Cache hit percentage: $(\text{cache hits} / \text{cache miss}) * 100$ where cache hits is the sum of all columnstore segments hits in the local SSD cache and cache miss is the columnstore segments misses in the local SSD cache summed across all nodes

B: $(\text{cache used} / \text{cache capacity}) * 100$ where cache used is the sum of all bytes in the local SSD cache across all nodes and cache capacity is the sum of the storage capacity of the local SSD cache across all nodes

Incorrect Answers:

C: DWU limit: Service level objective of the data warehouse.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-resource-utilization-query-activity>

✉  **kkk5566** 4 months, 1 week ago

Selected Answer: BD

go b &D

upvoted 1 times

✉  **Matt2000** 5 months ago

This link might be useful. It explains cache hit percentage and cache used percentage:

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-resource-utilization-query-activity>

upvoted 1 times

✉  **yogiazaad** 11 months, 3 weeks ago

This article is more relevant here.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monitor-cache>

upvoted 3 times

✉  **Deeksha1234** 1 year, 5 months ago

Selected Answer: BD

seems correct

upvoted 4 times

✉  **PallaviPatel** 1 year, 11 months ago

Selected Answer: BD

Correct Answer

upvoted 3 times

✉  **HaBroNounen** 2 years ago

correct

upvoted 3 times

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 percentage
- B. Cache hit percentage
- C. DWU limit 銷:
- D. Data IO percentage

Correct Answer: B

Monitor and troubleshoot slow query performance by determining whether your workload is optimally leveraging the adaptive cache for dedicated SQL pools.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monitor-cache>

✉  **HaBroNounen** Highly Voted 2 years ago

correct

upvoted 6 times

✉  **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: B

repeated

upvoted 1 times

✉  **Deeksha1234** 1 year, 5 months ago

correct

upvoted 4 times

You have an Azure Databricks resource.

You need to log actions that relate to changes in compute for the Databricks resource.

Which Databricks services should you log?

- A. clusters
- B. workspace
- C. DBFS
- D. SSH
- E. jobs

Correct Answer: B

Databricks provides access to audit logs of activities performed by Databricks users, allowing your enterprise to monitor detailed Databricks usage patterns.

There are two types of logs:

- ☞ Workspace-level audit logs with workspace-level events.
- ☞ Account-level audit logs with account-level events.

Reference:

<https://docs.databricks.com/administration-guide/account-settings/audit-logs.html>

✉  **azure9876** Highly Voted 2 years ago

It shall be A:Clusters, workspace logs does not have any cluster related resource change.

upvoted 23 times

✉  **Abakwagirl** 6 months, 2 weeks ago

A cluster is defined within the workspace and cluster events are logged at the workspace level. See "Cluster Events" in the following doc:<https://docs.databricks.com/administration-guide/account-settings/audit-logs.html>

upvoted 3 times

✉  **Deeksha1234** Highly Voted 1 year, 5 months ago

Selected Answer: A

A is correct

upvoted 5 times

✉  **matiandal** Most Recent 2 months ago

Answer : A.Clusters

why not workspace ?

Workspace is not a service that you should log to track changes in compute for the Databricks resource because it does not record events related to creating, editing, deleting, starting, or stopping clusters or jobs. Workspace events are related to actions performed on the workspace itself, such as creating, renaming, deleting, or importing notebooks, folders, libraries, or repos. These events do not affect the compute resources used by the Databricks resource, but rather the workspace content and configuration.

Therefore, workspace is not a relevant service for logging compute changes.

upvoted 1 times

✉  **kkk5566** 4 months, 1 week ago

Selected Answer: A

A is correct

upvoted 1 times

✉  **Ram9198** 4 months, 4 weeks ago

<https://learn.microsoft.com/en-us/azure/databricks/administration-guide/account-settings/audit-log-delivery>

MS document says cluster

upvoted 1 times

✉  **Ram9198** 6 months ago

Selected Answer: A

Cluster is the only compute first class resource

upvoted 3 times

✉  **vctrhugo** 6 months, 3 weeks ago

Selected Answer: A

A. Clusters: Databricks clusters are the primary compute resources in Azure Databricks. Monitoring and logging cluster-related actions will help you track changes in cluster creation, termination, resizing, and other cluster-related activities.

upvoted 2 times

 **Ast999** 10 months, 1 week ago

Selected Answer: A

100% SURE A IS A CORRECT ANSWER.

upvoted 3 times

 **demirsamuel** 1 year, 7 months ago

Selected Answer: A

definitely A

upvoted 3 times

 **upliftinghut** 1 year, 7 months ago

Selected Answer: B

Workspace is correct. Detail is here:

Set-AzDiagnosticSetting -ResourceId \$databricks.ResourceId -WorkspaceId \$logAnalytics.ResourceId -Enabled \$true -name "<diagnostic setting name>" -Category <comma separated list>

Link: <https://docs.microsoft.com/en-us/azure/databricks/administration-guide/account-settings/azure-diagnostic-logs#configure-diagnostic-log-delivery>

upvoted 1 times

 **Mckay_** 1 year, 7 months ago

I thought compute is related to cluster.

upvoted 3 times

 **KashRaynardMorse** 1 year, 8 months ago

Selected Answer: A

Answer: A (clusters)

Despite using workspace to enable logging, from there you need to select clusters from the list if you want to satisfy the "changes in compute for the Databricks resource" question, hence the service you should log is clusters. See link from Amsterliese.

Beware of links to databricks.com vs links to microsoft because they are two slightly different products (i.e. Databricks (on AWS) vs Azure Databricks).

For the other comment referencing dp200; the answer description only gives the definitions but no explanation.

upvoted 4 times

 **Deeksha1234** 1 year, 5 months ago

agree A should be the answer

upvoted 2 times

 **Amsterliese** 1 year, 9 months ago

From what I understand from MS documentation, it should be

A - clusters

<https://docs.microsoft.com/en-us/azure/databricks/administration-guide/account-settings/azure-diagnostic-logs#configure-diagnostic-log-delivery>

The links in previous comments here which support answer B - workspace refer to AWS databricks. I tried to find a similar setup in the MS documentation, but couldn't find anything. Please tell me if my thinking is wrong. (Always happy to learn ;)

upvoted 2 times

 **ovokpus** 1 year, 10 months ago

Selected Answer: A

Agreed with clusters!

upvoted 2 times

 **kanak01** 1 year, 11 months ago

A clusters

upvoted 1 times

 **svik** 1 year, 11 months ago

Selected Answer: A

compute is related to the cluster

upvoted 3 times

 **PallaviPatel** 1 year, 11 months ago

Selected Answer: A

A is correct answer, as compute relates to clusters.

upvoted 4 times

You are designing a highly available Azure Data Lake Storage solution that will include geo-zone-redundant storage (GZRS).

You need to monitor for replication delays that can affect the recovery point objective (RPO).

What should you include in the monitoring solution?

- A. 5xx: Server Error errors
- B. Average Success E2E Latency
- C. availability
- D. Last Sync Time

Correct Answer: D

Because geo-replication is asynchronous, it is possible that data written to the primary region has not yet been written to the secondary region at the time an outage occurs. The Last Sync Time property indicates the last time that data from the primary region was written successfully to the secondary region. All writes made to the primary region before the last sync time are available to be read from the secondary location. Writes made to the primary region after the last sync time property may or may not be available for reads yet.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/common/last-sync-time-get>

 **PallaviPatel** Highly Voted 1 year, 11 months ago

Selected Answer: D

D is correct as we need to see impact on rpo to know that we need to see when was last sync carried out.

upvoted 5 times

 **Nivas2401** Highly Voted 1 year, 10 months ago

Selected Answer: D

<https://docs.microsoft.com/en-us/azure/storage/common/last-sync-time-get?tabs=azure-powershell>

upvoted 5 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: D

last sync time

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: D

D. Last Sync Time: Monitoring the Last Sync Time metric will provide you with the information about the timestamp of the last successful synchronization between the primary and secondary regions. By monitoring this metric, you can determine if there are any replication delays impacting the RPO. If the Last Sync Time is significantly behind the current time, it indicates a replication delay and potential RPO impact.

upvoted 1 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: D

agree Last sync time is right

upvoted 4 times

 **ovokpus** 1 year, 10 months ago

Selected Answer: D

last sync time

upvoted 4 times

 **ANath** 1 year, 11 months ago

Answer is D.

<https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy?toc=/azure/storage/blobs/toc.json#check-the-last-sync-time-property>

upvoted 5 times

 **Fer079** 2 years ago

Selected Answer: B

The key word in this question is "monitor", It means that we would have to see the output over time, so the correct answer should be B. Average Success E2E Latency. In this way we can monitor the spent time for each replication

<https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blob-scalable-app-verify-metrics>

upvoted 3 times

 **jv2120** 2 years ago

Answer is D. See below why not B.

Any blob, file, queue, or table operation latency can cause cascading slowdowns in your application. The Success E2E Latency metric measures the total amount of time it takes for requests to be processed by the storage account APIs, sent to the client, and then acknowledged by the client.

upvoted 4 times

You configure monitoring for an Azure Synapse Analytics implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Storage Gen2 using an external table.

Files with an invalid schema cause errors to occur.

You need to monitor for an invalid schema error.

For which error should you monitor?

- A. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [com.microsoft.polybase.client.KerberosSecureLogin] occurred while accessing external file.'
- B. Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11" for linked server "(null)". Query aborted- the maximum reject threshold (0 rows) was reached while reading from an external source: 1 rows rejected out of total 1 rows processed.
- C. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [Unable to instantiate LoginClass] occurred while accessing external file.'
- D. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [No FileSystem for scheme: wasbs] occurred while accessing external file.'

Correct Answer: B

Error message: Cannot execute the query "Remote Query"

Possible Reason:

The reason this error happens is because each file has different schema. The PolyBase external table DDL when pointed to a directory recursively reads all the files in that directory. When a column or data type mismatch happens, this error could be seen in SSMS.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-errors-and-possible-solutions>

 **kilowd** Highly Voted 1 year, 11 months ago

Selected Answer: B

<https://techcommunity.microsoft.com/t5/datacat/polybase-setup-errors-and-possible-solutions/ba-p/305297>

upvoted 6 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: B

B is correct

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: B

To monitor for an invalid schema error when using PolyBase to load data from CSV files in Azure Data Lake Storage Gen2 using an external table in Azure Synapse Analytics, you should monitor the following error:

B. Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11" for linked server "(null)". Query aborted- the maximum reject threshold (0 rows) was reached while reading from an external source: 1 rows rejected out of total 1 rows processed.

This error indicates that a row in the CSV file has been rejected due to an invalid schema. By monitoring for this error, you can identify when data loading fails due to an incompatible or incorrect schema in the CSV files.

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

Selected Answer: B

B is correct

upvoted 3 times

 **Raghul108** 1 year, 11 months ago

Selected Answer: B

Correct

upvoted 4 times

 **PallaviPatel** 1 year, 11 months ago

Selected Answer: B

correct

upvoted 4 times

 **HaBroNounen** 2 years ago

correct

upvoted 3 times

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

You have an Azure Synapse Analytics dedicated SQL pool.

You run PDW_SHOWSPACEUSED('dbo.FactInternetSales'); and get the results shown in the following table.

| ROWS | RESERVED_SPACE | DATA_SPACE | INDEX_SPACE | UNUSED_SPACE | PDW_NODE_ID | DISTRIBUTION_ID |
|------|----------------|------------|-------------|--------------|-------------|-----------------|
| 694 | 2776 | 616 | 48 | 2112 | 1 | 1 |
| 407 | 2704 | 576 | 48 | 2080 | 1 | 2 |
| 53 | 2376 | 512 | 16 | 1848 | 1 | 3 |
| 58 | 2376 | 512 | 16 | 1848 | 1 | 4 |
| 168 | 2632 | 528 | 32 | 2072 | 1 | 5 |
| 195 | 2696 | 536 | 32 | 2128 | 1 | 6 |
| 5995 | 3464 | 1424 | 32 | 2008 | 1 | 7 |
| 0 | 2232 | 496 | 0 | 1736 | 1 | 8 |
| 264 | 2576 | 544 | 40 | 1992 | 1 | 9 |
| 3008 | 3016 | 960 | 32 | 2024 | 1 | 10 |
| ... | ... | ... | ... | ... | ... | ... |
| 1550 | 2832 | 752 | 48 | 2032 | 1 | 50 |
| 1238 | 2832 | 696 | 40 | 2096 | 1 | 51 |
| 192 | 2632 | 528 | 32 | 2072 | 1 | 52 |
| 1127 | 2768 | 680 | 48 | 2040 | 1 | 53 |
| 1244 | 3032 | 704 | 64 | 2264 | 1 | 54 |
| 409 | 2632 | 568 | 32 | 2032 | 1 | 55 |
| 0 | 2232 | 496 | 0 | 1736 | 1 | 56 |
| 1437 | 2832 | 728 | 40 | 2064 | 1 | 57 |
| 0 | 2232 | 496 | 0 | 1736 | 1 | 58 |
| 384 | 2632 | 560 | 32 | 2040 | 1 | 59 |
| 225 | 2768 | 544 | 40 | 2184 | 1 | 60 |

Which statement accurately describes the dbo.FactInternetSales table?

- A. All distributions contain data.
- B. The table contains less than 10,000 rows.
- C. The table uses round-robin distribution.
- D. The table is skewed.

Correct Answer: D

Data skew means the data is not distributed evenly across the distributions.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>

 Deeksha1234 Highly Voted 1 year, 5 months ago

Selected Answer: D

D is correct

upvoted 5 times

 ANath Highly Voted 1 year, 12 months ago

I think the answer is correct because in some cases the rows are zero.

upvoted 5 times

 kkk5566 Most Recent 4 months ago

Selected Answer: D

D is correct

upvoted 1 times

 kkk5566 4 months, 1 week ago

Selected Answer: D

You can use DBCC PDW_SHOWSPACEUSED to find the skew, however only on dedicated pools.

upvoted 2 times

 StudentFromAus 1 year, 6 months ago

Selected Answer: D

Answer is correct
upvoted 4 times

 agar 1 year, 10 months ago

Selected Answer: D

correct

upvoted 3 times

 Raghu108 1 year, 11 months ago

Selected Answer: D

Correct

upvoted 3 times

 PallaviPatel 1 year, 11 months ago

Selected Answer: D

correct as few distributions have more data and few have no data at all. The data should be evenly distributed across all the distributions.
upvoted 3 times

 Deeksha1234 1 year, 5 months ago

Agree !

upvoted 2 times

You have two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.

| Table | Column |
|---------|------------------|
| Flight | ArrivalAirportID |
| | ArrivalDateTime |
| Weather | AirportID |
| | ReportDateTime |

You need to recommend a solution that maximizes query performance.

What should you include in the recommendation?

- A. In the tables use a hash distribution of ArrivalDateTime and ReportDateTime.
- B. In the tables use a hash distribution of ArrivalAirportID and AirportID.
- C. In each table, create an IDENTITY column.
- D. In each table, create a column as a composite of the other two columns in the table.

Correct Answer: B

Hash-distribution improves query performance on large fact tables.

Incorrect Answers:

A: Do not use a date column for hash distribution. 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.

✉  **PallaviPatel** Highly Voted 1 year, 11 months ago

Selected Answer: B

correct

upvoted 5 times

✉  **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: B

correct

upvoted 1 times

✉  **75082SN** 1 year ago

Why not D?

upvoted 2 times

✉  **shoottheduck** 10 months, 2 weeks ago

Also, a composite key does not improve performance on its own.

Distributing on the two columns that are joined, will

upvoted 2 times

✉  **sensaint** 1 year ago

Then you are partly distributing on a date column which is very bad for performance.

upvoted 2 times

✉  **Deeksha1234** 1 year, 5 months ago

B seems correct but not sure what's wrong with D ?

upvoted 3 times

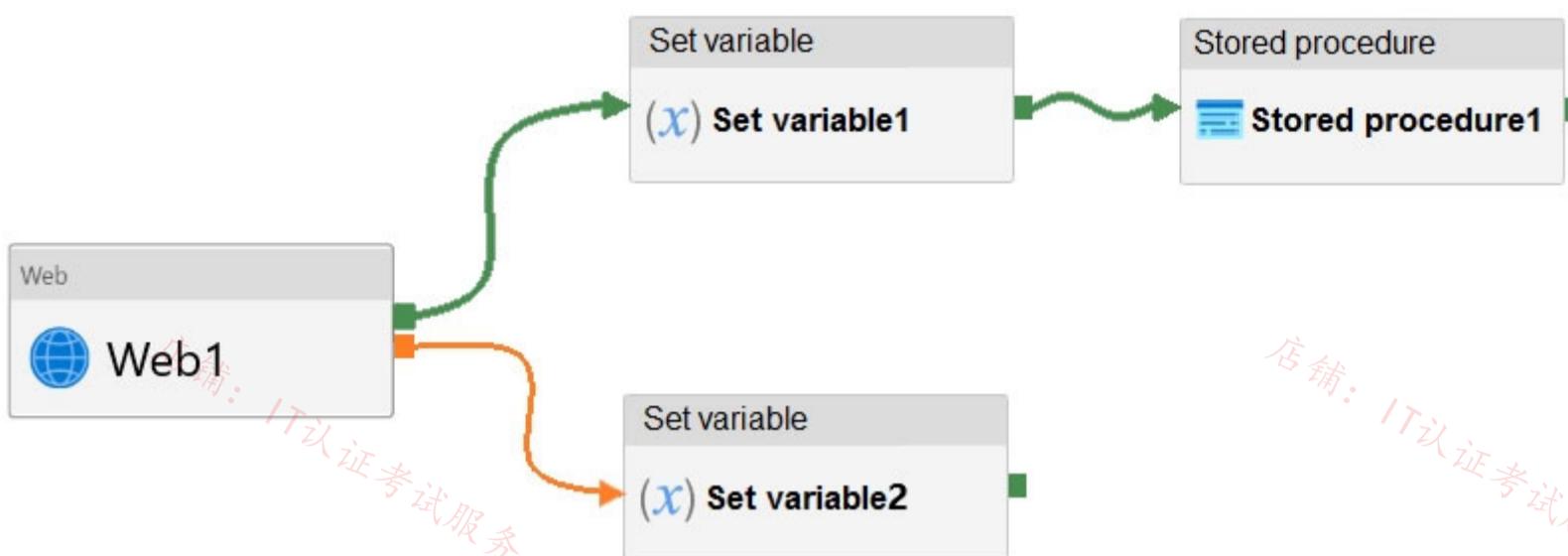
✉  **bad_atitude** 2 years ago

B is correct

upvoted 3 times

HOTSPOT -

You have an Azure Data Factory pipeline that has the activities 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:

Answer Area

Stored procedure1 will execute Web1 and Set variable1 [answer choice]

| |
|----------------------------------|
| <input type="button" value="▼"/> |
| complete |
| fail |
| succeed |

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]

| |
|----------------------------------|
| <input type="button" value="▼"/> |
| Canceled |
| Failed |
| Succeeded |

Correct Answer:

Answer Area

Stored procedure1 will execute Web1 and Set variable1 [answer choice]

| |
|----------------------------------|
| <input type="button" value="▼"/> |
| complete |
| fail |
| succeed |

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]

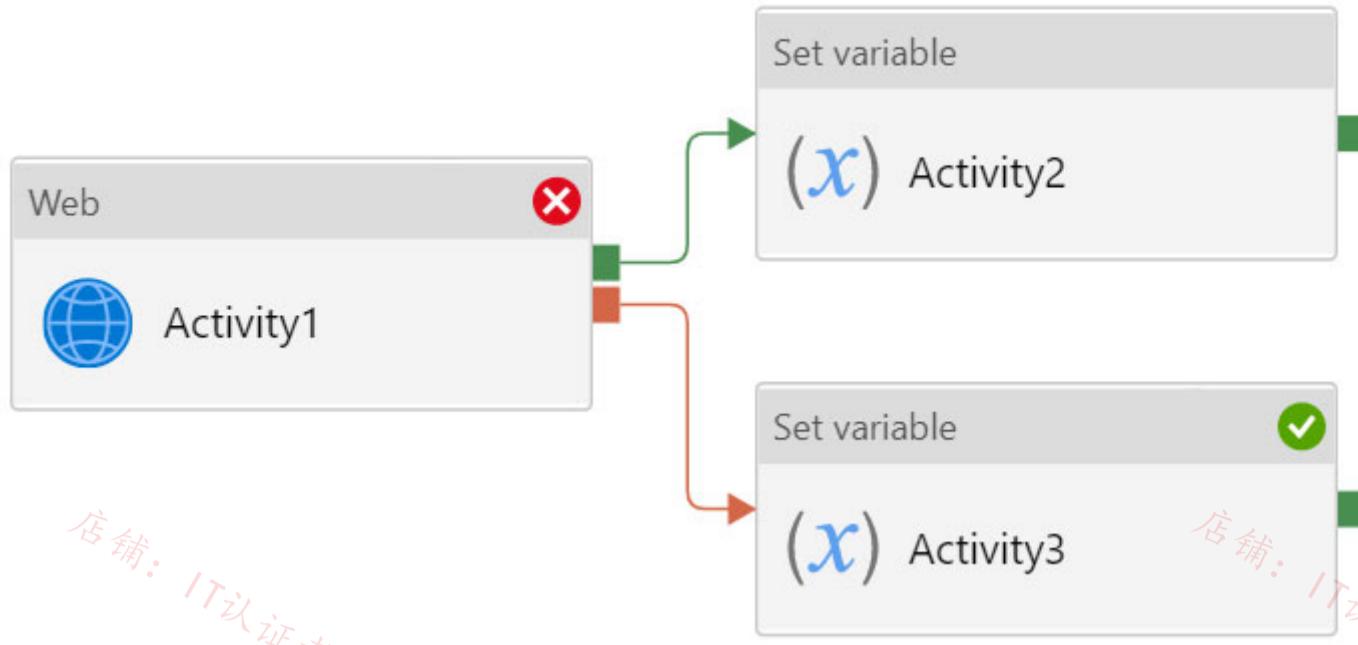
| |
|----------------------------------|
| <input type="button" value="▼"/> |
| Canceled |
| Failed |
| Succeeded |

Box 1: succeed

Box 2: failed

Example:

Now let's say we have a pipeline with 3 activities, where Activity1 has a success path to Activity2 and a failure path to Activity3. If Activity1 fails and Activity3 succeeds, the pipeline will fail. The presence of the success path alongside the failure path changes the outcome reported by the pipeline, even though the activity executions from the pipeline are the same as the previous scenario.



Activity1 fails, Activity2 is skipped, and Activity3 succeeds. The pipeline reports failure.

Reference:

<https://datasavvy.me/2021/02/18/azure-data-factory-activity-failures-and-pipeline-outcomes/>

□ **ItHYMeRish** Highly Voted 2 years ago

The answers are correct.

The second question is "failed" because web1 has both a success and failed path. web1 would have to have only a failed path for the second question to be considered successful.

upvoted 41 times

□ **XiltroX** 1 year, 1 month ago

The second answer should be "Succeeded". You are providing false information to other members. The reason why it is a success is because Set Variable 2 happened because of the failure of Web 1. Therefore, this red pipeline is deemed a success.

upvoted 21 times

□ **y154707** 2 months ago

Yes, as Wanchihh mentions, you are wrong.

Upper branch status are fail => skipped => skipped. According to the logic in the url below, this is deemed as a failed pipeline.

upvoted 2 times

□ **wanchihh** 3 months, 2 weeks ago

You are incorrect.

<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-pipeline-failure-error-handling#do-if-else-block>

upvoted 3 times

□ **Avi_Bdj** 1 year, 9 months ago

Second should also be succeeded.

upvoted 10 times

□ **a03** 2 years ago

Agree. Second is "Fail" because Success connector presented.

upvoted 4 times

□ **HaBroNounen** 2 years ago

I just tested it myself. Provided answers are correct

upvoted 10 times

□ **RajBathani** Highly Voted 2 years ago

The second answer should be Succeeded as 'Set Variable 2' has failed dependency on Web1.

upvoted 38 times

□ **dakku987** Most Recent 1 week ago

I think both are successful

bcz i think when web activity fail it will pass to the set variable and the purpose of the set variable will be COMPLETED so pipeline will be success
upvoted 1 times

□ **Abdulwahab1983** 1 month, 2 weeks ago

Both should be succeed

<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-pipeline-failure-error-handling#summary-table>

upvoted 1 times

□ **Abdulwahab1983** 1 month, 4 weeks ago

Do If Skip Else block

In this approach, customer defines the business logic, and defines both the Upon Failure path, and Upon Success path, with a dummy Upon Skipped activity attached. This approach renders pipeline succeeds, if Upon Failure path succeeds.

upvoted 1 times

□ **kkk5566** 4 months, 1 week ago

1. Success and 2. Failed

upvoted 2 times

□ **chryckie** 8 months, 2 weeks ago

The answer is correct! It's actually pretty neat how ADF determines that.

If an activity fails but there was a subsequent OnSuccess activity that never runs, it's a fail. To handle that, you also need an OnSkipped activity to follow the OnSuccess activity in case it never ran!

<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-pipeline-failure-error-handling#do-if-else-block>

upvoted 8 times

□ **Vanq69** 3 months ago

There are so many wrong high voted answers. READ THIS.

upvoted 1 times

□ **JoannaMar** 3 months, 2 weeks ago

Thanks @chryckie for this explanation. Finally it's clear!

upvoted 1 times

□ **AHUI** 9 months, 1 week ago

second box should be succeeded

<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-pipeline-failure-error-handling#do-if-skip-else-block>

upvoted 4 times

□ **vrodriguesp** 11 months, 4 weeks ago

Using this microsft doc: <https://learn.microsoft.com/en-us/azure/data-factory/tutorial-pipeline-failure-error-handling#try-catch-block> that claims

""We determine pipeline success and failures as follows:

-)Evaluate outcome for all leaves activities. If a leaf activity was skipped, we evaluate its parent activity instead

-)Pipeline result is success if and only if all nodes evaluated succeed""

I used this logic

When web1 activity fails: node setVariable2 succeeds and setVariable1 is skipped and its parent node web1 failed; overall pipeline fails

upvoted 6 times

□ **csd** 1 year, 4 months ago

In any scenario pipeline will show success status, cause we are catching the failure

upvoted 2 times

□ **StudentFromAus** 1 year, 6 months ago

The answers are correct.

upvoted 2 times

□ **datnguye** 2 years ago

It should be Succeeded in both.

The reference article says: The failure dependency means this pipeline reports success.

upvoted 14 times

□ **datnguye** 2 years ago

Updated: Correct ans as 1. Success and 2. Failed

The failure dependency means this pipeline reports success.

But, the presence of the success path alongside the failure path changes the outcome reported by the pipeline: Web-1 fails, Set-var-1 is skipped, and Set-var-2 succeeds --> The pipeline reports failure.

upvoted 14 times

□ **Remedios79** 1 year, 6 months ago

I agree with you too

upvoted 1 times

□ **ladywhiteadder** 1 year, 9 months ago

See <https://docs.microsoft.com/en-us/azure/data-factory/tutorial-pipeline-failure-error-handling#do-if-else-block>

upvoted 6 times

□ **ROLLINGROCKS** 1 year, 5 months ago

This is all you need for the right answer. Its well explained in the link.

upvoted 1 times

 **Yohannesmulu** 1 year, 9 months ago

Agreed!
upvoted 1 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

You have several Azure Data Factory pipelines that contain a mix of the following types of activities:

- Wrangling data flow
- Notebook
- Copy
- Jar

Which two Azure services should you use to debug the activities? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point

- A. Azure Synapse Analytics
- B. Azure HDInsight
- C. Azure Machine Learning
- D. Azure Data Factory
- E. Azure Databricks

Correct Answer: AC

 KrishIC Highly Voted 2 years ago

Selected Answer: DE

Notebook- azure databricks, managing activities in pipeline-datafactory
upvoted 36 times

 ElHomo2222 Highly Voted 1 year, 12 months ago

Selected Answer: DE

D & E; Databricks for Wrangling and Notebooks; ADF for Copy and Jar
upvoted 16 times

 kilowd 1 year, 11 months ago

Wrangling and Copy = ADF
Jar and Notebooks = Databricks
upvoted 11 times

 kkk5566 Most Recent 4 months, 1 week ago

Selected Answer: DE

is correct
upvoted 1 times

 auwia 6 months, 2 weeks ago

Selected Answer: DE

Wrangling and Copy = ADF
Jar and Notebooks = Databricks
upvoted 2 times

 vctrhugo 6 months, 3 weeks ago

Selected Answer: DE

D. Azure Data Factory: Azure Data Factory itself provides debugging capabilities for its activities. You can monitor and debug the execution of pipeline activities directly within the Azure Data Factory interface. It allows you to view activity run details, input/output data, logs, and diagnose any errors or issues encountered during execution.

E. Azure Databricks: Azure Databricks is a powerful analytics platform that integrates well with Azure Data Factory. You can use it to debug and analyze Notebook activities within the Data Factory pipelines. Azure Databricks provides an interactive environment to run and debug notebooks, allowing you to inspect intermediate data, execute code step-by-step, and troubleshoot any issues.

upvoted 3 times

 janaki 7 months, 2 weeks ago

Selected Answer: DE

D - Azure Data Factory
E - Azure Databricks
upvoted 3 times

 pavankr 7 months, 2 weeks ago

You "de-bug" the activity with ML??? Seriously??? come on man??? from where you are getting these answers???

upvoted 2 times

 **Mohamedali.Cintellic** 8 months, 2 weeks ago

Selected Answer: DE

D & E are correct
upvoted 2 times

 **vrodriguesp** 11 months, 4 weeks ago

Selected Answer: DE

Notebook on azure databricks, rest on pipeline data factory. No sense for AandC
upvoted 3 times

 **nicky87654** 1 year ago

Selected Answer: DE

Wrangling and Copy = ADF
Jar and Notebooks = Databricks
upvoted 4 times

 **Deeksha1234** 1 year, 4 months ago

Selected Answer: DE

should be DE
upvoted 2 times

 **martinamartina** 1 year, 5 months ago

Couldn't be AD?
upvoted 1 times

 **dsp17** 1 year, 5 months ago

Selected Answer: DE

DE - correct
upvoted 1 times

 **dsp17** 1 year, 5 months ago

Selected Answer: DE

Wrangling and Copy -> ADF
Jar and Notebooks -> Databricks
upvoted 1 times

 **Remedios79** 1 year, 6 months ago

Selected Answer: DE

absolutely D&E
upvoted 2 times

 **Remedios79** 1 year, 6 months ago

D and E absolutely!!
upvoted 1 times

 **VenkataPolepalli** 1 year, 7 months ago

Selected Answer: DE

Answer is D&E
upvoted 2 times

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1. You need to identify the extent of the data skew in Table1. What should you do in Synapse Studio?

- A. Connect to the built-in pool and run sys.dmv_nodes_db_partition_stats.
- B. Connect to Pool1 and run DBCC CHECKALLOC.
- C. Connect to the built-in pool and run DBCC CHECKALLOC.
- D. Connect to Pool1 and query sys.dmv_nodes_db_partition_stats.

Correct Answer: D

Microsoft recommends use of sys.dmv_nodes_db_partition_stats to analyze any skewness in the data.

Reference:

<https://docs.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sys-dm-db-partition-stats-transact-sql>

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

✉ **Lotusss** Highly Voted 1 year, 8 months ago

Correct. See Question 12 topic 4

upvoted 10 times

✉ **kkk5566** 4 months, 1 week ago

Question 8 topic 4

upvoted 1 times

✉ **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: D

Correct. See Question 12 topic 4

upvoted 1 times

✉ **Matt2000** 4 months, 4 weeks ago

Does sys.dmv_nodes_db_partition_stats exist? I found, however, found a reference for sys.dmv_db_partition_stats that seems to do the trick.
upvoted 1 times

✉ **niaspa** 6 months, 1 week ago

Selected Answer: D

D .See Question 12 topic 4

upvoted 2 times

✉ **vctrhugo** 6 months, 3 weeks ago

Selected Answer: D

D. Connect to Pool1 and query sys.dmv_nodes_db_partition_stats.

By connecting to Pool1, which represents the dedicated SQL pool, and querying the sys.dmv_nodes_db_partition_stats system view, you can obtain information about the distribution of data across the compute nodes in the SQL pool. This view provides details on the number of rows and the size of data partitions on each node, allowing you to identify any significant data skew in Table1.

upvoted 2 times

✉ **bulutfet** 7 months, 1 week ago

D correct

upvoted 1 times

✉ **janaki** 7 months, 2 weeks ago

Selected Answer: A

Option A is correct

upvoted 1 times

✉ **vctrhugo** 6 months, 3 weeks ago

How could you use built-in (serverless) to query dedicated pool?

upvoted 2 times

✉ **janaki** 7 months, 2 weeks ago

Sorry, option D is correct.

upvoted 3 times

Deeksha1234 1 year, 4 months ago
correct
upvoted 3 times

Question #25

Topic 4

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. Local tempdb percentage
- B. Cache used percentage
- C. Data IO percentage
- D. CPU percentage

Correct Answer: B

Monitor and troubleshoot slow query performance by determining whether your workload is optimally leveraging the adaptive cache for dedicated SQL pools.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monitor-cache>

juanlu46 Highly Voted 1 year, 8 months ago

Selected Answer: B

Is correct
upvoted 6 times

kkk5566 Most Recent 4 months, 1 week ago

Selected Answer: B

repeated
upvoted 1 times

vctrhugo 6 months, 3 weeks ago

Selected Answer: B

Repeated question.
upvoted 2 times

StudentFromAus 1 year, 6 months ago

Selected Answer: B

For already used queries, we need to monitor the adaptive caching
upvoted 4 times

You have an Azure data factory.

You need to examine the pipeline failures from the last 180 days.

What should you use?

- A. the Activity log blade for the Data Factory resource
- B. Pipeline runs in the Azure Data Factory user experience
- C. the Resource health blade for the Data Factory resource
- D. Azure Data Factory activity runs in Azure Monitor

Correct Answer: D

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor>

juanlu46 Highly Voted 1 year, 8 months ago

Selected Answer: D

Correct!

upvoted 7 times

kkk5566 Most Recent 4 months, 1 week ago

Selected Answer: D

repeated

upvoted 1 times

vctrhugo 6 months, 3 weeks ago

Selected Answer: D

If you see anything above 45 days involving logs on ADF, it won't be ADF itself.

upvoted 2 times

Jerrie86 11 months, 2 weeks ago

Asking to monitor Pipeline failures and D is activity runs. so Can't be D. Looks like they are missing an answer here
upvoted 1 times

dom271219 1 year, 4 months ago

Selected Answer: D

Redundant question

upvoted 1 times

Deeksha1234 1 year, 4 months ago

Selected Answer: D

correct

upvoted 3 times

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices.

The company must be able to monitor the devices in real-time.

You need to design the solution.

What should you recommend?

- A. Azure Analysis Services using Azure PowerShell
- B. Azure Stream Analytics Edge application using Microsoft Visual Studio
- C. Azure Analysis Services using Microsoft Visual Studio
- D. Azure Data Factory instance using Azure Portal

Correct Answer: B

Azure Stream Analytics on IoT Edge empowers developers to deploy near-real-time analytical intelligence closer to IoT devices so that they can unlock the full value of device-generated data.

You can use Stream Analytics tools for Visual Studio to author, debug, and create your Stream Analytics Edge jobs. After you create and test the job, you can go to the Azure portal to deploy it to your devices.

Incorrect:

Not A, not C: Azure Analysis Services is a fully managed platform as a service (PaaS) that provides enterprise-grade data models in the cloud. Use advanced mashup and modeling features to combine data from multiple data sources, define metrics, and secure your data in a single, trusted tabular semantic data model.

Reference:

<https://docs.microsoft.com/en-us/azure/iot-hub/monitor-iot-hub> <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-tools-for-visual-studio-edge-jobs>

 **pperf** 3 months ago

Q14 Topic4

Same question but different options.

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: B

repeated

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: B

Azure IoT Hub = Stream

upvoted 1 times

 **nicky87654** 11 months, 4 weeks ago

Selected Answer: B

Azure Stream Analytics Edge application using Microsoft Visual Studio

Azure Stream Analytics is a real-time data streaming service that allows you to analyze and process data streams in near real-time. The Stream Analytics Edge application can be deployed on IoT devices, such as those used to monitor manufacturing machinery, to enable real-time monitoring and analysis of the data generated by the devices. Stream Analytics Edge allows you to run Stream Analytics jobs on IoT

upvoted 4 times

 **Shanmahi** 1 year, 1 month ago

Selected Answer: B

Reasons for choosing option B --> IoT devices, streaming data, real-time data requirement

upvoted 2 times

 **MadhuMDLK1055** 1 year, 1 month ago

Ans is D

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Azure Data Factory is primarily a data integration service designed for orchestrating and managing data workflows, such as data movement and transformation.

While ADF can be used for data ingestion and processing, it is not optimized for real-time scenarios. Azure Data Factory works based on scheduled or triggered data pipelines rather than real-time streaming data processing.

upvoted 1 times

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

店铺：IT认证考试服务

You have an Azure Synapse Analytics dedicated SQL pool named SA1 that contains a table named Table1.

You need to identify tables that have a high percentage of deleted rows.

What should you run?

- A. sys.pdw_nodes_column_store_segments
- B. sys.dm_db_column_store_row_group_operational_stats
- C. sys.pdw_nodes_column_store_row_groups
- D. sys.dm_db_column_store_row_group_physical_stats

Correct Answer: C

Use sys.pdw_nodes_column_store_row_groups to determine which row groups have a high percentage of deleted rows and should be rebuilt.

Note: sys.pdw_nodes_column_store_row_groups provides clustered columnstore index information on a per-segment basis to help the administrator make system management decisions in Azure Synapse Analytics. sys.pdw_nodes_column_store_row_groups has a column for the total number of rows physically stored

(including those marked as deleted) and a column for the number of rows marked as deleted.

Incorrect:

Not A: You can join sys.pdw_nodes_column_store_segments with other system tables to determine the number of columnstore segments per logical table.

Not B: Use sys.dm_db_column_store_row_group_operational_stats to track the length of time a user query must wait to read or write to a compressed rowgroup or partition of a columnstore index, and identify rowgroups that are encountering significant I/O activity or hot spots.

 **greenlever** Highly Voted 1 year, 2 months ago

Selected Answer: C

has a column for the total number of rows physically stored (including those marked as deleted) and a column for the number of rows marked as deleted. Use sys.pdw_nodes_column_store_row_groups to determine which row groups have a high percentage of deleted rows and should be rebuilt

upvoted 5 times

 **d046bc0** Most Recent 3 weeks, 2 days ago

Selected Answer: D

<https://learn.microsoft.com/en-us/sql/relational-databases/system-catalog-views/sys-pdw-nodes-column-store-row-groups-transact-sql?view=aps-pdw-2016-au7>

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: D

D is correct

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

change to C

<https://learn.microsoft.com/en-us/sql/relational-databases/system-catalog-views/sys-pdw-nodes-column-store-row-groups-transact-sql?view=aps-pdw-2016-au7>

upvoted 3 times

 **andie123** 5 months, 2 weeks ago

Selected Answer: D

D is correct answer

upvoted 1 times

 **andie123** 5 months, 2 weeks ago

The sys.dm_db_column_store_row_group_physical_stats dynamic management view provides information about the physical characteristics of row groups in columnstore indexes, including the number of deleted rows in each row group. You can use this view to identify tables that have a high percentage of deleted rows by calculating the ratio of deleted rows to total rows for each table. -> D is the answer

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: C

Use sys.pdw_nodes_column_store_row_groups to determine which row groups have a high percentage of deleted rows and should be rebuilt.

<https://learn.microsoft.com/en-us/sql/relational-databases/system-catalog-views/sys-pdw-nodes-column-store-row-groups-transact-sql?view=aps-pdw-2016-au7>

upvoted 2 times

 **dimbrici** 1 year, 1 month ago

Selected Answer: C

C is the correct Answer !

upvoted 3 times

 **anks84** 1 year, 4 months ago

Selected Answer: C

C is the correct Answer !

upvoted 3 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

You have an enterprise data warehouse in Azure Synapse Analytics.

You need to monitor the data warehouse to identify whether you must scale up to a higher service level to accommodate the current workloads.

Which is the best metric to monitor?

More than one answer choice may achieve the goal. Select the BEST answer.

- A. DWU used
- B. CPU percentage
- C. DWU percentage
- D. Data IO percentage

Correct Answer: A

DWU used: $DWU\ limit * DWU\ percentage$

DWU used represents only a high-level representation of usage across the SQL pool and is not meant to be a comprehensive indicator of utilization. To determine whether to scale up or down, consider all factors which can be impacted by DWU such as concurrency, memory, tempdb, and adaptive cache capacity. We recommend running your workload at different DWU settings to determine what works best to meet your business objectives.

Azure Synapse Analytics monitor metric "DWU used"

Incorrect:

- * CPU percentage: CPU utilization across all nodes for the data warehouse.
- * DWU percentage: Maximum between CPU percentage and Data IO percentage
- * Data IO percentage: IO Utilization across all nodes for the data warehouse

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-resource-utilization-query-activity>

 **chryckie** Highly Voted 8 months, 2 weeks ago

Selected Answer: C

It must be DWU percentage. e.g. 95% is bad and 99% is very bad, and you don't need to look at anything else.

If you looked at DWU used, what can you infer without also knowing the DWU limit (or DWU percentage)?

upvoted 12 times

 **markpumc** Highly Voted 9 months, 4 weeks ago

C. DWU percentage is the best metric to monitor to identify whether you must scale up to a higher service level to accommodate the current workloads in Azure Synapse Analytics. DWU percentage measures the percentage of Data Warehouse Units (DWUs) in use, which indicates how much processing power is being used. If the DWU percentage consistently exceeds a certain threshold, it may be necessary to scale up to a higher service level to accommodate the workload. DWU used, CPU percentage, and Data IO percentage are also important metrics to monitor, but they do not directly reflect the overall processing power available in the data warehouse.

upvoted 8 times

 **Momoanwar** Most Recent 2 weeks, 1 day ago

Selected Answer: C

Chatgpt :

Knowing the absolute number of DWUs used (option A) doesn't provide complete information unless you also know the total DWUs available. On the other hand, the DWU percentage directly indicates how much of the available compute capacity is being used, which is a more informative metric for deciding whether scaling is needed.

Therefore, the best metric to monitor to identify whether you need to scale up to a higher service level to accommodate the current workloads would indeed be:

- C. DWU percentage

upvoted 1 times

 **jiriz** 2 months, 4 weeks ago

Selected Answer: C

C - DWU Percentage
upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: C

C is the best.

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

DWU percentage is the percentage of Data Warehouse Units (DWUs) used by the data warehouse. It is calculated as the maximum of CPU percentage and Data IO percentage. If the DWU percentage is consistently high, it may indicate that you need to scale up to a higher service level to accommodate the current workloads

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: C

Percentage should be way more useful than units itself.

upvoted 2 times

 **henryphchan** 7 months, 3 weeks ago

Selected Answer: C

i vote for C because only the % used is meaningful

upvoted 4 times

 **rohitbinnani** 8 months, 1 week ago

Selected Answer: C

I 100% agree with C. How will you know by a UNIT value if it's sufficient or not? You would need to check the percentage consumed out of total capacity, right? Hence, in my logical and design views it must be C --> DWU Percentage.

upvoted 4 times

 **Shanmahi** 1 year, 1 month ago

Selected Answer: A

DWU used is the metric to use, if only one best answer is expected. option A.

upvoted 2 times

 **shaileshutd** 1 year, 1 month ago

Selected Answer: A

As given in the document and explanation, DWU used = DWU limit * DWU percentage, it comprises limit and percentage.

The question also states that more than one answer may achieve the goal and we are supposed to select the best answer, I think DWU used gives the best metric.

upvoted 4 times

 **Tickxit** 1 year, 1 month ago

Which is the best one, DWU used or DWU percentage? We need to select one.

upvoted 1 times

 **CodingOwl** 1 year, 2 months ago

AC Both are answers

upvoted 1 times

 **Phund** 1 year, 3 months ago

should be both DWU metrics

upvoted 2 times

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices.

The company must be able to monitor the devices in real-time.

You need to design the solution.

What should you recommend?

- A. Azure Analysis Services using Azure PowerShell
- B. Azure Data Factory instance using Azure PowerShell
- C. Azure Stream Analytics cloud job using Azure Portal
- D. Azure Data Factory instance using Microsoft Visual Studio

Correct Answer: C

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. Your Stream Analytics job would ingest these events from Event Hubs and run real-time analytics queries against the streams.

Create a Stream Analytics job:

In the Azure portal, select + Create a resource from the left navigation menu. Then, select Stream Analytics job from Analytics.

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-get-started-with-azure-stream-analytics-to-process-data-from-iot-devices>

 **kkk5566** 4 months, 1 week ago

Selected Answer: C

correct

upvoted 1 times

 **maximilianogarcia6** 1 year, 1 month ago

this question is not repeated as options are different. It could appear the first one or this.

upvoted 3 times

 **AdarshKumarKhare** 1 year, 2 months ago

Question is a repeat

upvoted 1 times

 **sensaint** 1 year, 2 months ago

Selected Answer: C

Correct. Repeated question.

upvoted 3 times

HOTSPOT -

You have an Azure event hub named retailhub that has 16 partitions. Transactions are posted to retailhub. Each transaction includes the transaction ID, the individual line items, and the payment details. The transaction ID is used as the partition key.

You are designing an Azure Stream Analytics job to identify potentially fraudulent transactions at a retail store. The job will use retailhub as the input. The job will output the transaction ID, the individual line items, the payment details, a fraud score, and a fraud indicator.

You plan to send the output to an Azure event hub named fraudhub.

You need to ensure that the fraud detection solution is highly scalable and processes transactions as quickly as possible.

How should you structure the output of the Stream Analytics job? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

店铺: IT认证考试服务

Answer Area

Number of partitions:

| |
|----|
| 1 |
| 8 |
| 16 |
| 32 |

Partition key:

| |
|-----------------------|
| Fraud indicator |
| Fraud score |
| Individual line items |
| Payment details |
| Transaction ID |

Answer Area

Number of partitions:

| |
|----|
| 1 |
| 8 |
| 16 |
| 32 |

Correct Answer:

Partition key:

| |
|-----------------------|
| Fraud indicator |
| Fraud score |
| Individual line items |
| Payment details |
| Transaction ID |

Box 1: 16 -

For Event Hubs you need to set the partition key explicitly.

An embarrassingly parallel job is the most scalable scenario in Azure Stream Analytics. It connects one partition of the input to one instance of the query to one partition of the output.

Box 2: Transaction ID -

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features#partitions>

 **Preben** Highly Voted 2 years, 7 months ago

Correct.

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization>

Embarrassingly parallel jobs

Step 3 and 4.

upvoted 42 times

✉ **Liz42** 2 years, 2 months ago

The step 4 you've mentioned, @Preben, says: "The number of input partitions must equal the number of output partitions". The documentation continues to talk about scenarios that are not embarrassingly parallel like @Maunik has mentioned below

upvoted 1 times

✉ **Liz42** 2 years, 2 months ago

Disregard my above comment... meant to respond to another

upvoted 2 times

✉ **Momoanwar** [Most Recent] 2 weeks, 1 day ago

Correct cgtapgt :

For high scalability and quick processing in Azure Stream Analytics, it's important to align the output event hub partitions with the input source. Since the input event hub `retailhub` has 16 partitions, the output event hub `fraudhub` should also have 16 partitions to match. This ensures that the partitioning scheme is consistent and can handle the volume of transactions efficiently.

The partition key should be the 'Transaction ID', as this will ensure that all the events for a particular transaction will go to the same partition, maintaining the order of events which is crucial for transactional data and fraud detection scenarios.

So the correct answers are:

Number of partitions: 16

Partition key: Transaction ID

upvoted 1 times

✉ **kkk5566** 4 months, 1 week ago

correct

upvoted 1 times

✉ **_Lukas_** 5 months, 1 week ago

Number of partitions: Since the input event hub retailhub has 16 partitions, it makes sense to have the same number of partitions in the output event hub fraudhub to align the partitions. So the number of partitions should be 16.

Partition key: Since the transaction ID was used as the partition key in the input event hub, using the same partition key in the output event hub ensures that the data for the same transaction ID is processed by the same partition in both event hubs. This makes the flow of data from one event hub to the other more efficient. So the partition key should be the Transaction ID.

upvoted 1 times

✉ **Deeksha1234** 1 year, 5 months ago

correct

upvoted 3 times

✉ **nelineli** 1 year, 6 months ago

"A per-device or user unique identity makes a good partition key, but other attributes such as geography can also be used to group related events into a single partition."

upvoted 3 times

✉ **sdokmak** 1 year, 7 months ago

Event Hub -> Event Hub: x:x partitions

Event Hub -> Blob Storage: x:1 partitions or x:y partitions

Blob Storage -> Event Hub: x:x partitions

Blob Storage -> Blob Storage: x:1 partitions

upvoted 2 times

✉ **Maunik** 2 years, 4 months ago

Example of scenarios that are not embarrassingly parallel

Mismatched partition count

Input: Event hub with 8 partitions

Output: Event hub with 32 partitions

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization>

Should be 8 partitions based on link above

upvoted 2 times

✉ **Aditya0891** 1 year, 7 months ago

Maunik it did mention there that it results to "some level of parallelization". So I don't think this is the best option to choose if you have equal number of partitions (i.e 16 here) in your options

upvoted 1 times

✉ **nichag** 2 years, 5 months ago

Shouldn't the number of partitions only be 8, since the question only asks about the output?

upvoted 1 times

✉ **rumosgf** 2 years, 7 months ago

Why 16? Don't understand...

upvoted 2 times

✉  **wwdba** 1 year, 10 months ago

An embarrassingly parallel job is the most scalable scenario in Azure Stream Analytics. It connects one partition of the input to one instance of the query to one partition of the output.

The number of input partitions must equal the number of output partitions.

upvoted 2 times

✉  **mbravo** 2 years, 7 months ago

Embarrassingly parallel jobs

upvoted 10 times

✉  **captainbee** 2 years, 6 months ago

It's not THAT embarrassing

upvoted 10 times

✉  **Davico93** 1 year, 6 months ago

There are 2 eventhub, first has 16 partitions and the number of partitions asked is for the second eventhub, and both must be equals for better performance

upvoted 2 times

HOTSPOT -

You have an on-premises data warehouse that includes the following fact tables. Both tables have the following columns: DateKey, ProductKey, RegionKey.

There are 120 unique product keys and 65 unique region keys.

| Table | Comments |
|---------|---|
| Sales | The table is 600 GB in size. DateKey is used extensively in the WHERE clause in queries. ProductKey is used extensively in join operations. RegionKey is used for grouping. Severity-five percent of records relate to one of 40 regions. |
| Invoice | The table is 6 GB in size. DateKey and ProductKey are used extensively in the WHERE clause in queries. RegionKey is used for grouping. |

Queries that use the data warehouse take a long time to complete.

You plan to migrate the solution to use Azure Synapse Analytics. You need to ensure that the Azure-based solution optimizes query performance and minimizes processing skew.

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

NOTE: Each correct selection is worth one point

Hot Area:

Answer Area

| Table | Distribution type | Distribution column |
|-----------|---|---|
| Sales: | <input type="checkbox"/> Hash-distributed
<input type="checkbox"/> Round-robin | <input type="checkbox"/> DateKey
<input type="checkbox"/> ProductKey
<input type="checkbox"/> RegionKey |
| Invoices: | <input type="checkbox"/> Hash-distributed
<input type="checkbox"/> Round-robin | <input type="checkbox"/> DateKey
<input type="checkbox"/> ProductKey
<input type="checkbox"/> RegionKey |

Answer Area

| Table | Distribution type | Distribution column |
|-----------|--|--|
| Sales: | <input checked="" type="checkbox"/> Hash-distributed
<input type="checkbox"/> Round-robin | <input checked="" type="checkbox"/> DateKey
<input checked="" type="checkbox"/> ProductKey
<input checked="" type="checkbox"/> RegionKey |
| Invoices: | <input checked="" type="checkbox"/> Hash-distributed
<input type="checkbox"/> Round-robin | <input checked="" type="checkbox"/> DateKey
<input checked="" type="checkbox"/> ProductKey
<input checked="" type="checkbox"/> RegionKey |

Box 1: Hash-distributed -

Box 2: ProductKey -

ProductKey is used extensively in joins.

Hash-distributed tables improve query performance on large fact tables.

Box 3: Hash-distributed -

Box 4: RegionKey -

Round-robin tables are useful for improving loading speed.

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 not 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
- ☞ When the table is a temporary staging table

Note: A distributed table appears as a single table, but the rows are actually stored across 60 distributions. The rows are distributed with a hash or round-robin algorithm.

Reference:

<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute>

□  **lara_mia1** Highly Voted  2 years, 7 months ago

1. Hash Distributed, ProductKey because >2GB and ProductKey is extensively used in joins
2. Hash Distributed, RegionKey because "The table size on disk is more than 2 GB." and you have to chose a distribution column which: "Is not used in WHERE clauses. This could narrow the query to not run on all the distributions."

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

upvoted 101 times

□  **vblissings** 2 years, 5 months ago

i agree

upvoted 3 times

□  **Marcello83** 2 years, 6 months ago

I agree with lara_mia1

upvoted 3 times

□  **niceguy0371** 2 years, 4 months ago

Disagree on nr. 1 because of the reason you give for nr. 2. (choose a distribution column that is not used in where clauses. A join is also a where clause

upvoted 4 times

□  **sdokmak** 1 year, 7 months ago

nah mate, check out his link:

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 that is frequently in the GROUP BY clause.

Is not used in WHERE clauses. This could narrow the query to not run on all the distributions.

Is not a date column. WHERE clauses often filter by date. When this happens, all the processing could run on only a few distributions.

upvoted 3 times

□  **Rob77** Highly Voted  2 years, 7 months ago

Both hash as both are > 2GB. In the 2nd table RegionKey cannot be used with round_robin distribution as round_robin does not take a distribution key...

upvoted 28 times

□  **ploer** 1 year, 11 months ago

Correct: "A round-robin distributed table distributes table rows evenly across all distributions. The assignment of rows to distributions is random. Unlike hash-distributed tables, rows with equal values are not guaranteed to be assigned to the same distribution."

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>

upvoted 1 times

□  **kkk5566** Most Recent  4 months, 1 week ago

1. Hash Distributed, ProductKey
2. Hash Distributed, RegionKey

upvoted 2 times

□  **Ram9198** 4 months, 3 weeks ago

When two large fact tables have frequent joins - in this case one is large and another is a small dimension table. Hence highlighted answer is correct

upvoted 1 times

□  **dom271219** 1 year, 3 months ago

"Choose a distribution column with data that distributes evenly"

ProductKey is more relevant in both cases

upvoted 4 times

□ **Deeksha1234** 1 year, 5 months ago

1. Hash Distributed, ProductKey because table size >2GB and ProductKey is extensively used in joins . another, region key could have been considered (after join key which is product key) since its being used in grouping but 75% records belongs to one region so - NO for region key.

2. Hash Distributed, RegionKey because the table size on disk is more than 2 GB and Its being used in grouping (for this table more than 75% record doesn't fall in same region) and you have to chose a distribution column which is not used in WHERE clause.

upvoted 5 times

□ **Nishikag** 1 year, 6 months ago

To minimize data movement, select a distribution column that:

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 that is frequently in the GROUP BY clause.

Is not used in WHERE clauses. This could narrow the query to not run on all the distributions.

Is not a date column. WHERE clauses often filter by date. When this happens, all the processing could run on only a few distributions.

upvoted 2 times

□ **Remedios79** 1 year, 6 months ago

the provided answers are correct

upvoted 1 times

□ **kiranSargar** 1 year, 10 months ago

Generally facts table are hash distributed. so both the table should use hash distribution and distribution key would be product_key for both.

upvoted 1 times

□ **DarioEtna** 2 years, 5 months ago

as for me i guess this is the right choice:

1. Hash Distributed, RegionKey because

2. Hash Distributed, RegionKey because

"When two large fact tables have frequent joins, query performance improves when you distribute both tables on one of the join columns"
[Microsoft Documentation]

If we use for one ProductKey and for one RegionKey maybe the data movements would increase...or not?

upvoted 3 times

□ **DarioEtna** 2 years, 5 months ago

But we cannot use ProductKey in both because in Invoice table it is used in WHERE condition

upvoted 4 times

□ **Lucky_me** 2 years ago

If we choose RegionKey for Sales, we would have a processing skew.

upvoted 4 times

□ **Aditya0891** 1 year, 6 months ago

DarioEtna where in the question is it mentioned that both tables will be used together in a join query? They have different set of columns in where and group by, so why are you so sure that they will be used together? Answers provided are correct here

upvoted 1 times

□ **Amalbenrebai** 2 years, 5 months ago

Regarding the invoices table, we can use the Round-robin distribution because there is no obvious joining key in the table

upvoted 2 times

□ **zarga** 2 years, 6 months ago

1. Hash on product key

2. Hash on region key (used on group by and have 65 unique values)

upvoted 9 times

□ **BrennaFrenna** 2 years, 7 months ago

The sales table makes sense with hashing distribution on ProductKey and since there is no obvious joining key for invoices, you should use round robin distribution on RegionKey. When it would be a smaller table you should use replicated.

upvoted 3 times

□ **tubis** 2 years, 7 months ago

When it says 75% of records related to one of the 40 regions, if we partition the Sales by Region, isn't it improve the reading process drastically in compare to productKey?

upvoted 1 times

□ **Preben** 2 years, 7 months ago

That's 75 % of 61 % of the regions that will be done effectively. That's only efficient for 45 % of the queries. Not a whole lot.

upvoted 2 times

□ **patricka95** 2 years, 5 months ago

No, if 75% relate to one region and we hash on region, that means that those will all be on one node and there will be skew. Correct answers are Hash, Product, Hash, Region.

upvoted 3 times

 **bc5468521** 2 years, 7 months ago
I AGREE WITH BOTH HASH WITH PRODUCT KEY
upvoted 10 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

You have a partitioned table in an Azure Synapse Analytics dedicated SQL pool.

You need to design queries to maximize the benefits of partition elimination.

What should you include in the Transact-SQL queries?

- A. JOIN
- B. WHERE
- C. DISTINCT
- D. GROUP BY

Correct Answer: B

 **elimey** Highly Voted 2 years, 5 months ago

correct

upvoted 8 times

 **SG1705** Highly Voted 2 years, 7 months ago

Why ??

upvoted 6 times

 **okechi** 2 years, 6 months ago

Why ?? Because When you add the "WHERE" clause to your T-SQL query it allows the query optimizer accesses only the relevant partitions to satisfy the filter criteria of the query - which is what partition elimination is all about.

upvoted 43 times

 **noranathalie** 2 years, 2 months ago

In question 2, we just mentionned to not use the where condition columns to create partitions.. so the logic is unclear for me..

upvoted 2 times

 **noranathalie** 2 years, 2 months ago

please disregard my comment above. Partitioning is different from hash-column, so the criterias are different

upvoted 4 times

 **IgorLacik** 2 years, 6 months ago

Maybe this? <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization>

I think I read somewhere in the docs that you cannot apply complex queries on partition filtering, cannot find it though (not much help I guess, but hopefully better than nothing)

upvoted 1 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: B

correct

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: B

To maximize the benefits of partition elimination in Azure Synapse Analytics dedicated SQL pool, you should include the WHERE clause in your Transact-SQL queries.

The WHERE clause allows you to specify conditions that filter the rows returned by a query. When designing queries for partitioned tables, you can include predicates in the WHERE clause that align with the partitioning scheme. By doing so, the query optimizer can leverage partition elimination to exclude unnecessary partitions from the query execution plan.

Partition elimination is the process of excluding partitions from query processing based on the predicates specified in the WHERE clause. By eliminating partitions that do not contain relevant data, the query performance can be significantly improved.

upvoted 2 times

 **Deeksha1234** 1 year, 5 months ago

correct, agree with okechi

upvoted 1 times

 **dsp17** 1 year, 6 months ago

100% Correct. Think of it this way, you have 36 partitions over Month column for a table. You are interested in a specific month. so in WHERE clause of your select statement, you will give specific month to "eliminate" other 35 partitions scan.

upvoted 4 times

⊕  **ploer** 1 year, 11 months ago

A is surely true. But B also. If you have two tables small a and big B and you're joining them on condition a.some_column = b.some_column big table B would be filtered by the values found in a. An if B is partitioned on "some_column" we have the same effect as with the where clause.

upvoted 1 times

⊕  **kilowd** 1 year, 11 months ago

Selected Answer: B

B is Correct

Data partition elimination refers to the database server's ability to determine, based on query predicates

upvoted 1 times

⊕  **Canary_2021** 2 years ago

what's the difference between distribution and partition? I don't find any doc online to describe it clearly.

- Horizontal partitioning divides a table into multiple tables that contain the same number of columns.
- A distributed table appears as a single table, but the rows are actually stored across 60 distributions.

If a table have both distribution and Horizontal partition, how are data stored in SQL? For example a customer table, hash-distributed by region and Horizontal Partitioned by year of the activation data.

upvoted 2 times

⊕  **Lucky_me** 2 years ago

<https://stackoverflow.com/questions/51677471/what-is-a-difference-between-table-distribution-and-table-partition-in-sql/51677595>

upvoted 4 times

⊕  **sparkchu** 1 year, 9 months ago

distribution is a generally used technique for Massive Distributed Computing. we explicitly decide which distribution pattern to be used in Azure DWH, while Hadoop/Hive automatically distributes the table when created.

upvoted 1 times

You have an Azure Stream Analytics query. The query returns a result set that contains 10,000 distinct values for a column named clusterID.

You monitor the Stream Analytics job and discover high latency.

You need to reduce the latency.

Which two actions should you perform? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Add a pass-through query.
- B. Increase the number of streaming units.
- C. Add a temporal analytic function.
- D. Scale out the query by using PARTITION BY.
- E. Convert the query to a reference query.

Correct Answer: BD

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption> <https://docs.microsoft.com/en-us/azure/stream-analytics/repartition>

 **allagowf** Highly Voted 1 year, 2 months ago

Selected Answer: BD

key word: contains 10,000 distinct values for a column named clusterID --> PARTITION.
reduce the latency --> Increase SU + it refer to PARTITION too.

upvoted 8 times

 **vctrhugo** Highly Voted 6 months, 3 weeks ago

Selected Answer: BD

To reduce latency in an Azure Stream Analytics job with a query returning a result set containing 10,000 distinct values for a column named clusterID, you should perform the following actions:

B. Increase the number of streaming units:

Increasing the number of streaming units allocates more resources to your Stream Analytics job, allowing it to handle higher data volumes and processing loads. By increasing the streaming units, you can improve the job's throughput and reduce latency.

D. Scale out the query by using PARTITION BY:

Using the PARTITION BY clause in your query allows you to distribute the workload across multiple partitions or parallel processes. By partitioning the data based on relevant criteria, such as clusterID in this case, you can distribute the processing load and reduce latency by enabling parallel processing.

upvoted 6 times

 **kkk5566** Most Recent 4 months ago

Selected Answer: BD

correct

upvoted 1 times

 **hassexat** 4 months ago

Selected Answer: BD

B & D are correct

upvoted 1 times

 **rzeng** 1 year, 2 months ago

correct

upvoted 3 times

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1. You need to identify the extent of the data skew in Table1. What should you do in Synapse Studio?

- A. Connect to the built-in pool and query sys.dmw_nodes_db_partition_stats.
- B. Connect to the built-in pool and run DBCC CHECKALLOC.
- C. Connect to Pool1 and query sys.dmw_node_status.
- D. Connect to Pool1 and query sys.dmw_nodes_db_partition_stats.

Correct Answer: A

Use sys.dmw_nodes_db_partition_stats to analyze any skewness in the data.

Use it on the built-in pool, not on Pool1.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

⊕ **ank84** Highly Voted 1 year, 4 months ago

Selected Answer: D

Correct answer is D.

upvoted 6 times

⊕ **Ngol** Highly Voted 11 months, 4 weeks ago

I don't understand why Exam Topics should be giving different answers for questions they have repeated...like this one!

upvoted 5 times

⊕ **ShrikantW** Most Recent 3 weeks ago

D is the correct answer!

upvoted 1 times

⊕ **kkk5566** 4 months, 1 week ago

Selected Answer: D

repetd

upvoted 1 times

⊕ **bp_a_user** 8 months, 2 weeks ago

Its D!

Official Learning path>Returns page and row-count information for every partition in the current database.
nodes_db_partition_stats

<https://learn.microsoft.com/en-us/training/modules/analyze-optimize-data-warehouse-storage-azure-synapse-analytics/2-understand-skewed-data-space-usage>

upvoted 1 times

⊕ **zizonesol** 9 months, 4 weeks ago

We had the same question before. The correct answer is D

upvoted 4 times

⊕ **Vikram1710** 10 months, 2 weeks ago

Option A is confusing as we have different answer for same question.

upvoted 1 times

⊕ **vrodriguesp** 11 months, 3 weeks ago

Selected Answer: D

Answer is not so clear!, because I can't see any refernece on built-in pool. What is built-in pool?

Anyway looking at the doc here:

<https://learn.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sys-dm-db-partition-stats-transact-sql?view=sql-server-ver16>

that claims: "This syntax is not supported by serverless SQL pool in Azure Synapse Analytics."

So if built-in pool is serverless SQL pool the correct answer should be D (Connect to Pool1 and query sys.dmw_nodes_db_partition_stats).
upvoted 5 times

⊕ **OldSchool** 1 year, 1 month ago

Selected Answer: D

It can't be A and B because those two are connecting to Built-In pool (serverless) and the Q is about dedicated pool.
upvoted 2 times

 **OldSchool** 1 year, 1 month ago

If it is the mistake in wording the question and instead of dedicated is serverless, then the answer is A.
upvoted 1 times

 **dimbrici** 1 year, 1 month ago

Selected Answer: D

Question already seen
upvoted 3 times

 **AdarshKumarKhare** 1 year, 2 months ago

Question repeated
upvoted 2 times

 **SD4592** 1 year, 3 months ago

Selected Answer: D
Absolutely D
upvoted 4 times

 **debarun** 1 year, 3 months ago

Correct answer is D.
upvoted 3 times

 **federC** 1 year, 4 months ago

Agree with anks84. Correct answer should be D, built-in pool comes from a Synapse Serverless pool and here it says Dedicated
upvoted 2 times

 **pangas2567** 1 year, 4 months ago

Selected Answer: D
The same question as #8 Topic #4, but different answer. Should be D.
upvoted 2 times

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

- A. Connect to Pool1 and DBCC PDW_SHOWSPACEUSED.
- B. Connect to the built-in pool and run DBCC PDW_SHOWSPACEUSED.
- C. Connect to the built-in pool and run DBCC CHECKALLOC.
- D. Connect to the built-in pool and query sys.dmv_pdw_sys_info.

Correct Answer: D

Use sys.dmv_nodes_db_partition_stats to analyze any skewness in the data.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

✉ **Leyya11111** Highly Voted 1 year, 4 months ago

Selected Answer: A

<https://github.com/rgl/azure-content/blob/master/articles/sql-data-warehouse/sql-data-warehouse-manage-distributed-data-skew.md>
upvoted 8 times

✉ **anks84** 1 year, 4 months ago

Correct, answer is A !

upvoted 4 times

✉ **Momoanwar** Most Recent 2 weeks, 1 day ago

Selected Answer: A

Chatgpt : A

Option D suggests querying `sys.dmv_pdw_sys_info`, which provides information about the SQL pool nodes and their characteristics, rather than details about data distribution or skew within a table. To investigate data skew specifically, you need to understand how the data is distributed across the distributions or partitions of a table, which is not information that `sys.dmv_pdw_sys_info` would provide.

Therefore, while `sys.dmv_pdw_sys_info` could give you insights into the overall system, it would not be the right choice for diagnosing data skew within a specific table. For that purpose, `DBCC PDW_SHOWSPACEUSED` is more appropriate, despite it not being a direct indicator of skew, it can still give you an initial indication based on space usage which might suggest further investigation if there are anomalies.

upvoted 1 times

✉ **ShrikantW** 3 weeks ago

it's dedicated Pool which basically eliminates all the other options except A

upvoted 2 times

✉ **kkk5566** 4 months, 1 week ago

Selected Answer: A

concept repeated , A is correct

upvoted 1 times

✉ **vctrhugo** 6 months, 3 weeks ago

Selected Answer: A

sys.dmv_pdw_sys_info actually provides a set of appliance-level counters that reflect overall activity on the appliance. DBCC PDW_SHOWSPACEUSED should be used instead since it displays the number of rows, disk space reserved, and disk space used for a specific table, or for all tables in a Azure Synapse Analytics or Analytics Platform System (PDW) database.

upvoted 1 times

✉ **pavankr** 7 months, 2 weeks ago

Ok, he did the typo for printing D. It should be "Connect to the built-in pool and use sys.dmv_nodes_db_partition_stats"

upvoted 2 times

✉ **janaki** 7 months, 2 weeks ago

Read Question 20, Topic 4 - Why examtopics giving 2 different answers for the same question?

For Q.20, Topic 4 - it says answer is B

here for Q.36, Topic 4 - it says answer is D

Examtopics, you first decide what you want to answer.

upvoted 4 times

✉ **Internal_Koala** 5 months, 2 weeks ago

Similar to Question 24, Topic 4 - and Q.36, Topic 4 - Answer is different for same question.

upvoted 1 times

✉ **vrodriguesp** 11 months, 3 weeks ago

Selected Answer: A

-Use DBCC PDW_SHOWSPACEUSED for seeing the skewness (each size in distributions, etc) in a table.

-By using sys.dm_pdw_request_steps table (dynamic management view, DMV) you can see how the operation is really executed and how long it took.

ref: <https://tsmatz.wordpress.com/2020/10/07/azure-synapse-analytics-sql-dedicated-pool-performance-distribution-hash/>
upvoted 4 times

✉ **brzhanyu** 1 year, 1 month ago

Selected Answer: A

need to connect Azure Synapse Analytics dedicated SQL pool1 not built-in pool (serverless pool)

upvoted 3 times

✉ **OldSchool** 1 year, 1 month ago

Selected Answer: A

A is the answer.

Read Question 20, Topic 4

upvoted 1 times

✉ **rzeng** 1 year, 2 months ago

A is the right one!

upvoted 1 times

✉ **igormmpinto** 1 year, 2 months ago

Selected Answer: A

Answer is A

A quick way to check for data skew is to use DBCC PDW_SHOWSPACEUSED. The following SQL code returns the number of table rows that are stored in each of the 60 distributions. For balanced performance, the rows in your distributed table should be spread evenly across all the distributions.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>

upvoted 3 times

✉ **walidazure** 1 year, 3 months ago

Answer A

upvoted 3 times

✉ **momani** 1 year, 3 months ago

Answer A is correct

upvoted 1 times

✉ **walidazure** 1 year, 3 months ago

Selected Answer: A

Answer A

upvoted 3 times

✉ **federic** 1 year, 4 months ago

answer A is the correct one.

upvoted 3 times

✉ **pangas2567** 1 year, 4 months ago

Selected Answer: C

I think it should be rather C.

<https://docs.microsoft.com/en-us/sql/t-sql/database-console-commands/dbcc-checkalloc-transact-sql?view=sql-server-ver16#:~:text=summary%20describes%20the-,distribution,-of%20the%20data>

The answer doesn't even correspond with the explanation.

upvoted 1 times

You use Azure Data Lake Storage Gen2.

You need to ensure that workloads can use filter predicates and column projections to filter data at the time the data is read from disk.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Reregister the Azure Storage resource provider.
- B. Create a storage policy that is scoped to a container.
- C. Reregister the Microsoft Data Lake Store resource provider.
- D. Create a storage policy that is scoped to a container prefix filter.
- E. Register the query acceleration feature.

Correct Answer: AE

Prerequisites -

To access Azure Storage, you'll need an Azure subscription. If you don't already have a subscription, create a free account before you begin.

A general-purpose v2 storage account.

Query acceleration accepts filtering predicates and column projections which enable applications to filter rows and columns at the time that data is read from disk.

Only the data that meets the conditions of a predicate are transferred over the network to the application. This reduces network latency and compute cost.

Note: Query acceleration enables applications and analytics frameworks to dramatically optimize data processing by retrieving only the data that they require to perform a given operation. This reduces the time and processing power that is required to gain critical insights into stored data.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-query-acceleration-how-to>

 **Sima_al** Highly Voted 1 year ago

- E. Register the query acceleration feature.
- D. Create a storage policy that is scoped to a container prefix filter.

To filter data at the time it is read from disk, you need to use the query acceleration feature of Azure Data Lake Storage Gen2. To enable this feature, you need to register the query acceleration feature in your Azure subscription.

In addition, you can use storage policies scoped to a container prefix filter to specify which files and directories in a container should be eligible for query acceleration. This can be used to optimize the performance of the queries by only considering a subset of the data in the container.

upvoted 17 times

 **esaade** Highly Voted 10 months ago

Selected Answer: BE

Option A, reregistering the Azure Storage resource provider, and Option C, reregistering the Microsoft Data Lake Store resource provider, are not necessary to enable filter predicates and column projections in Azure Data Lake Storage Gen2.

Option D, creating a storage policy that is scoped to a container prefix filter, is not a valid option as Azure Data Lake Storage Gen2 does not support storage policies scoped to container prefix filters.

upvoted 5 times

 **Nidie** 5 months, 1 week ago

It has, I think

upvoted 2 times

 **Momoanwar** Most Recent 2 weeks, 1 day ago

Selected Answer: DE

Correction - Chatgpt : DE

Option A, which suggests re-registering the Azure Storage resource provider, is typically not related to performance tuning or enabling specific features like query acceleration within a storage solution. Re-registering a resource provider is an administrative task that may be necessary when there are issues with the Azure subscription or the resource provider itself, which could affect the provisioning and management of Azure services.

For the scenario described, where the goal is to filter data at the time it is read from disk to optimize query performance, re-registering the Azure Storage resource provider would not directly impact the ability to use filter predicates and column projections. Instead, enabling features that allow for such optimizations, like query acceleration (E), and setting up policies for how data is stored and accessed (D), are the relevant actions to take.

upvoted 1 times

⊕  **Momoanwar** 2 weeks, 1 day ago

Selected Answer: AE

Chatgpt : AE

Option A, which suggests re-registering the Azure Storage resource provider, is typically not related to performance tuning or enabling specific features like query acceleration within a storage solution. Re-registering a resource provider is an administrative task that may be necessary when there are issues with the Azure subscription or the resource provider itself, which could affect the provisioning and management of Azure services.

For the scenario described, where the goal is to filter data at the time it is read from disk to optimize query performance, re-registering the Azure Storage resource provider would not directly impact the ability to use filter predicates and column projections. Instead, enabling features that allow for such optimizations, like query acceleration (E), and setting up policies for how data is stored and accessed (D), are the relevant actions to take.

upvoted 1 times

⊕  **EliteAllen** 1 month, 1 week ago

Selected Answer: DE

D. Create a storage policy that is scoped to a container prefix filter.
E. Register the query acceleration feature.

D&E are correct

upvoted 2 times

⊕  **pperf** 3 months ago

Its D & E

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-query-acceleration>

upvoted 1 times

⊕  **pperf** 3 months ago

Ignore this chatgpt pointing to B& E

upvoted 1 times

⊕  **pperf** 3 months ago

mod kindly remove both the replies, not relevant.

upvoted 1 times

⊕  **kkk5566** 4 months, 1 week ago

Selected Answer: DE

should be correct

upvoted 1 times

⊕  **kkk5566** 4 months ago

go to BE

upvoted 1 times

⊕  **Ast999** 10 months, 1 week ago

Selected Answer: DE

D + E = correct

upvoted 3 times

⊕  **nicky87654** 11 months, 4 weeks ago

Selected Answer: DE

E. Register the query acceleration feature.

D. Create a storage policy that is scoped to a container prefix filter.

upvoted 3 times

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

- A. Connect to Pool1 and run DBCC PDW_SHOWSPACEUSED.
- B. Connect to the built-in pool and run DBCC PDW_SHOWSPACEUSED.
- C. Connect to Pool1 and run DBCC CHECKALLOC.
- D. Connect to the built-in pool and query sys.dm_pdw_sys_info.

Correct Answer: B

 pavankr Highly Voted 7 months, 2 weeks ago

For the "Exam Topics" team:

To begin with, your questions vs answers are completely wrong., period. Check your answer for the question#36 in the same page itself!!! Why you are misleading us who are preparing seriously for the exam?? I need an immediate explanation why these questions Q#36 and Q#38 with different answers being at the same question pattern??? Seriously.

upvoted 9 times

 Jerrie86 Highly Voted 11 months, 3 weeks ago

This is repeated way too many times.

upvoted 8 times

 kkk5566 Most Recent 4 months, 1 week ago

Selected Answer: A

A is correct

upvoted 1 times

 kkk5566 4 months ago

PDW_SHOWSPACEUSED
upvoted 1 times

 OfficeSaracus 8 months, 1 week ago

Selected Answer: A

A for sure

upvoted 1 times

 duzi 11 months, 2 weeks ago

Question 36 from the same topic has the same question but as right answer D. So what is the right answer here?

upvoted 1 times

 pavankr 7 months, 2 weeks ago

Looks like he is misleading us?
upvoted 1 times

 pk07 11 months, 4 weeks ago

Selected Answer: A
(H)Agreed!
upvoted 2 times

 Mouli10 11 months, 4 weeks ago

Selected Answer: A

Its A we need to connect to Pool1
upvoted 4 times

 nicky87654 11 months, 4 weeks ago

Selected Answer: A

Connect to Pool1 and run DBCC PDW_SHOWSPACEUSED

Azure Synapse Analytics dedicated SQL pool (formerly known as Azure Synapse Analytics Parallel Data Warehouse) uses a Massively Parallel Processing (MPP) architecture and DBCC PDW_SHOWSPACEUSED is a system stored procedure that can be used to check the distribution of data across the compute nodes. By running this command on Pool1 and specifying the fact table Table1, you can identify the extent of data skew in Table1 and determine if the data is evenly distributed across the compute nodes or if it is skewed towards a specific node

upvoted 5 times

 ZIMARAKI 12 months ago

Selected Answer: A

It's A

upvoted 4 times

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

店铺: IT认证考试服务

You have an Azure Data Lake Storage Gen2 account that contains two folders named Folder1 and Folder2.

You use Azure Data Factory to copy multiple files from Folder1 to Folder2.

You receive the following error.

Operation on target Copy_sks failed: Failure happened on 'Sink' side.

ErrorCode=DelimitedTextMoreColumnsThanDefined,

'Type=Microsoft.DataTransfer.Common.Snared.HybridDeliveryException,

Message=Error found when processing 'Csv/Tsv Format Text' source

'0_2020_11_09_11_43_32.avro' with row number 53: found more columns than expected column count 27.,

Source=Microsoft.DataTransfer.Common,'

What should you do to resolve the error?

- A. Change the Copy activity setting to Binary Copy.
- B. Lower the degree of copy parallelism.
- C. Add an explicit mapping.
- D. Enable fault tolerance to skip incompatible rows.

Correct Answer: C

Yemeral Highly Voted 8 months, 1 week ago

Selected Answer: A

Correct answer is A. We are just copying files between folders. Selecting binary copy, ADF will not check schema.

With D we would discard data

With C we would change file contents

upvoted 10 times

chryckie Highly Voted 8 months, 2 weeks ago

Selected Answer: A

It's tricky.

Not D, because you don't just throw away data.

Likely not C, because it doesn't solve for future schema variability. (Avro formats are usually chosen in situations where the schema may evolve over time, because they store both the data and schema in the file itself.)

A makes most sense, since you're just trying to move files over. Binary preserves everything as-is, and you can read/interpret them as ASCII/UTF-8/whatever later.

upvoted 7 times

chryckie 8 months, 2 weeks ago

Oh! Also, the message says it's trying to process the Avro file as a Csv/Tsv Format Text. That's likely the issue.

upvoted 1 times

SATHTECH Most Recent 1 month, 2 weeks ago

C. Add an explicit mapping.

Explicit mapping involves specifying the mapping between source and destination columns explicitly. By doing this, you can ensure that each column in the source file is correctly mapped to its corresponding column in the destination file, which helps to address issues related to column count mismatches.

While other options may have their use cases, such as changing the copy activity setting to Binary Copy or enabling fault tolerance to skip incompatible rows, adding an explicit mapping (Option C) is specifically designed to handle issues where the source and destination structures do not match in terms of column count or order.

Therefore, in the context of resolving a "DelimitedTextMoreColumnsThanDefined" error, adding an explicit mapping is the most appropriate action.

upvoted 4 times

matiandal 2 months ago

Vote for C

we have a schema mismatch -)

Also

- Option A: Binary Copy is used for copying non-parseable files like images or videos, not for structured data like CSV.

upvoted 2 times

 **pperf** 3 months ago

Selected Answer: A

<https://sqlwithmanoj.com/2020/07/29/azure-data-factory-adf-pipeline-failure-found-more-columns-than-expected-column-count-delimitedtextmorecolumnsthandedefined/>

upvoted 2 times

 **EliteAllen** 3 months, 4 weeks ago

Selected Answer: A

A. Change the Copy activity setting to Binary Copy: This would bypass the error by copying the files as-is without interpreting the contents. This method might be suitable if the files are not strictly delimited text files or if you plan to handle the data inconsistency at a later stage or in a different part of the pipeline.

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: A

A is correct

upvoted 1 times

 **Tightbot** 4 months, 3 weeks ago

Selected Answer: C

I would go with Option C- Add an explicit mapping.

Laying out possible derivations from the question

1. the actual error says - column mismatch .

2. Even though the filename is "filename.avro" , it could just be a filename, the source file type is CSV/TSV.

Possible answers

1. Add an explicit mapping

2. Enabling Fault tolerance to skip incompatible rows

I think both would be a possible solution, but to me, skipping incompatible rows is more of a temporary solution and explicit mapping would be more permanent for this error. I'm also excluding future schema issues that arise after this as there is no information about it.

upvoted 1 times

 **Ram9198** 4 months, 3 weeks ago

Selected Answer: A

It says CSV/tsv source but file is avro so A is the answer

upvoted 1 times

 **andjurovicela** 6 months, 2 weeks ago

Selected Answer: D

I was pondering a bit about this one, and decided to go with D. Reasoning behind this is because the question was "how to resolve this error?" and 100% preservation of source data hasn't been a condition, hence D is the most straightforward.

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: C

Binary Copy is a setting that can be used in Azure Data Factory to improve performance when copying binary data, such as Avro or Parquet files. It optimizes the data transfer by copying the data as-is without parsing or transforming it. However, in this case, the error is related to the mismatch in the column structure, which cannot be resolved by changing the copy setting to Binary Copy.

upvoted 1 times

 **azure_user11** 7 months, 3 weeks ago

Selected Answer: A

I think the purpose here is to just copy files as-is from one folder to another. <https://learn.microsoft.com/en-us/azure/data-factory/format-binary>

upvoted 1 times

 **levto** 8 months ago

Selected Answer: A

agree with Yemeral

upvoted 1 times

 **sk20** 8 months, 3 weeks ago

Correct Answer D . It makes sense to use Fault Tolerance . Refer link below.

<https://learn.microsoft.com/en-us/answers/questions/1178682/found-more-columns-than-expected-column-count-35>

upvoted 2 times

 **shakes103** 9 months ago

Selected Answer: C

Correct answer is C

upvoted 2 times

 **AscentAcademy** 10 months, 1 week ago

It appears we're trying to copy an avro file. This should be done as a binary copy, so we should select A. In fact, you I found someone who had this exact issue here: <https://sqlwithmanoj.com/2020/07/29/azure-data-factory-adf-pipeline-failure-found-more-columns-than-expected-column-count-delimitedtextmorecolumns than defined/>

upvoted 6 times

 **shoottheduck** 10 months, 2 weeks ago

Selected Answer: D

I have checked this in ADF. Also see doc:

<https://learn.microsoft.com/nl-nl/azure/data-factory/copy-activity-fault-tolerance#copying-tabular-data>

upvoted 4 times

Question #40

Topic 4

A company plans to use Apache Spark analytics to analyze intrusion detection data.

You need to recommend a solution to analyze network and system activity data for malicious activities and policy violations. The solution must minimize administrative efforts.

What should you recommend?

- A. Azure HDInsight
- B. Azure Data Factory
- C. Azure Data Lake Storage
- D. Azure Databricks

Correct Answer: D

 **Mouli10** Highly Voted  11 months, 4 weeks ago

Selected Answer: D

Azure databricks

upvoted 6 times

 **kkk5566** Most Recent  4 months, 1 week ago

Selected Answer: D

correct

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: D

By leveraging Azure Databricks, you can easily perform advanced analytics on the intrusion detection data using Spark's powerful distributed processing capabilities. Databricks provides an interactive and collaborative environment where you can write Spark code, explore and visualize data, and build machine learning models. It also integrates with popular data sources, including Azure Data Lake Storage, for efficient data ingestion and processing.

upvoted 2 times

 **Stefan94** 12 months ago

Correct

upvoted 3 times

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool.

You need to monitor the database for long-running queries and identify which queries are waiting on resources.

Which dynamic management view should you use for each requirement? To answer, select the appropriate options in the answer area.

NOTE: Each correct answer is worth one point.

Answer Area

Monitor the database for long-running queries:

| | |
|---|---|
| | ▼ |
| sys.dm_pdw_exec_requests
sys.dm_pdw_sql_requests
sys.dm_pdw_exec_sessions | ▼ |

Identify which queries are waiting on resources:

| | |
|--|---|
| | ▼ |
| sys.dm_pdw_waits
sys.dm_pdw_lock_waits
sys.resource_governor_workload_groups | ▼ |

Answer Area

Monitor the database for long-running queries:

| | |
|---|---|
| | ▼ |
| sys.dm_pdw_exec_requests
sys.dm_pdw_sql_requests
sys.dm_pdw_exec_sessions | ▼ |

Correct Answer:

Identify which queries are waiting on resources:

| | |
|--|---|
| | ▼ |
| sys.dm_pdw_lock_waits
sys.dm_pdw_waits
sys.resource_governor_workload_groups | ▼ |

 **auwia** Highly Voted 6 months, 2 weeks ago

The sys.dm_pdw_lock_waits view is specific to SQL Server and is used to monitor lock waits and lock resources in regular SQL Server environments, not in Azure Synapse Analytics dedicated SQL pools.

My answers are:

1. sys.dm_pdw_exec_requests
2. sys.dm_pdw_waits

There is a similar question in the Microsoft official practice assessment and the explanation is the following:

The sys.dm_pdw_waits view holds information about all wait stats encountered during the execution of a request or query, including locks and waits on a transmission queue

upvoted 15 times

 **bp_a_user** Highly Voted 8 months, 2 weeks ago

Its dm_pdw_waits:

Queries in the Suspended state can be queued due to a large number of active running queries. These queries also appear in the sys.dm_pdw_waits view with a type of UserConcurrencyResourceType from the official learning path: <https://learn.microsoft.com/en-us/training/modules/manage-monitor-data-warehouse-activities-azure-synapse-analytics/6-use-dynamic-management-views-to-identify-troubleshoot-query-performance>

upvoted 8 times

 **matiandal** Most Recent 2 months ago

provided answers are correct

upvoted 1 times

□ **kkk5566** 4 months, 1 week ago

1. sys.dm_pdw_exec_requests
2. sys.dm_pdw_waits

upvoted 6 times

□ **vctrhugo** 6 months, 3 weeks ago

"Queries in the Suspended state can be queued due to a large number of active running queries. These queries also appear in the sys.dm_pdw_waits waits query with a type of UserConcurrencyResourceType."

upvoted 2 times

□ **AHUI** 9 months, 1 week ago

correct

<https://learn.microsoft.com/en-us/azure/azure-sql/database/monitoring-with-dmvs?view=azuresql>

upvoted 3 times

□ **AHUI** 9 months, 1 week ago

box 1: is correct

box 2: sys.dm_pdw_waits

<https://learn.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sys-dm-pdw-waits-transact-sql?view=aps-pdw-2016-au7>

upvoted 21 times

You have an Azure Data Factory pipeline named pipeline1 that includes a Copy activity named Copy1. Copy1 has the following configurations:

- The source of Copy1 is a table in an on-premises Microsoft SQL Server instance that is accessed by using a linked service connected via a self-hosted integration runtime.
- The sink of Copy1 uses a table in an Azure SQL database that is accessed by using a linked service connected via an Azure integration runtime.

You need to maximize the amount of compute resources available to Copy1. The solution must minimize administrative effort.

What should you do?

- A. Scale out the self-hosted integration runtime.
- B. Scale up the data flow runtime of the Azure integration runtime and scale out the self-hosted integration runtime.
- C. Scale up the data flow runtime of the Azure integration runtime.

Correct Answer: C

✉  **azure_user11** Highly Voted 7 months, 3 weeks ago

Why not B?

<https://learn.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime>

Azure integration runtime provides the native compute to move data between cloud data stores in a secure, reliable, and high-performance manner. You can set how many data integration units to use on the copy activity, and the compute size of the Azure IR is elastically scaled up accordingly without requiring you to explicitly adjust the size of the Azure Integration Runtime.

For high availability and scalability, you can scale out the self-hosted IR by associating the logical instance with multiple on-premises machines in active-active mode.

upvoted 12 times

✉  **BillMyI** Highly Voted 7 months, 2 weeks ago

I would answer A.

<https://learn.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime>

Copying between a cloud data source and a data source in a private network: if either the source or sink linked service points to a self-hosted IR, the copy activity is executed on the self-hosted IR.

upvoted 9 times

✉  **jongert** Most Recent 21 hours, 36 minutes ago

Integration runtime is hosted on the location of the sink for copy activity if I am not mistaken.

upvoted 1 times

✉  **dakku987** 1 week ago

Selected Answer: C

chat gpt

C. Scale up the data flow runtime of the Azure integration runtime.

Explanation:

In Azure Data Factory, when you're copying data between different data stores, the compute resources used by the Copy activity are mainly determined by the data flow involved in the copying process. Azure Data Factory provides two types of integration runtimes:

upvoted 1 times

✉  **Momoanwar** 2 weeks, 1 day ago

Selected Answer: A

Chatgpt

The self-hosted integration runtime can be scaled out by adding additional nodes, which allows it to process more activities simultaneously. This is a way to increase compute resources without a significant administrative overhead since it involves configuration changes rather than physical infrastructure changes.

Options B and C involve scaling up the data flow runtime, which is not applicable in this context since the Copy activity does not use data flow runtime; it uses the integration runtime for data movement. Therefore, the correct answer to maximize compute resources for Copy1 with minimal administrative effort is:

A. Scale out the self-hosted integration runtime.

upvoted 1 times

 **SATHTECH** 1 month, 2 weeks ago

For maximizing the amount of compute resources available to the Copy activity in Azure Data Factory, you should consider scaling up the data flow runtime of the Azure integration runtime.

Option C. Scale up the data flow runtime of the Azure integration runtime.

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: A

<https://learn.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime#self-hosted-ir-compute-resource-and-scaling>
A should be correct.

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

correct it, C , Option A, "Scale out the self-hosted integration runtime," is not the best solution to maximize the amount of compute resources available to Copy1 because it would not minimize administrative effort. Scaling out the self-hosted integration runtime would involve adding more nodes to the runtime pool, which would require allocating new virtual machines and registering new nodes on the integration runtime. This process can be time-consuming and would require additional administrative effort1.

upvoted 3 times

 **Ram9198** 4 months, 3 weeks ago

Selected Answer: A

if either the source or sink linked service points to a self-hosted IR, the copy activity is executed on the self-hosted IR.

upvoted 2 times

 **alegiordx** 6 months ago

Selected Answer: A

My answer is A due to the precedence criteria among Integration runtimes selection when source and sink linked services are linked to different IRs, as described here

<https://learn.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime#determining-which-ir-to-use>

upvoted 2 times

 **tsmk** 6 months ago

Selected Answer: C

The important point is - "The solution must minimize administrative effort."

Azure integration runtime - Cloud-based service - Not as scalable as a SHIR

Self-hosted integration runtime - More scalable - Requires more administrative effort

Scaling up the Azure IR will give us more compute resources without increasing the administrative effort.

upvoted 1 times

 **andjurovicela** 6 months, 1 week ago

Selected Answer: A

According to the MS document, A seems to be the correct answer.

upvoted 2 times

 **Azure_2023** 7 months ago

Selected Answer: B

I would answer B

upvoted 1 times

 **JG1984** 6 months, 3 weeks ago

scaling out the self hosted will not increase the amount of compute resources becoz its already running on the physical machine (on-premises) . However, the Azure integration runtime is a managed service, so scaling up its data flow runtime will increase the amount of compute resources available to Copy1.

upvoted 1 times

You are designing a solution that will use tables in Delta Lake on Azure Databricks.

You need to minimize how long it takes to perform the following:

- Queries against non-partitioned tables
- Joins on non-partitioned columns

Which two options should you include in the solution? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. the clone command
- B. Z-Ordering
- C. Apache Spark caching
- D. dynamic file pruning (DFP)

Correct Answer: BD

 **OfficeSaracus** Highly Voted 8 months ago

Selected Answer: BD

Seems correct:

<https://learn.microsoft.com/en-us/azure/databricks/optimizations/dynamic-file-pruning>

<https://learn.microsoft.com/en-us/azure/databricks/delta/data-skipping>

upvoted 9 times

 **vctrhugo** Highly Voted 6 months, 2 weeks ago

Selected Answer: BD

Dynamic file pruning, can significantly improve the performance of many queries on Delta Lake tables. Dynamic file pruning is especially efficient for non-partitioned tables, or for joins on non-partitioned columns. The performance impact of dynamic file pruning is often correlated to the clustering of data so consider using Z-Ordering to maximize the benefit.

upvoted 6 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: BD

correct

upvoted 2 times

You have an Azure Data Lake Storage Gen2 account named account1 that contains a container named container1.

You plan to create lifecycle management policy rules for container1.

You need to ensure that you can create rules that will move blobs between access tiers based on when each blob was accessed last.

What should you do first?

- A. Configure object replication
- B. Create an Azure application
- C. Enable access time tracking
- D. Enable the hierarchical namespace

Correct Answer: C

cloud_lady Highly Voted 8 months ago

Selected Answer: C

Answer is correct.

Customers stores huge amount of data in Azure blob storage. Sometimes this data is accessed frequently and other times infrequently. Last access time tracking integrates with the lifecycle of Azure blob storage to allow automatic tiering and deletion of data based on when individual blobs are accessed last.

upvoted 8 times

kkk5566 Most Recent 4 months, 1 week ago

Selected Answer: C

correct

upvoted 1 times

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. Data IO percentage

C. Cache hit percentage

D. CPU percentage

Correct Answer: C

 **kkk5566** 4 months, 1 week ago

Selected Answer: C

Repeated

upvoted 1 times

 **darshilparmar** 7 months ago

Repeat Questons

upvoted 4 times

 **henryphchan** 8 months ago

Selected Answer: C

Answer is C, and it's a repeated question

upvoted 4 times

HOTSPOT

You have an Azure data factory named DF1 that contains 10 pipelines.

The pipelines are executed hourly by using a schedule trigger. All activities are executed on an Azure integration runtime.

You need to ensure that you can identify trends in queue times across the pipeline executions and activities. The solution must minimize administrative effort.

How should you configure the Diagnostic settings for DF1? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Collect:

| | |
|----------------------------|---|
| Pipeline activity runs log | ▼ |
| Pipeline runs log | ▼ |
| Trigger runs log | ▼ |

Send to:

| | |
|-------------------------|---|
| Event hub | ▼ |
| Log Analytics workspace | ▼ |
| Storage account | ▼ |

Answer Area

Collect:

| | |
|----------------------------|---|
| Pipeline activity runs log | ▼ |
| Pipeline runs log | ▼ |
| Trigger runs log | ▼ |

Correct Answer:

Send to:

| | |
|-------------------------|---|
| Event hub | ▼ |
| Log Analytics workspace | ▼ |
| Storage account | ▼ |

vk880 Highly Voted 8 months ago

1. To identify trends in queue times, you should focus on the Pipeline activity run logs rather than the Pipeline run logs. Pipeline activity run logs allows you to track the queue times for individual activities within the pipeline. While Pipeline run logs logs may provide some information about queue times, they do not provide granular details for each activity within the pipeline.

upvoted 6 times

henryphchan 7 months, 3 weeks ago

so the provided answer is correct.

upvoted 5 times

Momoanwar Most Recent 2 weeks, 1 day ago

Wrong activity runs dont minimize efforts, chatgpt :

To minimize administrative effort while still being able to identify trends in queue times across pipeline executions and activities, you should collect:

- Pipeline runs log: This log provides a high-level overview of each pipeline execution, which is sufficient for identifying trends in queue times

without the need for the more granular detail that would come from collecting activity runs logs.

And send to:

- Log Analytics workspace: This will allow for centralized logging and analytics, which is effective for trend analysis with minimal administrative effort.

So the settings should be:

Collect: Pipeline runs log
Send to: Log Analytics workspace
upvoted 1 times

- ✉ **shreembreeze** 2 months, 3 weeks ago
Did any of you completed DP-203 exam here
upvoted 1 times
- ✉ **dakku987** 1 week ago
what about you i am preparing for now
upvoted 1 times
- ✉ **kkk5566** 4 months, 1 week ago
correct
upvoted 2 times

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 percentage
- B. Cache hit percentage
- C. Data Warehouse Units (DWU) used
- D. Data IO percentage

Correct Answer: B

✉  **darshilparmarr** Highly Voted 7 months ago

Repeated 4 times
upvoted 11 times

✉  **matiandal** Most Recent 2 months ago

Correct Answer: C

Why not A ?

While the DWU percentage can provide insights into whether your workload is more CPU or IO intensive, the DWU used can provide a more direct measure of the overall resource utilization¹. This can be more helpful in identifying if resource contention (i.e., your workload is demanding more resources than are available) is causing the slow performance of commonly used queries

R: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-resource-utilization-query-activity>
upvoted 1 times

✉  **EliteAllen** 3 months, 4 weeks ago

Selected Answer: B

Monitoring DWU used (Option C) can certainly be part of a comprehensive approach to diagnosing the performance issues, focusing on the cache hit percentage (Option B) might offer a more targeted way to address the specific problem described in the scenario.
upvoted 1 times

✉  **kkk5566** 4 months ago

Selected Answer: B

Repeated
upvoted 1 times

✉  **kkk5566** 4 months, 1 week ago

Selected Answer: C

Repeated
upvoted 1 times

✉  **AzureSQLDatabase** 3 months, 3 weeks ago

says the guy with the wrong answer xD

I know you know the answer. I know you chose C by mistake which you don't know.

AAAHHHH how do I know so much yet so little!!!

upvoted 2 times

HOTSPOT

You have an Azure subscription that contains the resources shown in the following table.

| Name | Type | Description |
|-------|--|---------------------|
| ws1 | Azure Synapse Analytics workspace | None |
| kv1 | Azure Key Vault | None |
| UAMI1 | User-assigned managed identity | Associated with ws1 |
| sp1 | Apache Spark pool in Azure Synapse Analytics | Associated with ws1 |

You need to ensure that you can run Spark notebooks in ws1. The solution must ensure that you can retrieve secrets from kv1 by using UAMI1.

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

NOTE: Each correct selection is worth one point.

Answer Area

In the Azure portal:

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.
- Create a linked service to kv1.

In Synapse Studio:

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.
- Create a linked service to kv1.

Answer Area

In the Azure portal:

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.
- Create a linked service to kv1.**

Correct Answer:

In Synapse Studio:

- Add a role-based access control (RBAC) role to kv1.**
- Add a role-based access control (RBAC) role to ws1.**
- Create a linked service to kv1.

 **hassexat** Highly Voted 4 months ago

In the Azure Portal --> Add a role-based access control (RBAC) role to kv1

In Synapse Studio --> Create a linked service to kv1

upvoted 16 times

 **Momoanwar** Most Recent 2 weeks, 1 day ago

Chatgpt:

To enable Spark notebooks in Azure Synapse Analytics workspace `ws1` to retrieve secrets from Azure Key Vault `kv1` using the user-assigned managed identity `UAMI1`, you need to set the appropriate permissions and configurations:

In the Azure portal:

- Add a role-based access control (RBAC) role to `kv1`. Assign the user-assigned managed identity `UAMI1` the necessary role, like "Key Vault Secrets User", to retrieve secrets from the Key Vault.

In Synapse Studio:

- Create a linked service to `kv1`. This linked service should use the user-assigned managed identity `UAMI1` for authentication, allowing the Spark notebooks to use this linked service to access Key Vault secrets.

So the selections should be:

In the Azure portal: Add a role-based access control (RBAC) role to 'kv1'.

In Synapse Studio: Create a linked service to 'kv1'.

upvoted 2 times

 **y154707** 2 months ago

I'm reporting this comment on purpose so the ExamTopics team can review the answer, as requested below.

ExamTopics team, please, do one of the following:

- 1) clearly explain and provide the support link or material for the answer you chose, or
- 2) correct the answers selected.

As stated in previous answers in this discussion, the logic says that first you have to add the role in RBAC to kv1 and then associate the kv1 source as an linked service in synapse.

upvoted 1 times

 **akshy** 4 months ago

The boxes should be reversed for the answers as it does not make sense currently

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Box1. Add a role-based access control (RBAC) role to kv

Box2. Create a linked service to kv1

upvoted 2 times

 **_Lukas_** 5 months, 1 week ago

In the Azure portal:

Add a role-based access control (RBAC) role to kv1 - You need to assign the 'Key Vault Secrets User' role to UAMI1 on kv1. This will grant the managed identity the necessary permissions to retrieve secrets from Key Vault.

In Synapse Studio:

Create a linked service to kv1 - You need to create a linked service in Azure Synapse Studio to connect to kv1. The linked service will use the User-Assigned Managed Identity (UAMI1) to authenticate to the Azure Key Vault.

upvoted 4 times

HOTSPOT

You have an Azure Data Factory pipeline shown in the following exhibit.



The execution log for the first pipeline run is shown in the following exhibit.

Activity runs

Pipeline run ID 87f89922-14fa-468f-b13f-2fb67606f4ff

All status ▾

Showing 1 - 2 items

| Activity name | Activity type | Run start | Duration | Status |
|----------------|------------------|--------------------------|----------|-------------|
| Web_GetIP | Web | Nov 10, 2022, 11:11:36 a | 00:00:02 | ✖ Failed |
| Exec_COPY_BLOB | Execute Pipeline | Nov 10, 2022, 11:11:25 a | 00:00:11 | ✔ Succeeded |

The execution log for the second pipeline run is shown in the following exhibit.

Activity runs

Pipeline run ID a7b5b522-cfaf-4c09-b3a9-f842986be984

All status ▾

Showing 1 - 3 items

| Activity name | Activity type | Run start | Duration | Status |
|----------------|------------------|--------------------------|----------|-------------|
| Set status | Set variable | Nov 10, 2022, 11:13:17 a | 00:00:01 | ✔ Succeeded |
| Web_GetIP | Web | Nov 10, 2022, 11:12:59 a | 00:00:16 | ✔ Succeeded |
| Exec_COPY_BLOB | Execute Pipeline | Nov 10, 2022, 11:12:48 a | 00:00:11 | ⌚ Skipped |

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**

The Retry property of the Web_GetIP activity is set to 1.

Yes No

The waitOnCompletion property of the Exec_COPY_BLOB activity is set to true.

Yes No

The Exec_COPY_BLOB activity was skipped during the second run due to pipeline dependencies.

Yes No

Answer Area**Statements**

The Retry property of the Web_GetIP activity is set to 1.

| | |
|--------------------------------------|----------------------------------|
| <input checked="" type="radio"/> Yes | <input type="radio"/> No |
| <input type="radio"/> | <input checked="" type="radio"/> |
| <input checked="" type="radio"/> | <input type="radio"/> |

Correct Answer:

The waitOnCompletion property of the Exec_COPY_BLOB activity is set to true.

The Exec_COPY_BLOB activity was skipped during the second run due to pipeline dependencies.

⊕ **Ram9198** Highly Voted 4 months ago

No yes no

upvoted 6 times

⊕ **Matt2000** Highly Voted 4 months, 4 weeks ago

No, No, No

The Retry Property is not set to one for Web_GetIP: Otherwise, we would see a retry of that activity in the first run.

waitOnCompletion property is not set to true: In the second run, Exec_COPY_BLOB takes as long as in the first one, despite being skipped. So, it could not have been waiting for the pipeline that it had triggered to complete.

Exec_COPY_BLOB cannot be skipped due to a pipeline dependency since it is the first activity in the pipeline. Most likely, its activity state was manually set to 'skipped'.

upvoted 5 times

⊕ **dakku987** 1 week ago

i think waitOnCompletion is yes bcz only after 11 sec next activity get started if it was not set to true all first and second activity both will be started at same time

upvoted 1 times

⊕ **y154707** Most Recent 2 months ago

ExamTopics team, please, do one of the following:

- 1) clearly explain and provide the support link or material for the answer you chose, or
- 2) correct the answers selected.

As stated in answers below in this discussion, the response are N-Y-N or N-N-N, but there's nothing that can point us to the answer given in the resolution.

upvoted 3 times

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

- A. Connect to the built-in pool and query sys.dmv_nodes_db_partition_stats.
- B. Connect to Pool1 and run DBCC PDW_SHOWSPACEUSED.
- C. Connect to Pool1 and query sys.dmv_node_status.
- D. Connect to the built-in pool and query sys.dmv_sys_info.

Correct Answer: A

Trulysme 1 month, 1 week ago

Answer B
Topic 4 #36 and #38
upvoted 1 times

kkk5566 4 months, 1 week ago

Selected Answer: B
REPETED,B is correct
upvoted 1 times

patjoo 4 months, 3 weeks ago

Selected Answer: B
It is indeed B:
<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute#determine-if-the-table-has-data-skew>
upvoted 1 times

Nidie 5 months, 1 week ago

Selected Answer: B
I think it is B , same question in the previous pages
upvoted 4 times

You have several Azure Data Factory pipelines that contain a mix of the following types of activities:

- Power Query
- Notebook
- Copy
- Jar

Which two Azure services should you use to debug the activities? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

- A. Azure Machine Learning
- B. Azure Data Factory
- C. Azure Synapse Analytics
- D. Azure HDInsight
- E. Azure Databricks

Correct Answer: BE

 **Sdevi49** Highly Voted 5 months ago

The Copy activity is native to Azure Data Factory, so you would use Azure Data Factory to debug it.

The Notebook and Jar activities are related to Databricks jobs, so you would use Azure Databricks to debug them.

Power Query is more associated with data wrangling and transformation, and while it can be used in various services, in the context of Azure Data Factory, you'd likely be debugging within Azure Data Factory or Azure Synapse Analytics. However, given the other activities listed, Azure Data Factory is the more probable choice for this scenario.

Therefore, the correct answers are:

- B. Azure Data Factory
 - E. Azure Databricks
- upvoted 6 times

 **Momoanwar** Most Recent 2 weeks, 1 day ago

Selected Answer: BE

Chatgpt BE :

To debug the various types of activities in Azure Data Factory pipelines, you would use:

- B. Azure Data Factory - This is the primary environment where you orchestrate and monitor the pipelines and activities including Power Query, Copy, and others.
- E. Azure Databricks - For debugging activities related to Notebooks and Jar files, Azure Databricks provides an environment to run and debug these types of activities.

These services offer the necessary tools and environments to debug the specified activities within Azure Data Factory pipelines.
upvoted 1 times

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices.

The company must be able to monitor the devices in real-time.

You need to design the solution.

What should you recommend?

- A. Azure Analysis Services using Microsoft Visual Studio
- B. Azure Data Factory instance using Azure PowerShell
- C. Azure Analysis Services using Azure PowerShell
- D. Azure Stream Analytics cloud job using Azure Portal

Correct Answer: D

 **kkk5566** 4 months, 1 week ago

Selected Answer: D

repeated

upvoted 3 times

You have an Azure Synapse Analytics dedicated SQL pool named pool1.

You need to perform a monthly audit of SQL statements that affect sensitive data. The solution must minimize administrative effort.

What should you include in the solution?

- A. workload management
- B. sensitivity labels
- C. dynamic data masking
- D. Microsoft Defender for SQL

Correct Answer: B

 **Momoanwar** 2 weeks, 1 day ago

Selected Answer: B

Correct, chatgpt

To conduct a monthly audit of SQL statements impacting sensitive data in Azure Synapse Analytics while minimizing administrative effort, include the use of Sensitivity Labels to classify sensitive data and utilize Microsoft Defender for SQL for advanced threat detection and monitoring of related activities. These tools will aid in governance and provide alerts for any suspicious access or manipulation of sensitive data. For a comprehensive approach to setting up these features, refer to the official Microsoft documentation for guidance.<https://learn.microsoft.com/en-us/azure/synapse-analytics/>

upvoted 1 times

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub to communicate with the IoT devices.

The company must be able to monitor the devices in real-time.

You need to design the solution.

What should you recommend?

- A. Azure Analysis Services using Azure Portal
- B. Azure Stream Analytics Edge application using Microsoft Visual Studio
- C. Azure Analysis Services using Azure PowerShell
- D. Azure Analysis Services using Microsoft Visual Studio

Correct Answer: B

 **Momoanwar** 2 weeks, 1 day ago

Selected Answer: B

Like 52...

upvoted 1 times

HOTSPOT

You have an Azure data factory.

You execute a pipeline that contains an activity named Activity1. Activity1 produces the following output.

```
{
  ...
  "dataRead": 1208,
  "dataWritten": 1208,
  "filesRead": 1,
  "filesWritten": 1,
  "sourcePeakConnections": 3,
  "sinkPeakConnections": 2,
  "copyDuration": 13,
  "throughput": 0.147,
  "effectiveIntegrationRuntime": "AutoResolveIntegrationRuntime (West Central US)",
  "usedDataIntegrationUnits": 4,
  "reportLineageToPurview": {
    "status": "Succeeded",
    "durationInSecond": "4"
  }
}
...
```

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 |
|--|-----------------------|-----------------------|----|
| Activity1 is a Copy activity. | <input type="radio"/> | <input type="radio"/> | |
| Activity1 is executed by using a self-hosted integration runtime. | <input type="radio"/> | <input type="radio"/> | |
| The data factory that executed the pipeline is connected to Microsoft Purview. | <input type="radio"/> | <input type="radio"/> | |

| Answer Area | Statements | Yes | No |
|--|-------------------------------------|-------------------------------------|----|
| Activity1 is a Copy activity. | <input checked="" type="checkbox"/> | <input type="radio"/> | |
| Activity1 is executed by using a self-hosted integration runtime. | <input type="radio"/> | <input checked="" type="checkbox"/> | |
| The data factory that executed the pipeline is connected to Microsoft Purview. | <input checked="" type="checkbox"/> | <input type="radio"/> | |

y154707 Highly Voted 2 months ago

At last an answer from Examtopics team that makes sense with the problem stated and is not based on an alternate wacky multiverse where all the rules and good practices of Azure are broken.

upvoted 5 times

MielniByczq Most Recent 5 days, 15 hours ago

Alternate wacky multiverse. Lesson Learned from this comment.

upvoted 1 times

Momoanwar 2 weeks, 1 day ago

Correct, chatgpt:

1. Activity1 appears to be a Copy activity given the "dataRead," "dataWritten," and "copyDuration" fields.
 2. Activity1 does not use a self-hosted integration runtime; it uses an Azure integration runtime as indicated by "AutoResolveIntegrationRuntime."
 3. The data factory is connected to Microsoft Purview, as evidenced by the "reportLineageToPurview" section indicating a successful status.
- Hence, the correct answers are:
- Yes, Activity1 is a Copy activity.
 - No, Activity1 is not executed using a self-hosted integration runtime.
 - Yes, the data factory that executed the pipeline is connected to Microsoft Purview.
- upvoted 2 times

Question #56

Topic 4

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 percentage
- B. Cache hit percentage
- C. DWU limit
- D. Data Warehouse Units (DWU) used

Correct Answer: B

 **kam1122** 3 weeks, 5 days ago

this question repeat like 6 or 7 times.....

upvoted 1 times

 **kam1122** 1 month, 2 weeks ago

correct answer B, (repeat to many times)

upvoted 2 times

 **matiandal** 2 months ago

correct answer A

upvoted 1 times

You have an Azure subscription that contains an Azure Synapse Analytics workspace and a user named User1.

You need to ensure that User1 can review the Azure Synapse Analytics database templates from the gallery. The solution must follow the principle of least privilege.

Which role should you assign to User1?

- A. Storage Blob Data Contributor.
- B. Synapse Administrator
- C. Synapse Contributor
- D. Synapse User

Correct Answer: C

RK710 1 day, 20 hours ago

Selected Answer: D

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.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/database-designer/create-lake-database-from-lake-database-templates>

upvoted 2 times

ExamDestroyer69 5 days, 10 hours ago

Selected Answer: D

I believe the answer is D. Synapse User as it has less priviledges than contributor.

I am unaware what roles can and cant view templates, some one will have to test this. But as you can just google the templates, i assume its the least priveledged role

See link showing priveledges:

<https://learn.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-synapse-rbac-roles#:~:text=Can%20read%20and%20write%20artifacts%0ACan%20view%20saved%20notebook%20and%20pipeline%20output%0ACan%20do%20all%20actions%20on%20Spark%20activities%0ACan%20view%20Spark%20pool%20logs>

upvoted 2 times

You have a Log Analytics workspace named la1 and an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 sends logs to la1.

You need to identify whether a recently executed query on Pool1 used the result set cache.

What are two ways to achieve the goal? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

- A. Review the sys.dm_pdw_sql_requests dynamic management view in Pool1.
- B. Review the sys.dm_pdw_exec_requests dynamic management view in Pool1.
- C. Use the Monitor hub in Synapse Studio.
- D. Review the AzureDiagnostics table in la1.
- E. Review the sys.dm_pdw_request_steps dynamic management view in Pool1.

Correct Answer: BC

 **jonpert** 4 days, 16 hours ago

Selected Answer: BC

Correct

<https://learn.microsoft.com/en-us/azure/synapse-analytics/monitoring/how-to-monitor-using-azure-monitor>

<https://learn.microsoft.com/en-us/sql/t-sql/statements/set-result-set-caching-transact-sql?view=azuresqldw-latest>

<https://learn.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sys-dm-pdw-exec-requests-transact-sql?view=aps-pdw-2016-au7>

upvoted 1 times

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Sales.Orders. Sales.Orders contains a column named SalesRep.

You plan to implement row-level security (RLS) for Sales.Orders.

You need to create the security policy that will be used to implement RLS. The solution must ensure that sales representatives only see rows for which the value of the SalesRep column matches their username.

How should you complete the code? To answer, select the appropriate options in the answer area.

Answer Area

```
CREATE SCHEMA Security;
GO
CREATE FUNCTION Security.tvf_securitypredicate
(@SalesRep AS nvarchar(50))
RETURNS TABLE
WITH
    ENCRYPTION
    RETURNS NULL ON NULL INPUT
    SCHEMABINDING
AS
    RETURN SELECT 1 AS tvf_securitypredicate_result
WHERE @SalesRep = USER_NAME();
GO
CREATE SECURITY POLICY SalesFilter
    ADD BLOCK PREDICATE Security.tvf_securitypredicate(SalesRep)
    ADD BLOCK PREDICATE tvf_securitypredicate_result
    ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)
ON Sales.Orders
WITH (STATE = ON);
```

Answer Area

```
CREATE SCHEMA Security;
GO
CREATE FUNCTION Security.tvf_securitypredicate
(@SalesRep AS nvarchar(50))
RETURNS TABLE
WITH
    ENCRYPTION
    RETURNS NULL ON NULL INPUT
    SCHEMABINDING
```

Correct Answer: **AS**

```
    RETURN SELECT 1 AS tvf_securitypredicate_result
    WHERE @SalesRep = USER_NAME();
    GO
    CREATE SECURITY POLICY SalesFilter
        ADD BLOCK PREDICATE Security.tvf_securitypredicate(SalesRep)
        ADD BLOCK PREDICATE tvf_securitypredicate_result
        ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)
    ON Sales.Orders
    WITH (STATE = ON);
```

 **jongert** 4 days, 16 hours ago

Correct.

upvoted 2 times

店铺: IT认证考试服务

店铺: IT认证考试服务

Question #60

Topic 4

You have an Azure data factory named DF1. DF1 contains a single pipeline that is executed by using a schedule trigger.

From Diagnostics settings, you configure pipeline runs to be sent to a resource-specific destination table in a Log Analytics workspace.

You need to run KQL queries against the table.

Which table should you query?

- A. ADFPipelineRun
- B. ADFTtriggerRun
- C. ADFActivityRun
- D. AzureDiagnostics

Correct Answer: B

 **Oleh777** 1 day, 13 hours ago

IMHO activity run should show all the details
upvoted 1 times

店铺: IT认证考试服务

店铺: IT认证考试服务

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool named sqlpool1 that contains a table named Sales1.

Each row in the Sales table contains regional sales data and a field that lists the username of a sales analyst.

You need to configure row-level security (RLS) to ensure that the analysts can view only the rows containing their respective data.

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

NOTE: Each correct selection is worth one point.

Answer Area

To configure RLS, create:

- A materialized view in sqlpool1
- A security policy in the Sales table
- Database scoped credentials in sqlpool1

To designate which rows each analyst can access, use:

- A masking rule
- A table-valued function
- The CONTAINS predicate

Answer Area

To configure RLS, create:

- A materialized view in sqlpool1
- A security policy in the Sales table**
- Database scoped credentials in sqlpool1

To designate which rows each analyst can access, use:

- A masking rule**
- A table-valued function**
- The CONTAINS predicate

 **jonpert** 4 days, 16 hours ago

Correct

<https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16>

upvoted 1 times

You have an Azure subscription that contains an Azure Synapse workspace named WS1 and an Azure Monitor action group named Group1. WS1 has a dedicated SQL pool.

You plan to archive monitoring data for integration activity runs.

You need to ensure that you can configure custom alerts based on the archived data that will execute Group1. The solution must minimize administrative effort.

Which diagnostic setting should you select?

- A. Send to Log Analytics workspace
- B. Archive to a storage account
- C. Stream to an event hub
- D. Send to a partner solution

Correct Answer: A

 **jonpert** 4 days, 16 hours ago

Correct

upvoted 1 times

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1.

You have the queries shown in the following table.

| Name | Users | Result set size |
|--------|-----------------------------------|-----------------|
| Query1 | Deterministic runtime expressions | 25 MB |
| Query2 | Deterministic built-in functions | 1 GB |
| Query3 | User-defined functions (UDFs) | 50 MB |
| Query4 | Row-level security (RLS) | 15 GB |

You are evaluating whether to enable result set caching for Pool1.

Which query results will be cached if result set caching is enabled?

- A. Query1 only
- B. Query2 only
- C. Query1 and Query2 only
- D. Query1 and Query3 only
- E. Query1, Query2, and Query3 only

Correct Answer: C

 **Tapaskaro** 4 days, 1 hour ago

correct

What's not cached

Once result set caching is turned ON for a database, results are cached for all queries until the cache is full, except for these queries:

Queries with built-in functions or runtime expressions that are non-deterministic even when there's no change in base tables' data or query. For example, DateTime.Now(), GetDate().

Queries using user defined functions

Queries using tables with row level security

Queries returning data with row size larger than 64KB

Queries returning large data in size (>10GB)

upvoted 1 times

 **jongert** 4 days, 16 hours ago

Correct.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-caching#whats-not-cached>

upvoted 1 times

You have an Azure subscription that contains an Azure Synapse Analytics workspace name workspace1, workspace1 contains an Azure Synapse Analytics dedicated SQL pool named Pool1.

You create a mapping data flow in an Azure Synapse pipeline that writes data to Pool1.

You execute the data flow and capture the execution information.

You need to identify how long it takes to write the data to Pool1.

Which metric should you use?

- A. the rows written
- B. the sink processing time
- C. the transformation processing time
- D. the post processing time

Correct Answer: B

 **jongert** 4 days, 16 hours ago

Correct:

Each transformation stage includes a total time for that stage to complete with each partition execution time totaled together. When you select the Sink, you see "Sink Processing Time". This time includes the total of the transformation time plus the I/O time it took to write your data to your destination store. The difference between the Sink Processing Time and the total of the transformation is the I/O time to write the data.

<https://learn.microsoft.com/en-us/azure/data-factory/concepts-data-flow-monitoring#total-sink-processing-time-vs-transformation-processing-time>

upvoted 1 times

Topic 5 - Testlet 1

Introductory Info

Case study -

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case.

However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview -

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly. You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

Purge Twitter feed data records that are older than two years.

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse

Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question

HOTSPOT -

You need to design a data storage structure for the product sales transactions. The solution must meet the sales transaction dataset requirements.

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

Table type to store the product sales transactions:

| |
|-------------|
| Hash |
| Round-robin |
| Replicated |

When creating the table for sales transactions:

| |
|--|
| Configure a clustered index. |
| Set the distribution column to product ID. |
| Set the distribution column to the sales date. |

Correct Answer:

Answer Area

Table type to store the product sales transactions:

| |
|-------------|
| Hash |
| Round-robin |
| Replicated |

When creating the table for sales transactions:

| |
|--|
| Configure a clustered index. |
| Set the distribution column to product ID. |
| Set the distribution column to the sales date. |

Box 1: Hash -

Scenario:

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

A hash distributed table can deliver the highest query performance for joins and aggregations on large tables.

Box 2: Set the distribution column to the sales date.

Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Reference:

<https://rajanieshkaushikk.com/2020/09/09/how-to-choose-right-data-distribution-strategy-for-azure-synapse/>

✉  **Jerrie86** Highly Voted 11 months, 2 weeks ago

This case study was in my exam and I scored 970. I chose productid.

upvoted 43 times

✉  **jongert** 4 days, 14 hours ago

Congrats! The answer is productid, since ms documentation states NOT to distribute by a date column. When doing so, all data for a given date is partitioned into one distribution. When processing, this hinders parallelism.

upvoted 2 times

✉  **RoyP654** 7 months ago

Good Job, Congrats!

upvoted 4 times

✉  **Julia01** Highly Voted 1 year, 4 months ago

I'd choose product id as well since it will be used in joins "Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible."

upvoted 21 times

✉  **mokrani** 1 year, 2 months ago

Why not sales date for distribution column ?

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right...

upvoted 1 times

✉  **kl8585** 1 year, 1 month ago

because it's asking about distribution, not partition. The requirements say "ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible". The best way to do so is hash distributing on product ID, this way all rows with the same product id will be on the same node and there will be no data shuffling, hence fast queries

upvoted 15 times

✉  **kkk5566** Most Recent 4 months, 1 week ago

Hash and Distribution on Product ID

upvoted 2 times

✉  **XiltroX** 1 year, 1 month ago

In MS's own documentation, it is not recommended to use a date column for distribution. Therefore, the second option should be ProductID

upvoted 8 times

✉  **pavankr** 7 months, 2 weeks ago

So then why this guy is misleading us?? I find lot of answers misleading us.

upvoted 3 times

□ **OldSchool** 1 year, 1 month ago

Hash and Distribution on Product ID, never make distribution on Date.:

<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>

Partition on Date as explained here:

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition>

upvoted 11 times

□ **kornat** 9 months, 1 week ago

True! !!

upvoted 2 times

□ **berend1** 1 year, 2 months ago

Partition column: date, distribution column: ProductID

upvoted 5 times

□ **greenlever** 1 year, 3 months ago

I think so, Set distribution to Product ID

upvoted 3 times

□ **pangas2567** 1 year, 4 months ago

Why not Set distribution to Product ID? With the date as the distribution column we lose the advantage of using all 60 nodes, right?

upvoted 9 times

Introductory Info

Case study -

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case.

However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview -

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly. You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

Purge Twitter feed data records that are older than two years.

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse

Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question

DRAG DROP -

You need to ensure that the Twitter feed data can be analyzed in the dedicated SQL pool. The solution must meet the customer sentiment analytics requirements.

Which three Transact-SQL DDL commands should you run in sequence? To answer, move the appropriate commands from the list of commands 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:

Commands

CREATE EXTERNAL DATA SOURCE
CREATE EXTERNAL FILE FORMAT
CREATE EXTERNAL TABLE
CREATE EXTERNAL TABLE AS SELECT
CREATE DATABASE SCOPED CREDENTIAL

Answer Area

Correct Answer:

Commands:

CREATE EXTERNAL DATA SOURCE
CREATE EXTERNAL FILE FORMAT
CREATE EXTERNAL TABLE
CREATE EXTERNAL TABLE AS SELECT
CREATE DATABASE SCOPED CREDENTIAL

Answer Area

CREATE EXTERNAL DATA SOURCE
CREATE EXTERNAL FILE FORMAT
CREATE EXTERNAL TABLE AS SELECT

Scenario: Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds. Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Box 1: CREATE EXTERNAL DATA SOURCE

External data sources are used to connect to storage accounts.

Box 2: CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL FILE FORMAT creates an external file format object that defines external data stored in Azure Blob Storage or Azure Data Lake Storage.

Creating an external file format is a prerequisite for creating an external table.

Box 3: CREATE EXTERNAL TABLE AS SELECT

When used in conjunction with the CREATE TABLE AS SELECT statement, selecting from an external table imports data into a table within the SQL pool. In addition to the COPY statement, external tables are useful for loading data.

Incorrect Answers:

CREATE EXTERNAL TABLE -

The CREATE EXTERNAL TABLE command creates an external table for Synapse SQL to access data stored in Azure Blob Storage or Azure Data Lake Storage.

Reference:

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

□ **AzureJobsTillRetire** Highly Voted 1 year ago

Given answers are correct

Box 1: CREATE EXTERNAL DATA SOURCE

Box 2: CREATE EXTERNAL FILE FORMAT

Box 3: CREATE EXTERNAL TABLE AS SELECT

Requirements: Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds. Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Why CREAT DATABASE SCOPED CREDENTIAL is not required?

Requirement: The users must be authenticated by using their own Azure AD credentials

Why not CREATE EXTERNAL TABLE?

Requirement: Allow Contoso users to use PolyBase ... to query ...

PolyBase has limitations. CREATE EXTERNAL TABLE AS SELECT stored the data within the SQL pool and avoids those limitations.

<https://learn.microsoft.com/en-us/sql/relational-databases/polybase/polybase-versioned-feature-summary?view=sql-server-ver16>

upvoted 21 times

□ **vrodriguesp** 10 months, 4 weeks ago

are you sure we can create EXTERNAL DATA SOURCE without DATABASE SCOPED CREDENTIAL?

upvoted 3 times

□ **JG1984** 6 months, 3 weeks ago

It is not necessary if the users are already authenticated by using their own Azure AD credentials.

upvoted 3 times

□ **JasonVu** 1 year ago

CETAS is not available in dedicated SQL pool

upvoted 2 times

□ **matiandal** 4 months, 1 week ago

your r making a mistake mate.

Check the following link from MS Learn

R: <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>

"You can use a CREATE EXTERNAL TABLE AS SELECT (CETAS) statement in
--> a dedicated SQL pool

OR

--> serverless SQL pool to persist the results of a query in an external table, which stores its data in a file in the data lake."

upvoted 3 times

□ **Ram9198** 4 months, 3 weeks ago

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop> external tables are supported in both SQL pools

upvoted 2 times

□ **AzureJobsTillRetire** 1 year ago

Also this one.

CREATE EXTERNAL TABLE AS SELECT (Transact-SQL)

Applies to: SQL Server 2022 (16.x) and later, Azure Synapse Analytics, Analytics Platform System (PDW)

<https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-as-select-transact-sql?view=aps-pdw-2016-au7>

upvoted 2 times

□ **juanlu46** Highly Voted 1 year, 3 months ago

1. Scoped Database Credential

2. External Data Source

3 External File Format

upvoted 7 times

scarycat 1 year, 1 month ago

Scoped Database Credencial is a DCL command, not DDL

upvoted 3 times

OldSchool 1 year, 1 month ago

Correct

upvoted 2 times

Abdulwahab1983 [Most Recent] 1 month, 2 weeks ago

twitter feeds are going to be stored in azure storage which also going to need data life cycle management. If we are not storing the data in the dedicated sql pool table then we do not use CETAS we only create an external table to query the data in the azure storage.

upvoted 1 times

kkk5566 4 months, 1 week ago

DS,format,CETAS

upvoted 1 times

patjoo 4 months, 3 weeks ago

According to Microsoft documentation:

You can create external tables in Synapse SQL pools via the following steps:

CREATE EXTERNAL DATA SOURCE to reference an external Azure storage and specify the credential that should be used to access the storage.

CREATE EXTERNAL FILE FORMAT to describe format of CSV or Parquet files.

CREATE EXTERNAL TABLE on top of the files placed on the data source with the same file format.

<https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables?tabs=hadoop#external-tables-in-dedicated-sql-pool-and-serverless-sql-pool>

upvoted 2 times

Ram9198 4 months, 3 weeks ago

Answer should CET - RLS is supported on external tables and you do not need CETAS to implement RLS refer <https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16>

upvoted 1 times

Ram9198 4 months, 3 weeks ago

<https://learn.microsoft.com/en-us/answers/questions/739341/rowlevelsecurity-on-external-table>. RLS is not supported on an external table, then how CETAS be an answer

upvoted 1 times

Matt2000 5 months ago

Concerning not needing CREATE DATABASE SCOPED CREDENTIAL for CREATE EXTERNAL DATA SOURCE: "External data source without credential can access public storage account or use the caller's Azure AD identity to access files on Azure storage."

Ref: <https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azuresqldw-latest&tabs=dedicated>

upvoted 1 times

BPW 8 months, 2 weeks ago

Box 1: CREATE EXTERNAL DATA SOURCE

Box 2: CREATE EXTERNAL FILE FORMAT

Box 3: CREATE EXTERNAL TABLE

upvoted 5 times

MartianNC 9 months, 2 weeks ago

The reason you use CTAS is that you must implement row level security.

upvoted 2 times

Jerrie86 11 months, 2 weeks ago

Starting with SQL Server 2022 (16.x), Create External Table as Select (CETAS) is supported to create an external table and then export, in parallel, the result of a Transact-SQL SELECT statement to Azure Data Lake Storage (ADLS) Gen2, Azure Storage Account V2, and S3-compatible object storage.

So shouldnt third be Create External TABLE ?

We dont want to write data to ADLS. We want to read.

<https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-as-select-transact-sql?view=azuresqldw-latest&preserve-view=true>

upvoted 5 times

JitBiswas 8 months ago

You are right. The question is asking to "read" the tweeter feed stored as parquet file in ADLS via PolyBase. This is supported with CREATE EXTERNAL TABLE - which in turn reads data from ADLS. Please refer <https://learn.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql?view=sql-server-ver16&tabs=dedicated>

It is mentioned - "This command creates an external table for PolyBase to access data stored in a Hadoop cluster or Azure Blob Storage PolyBase external table that references data stored in a Hadoop cluster or Azure Blob Storage."

upvoted 1 times

youngbug 11 months, 3 weeks ago

PolyBase is a technology that accesses external data stored in Azure Blob storage or Azure Data Lake Store via the T-SQL language. So no need to copy table into Dedicated SQL Pool.

upvoted 1 times

bigw 1 year ago

why use CETAS instead of Create External Table?

upvoted 1 times

Pais 1 year ago

<https://learn.microsoft.com/en-us/sql/t-sql/statements/create-table-as-select-azure-sql-data-warehouse?toc=%2Fazure%2Fsynapse-analytics%2Fsql-data-warehouse%2Fbreadcrumb%2Ftoc.json&bc=%2Fazure%2Fsynapse-analytics%2Fsql-data-warehouse%2Flatest&view=azure-sqldw-latest&preserve-view=true#examples-using-ctas-to-replace-sql-server-code>

upvoted 2 times

JasonVu 1 year ago

your link points to CTAS, which is a different topic

upvoted 1 times

Igor85 1 year ago

CREATE DATABASE SCOPED CREDENTIALS should be run before all other steps in the given answer

upvoted 1 times

7yut 1 year, 1 month ago

I think the provided answer in answer are is correct

upvoted 2 times

greenlever 1 year, 3 months ago

External file format is required when external table needs to refer to Hadoop files

upvoted 1 times

Introductory Info

Case study -

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case.

However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview -

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly. You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

- Purge Twitter feed data records that are older than two years.

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse

Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question

HOTSPOT -

You need to design the partitions for the product sales transactions. The solution must meet the sales transaction dataset requirements.

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

Partition product sales transactions data by:

| |
|--------------|
| Sales date |
| Product ID |
| Promotion ID |

Store product sales transactions data in:

| |
|--|
| An Azure Synapse Analytics dedicated SQL pool |
| An Azure Synapse Analytics serverless SQL pool |
| An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace |
| An Azure Data Lake Storage Gen2 account |

Answer Area

Partition product sales transactions data by:

| |
|--------------|
| Sales date |
| Product ID |
| Promotion ID |

Correct Answer:

Store product sales transactions data in:

| |
|--|
| An Azure Synapse Analytics dedicated SQL pool |
| An Azure Synapse Analytics serverless SQL pool |
| An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace |

Box 1: Sales date -

Scenario: Contoso requirements for data integration include:

- Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Box 2: An Azure Synapse Analytics Dedicated SQL pool

Scenario: Contoso requirements for data integration include:

- Ensure that data storage costs and performance are predictable.

The size of a dedicated SQL pool (formerly SQL DW) is determined by Data Warehousing Units (DWU).

Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage. This format significantly reduces the data storage costs, and improves query performance.

Synapse analytics dedicated sql pool

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-overview-what-is>

✉  **Jerrie86** Highly Voted 11 months, 3 weeks ago

Partition is different than distribution. Distribution=ProductID and partition by Date.

Distribution:

When you store a table on Azure DW you are storing it amongst 60 nodes. Your table data is distributed across these nodes (using Hash distribution or Round Robin distribution depending on your needs). You can also choose to have your table (preferably a very small table) replicated across these nodes.

Partition : Partitioning is completely divorced from this concept of distribution. When we partition a table we decide which rows belong into which partitions based on some scheme (like date in this case) Chunk of records for that date range gets its own space in the backend behind the scenes. we can partition data based on anything as long as we know how the data is in our system.

And when we put both in use together, all the partitions are horizontally partitioned so that the incoming data is divided into 60 nodes to provide extreme parallelization to the queries.

<https://www.linkedin.com/pulse/partitioning-distribution-azure-synapse-analytics-swapnil-mule>

upvoted 17 times

✉  **DataEngDP** Most Recent 3 months, 3 weeks ago

Load the sales transaction dataset to Azure Synapse Analytics---HERE you have the answer on where to store the "transactional" data---ONLY POSSIBILITY is Azure Synapse Analytics Dedicated SQL Pool.

upvoted 2 times

✉  **kkk5566** 4 months, 1 week ago

Partition by date &dedicated pool

upvoted 2 times

✉  **gerrie1979** 1 year, 2 months ago

As far as I see it, we need to distribute the fact table accross the 60 distributions of a dedicated sql pool which means using NO date key (because of MPP) so using the productId key and within each distribution we need to partition the data by the date column so that data can quickly be deleted and queried by all 60 distributions at once

upvoted 2 times

✉  **Jerrie86** 11 months, 2 weeks ago

First question is partition not distribution. So Date is correct

upvoted 3 times

✉  **sensaint** 1 year, 2 months ago

I would partition by ProductID since joins and filtering must be optimized for that column

upvoted 1 times

✉  **mokrani** 1 year, 2 months ago

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Also we will delete data using sales date

I think distribution = ProductID , Partition = Sales_date

upvoted 16 times

✉  **sensaint** 1 year ago

Correct. Forget above statement. Partition should be Sales Date!!

upvoted 7 times

✉  **Igor85** 1 year ago

don't confuse partitions and distribution for hash-distributed table

upvoted 4 times

Introductory Info

Case study -

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case.

However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview -

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises

Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL

Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

Purge Twitter feed data records that are older than two years.

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse

Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question

You need to implement the surrogate key for the retail store table. The solution must meet the sales transaction dataset requirements.

What should you create?

- A. a table that has an IDENTITY property
- B. a system-versioned temporal table
- C. a user-defined SEQUENCE object
- D. a table that has a FOREIGN KEY constraint

Correct Answer: A

Scenario: Implement a surrogate key to account for changes to the retail store addresses.

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>

 **kkk5566** 4 months, 1 week ago

Selected Answer: A

is correct

upvoted 2 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: A

A surrogate key is a system-generated unique identifier that is used as a substitute for a natural key. In this case, the surrogate key will be used to account for changes to the retail store addresses.

By creating a table with an IDENTITY property, you can ensure that a unique surrogate key is automatically generated for each row inserted into the table. The IDENTITY property assigns a unique value to the column automatically, incrementing by one for each new row.

Using an IDENTITY column as the surrogate key will provide an efficient way to join and filter sales transaction records based on product ID, as required by the sales transaction dataset requirements.

upvoted 4 times

 **sntlkumar** 8 months, 1 week ago

Given answer is correct

upvoted 2 times

 **uira** 1 year, 1 month ago

Selected Answer: A

Identity should be used.

upvoted 3 times

 **7yut** 1 year, 1 month ago

Selected Answer: A

Correct

upvoted 3 times

 **anks84** 1 year, 4 months ago

Selected Answer: A

A is the correct Answer !

upvoted 4 times

Introductory Info

Case study -

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case.

However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview -

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly. You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

Purge Twitter feed data records that are older than two years.

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse

Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question

HOTSPOT -

You need to design an analytical storage solution for the transactional data. The solution must meet the sales transaction dataset requirements.

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

Table type to store retail store data:

| |
|-------------|
| Hash |
| Replicated |
| Round-robin |

Table type to store promotional data:

| |
|-------------|
| Hash |
| Replicated |
| Round-robin |

Answer Area

Table type to store retail store data:

Correct Answer:

| |
|-------------|
| Hash |
| Replicated |
| Round-robin |

Table type to store promotional data:

| |
|-------------|
| Hash |
| Replicated |
| Round-robin |

Box 1: Round-robin -

Round-robin tables are useful for improving loading speed.

Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month.

Box 2: Hash -

Hash-distributed tables improve query performance on large fact tables.

Scenario:

⇒ You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

⇒ Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>

□  **greenlever** Highly Voted 1 year, 3 months ago

replicated
hash
upvoted 41 times

□  **Jerrie86** Highly Voted 11 months, 2 weeks ago

Data is more than 100GB : hash
Dimension data less than 2GB: replicated
Staging table data less than 5Gb:Round Robin

So replicated and Hash
upvoted 17 times

□  **hassexat** Most Recent 4 months ago

replicated & hash
upvoted 1 times

□  **hassexat** 4 months ago

Replicated --> Because is not a staging table and is moreless 2GB
Hash --> Because is 200GB
upvoted 1 times

□  **kkk5566** 4 months, 1 week ago

replicated ,hash tables are best for queries with joins and aggregations.
upvoted 1 times

□  **peacejh** 5 months, 1 week ago

In the text it says that the table is 200GB, so hash. In the answer explanation it suddenly is only 5 GB
upvoted 1 times

□  **andjurovicela** 5 months, 2 weeks ago

Box1: (clearly) replicated
Box2: I can see why someone would say round-robin since it is not uncommon for large dim_tables (and that is what this promotions table will essentially be) to use this distribution BUT per Microsoft doc below using round-robin makes sense in situation tht simply do not apply here:
- When getting started as a simple starting point since it is the default --> NOT THE CASE
- If there is no obvious joining key --> THERE IS, product id which will be present in the fact transactions table as well
If there is no good candidate column for hash distributing the table - THERE IS, PromotionID
- If the table does not share a common join key with other tables - IT DOES, ProductID
- If the join is less significant than other joins in the query - NO INF on this
- When the table is a temporary staging table - IT IS NOT

source: <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>
therefore, box2: Hash
upvoted 2 times

✉ **auwia** 6 months, 2 weeks ago

Retail Store contains info about store (like address), it's clearly a dimension table, by the consequence it is REPLICATED.
The second is correct: HASH.
upvoted 3 times

✉ **pavankr** 7 months, 2 weeks ago

So on which answer we should reply on?????? Why this web site guy is guiding us all wrong answers?????
upvoted 3 times

✉ **JosephVishal** 11 months, 3 weeks ago

Box1: Replicated
Box2: Hash. Since, the Retail store table, will be used in queries and there is no mention of data loads to this table. It should be replicated and not Round-Robin.
upvoted 2 times

✉ **Taou** 1 year ago

1st is Replicated
upvoted 2 times

✉ **AzureJobsTillRetire** 1 year, 1 month ago

Box1: Replicated. As the Retail Store is going to be replicated in each distribution to facilitate SQL queries.

Box2: Hash for large fact tables
upvoted 1 times

✉ **smsme323** 1 year, 3 months ago

replicated
HAsh
upvoted 2 times

✉ **juanlu46** 1 year, 3 months ago

-Replicated
-Hash
upvoted 3 times

✉ **anks84** 1 year, 4 months ago

Looks like "Retail Store" is a dimension table with 2MB size.
So, Replicated should be better option in my opinion,
upvoted 5 times

✉ **federc** 1 year, 4 months ago

Agree with you Julia01, Replicated would be more reasonable for a 2MB table
upvoted 2 times

✉ **R12346** 1 year, 4 months ago

The retail table should be of replicated type since it is just 2 MB
upvoted 3 times

Introductory Info

Case study -

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case.

However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview -

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly. You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

Purge Twitter feed data records that are older than two years.

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse

Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question

HOTSPOT -

You need to implement an Azure Synapse Analytics database object for storing the sales transactions data. The solution must meet the sales transaction dataset requirements.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Transact-SQL DDL command to use:

| |
|-----------------------|
| ▼ |
| CREATE EXTERNAL TABLE |
| CREATE TABLE |
| CREATE VIEW |

Partitioning option to use in the WITH clause of the DDL statement:

| |
|------------------------|
| ▼ |
| FORMAT_OPTIONS |
| FORMAT_TYPE |
| RANGE LEFT FOR VALUES |
| RANGE RIGHT FOR VALUES |

Correct Answer:

Answer Area

Transact-SQL DDL command to use:

| |
|-----------------------|
| ▼ |
| CREATE EXTERNAL TABLE |
| CREATE TABLE |
| CREATE VIEW |

Partitioning option to use in the WITH clause of the DDL statement:

| |
|------------------------|
| ▼ |
| FORMAT_OPTIONS |
| FORMAT_TYPE |
| RANGE LEFT FOR VALUES |
| RANGE RIGHT FOR VALUES |

Box 1: Create table -

Scenario: Load the sales transaction dataset to Azure Synapse Analytics

Box 2: RANGE RIGHT FOR VALUES -

Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

RANGE RIGHT: Specifies the boundary value belongs to the partition on the right (higher values).

FOR VALUES (boundary_value [,...n]): Specifies the boundary values for the partition.

Scenario: Load the sales transaction dataset to Azure Synapse Analytics.

Contoso identifies the following requirements for the sales transaction dataset:

- ☞ Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.
- ☞ Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.
- ☞ Implement a surrogate key to account for changes to the retail store addresses.
- ☞ Ensure that data storage costs and performance are predictable.
- ☞ Minimize how long it takes to remove old records.

Reference:

<https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse>

✉  **XiltroX**  1 year, 1 month ago

Its funny cause in the scenario, there is a BIG hint on what to use for box 2. Read it up.

upvoted 11 times

✉  **LijuLJ** 5 months, 4 weeks ago

its not fun , you will get it only when you read it fully and carefully :) :P

upvoted 3 times

✉  **Azurre**  9 months, 3 weeks ago

Hint as per XiltroX:

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

upvoted 10 times

✉  **SenMia**  3 weeks, 3 days ago

can someone help me with on first one, why is it create table and not external table?

upvoted 1 times

✉  **jongert** 4 days, 3 hours ago

External table is used when using an external data source such as data stored in Azure Data Lake Storage Gen2. In this case, as seen in a previous question, the data will be stored in the columnar store of the dedicated SQL pool.

upvoted 2 times

✉  **meatpoof** 1 week, 1 day ago

as far as i see, syntax only belongs to create table

<https://learn.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse?view=aps-pdw-2016-au7#PartitionedTable>

upvoted 1 times

✉  **AndreiG** 2 months, 3 weeks ago

Why create table and not create external table ?

upvoted 2 times

✉  **kkk5566** 4 months, 1 week ago

<https://learn.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse?view=aps-pdw-2016-au7#ExTablePartitions>

upvoted 2 times

✉  **kkk5566** 4 months ago

create table ,right

upvoted 2 times

✉  **Xinyuehong** 1 year, 2 months ago

agreed

upvoted 2 times

✉  **federC** 1 year, 4 months ago

correct

upvoted 2 times

Introductory Info

Case study -

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case.

However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview -

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly. You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

- Purge Twitter feed data records that are older than two years.

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse

Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question

You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements.

Which Azure Storage functionality should you include in the solution?

- A. change feed
- B. soft delete
- C. time-based retention
- D. lifecycle management

Correct Answer: D

Scenario: Purge Twitter feed data records that are older than two years.

Data sets have unique lifecycles. Early in the lifecycle, people access some data often. But the need for access often drops drastically as the data ages. Some data remains idle in the cloud and is rarely accessed once stored. Some data sets expire days or months after creation, while other data sets are actively read and modified throughout their lifetimes. Azure Storage lifecycle management offers a rule-based policy that you can use to transition blob data to the appropriate access tiers or to expire data at the end of the data lifecycle.

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-overview>

 **yogiazaad** Highly Voted 11 months, 2 weeks ago

Selected Answer: D

Given answer is correct.

Time bases retention is to retain data for a specific time. it wont delete the data. The requirement is to deleted the data after 2 Years. Which can be accomplished by Data life cycle management.

A time-based retention policy stores blob data in a Write-Once, Read-Many (WORM) format for a specified interval. When a time-based retention policy is set, clients can create and read blobs, but can't modify or delete them. After the retention interval has expired, blobs can be deleted but not overwritten.

<https://learn.microsoft.com/en-us/azure/storage/blobs/immutable-time-based-retention-policy-overview>

upvoted 8 times

 **yogiazaad** 11 months, 2 weeks ago

A time-based retention policy protects against deletion of blob while it is in effect. Note that it will not automatically delete the blob after the retention period.

upvoted 2 times

 **Momoanwar** Most Recent 2 weeks, 1 day ago

Selected Answer: D

Chatgpt say d

Based on the case study information provided, the best Azure Storage functionality to include in the solution for data retention of Twitter feed data records is:

D. Lifecycle Management

This feature enables the creation of rules to manage the lifecycle of the data stored in Azure Blob Storage. It can automate the process of purging Twitter feed data records that are older than two years, aligning with the requirement for minimal administrative effort and ensuring compliance with the data retention policy.

upvoted 1 times

 **kkk5566** 4 months, 1 week ago

Selected Answer: D

is correct

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: D

Azure Storage provides a feature called lifecycle management, which allows you to define rules to manage the retention and deletion of data in your storage account. Lifecycle management enables you to automatically transition, delete, or take other actions on objects in your storage account based on specified conditions.

upvoted 2 times

 **cale** 9 months, 1 week ago

Selected Answer: D

Answer is D

upvoted 3 times

 **haidebelognime** 10 months, 2 weeks ago

Selected Answer: C

do the research. it is C time-based retention

upvoted 1 times

 **JG1984** 6 months, 3 weeks ago

Time-based retention is a feature of Azure Blob storage that allows you to automatically delete blobs that have reached a certain age. While this feature can be useful for automatically deleting old data, it is not the most appropriate solution for the scenario you described because it only applies to block blobs and append blobs, and it is not available for other types of data in Azure Storage, such as files or tables.

In contrast, lifecycle management is a more comprehensive solution that allows you to define rules to automatically transition data to different access tiers or expire data at the end of its lifecycle. This functionality applies to block blobs, append blobs, and versioned block blobs, and it provides more flexibility in defining data retention policies.

upvoted 2 times

 **vctrhugo** 6 months, 1 week ago

Time-based retention enables users to store business-critical data in a WORM (Write Once, Read Many) state. While in a WORM state, data cannot be modified or deleted for a user-specified interval.

upvoted 2 times

 **Ast999** 10 months, 1 week ago

You are wrong. As it was said few times. Time-based retention will protect the data during set period from deletion but it won't delete it automatically after set time.

upvoted 4 times

 **MrWood47** 11 months, 3 weeks ago

Selected Answer: C

Sima_al explanation is correct

upvoted 1 times

 **nicky87654** 11 months, 3 weeks ago

Selected Answer: C

C: time-based retention

upvoted 2 times

 **Sima_al** 1 year ago

C: time-based retention

Based on the customer sentiment analytics requirements, you should include time-based retention in the data retention solution for the Twitter feed data records. Time-based retention allows you to specify a retention period for data in Azure Storage and ensures that data is not deleted before its retention period expires. This functionality can be used to meet the requirement to purge Twitter feed data records that are older than two years.

Option A (change feed) is a feature of Azure Table Storage and Azure Cosmos DB that provides a stream of change events on a table or container.

Option B (soft delete) is a feature of Azure Table Storage and Azure Cosmos DB that allows you to mark an entity as deleted without permanently deleting it. This allows you to recover deleted data if necessary.

Option D (lifecycle management) is a feature of Azure Blob Storage that allows you to specify policies for automatically transitioning blobs to different storage tiers or deleting them based on their age or access patterns.

upvoted 3 times

yogiazzaad 11 months, 2 weeks ago

Time bases retention is not the correct answer.

A time-based retention policy protects against deletion of blob while it is in effect. Note that it will not automatically delete the blob after the retention period.

upvoted 3 times

Snomax 1 year, 2 months ago

Selected Answer: D

Agreed.

upvoted 3 times

Topic 6 - Testlet 2

Introductory Info

Case study -

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case.

However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview -

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements -

Business Goals -

Litware wants to create a new analytics environment in Azure to meet the following requirements:

See inventory levels across the stores. Data must be updated as close to real time as possible.

Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.

Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements -

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals.

Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.

Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

Use Azure Active Directory (Azure AD) authentication whenever possible.

Use the principle of least privilege when designing security.

Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data

Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment -

Litware plans to implement the following environment:

The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure. Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Daily inventory data comes from a Microsoft SQL server located on a private network.

Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.

Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.

Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

Question

HOTSPOT -

Which Azure Data Factory components should you recommend using together to import the daily inventory data from the SQL server to Azure Data Lake Storage?

To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Integration runtime type:

- Azure integration runtime
- Azure-SSIS integration runtime
- Self-hosted integration runtime

Trigger type:

- Event-based trigger
- Schedule trigger
- Tumbling window trigger

Activity type:

- Copy activity
- Lookup activity
- Stored procedure activity

Answer Area

Integration runtime type:

- Azure integration runtime
- Azure-SSIS integration runtime
- Self-hosted integration runtime

Trigger type:

- Event-based trigger
- Schedule trigger
- Tumbling window trigger

Activity type:

- Copy activity
- Lookup activity
- Stored procedure activity

Box 1: Self-hosted integration runtime

A self-hosted IR is capable of running copy activity between a cloud data stores and a data store in private network.

Box 2: Schedule trigger -

Schedule every 8 hours -

Box 3: Copy activity -

Scenario:

- ☞ Customer data, including name, contact information, and loyalty number, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
- ☞ Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

□ **ArvindK06** Highly Voted 2 years, 2 months ago

Should be Tumbling Window in my opinion. Since Inventory data should be updated in real time as close as possible. Only Customer & Product data are available every 8 hours.

upvoted 23 times

□ **andjurovicela** 5 months, 2 weeks ago

I think you are mixing up different requirements. Daily inventory, should be ingested daily, I would say and for that schedule trigger makes sense. It would make less sense to use a thumbling window, not to mention that you can't even use "day" as the first argument (time interval) for the tumbling window function when coding, only hour as the biggest measurement unit

upvoted 2 times

□ **ANath** 1 year, 10 months ago

I also think it should be a Tumbling Window. Because it said 'Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.'

upvoted 3 times

□ **AzureJobsTillRetire** Highly Voted 1 year, 1 month ago

Agreed with the given answers.

Box1: Self-hosted integration runtime

Why not Azure-SSIS integration rutime? SSIS is not mentioned, and the ETL tool in use is ADF.

Why not Azure integration runtime? On-premise SQL Server database is used.

Box2: Schedule trigger

Why not event-based trigger? Schedule runs every 8 hours

Why not tumbling window schedule? There is no requirement for a tumbling window schedule. If the ETL jobs run close to 8 hours, a tumbling window schedule may be required. If jobs need to automatically re-run on failures, a tumbling window schedule may be required. Those requirements are not there. Schedule trigger fits for purpose.

Box3: Copy activity

No need for explanation

upvoted 22 times

□ **vrodriguesp** 10 months, 3 weeks ago

agree with you

upvoted 1 times

□ **kkk5566** Most Recent 4 months ago

Self-hosted,Schedule ,Copy data

upvoted 1 times

□ **Bhuvanesh2104** 12 months ago

The below link refers to the (a) Azure Integration Runtime: <https://learn.microsoft.com/en-us/azure/data-factory/tutorial-managed-virtual-network-on-premises-sql-server>

upvoted 2 times

□ **andjurovicela** 6 months ago

the link you provided clearly states that virtual networks should be in place whereas the task says "Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure." Therefore self-hosted integration runtime is the answer (as ExamTopics suggested)

upvoted 1 times

□ **sunil_smile** 1 year, 3 months ago

I think it should be SSIS integration runtime... Because currently there are SSIS pipelines which does the data integration

upvoted 2 times

□ **Deeksha1234** 1 year, 4 months ago

In opinion the given answer is correct since its daily inventory data, i.e. will be loaded once daily.

upvoted 3 times

□ **Canary_2021** 2 years ago

Answers are correct because 'Daily inventory data comes from a Microsoft SQL server located on a private network.'

upvoted 2 times

□ **dija123** 2 years, 1 month ago

I believe a Microsoft SQL server located on a private network means on Azure not on premises, which means the integration run time should be azure not self hosted.

upvoted 8 times

□ **Davico93** 1 year, 6 months ago

maybe.... even if it is an azure resource, but in private network, we need SelfHosted

upvoted 2 times

□ **datnguye** 2 years ago

Should it be Self-hosted as Microsoft SQL server, not Azure though?

upvoted 1 times

□ **azurearmy** 2 years, 2 months ago

The answers are correct.

upvoted 18 times

□ **AppleVan** 2 years, 3 months ago

Shouldn't it be event based?

upvoted 3 times

□ **rikku33** 2 years, 3 months ago

Schedule trigger - because daily. so the given answer is correct

upvoted 5 times

□ **samko92** 2 years, 2 months ago

It is confusing cos at the top it says they want the inventory as real time as possible , but then further down it says every 8 hours. Conflicting info

upvoted 3 times

□ **kl8585** 1 year, 1 month ago

read carefully - import every 8 hours for customer date, not inventory data. Event triggers can be used only with storage account, so Event based is for sure wrong. It's tumbling windows

upvoted 1 times

□ **OldSchool** 1 year, 1 month ago

It says: "Daily inventory data comes from a Microsoft SQL server located on a private network." So the answer is as given: Self-hosted, Schedule, Copy

upvoted 1 times

Topic 7 - Testlet 3

Introductory Info

Case study -

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case.

However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview -

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly. You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

Purge Twitter feed data records that are older than two years.

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse

Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question

DRAG DROP -

You need to implement versioned changes to the integration pipelines. The solution must meet the data integration requirements.

In which order should you perform the actions? 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

- Merge changes
- Create a pull request
- Create a feature branch
- Publish changes
- Create a repository and a main branch

Answer Area



Correct Answer:

Actions

-
-
-
-
-

Answer Area

- Create a repository and a main branch
- Create a feature branch
- Create a pull request
- Merge changes
- Publish changes

Scenario: Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Step 1: Create a repository and a main branch

You need a Git repository in Azure Pipelines, TFS, or GitHub with your app.

Step 2: Create a feature branch -

Step 3: Create a pull request -

Step 4: Merge changes -

Merge feature branches into the main branch using pull requests.

Step 5: Publish changes -

Reference:

<https://docs.microsoft.com/en-us/azure/devops/pipelines/repos/pipeline-options-for-git>

□  **ItHYMeRish** Highly Voted 2 years ago

The answer provided is correct
upvoted 37 times

□  **SameerL** 1 year, 5 months ago

The provided sequence is correct per below link:

<https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-delivery>
upvoted 9 times

□  **NaiCob** Highly Voted 2 years ago

Before creating a pull request, it is required to save our changes on a feature branch (publish our local changes). So the correct order is:

1. Create a repository and a main branch
2. Create a feature branch
3. Publish changes
4. Create a pull request
5. Merge changes

upvoted 18 times

□  **Igor85** 1 year ago

no, publish you can only do from the main branch. to publish changes from main you first have to create a PR, get approval, merge to main.
upvoted 5 times

□  **wwdba** 1 year, 10 months ago

I agree. This was my order too. My understanding is: Publish changes = Push the changes to the remote repository
upvoted 1 times

□  **corebit** 2 years ago

@NaiCob I believe the given answer is correct. What changes are published before creating a PR?
upvoted 3 times

□  **NaiCob** 2 years ago

Before Pull Request you have to publish your local changes
upvoted 1 times

□  **xeti** 1 year, 10 months ago

No, the given answer is correct.

□  **dev2dev** 1 year, 11 months ago
Publish is done after the merge to collaboration (main) branch and is essentially a CI trigger to update the adf_publish branch.
upvoted 4 times

□  **dev2dev** 1 year, 11 months ago

Nope. Given answers are correct.
upvoted 2 times

□  **Adediwura** Most Recent 2 months, 2 weeks ago

The given answer is correct. Please note that publish change is not same as committing.
upvoted 1 times

□  **kkk5566** 4 months, 1 week ago

correct
upvoted 1 times

□  **Jerrie86** 11 months, 2 weeks ago

This case study was in my exam word to word. Thanks guys. Passed at 970.
Just don't do dumps but try to understand the logic via some YouTube DP-203 tutorials. There is a series by databag.ai. So please study to excel in exam as well in prof life.
upvoted 3 times

 **anks84** 1 year, 4 months ago

Given answer and sequence is absolutely correct !!

upvoted 2 times

 **Deeksha1234** 1 year, 4 months ago

given answer is correct

upvoted 2 times

 **hbad** 1 year, 8 months ago

Given answer is correct:

1. Create a repository and a main branch -You need a Git repository in Azure Pipelines, TFS, or GitHub with your app.
2. Create a feature branch -
3. Create a pull request - you propose that changes you've made on a head branch should be merged
4. Merge changes -merge feature branches into the main/collaboration branch using pull requests.
5. Publish changes -after you merged changes to the collaboration branch (main is default), click Publish to manually publish.

<https://docs.microsoft.com/en-us/azure/data-factory/source-control>

<https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/getting-started/about-collaborative-development-models>

<https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/proposing-changes-to-your-work-with-pull-requests/about-pull-requests>

upvoted 2 times

 **Send2** 1 year, 8 months ago

<https://www.atlassian.com/git/tutorials/comparing-workflows/feature-branch-workflow>

upvoted 1 times

 **Kondzio** 1 year, 10 months ago

I think it's correct. Publish step is the last one, because there is no auto-publish on the master branch by default

upvoted 2 times

 **edba** 2 years ago

I think answer is correct. Pls refer to <https://docs.microsoft.com/en-us/azure/data-factory/source-control#version-control>

upvoted 4 times

 **Canary_2021** 2 years ago

The answers are correct.

<https://www.youtube.com/watch?v=cLf3nAiGG3Q>

upvoted 2 times

Topic 8 - Testlet 4

Introductory Info

Case study -

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case.

However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview -

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest is integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment -

Transactional Data -

Contoso has three years of customer, transactional, operational, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL Server instances contain data from various operational systems. The data is loaded into the instances by using SQL Server Integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time periods. Sales transaction data that is older than three years will be removed monthly. You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 200 GB.

Streaming Twitter Data -

The ecommerce department at Contoso develops an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

Planned Changes -

Contoso plans to implement the following changes:

Load the sales transaction dataset to Azure Synapse Analytics.

Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.

Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.

Sales Transaction Dataset Requirements

Contoso identifies the following requirements for the sales transaction dataset:

Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Customer Sentiment Analytics Requirements

Contoso identifies the following requirements for customer sentiment analytics:

Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds.

Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.

Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

Ensure that the data store supports Azure AD-based access control down to the object level.

Minimize administrative effort to maintain the Twitter feed data records.

Purge Twitter feed data records that are older than two years.

Data Integration Requirements -

Contoso identifies the following requirements for data integration:

Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synapse

Analytics and transform the data.

Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Question

HOTSPOT -

You need to design a data ingestion and storage solution for the Twitter feeds. The solution must meet the customer sentiment analytics requirements.

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

To increase the throughput of ingesting the Twitter feeds:

- Configure Event Hubs partitions.
- Enable Auto-Inflate in Event Hubs.
- Use Event Hubs Dedicated.

To store the Twitter feed data, use:

- An Azure Data Lake Storage Gen2 account
- An Azure Databricks high concurrency cluster
- An Azure General-purpose v2 storage account in the Premium tier

Correct Answer:

Answer Area

To increase the throughput of ingesting the Twitter feeds:

| |
|------------------------------------|
| Configure Event Hubs partitions. |
| Enable Auto-Inflate in Event Hubs. |
| Use Event Hubs Dedicated. |

To store the Twitter feed data, use:

| |
|---|
| An Azure Data Lake Storage Gen2 account |
| An Azure Databricks high concurrency cluster |
| An Azure General-purpose v2 storage account in the Premium tier |

Box 1: Configure Event Hubs partitions

Scenario: Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Event Hubs is designed to help with processing of large volumes of events. Event Hubs throughput is scaled by using partitions and throughput-unit allocations.

Incorrect Answers:

⊖ Event Hubs Dedicated: Event Hubs clusters offer single-tenant deployments for customers with the most demanding streaming needs.

This single-tenant offering has a guaranteed 99.99% SLA and is available only on our Dedicated pricing tier.

⊖ Auto-Inflate: The Auto-inflate feature of Event Hubs automatically scales up by increasing the number of TUs, to meet usage needs.

Event Hubs traffic is controlled by TUs (standard tier). Auto-inflate enables you to start small with the minimum required TUs you choose.

The feature then scales automatically to the maximum limit of TUs you need, depending on the increase in your traffic.

Box 2: An Azure Data Lake Storage Gen2 account

Scenario: Ensure that the data store supports Azure AD-based access control down to the object level.

Azure Data Lake Storage Gen2 implements an access control model that supports both Azure role-based access control (Azure RBAC) and POSIX-like access control lists (ACLs).

Incorrect Answers:

⊖ Azure Databricks: An Azure administrator with the proper permissions can configure Azure Active Directory conditional access to control where and when users are permitted to sign in to Azure Databricks.

⊖ Azure Storage supports using Azure Active Directory (Azure AD) to authorize requests to blob data.

You can scope access to Azure blob resources at the following levels, beginning with the narrowest scope:

- An individual container. At this scope, a role assignment applies to all of the blobs in the container, as well as container properties and metadata.
- The storage account. At this scope, a role assignment applies to all containers and their blobs.
- The resource group. At this scope, a role assignment applies to all of the containers in all of the storage accounts in the resource group.
- The subscription. At this scope, a role assignment applies to all of the containers in all of the storage accounts in all of the resource groups in the subscription.
- A management group.

Reference:

<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features> <https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

✉ noobprogrammer 1 year, 8 months ago

Answer looks correct to me:

1) Configure Event Hubs partition - The description says: "Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units."

2) An Azure Data Lake Storage Gen2 account.

Databricks cluster has nothing to do with storage, and a Data lake fits the needs

upvoted 14 times

✉ Deeksha1234 1 year, 4 months ago

correct!

Topic 9 - Testlet 5

✉ kkk5566 4 months, 1 week ago

correct

upvoted 1 times

Introductory Info

Case study -

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case.

However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview -

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements -

Business Goals -

Litware wants to create a new analytics environment in Azure to meet the following requirements:

See inventory levels across the stores. Data must be updated as close to real time as possible.

Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.

Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements -

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals.

Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.

Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

Use Azure Active Directory (Azure AD) authentication whenever possible.

Use the principle of least privilege when designing security.

Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data

Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment -

Litware plans to implement the following environment:

The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure. Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Daily inventory data comes from a Microsoft SQL server located on a private network.

Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.

Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.

Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

Question

What should you recommend to prevent users outside the Litware on-premises network from accessing the analytical data store?

- A. a server-level virtual network rule
- B. a database-level virtual network rule
- C. a server-level firewall IP rule
- D. a database-level firewall IP rule

Correct Answer: C

Scenario:

- ⇒ Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.
- ⇒ Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

Since Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure, they will have to create firewall IP rules to allow connection from the IP ranges of the on-premise network. They can also use the firewall rule 0.0.0.0 to allow access from Azure services.

Reference:

<https://docs.microsoft.com/en-us/azure/sql-database/sql-database-vnet-service-endpoint-rule-overview>

✉ **Kyle1** Highly Voted 2 years, 3 months ago

I think it should be C. The company doesn't want any virtual network stuff and server-level is more comprehensive, thus safer than just database-level rule.

upvoted 35 times

✉ **Marcus1612** Highly Voted 2 years, 3 months ago

The answer is C. Since there is no VPN between on-premises machines and Azure SQL server, communications use a public endpoint. You can limit the public access to databases through a Server Level IP Firewall rules. <https://docs.microsoft.com/en-us/azure/sql-database/network-access-controls-overview>

upvoted 12 times

✉ **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: C

is correct

upvoted 1 times

✉ **vctrhugo** 6 months, 3 weeks ago

Selected Answer: C

Option C, a server-level firewall IP rule, is the correct choice in this scenario. It provides a firewall rule at the server level, which means that any connections to the analytical data store will be filtered based on the specified IP addresses. By configuring the server-level firewall rule to only allow access from the IP addresses within the Litware on-premises network, you can effectively block access from users outside the network.

Options A and B, virtual network rules, are not the most appropriate choices in this case because they apply to the network level rather than the server or database level. Virtual network rules are used to control access to the Azure SQL server or database from specific virtual networks or subnets.

upvoted 2 times

✉ **AzureJobsTillRetire** 1 year, 1 month ago

Selected Answer: C

Actually, my preferred option is "database-level firewall IP rules for each and every database", but that option is not there, so I will have to choose C (a server-level firewall IP rule). Option D (a database-level firewall IP rule) is not sufficient, since we will have at least two databases, including Master and the Data Store database, to protect.

"We recommend that you use database-level IP firewall rules whenever possible. This practice enhances security and makes your database more portable. Use server-level IP firewall rules for administrators. Also use them when you have many databases that have the same access requirements, and you don't want to configure each database individually."

<https://learn.microsoft.com/en-us/azure/azure-sql/database/firewall-configure?view=azuresql>

upvoted 4 times

✉ **Deeksha1234** 1 year, 4 months ago

Selected Answer: C

correct
upvoted 1 times

□ **StudentFromAus** 1 year, 6 months ago

Selected Answer: C

Answer is correct
upvoted 1 times

□ **MvanG** 1 year, 7 months ago

Synapse Analytics has built in firewall. That combined with "least privileged, answer should be D. a Database -level firewall IP rule.
upvoted 1 times

□ **parx** 1 year, 9 months ago

Option A seems correct. Virtual Network (VNET) not to be confused with VPN. When setting up a IP rule at VNET, any resource within this VNET will be accessible only to that IP address. In this case On-Premises IP.

upvoted 1 times

□ **sdokmak** 1 year, 7 months ago

This is a tough one! Their on-prem data is in a private network, so in hindsight we should vNet peering to azure because server-level ip is not enough, yet... vNet is not enough for On-Prem either and would VPN/Express Routes involved. Best answer is C.
upvoted 1 times

□ **dev2dev** 1 year, 11 months ago

Selected Answer: C

read the last line "Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure." so C is correct instead of A
upvoted 1 times

□ **PallaviPatel** 1 year, 11 months ago

Selected Answer: C

C looks correct.
upvoted 1 times

□ **SabaJamal2010AtGmail** 2 years ago

A server-level firewall IP rule is correct
upvoted 1 times

□ **Canary_2021** 2 years ago

<https://docs.microsoft.com/en-us/azure/azure-sql/database/firewall-configure>
• Server-level IP firewall rules: These rules enable clients to access your entire server, that is, all the databases managed by the server.
• Database-level IP firewall rules: Database-level IP firewall rules enable clients to access certain (secure) databases. You create the rules for each database (including the master database), and they're stored in the individual database.
• We recommend that you use database-level IP firewall rules whenever possible.
So if target analytical data store is SQL data base in Azure, it is better to use database-level IP firewall rules.

For this question, the target analytical data store is Power BI, Ingestion data store is Data Lake Gen2. Not sure if this is the reason to select C?
upvoted 2 times

□ **vanrell** 1 year, 9 months ago

Remember, the question is on how to PREVENT outside users to gain access, combined with no VPN, answer should be C.
If the question was about GRANTING access, then following principle of least privilege would be as you mentioned a database level IP Firewall.
upvoted 3 times

□ **alexleonvalencia** 2 years ago

Selected Answer: C

Correcta
upvoted 1 times

□ **FredNo** 2 years, 1 month ago

Selected Answer: C

Answer is C because the company doesn't want to use any virtual network tools.
upvoted 3 times

□ **jefvaen** 2 years, 3 months ago

Answer should be D, I think.
"Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure"
upvoted 3 times

□ **YipingRuan** 2 years, 2 months ago

How to add rule at database level?
upvoted 1 times

□ **MarkJoh** 5 days, 15 hours ago

sp_set_database_firewall_rule
upvoted 1 times

Introductory Info

Case study -

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case.

However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview -

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements -

Business Goals -

Litware wants to create a new analytics environment in Azure to meet the following requirements:

See inventory levels across the stores. Data must be updated as close to real time as possible.

Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.

Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements -

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals.

Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.

Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

Use Azure Active Directory (Azure AD) authentication whenever possible.

Use the principle of least privilege when designing security.

Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data

Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment -

Litware plans to implement the following environment:

The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure. Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Daily inventory data comes from a Microsoft SQL server located on a private network.

Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.

Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.

Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

Question

What should you recommend using to secure sensitive customer contact information?

- A. Transparent Data Encryption (TDE)
- B. row-level security
- C. column-level security
- D. data sensitivity labels

Correct Answer: D

Scenario: Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Labeling: You can apply sensitivity-classification labels persistently to columns by using new metadata attributes that have been added to the SQL Server database engine. This metadata can then be used for advanced, sensitivity-based auditing and protection scenarios.

Incorrect Answers:

A: Transparent Data Encryption (TDE) encrypts SQL Server, Azure SQL Database, and Azure Synapse Analytics data files, known as encrypting data at rest. TDE does not provide encryption across communication channels.

Reference:

<https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview> <https://docs.microsoft.com/en-us/azure/sql-database/sql-database-security-overview>

✉  **echerish** Highly Voted 2 years, 4 months ago

Answer is C

<https://azure.microsoft.com/en-ca/updates/column-level-security-is-now-supported-in-azure-sql-data-warehouse/>

You can use CLS to manage user access to specific columns in your tables in a simpler manner, without having to redesign your data warehouse. CLS eliminates the need to maintain access restriction logic away from the data in another application or introduce views for filtering out sensitive columns for a subset of users.

upvoted 42 times

✉  **FredNo** Highly Voted 2 years, 1 month ago

Selected Answer: C

Answer is C

upvoted 8 times

✉  **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: C

is correct

upvoted 1 times

✉  **vctrhugo** 6 months, 3 weeks ago

Selected Answer: C

Option C, column-level security, allows you to control access to specific columns within a table based on user permissions. This means you can restrict access to sensitive customer contact information, such as phone numbers, only to authorized users or roles within the organization. By implementing column-level security, you can ensure that business analysts are limited in their access to customer contact information, preventing them from viewing or analyzing the phone numbers, which are not analytically relevant in this case.

upvoted 3 times

✉  **XiltroX** 1 year, 1 month ago

Answer is 100% C. The only way you can prevent users from seeing sensitive information is either through partial masking (partially visible) or complete blocking by using column level security.

upvoted 4 times

✉  **anks84** 1 year, 4 months ago

Selected Answer: C

Correct Answer is C i.e. Column-level security !!

upvoted 3 times

✉  **yyhhb** 1 year, 4 months ago

Selected Answer: D

The answer is D.

Between C and D, D can "minimize the number of different Azure services needed to achieve the business goals." But C needs to distribute role to the user, that is more complicated to apply.

upvoted 2 times

 **Deeksha1234** 1 year, 4 months ago

Selected Answer: C

C - column level security is correct

upvoted 2 times

 **dmgArtyco** 1 year, 5 months ago

La verdad que no queda nada claro

upvoted 1 times

 **Remedios79** 1 year, 6 months ago

Selected Answer: C

It's C because of this requirement : "Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant."

upvoted 3 times

 **Davico93** 1 year, 6 months ago

It's tricky, because it says "secure" and not "not to access"

upvoted 1 times

 **Arunava05** 1 year, 7 months ago

Even in udemy also the answer is same as ' Data sensitivity labels'

upvoted 2 times

 **AIcubeHead** 1 year, 9 months ago

Selected Answer: C

You don't secure data with sensitivity labels. They can only be used to identify who has accessed sensitive data. So it has to be column level security. There should really have been a Data Masking option here instead of column level security.

upvoted 4 times

 **coulia** 1 year, 11 months ago

It should be C, because only analysts are might not see those columns but others yes

upvoted 3 times

 **Remedios79** 1 year, 6 months ago

I agree with you!

upvoted 1 times

 **PallaviPatel** 1 year, 11 months ago

Selected Answer: C

column level security is correct answer

upvoted 2 times

 **Raghu108** 1 year, 11 months ago

Selected Answer: C

I fee it's C as we can limit access via CLS.

upvoted 1 times

 **datnguye** 2 years ago

"Limit access" confusing the selection between C and D.

In this case, I would choose C because it doesn't say to eliminate column(s) e.g. Phone from querying data

upvoted 2 times

 **datnguye** 2 years ago

Oh I mean to choose D (can't edit the comment)

upvoted 2 times

Topic 10 - Testlet 6

Introductory Info

Case study -

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case.

However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study -

To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.

Overview -

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements -

Business Goals -

Litware wants to create a new analytics environment in Azure to meet the following requirements:

See inventory levels across the stores. Data must be updated as close to real time as possible.

Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.

Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements -

Litware identifies the following technical requirements:

Minimize the number of different Azure services needed to achieve the business goals.

Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.

Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

Use Azure Active Directory (Azure AD) authentication whenever possible.

Use the principle of least privilege when designing security.

Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data

Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.

Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment -

Litware plans to implement the following environment:

The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure. Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

Daily inventory data comes from a Microsoft SQL server located on a private network.

Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.

Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.

Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

Question

What should you do to improve high availability of the real-time data processing solution?

- A. Deploy a High Concurrency Databricks cluster.
- B. Deploy an Azure Stream Analytics job and use an Azure Automation runbook to check the status of the job and to start the job if it stops.
- C. Set Data Lake Storage to use geo-redundant storage (GRS).
- D. Deploy identical Azure Stream Analytics jobs to paired regions in Azure.

Correct Answer: D

Guarantee Stream Analytics job reliability during service updates

Part of being a fully managed service is the capability to introduce new service functionality and improvements at a rapid pace. As a result, Stream Analytics can have a service update deploy on a weekly (or more frequent) basis. No matter how much testing is done there is still a risk that an existing, running job may break due to the introduction of a bug. If you are running mission critical jobs, these risks need to be avoided. You can reduce this risk by following Azure's paired region model.

Scenario: The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure

Reference:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-job-reliability>

✉  **petulda**  2 years, 4 months ago

There is a request 'Minimize number of Azure services'. With <https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-capture-overview> Event capture, data can be stored in DL without using Stream Analytics. In this case just Regional redundancy for DL would be needed.

upvoted 12 times

✉  **subhub** 3 months, 3 weeks ago

I passed today.. 820.. Thoughts.. 80% of exam questions were in Examtopics. Other 20% were not difficult. I got the Contoso case study... Good Luck.

upvoted 6 times

✉  **sachabess79** 2 years, 3 months ago

NB : it's an asynchronous copy.

upvoted 1 times

✉  **ian_viana** 2 years, 3 months ago

Agree, they also want a stage on data lake 2.

"Stage Inventory data in Azure Data Lake Storage Gen2"

we don't need Stream Analytics to do that. Event Hub enables you to automatically capture the streaming data in Event Hubs in an Azure Blob storage or Azure Data Lake Storage Gen 1 or Gen 2 account of your choice, with the added flexibility of specifying a time or size interval.

upvoted 1 times

✉  **ian_viana** 2 years, 3 months ago

Please desconsider my answer!

Event Hub can capture data to Data Lake and Blob. But I think the key word in the question is: real-time data PROCESSING solution azure. Event hub is just for capture. Stream Analytics do the processing so I'm going with answer D

upvoted 9 times

✉  **Marcus1612** 2 years, 3 months ago

I agree, Regional redundancy will be great for data but the processing would be lost. We need a solution for High Availability for PROCESSING and DATA.

upvoted 8 times

✉  **GDJ2022** 1 year, 11 months ago

The question is asking "improve high availability of the real-time data processing solution" and not high availability of data. Hence the correct answer is D

upvoted 2 times

 **kkk5566** Most Recent 4 months, 1 week ago

Selected Answer: D

should be D

upvoted 1 times

 **vctrhugo** 6 months, 3 weeks ago

Selected Answer: D

By deploying identical Azure Stream Analytics jobs to paired regions in Azure, you ensure redundancy and fault tolerance for the real-time data processing solution. Paired regions in Azure are geographically separated and designed to provide resilience and data protection in the event of a regional outage or failure. If one region becomes unavailable, the other paired region can seamlessly take over the processing workload, ensuring continuous availability of the real-time data processing solution.

upvoted 1 times

 **Deeksha1234** 1 year, 4 months ago

Selected Answer: D

answer D is correct

upvoted 4 times

 **StudentFromAus** 1 year, 6 months ago

Selected Answer: D

Answer is correct

upvoted 3 times

 **GDJ2022** 1 year, 11 months ago

D is correct.

The question is asking "improve high availability of the real-time data processing solution" and not high availability of data. Hence the correct answer is

upvoted 3 times

 **PallaviPatel** 1 year, 11 months ago

Selected Answer: D

I go with D and info provided by Canary_2021 is correct.

upvoted 2 times

 **HaBroNounen** 2 years ago

guys, the correct answer is A. It says to limit the amount of different services to use. Databricks is being used as a analytical tool for the data scientist already, so it can also be used for processing jobs.

upvoted 2 times

 **Davico93** 1 year, 6 months ago

You are right, but HC doesn't improve one shot processing, this would work better with multiple users

upvoted 1 times

 **edba** 2 years ago

I think the answer is correct!

upvoted 2 times

 **Canary_2021** 2 years ago

The answer should be D if the real time data load solution to move data from Azure Data Lake Storage Gen2 to Data Lake Gen2 to Azure SQL DB or Synapse Analytics as analytical data store.

If this way, Power BI and Azure Databricks notebooks will run query against Azure SQL DB or Synapse Analytics.

- Daily inventory data comes from a Microsoft SQL server located on a private network.
- Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store.
- See inventory levels across the stores. Data must be updated as close to real time as possible.
- Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

upvoted 4 times

 **jx1982** 2 years ago

I think the answer C is correct, high availability of "the real-time data processing", not high availability of "the data storage"

upvoted 4 times

 **jx1982** 2 years ago

sorry, typo, right answer is D

upvoted 3 times

 **FredNo** 2 years, 1 month ago

What is the correct answer?

upvoted 2 times

