

On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds.  
upvoted 5 times

□ **certstowinirl** 1 year, 7 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 4 times

□ **brendy** 1 year, 9 months ago

Is this correct?

upvoted 2 times

□ **Sudheer\_K** 1 year, 8 months ago

Not sure, what about the cost factor and premium doesn't minimise cost.

upvoted 1 times

□ **husseyn** 2 years ago

Concurrent Jobs should be raised - There is less cpu utilization

upvoted 2 times

□ **husseyn** 2 years ago

please ignore this, it was meant for the question before

upvoted 9 times

Question #37

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?

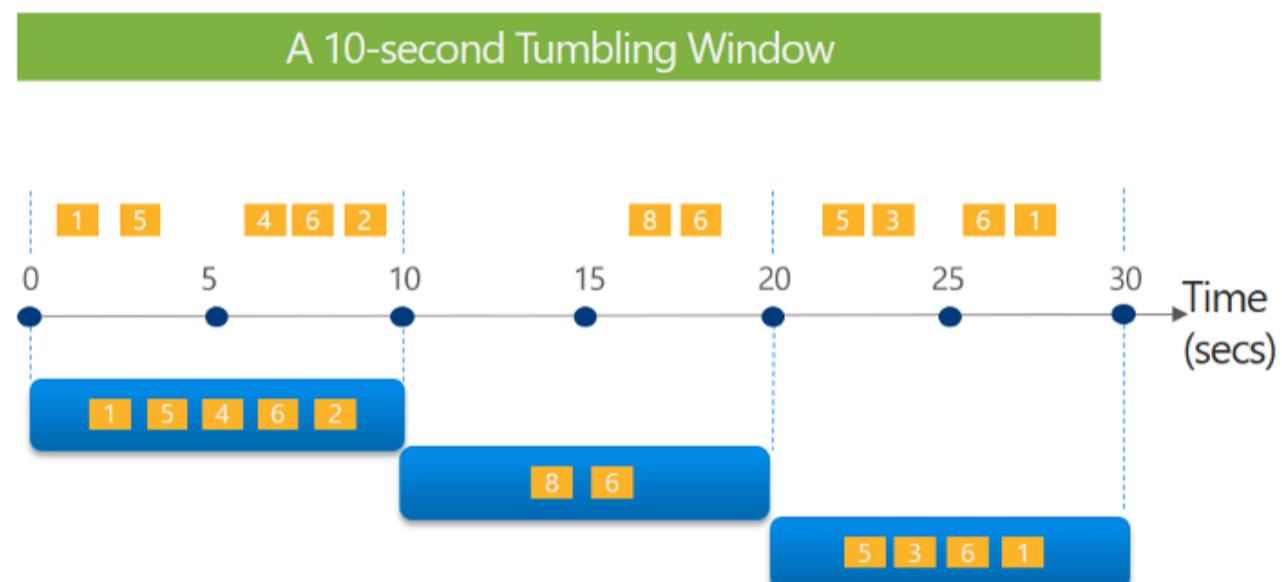
A. Yes

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 ago

correct answer

upvoted 37 times

□ **Deeksha1234** Most Recent 10 months, 1 week ago

correct

upvoted 1 times

□ **practicewizards** 11 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** 8 months ago

Both are correct. A Hopping window with hop-size = window-size is identical to a Tumbling window.

upvoted 4 times

 **sarapaisley** 1 year, 2 months ago

淘宝店铺 : <https://shop63989109.taobao.com/>

**Selected Answer: A**

correct

upvoted 2 times

 **agar** 1 year, 3 months ago

correct "D cholo

upvoted 1 times

 **anto69** 1 year, 4 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** 1 year, 5 months ago

A is correct

upvoted 1 times

 **lukeonline** 1 year, 5 months ago

**Selected Answer: A**

correct

upvoted 1 times

 **Canary\_2021** 1 year, 5 months ago

**Selected Answer: A**

A is Correct Answer

upvoted 1 times

 **rashjan** 1 year, 6 months ago

**Selected Answer: A**

correct

upvoted 1 times

 **paoloscott** 1 year, 6 months ago

Correct answer !

upvoted 1 times

 **AnandEMani** 1 year, 8 months ago

correct

upvoted 1 times

 **hugoborda** 1 year, 8 months ago

Answer is correct

upvoted 1 times

 **damaldon** 1 year, 11 months ago

Fully agree

upvoted 2 times

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

## Question #38

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 ago

answer correct

upvoted 22 times

 **MoDar** Highly Voted 1 year, 9 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 9 times

 **mamahani** Most Recent 4 weeks, 1 day ago

Selected Answer: B

B is correct answer

upvoted 1 times

 **MScapris** 5 months, 1 week ago

Selected Answer: B

NO is correct

upvoted 1 times

 **Deeksha1234** 10 months, 1 week ago

B is correct

upvoted 2 times

 **sarapaisley** 1 year, 2 months ago

Selected Answer: B

session window wouldn't count periods with no tweets

upvoted 3 times

 **Sandip4u** 1 year, 5 months 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** 1 year, 5 months ago

B - it has to be tumbling window

upvoted 1 times

 **rashjan** 1 year, 6 months ago

Selected Answer: B

correct: no

upvoted 1 times

 **dragos\_dragos62000** 1 year, 11 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 1 year, 10 months ago

淘宝店铺：<https://shop63989109.taobao.com/>

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 1 year, 10 months ago

I Disagree. The session could be extended if the maximum duration is set longer than the timeout.  
upvoted 4 times

□  EddyRoboto 1 year, 11 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?

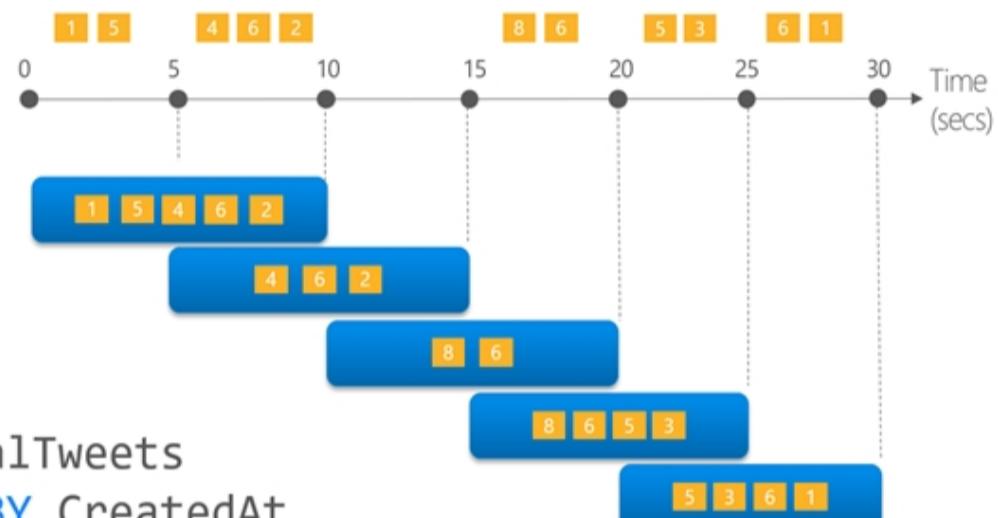
- A. Session
- B. Tumbling
- C. Sliding
- D. Hopping

**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 1 year, 6 months ago

Respuesta Correcta: Hopping

upvoted 13 times

 **Rossana** Most Recent 1 month, 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** 10 months, 1 week ago

Selected Answer: D

Hopping is right

upvoted 3 times

 **StudentFromAus** 11 months, 4 weeks ago

Why shouldn't it be sliding?

upvoted 1 times

淘宝店铺 : <https://shop63989109.taobao.com/>

□ **C1995** 1 year, 1 month ago

Why is sliding not correct?

upvoted 1 times

□ **Davico93** 11 months, 3 weeks ago

I want to know too

upvoted 1 times

□ **[Removed]** 7 months, 1 week 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 1 times

□ **SebK** 1 year, 2 months ago

**Selected Answer: D**

Correct

upvoted 2 times

□ **wwdba** 1 year, 3 months ago

Hopping is correct!

upvoted 1 times

□ **metallicjade** 1 year, 3 months ago

**Selected Answer: D**

hopping window

upvoted 1 times

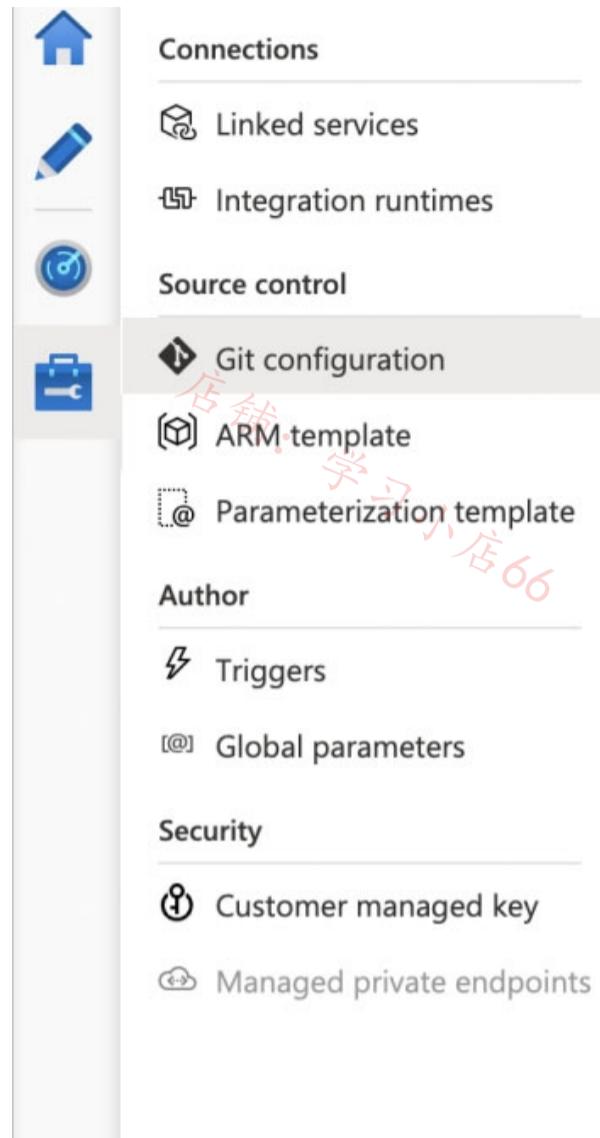
□ **Teraflow** 1 year, 5 months ago

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 12 months ago

Correct Answer i test it in devops

upvoted 23 times

✉  **VyshakhUnnikrishnan** Highly Voted 1 year, 3 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, 2 months ago

could you please paste the source of that? I can't find it

upvoted 2 times

✉  **mamahani** Most Recent 4 weeks, 1 day ago

Adf\_publish

/Dwh\_batchetl/adf\_publish/contososales

upvoted 1 times

✉  **Deeksha1234** 10 months, 1 week ago

correct

upvoted 1 times

✉  **SameerL** 11 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, 2 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, 3 months ago

Can someone confirm the correct answer? Is it:

1. adf\_publish

2. /.

Please let me know

upvoted 3 times

✉  **ANath** 1 year, 4 months ago

The answers are correct.

upvoted 2 times

✉  **Rohan21** 1 year, 4 months ago

Second answer should be contososales

upvoted 2 times

淘宝店铺：<https://shop63989109.taobao.com/>

□ **varmal** 1 year, 5 months 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, 4 months ago

I would say so as well

upvoted 3 times

□ **Davico93** 11 months, 2 weeks ago

You are thinking in an agnostic way and this is Azure DevOps

upvoted 2 times

□ **anto69** 1 year, 4 months ago

Yeah, totally agree. Nobody uses this notation

upvoted 3 times

□ **VeroDon** 1 year, 5 months ago

correct

upvoted 2 times

□ **Skeinofi** 1 year, 5 months 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 ~~Timezone~~: **学习小店66** 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	▼

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)

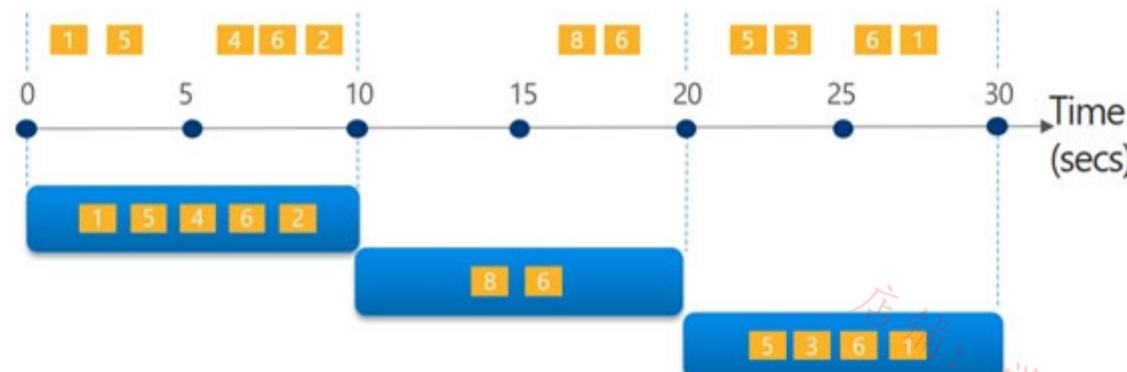
Box 1: timestamp by -

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, 4 months ago

The answers are correct  
upvoted 20 times

□ **seba020** Most Recent 2 months, 2 weeks ago

reason for "TIMESTAMPBY":  
<https://learn.microsoft.com/en-us/stream-analytics-query/timestamp-by-azure-stream-analytics>  
upvoted 3 times

□ **Dhaval\_Azure** 2 months, 3 weeks 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

□ **Deeksha1234** 10 months, 1 week ago

correct answer  
upvoted 2 times

□ **May99** 1 year, 4 months ago

I think it's system.timestamp()  
upvoted 3 times

□ **sdokmak** 1 year ago

From examples, I can only see system.timestamp() used after SELECT, not FROM.  
upvoted 1 times

□ **jv2120** 1 year, 5 months ago

It only says about window size, not sure why tumbling window not hopping.  
upvoted 2 times

□ **Andreas\_K** 1 year, 5 months 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 20 times

□ **TestMitch** 1 year, 5 months ago

Correcto  
upvoted 2 times

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

**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:

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	▼

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] 1 year, 5 months 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 67 times

Canary\_2021 [Highly Voted] 1 year, 5 months 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 38 times

Rossana [Most Recent] 1 month, 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 3 months, 4 weeks 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 3 times

DAYENKAR 4 months, 3 weeks ago

Both answer are polybase

upvoted 1 times

XiltroX 6 months, 1 week 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 10 months, 1 week ago

Agree with marcin1212

it should be

P1: Set the partition option to "Dynamic range "

P2: PolyBase

upvoted 4 times

 **NamitSehgal** 11 months, 3 weeks ago

淘宝店铺：<https://shop63989109.taobao.com/>

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 1 times

 **Towin** 1 year 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

 **AIcubeHead** 1 year, 2 months ago

Answer is completely wrong. If writing to a non-partitioned table in a dedicated SQL Pool you ALWAYS want to choose Polybase whenever possible. So the answers are:

P1: "Dynamic Range"

P2: "Polybase"

upvoted 3 times

 **AIcubeHead** 1 year, 2 months ago

P1: Set the partition option to "Dynamic range "

P2: PolyBase - PolyBase is significantly faster than BulkInsert

upvoted 3 times

 **alex1491** 1 year, 2 months ago

From Synapse to Data lake it's not even option bulk insert or polybase. The only way to use those option are from Data lake to Synapse

P1: Set the partition option to "Dynamic range "

P2: Polybase

upvoted 2 times

 **Oldrich22** 1 year, 3 months ago

P1: Set the partition option to "Dynamic range "

P2: PolyBase

upvoted 3 times

 **ItHYMeRish** 1 year, 5 months ago

I believe both answers are PolyBase. PolyBase supports both export to and import from ADLS as documented here:

<https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-versioned-feature-summary>

PolyBase does support delimited text files, which contradicts the question's official answer. "Currently PolyBase can load data from UTF-8 and UTF-16 encoded delimited text files as well as the popular Hadoop file formats RC File, ORC, and Parquet (non-nested format)."

<https://techcommunity.microsoft.com/t5/datacat/azure-sql-data-warehouse-loading-patterns-and-strategies/ba-p/305456#:~:text=Currently%20PolyBase%20can%20load%20data%20from%20UTF-8%20and,data%20from%20gzip%2C%20zlib%20and%20Snappy%20compressed%20files.>

upvoted 12 times

 **Andreas\_K** 1 year, 5 months ago

Right answer should be PolyBase in both cases. It provides the highest performance and supports delimited text files.

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/load-data-overview>

<https://docs.microsoft.com/en-us/azure/data-factory/load-azure-sql-data-warehouse?tabs=data-factory>

upvoted 3 times

 **marcin1212** 1 year, 5 months ago

@Andreas

how to use PolyBase when copy data from Synapse to file ?

upvoted 4 times

 **temacc** 5 months ago

PolyBase uses

PolyBase enables the following scenarios in SQL Server:

Query data stored in Azure Blob Storage. Azure Blob Storage is a convenient place to store data for use by Azure services. PolyBase makes it easy to access the data by using T-SQL.

Query data stored in Hadoop from a SQL Server instance or PDW. Users are storing data in cost-effective distributed and scalable systems, such as Hadoop. PolyBase makes it easy to query the data by using T-SQL.

Import data from Hadoop, Azure Blob Storage, or Azure Data Lake Store. Leverage the speed of Microsoft SQL's columnstore technology

and analysis capabilities by importing data from Hadoop, Azure Blob Storage, or Azure Data Lake Store into relational tables. There is no need for a separate ETL or import tool.

Export data to Hadoop, Azure Blob Storage, or Azure Data Lake Store. Archive data to Hadoop, Azure Blob Storage, or Azure Data Lake Store to achieve cost-effective storage and keep it online for easy access.

Integrate with BI tools. Use PolyBase with Microsoft's business intelligence and analysis stack, or use any third-party tools that are compatible with SQL Server.

upvoted 2 times

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

**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:**

<input type="checkbox"/>	▼

**Trigger:**

<input type="checkbox"/>	▼

**Answer Area**

**Load methodology:**

<input type="checkbox"/>	▼

**Correct Answer:**

**Trigger:**

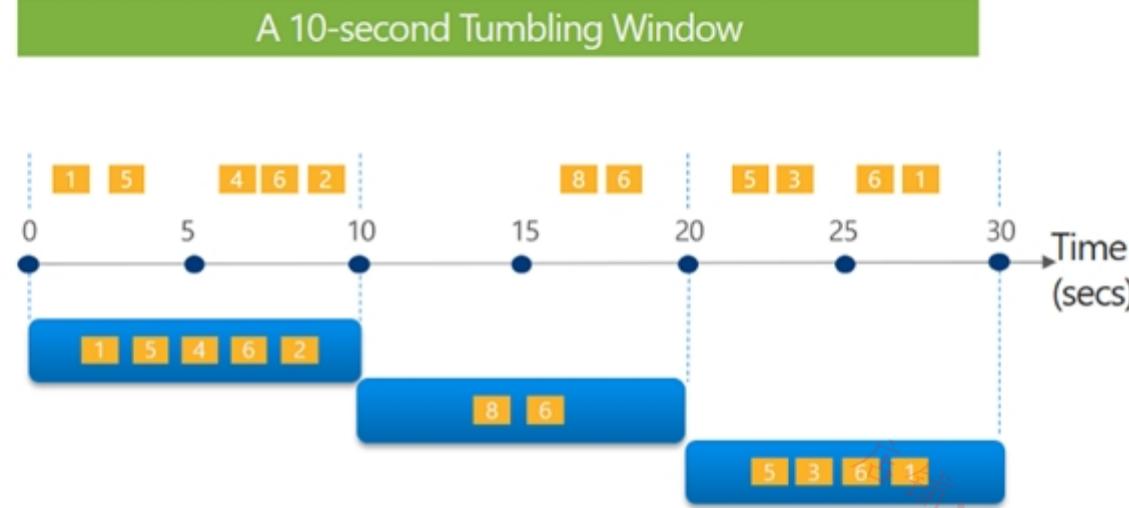
<input type="checkbox"/>	▼

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



```
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 1 year, 5 months 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 23 times

⊕ **xcsakubara** Highly Voted 1 year, 3 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 11 times

⊕ **vedantnj** Most Recent 1 month ago

Hi there

upvoted 1 times

⊕ **Rossana** 1 month, 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 2 times

⊕ **martcerv** 5 months, 2 weeks 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 5 times

⊕ **Deeksha1234** 10 months, 1 week ago

- 1) Incremental Load
- 2) Tumbling Window

upvoted 3 times

⊕ **jskibick** 1 year 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 9 times

□ **Massy** 1 year, 2 months ago

淘宝店铺：<https://shop63989109.taobao.com/>

both Tumbling Window and Schedule trigger will reach the goal. Which one is more cost effective?

upvoted 1 times

□ **Boompiee** 1 year 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 1 times

□ **xcsakubara** 1 year, 3 months ago

why not schedule trigger?

upvoted 1 times

□ **sparkchu** 1 year, 2 months ago

for backfill purpose? just guessing.

upvoted 1 times

□ **jv2120** 1 year, 5 months ago

incremental, fixed schedule every hour.

upvoted 5 times

□ **jv2120** 1 year, 5 months ago

correct answer..tumbling window

upvoted 1 times

□ **Ayan3B** 1 year, 6 months ago

As a input we are receiving csv files so why not trigger mechanism to the pipeline when file arrived.

upvoted 2 times

□ **ItHYMeRish** 1 year, 5 months 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 32 times

□ **ANath** 1 year, 4 months ago

That's correct. Well explained

upvoted 1 times

## Question #44

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 (94%)	6%
---------	----

✉  lukeonline Highly Voted 1 year, 5 months 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 26 times

✉  kamil\_k 1 year, 2 months ago

or rather Scala does not support concurrent instances (but yes, it implies HC cluster will not support Scala)  
upvoted 2 times

✉  Rossana Most Recent 1 month, 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 6 months ago

<https://docs.databricks.com/clusters/configure.html>

upvoted 1 times

✉  Igor85 6 months, 2 weeks ago

high concurrency cluster is already a legacy cluster mode. question is not relevant anymore

upvoted 2 times

淘宝店铺：<https://shop63989109.taobao.com/>

□ **greenlever** 7 months, 4 weeks 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** 8 months, 3 weeks ago

NO IS CORRECT ANSWER

upvoted 1 times

□ **Deeksha1234** 10 months, 1 week ago

correct , answer B, agree with lukeonline

upvoted 1 times

□ **mkthoma3** 11 months, 3 weeks ago

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>

upvoted 1 times

□ **Hanse** 1 year, 3 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** 1 year, 5 months ago

B is correct

upvoted 2 times

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

**Selected Answer: B**

Respuesta correcta; Standar para Cientificos y jobs. Alta concurrencia para ingenieros de datos.

upvoted 3 times

□ **Sanand** 1 year, 5 months ago

Agree! - Correct answer; Standard for Scientists and jobs. High concurrency for data engineers.

upvoted 1 times

## Question #45

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?

- 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 1 year, 5 months 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 53 times

 **lukeonline** 1 year, 5 months ago

Lol, I searched in the internet for the "patient pipeline".... should have read the comments first :)

upvoted 17 times

 **Remedios79** 11 months, 3 weeks ago

Thank you. I was wondering about "patient" and related it on my poor english!

upvoted 2 times

 **dsp17** 11 months ago

Big thanks onyerleft :)

upvoted 2 times

 **vigilante89** Most Recent 5 months, 3 weeks 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 1 times

 **Deeksha1234** 10 months, 1 week ago

right C

upvoted 1 times

 **DrTaz** 1 year, 5 months ago

what the hk is a patient pipeline?

upvoted 2 times

 **anto69** 1 year, 4 months ago

lol I think they mean "parent"  
upvoted 1 times

淘宝店铺：<https://shop63989109.taobao.com/>

□  **jv2120** 1 year, 5 months ago

It should be tumbling window since 2 dependent pipelines run state. from given option only schedule event fits but its not correct.  
upvoted 3 times

□  **AzureJobsTillRetire** 6 months, 2 weeks 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** 1 year, 5 months ago

Shouldn't the answer be A/D?  
upvoted 1 times

□  **TashaP** 9 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** 1 year, 5 months ago

I think it should be Tumbling window  
upvoted 2 times

□  **corebit** 1 year, 5 months ago

The question or answers do not mention Tumbling Window. What is the basis for the response? Any more context?  
upvoted 2 times

Question #46

## 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	Answer Area
a database scoped credential	 
an asymmetric key	
an external data source	
a database encryption key	
an external file format	

Correct Answer:

Components	Answer Area
 	an asymmetric key
	a database scoped credential
	an external data source
	a database encryption key
	an external file format

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>

✉  **aleleonvalencia** Highly Voted  1 year, 6 months ago

- 1.- A database scoped credential
- 2.- an External data sorce
- 3.- a external file format

upvoted 146 times

✉  **DiscussoR** 2 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 1 times

淘宝店铺：<https://shop63989109.taobao.com/>

□ **Franz58** 10 months, 2 weeks 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

□ **DiscussoR** 2 months, 2 weeks ago

File format is not related to a specific file

upvoted 1 times

□ **Bilal2** 5 months ago

agreed.

<https://www.sqlshack.com/sql-server-polybase-external-tables-with-azure-blob-storage/>

upvoted 1 times

□ **engrbrain** Highly Voted 1 year, 5 months 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 48 times

□ **kamil\_k** 1 year, 2 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, 2 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

□ **vanrell** 1 year, 2 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 3 times

□ **sdokmak** 1 year ago

access key is for storage account so you still need a master/asymmetric key for the database.

upvoted 3 times

□ **sdokmak** 1 year ago

\*sorry, not asymmetric  
upvoted 3 times

□ **rocky48** Most Recent 1 week, 4 days ago

- 1.- A database scoped credential
- 2.- an External data source
- 3.- a external file format

upvoted 1 times

□ **Rossana** 1 month, 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** 1 month, 3 weeks ago

- 1.A database scoped credential
- 2.an External data source
- 3 a external file format

upvoted 2 times

□ **DiscussoR** 2 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 1 times

□ **esaade** 2 months, 3 weeks ago

o 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** 3 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 (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

□ **SophieM** 4 months, 3 weeks ago

The database master key is a SYMMETRIC key that is used to protect the private keys of certificates and asymmetric keys that are present in the database.

We should start with Database Encryption Key, which is a Master Key. Not the assymetric key, as this can not be a master key. Agreed with engrbrain:

1. Database Encryption Key
2. Database Scoped Credential
3. External Data Source

upvoted 2 times

□ **OldSchool** 6 months, 2 weeks ago

-- Create a database master key if one does not already exist, using your own password. This key is used to encrypt the credential secret in next step.

```
CREATE MASTER KEY ENCRYPTION BY PASSWORD = '<password>';
```

-- Create a database scoped credential with Azure storage account key as the secret.

```
CREATE DATABASE SCOPED CREDENTIAL AzureStorageCredential
```

WITH

```
IDENTITY = '<my_account>',  
SECRET = '<azure_storage_account_key>';
```

-- Create an external data source with CREDENTIAL option.

```
CREATE EXTERNAL DATA SOURCE MyAzureStorage
```

WITH

```
( LOCATION = 'wasbs://daily@logs.blob.core.windows.net/' ,
```

```
CREDENTIAL = AzureStorageCredential ,
```

```
TYPE = HADOOP
```

```
) ;
```

upvoted 5 times

□ **pmc08** 8 months, 2 weeks ago

- 1.- An external data source
- 2.- An External File Format
- 3.- A database scoped credential

First you have to create a db master key and db scoped credential,

then you have to create an external data source and den you need to configure the external data format  
(answer given from the skillcertpro platform)

upvoted 2 times

□ **anks84** 9 months ago

- 1.Create a master key on the database. The master key is required to encrypt the credential secret.
  - 2.Create a database scoped credential for Azure blob storage; IDENTITY can be anything as it's not used.
  - 3.Create an external data source with CREATE EXTERNAL DATA SOURCE.
  - 4.Create an external file format with CREATE EXTERNAL FILE FORMAT.
  - 5.Create an external table pointing to data stored in Azure storage with CREATE EXTERNAL TABLE.
- Refer <https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-configure-azure-blob-storage?view=sql-server-ver15>

upvoted 2 times

□ **Deeksha1234** 10 months, 1 week ago

correct, agree with engrbrain

upvoted 2 times

□ **Deeksha1234** 10 months, 1 week ago

- 1- Create a Database Encryption Key
- 2 - Create a Database Scoped Credential
- 3 - Create an External Data Source

upvoted 3 times

淘宝店铺：<https://shop63989109.taobao.com/>

□ **Franz58** 10 months, 3 weeks ago

It requires access to the DW, not to a single file, therefore external file format is not needed here.  
1- Create a Database Encryption Key  
2 - Create a Database Scoped Credential  
3 - Create an External Data Source  
upvoted 3 times

□ **VM\_GCP** 11 months, 1 week ago

I think, the correct reference to this question is <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#c-create-external-data-source-to-reference-azure-data-lake-store-gen-2-using-the-storage-account-key>

1. Create encryption master key  
2. Create Database Scoped Credentials  
3. Create External Data Source.

I see some are ref to this doc <https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-configure-azure-blob-storage?view=sql-server-ver15>, but check, its only applicable to SQL Server, and not to Synapse.

upvoted 3 times

□ **ads5891** 10 months ago

I guess the reference should be <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#b-create-external-data-source-to-reference-azure-data-lake-store-gen-1-or-2-using-a-service-principal>  
upvoted 1 times

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

1.Create a master key on the database. The master key is required to encrypt the credential secret.  
2.Create a database scoped credential for Azure blob storage; IDENTITY can be anything as it's not used.  
3.Create an external data source with CREATE EXTERNAL DATA SOURCE.  
4.Create an external file format with CREATE EXTERNAL FILE FORMAT.  
5.Create an external table pointing to data stored in Azure storage with CREATE EXTERNAL TABLE.  
Refer <https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-configure-azure-blob-storage?view=sql-server-ver15>

upvoted 15 times

□ **ANath** 1 year, 5 months ago

Why not a asymmetric key at first? The following link says we should use one:  
<https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/sql-server-and-database-encryption-keys-database-engine?view=sql-server-ver15>

Can anyone clarify.

upvoted 2 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 1 year, 5 months ago

**Selected Answer: D**

The link provided is the source of truth  
upvoted 12 times

 **Teraflow** Highly Voted 1 year, 5 months ago

**Selected Answer: D**

D is correct  
upvoted 6 times

 **Deeksha1234** Most Recent 10 months, 1 week ago

correct  
upvoted 1 times

 **austin06112000** 1 year, 2 months ago

D is correct.  
upvoted 2 times

 **DrTaz** 1 year, 5 months ago

**Selected Answer: D**  
Correct. D is the one that makes most sense.  
upvoted 2 times

**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

<input type="checkbox"/> Collect(Score)
<input checked="" type="checkbox"/> CollectTop(1) OVER(ORDER BY Score Desc)
<input type="checkbox"/> Game, MAX(Score)
<input type="checkbox"/> TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)

as HighestScore

FROM input TIMESTAMP BY CreatedAt

GROUP BY

<input type="checkbox"/> Game
<input type="checkbox"/> Hopping(minute,5)
<input type="checkbox"/> Tumbling(minute,5)
<input type="checkbox"/> Windows(TumblingWindow(minute,5),Hopping(minute,5))

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

SELECT

<input type="checkbox"/> Collect(Score)
<input checked="" type="checkbox"/> CollectTop(1) OVER(ORDER BY Score Desc)
<input type="checkbox"/> Game, MAX(Score)
<input type="checkbox"/> TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)

as HighestScore

FROM input TIMESTAMP BY CreatedAt

GROUP BY

<input type="checkbox"/> Game
<input checked="" type="checkbox"/> Hopping(minute,5)
<input type="checkbox"/> Tumbling(minute,5)
<input type="checkbox"/> 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.

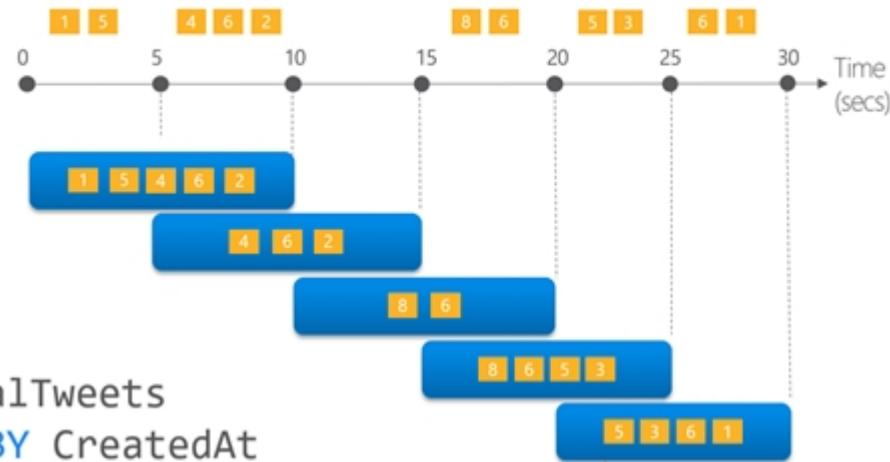
Ordering/ranking is based on event columns and can be specified in ORDER BY clause.

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/stream-analytics-window-functions>

✉ **alexleonvalencia** Highly Voted 1 year, 6 months ago

TopOne() / Tumbling  
upvoted 100 times

✉ **MarcelIT** 4 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 6 times

✉ **\_lene\_** 2 months ago

The question was "the highest scoring record of each game", so that's what we need - one row for each game  
upvoted 1 times

✉ **cr727** 4 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 1 times

✉ **gf2tw** Highly Voted 1 year, 6 months ago

Syntax for Hopping window requires 3 arguments, seems this should be Tumbling Window which fulfils the exact same requirements.  
upvoted 34 times

✉ **anto69** 1 year, 4 months ago

Yeah sure  
upvoted 2 times

✉ **akk\_1289** Most Recent 5 months 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 5 times

✉ **mroova** 3 months, 3 weeks 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 6 times

✉ **Achu24** 5 months, 2 weeks ago

Game(max) and tumbling window is the correct answer  
upvoted 7 times

✉ **mesloth** 4 months, 4 weeks ago

Correct. Here top score is being asked, instead of Rank  
upvoted 3 times

✉ **JosephVishal** 5 months, 4 weeks ago

Tumbling window seems to be correct, in question there is no fixed time interval specified.  
upvoted 1 times

✉ **THAYTRUONG** 6 months, 3 weeks ago

TopOne() / Tumbling is correct answer  
upvoted 2 times

用户: **bakstorage00001** 8 months, 3 weeks ago

淘宝店铺: <https://shop63989109.taobao.com/>

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

用户: **allagowf** 8 months ago

true, if tumbling is not the correct answer in the exam then we fu\*\*-up really fu\*\*-up hahaha  
upvoted 2 times

用户: **Deeksha1234** 10 months, 1 week ago

It should be TopOne() and Tumbling  
upvoted 1 times

用户: **jainparag1** 10 months, 3 weeks ago

It should be Tumbling window.  
upvoted 3 times

用户: **agar** 1 year, 3 months ago

it is Tumbling "D"  
upvoted 1 times

用户: **chxzqw** 1 year, 4 months ago

pls why not game, max(score) ?  
upvoted 7 times

用户: **svik** 1 year, 4 months ago

If max(score) is used then we have to have Game in the group by clause  
upvoted 9 times

用户: **assU2** 1 year, 4 months ago

pls can someone explain in details why not game, max(score) ?  
upvoted 3 times

用户: **adfgasd** 1 year, 4 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 8 times

用户: **stunner85\_** 1 year, 4 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 2 times

用户: **Teraflow** 1 year, 5 months ago

It should be tumbling window  
upvoted 3 times

用户: **ayush188** 1 year, 5 months ago

TUMBLING will be the answer folks  
upvoted 4 times

用户: **m2shines** 1 year, 5 months ago

Incorrect, answer shud be Tumbling  
upvoted 2 times

用户: **jxj770** 1 year, 5 months ago

Tumbling is right  
upvoted 4 times

## Question #49

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 (92%) 8%

✉ **NaiCob** Highly Voted 1 year, 5 months ago

Correct answer: No - you cannot execute the R script using a stored procedure activity  
upvoted 46 times

✉ **Daemon69** Highly Voted 1 year, 4 months ago

I select A because you can use R script in sp\_execute\_external\_script  
upvoted 12 times

✉ **Rossana** 1 month, 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 1 times

✉ **sparkchu** 1 year, 2 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

✉ **FRANCIS\_A\_M** Most Recent 1 month, 4 weeks 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 1 times

✉ **Kamekung** 3 months 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** 2 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** 3 months, 3 weeks ago

Cant we fix answers correctly in the portal, instead of relying on votes  
upvoted 2 times

✉ **mckovin** 4 months, 1 week ago

Correct  
upvoted 1 times

淘宝店铺：<https://shop63989109.taobao.com/>

 **millusmiley** 4 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** 4 months, 1 week ago

I think this is not possible we have to pay :(  
upvoted 2 times

 **Dusica** 4 months, 4 weeks 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 1 times

 **akk\_1289** 5 months 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** 10 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** 10 months, 1 week ago

should be B  
upvoted 2 times

 **nilubabu** 12 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 4 times

 **rafaelptu** 1 year, 2 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, 4 months ago

**Selected Answer: B**  
Should be NO  
upvoted 1 times

 **dev2dev** 1 year, 4 months ago

The need is to transform using R script and load into synapse. So answer is no.  
upvoted 1 times

 **Roewe** 1 year, 5 months ago

**Selected Answer: B**  
Should be no  
upvoted 3 times

 **engrbrain** 1 year, 5 months ago

The Answer is NO. Stored Procedure Activity cannot run R Script  
upvoted 3 times

## Question #50

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.azure.databricks.net/clusters/configure.html>

*Community vote distribution*

B (100%)

 **djinchege**  1 year, 9 months ago

data scientist need scala so standard, jobs need scala so standard, so B but for different reasons  
upvoted 43 times

 **Gina8008** 1 year, 4 months ago

engineer has to share the cluster so high -concurrency is correct. the answer should be A  
upvoted 2 times

 **Aditya0891** 1 year 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 5 times

 **111222333**  2 years ago

Correct is A  
upvoted 16 times

 **dfdsfdsfsd** 2 years ago

Agree. Jobs cannot use a high-concurrency cluster because it does not support Scala.  
upvoted 5 times

 **Aditya0891** 1 year 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 is No  
upvoted 5 times

 **Hanse**  1 year, 3 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 1 times

□ **avijitd** 1 year, 5 months ago

**Selected Answer: B**

NO - as High concurrency not support Scala  
upvoted 6 times

□ **rashjan** 1 year, 6 months ago

**Selected Answer: B**

correct: no  
upvoted 5 times

□ **arjunbhai** 1 year, 6 months ago

Like djinchev said, Data scientists need scala so B.

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure>  
upvoted 2 times

□ **Julius7000** 1 year, 8 months ago

-Data Engineers: Correct, they are working together, they need High-Concurrence cluster  
-Jobs: Correct, Standard 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 13 times

□ **Julius7000** 1 year, 8 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-concurrency 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** 1 year, 9 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** 1 year, 11 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** 1 year, 8 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 ago

A is the right answer because Standard cluster supports scala  
upvoted 2 times

□ **Wisent** 2 years 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

## Question #51

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 (70%)

C (30%)

□  **Canary\_2021** Highly Voted  1 year, 5 months 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 11 times

□  **HaBroNounen** 1 year, 5 months 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 13 times

□  **Reloadedvn** Most Recent  3 weeks, 5 days ago

**Selected Answer: B**

Agree to B

upvoted 1 times

□  **\_lene\_** 2 months ago

**Selected Answer: C**

The cluster does auto-scaling by default. Auto-termination should be set up manually

upvoted 2 times

□  **Deeksha1234** 10 months, 1 week ago

**Selected Answer: B**

B is correct

upvoted 1 times

□  **jz10** 1 year, 1 month 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 1 times

□ **Amsterliese** 1 year, 2 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 3 times

□ **sunithagsk** 1 year, 2 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 3 times

□ **alex1491** 1 year, 3 months ago

**Selected Answer: C**  
i try it and it's possible to create a cluster with auto termination.  
upvoted 2 times

□ **AngelJP** 1 year, 3 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, 3 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, 3 months ago

B is correct answer.  
High concurrency cluster cannot AUTO terminated  
upvoted 2 times

□ **ANath** 1 year, 5 months 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 4 times

□ **kamil\_k** 1 year, 2 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

□ **venkatibm** 1 year, 5 months 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**  1 year, 5 months 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 75 times

✉  **Bro111** 6 months, 1 week ago

Why not /{deviceType}/out/{YYYY}/{MM}/{DD}/{HH}.json ?

upvoted 2 times

✉  **sensaint** 6 months, 1 week ago

It is not an option. It says /{deviceID}/out/{YYYY}/{MM}/{DD}/{HH}.json

upvoted 4 times

✉  **onyerleft**  1 year, 5 months 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 67 times

□ **Davico93** 11 months, 2 weeks ago

but, the solution must minimize read times, I think is @trigger().startTime

upvoted 2 times

□ **rocky48** [Most Recent] 1 week, 4 days ago

- 1) @trigger().outputs.windowStartTime
- 2) /{YYYY}/{MM}/{DD}/{HH}\_{deviceType}.json
- 3) Merge

upvoted 3 times

□ **rzeng** 7 months, 2 weeks ago

1. windowstarttime
2. yyyy/mm/dd/hh\_devicetype.json
3. Merge

upvoted 5 times

□ **Deeksha1234** 10 months, 1 week ago

- 1) @trigger().outputs.windowStartTime
- 2) /{YYYY}/{MM}/{DD}/{HH}\_{deviceType}.json
- 3) Merge

agree with onyer

upvoted 2 times

□ **Rafafouille76** 1 year, 3 months ago

Of course it is a merge, can't believe the official provided answers are so wrong ... Who wrote that

upvoted 5 times

□ **kamil\_k** 1 year, 2 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 2 times

□ **Jaws1990** 1 year, 4 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** 1 year, 5 months 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 2 times

□ **jv2120** 1 year, 5 months ago

data is generated every 5 min but output needs every 1 hour/device it, it needs to merge files to achieve this.

upvoted 1 times

□ **tony4fit** 1 year, 5 months ago

The answers are correct. Flatten Hierarchy. <https://vmfocus.com/2019/01/09/using-azure-data-factory-to-copy-data-between-azure-file-shares-part-1/>

upvoted 2 times

□ **Aditya0891** 1 year ago

think logically what flatten and merge means and what is asked in the question

upvoted 2 times

Question #53

## 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>

 **ItHYMeRish**  1 year, 5 months 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 169 times

 **nmmn22** 2 months ago

thats such a cool explanation, i aspire to have the same critical thinking skills u have

upvoted 1 times

 **DataEX** 4 months ago

The correct structure answer will have 561.600 folders per year.

upvoted 1 times

□ **ML\_Novice** 9 months, 1 week ago

ItHYMeRIsh you re a genius man  
upvoted 3 times

□ **Deeksha1234** 10 months, 1 week ago

Agree, correct answer  
upvoted 1 times

□ **gf2tw** Highly Voted 1 year, 6 months 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 13 times

□ **SAli12** 1 year, 5 months ago

Yes I agree, raw/regionId --> timestamp --> deviceId.json  
upvoted 3 times

□ **rocky48** Most Recent 1 week, 4 days ago

The correct answer is  
{raw/regionID}/{YYYY}/{MM}/{DD}/{HH}/{mm}/{deviceID}.json  
upvoted 1 times

□ **georgich87** 1 year, 2 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 2 times

□ **wwdba** 1 year, 2 months ago

{raw/regionID}/{YYYY}/{MM}/{DD}/{HH}/{mm}/{deviceID}.json  
upvoted 1 times

□ **staniopolis** 1 year, 3 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 2 times

□ **staniopolis** 1 year, 3 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, 4 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, 4 months ago

Still feel {raw/RegionID} / {YYYY/MM/DD/mm} /{DeviceID} is correct. Just have some questions after compare answers of question 54.  
upvoted 1 times

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

/{SubjectArea}/{DataSource}/{YYYY}/{MM}/{DD}/{FileData}\_{YYYY}\_{MM}\_{DD}.  
upvoted 2 times

□ **PA7** 1 year, 5 months ago

raw/regionid -> DeviceId -> YYYY/MM/dd/HH-mm  
upvoted 4 times

□ **mr\_corte** 1 year, 5 months ago

{raw/regionID}/{deviceID}/{YYYY}/{MM}/{DD}/{HH}{mm} imo.  
upvoted 4 times

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

**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
FROM input TIMESTAMP BY EventTime

```

**Correct Answer:**

```

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'

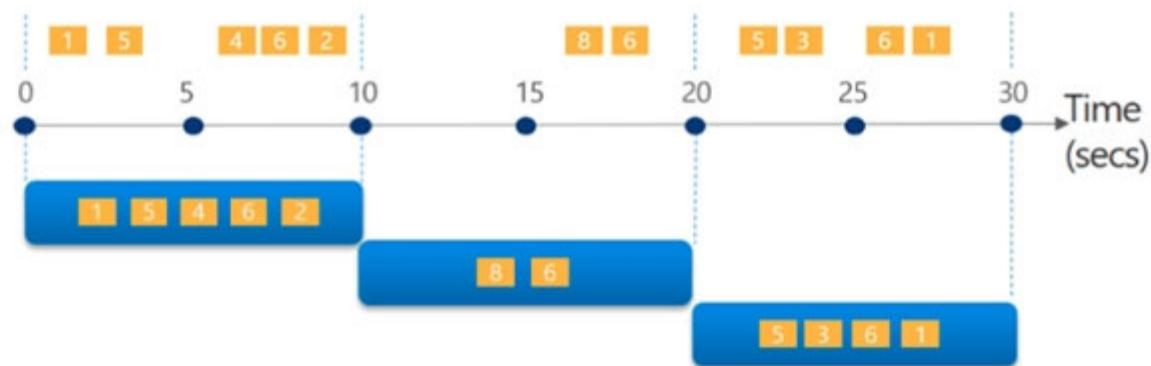
Box 2: ,TumblingWindow(Second, 5)

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 1 year, 5 months 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 66 times

□ **yogiazaad** 3 months, 4 weeks ago

淘宝店铺 : <https://shop63989109.taobao.com/>

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, 3 months ago

I concur!

upvoted 1 times

□ **onyerleft** 1 year, 5 months ago

Yes this sounds right

upvoted 2 times

□ **Davico93** 11 months, 3 weeks 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** 11 months, 3 weeks ago

Sorry, you are right @Fer079!

upvoted 2 times

□ **Canary\_2021** [Highly Voted] 1 year, 5 months 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 20 times

□ **Fer079** 1 year, 4 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** 1 month, 2 weeks ago

you do not need partition by with LAG function; its 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 its 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 dont have these records anymore; we need to have 'startting' record for correctly operating device with heartbeat event , and this for every singe "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, 1 month ago

LAG does not require the PARTITION BY this is optional..

upvoted 2 times

□ **rocky48** [Most Recent] 1 week, 4 days ago

1. Where EventType = 'HeartBeat'

2. SessionWindow

upvoted 1 times

□ **SinSS** 1 month, 1 week ago

1. Where EventType = 'HeartBeat'

2. SessionWindow

upvoted 1 times

□ **dom271219** 9 months, 2 weeks ago

The where clause must be EventType != 'HeartBeat' otherwise you're not counting the uptime between the fault

upvoted 1 times

□ **dom271219** 9 months, 2 weeks ago

Sorry ignore it

upvoted 1 times

□ **Deeksha1234** 10 months, 1 week ago

Agree with Fer079 , EventType='HeartBeat' and Session window is correct

upvoted 1 times

□ **uzairahm** 11 months, 2 weeks 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

淘宝店铺：<https://shop63989109.taobao.com/>

□ **sparkchu** 1 year, 2 months ago

tricky but instructive question.

upvoted 2 times

□ **iooj** 1 year, 4 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 12 times

□ **MadEgg** 1 year 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, 4 months ago

B and A?

upvoted 1 times

□ **engrbrain** 1 year, 5 months 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** 1 year, 5 months 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 3 times

## Question #55

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  1 year, 5 months ago

I wish you a DP203 as easy as this question folks  
upvoted 42 times

 **CodingOwl** 7 months, 1 week ago

Username suits youe wis! :D  
upvoted 2 times

 **anto69** 1 year, 4 months ago

We all hope man  
upvoted 2 times

 **Deeksha1234** Most Recent  10 months, 1 week ago

**Selected Answer: A**

A is correct  
upvoted 3 times

 **romanzdk** 1 year, 4 months ago

**Selected Answer: A**  
Correct  
upvoted 3 times

 **leandrors** 1 year, 5 months ago

**Selected Answer: A**  
Correct  
upvoted 3 times

 **Will\_KaiZuo** 1 year, 5 months 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 1 year, 5 months 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 21 times

**Ankit\_Az** Most Recent 1 week, 2 days ago

**Selected Answer: D**

Correct. As soon as you see backfill, its tumbling.

upvoted 1 times

**Deeksha1234** 10 months, 1 week ago

**Selected Answer: D**

D is correct

upvoted 1 times

**Sriramiyer92** 10 months, 2 weeks 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 2 times

**dev2dev** 1 year, 4 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 (73%)	AD (17%)	10%
----------	----------	-----

Philippe Highly Voted 1 year, 4 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 12 times

necktru Highly Voted 1 year, 1 month 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

Spinozabubble Most Recent 1 month 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 2 times

Deeksha1234 10 months, 1 week ago

**Selected Answer: AC**

should be AC

upvoted 3 times

Boumisasound 1 year, 3 months ago

I will go for AC :

Why not D? Cause they are not mentioned some cost optimisation  
upvoted 3 times

boopathi 1 year, 3 months ago

AD are correct ?

upvoted 1 times

Istiaque 1 year, 4 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 minimize 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, 4 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, 4 months ago

淘宝店铺 : <https://shop63989109.taobao.com/>

is A,D correct

upvoted 2 times

rainbowyu 1 year, 5 months 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, 3 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 1 year, 5 months 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 1 year, 5 months ago

should it be C, D?

upvoted 2 times

Dusica 4 months, 4 weeks ago

YOU CAN'T GO WITHOUT A

upvoted 1 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, 4 months ago

Correct Answer.

upvoted 6 times

 **Jiaa** Most Recent 5 months, 1 week ago

C is correct

upvoted 2 times

 **Daniko** 7 months, 3 weeks ago

Selected Answer: C

C is correct

upvoted 2 times

 **Deeksha1234** 10 months, 1 week ago

C is correct

upvoted 1 times

 **Remedios79** 11 months, 3 weeks ago

Append is correct

upvoted 1 times

 **bad\_atitute** 1 year, 5 months ago

agree with append

upvoted 4 times

## Question #59

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 (77%)

B (23%)

corebit [Highly Voted] 1 year, 5 months 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 19 times

Canary\_2021 [Highly Voted] 1 year, 4 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

Jerrie86 4 months, 2 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 1 times

Canary\_2021 1 year, 4 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 6 times

Deeksha1234 [Most Recent] 10 months, 1 week ago

correct

upvoted 1 times

Anandtr 10 months, 1 week ago

**Selected Answer: A**

Correct

upvoted 1 times

mkthoma3 11 months, 3 weeks 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

Question #60

Topic 2

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution. After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You 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 1 year, 5 months 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 12 times

 **Deeksha1234** Most Recent 10 months, 1 week ago

**Selected Answer: B**

B is correct

upvoted 1 times

 **youngbug** 10 months, 1 week 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 2 times

 **sdokmak** 1 year ago

**Selected Answer: B**

serverless works for data lake

dedicated doesn't

upvoted 2 times

 **GDJ2022** 1 year, 4 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** 1 year, 5 months ago

If using dedicated SQL pool, after creating an external table, need a further CTAS for adding derived columns.

upvoted 3 times

## Question #61

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

A (71%)

B (29%)

✉  **rainbowyu** Highly Voted 1 year, 4 months ago

You can't use serverless pool to create table in dedicate pool  
upvoted 15 times

✉  **OldSchool** Most Recent 6 months, 1 week 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 2 times

✉  **Knoushore1** 7 months, 1 week ago

**Selected Answer: B**

Table1 is in dedicated sql pool  
upvoted 2 times

✉  **berend1** 7 months, 2 weeks ago

if table 1 would be serverless, yes, now no  
upvoted 1 times

✉  **emna2022** 8 months, 3 weeks 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 2 times

✉  **Deeksha1234** 10 months, 1 week ago

**Selected Answer: A**

yes, with serverless pool we can add a new column while creating an external table  
upvoted 1 times

✉  **youngbug** 10 months, 1 week 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

淘宝店铺：<https://shop63989109.taobao.com/>

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

Serverless pool works for data lake  
Dedicated doesn't

upvoted 1 times

□ **nefarious\_smalls** 1 year 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 2 times

□ **Andushi** 1 year, 1 month ago

**Selected Answer: A**

Answer should be Yes

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-cetas#examples>

upvoted 2 times

□ **Billybob0604** 6 months, 1 week ago

it doesn't say in the link you can add a column using external table, so no.

upvoted 2 times

□ **ranjsi01** 1 year, 2 months ago

answer is Yes

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

upvoted 1 times

□ **g2000** 1 year, 1 month ago

Table1 is not an external table

upvoted 1 times

□ **edba** 1 year, 5 months ago

correct to me.

upvoted 4 times

□ **edba** 1 year, 5 months 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 ago

edba can you please suggest where in the link is it mentioned that you can use extra columns ?

upvoted 1 times

□ **Aditya0891** 12 months ago

Ignore my comments, I got your point thanks :)

upvoted 1 times

□ **alex623** 1 year, 4 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 the answer is no

upvoted 1 times

## Question #62

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 (62%)

A (38%)

juanlu46 Highly Voted 1 year, 1 month ago

**Selected Answer: B**

Is part of a possible solution, but it isn't sufficient to meet the goal, you 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 11 times

oldpony Highly Voted 1 year 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 8 times

esaade Most Recent 3 months 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 4 times

OldSchool 6 months, 2 weeks ago

If DateTime is part of data in files in container1 then 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 1 times

Deeksha1234 10 months, 1 week 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 4 months, 4 weeks ago

AGREED

upvoted 1 times

Strix 10 months, 2 weeks ago

**Selected Answer: B**

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables>

upvoted 2 times

淘宝店铺：<https://shop63989109.taobao.com/>

□ **Davico93** 11 months, 2 weeks 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** 12 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, 1 month 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 (93%)	7%
---------	----

juanlu46 Highly Voted 1 year, 1 month 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 15 times

esaade Most Recent 3 months 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 3 times

vrodriguesp 5 months ago

**Selected Answer: A**

yes, you can execute R script in notebook and call it via adf

upvoted 3 times

urielramoss 6 months, 2 weeks ago

**Selected Answer: A**

the answer is YES. I already used this solution in a previous project.

upvoted 3 times

rzeng 7 months, 2 weeks ago

should be YES

upvoted 2 times

dom271219 9 months, 2 weeks ago

**Selected Answer: A**

We do sth like it in my company

upvoted 2 times

Deeksha1234 10 months, 1 week ago

**Selected Answer: A**

answer should be A

upvoted 3 times

 **Sriramiyer92** 10 months, 2 weeks ago

淘宝店铺 : <https://shop63989109.taobao.com/>

**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 2 times

 **Davico93** 11 months, 3 weeks 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 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 ago

**Selected Answer: A**

R in notebook and call via Data Factory

upvoted 4 times

 **MS\_Nikhil** 1 year ago

**Selected Answer: A**

You can execute R code in a notebook.

upvoted 4 times

 **hbad** 1 year 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 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 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** 6 months, 2 weeks 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, 1 month ago

**Selected Answer: A**

I agree that Yes

upvoted 2 times

 **gauravgogs** 1 year, 1 month 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,<sup>5</sup> 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%)

✉  **dom271219** 9 months, 2 weeks ago

**Selected Answer: B**

There is no R in ADF dataflow  
upvoted 2 times

✉  **Deeksha1234** 10 months, 1 week ago

**Selected Answer: B**

B is right  
upvoted 1 times

✉  **Remedios79** 11 months, 3 weeks ago

The answer is no.  
upvoted 2 times

✉  **juanlu46** 1 year, 1 month ago

**Selected Answer: B**

Is correct.  
Mapping Dataflows can't execute R code that is a requirement, so not meet the goal.  
upvoted 4 times

## Question #65

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, 1 month 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/introduction>

upvoted 11 times

bp\_a\_user 1 month, 1 week ago

...but where is the ingest done?

upvoted 1 times

juanlu46 1 year, 1 month 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 9 times

esaaide Most Recent 3 months 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 2 times

vrodriguesp 5 months ago

**Selected Answer: A**

should be yes, you can schedule notebook directly from databricks

upvoted 3 times

lemonpotato 5 months, 2 weeks ago

**Selected Answer: A**

Has to be Yes

upvoted 1 times

XiltroX 6 months, 1 week 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 7 months, 4 weeks 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 1 times

淘宝店铺：<https://shop63989109.taobao.com/>

□ **Deeksha1234** 10 months, 1 week ago

**Selected Answer: A**

yes, its possible

upvoted 1 times

□ **demirsamuel** 1 year ago

**Selected Answer: A**

I go for A as well

upvoted 2 times

□ **observador081** 1 year 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?

Seleciona somente uma resposta.

VNet1

Subnet1

WebApp1

NSG1

upvoted 2 times

□ **allagowf** 7 months, 2 weeks ago

the answer is : VNet1

upvoted 1 times

□ **cuongthh** 1 year ago

**Selected Answer: A**

I go for A.

upvoted 2 times

□ **HoangTr** 1 year 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, 1 month 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** 11 months, 2 weeks ago

you made me doubt about it

upvoted 1 times

□ **Andushi** 1 year, 1 month ago

**Selected Answer: A**

The solution meet the goal

upvoted 2 times

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, 1 month ago

Correct

upvoted 7 times

✉  **Deeksha1234** Most Recent 10 months, 1 week ago

**Selected Answer: D**

D is correct

upvoted 3 times

✉  **juanlu46** 1 year, 1 month ago

**Selected Answer: D**

Is correct

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-flatten>

upvoted 4 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.

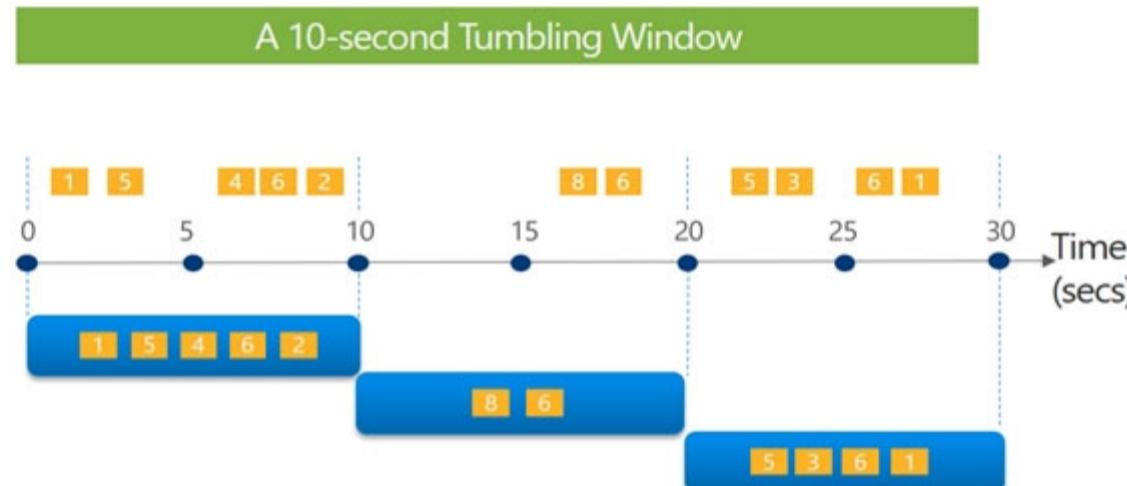
Which windowing function should you use?

- 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 11 months, 3 weeks ago

corrett. It would be corret also a hopping window with hop and size both to 5 seconds  
upvoted 5 times

**Kezzah** Most Recent 9 months, 3 weeks ago

**Selected Answer: D**

correct

upvoted 3 times

**Deeksha1234** 10 months, 1 week ago

**Selected Answer: D**

correct

upvoted 2 times

**nefarious\_smalls** 1 year ago

correct

upvoted 2 times

**juanlu46** 1 year, 1 month ago

**Selected Answer: D**

Is correct

upvoted 3 times

淘宝店铺：<https://shop63989109.taobao.com/>

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

## Question #68

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 (57%)

A (43%)

✉  **necktru** Highly Voted 1 year, 1 month 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 14 times

✉  **Raminn** Highly Voted 4 months, 3 weeks ago

**Selected Answer: C**

Using chatgpt : Append

upvoted 7 times

✉  **Ankit\_Az** Most Recent 1 week, 1 day ago

I feel Append is correct here

upvoted 1 times

✉  **janaki** 1 week, 6 days 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 1 times

✉  **esaade** 3 months 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 5 times

✉  **Rakrah** 4 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 4 times

✉  **Okea** 4 months, 1 week ago

Update mode is the answer:

淘宝店铺：<https://shop63989109.taobao.com/>

It involves writing the data records that are either new or for which the old value is updated. So this mode can be used when it is required to have the "upsert" mode of operation doing some aggregation. If no aggregation is applied, the update mode works the same as the append mode.

<https://medium.com/analytics-vidhya/spark-streaming-output-modes-600c689b6bf9>

upvoted 1 times

□ **agold96** 4 months, 3 weeks ago

**Selected Answer: C**

New rows will be added suggest "Append", correct is C for me

upvoted 3 times

□ **hanzocuk** 5 months, 1 week ago

Correct is A

Focus on the task-> "Sales >> transactions will never be updated<<. Instead, >>new rows will be added<< to adjust a sale" (Yes, very poorly formulated as usual, who is responsible for this adjustment??)

From spark docs -> "append: Only the new rows in the streaming DataFrame/Dataset will be written to the sink"

update would be similar to append if no aggregations were involved, but in our case we have aggregations.

<https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.streaming.DataStreamWriter.outputMode.html#pyspark.sql.streaming.DataStreamWriter.outputMode>:

upvoted 1 times

□ **hanzocuk** 5 months, 1 week ago

Sorry made a typo, C: Append....

upvoted 1 times

□ **hanzocuk** 5 months, 1 week ago

Moderator please dont include any of above, I feel it could mislead people as I am not even sure myself...

- 1) transactions never updated -> suggests Append
- 2) new rows added -> suggests Append.... to adjust a sale -> suggests Update
- 3) rows minimized - suggests Update

This is altogether poorly formulated... I think as a whole A: Update is a better choice

upvoted 2 times

□ **vikasptl07** 5 months, 1 week ago

As per below article answer should be complete mode

<https://medium.com/analytics-vidhya/spark-streaming-output-modes-600c689b6bf9>

upvoted 1 times

□ **vikasptl07** 5 months, 1 week ago

update in outputmode in streaming does not work without watermark on timestamp ,append is the answer

upvoted 1 times

□ **Igor85** 6 months, 2 weeks ago

the fact that aggregations are mentioned here is clearly pointing to 'update', without them it would be obviously 'append' mode

upvoted 2 times

□ **Dusica** 6 months, 3 weeks ago

APPEND

upvoted 1 times

□ **bokLuci** 7 months, 1 week ago

**Selected Answer: C**

Certainly 'C'. Line total will be aggregated at reporting time and those aggs will be run on the cumulative delta of sales transaction amounts. It's normal design of transactional delta for end of period reporting.

You don't have anything to 'Update', you are only appending the delta from the previous transaction.

upvoted 4 times

□ **rzeng** 7 months, 2 weeks ago

A - update, this reduce duplicate data

upvoted 2 times

□ **ads5891** 10 months ago

**Selected Answer: A**

I think this should be "Update" mode because the key is "minimize duplicates". Please check <https://sparkbyexamples.com/spark/spark-streaming-outputmode/>

upvoted 2 times

□ **Fidel\_104** 10 months ago

**Selected Answer: C**

I'm still not convinced about 'update'. The DB docs doesn't even mention append for writeStream outputmodes, just complete and append:  
<https://docs.databricks.com/delta/delta-streaming.html>  
Lmk if I'm missing something!

upvoted 1 times

≡  **Genere** 9 months, 3 weeks ago

Mode

Example

Notes

Complete

```
.outputMode("complete")
```

The entire updated Result Table is written to the sink. The individual sink implementation decides how to handle writing the entire table.

Append

```
.outputMode("append")
```

Only the new rows appended to the Result Table since the last trigger are written to the sink.

Update

```
.outputMode("update")
```

Only the rows in the Result Table that were updated since the last trigger will be outputted to the sink. Since Spark 2.1.1

upvoted 1 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 (74%)

A (26%)

**Saransundar** Highly Voted 1 year 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 11 times
```

**Davico93** 11 months, 2 weeks ago

but you don't need it from master, just DW1

upvoted 4 times

**Saim8711** Highly Voted 11 months, 2 weeks 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 7 times

**TestingCRM** Most Recent 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** 4 months, 3 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** 6 months, 2 weeks ago

**Selected Answer: D**

i think "D"

upvoted 1 times

**allagowf** 7 months, 2 weeks 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
```

由 **ads5891** 10 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

由 **youngbug** 10 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

由 **zxc01** 10 months, 1 week 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

由 **Saim8711** 11 months, 3 weeks 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

由 **MS2710** 5 months, 3 weeks 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

由 **jihenTR13** 11 months, 3 weeks ago

**Selected Answer: D**

i agree with Saransundar

upvoted 2 times

由 **Feljoud** 1 year, 1 month ago

**Selected Answer: A**

A: Provided source gives the solution

upvoted 3 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 ago

**Selected Answer: B**

Obviously B, IoT is event hub of stream data so we need stream analytics for sure.

upvoted 7 times

□  **Ankit\_Az** Most Recent 1 week, 1 day ago

**Selected Answer: B**

Correct

upvoted 1 times

□  **Coderhbti** 1 month, 3 weeks ago

**Selected Answer: B**

B is correct

upvoted 2 times

□  **Deeksha1234** 10 months, 1 week ago

B is correct

upvoted 1 times

□  **Remedios79** 11 months, 3 weeks 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** 4 months, 3 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 2 times

✉ **MScapris** 5 months, 2 weeks ago

**Selected Answer: C**

correct !

upvoted 2 times

✉ **Deeksha1234** 10 months, 1 week ago

correct

upvoted 2 times

✉ **dsp17** 11 months ago

Watermark delay - correct

upvoted 2 times

✉ **sagur** 1 year 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：

**Distribution:**

Hash
Round robin
Replicated

**Index:**

Clustered columnstore
Clustered
Nonclustered

店铺：学习小店66

**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 10 months, 1 week ago

Hash and clustered columnstore..right

upvoted 7 times

 **Ankit\_Az** Most Recent 1 week, 1 day ago

Correct

hash and CCI

upvoted 1 times

objecto 11 months, 3 weeks ago

淘宝店铺：<https://shop63989109.taobao.com/>

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

henryphchan 1 month ago

Round Robin is used for Staging Table

upvoted 1 times

Revave2 11 months, 2 weeks ago

I would go with Hash as the table is >2gb and is a fact table...

upvoted 8 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 ago

**Selected Answer: AB**

A and B are correct  
upvoted 9 times

 **Ankit\_Az** Most Recent 1 week, 1 day ago

**Selected Answer: AB**

Correct  
upvoted 1 times

 **Coderhbt1** 1 month, 3 weeks ago

**Selected Answer: AB**

Correct Answers  
upvoted 1 times

 **Deeksha1234** 10 months, 1 week ago

**Selected Answer: AB**

correct answer  
upvoted 1 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 search bar, there's a time filter set to 'Pacific Time (US & C...)' and 'Last 7 days'. The main table displays 21 items, showing the 'Pipeline name', 'Run start', and 'Run end' for each run. One row, 'S3ToDataLakeCopy' from 11/5/20, 6:00:18 AM, is highlighted with a red box around its pipeline name column.

Pipeline name	Run start ↑↓	Run end
S3ToDataLakeCopy	11/5/20, 6:00:18 AM	11/5/20, 6:03:18 AM
DatabricksJarPipeline	11/4/20, 6:04:11 PM	11/4/20, 6:10:11 PM
S3ToDataLakeCopy	11/4/20, 6:00:18 PM	11/4/20, 6:03:18 PM
S3ToDataLakeCopy	11/4/20, 6:00:19 AM	11/4/20, 6:04:19 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.

淘宝店铺 : <https://shop63989109.taobao.com/>

**SalesAnalyticsMLPipeline**

**Activity runs**

Pipeline run ID: a600eabe-19fb-4d0b-bd8d-d20b21223923

Activity name	Activity type	Run start	Duration	Status
Location_HTTP	Copy	11/5/20, 12:12:44 PM	00:00:15	<span style="color: green;">✓ Succeeded</span>
Clickstream_S3	Copy	11/5/20, 12:12:44 PM	00:00:27	<span style="color: red;">✗ Failed</span>
Customer_Salesforce	Copy	11/5/20, 12:12:44 PM	00:00:10	<span style="color: green;">✓ Succeeded</span>
POS_SQL	Copy	11/5/20, 12:12:44 PM	00:00:36	<span style="color: red;">✗ Failed</span>
Products_SAP	Copy	11/5/20, 12:12:44 PM	00:00:08	<span style="color: green;">✓ Succeeded</span>

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-visually>

*Community vote distribution*

A (100%)

Deeksha1234 Highly Voted 10 months, 1 week ago

**Selected Answer: A**

Monitor, in ADF we have monitor to check all activity runs  
upvoted 5 times

TechMgr Most Recent 1 month, 1 week ago

**Selected Answer: A**

A is correct  
upvoted 1 times

MScapris 5 months, 2 weeks ago

**Selected Answer: A**

is correct answer  
upvoted 4 times

nicky87654 5 months, 3 weeks ago

**Selected Answer: A**

A IS CORRECT ANSWER  
upvoted 2 times

Franz58 10 months, 2 weeks ago

I'd go with A using this:  
<https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor>

upvoted 3 times

淘宝店铺：<https://shop63989109.taobao.com/>

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

Azure monitor is different from the Monitor option shown in the screenshot.

upvoted 3 times

□ **demirsamuel** 1 year ago

answer is correct. screenshot shows azure data factory pipeline run

upvoted 2 times

□ **demirsamuel** 1 year ago

\* under monitor section in ADF

upvoted 2 times

□ **g2000** 1 year, 1 month 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 ago

Looks like Data Factory to me. If Data Factory was there I would have picked it.

upvoted 1 times

□ **upliftinghut** 1 year ago

answer is correct, monitoring is under Monitor and Dashboard

upvoted 2 times

## Question #75

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=UserErrorFileNotFound,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 (92%)	8%
---------	----

✉  **KashRaynardMorse** Highly Voted 1 year, 1 month 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 21 times

✉  **Billybob0604** 6 months, 1 week 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** 5 months, 3 weeks ago

Answer A. The path in error message shows hour=06 whereas the hour of retry run is 07.

upvoted 3 times

✉  **Remedios79** 11 months, 3 weeks ago

Thank you for the detail

upvoted 1 times

✉  **sugiats** Most Recent 3 months ago

For 7 days, this job was succeeding.

So path rule seems to be right.

upvoted 2 times

✉  **esaade** 3 months 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 2 times

✉  **raphasc** 5 months, 1 week 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** 4 months, 3 weeks ago

question says that it ran fine earlier. Parameter must have been set correctly. Answer is B

upvoted 1 times

✉  **dmitriyopo** 7 months, 1 week ago

**Selected Answer: B**

No file

upvoted 1 times

 **Rohan** 8 months ago**Selected Answer: B**

B is correct

upvoted 2 times

 **dom271219** 9 months, 2 weeks ago**Selected Answer: B**

B of course

upvoted 2 times

 **CloudixExamTopics** 9 months, 2 weeks 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 4 times

 **pangas2567** 9 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 7 times

 **Deeksha1234** 10 months, 1 week ago**Selected Answer: B**

B is correct

upvoted 2 times

 **ROLLINGROCKS** 10 months, 2 weeks 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 4 times

 **Sriramiyer92** 10 months, 2 weeks ago

B

Reason : Operation returned an invalid status code 'NotFound'. &amp; 'Message: 'The specified path does not exist.''

upvoted 1 times

 **virendrapsingh** 1 year ago**Selected Answer: B**

Answer should be B.

upvoted 2 times

 **demirsamuel** 1 year 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 intervall. Additionally the stacktrace shows that no filepath exists.

upvoted 3 times

 **RamboRinky** 1 year 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 ago

it was correct for 7 days so there's no way that's true.

upvoted 5 times

 **chuckas** 1 year, 1 month ago

I agree B for me

upvoted 3 times

 **g2000** 1 year, 1 month ago

B seems a better fit. no data

upvoted 3 times

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 Spark 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 ago

I think this is one is correct.

upvoted 8 times

 **Ankit\_Az** Most Recent 1 week, 1 day ago

**Selected Answer: C**

Correct

upvoted 1 times

 **akk\_1289** 4 months ago

The correct answer is C. From Synapse Studio, select the workspace. From Monitor, select Apache Spark applications.

upvoted 2 times

 **pangas2567** 9 months ago

**Selected Answer: C**

Correct

<https://docs.microsoft.com/en-us/azure/synapse-analytics/monitoring/how-to-monitor-spark-applications>

upvoted 3 times

 **Deeksha1234** 10 months, 1 week ago

**Selected Answer: C**

C is correct

upvoted 1 times

Question #77

**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, 1 month ago

Similar to another question in this dump. Seems correct!

upvoted 12 times

 **rzeng** Most Recent 7 months, 2 weeks ago

correct

upvoted 3 times

 **dom271219** 9 months, 2 weeks ago

"Specify a temporary folder to stage the data" must be before creating the DF : I am wrong ?

upvoted 1 times

淘宝店铺：<https://shop63989109.taobao.com/>

□ **Karl\_Cen** 5 months 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** 10 months, 1 week ago

correct

upvoted 1 times

□ **nefarious\_smalls** 1 year ago

correct

upvoted 1 times

□ **demirsamuel** 1 year 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%)	13%	13%
----------	-----	-----

✉ dom271219 Highly Voted 9 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 15 times

✉ Ankit\_Az Most Recent 1 week, 1 day ago

**Selected Answer: AB**

Correct

upvoted 1 times

✉ Debasish93 1 month 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 2 times

✉ azure\_user11 1 month 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

✉ esaade 3 months 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 4 times

ak\_1289 4 months ago

淘宝店铺：<https://shop63989109.taobao.com/>

The correct answers are B. Create a Git repository and A. From the UX Authoring canvas, select Set up code repository.  
upvoted 1 times

Raminn 4 months, 3 weeks 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 6 months, 1 week ago

Selected Answer: AF

Since there is no mention of GitHub or DevOps the solution that works for both is A & F

upvoted 1 times

dmitriypo 7 months, 1 week 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 7 months, 4 weeks 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 8 months, 4 weeks ago

F A is the correct order. Save changes then set up code repos

upvoted 2 times

anks84 9 months ago

Selected Answer: AF

Should be AF as we want to achieve the version control for the code changes.

upvoted 3 times

federic 9 months ago

why not A and B ? I would set up the code repository after creating the git repo

upvoted 3 times

yuorrik 3 months, 1 week 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 4 months ago

If it's hard to remember, memorize it as CAMP.

upvoted 15 times

**azure\_user11** Most Recent 1 month ago

Correct. I've worked with this many times. It's the right order.

upvoted 2 times

**Karl\_Cen** 5 months 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** 5 months 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** 3 months, 3 weeks 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 1 times

✉ **vrodriguesp** 5 months 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 2 times

✉ **dmitriypo** 7 months, 1 week ago

Correct

upvoted 2 times

✉ **TiredDad** 7 months, 1 week ago

Should it not be merge, then publish, then create a schedule trigger, finally associate the schedule trigger with pipeline1?

upvoted 3 times

✉ **TiredDad** 7 months, 1 week ago

Pls ignore, I agree with the suggested answer

upvoted 2 times

✉ **rzeng** 7 months, 2 weeks ago

correct

upvoted 1 times

✉ **anks84** 9 months ago

Looks correct

upvoted 2 times

Question #80

**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 =
    (
        abfs
        abfss
        wasb
        wasbs
    )
    ://data@newyorktaxidataset.dfs.core.windows.net' ,
credential = ADLS_credential ,
TYPE -
);
    (
        BLOB_STORAGE
        HADOOP
        RDBMS
        SHARP MAP MANAGER
)
```

**Correct Answer:**

**Answer Area**

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore
WITH
( Location1 =
    (
        abfs
        abfss
        wasb
        wasbs
    )
    ://data@newyorktaxidataset.dfs.core.windows.net' ,
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**  9 months, 1 week 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 19 times

✉ **vigilante89** Highly Voted 5 months, 3 weeks ago

abfss and Hadoop

upvoted 5 times

✉ **ZIMARAKI** Most Recent 4 months, 4 weeks ago

abfss & hadoop

upvoted 3 times

✉ **greenlever** 7 months, 4 weeks ago

abfss

Hadoop

abfss endpoint when your account has secure transfer enabled

upvoted 5 times

✉ **sesank** 8 months, 2 weeks ago

abfss

Hadoop

upvoted 3 times

✉ **anks84** 9 months ago

-abfss

-hadoop

upvoted 5 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 (83%)      B (17%)

 **yogiazaad** Highly Voted 4 months, 1 week 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 11 times

 **esaade** Highly Voted 3 months 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 5 times

 **Ast999** Most Recent 3 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** 5 months, 3 weeks ago

**Selected Answer: B**

Correct

upvoted 2 times

 **dom271219** 9 months ago

**Selected Answer: B**

Agreed

upvoted 2 times

 **kumarrahul1107** 9 months, 1 week 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 (100%)

 **dmitriypo** 7 months, 1 week ago

**Selected Answer: BE**

Agree with the answer

upvoted 4 times

 **allagowf** 7 months, 2 weeks ago

**Selected Answer: BE**

Answer is correct : BE

upvoted 4 times

 **amitshinde14** 8 months, 3 weeks ago

Correct ans.

upvoted 3 times

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%)

 **Ankit\_Az** 1 week, 1 day ago

**Selected Answer: C**

Correct

upvoted 1 times

 **GodfreyMbizo** 4 months, 1 week ago

correct answer

upvoted 2 times

 **GodfreyMbizo** 4 months, 1 week ago

database scoped credentials first

upvoted 2 times

 **anks84** 9 months ago

**Selected Answer: C**

Correct Answer !

upvoted 4 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>

✉️  **kim32** 4 weeks, 1 day ago

I selected B, pull request  
upvoted 1 times

✉️  **Xinyuehong** 7 months, 4 weeks ago

I had heard "publish to", never heard of "publish from". So confused.  
upvoted 2 times

✉️  **Igor85** 6 months, 2 weeks ago

probably it was meant to publish from collaboration branch to adf\_publish branch  
upvoted 2 times

✉️  **debarun** 9 months, 1 week ago

Why not B ?  
upvoted 1 times

✉️  **DataEX** 9 months, 1 week 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 7 times

## Question #85

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: "-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** 2 months, 3 weeks 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 3 times

淘宝店铺：<https://shop63989109.taobao.com/>

 **Okea** 5 months, 1 week 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 3 times

 **dom271219** 9 months ago

**Selected Answer: D**

```
"dependsOn": [  
  {  
    "type": "SelfDependencyTumblingWindowTriggerReference",  
    "size": "01:00:00",  
    "offset": "-01:00:00"  
  }]
```

upvoted 4 times

Question #86

**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**

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
<b>A Get Metadata activity</b>
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  9 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 6 times

 **dmitriypo** Highly Voted  7 months, 1 week ago

Looks correct. Checked in the doc.

upvoted 5 times

 **anks84** Most Recent  9 months ago

Looks Correct !!

upvoted 2 times

## Question #87

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 (81%)

B (19%)

 **Amalbenrebai** Highly Voted 1 year, 9 months ago

- data engineers: high concurrency cluster
  - jobs: Standard cluster
  - data scientists: Standard cluster
- upvoted 80 times

 **gogosgh** 1 month ago

The issue is the jobs are going to be ran by multiple users i.e. engineers and scientists? So it needs to be hugi concurrency cluster?  
upvoted 1 times

 **supriyako** 8 months, 3 weeks ago

Correct. Because jobs could be for Scala notebook, which is supported by Standard cluster mode  
upvoted 1 times

 **Egocentric** 1 year, 1 month ago

agreed  
upvoted 1 times

 **Julius7000** 1 year, 8 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 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 1 year, 9 months ago

The answer must be A!

upvoted 33 times

□ **Ast999** Most Recent 3 months, 1 week ago

**Selected Answer: A**

SCALA = STANDARD

upvoted 2 times

□ **allagowf** 7 months, 2 weeks ago

**Selected Answer: A**

data scientists and Job --> Scala --> Standard cluster .

upvoted 1 times

□ **greenlever** 7 months, 4 weeks ago

**Selected Answer: A**

Correct

upvoted 1 times

□ **anks84** 9 months ago

**Selected Answer: A**

We would need a Standard cluster for the jobs to support Scala. High-concurrecy cluster does not support Scala.  
Hence, the Answer is A !

upvoted 1 times

□ **Deeksha1234** 10 months, 2 weeks ago

**Selected Answer: B**

the answer should be No

upvoted 1 times

□ **Deeksha1234** 9 months, 4 weeks ago

sorry 'A' should be correct

upvoted 2 times

□ **sethuramansp** 11 months, 2 weeks 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

□ **Eyepatch993** 1 year, 2 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

□ **Boompiee** 1 year, 1 month 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

□ **Aditya0891** 12 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 1 times

□ **Hanse** 1 year, 3 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 5 times

□ **ovokpus** 1 year, 3 months ago

**Selected Answer: A**

Yes it seems to be!

upvoted 2 times

淘宝店铺：<https://shop63989109.taobao.com/>

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

**Selected Answer: A**

correct

upvoted 2 times

□ **kilowd** 1 year, 4 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

□ **let\_88** 1 year, 4 months ago

As per the doc in Microsoft the High Concurrency cluster doesn't support 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.

<https://docs.microsoft.com/en-us/azure/databricks/clusters/configure#cluster-mode>

upvoted 6 times

□ **tesen\_tolga** 1 year, 4 months ago

**Selected Answer: A**

The answer must be A!

upvoted 2 times

□ **SabaJamal2010AtGmail** 1 year, 5 months ago

The solution does not meet the requirement because: "High Concurrency clusters work only for 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.

upvoted 2 times

□ **FredNo** 1 year, 6 months ago

**Selected Answer: A**

Data scientists and jobs use Scala so they need standard cluster

upvoted 9 times

## Question #88

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.azure.databricks.net/clusters/configure.html>

Community vote distribution

B (100%)

 **dfdsfsfsd** Highly Voted 2 years 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 42 times

 **Preben** 2 years 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** 1 year, 11 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 1 year, 11 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 13 times

 **Pais** Most Recent 6 months, 1 week ago

**Selected Answer: B**  
 High-concurrency cluster does not support Scala.  
 upvoted 1 times

 **OldSchool** 6 months, 3 weeks ago

**Selected Answer: B**

Jobs require Scala so the answer is B) No.  
upvoted 1 times

**greenlever** 7 months, 4 weeks ago

**Selected Answer: B**

Cluster for Jobs should support scala - STANDARD  
upvoted 1 times

**anks84** 9 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** 9 months, 1 week 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** 9 months, 4 weeks ago

No is correct  
upvoted 1 times

**ClassMistress** 1 year 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, 2 months ago

1 and 2 are same questions but answers are different why?  
upvoted 2 times

**Hanse** 1 year, 3 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 Standard  
upvoted 2 times

**lukeonline** 1 year, 5 months ago

**Selected Answer: B**

high concurrency does not support scala  
upvoted 2 times

**rashjan** 1 year, 6 months ago

**Selected Answer: B**

wrong: no  
upvoted 1 times

**FredNo** 1 year, 6 months ago

**Selected Answer: B**

Answer is no because high concurrency does not support scala  
upvoted 5 times

**Aslam208** 1 year, 7 months ago

Answer is No  
upvoted 2 times

**damaldon** 1 year, 11 months ago

Answer: NO  
-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

**nas28** 2 years ago

Answer correct : No. but the reason is wrong, They want data scientists cluster to shut down automatically after 120 minutes so Standard cluster not high concurrency

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 (100%)

 **anks84** Highly Voted 9 months ago

**Selected Answer: A**

Answer is Correct !

upvoted 7 times

 **dmitriypo** Highly Voted 7 months, 1 week ago

**Selected Answer: A**

Of course A

upvoted 5 times

 **rzeng** Most Recent 7 months, 2 weeks ago

A is right

upvoted 4 times

## Question #90

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 (72%)	DE (22%)	6%
----------	----------	----

✉ **dumbled** 1 month, 2 weeks ago

**Selected Answer:** BD

correct

upvoted 2 times

✉ **esaade** 3 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 4 times

□ **Lestrang** 4 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** 4 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** 4 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** 6 months, 2 weeks ago

**Selected Answer: BD**

need to first change UDF to a calculated column and then enable result set caching. agreed with the answer

upvoted 3 times

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

□ **kl8585** 7 months 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 2 times

□ **rzeng** 7 months, 2 weeks 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** 7 months, 4 weeks 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** 9 months ago

**Selected Answer: BD**

Answer is Correct !

upvoted 4 times

Question #91

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 (50%)

A (50%)

✉ **vadiminski\_a** 2 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 1 times

✉ **vadiminski\_a** 2 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

✉ **esaade** 2 months, 3 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 4 times

✉ **Kate0204** 3 months ago

**Selected Answer: A**

An Azure Stream Analytics job consists of an input, query, and an output.  
 upvoted 1 times

✉ **Karl\_Cen** 4 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** 4 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** 4 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

 **alexnicolita** 4 months, 3 weeks ago

淘宝店铺 : <https://shop63989109.taobao.com/>

**Selected Answer: A**

Why not Azure Stream Analytics and Azure Synapse Analytics?  
upvoted 2 times

Question #92

Topic 2

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** 4 months, 3 weeks ago

**Selected Answer: A**

<https://learn.microsoft.com/en-us/azure/databricks/kb-scala/flatten-nested-columns-dynamically>  
upvoted 3 times

 **Raminn** 4 months, 3 weeks 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>	

Correct Answer:

```
df_sales.filter(col('Region')=='HQ').
    .groupBy(col('SalesPerson'))
    .agg(sum('Amount').alias('TotalAmount')).
    .orderBy(desc('TotalAmount'))
    .limit(3)
```

 **esaade** Highly Voted 3 months ago

```
df_sales.filter(col("Region") == "HQ")
    .groupBy(col('SalesPerson'))
    .agg(sum('Amount').alias('TotalAmount'))
    .orderBy(desc('TotalAmount'))
    .limit(3)
```

upvoted 5 times

 **aurorafang** Most Recent 4 months, 2 weeks ago

for the sequence, group\_by usually put before the order by operations  
upvoted 3 times

 **Raminn** 4 months, 3 weeks ago

correct

upvoted 1 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%)

 **haythemsi** 1 month ago

Correct, D

upvoted 1 times

 **AHUI** 2 months, 1 week ago

ans is correct

upvoted 1 times

 **FRANCIS\_A\_M** 2 months, 1 week ago

**Selected Answer: D**

Correct, D

upvoted 4 times

## Question #95

## 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.

FRANCIS\_A\_M Highly Voted 2 months, 1 week ago

Correct

upvoted 6 times

SinSS Most Recent 2 weeks, 6 days 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 1 times

mhi 3 weeks, 4 days ago

Shouldn't you merge the new branch into the main branch?

upvoted 4 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%)

 **aemilka** 1 month, 3 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 4 times

 **shakes103** 2 months ago

**Selected Answer: BC**

Answer is B and C.  
upvoted 4 times

 **Sibaprasad** 2 months 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 2 times

 **FRANCIS\_A\_M** 2 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 4 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** Highly Voted 2 months 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 7 times

✉ **aemilka** 1 month ago

Correct.

"The lineage data will automatically be captured during the activities execution."

upvoted 1 times

✉ **Gopinath123** Most Recent 1 month, 1 week ago

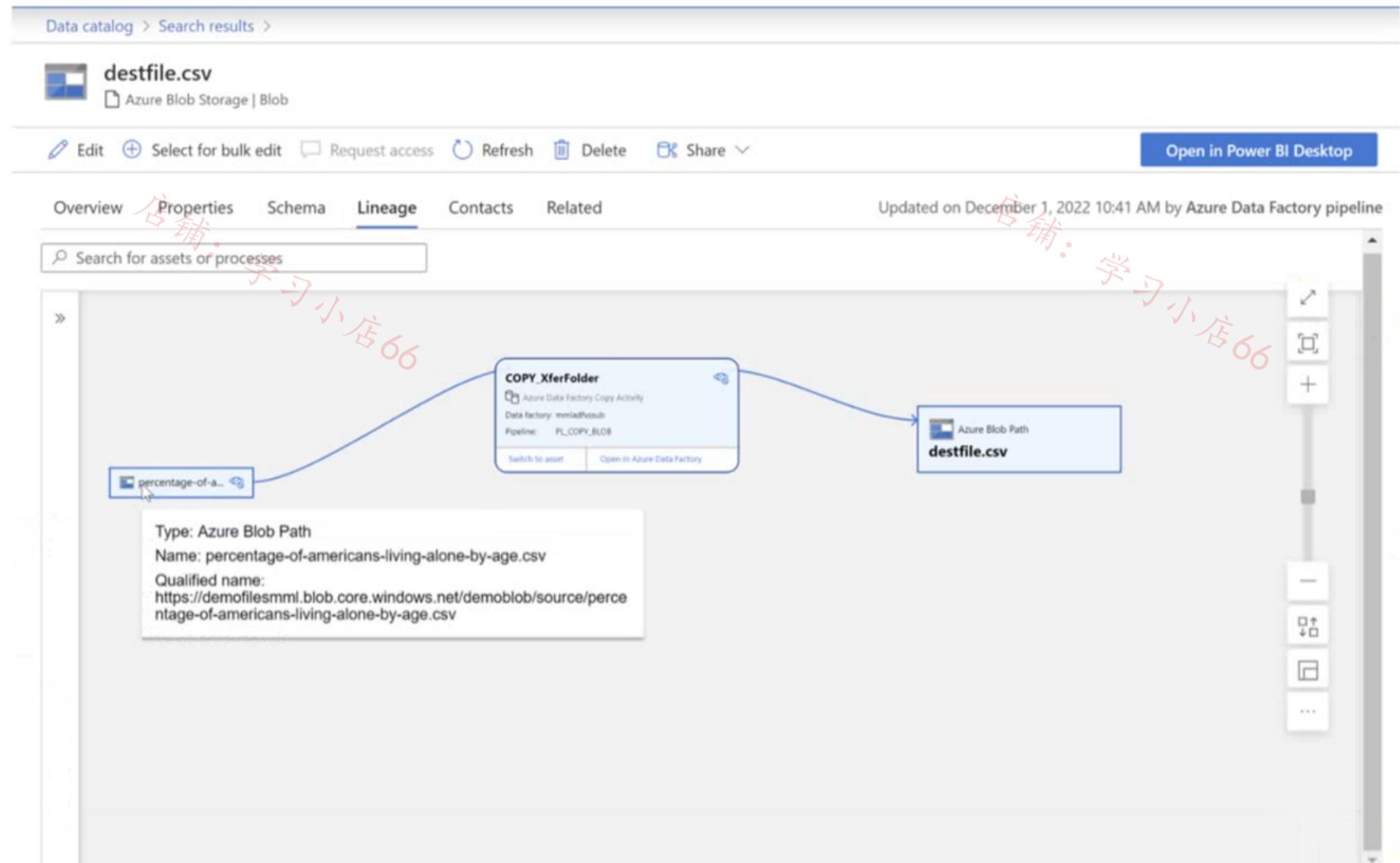
**Selected Answer: B**

<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 1 month, 2 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 4 times

shakes103 1 month, 2 weeks ago

The answer is also displayed on the top right corner of the image displayed.

upvoted 5 times

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

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 (100%)

 **Sibaprasad** Highly Voted  2 months 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 9 times

 **peches** Most Recent  5 days, 2 hours 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 1 times

 **chryckie** 1 month, 2 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 4 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 2 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 24 times

 **vegeta379** Most Recent 1 week, 5 days 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** 2 weeks, 2 days ago

I think the requirement is to select specific columns, hence CSV?  
upvoted 1 times

 **verisdev** 3 weeks, 3 days ago

淘宝店铺：<https://shop63989109.taobao.com/>

it supposed to be Parquet instead of CSV  
upvoted 2 times

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

## 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>COMMIT TRAN</code>	
<code>ROLLBACK TRAN</code>	
<code>SET RESULT_SET_CACHING ON</code>	
...	
...	
...	
...	
...	
...	

```

BEGIN TRY
    INSERT INTO dbo.Table1 (col1, col2, col3)
    SELECT col1, col2, col3 FROM stage.Table1;
END TRY
BEGIN CATCH
    IF @@TRANCOUNT > 0
        BEGIN
            ROLLBACK TRAN;
        END
    END CATCH;
    IF @@TRANCOUNT > 0
        BEGIN
            COMMIT TRAN;
        END

```

## Answer Area

```

BEGIN DISTRIBUTED TRANSACTION
BEGIN TRY
    INSERT INTO dbo.Table1 (col1, col2, col3)
    SELECT col1, col2, col3 FROM stage.Table1;
END TRY
BEGIN CATCH
    IF @@TRANCOUNT > 0

```

Correct Answer:

```

        BEGIN
            ROLLBACK TRAN;
        END
    END CATCH;
    IF @@TRANCOUNT > 0
        BEGIN
            COMMIT TRAN;
        END

```

Given answer is wrong. It should be BEGIN TRAN 淘宝店铺：<https://shop63989109.taobao.com/> 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 8 times

 **janaki** Most Recent 1 week, 1 day ago

Its BEGIN TRAN  
then ROLLBACK TRAN  
upvoted 2 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:

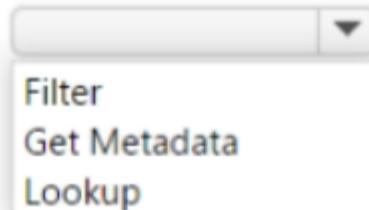
- Minimize the effort to author the pipeline.
- Ensure that the number of data integration units allocated to the upload operation can be controlled.

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

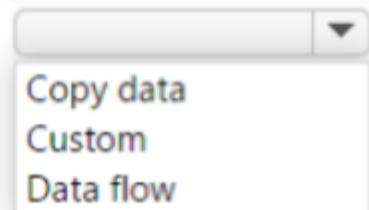
NOTE: Each correct answer is worth one point.

**Answer Area**

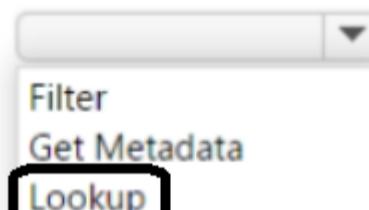
To retrieve the watermark value, use:



To perform the upload, use:

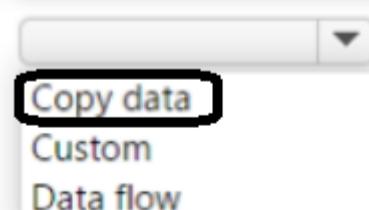
**Answer Area**

To retrieve the watermark value, use:



**Correct Answer:**

To perform the upload, use:



**OfficeSaracus** Highly Voted 1 month ago

Seems correct to me

upvoted 6 times

**DarKru** Most Recent 1 week, 1 day ago

Correct. The example is here

<https://learn.microsoft.com/en-us/azure/data-factory/tutorial-incremental-copy-portal>

upvoted 1 times

**haythemsi** 1 month 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 2 times

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

## 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 =   
              
              
              
    FIELDTERMINATOR = '0x0b',  
    FIELDQUOTE =   
                  
                  
                  
    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：学习小店66

 **Yemeral** Highly Voted  1 month 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 8 times

店铺：学习小店66

店铺：学习小店66

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 (67%)

BC (33%)

✉ **akk\_1289** Highly Voted 1 month 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 6 times

✉ **henryphchan** 1 month ago

agree with you  
upvoted 1 times

✉ **aemilka** Most Recent 1 month ago

Correct  
upvoted 2 times

✉ **haythemsi** 1 month ago

**Selected Answer: CD**  
Correct  
upvoted 2 times

✉ **Aninina** 1 month ago

**Selected Answer: BC**  
I think B and C  
upvoted 1 times

Question #105

## 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] 1 month ago

Correct! It's a easy one.

upvoted 5 times

✉  **shakes103** [Most Recent] 1 month 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**

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**

**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")
```

Correct Answer:

```
%%local  
%%spark  
%%sql  
  
val scala_df = spark.sqlContext.sql ("select * from pysparkdftemptable")  
scala_df.write.  
    jdbc  
    saveAsTable  
    synapsesql
```

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 ADFPipeline1.

You need to invoke SynPipeline1 from ADFPipeline1.

Which type of activity should you use?

- A. Web
- B. Spark
- C. Custom
- D. Notebook

**Correct Answer: A**

## 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

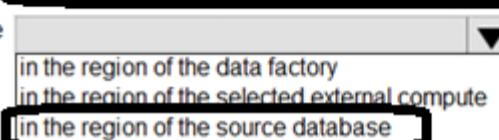
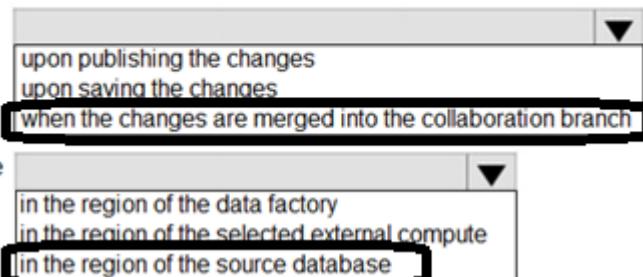
in the region of the data factory  
in the region of the selected external compute  
in the region of the source database

### 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



店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

**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.

▼

one time
two times
zero times

▼

is unaffected by
will automatically adjust for
will require an adjustment for

Question #112

Topic 2

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

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

**Topic 3 - Question Set 3**

## Question #1

## 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 ago

1. create user from external provider for Group1
2. create Role1 with select on schema1
3. add user to the Role1

upvoted 84 times

SameerL 11 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 5 times

AlexLo 1 year, 4 months ago

- Sorry, but "add user to the Role1" is not part of the answers. Or, which option is that?

upvoted 3 times

thomas02 1 year, 3 months ago

□ **PallaviPatel** 1 year, 4 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, 1 month ago

UDEMY says this as well. So correct

upvoted 1 times

□ **patricka95** Highly Voted 1 year, 10 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** 1 year, 5 months ago

Can somebody explain why we have to create the user first and not the role?

upvoted 2 times

□ **vanrell** 1 year, 2 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 2 times

□ **SQLDev0000** 1 year, 2 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 2 times

□ **sachabess79** 1 year, 8 months ago

Agreed 100%

upvoted 3 times

□ **Aditya0891** 11 months, 4 weeks ago

Answer is not wrong. Read the question properly

upvoted 1 times

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

correct

upvoted 1 times

□ **Deeksha1234** 10 months, 1 week ago

given answer is correct

upvoted 2 times

□ **SabaJamal2010AtGmail** 1 year, 5 months 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 5 times

□ **ADHDBA** 1 year, 2 months ago

it should be least privileged so select on schema is correct not on dw1

upvoted 1 times

□ **eng1** 1 year, 11 months ago

It should be D-E-A

upvoted 1 times

□ **eng1** 1 year, 11 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** 1 year, 11 months ago

It should be C-A-E

upvoted 1 times

□ **SG1705** 1 year, 12 months ago

Is the answer correct ??

upvoted 1 times

淘宝店铺：<https://shop63989109.taobao.com/>

□  **Marcello83** 1 year, 11 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** 11 months, 3 weeks 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
<b>TDE with customer-managed keys</b>
TDE with platform-managed keys

To maintain client app access in the event of a datacenter outage:

<b>Create and configure Azure key vaults in two Azure regions.</b>
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:

淘宝店铺：<https://shop63989109.taobao.com/>

<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 1 year, 11 months ago

Guys the answers are correct: <https://docs.microsoft.com/en-us/azure/azure-sql/database/transparent-data-encryption-byok-overview>  
upvoted 53 times

Slena 1 year, 8 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 4 times

bhavesh\_wadhwani Highly Voted 1 year, 8 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 1 year, 8 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

Deeksha1234 Most Recent 10 months, 1 week ago

correct answer  
upvoted 1 times

SabaJamal2010AtGmail 1 year, 5 months ago

Both answers Correct 1) Transparent Data Encryption with customer-managed key 2) key vault in 2 regions  
upvoted 1 times

Skeinofi 1 year, 5 months 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 1 year, 8 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 1 year, 12 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 1 year, 12 months ago

Moreover, always encrypted is NOT TDE option. The question asks to enable TDE.  
upvoted 3 times

Reel 6 months, 3 weeks 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 1 year, 12 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 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 in 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** 1 year, 12 months ago

TDE doesn't operate with customer keys but always encrypted does

upvoted 1 times

□ **Rob77** 2 years 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** 1 year, 5 months 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

Question #3

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	Customer	PasswordSalt
dbo	ErrorLog	UserName
SalesLT	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 1 year, 11 months ago

Correct!

upvoted 21 times

 **saty\_nl** Highly Voted 1 year, 11 months ago

淘宝店铺：<https://shop63989109.taobao.com/>

Answer is correct. Dynamic data masking will limit the exposure of sensitive data.  
upvoted 7 times

 **anks84** Most Recent 9 months ago

**Selected Answer: AC**  
Given Answers are correct!  
upvoted 4 times

 **Deeksha1234** 10 months, 1 week ago

correct  
upvoted 2 times

 **Remedios79** 11 months, 2 weeks ago

Also for me is correct  
upvoted 2 times

 **sparkchu** 1 year, 2 months ago

log auditing & tracing is important for data governance, therefore necessary for any data solution.  
upvoted 2 times

 **rashjan** 1 year, 6 months ago

**Selected Answer: AC**  
correct  
upvoted 2 times

 **rashjan** 1 year, 6 months 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 4 times

 **dduque10** 1 year, 8 months ago

Is it really C correct?  
upvoted 1 times

**店铺：学习小店66**

 **rashjan** 1 year, 6 months ago  
Yes, the logs are used to identify the user who executed the query.  
upvoted 2 times

 **Dizzystar** 1 year, 7 months ago

wondering the same thing.  
upvoted 1 times

**店铺：学习小店66**

**店铺：学习小店66**

**店铺：学习小店66**

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 1 year, 12 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 31 times

✉ **anto69** 1 year, 4 months ago

yeah, from ms docs: "ensuring that specific users can access only certain columns of a table pertinent to their department"

upvoted 2 times

✉ **mikerss** 1 year, 11 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 8 times

✉ **Deeksha1234** 10 months, 1 week ago

I agree with the logic provided

upvoted 2 times

✉ **Marcus1612** 1 year, 8 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 2 times

✉ **Tracy\_Anderson** 1 year, 10 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 1 year, 6 months ago

**Selected Answer: C**

Data masking does not protect against inferring with the data

upvoted 10 times

✉ **SinSS** Most Recent 2 weeks, 6 days ago

Only with DDM, you can guess with trying some queries

upvoted 1 times

✉ **Okea** 4 months, 1 week 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** 9 months ago

**Selected Answer: C**

Column level security is the correct answer !!

upvoted 2 times

 **Deeksha1234** 10 months, 1 week ago

correct

upvoted 1 times

 **orm33** 1 year 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** 12 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, 1 month 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, 4 months ago

There are 2 parts to it:

1. provide salespeople with the ability to **< b >view all the entries</b>** 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, 4 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, 4 months ago**Selected Answer: A**

You get 'The SELECT permission was denied on the colum...' 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** 1 year, 5 months 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** 1 year, 5 months 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** 1 year, 6 months ago

Selected Answer: C

upvoted 1 times

 **Amalbenrebai** 1 year, 9 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

 **patricka95** 1 year, 10 months ago

Column level security is the correct answer.

It is obvious based on "The solution must prevent all the salespeople from viewing or inferring the credit card information.". If masking was used, they could still view or infer the credit card data. Also, I interpret "Entries" to imply rows.

upvoted 1 times

 **Himlo24** 2 years ago

Shouldn't the answer be C? Because the salesperson will get an error when trying to query credit card info.

upvoted 3 times

 **mvisca** 2 years ago

Nope, the salesperson, generally, uses the last 4 digits of the card to validate, in a pickup for example. They don't need to know all the others numbers, so data masking is correct.

upvoted 10 times

 **mbravo** 2 years ago

It is not because there is a requirement that the data should be protected not only from viewing but also inferring. Masked data can still be inferred using brute force techniques. The only option in this case is C (Column level encryption).

upvoted 5 times

 **terajuana** 1 year, 12 months ago

nope - the question contains

"You need to recommend a solution to provide salespeople with the ability to view all the entries in Customers" if you implement column-level security then they cannot view all items i.e. select \* from the table because it will give them an error. The only way to fulfil the requirement therefore is masking

upvoted 10 times

 **captainbee** 1 year, 11 months ago

Ironically DP-200 has the exact same question and everyone was leaning toward Column Level Security. I think being able to look at all entries means looking at all ROWS, rather than columns. They're able to do that still with CLS, just can't see all columns. You can still infer when there's data masking.

upvoted 2 times

 **ML\_Novice** 9 months, 1 week ago

please, what do you mean by "still Infer"?

upvoted 1 times

 **Preben** 2 years ago

"You need to recommend a solution to provide salespeople with the ability to view all the entries in Customers."

Credit card data is an entry in the Customers table. How can they view that entry if you use column level encryption?

upvoted 3 times

Question #5

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 1 year, 6 months ago

**Selected Answer: ACE**

correct

upvoted 14 times

 **Nathan\_W** Highly Voted 1 year, 8 months ago

nice question!

upvoted 6 times

 **francocalvo** Most Recent 3 weeks, 6 days ago

Isn't this a ACL model instead of RBAC?

upvoted 1 times

 **Jerrie86** 4 months, 2 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** 9 months ago

**Selected Answer: ACE**

CORRECT !!

upvoted 3 times

 **ML\_Novice** 9 months, 1 week ago

E-> A->C

is the order right ?

upvoted 2 times

 **Deeksha1234** 10 months, 1 week ago

**Selected Answer: ACE**

correct

upvoted 1 times

 **juanlu46** 1 year, 1 month ago

**Selected Answer: ACE**

Is correct!

upvoted 1 times

 **nss8500** 1 year, 4 months ago

**Selected Answer: ACE**

correct

upvoted 1 times

 **Podavenna** 1 year, 8 months ago  
Correct answer!  
upvoted 4 times

淘宝店铺 : <https://shop63989109.taobao.com/>

店铺: 学习小店66

店铺: 学习小店66

店铺: 学习小店66

店铺: 学习小店66

## Question #6

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 1 year, 9 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 34 times

✉ **hanzocuk** 5 months, 1 week 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

✉ **rzeng** Most Recent 7 months, 2 weeks ago

so you need to encrypt the df, you need to remove the bonded service first , answer is correct

upvoted 1 times

✉ **RajashekharC** 9 months, 3 weeks ago

Its C:

Your ADF should be empty during encryption process using a KEY

upvoted 3 times

✉ **Deeksha1234** 10 months, 1 week ago

**Selected Answer: C**

correct answer

upvoted 2 times

✉ **juanlu46** 1 year, 1 month 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, 4 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** 1 year, 5 months 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** 1 year, 5 months 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** 1 year, 6 months 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** 1 year, 7 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** 1 year, 9 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 1 year, 8 months ago

Correct answer!

upvoted 27 times

 **kornat** Most Recent 2 months ago

correct

upvoted 1 times

 **TimboobmiT** 7 months, 3 weeks ago

Why not dynamic data masking?

upvoted 2 times

 **Deeksha1234** 10 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 4 times

 **juanlu46** 1 year, 1 month 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, 2 months ago

Correct!

upvoted 1 times

**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 file system. 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.

Hot Area:

**Answer Area**

Use

- |                                   |
|-----------------------------------|
| Azure Active Directory (Azure AD) |
| 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

- |                                   |
|-----------------------------------|
| Azure Active Directory (Azure AD) |
| a shared access signature (SAS)   |
| a shared key                      |

to authenticate by using

- |                         |
|-------------------------|
| a managed identity      |
| a stored access policy  |
| an Authorization header |

Box 1: Azure Active Directory (Azure AD)

On Azure, managed identities eliminate the need for developers having to manage credentials by providing an identity for the Azure resource in Azure AD and using it to obtain Azure Active Directory (Azure AD) tokens.

Box 2: a managed identity -

A data factory can be associated with a managed identity for Azure resources, which represents this specific data factory. You can directly use this managed identity for Data Lake Storage Gen2 authentication, similar to using your own service principal. It allows this designated factory to access and copy data to or from your Data Lake Storage Gen2.

Note: The Azure Data Lake Storage Gen2 connector supports the following authentication types.

- Account key authentication
- Service principal authentication
- Managed identities for Azure resources authentication

Reference:

<https://docs.microsoft.com/en-us/azure/active-directory/managed-identities-azure-resources/overview> <https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage>

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

Correct Answer!

upvoted 26 times

 **Deeksha1234** Most Recent 10 months ago

correct

upvoted 2 times

 **juanlu46** 1 year, 1 month ago

It's make sense. You only authorised Data Factory instance by Azure Active Directory, you don't need to share keys that can be retrieved by users. And this option meet all of the requirements.

upvoted 2 times

淘宝店铺 : <https://shop63989109.taobao.com/>

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

**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>

 steee Highly Voted 1 year, 9 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 111 times

 Amalbenrebai 1 year, 9 months ago

I agree: NO NO NO

upvoted 13 times

Seansmyke 1 year, 3 months ago

淘宝店铺：<https://shop63989109.taobao.com/>

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 1 week, 6 days ago

but the questions is asked about region B and region C.  
upvoted 1 times

ADHDBA 1 year, 1 month 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, 1 month 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 12 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

AzureJobsTillRetire 6 months, 1 week ago

I do not agree. I think HaBroNouen says it well as below.  
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.  
upvoted 3 times

Shanmahi 6 months ago

Did this question appear in any of the exam attempts ?  
upvoted 4 times

dsp17 11 months ago

100 % Agreed.  
upvoted 1 times

HaBroNouen Highly Voted 1 year, 8 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 70 times

Julius7000 1 year, 8 months ago

I think You are correct  
upvoted 6 times

noranathalie 1 year, 7 months ago

I would go for this answer as well.. otherwise the double question 2 and 3 would be useless..  
upvoted 2 times

YLiu 1 year, 7 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

janaki Most Recent 1 week, 1 day 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

chryckie 1 month, 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** 1 month, 1 week 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 1 times

□ **chryckie** 1 month, 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** 1 month, 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** 1 month, 1 week 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** 2 months, 2 weeks ago

after reading discussion very confused. What could be the answer.

upvoted 6 times

□ **rcpaudel** 2 weeks, 3 days ago

Correct answer is YES, NO & YES, look at the explanation from essaade 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

□ **essaade** 2 months, 4 weeks 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 3 times

□ **Billybob0604** 6 months ago

This answer is clearly NO, NO, NO

upvoted 1 times

□ **XiltroX** 6 months 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** 7 months, 1 week 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

淘宝店铺：<https://shop63989109.taobao.com/>

✉ **rzeng** 7 months, 2 weeks ago

agree with No, No, No

upvoted 1 times

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

No no no

Obviously

upvoted 1 times

✉ **Deeksha1234** 10 months ago

Correct, Agree with Habronounen

upvoted 1 times

✉ **Saddu1** 10 months, 2 weeks ago

The answer must be yes yes yes as the engineers have access to sensitive data across regions, they need masking

upvoted 1 times

✉ **wikicog** 11 months, 1 week ago

Answer is correct

upvoted 2 times

✉ **wikicog** 11 months, 1 week ago

1)

Analysts have access to in-region sensitive data (table1)

Hence, analysts in RegionA have access to sensitive data in regionA.

Sensitive data in regionA is Financial and PII data (table2)

Financial and PII data translates to CardOnFile and ContactEmail column (table3)

Hence, analysts in RegionA have access to CardOnFile and ContactEmail

Conclusion: Analysts in RegionA have access to sensitive data, so masking should be applied.

upvoted 1 times

✉ **wikicog** 11 months, 1 week ago

2)

Engineers have access to all numeric sensitive data (table1)

Hence, engineers in RegionC have access to numeric sensitive data in regionA.

Sensitive data in regionA is Financial and PII data (table2)

Financial and PII data translates to CardOnFile and ContactEmail column (table3)

Of these, only CardOnFile is numeric

Hence, engineers in RegionC have only access to CardOnFile in RegionA

Conclusion: no masking is required for the Height column, since engineers don't have access to it.

upvoted 1 times

✉ **wikicog** 11 months, 1 week ago

3)

Engineers have access to all numeric sensitive data (table1)

Hence, engineers in RegionB have access to numeric sensitive data in RegionB.

Sensitive data in RegionB is Financial, PII data and Medical (table2)

Financial, PII and Medical translates to CardOnFile, Height and ContactEmail columns (table3)

Of these, CardOnFile and Height are numeric.

Hence, engineers in RegionB have access to CardOnFile and Height

Conclusion: Engineers in RegionB have access to Height column in Patients\_RegionB, so masking should be applied

upvoted 1 times

Question #10

## 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.

Assign a managed identity to Server1.

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.



## 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.

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 1 year, 8 months ago

Answer is right!

upvoted 15 times

Liz42 Highly Voted 1 year, 8 months ago

Shouldn't the last two be switched? Enable TDE then configure the key?

upvoted 7 times

淘宝店铺：<https://shop63989109.taobao.com/>

□ **anto69** 1 year, 4 months ago

I also think so, but not sure

upvoted 1 times

□ **noranathalie** 1 year, 7 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 8 times

□ **shoottheduck** 3 months, 1 week ago

Checked this link and it supports the answer given

upvoted 2 times

□ **hanzocuk** [Most Recent] 5 months, 1 week ago

1. Get a KV

2. Add key to KV

3. Assign MI to server

4. Enable TDE

5. Config TDE

upvoted 2 times

□ **Okea** 4 months, 1 week ago

4 and 5 should be swapped

upvoted 1 times

□ **rzeng** 7 months, 2 weeks ago

correct

upvoted 1 times

□ **Deeksha1234** 9 months, 4 weeks ago

given ans is correct

upvoted 1 times

□ **dev2dev** 1 year, 4 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 1 year, 8 months ago

Correct!

upvoted 21 times

⊕  **juanlu46** Highly Voted 1 year, 1 month ago

**Selected Answer: B**

Correct!

upvoted 5 times

⊕  **anks84** Most Recent 9 months ago

**Selected Answer: B**

Correct !

upvoted 3 times

⊕  **Deeksha1234** 10 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>

✉ **mshakir** Highly Voted 1 year, 8 months ago

Answer is Correct according to given link  
upvoted 12 times

✉ **Canary\_2021** Highly Voted 1 year, 5 months 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 11 times

✉ **esaade** Most Recent 2 months, 4 weeks 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 2 times

✉ **shoottheduck** 3 months, 1 week ago

**Selected Answer: A**

Correct

upvoted 2 times

✉ **Deeksha1234** 10 months ago

**Selected Answer: A**

A is correct

upvoted 2 times

✉ **Doty** 1 year ago

As A is correct

upvoted 3 times

✉ **BerendJan** 1 year, 8 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** 1 year, 7 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 2 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 7 months, 2 weeks ago

**Selected Answer: B**

correct B

upvoted 7 times

 **anks84** Highly Voted 9 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
```

<b>.bucketBy</b> <b>.format</b> <b>.partitionBy</b> <b>.sortBy</b>	<b>(“*”)</b> (“GeographyRegionID”) (“GeographyRegionID”, “Year”, “Month”, “Day”) (“Year”, “Month”, “Day”, “GeographyRegionID”)
---	---

<b>.mode (“append”)</b>	<b>.csv(“/DBTBL1”)</b> <b>.json(“/DBTBL1”)</b> <b>.parquet(“/DBTBL1”)</b> <b>.saveAsTable(“/DBTBL1”)</b>
-------------------------	---

**Answer Area**

```
df.write
```

Correct Answer:

<b>.bucketBy</b> <b>.format</b> <b>.partitionBy</b> <b>.sortBy</b>	<b>(“*”)</b> (“GeographyRegionID”) (“GeographyRegionID”, “Year”, “Month”, “Day”) (“Year”, “Month”, “Day”, “GeographyRegionID”)
---	---

<b>.mode (“append”)</b>	<b>.csv(“/DBTBL1”)</b> <b>.json(“/DBTBL1”)</b> <b>.parquet(“/DBTBL1”)</b> <b>.saveAsTable(“/DBTBL1”)</b>
-------------------------	---

Box 1: .partitionBy -

Incorrect Answers:

- .format:

Method: format():

淘宝店铺：<https://shop63989109.taobao.com/>

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 1 year, 5 months ago

1. Partition by
2. GeographyRegionID, Year, Month, Day as the pipelines are per region this seems right choice
3. Parquet

upvoted 69 times

□ **uzairahm** 11 months, 2 weeks ago

regarding point 2 Solution needs to support daily incremental load so having Year, Month, Day first would be more useful  
upvoted 3 times

□ **petilda** Highly Voted 1 year, 9 months ago

I suggest storing the data in parquet  
upvoted 48 times

□ **JosephVishal** Most Recent 5 months, 3 weeks ago

For 3.) if parquet with partitions, then it should "overwrite" mode instead of "append". Since, it is "append" mode, I think saveAsTable sis more appropriate.  
upvoted 1 times

□ **Deeksha1234** 10 months ago

Agree with Pallavi  
1. Partition by  
2. GeographyRegionID, Year, Month, Day  
3. Parquet  
upvoted 5 times

□ **OldSchool** 6 months, 2 weeks 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

□ **OldSchool** 6 months, 1 week ago

Disregard my comment on 2). Provided answer is the correct one.

upvoted 1 times

□ **dsp17** 11 months ago

Parquet is must (offer higher compression rates)- "The solution must minimize storage costs."  
upvoted 2 times

□ **Aurelkb** 11 months, 2 weeks ago

it is the same question on Topic 1 Question 36.  
Then  
1. Partition by  
2. GeographyRegionID, Year, Month, Day  
3. Parquet  
upvoted 7 times

□ **Backy** 1 year 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** 11 months, 2 weeks ago

Agree, but if you choose the first one, you won't have the daily data  
upvoted 1 times

□ **allagowf** 7 months, 2 weeks ago

no mentioning for daily data in the question  
upvoted 2 times

□ **Spinozabubble** 3 weeks, 4 days ago

daily incremental load pipelines  
upvoted 1 times

□ **Amsterliese** 1 year, 1 month 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, 4 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 Parquet are answers

upvoted 8 times

□ **bad\_atitude** 1 year, 5 months ago

Mes chers amis:  
1.Sortby  
2.GeographyRegionId, Year, Month, Day  
3.Parquet  
upvoted 6 times

□ **jv2120** 1 year, 5 months ago

only reason for using .parquet is option seems to be dataset path not table else saveable is right.  
upvoted 2 times

□ **Aslam208** 1 year, 7 months ago

I agreed with @hrynewka, saveAsTable takes db name and table name not path.  
upvoted 4 times

□ **hrynewka** 1 year, 7 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, 2 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** 1 year, 8 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

□ **A1000** 1 year, 8 months ago

saveAsTable is the right option

DataFrameWriter.format(args)  
.option(args)  
.bucketBy(args)  
.partitionBy(args)  
.save(path)

DataFrameWriter.format(args).option(args).sortBy(args).saveAsTable(table)  
upvoted 1 times

□ **kimalto452** 1 year, 8 months ago

nop nop and nop  
upvoted 10 times

MoDar 1 year, 9 months ago

淘宝店铺：<https://shop63989109.taobao.com/>

saveAsTable() creates a permanent, physical table stored in S3 using the Parquet format  
upvoted 4 times

yyhhh 9 months, 2 weeks ago

I agree with you. the default format is Parquet when .format() is ignored.  
ref: <https://www.oreilly.com/library/view/learning-spark-2nd/9781492050032/ch04.html>  
- "If you don't specify this method, then the default is Parquet or whatever is set in spark.sql.sources.default."  
- saveAsTable() : The table to save to.  
upvoted 1 times

yyhhh 9 months, 2 weeks ago

.saveAsTable() is used in spark. for df, the answer is .parquet('...')  
upvoted 1 times

SaferSephy 1 year, 9 months ago

Also for the partitions, i'd say do {sourcetype}always before yyymmdd so you can easily isolate it for security purposes  
upvoted 6 times

## Question #15

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>

✉  **alexleonvalencia** Highly Voted 1 year, 6 months ago

**Selected Answer: AC**

Respuesta A/C

upvoted 14 times

✉  **VJPR** 1 year, 5 months ago

why not RBAC?

upvoted 6 times

✉  **sensaint** 5 months, 1 week 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

✉  **zizonesol** 2 months, 4 weeks 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 2 times

✉  **lukeonline** Highly Voted 1 year, 5 months ago

**Selected Answer: AB**

A and B

upvoted 14 times

✉  **mamahani** Most Recent 1 month, 1 week 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 Apache 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** 2 months, 4 weeks ago

**Selected Answer: BC**

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** 2 months, 2 weeks ago

ChatGPT isn't always reliable  
upvoted 4 times

 **janaki** 1 week, 6 days 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 1 times

 **AHUI** 3 months 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** 3 months, 2 weeks ago

**Selected Answer: B**

The answer is B  
upvoted 1 times

 **yogiazzaad** 4 months, 4 weeks ago

Given answer is correct.

Below from Microsoft documentation:

"A multi-tenant application can create a policy to enforce a logical separation of each tenant's data rows from every other tenant's rows. Efficiencies are achieved by the storage of data for many tenants in a single table. Each tenant can see only its data rows."  
<https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16>

upvoted 1 times

 **Taou** 5 months, 1 week ago

**Selected Answer: AC**

A and C must go together, so i think the right answer is AC  
upvoted 1 times

 **juamd** 6 months ago

According to Microsoft documentation:

```
CREATE SECURITY POLICY SalesFilter
ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)
ON Sales.Orders
WITH (STATE = ON);
GO
```

So the answer are A and C

upvoted 5 times

 **AzureJobsTillRetire** 6 months, 1 week ago

**Selected Answer: AB**

Given answer is correct.

D & E are obviously wrong.

C (in Row Level Security) is not necessary and may not be the right solution either. The best way to secure data is not to allow users to access the data at all. For example, we can store data in different databases or schemas and use RBAC to control user access. Row level security first gives users access to the data (in the table that contains all the data for all users) and then restrict data access to a particular part of the table. This is always less secure than not giving user access to the tables that do not contain any data the user should not have access to. Furthermore, Row Level Security may be breached by guessing work queries. I have done that before and I'm quite confident that I can breach any Row Level Security in SQL database but do not want to elaborate here.

upvoted 4 times

 **AzureJobsTillRetire** 5 months, 3 weeks ago

Please disregard my previous comments. For purpose of the exam, the answer to the question is AC.

upvoted 1 times

 **dmitriypo** 7 months, 1 week ago

**Selected Answer: AC**

```
CREATE SECURITY POLICY SalesFilter
ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)
```

```
ON Sales.Orders  
WITH (STATE = ON);  
GO
```

淘宝店铺：<https://shop63989109.taobao.com/>

<https://learn.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=azure-sqldw-latest>  
upvoted 1 times

□ **allagowf** 8 months ago

**Selected Answer: AB**

for those who select C : this function is to apply RLS on inserting and updating, but not selecting so other user can read the data.

[https://azure.microsoft.com/en-gb/blog/sql-database-row-level-security-block-predicates/#:~:text=Block%20predicates%20address%20a%20common,SQL%20Database%20\(V12\)%20server.](https://azure.microsoft.com/en-gb/blog/sql-database-row-level-security-block-predicates/#:~:text=Block%20predicates%20address%20a%20common,SQL%20Database%20(V12)%20server.)  
upvoted 5 times

□ **k18585** 6 months, 3 weeks ago

the link you provided talks about BLOCK predicates, the option in this question says "filter predicates". And in the article you linked it states "Whereas filter predicates apply to read operations, block predicates apply to write operations". Please don't give wrong answer/reference  
upvoted 2 times

□ **Phund** 9 months ago

**Selected Answer: AB**

scope is database in pool not table in database  
upvoted 4 times

□ **Deeksha1234** 10 months ago

**Selected Answer: AC**

Implement RLS by using the CREATE SECURITY POLICYTransact-SQL statement, and predicates created as inline table-valued functions.

<https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16#CodeExamples>  
upvoted 1 times

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

**Selected Answer: AC**

ANSWER IS A AND C  
upvoted 1 times

□ **Remedios79** 11 months, 2 weeks ago

a and c! C because RLS is implemented by the creation of a function on the column you want secure.  
upvoted 2 times

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

**Selected Answer: AC**

<https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security?view=sql-server-ver16#CodeExamples>  
upvoted 2 times

## Question #16

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 1 year, 5 months ago

I think it's a terrible question, both A(using T-SQL) and B (via GUI) can do the job.

upvoted 16 times

✉ **rzeng** Highly Voted 7 months, 2 weeks 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 8 times

✉ **Shanmahi** Most Recent 6 months ago

**Selected Answer: B**

email masking option via ssms

upvoted 1 times

✉ **OldSchool** 6 months, 3 weeks 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** 8 months, 3 weeks ago

B correct

upvoted 1 times

✉ **Deeksha1234** 10 months ago

both A and B are correct

upvoted 1 times

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

**Selected Answer: A**

Occams razor with this one

upvoted 1 times

✉ **Glen711** 10 months, 4 weeks 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 2 times

 **StudentFromAus** 11 months, 2 weeks ago

淘宝店铺 : <https://shop63989109.taobao.com/>

The answer should be B as it's not the default email mask format.  
upvoted 1 times

 **Navthing** 11 months, 3 weeks ago

**Selected Answer: A**  
Both A & B are correct But I will prefer A.  
upvoted 1 times

 **NamitSehgal** 11 months, 3 weeks ago

Sorry A is my proffered way, I can not edit my earlier comment.  
upvoted 1 times

 **NamitSehgal** 11 months, 3 weeks ago

Both A and B are correct  
For SQLMI, it can not be done for portal from for Azure SQL and Azure Synapse, we just need to click Add Mask for one or more columns and then Save to apply a mask for these fields.  
Yes, using SSMS is my preferred way as we have handoff policy from portal, all should be automated. So I go for B.  
upvoted 1 times

 **AIcubeHead** 1 year, 2 months ago

**Selected Answer: A**  
However, you can also do it with B as you can do an ALTER TABLE statement and an email mask is just a custom text function of a specific format  
upvoted 1 times

 **dev2dev** 1 year, 4 months ago

**Selected Answer: B**  
both B and A are correct  
upvoted 1 times

 **adel182ff** 1 year, 4 months ago

From the Azure portal, set a mask on the Email column.  
upvoted 1 times

 **Canary\_2021** 1 year, 5 months ago

Both A and B mask email as aXX@XXXX.com. How to mask email as XXX@XXXX.com?  
upvoted 1 times

**Both A and B mask email as aXX@XXXX.com by default.**

**if we don't want the default format for data masking and want to change the format, we can click on the masking rule applied -> select Masking Field Format as custom string -> update and save.**  
Both A and B support it. Which one should be selected?  
upvoted 2 times

 **PallaviPatel** 1 year, 5 months ago

we want to set email mask on email column hence option A is correct than option B.  
upvoted 4 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, 4 months ago

**Selected Answer: D**

D is the way we do it in our company. So it works at least.

upvoted 8 times

 **PallaviPatel** Highly Voted 1 year, 5 months ago

the answer and explanation given is correct.

upvoted 7 times

 **janaki** Most Recent 1 week, 6 days ago

Vnet = managed identity

upvoted 2 times

 **yogiazaad** 4 months, 4 weeks 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** 10 months ago

**Selected Answer: D**

yes, correct

upvoted 2 times

 **EmmettBrown** 1 year, 1 month ago

**Selected Answer: D**

Managed identity is correct

upvoted 3 times

 **anto69** 1 year, 4 months ago

I too I think is correct, anyway for sure it's possible

upvoted 4 times

 **bad\_atitude** 1 year, 5 months ago

I believe so

upvoted 4 times

**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** Highly Voted 1 year, 5 months ago

淘宝店铺：<https://shop63989109.taobao.com/>

Correct Answer

upvoted 17 times

□ **HaBroNounen** Highly Voted 1 year, 5 months 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 10 times

□ **janaki** 1 week, 6 days ago

You are correct! :-)

upvoted 2 times

□ **vrodriguesp** Most Recent 3 months, 3 weeks 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

□ **Deeksha1234** 10 months ago

correct

upvoted 1 times

□ **Remedios79** 11 months, 2 weeks ago

Correct!

upvoted 1 times

□ **juanlu46** 1 year, 1 month ago

Totally correct!

upvoted 1 times

Question #19

**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:

<input type="checkbox"/>	▼
<input checked="" type="checkbox"/>	▼
<input type="checkbox"/>	▼

Premium  
Standard

Advanced option to enable:

<input type="checkbox"/>	▼
<input checked="" type="checkbox"/>	▼
<input type="checkbox"/>	▼

Azure Data Lake Storage Credential Passthrough  
Table Access Control

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

Tier:

<input type="checkbox"/>	▼
<input checked="" type="checkbox"/>	▼
<input type="checkbox"/>	▼

Premium  
Standard

Advanced option to enable:

<input type="checkbox"/>	▼
<input checked="" type="checkbox"/>	▼
<input type="checkbox"/>	▼

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, 4 months ago

Correct

upvoted 10 times

**HaBroNounen** Highly Voted 1 year, 5 months ago

Provided answer is correct

<https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough>

upvoted 6 times

**anks84** Most Recent 9 months, 1 week ago

Given Answer is correct

upvoted 3 times

**Deeksha1234** 10 months ago

correct

upvoted 3 times

淘宝店铺：<https://shop63989109.taobao.com/>

edba 1 year, 5 months ago

I think answer is correct!

upvoted 3 times

Question #20

Topic 3

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] 1 year, 5 months ago

correct

upvoted 11 times

Jerrie86 [Most Recent] 4 months, 2 weeks ago

Whenever you see Vnet , answer is usually managed Identity

upvoted 3 times

anks84 9 months, 1 week ago

Correct, Managed Identity authentication is required when your storage account is attached to a VNet.

upvoted 4 times

Deeksha1234 10 months ago

correct

upvoted 1 times

ravi2931 1 year, 2 months ago

Correct

upvoted 1 times

alex1491 1 year, 2 months ago

the key here is virtual network. Correct!

upvoted 1 times

ANath 1 year, 4 months ago

Correct

upvoted 3 times

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 1 year, 5 months ago

Agree with the answer => B

upvoted 17 times

 **Deeksha1234** Most Recent 10 months ago

correct

upvoted 2 times

 **Remedios79** 11 months, 2 weeks ago

the key here is "limited time period", so SAS.

upvoted 3 times

 **juanlu46** 1 year, 1 month 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>

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 43 times

Billybob0604 5 months, 3 weeks ago

This is it. Correct

upvoted 1 times

edba 1 year, 4 months ago

To be more clear, for box it shall be user delegation SAS which is secured with ADD credentials.

upvoted 2 times

vivekazure Highly Voted 1 year, 4 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 13 times

gogosgh Most Recent 1 month 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

OldSchool 6 months, 1 week 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 6 months, 1 week 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 3 times

KR8055 7 months, 3 weeks 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 2 times

sunil\_smile 8 months, 3 weeks 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 4 times

vrodriguesp 4 months, 1 week 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 4 months, 1 week 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

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** 10 months ago

Given answer seems correct, agree with HaBroNounen's explanation

upvoted 1 times

□  **vishal10** 10 months, 2 weeks 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** 10 months, 3 weeks 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** 1 year, 5 months 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

□  **Canary\_2021** 1 year, 5 months 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** 1 year, 5 months 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** 1 year, 5 months 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

## Question #23

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 1 year, 5 months ago

correct

upvoted 11 times

✉  **borinot** Most Recent 6 months, 3 weeks ago

And Topic 3 question 24 is column-level encryption?

upvoted 2 times

✉  **vrodriguesp** 3 months, 3 weeks 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 2 times

✉  **Deeksha1234** 10 months ago

correct!

upvoted 1 times

✉  **juanlu46** 1 year, 1 month ago

Selected Answer: E

Correct!

upvoted 3 times

✉  **wwdba** 1 year, 3 months ago

Correct

upvoted 3 times

✉  **ANath** 1 year, 4 months ago

Correct

upvoted 4 times

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, 1 month 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 11 times

□  yogiazaad Most Recent 4 months, 4 weeks 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 1 times

□  yogiazaad 4 months, 4 weeks ago

IS column level encryption supported on Dedicated SQL Pools? The question is relate to Dedicated Pool?

upvoted 1 times

□  Deeksha1234 10 months ago

correct

upvoted 1 times

## Question #25

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 6 months, 3 weeks 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 10 times

⊕  **yogiazaad** 4 months, 3 weeks 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 1 times

⊕  **bokLuci** Highly Voted 7 months, 2 weeks 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 8 times

⊕  **esaade** Most Recent 2 months, 4 weeks 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 1 times

 **yogiazaad** 4 months, 3 weeks 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 3 times

 **AzureJobsTillRetire** 6 months, 1 week 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 4 times

 **Deeksha1234** 10 months ago

**Selected Answer: DF**

so the answer is correct

upvoted 2 times

 **Deeksha1234** 10 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 3 times

 **Davico93** 11 months, 3 weeks 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 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** 12 months ago

Please ignore this. It's not correct. Examtopics should provide a delete option here.

upvoted 4 times

 **sdokmak** 1 year ago

Following principal 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 3 times

 **virendrapsingh** 1 year ago

Agreed with your comment on least privilege as it is mentioned specifically in the question.

Choices A & F should be the answer.

upvoted 4 times

 **Aditya0891** 1 year 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 ago

**Selected Answer: DF**

Correct

upvoted 1 times

 juanlu46 1 year, 1 month ago

**Selected Answer: DF**

Is correct!

upvoted 2 times

店铺: 学习小店66

店铺: 学习小店66

店铺: 学习小店66

店铺: 学习小店66

**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** Highly Voted  1 year, 5 months ago

Correct

upvoted 19 times

  **Amsterliese** Highly Voted  1 year, 1 month 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 5 times

□ **Deeksha1234** Most Recent 10 months, 1 week ago

answer is correct

upvoted 3 times

□ **dev2dev** 1 year, 4 months ago

B is wrong. Contained users not supported by synapse analytics. D is correct ('MS SQL Server logins')

upvoted 3 times

□ **PallaviPatel** 1 year, 4 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 14 times

□ **dev2dev** 1 year, 4 months ago

correct

upvoted 1 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:

<b>Actions</b>	<b>Answer Area</b>
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.	

<b>Actions</b>	<b>Answer Area</b>
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.
Correct Answer: 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 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 151 times

✉️  **datapc** 7 months, 1 week ago

<https://learn.microsoft.com/en-us/azure/azure-monitor/essentials/tutorial-resource-logs?source=recommendations>

Above order is mentioned here.

upvoted 3 times

✉️  **RajashekharC** 9 months, 3 weeks ago

This is correct order, I have tried this on Azure portal.

upvoted 3 times

✉️  **Deeksha1234** 10 months, 1 week ago

seems correct to me

upvoted 1 times

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

Shouldn't it need to swap step 3 & 4?

upvoted 3 times

✉️  **herculian\_effort** Highly Voted 1 year, 11 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

✉️  **Igor85** 6 months, 1 week ago

steps 2 & 3 must be swapped. you can't send data to log analytics workspace that isn't created yet

upvoted 1 times

✉️  **KashRaynardMorse** 1 year, 1 month 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 2 times

✉️  **Armandoo** 1 year, 10 months ago

This is the correct answer

upvoted 1 times

✉️  **LiLy91** 1 year, 5 months ago

Don't you have to select PipelineRuns Category while adding a diagnostic setting?  
upvoted 1 times

□ **Sriramiyer92** [Most Recent] 10 months, 2 weeks 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** 11 months, 3 weeks 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

□ **steeee** 1 year, 9 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 11 times

□ **Amalbenrebai** 1 year, 9 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 8 times

□ **mss1** 1 year, 10 months ago

If you create diagnostics from the Datafactory you wil 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

To complete my answer. I also agree with "Sunnyb". There are more solutions to this question.

upvoted 2 times

□ **Marcus1612** 1 year, 8 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** 1 year, 11 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 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 ago

regarding the storage account, what is it for?  
upvoted 1 times

□ **sagga** 2 years 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, 1 month 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

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

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 1 year, 11 months ago

Correct!

upvoted 38 times

 **Deeksha1234** Most Recent 10 months, 1 week ago

**Selected Answer: B**

B is correct

upvoted 2 times

 **youlitai003** 1 year, 1 month 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** 6 months ago

you can only enable double encryption when you are creating a new workspace.

upvoted 2 times

 **djblue** 1 year, 3 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** Highly Voted 9 months ago

-Execute  
-Execute  
-Write  
upvoted 5 times

 **bakamon** Most Recent 1 week, 6 days 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** 6 months 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 4 times

□ **mamahani** 1 month, 1 week 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

□ **dom271219** 9 months ago

Correct : Execute to traverse the folders and Write to append the file

upvoted 3 times

Question #30

**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  9 months ago

I would say:

1. sql - this is correct
2. managed hive metastore

upvoted 5 times

淘宝店铺：<https://shop63989109.taobao.com/>

✉  **federc** 9 months ago

scrath that, given anwers are correct. sql + hive metastore

upvoted 11 times

✉  **TestingCRM** Most Recent 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.

淘宝店铺：<https://shop63989109.taobao.com/>

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 8 months, 3 weeks 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 25 times

□  **mamahani** 1 month, 1 week 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** 7 months, 2 weeks ago

Agree no mention for tiering in the question so LRS is the best option to minimize the cost

upvoted 3 times

□  **dom271219** Highly Voted 9 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 12 times

□  **BPW** Most Recent 2 weeks, 6 days 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 1 times

□  **mamahani** 1 month, 1 week 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** 1 month, 1 week 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** 5 months, 2 weeks ago

I strongly doubt they didn't offer the whole question. The question is not clear.

upvoted 1 times

□  **AzureJobsTillRetire** 6 months, 1 week 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

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

**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.

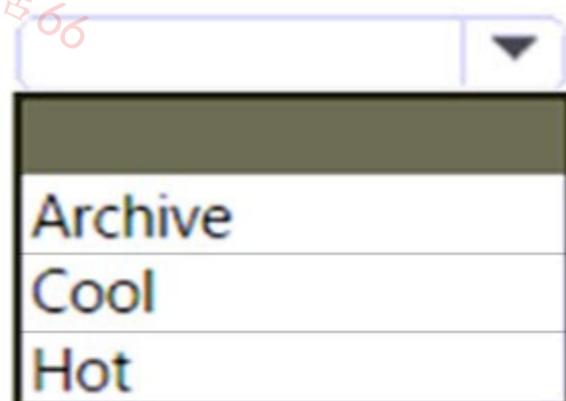
You need to recommend a data retention solution. The solution must minimize costs.

Which access tier should you recommend for each time frame? To answer, select the appropriate options in the answer area.

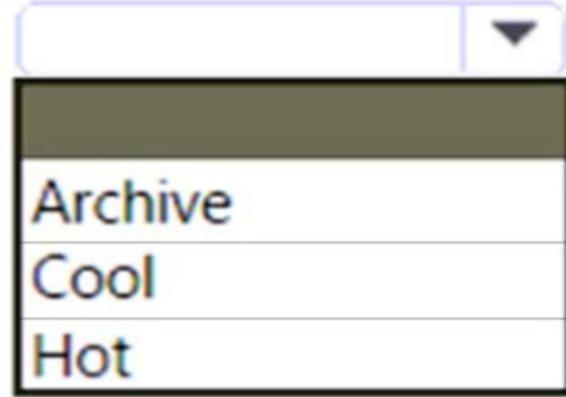
NOTE: Each correct selection is worth one point.

Hot Area:

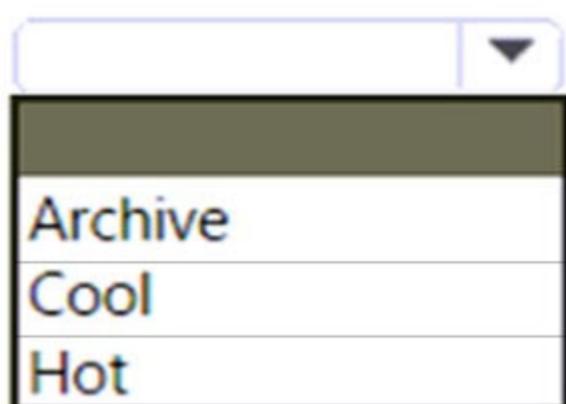
First 30 days:



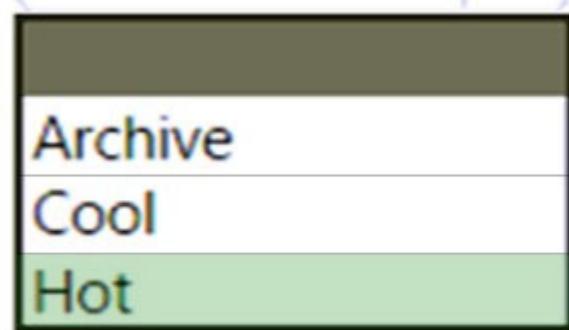
After 90 days:



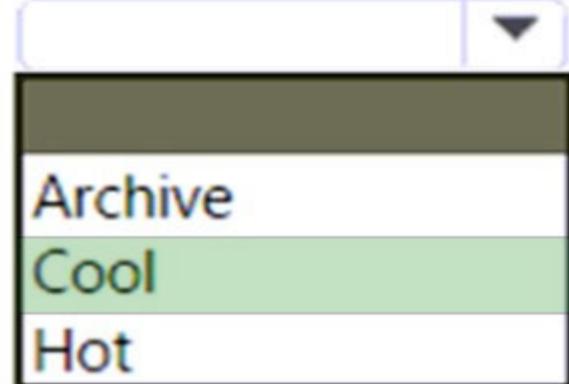
After 365 days:



First 30 days:



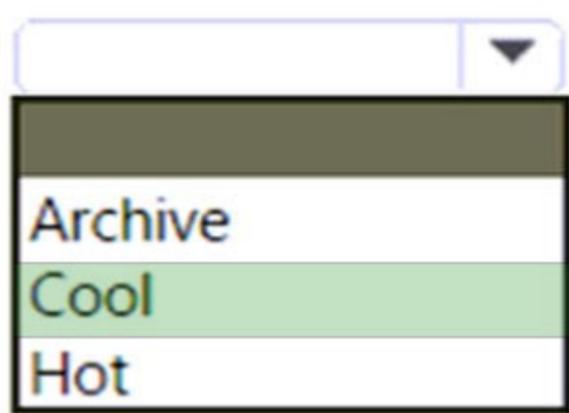
After 90 days:



Correct Answer:

店铺：学习小店66

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 4 months, 2 weeks ago

Hot, Cool, Cool is correct.

Ref: <https://learn.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview>

upvoted 6 times

曰  **Sima\_al** Most Recent 4 months, 3 weeks ago

淘宝店铺 : <https://shop63989109.taobao.com/>

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 1 times

曰  **shoottheduck** 3 months, 1 week ago

Cool has a response time of Milliseconds. So Hot, Cool, Cool

upvoted 4 times

曰  **gabrys1997** 8 months, 2 weeks ago

I think that 'cool' tier is just enough, it provides availability on 99.9%

upvoted 2 times

曰  **hanzocuk** 5 months, 1 week 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 2 times

曰  **Marcohcm** 8 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

曰  **Strix** 9 months, 1 week 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**

- Access control lists (ACLs)
- Role-based access control (RBAC) roles
- Shared access signatures (SAS)
- Shared account keys

**Answer Area**

To grant permissions at the storage account level:

To grant permissions at the object level:

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

To grant permissions at the storage account level:

To grant permissions at the object level:

 **SannPro** Highly Voted 4 months, 1 week ago

Correct

upvoted 6 times

 **aemilka** Most Recent 1 month, 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 3 times

 **Venub28** 4 months, 3 weeks ago

Given answer is correct

upvoted 3 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** 1 month, 1 week 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 3 times

✉  **AHUI** 2 months 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** 1 month, 3 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 6 times

✉  **shakes103** 1 month, 4 weeks ago

I have looked it up too. Answer is C  
 upvoted 2 times

✉  **aemilka** 1 month, 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 4 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	
<b>--X permissions</b>	

Correct Answer:

Directory1:

	▼
--- permissions	
-WX permissions	
<b>--X permissions</b>	

File1:

	▼
--- permissions	
<b>-WX permissions</b>	
--X permissions	

 **BPW**  1 month, 3 weeks ago

Answer is

--x/ -wx/ ---

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

upvoted 13 times

□  **mamahani** Most Recent 1 month, 1 week ago

淘宝店铺 : <https://shop63989109.taobao.com/>

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 2 times

□  **Ahmad\_Abukhater** 2 months ago

last box file1 should be --- (First option)

<https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control>

upvoted 4 times

□  **DataEngineer7331** 1 month, 3 weeks ago

According to this your Link, the Directory should have "-WX" and the File "---"

upvoted 1 times

Question #36

Topic 3

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**

□  **xymtyk** 1 month, 1 week ago

**Selected Answer: B**

Correct.

upvoted 1 times

□  **Iamthealpha** 1 month, 1 week 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 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**

 **RoyP654** 2 weeks, 2 days 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

 **lamthealpha** 1 month, 1 week 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 4 times

 **shakes103** 1 month, 4 weeks 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 4 times

 **halamgir15** 1 month, 4 weeks ago

I think it should be D:

dynamic data masking

upvoted 3 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)

kim32 4 weeks ago

It should be Row Level security not column since limited for some users  
upvoted 1 times

francocalvo 3 weeks, 5 days 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 3 times

haythemsi 1 month ago

Correct

upvoted 3 times

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 1 month ago

Given answer is correct.

Refer this link - <https://learn.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables>  
upvoted 5 times

#### Topic 4 - Question Set 4

## Question #1

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  1 year, 11 months ago

correct B

upvoted 36 times

✉  **aortega** 1 year, 8 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 4 times

✉  **temmytak** Most Recent  2 months ago

**Selected Answer: B**

Correct B

upvoted 2 times

✉  **Shanmahi** 6 months 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 1 times

淘宝店铺：<https://shop63989109.taobao.com/>

□ **dmitriypo** 7 months ago

**Selected Answer: B**

B is correct

upvoted 1 times

□ **Deeksha1234** 10 months ago

**Selected Answer: B**

yes, B is correct

upvoted 1 times

□ **Remedios79** 11 months, 2 weeks ago

correct

upvoted 1 times

□ **Lily91** 1 year, 4 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** 1 year, 5 months 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** 1 year, 8 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** 1 year, 11 months ago

Incoming queries use the primary key SaleKey column to retrieve data as displayed in the following table ..doesnt 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]** 1 year, 7 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** 11 months, 4 weeks 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** 1 year, 11 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 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

淘宝店铺：<https://shop63989109.taobao.com/>

Preben 2 years 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

## Question #2

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 1 year, 11 months ago

correct!

upvoted 26 times

 **Miris** Highly Voted 2 years ago

correct

upvoted 13 times

 **mamahani** Most Recent 1 month, 1 week 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 2 times

 **mamahani** 1 month, 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 1 times

 **AHUI** 2 months ago

Selected Answer: B

correct

upvoted 2 times

 **Rakrah** 4 months 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 2 times

 **OldSchool** 6 months, 1 week 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\_** 8 months, 2 weeks 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 9 months ago

淘宝店铺：<https://shop63989109.taobao.com/>

**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 9 months, 1 week 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

proserv 9 months, 2 weeks ago

It's Option C , Non Clustered .

Reason : Heap table is a table without clustered columns store index and we can create non cluster index on heap table. Furthermore , there is not non cluster column store index exist

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-index#heap-tables>

Thanks

upvoted 2 times

Deeksha1234 10 months ago

correct answer

upvoted 1 times

dsp17 11 months ago

**Selected Answer: B**

It is clearly Clustered columnstore index

upvoted 2 times

NamitSehgal 11 months, 2 weeks ago

Clustered Columnstore Index (CCI) A clustered columnstore index (CCI) is usually the best choice providing optimal query performance for almost large tables. By default, Synapse Analytics creates a clustered columnstore index (CCI), when no index options are specified.  
So here I would prefer a non cluster index on those two columns.

upvoted 1 times

StudentFromAus 11 months, 2 weeks ago

**Selected Answer: B**

It is clearly Clustered columnstore index

upvoted 2 times

Aditya0891 1 year ago

I found this and it states that clustered index would be best suited for this case. <https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet> . Please correct me if I'm wrong

upvoted 2 times

mamahani 1 month, 1 week ago

I believe you are correct; according to this cheat sheet from microsoft in this particular case (table up to 100 mln records with heavily used 1-2 columns and using a lot of aggregations) the best option recommended by microsoft is clustered index; if i get this question in the exam, that will be my answer ; thanks for the link btw; very very useful!

upvoted 1 times

Aditya0891 1 year ago

I don't know how are people are opting for nonclustered columnstore index. It's clearly mentioned the type of indexes supported in dedicated SQL pool - "Dedicated SQL pool offers several indexing options including clustered columnstore indexes, clustered indexes and nonclustered indexes, and a non-index option also known as heap."

upvoted 3 times

vrodriguesp 4 months, 3 weeks ago

Yes, there is no 'nonclustered columnstore index' at all. I can't understand

upvoted 1 times

sdokmak 1 year ago

**Selected Answer: C**

The question's not clear, is it just the index type or the queries as well?

If queries can change, then it might be A: Clustered Columnstore, but we also don't know how many columns we are using to query, if it's just one column out of 50, then yes A: <https://www.spotlightcloud.io/blog/when-to-use-clustered-or-non-clustered-indexes-in-sql-server>

Otherwise, it's C: nonclustered.

B's not an option for Synapse as mentioned in other discussions.

upvoted 1 times

## Question #3

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 1 year, 7 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 28 times

 **dragos\_dragos62000** Highly Voted 1 year, 11 months ago

Correct

upvoted 9 times

 **aemilka** Most Recent 1 month, 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 1 times

 **kornat** 2 months ago

**Selected Answer: C**

correct

upvoted 1 times

 **esaade** 2 months, 4 weeks 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 2 times

 **lafita** 4 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 1 times

淘宝店铺：<https://shop63989109.taobao.com/>

**youngbug** 4 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 1 times

**gerrie1979** 7 months 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** 7 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 1 times

**Raghul08** 1 year, 4 months ago

My Answer is B

upvoted 1 times

**edba** 1 year, 5 months ago

I think answer is B - Cluster Event logs. Because there are 3 ways to install a new library (<https://docs.microsoft.com/en-us/azure/databricks/libraries/cluster-libraries#--install-a-library-on-a-cluster>), using init script is just one of them.

upvoted 6 times

**Canary\_2021** 1 year, 5 months ago

B 'Cluster event logs' is the correct answer.

upvoted 5 times

**rashjan** 1 year, 6 months ago

**Selected Answer: B**

I would go with Cluster Event Logs.

upvoted 5 times

**Sudheer\_K** 1 year, 8 months ago

Shouldn't it be cluster-scoped Init scripts rather than global init scripts.

upvoted 3 times

**gf2tw** 1 year, 6 months ago

Yep, the answer even specifies cluster-scoped init scripts so it seems somehow the question doesn't match up.

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 1 year, 11 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 26 times

✉ **snna4** 1 year, 5 months 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 6 times

✉ **damaldon** Highly Voted 1 year, 11 months ago

Correct!

upvoted 6 times

✉ **dmitriypo** Most Recent 7 months ago

**Selected Answer: D**

Agree with D

upvoted 2 times

✉ **Deeksha1234** 10 months ago

correct

upvoted 1 times

✉ **KrishIC** 1 year, 6 months ago

**Selected Answer: D**

CORRECT

upvoted 3 times

✉ **FredNo** 1 year, 6 months ago

**Selected Answer: D**

Correct

upvoted 3 times

✉ **Jayant68** 1 year, 6 months ago

Correct..

upvoted 2 times

## Question #5

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 1 year, 11 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 17 times

□  **Deeksha1234** Most Recent 10 months ago

correct

upvoted 2 times

□  **juanlu46** 1 year, 1 month 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 1 year, 6 months ago

**Selected Answer: A**

Correct

upvoted 8 times

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

Correct answer!

upvoted 5 times

 **markpumc** Most Recent 3 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 3 times

 **Deeksha1234** 10 months ago

correct

upvoted 1 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 1 year, 8 months ago

Correct answer!

upvoted 19 times

✉  **AHUI** Most Recent 2 months ago

**Selected Answer: C**

agreed

upvoted 2 times

✉  **Deeksha1234** 10 months ago

correct

upvoted 1 times

✉  **juanlu46** 1 year, 1 month ago

**Selected Answer: C**

Is correct!

upvoted 2 times

✉  **aortega** 1 year, 8 months ago

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/resource-classes-for-workload-management>

upvoted 2 times

## Question #8

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 1 year, 8 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 26 times

✉  **wijaz789** Highly Voted 1 year, 9 months ago

-- Find data skew for a distributed table  
DBCC PDW\_SHOWSPACEUSED('dbo.FactInternetSales');

upvoted 16 times

✉  **ItHYMeRish** 1 year, 5 months 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 7 times

✉  **aemilka** Most Recent 1 month, 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 1 times

✉  **Okea** 4 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** 4 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 4 times

✉  **youngbug** 4 months, 3 weeks ago

A is a quicker way, but you can run DBCC in a serverless SQL pool, the built-in pool.

upvoted 1 times

淘宝店铺：<https://shop63989109.taobao.com/>

□ **steve7** 4 months, 3 weeks ago

Right answer is A. DBCC PDW\_SHOWSPACEUSED. google it

upvoted 1 times

□ **Deeksha1234** 10 months ago

**Selected Answer: D**

D is correct

upvoted 3 times

□ **Franz58** 10 months, 2 weeks 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** 11 months, 2 weeks 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, 1 month 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

□ **Felix1** 1 year, 1 month ago

**Selected Answer: A**

DBCC PDW\_SHOWSPACEUSED

upvoted 1 times

□ **AlCubeHead** 1 year, 2 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, 2 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

□ **AlCubeHead** 1 year, 2 months ago

The question specifies dedicated Pool NOT Built-in Pool, so it is A

upvoted 1 times

□ **Amsterliese** 1 year, 1 month 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

□ **ovokpus** 1 year, 3 months ago

**Selected Answer: A**

This is right from the learning material

<https://docs.microsoft.com/en-us/learn/modules/analyze-optimize-data-warehouse-storage-azure-synapse-analytics/2-understand-skewed-data-space-usage>

upvoted 2 times

□ **kilowd** 1 year, 4 months ago

**Selected Answer: D**

Use sys.dm\_pdw\_nodes\_db\_partition\_stats to analyze any skewness in the data.

upvoted 5 times

□ **kilowd** 1 year, 4 months ago

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>  
upvoted 3 times

 **LiLy91** 1 year, 4 months ago

**Selected Answer: A**

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute>  
upvoted 2 times

 **LiLy91** 1 year, 4 months ago  
Correction, the answer should be D

upvoted 3 times

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

**HOTSPOT -**

You need to collect application metrics, streaming query events, and application log messages for an Azure Databrick cluster.

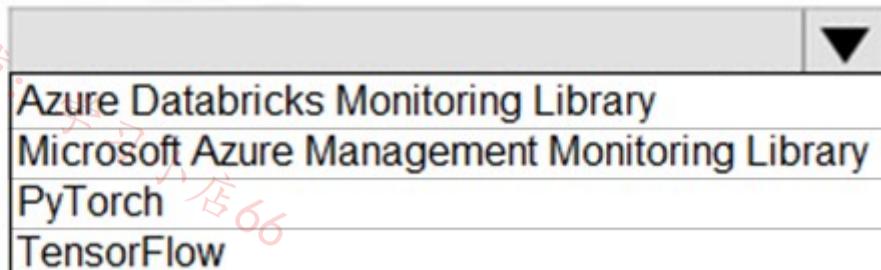
Which type of library and workspace should you implement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

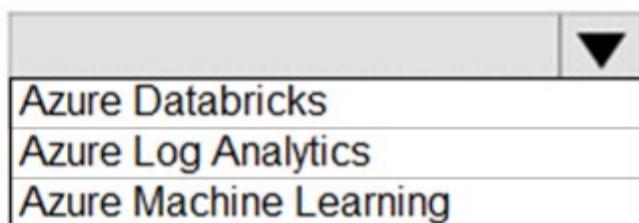
**Answer Area**

Library: 店铺



店铺：学习小店66

Workspace:

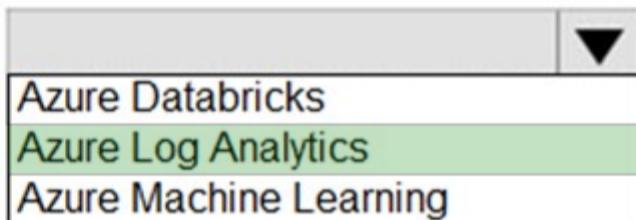
**Answer Area**

Correct Answer:

Library:



Workspace:



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>

□ **leandrors** Highly Voted 1 year, 7 months ago

Correct!

upvoted 7 times

□ **Igor85** Most Recent 6 months ago

the solution works for databricks runtime 10.x only, though.

newer version isn't supported yet

upvoted 2 times

□ **dmitriypo** 7 months ago

The given answer is correct

upvoted 1 times

□ **Deeksha1234** 10 months ago

Correct

upvoted 1 times

□ **wwdba** 1 year, 3 months ago

Correct!

upvoted 1 times

淘宝店铺：<https://shop63989109.taobao.com/>

 **Start** 1 year, 8 months ago

Answer is correct

<https://docs.microsoft.com/en-us/azure/architecture/databricks-monitoring/application-logs>

upvoted 3 times

 **MFO\_FM** 1 year, 8 months ago

is it correct

upvoted 1 times

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

## Question #10

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 1 year, 8 months ago

Correct Answer!  
upvoted 13 times

□  **Jerrie86** Most Recent 4 months, 2 weeks ago

Rollback works with transactions. answer B  
upvoted 2 times

□  **nicky87654** 4 months, 3 weeks ago

Selected Answer: B  
Correct Answer! B. sys.dm\_pdw\_nodes\_tran\_database\_transactions  
upvoted 2 times

□  **dmitriypo** 7 months ago

Selected Answer: B  
The given answer is correct  
upvoted 2 times

□  **allagowf** 7 months, 2 weeks ago

Selected Answer: B  
B. sys.dm\_pdw\_nodes\_tran\_database\_transactions  
upvoted 3 times

□  **Deeksha1234** 10 months ago

correct  
upvoted 1 times

□  **ladywhiteadder** 1 year, 2 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 1 times

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-cyrl-ba/azure/stream-analytics/stream-analytics-monitoring>

 **Lrng15** Highly Voted 1 year, 8 months ago

duplicate question. correct answer B

upvoted 15 times

 **Jerrie86** Highly Voted 4 months, 2 weeks ago

**Selected Answer: B**

Money is the answer to all problems. Answer B. increase SU units.

upvoted 7 times

 **yogiazaad** Most Recent 4 months, 2 weeks ago

<https://learn.microsoft.com/en-us/azure/stream-analytics/stream-analytics-job-metrics>

This link is useful

upvoted 1 times

 **Deeksha1234** 10 months ago

**Selected Answer: B**

correct!

upvoted 1 times

 **snna4** 1 year, 5 months ago

It's just a similar question. Proposed answers are different.

upvoted 1 times

 **Sudheer\_K** 1 year, 8 months ago

Repeated

upvoted 4 times

## Question #12

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 1 year, 8 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 22 times

 **yassine70** 1 year, 8 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** 1 year, 6 months 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 1 year, 5 months 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 filed should in where clause to improve query performance.

upvoted 13 times

 **Canary\_2021** 1 year, 5 months ago

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

upvoted 1 times

 **Karl\_Cen** Most Recent 4 months, 1 week ago

**Selected Answer: C**

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** 7 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** 7 months ago

Forget it. I agree with the provided answer D

upvoted 1 times

□ **dom271219** 9 months, 2 weeks ago

**Selected Answer: D**

Tables ? These are the columns, aren't they ?

upvoted 1 times

□ **Deeksha1234** 10 months ago

D is right

upvoted 1 times

□ **nefarious\_smalls** 1 year 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** 11 months, 4 weeks ago

It's about partition not distribution. Read the question carefully first

upvoted 2 times

□ **Dizzystar** 1 year, 7 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, 4 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** 1 year, 8 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** 1 year, 7 months ago

are you thinking of hash distributions instead of partitions?

upvoted 5 times

□ **rikku33** 1 year, 8 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** 1 year, 7 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** 1 year, 8 months ago

WHERE is applied on the WarehouseID, so D

upvoted 5 times

□ **YipingRuan** 1 year, 7 months ago

Nope, don't use WHERE

upvoted 2 times

□ **mbl** 1 year, 7 months ago

it does : "Analysts will most commonly analyze transactions for a warehouse"

upvoted 3 times

 **AppleVan** 1 year, 8 months ago

淘宝店铺：<https://shop63989109.taobao.com/>

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** 1 year, 8 months ago

can someone confirm this ?  
upvoted 1 times

 **Samanda** 1 year, 7 months ago

It's 100% D  
upvoted 3 times

 **rav009** 1 year, 8 months ago

I will go C  
upvoted 3 times

## Question #13

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 1 year, 9 months ago

Should be C and D

upvoted 55 times

 **anto69** 1 year, 4 months ago

Yup, that makes sense

upvoted 2 times

 **GervasioMontaNelas** Highly Voted 1 year, 9 months ago

100% CD

upvoted 12 times

 **BPW** Most Recent 2 weeks, 6 days ago

Should be A and E

upvoted 1 times

 **esaade** 2 months, 3 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** 1 month ago

we are not querying against time. the fact table has only dates

upvoted 1 times

 **XiltroX** 6 months, 1 week ago

For sure its CD

upvoted 1 times

 **Xinyuehong** 7 months, 3 weeks ago

**Selected Answer: CD**

CD with no doubt.

upvoted 1 times

 **Deeksha1234** 10 months ago

correct - C&D agree with StudentFromAus M

upvoted 2 times

 **StudentFromAus** 11 months, 2 weeks 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 3 times

 **Davico93** 11 months, 3 weeks ago

**Selected Answer: CD**

basic knowledge for fact and dim tables

upvoted 2 times

 **AIcubeHead** 1 year, 2 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 3 times

 **wwdba** 1 year, 2 months ago

Should be CD!

upvoted 1 times

 **Boumisasound** 1 year, 3 months ago

**Selected Answer: CD**

I'm agree for CD

upvoted 1 times

 **ovokpus** 1 year, 3 months ago

**Selected Answer: AE**

this makes the most sense

upvoted 2 times

 **kanak01** 1 year, 4 months ago

**Selected Answer: CD**

C & D should be correct

upvoted 1 times

 **bahamutedean** 1 year, 4 months ago

should be CD

upvoted 2 times

 **BusinessApps** 1 year, 5 months ago

**Selected Answer: CD**

Answer C and D

upvoted 3 times

 **alexleonvalencia** 1 year, 6 months ago

**Selected Answer: CD**

Respuesta correcta CD

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 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, 4 months ago

Repeated question

upvoted 10 times

 **PallaviPatel** Highly Voted 1 year, 4 months ago

**Selected Answer: C**

C is correct

upvoted 8 times

 **ZIMARAKI** Most Recent 4 months, 3 weeks ago

**Selected Answer: C**

C is correct

upvoted 2 times

 **Deeksha1234** 10 months ago

correct

upvoted 1 times

 **SabaJamal2010AtGmail** 1 year, 5 months ago

C is correct

upvoted 4 times

## Question #15

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>

✉  **yogiazzaad** 4 months, 2 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 1 times

✉  **Deeksha1234** 10 months ago

**Selected Answer: BD**

seems correct  
upvoted 3 times

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

**Selected Answer: BD**

Correct Answer  
upvoted 2 times

✉  **HaBroNounen** 1 year, 5 months 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. DWU limit
- 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 1 year, 5 months ago

correct

upvoted 5 times

 **Deeksha1234** Most Recent 10 months ago

correct

upvoted 3 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 1 year, 5 months ago

It shall be A:Clusters, workspace logs does not have any cluster related resource change.

upvoted 22 times

✉  **Ast999** Most Recent 3 months, 1 week ago

**Selected Answer: A**

100% SURE A IS A CORRECT ANSWER.

upvoted 2 times

✉  **Deeksha1234** 10 months ago

**Selected Answer: A**

A is correct

upvoted 4 times

✉  **demirsamuel** 1 year ago

**Selected Answer: A**

definitely A

upvoted 2 times

✉  **upliftinghut** 1 year 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 ago

I thought compute is related to cluster.

upvoted 2 times

✉  **KashRaynardMorse** 1 year 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 3 times

 **Deeksha1234** 10 months ago  
agree A should be the answer  
upvoted 1 times

淘宝店铺：<https://shop63989109.taobao.com/>

 **Amsterliese** 1 year, 1 month 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 1 times

 **ovokpus** 1 year, 3 months ago

**Selected Answer: A**  
Agreed with clusters!  
upvoted 2 times

 **kanak01** 1 year, 4 months ago

A clusters  
upvoted 1 times

 **svik** 1 year, 4 months ago

**Selected Answer: A**  
compute is related to the cluster  
upvoted 3 times

 **PallaviPatel** 1 year, 4 months ago

**Selected Answer: A**  
A is correct answer, as compute relates to clusters.  
upvoted 4 times

 **Canary\_2021** 1 year, 5 months ago

**Selected Answer: A**  
A should be correct answer.  
<https://www.examtopics.com/exams/microsoft/dp-200/view/17/>  
upvoted 4 times

 **Canary\_2021** 1 year, 5 months ago

What kind of changes belong to 'changes in compute for the Databricks resource'? Any example?  
upvoted 1 times

 **ItHYMeRish** 1 year, 5 months ago

**Selected Answer: B**  
The answer is correct. The workspace logs contain information about cluster events.  
<https://docs.databricks.com/administration-guide/account-settings/audit-logs.html#audit-events>  
upvoted 4 times

 **azure9876** 1 year, 5 months ago

Please check this table:  
<https://docs.databricks.com/administration-guide/account-settings/audit-logs.html#workspace-level-audit-log-events>  
In general, all logs are belongs to workspace-level audit log if you check the title of the table. But if you check in details, cluster related logs belongs to clusters part.  
upvoted 3 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>

 **Deeksha1234** 10 months ago

**Selected Answer: D**

agree Last sync time is right  
upvoted 2 times

 **Nivas2401** 1 year, 3 months ago

**Selected Answer: D**

<https://docs.microsoft.com/en-us/azure/storage/common/last-sync-time-get?tabs=azure-powershell>  
upvoted 4 times

 **ovokpus** 1 year, 3 months ago

**Selected Answer: D**

last sync time  
upvoted 3 times

 **ANath** 1 year, 4 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 3 times

 **PallaviPatel** 1 year, 4 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 4 times

 **Fer079** 1 year, 5 months 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** 1 year, 5 months 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>

□  **Deeksha1234** 10 months ago

**Selected Answer: B**

B is correct

upvoted 2 times

□  **kilowd** 1 year, 4 months ago

**Selected Answer: B**

<https://techcommunity.microsoft.com/t5/datacat/polybase-setup-errors-and-possible-solutions/ba-p/305297>

upvoted 4 times

□  **Raghu108** 1 year, 4 months ago

**Selected Answer: B**

Correct

upvoted 3 times

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

**Selected Answer: B**

correct

upvoted 3 times

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

correct

upvoted 2 times

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** 10 months ago

**Selected Answer: D**

D is correct

upvoted 2 times

 **StudentFromAus** 11 months, 2 weeks ago

**Selected Answer: D**

Answer is correct

upvoted 3 times

 **agar** 1 year, 3 months ago

**Selected Answer: D**

correct

upvoted 2 times

 **Raghul08** 1 year, 4 months ago

**Selected Answer: D**

Correct

upvoted 2 times

 **PallaviPatel** 1 year, 4 months ago

**Selected Answer: D**

淘宝店铺：<https://shop63989109.taobao.com/>

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** 10 months ago

Agree !

upvoted 1 times

 **ANath** 1 year, 5 months ago

I think the answer is correct because in some cases the rows are zero.

upvoted 3 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.

 **75082SN** 5 months, 1 week ago

Why not D?

upvoted 2 times

 **shoottheduck** 3 months, 1 week ago

Also, a composite key does not improve performance on its own.

Distributing on the two columns that are joined, will

upvoted 1 times

 **sensaint** 5 months, 1 week ago

Then you are partly distributing on a date column which is very bad for performance.

upvoted 1 times

 **Deeksha1234** 10 months ago

B seems correct but not sure what's wrong with D ?

upvoted 3 times

 **PallaviPatel** 1 year, 4 months ago

**Selected Answer: B**

correct

upvoted 4 times

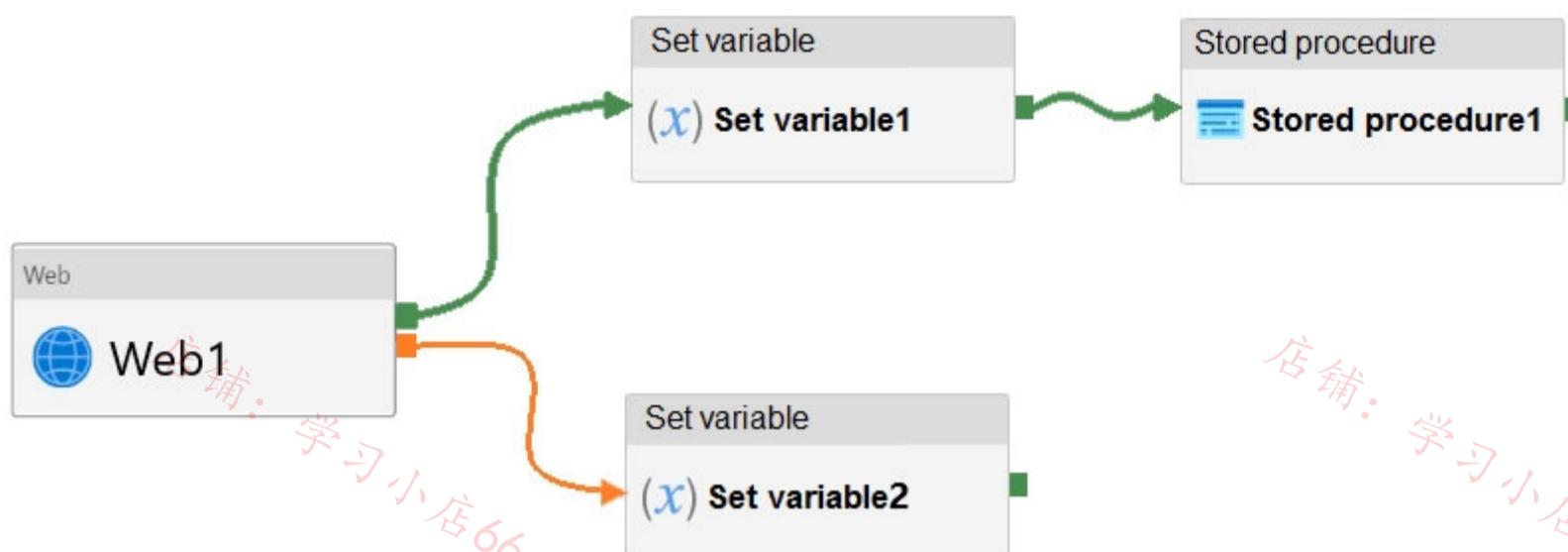
 **bad\_attitude** 1 year, 5 months ago

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]

	▼
complete	
fail	
succeed	

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]

	▼
Canceled	
Failed	
Succeeded	

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

Stored procedure1 will execute Web1 and Set variable1 [answer choice]

	▼
complete	
fail	
succeed	

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]

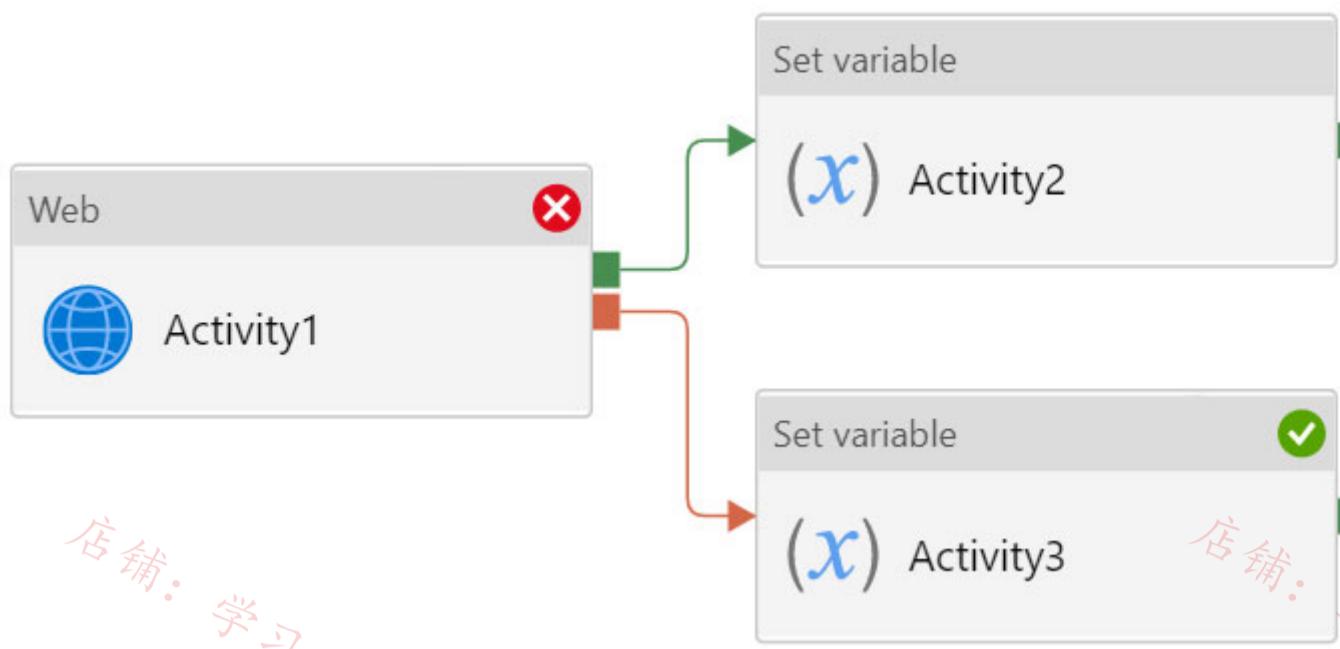
	▼
Canceled	
Failed	
Succeeded	

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 1 year, 5 months 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 34 times

□ **Avi\_Bdj** 1 year, 2 months ago

Second should also be succeeded.

upvoted 6 times

□ **a03** 1 year, 5 months ago

Agree. Second is "Fail" because Success connector presented.

upvoted 3 times

□ **XiltroX** 6 months, 1 week 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 10 times

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

I just tested it myself. Provided answers are correct

upvoted 7 times

□ **RajBathani** Highly Voted 1 year, 6 months ago

The second answer should be Succeeded as 'Set Variable 2' has failed dependency on Web1.

upvoted 33 times

□ **chryckie** Most Recent 1 month, 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 3 times

□ **AHUI** 2 months 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 3 times

□ **vrodriguesp** 4 months, 3 weeks ago

Using this Microsoft 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

淘宝店铺：<https://shop63989109.taobao.com/>

When web1 activity fails: node setVariable2 succeeds and setVariable1 is skipped and its parent node web1 failed; overall pipeline fails upvoted 6 times

□ **csd** 10 months ago

In any scenario pipeline will show success status, cause we are catching the failure  
upvoted 2 times

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

The answers are correct.  
upvoted 1 times

□ **datnguye** 1 year, 5 months ago

It should be Succeeded in both.  
The reference article says: The failure dependency means this pipeline reports success.  
upvoted 14 times

□ **datnguye** 1 year, 5 months 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 13 times

□ **Remedios79** 11 months, 1 week ago

I agree with you too  
upvoted 1 times

□ **ladywhiteadder** 1 year, 2 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** 10 months, 2 weeks ago

This is all you need for the right answer. Its well explained in the link.  
upvoted 1 times

□ **Yohannesmulu** 1 year, 2 months ago

Agreed!  
upvoted 1 times

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 1 year, 5 months ago

**Selected Answer: DE**

Notebook- azure databricks, managing activities in pipeline-datafactory  
upvoted 32 times

**ElHomo2222** Highly Voted 1 year, 4 months ago

**Selected Answer: DE**

D & E; Databricks for Wrangling and Notebooks; ADF for Copy and Jar  
upvoted 13 times

**kilowd** 1 year, 4 months ago

Wrangling and Copy = ADF  
Jar and Notebooks = Databricks  
upvoted 7 times

**janaki** Most Recent 2 weeks ago

**Selected Answer: DE**

D - Azure Data Factory  
E - Azure Databricks  
upvoted 1 times

**pavankr** 2 weeks ago

You "de-bug" the activity with ML??? Seriously??? come on man??? from where you are getting these answers???

upvoted 1 times

**Mohamedali.Cintellic** 1 month, 1 week ago

**Selected Answer: DE**

D & E are correct  
upvoted 1 times

**vrodriguesp** 4 months, 3 weeks ago

**Selected Answer: DE**

Notebook on azure databricks, rest on pipeline data factroy. No sense for AandC  
upvoted 3 times

**nicky87654** 5 months, 1 week ago

**Selected Answer: DE**

Wrangling and Copy = ADF  
Jar and Notebooks = Databricks  
upvoted 3 times

**Deeksha1234** 10 months ago

**Selected Answer: DE**

should be DE  
upvoted 2 times

martinamartina 10 months ago

淘宝店铺：<https://shop63989109.taobao.com/>

Couldn't be AD?  
upvoted 1 times

dsp17 10 months, 4 weeks ago

Selected Answer: DE

DE - correct  
upvoted 1 times

dsp17 11 months ago

Selected Answer: DE

Wrangling and Copy -> ADF  
Jar and Notebooks -> Databricks  
upvoted 1 times

Remedios79 11 months, 1 week ago

Selected Answer: DE

absolutely D&E  
upvoted 2 times

Remedios79 11 months, 1 week ago

D and E absolutely!!  
upvoted 1 times

VenkataPolepalli 1 year ago

Selected Answer: DE

Answer is D&E  
upvoted 1 times

Lucky\_me 1 year, 1 month ago

Selected Answer: DE

Notebook, jar: databricks  
Managing activities/pipelines: ADF  
upvoted 1 times

rafaelptu 1 year, 2 months ago

Selected Answer: AC

Azure synapse analytics já contempla o ADF e ADB.  
upvoted 1 times

kanak01 1 year, 4 months ago

Selected Answer: DE

Data Factory & Databricks  
upvoted 1 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.dmw\_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.dmw\_nodes\_db\_partition\_stats.

**Correct Answer: D**

Microsoft recommends use of sys.dmw\_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, 1 month ago

Correct. See Question 12 topic 4

upvoted 7 times

✉  **bulutfet** Most Recent 1 week, 2 days ago

D correct

upvoted 1 times

✉  **janaki** 2 weeks ago

**Selected Answer: A**

Option A is correct

upvoted 1 times

✉  **janaki** 2 weeks ago

Sorry, option D is correct.

upvoted 1 times

✉  **Deeksha1234** 10 months 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. 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>

 **StudentFromAus** 11 months, 2 weeks ago

**Selected Answer: B**

For already used queries, we need to monitor the adaptive caching  
upvoted 3 times

 **juanlu46** 1 year, 1 month ago

**Selected Answer: B**

Is correct  
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, 1 month ago

**Selected Answer: D**

Correct!

upvoted 5 times

✉  Jerrie86 Most Recent 4 months, 2 weeks ago

Asking to monitor Pipeline failures and D is activity runs. so Cant be D. Looks like they are missing an answer here

upvoted 1 times

✉  dom271219 9 months, 2 weeks ago

**Selected Answer: D**

Redundant question

upvoted 1 times

✉  Deeksha1234 10 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>

 nicky87654 4 months, 3 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 2 times

 Shanmahi 6 months ago

**Selected Answer: B**

Reasons for choosing option B --> IoT devices, streaming data, real-time data requirement

upvoted 2 times

 MadhuMDLK1055 6 months, 1 week ago

Ans is D

upvoted 1 times

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.

 **dimbrici** 6 months, 3 weeks ago

**Selected Answer: C**

C is the correct Answer !

upvoted 2 times

 **greenlever** 7 months, 3 weeks 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 2 times

 **anks84** 9 months ago

**Selected Answer: C**

C is the correct Answer !

upvoted 2 times

## Question #29

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>

✉  **henryphchan** 3 weeks, 2 days ago

**Selected Answer: C**

i vote for C because only the % used is meaningful

upvoted 2 times

✉  **rohitbinnani** 1 month 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 2 times

✉  **chryckie** 1 month, 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 3 times

✉  **markpumc** 2 months, 3 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 3 times

✉  **Shanmahi** 6 months ago

**Selected Answer: A**

DWU used is the metric to use, if only one best answer is expected. option A.

upvoted 2 times

✉  **shaileshutd** 6 months, 1 week 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

淘宝店铺：<https://shop63989109.taobao.com/>

□ **Tickxit** 6 months, 3 weeks ago

Which is the best one, DWU used or DWU percentage? We need to select one.

upvoted 1 times

□ **CodingOwl** 7 months ago

AC Both are answers

upvoted 1 times

□ **Phund** 9 months ago

should be both DWU metrics

upvoted 2 times

Question #30

Topic 4

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>

□ **maximilianogarcia6** 6 months, 2 weeks ago

this question is not repeated as options are different. It could appear the first one or this.

upvoted 2 times

□ **AdarshKumarKhare** 7 months, 1 week ago

Question is a repeat

upvoted 1 times

□ **sensaint** 7 months, 2 weeks 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:

**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 ago

Correct.

Embarrassingly parallel jobs

Step 3 and 4.

upvoted 38 times

□ **Liz42** 1 year, 8 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** 1 year, 8 months ago

Disregard my above comment... meant to respond to another

upvoted 2 times

□ **Deeksha1234** Most Recent 10 months ago

correct

upvoted 1 times

□ **nelineli** 11 months, 2 weeks 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 2 times

□ **sdokmak** 1 year 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** 1 year, 9 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 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** 1 year, 10 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 ago

Why 16? Don't understand...

upvoted 2 times

□ **mbravo** 2 years ago

Embarrassingly parallel jobs

upvoted 10 times

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

It's not THAT embarrassing

upvoted 10 times

□ **wwdba** 1 year, 3 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 1 times

□ **Davico93** 11 months, 2 weeks 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 1 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:	<div style="border: 1px solid black; padding: 5px; width: fit-content;"> <input type="checkbox"/> Hash-distributed  <input type="checkbox"/> Round-robin         </div>	<div style="border: 1px solid black; padding: 5px; width: fit-content;"> <input type="checkbox"/> DateKey  <input type="checkbox"/> ProductKey  <input type="checkbox"/> RegionKey         </div>
Invoices:	<div style="border: 1px solid black; padding: 5px; width: fit-content;"> <input type="checkbox"/> Hash-distributed  <input type="checkbox"/> Round-robin         </div>	<div style="border: 1px solid black; padding: 5px; width: fit-content;"> <input type="checkbox"/> DateKey  <input type="checkbox"/> ProductKey  <input type="checkbox"/> RegionKey         </div>

**Answer Area**

Table	Distribution type	Distribution column
Sales:	<div style="border: 1px solid black; padding: 5px; width: fit-content;"> <input checked="" type="checkbox"/> Hash-distributed  <input type="checkbox"/> Round-robin         </div>	<div style="border: 1px solid black; padding: 5px; width: fit-content;"> <input checked="" type="checkbox"/> DateKey  <input checked="" type="checkbox"/> ProductKey  <input type="checkbox"/> RegionKey         </div>
Invoices:	<div style="border: 1px solid black; padding: 5px; width: fit-content;"> <input checked="" type="checkbox"/> Hash-distributed  <input type="checkbox"/> Round-robin         </div>	<div style="border: 1px solid black; padding: 5px; width: fit-content;"> <input checked="" type="checkbox"/> DateKey  <input checked="" type="checkbox"/> ProductKey  <input type="checkbox"/> RegionKey         </div>

Box 1: Hash-distributed -

Box 2: ProductKey -

ProductKey is used extensively in joins.

Hash-distributed tables improve query performance on large fact tables.

## 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 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 90 times

□  **v blessings** 1 year, 10 months ago

i agree

upvoted 2 times

□  **Marcello83** 1 year, 11 months ago

I agree with lara\_mia1

upvoted 2 times

□  **niceguy0371** 1 year, 9 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 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 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, 4 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

□  **dom271219** Most Recent 9 months ago

"Choose a distribution column with data that distributes evenly"

ProductKey is more relevant in both cases

upvoted 3 times

□  **Deeksha1234** 10 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 2 times

 **Nishikag** 11 months, 1 week ago

淘宝店铺 : <https://shop63989109.taobao.com/>

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** 11 months, 2 weeks ago

the provided answers are correct

upvoted 1 times

 **kiranSargar** 1 year, 2 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** 1 year, 10 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** 1 year, 10 months ago

But we cannot use ProductKey in both because in Invoice table it is used in WHERE condition

upvoted 3 times

 **Lucky\_me** 1 year, 5 months ago

If we choose RegionKey for Sales, we would have a processing skew.

upvoted 3 times

 **Aditya0891** 11 months, 3 weeks 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** 1 year, 10 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** 1 year, 11 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** 1 year, 12 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 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 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** 1 year, 10 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 ago

I AGREE WITH BOTH HASH WITH PRODUCT KEY

upvoted 10 times

## Question #33

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 1 year, 10 months ago

correct

upvoted 7 times

SG1705 Highly Voted 1 year, 12 months ago

Why ??

upvoted 6 times

IgorLacik 1 year, 11 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

okechi 1 year, 11 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 38 times

noranathalie 1 year, 7 months ago

In question 2, we just mentioned to not use the where condition columns to create partitions.. so the logic is unclear for me..

upvoted 2 times

noranathalie 1 year, 7 months ago

please disregard my comment above. Partitioning is different from hash-column, so the criterias are different

upvoted 4 times

Deeksha1234 Most Recent 10 months ago

correct, agree with okechi

upvoted 1 times

dsp17 11 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 3 times

ploer 1 year, 4 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, 4 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 1 year, 5 months 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

淘宝店铺：<https://shop63989109.taobao.com/>

□ **Lucky\_me** 1 year, 5 months 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, 2 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

Question #34

Topic 4

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** 7 months, 2 weeks 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 4 times

□ **rzeng** 7 months, 2 weeks ago

correct

upvoted 2 times

## Question #35

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>

✉ **bp\_a\_user** 1 month, 1 week 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** 2 months, 3 weeks ago

We had the same question before. The correct answer is D

upvoted 3 times

✉ **Vikram1710** 3 months, 1 week ago

Option A is confusing as we have different answer for same question.

upvoted 1 times

✉ **vrodriguesp** 4 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 4 times

✉ **Ngol** 4 months, 3 weeks ago

I don't understand why Exam Topics should be giving different answers for questions they have repeated...like this one!

upvoted 4 times

✉ **OldSchool** 6 months, 3 weeks 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 1 times

✉ **OldSchool** 6 months 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** 6 months, 3 weeks ago

**Selected Answer: D**

Question already seen

upvoted 2 times

✉ **AdarshKumarKhare** 7 months, 1 week ago

Question repeated

upvoted 2 times

淘宝店铺：<https://shop63989109.taobao.com/>

 **SD4592** 8 months, 3 weeks ago

**Selected Answer: D**

Absolutely D

upvoted 3 times

 **debarun** 8 months, 4 weeks ago

Correct answer is D.

upvoted 3 times

 **federc** 9 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** 9 months ago

**Selected Answer: D**

The same question as #8 Topic #4, but different answer. Should be D.

upvoted 1 times

 **anks84** 9 months ago

**Selected Answer: D**

Correct answer is D.

upvoted 4 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.dm\_pdw\_sys\_info.

**Correct Answer: D**

Use 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>

□ **Leyya11111** Highly Voted 9 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** 9 months ago

Correct, answer is A !

upvoted 4 times

□ **pavankr** Most Recent 1 week, 5 days ago

Ok, he did the typo for printing D. It should be "Connect to the built-in pool and use sys.dm\_pdw\_nodes\_db\_partition\_stats"  
upvoted 1 times

□ **janaki** 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 2 times

□ **vrodriguesp** 4 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** 6 months, 1 week ago

**Selected Answer: A**

need to connect Azure Synapse Analytics dedicated SQL pool1 not built-in pool (serverless pool)

upvoted 3 times

□ **OldSchool** 6 months, 1 week ago

**Selected Answer: A**

A is the answer.

Read Question 20, Topic 4

upvoted 1 times

□ **rzeng** 7 months, 2 weeks ago

A is the right one!

upvoted 1 times

□ **igormmpinto** 7 months, 3 weeks 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** 8 months ago

Answer A

upvoted 3 times

 **momani** 8 months, 1 week ago

Answer A is correct

upvoted 1 times

 **walidazure** 8 months, 2 weeks ago

**Selected Answer: A**

Answer A

upvoted 3 times

 **feder** 9 months ago

answer A is the correct one.

upvoted 3 times

 **pangas2567** 9 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 5 months, 1 week 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 11 times

 **esaade** Most Recent 2 months, 4 weeks 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 1 times

 **Ast999** 3 months ago

Selected Answer: DE

D + E = correct  
upvoted 2 times

 **nicky87654** 4 months, 3 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 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 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**

□ **Jerrie86** Highly Voted 4 months, 2 weeks ago

This is repeated way too many times.

upvoted 6 times

□ **pavankr** Most Recent 1 week, 5 days 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 1 times

□ **OfficeSaracus** 1 month, 1 week ago

**Selected Answer: A**

A for sure

upvoted 1 times

□ **duzi** 4 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** 1 week, 5 days ago

Looks like he is misleading us?

upvoted 1 times

□ **pk07** 4 months, 3 weeks ago

**Selected Answer: A**

(H)Agreed!

upvoted 2 times

□ **Mouli10** 4 months, 3 weeks ago

**Selected Answer: A**

Its A we need to connect to Pool1

upvoted 4 times

□ **nicky87654** 4 months, 3 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 4 times

□ **ZIMARAKI** 4 months, 3 weeks ago

**Selected Answer: A**

It's A

upvoted 4 times

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 1 month 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 6 times

azure\_user11 Most Recent 3 weeks, 2 days 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 3 weeks, 5 days ago

**Selected Answer: A**

agree with Yemeral

upvoted 1 times

chryckie 1 month, 1 week 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 4 times

chryckie 1 month, 1 week 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

sk20 1 month, 2 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 1 month, 3 weeks ago

**Selected Answer: C**

Correct answer is C

upvoted 1 times

淘宝店铺：<https://shop63989109.taobao.com/>

AscentAcademy 3 months 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-delimitedtextmorecolumnsdefined/>

upvoted 4 times

shoottheduck 3 months, 1 week 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

raydoneaan 3 months, 3 weeks ago

C is correct

upvoted 1 times

vrodriguesp 3 months, 3 weeks ago

Selected Answer: D

mapping is correct because error is only on one row (row number 53) so the only acceptable should be D

upvoted 3 times

Jerrie86 4 months, 2 weeks ago

Selected Answer: D

The answer should be D. The error 'there are more columns in the source file '0\_2020\_11\_09\_11\_43\_32.avro' than expected could be because of one extra column delimiter. And that leads to the error.

Extra column error would have occurred at row 1 if there was actually an extra column.

Answer should be D to skip that row because the data coming from the source is not correct.

upvoted 4 times

Lestrang 4 months, 2 weeks ago

Who said it is not correct? It just has extra columns for this particular record. Why discard potentially valuable data when you can keep it by defining an explicit mapping?

Sure this seems like a 1 row only but you have no guarantee that this won't happen again.

upvoted 4 times

agold96 4 months, 2 weeks ago

Selected Answer: D

As the error happens on only one row, I guess the mapping is done correctly, there is just a mistake on the row 53. Then, the answer should be D.

upvoted 4 times

Lestrang 4 months, 2 weeks ago

Selected Answer: C

C. Add an explicit mapping.

The error message indicates that there are more columns in the source file '0\_2020\_11\_09\_11\_43\_32.avro' than expected. One way to resolve this issue is to add an explicit mapping in the Copy activity settings, which specifies the columns in the source file and their corresponding columns in the destination. This ensures that the correct columns are being copied and can help prevent issues with incompatible column counts.

upvoted 2 times

youngbug 4 months, 3 weeks ago

Why not D?

upvoted 2 times

vrodriguesp 4 months, 3 weeks ago

Selected Answer: C

I Agree

upvoted 2 times

Stefan94 4 months, 3 weeks ago

Correct

upvoted 3 times

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 4 months, 3 weeks ago

**Selected Answer: D**

Azure databricks  
upvoted 5 times

✉  **Stefan94** Most Recent 4 months, 3 weeks 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:

	▼
<input checked="" type="checkbox"/> sys.dm_pdw_exec_requests <input type="checkbox"/> sys.dm_pdw_sql_requests <input type="checkbox"/> sys.dm_pdw_exec_sessions	▼

Correct Answer:

Identify which queries are waiting on resources:

	▼
<input type="checkbox"/> sys.dm_pdw_waits <input checked="" type="checkbox"/> sys.dm_pdw_lock_waits <input type="checkbox"/> sys.resource_governor_workload_groups	▼

 **bp\_a\_user** 1 month, 1 week 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 query with a type of UserConcurrencyResourceTyp 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 2 times

 **AHUI** 2 months ago

correct

<https://learn.microsoft.com/en-us/azure/azure-sql/database/monitoring-with-dmvs?view=azuresql>

upvoted 2 times

 **AHUI** 2 months 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 15 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\_2023** 2 days, 22 hours ago

**Selected Answer: B**

I would answer B

upvoted 1 times

 **BillMy1** 2 weeks, 1 day 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 3 times

 **azure\_user11** 2 weeks, 5 days 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 4 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** 1 month 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 3 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** 1 month 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 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 limit
- B. Data IO percentage
- C. Cache hit percentage
- D. CPU percentage

**Correct Answer: C**

✉  **darshilparmar** 2 days ago

Repeat Questions

upvoted 1 times

✉  **henryphchan** 3 weeks, 6 days ago

**Selected Answer: C**

Answer is C, and it's a repeated question

upvoted 2 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 4 weeks, 1 day 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 2 times

henryphchan 3 weeks, 2 days ago

so the provided answer is correct.

upvoted 2 times

Question #47

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. Data Warehouse Units (DWU) used
- D. Data IO percentage

**Correct Answer: B**

 **darshilparmar** 2 days ago

Repeated 4 times  
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:

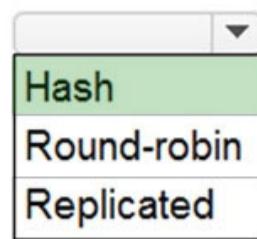
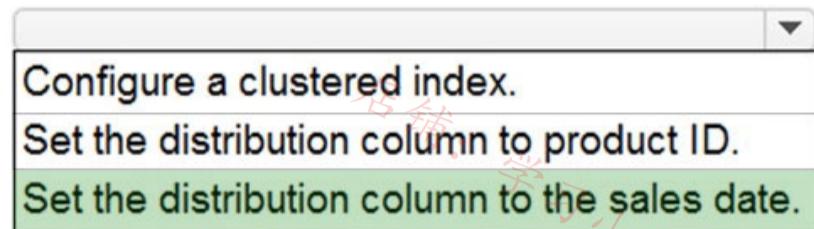
#### Answer Area

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:****When creating the table for sales transactions:**

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 4 months, 2 weeks ago

This case study was in my exam and I scored 970. I chose productid.

upvoted 21 times

✉ **Julia01** Highly Voted 9 months ago

Id 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 16 times

✉ **mokrani** 7 months, 1 week 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** 6 months, 3 weeks 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 7 times

✉ **XiltroX** Most Recent 6 months, 1 week 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 6 times

✉ **pavankr** 1 week, 5 days ago

So then why this guy is misleading us?? I find lot of answers misleading us.

upvoted 1 times

✉ **OldSchool** 6 months, 1 week 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 10 times

✉ **kornat** 2 months ago

True! !!

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

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 5 months, 3 weeks 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 13 times

□ **vrodriguesp** 3 months, 3 weeks ago

are you sure we can create EXTERNAL DATA SOURCE without DATABASE SCOPED CREDENTIAL?

upvoted 1 times

□ **JasonVu** 5 months, 2 weeks ago

CETAS is not available in dedicated SQL pool

upvoted 2 times

□ **AzureJobsTillRetire** 5 months, 1 week ago

Please see below.

CREATE TABLE AS SELECT (Azure Synapse Analytics)

<https://learn.microsoft.com/en-us/sql/t-sql/statements/create-table-as-select-azure-sql-data-warehouse?view=aps-pdw-2016-au7>

upvoted 1 times

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

□ **juanlu46** Highly Voted 8 months, 2 weeks ago

1. Scoped Database Credencial

2. External Data Source

3 External File Format

upvoted 6 times

□ **scarycat** 6 months ago

Scoped Database Credencial is a DCL command, not DDL

upvoted 2 times

□ **OldSchool** 6 months, 1 week ago

Correct

upvoted 2 times

□ **BPW** Most Recent 1 month, 2 weeks ago

Box 1: CREATE EXTERNAL DATA SOURCE  
Box 2: CREATE EXTERNAL FILE FORMAT  
Box 3: CREATE EXTERNAL TABLE

upvoted 3 times

淘宝店铺：<https://shop63989109.taobao.com/>

□ **MartianNC** 2 months, 1 week ago

The reason you use CTAS is that you must implement row level security.

upvoted 1 times

□ **Jerrie86** 4 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=azure-sqldw-latest&preserve-view=true>

upvoted 3 times

□ **JitBiswas** 4 weeks, 1 day 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** 4 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** 6 months ago

why use CETAS instead of Create External Table?

upvoted 1 times

□ **Pais** 5 months, 4 weeks 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%2Ftoc.json&bc=%2Fazure%2Fsynapse-analytics%2Fsql-data-warehouse%2Fbreadcrumb%2Ftoc.json&view=azure-sqldw-latest&preserve-view=true#examples-using-ctas-to-replace-sql-server-code>

upvoted 2 times

□ **JasonVu** 5 months, 2 weeks ago

your link points to CTAS, which is a different topic

upvoted 1 times

□ **Igor85** 6 months ago

CREATE DATABASE SCOPED CREDENTIALS should be run before all other steps in the given answer

upvoted 1 times

□ **7yut** 6 months, 1 week ago

I think the provided answer in answer are is correct

upvoted 1 times

□ **greenlever** 8 months, 1 week 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

**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**  4 months, 2 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 7 times

✉  **gerrie1979**  7 months, 1 week 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** 4 months, 2 weeks ago

First question is partition not distribution. So Date is correct

upvoted 1 times

✉  **sensaint** 7 months, 1 week ago

I would partition by ProductID since joins and filtering must be optimized for that column

upvoted 1 times

✉  **mokrani** 7 months, 1 week 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 11 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>

由  **sntlkumar** 1 month, 1 week ago

Given answer is correct

upvoted 1 times

由  **uira** 6 months ago

**Selected Answer: A**

Identity should be used.

upvoted 2 times

由  **7yut** 6 months, 1 week ago

**Selected Answer: A**

Correct

upvoted 2 times

由  **anks84** 9 months, 1 week ago

**Selected Answer: A**

A is the correct Answer !

upvoted 3 times

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

店铺：学习小店66

## 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>

曰  **greenlever** Highly Voted 8 months ago

replicated

hash

upvoted 24 times

曰  **Jerrie86** Highly Voted 4 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 6 times

曰  **pavankr** Most Recent 1 week, 5 days ago

So on which answer we should reply on?????? Why this web site guy is guiding us all wrong answers?????

upvoted 1 times

曰  **JosephVishal** 4 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 1 times

曰  **Taou** 5 months, 1 week ago

1st is Replicated

upvoted 1 times

曰  **AzureJobsTillRetire** 6 months, 1 week 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

曰  **smsme323** 8 months, 2 weeks ago

replicated

HASH

upvoted 2 times

曰  **juanlu46** 8 months, 2 weeks ago

-Replicated

-Hash

upvoted 3 times

淘宝店铺：<https://shop63989109.taobao.com/>

□ **anks84** 9 months ago

Looks like "Retail Store" is a dimension table with 2MB size.

So, Replicated should be better option in my opinion,

upvoted 4 times

□ **feder** 9 months ago

Agree with you Julia01, Replicated would be more reasonable for a 2MB table

upvoted 1 times

□ **R12346** 9 months ago

The retail table should be of replicated type since it is just 2 MB

upvoted 2 times

□ **Julia01** 9 months ago

Shouldn't it be replicated in the first one?

upvoted 3 times

□ **pangas2567** 9 months ago

Agree, only 2MB of size could definitely be replicated. I don't see any load speed requirements in the description for this table.

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

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 -

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>

□ **Azurre** 2 months, 2 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 2 times

□ **XiltroX** 6 months, 1 week ago

Its funny cause in the scenario, there is a BIG hint on what to use for box 2. Read it up.

upvoted 3 times

□ **Xinyuehong** 7 months, 3 weeks ago

agreed

upvoted 1 times

□ **federC** 9 months 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 -

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**  4 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 5 times

□  **yogiazaad** 4 months, 1 week 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 1 times

□  **cale**  2 months, 1 week ago

**Selected Answer: D**

Answer is D

upvoted 2 times

 **haidebelognime** 3 months, 1 week ago

**Selected Answer: C**

do the research. it is C time-based retention

upvoted 1 times

 **Ast999** 3 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 3 times

 **MrWood47** 4 months, 3 weeks ago

**Selected Answer: C**

Sim\_al explanation is correct

upvoted 1 times

 **nicky87654** 4 months, 3 weeks ago

**Selected Answer: C**

C: time-based retention

upvoted 2 times

 **Sima\_al** 5 months, 1 week 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** 4 months, 1 week 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 2 times

 **Snomax** 7 months ago

**Selected Answer: D**

Agreed.

upvoted 2 times

## Topic 6 - Testlet 2

Question #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 -

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

Correct Answer:

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 -

淘宝店铺：<https://shop63989109.taobao.com/>

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 1 year, 7 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 18 times

 **A\_Nath** 1 year, 3 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

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

The answers are correct.

upvoted 15 times

 **Bhuvanesh2104** Most Recent 4 months, 4 weeks 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 1 times

 **AzureJobsTillRetire** 6 months, 1 week ago

Agreed with the given answers.

Box1: Self-hosted integration runtime

Why not Azure-SSIS integration runtime? 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 8 times

 **vrodriguesp** 3 months, 3 weeks ago

agree with you

upvoted 1 times

 **sunil\_smile** 8 months, 3 weeks ago

I think it should be SSIS integration runtime... Because currently there are SSIS pipelines which does the data integration

upvoted 1 times

 **Deeksha1234** 10 months ago

In opinion the given answer is correct since its daily inventory data, i.e. will be loaded once daily.

upvoted 1 times

 **Canary\_2021** 1 year, 5 months ago

Answers are correct because 'Daily inventory data comes from a Microsoft SQL server located on a private network.'

upvoted 1 times

 **dija123** 1 year, 6 months 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

 **datnguye** 1 year, 5 months ago

Should it be Self-hosted as Microsoft SQL server, not Azure though?

upvoted 1 times

 **Davico93** 11 months, 2 weeks ago

maybe.... even if it is an azure resource, but in private network, we need SelfHosted

upvoted 1 times

 **AppleVan** 1 year, 8 months ago

Shouldn't it be event based?

upvoted 3 times

淘宝店铺 : <https://shop63989109.taobao.com/>

□ **rikku33** 1 year, 8 months ago

Schedule trigger - because daily. so the given answer is correct

upvoted 5 times

□ **samko92** 1 year, 7 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** 6 months, 3 weeks ago

read carefully - import every 8 hours for customer date, not inventory data. Evenet triggers can be used only with storage account, so Event based is for sure wrong. It's tumbling windows

upvoted 1 times

□ **OldSchool** 6 months, 1 week 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

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 1 year, 5 months ago

The answer provided is correct  
upvoted 26 times

□ **SameerL** 10 months, 4 weeks ago

The provided sequence is correct per below link:

<https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-delivery>  
upvoted 4 times

□ **NaiCob** Highly Voted 1 year, 5 months 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 17 times

□ **Igor85** 6 months 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 1 times

□ **wwdba** 1 year, 3 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** 1 year, 5 months ago

@NaiCob I believe the given answer is correct. What changes are published before creating a PR?  
upvoted 3 times

□ **NaiCob** 1 year, 5 months ago

Before Pull Request you have to publish you local changes  
upvoted 1 times

□ **xeti** 1 year, 3 months ago

No, the given answer is correct.

Publish is done after the merge to collaboration (main) branch and is essentially a CI trigger to update the adf\_publish branch.  
upvoted 3 times

□ **dev2dev** 1 year, 4 months ago

Nope. given answers are correct.  
upvoted 1 times

□ **Jerrie86** Most Recent 4 months, 2 weeks ago

This case study was in my exam word to word. Thanks guys. passed at 970.

Just dont do dumps but try to understand the logic via some youtube dp-203 tutorials. There is series by databag.ai. So please study to excel in exam as well in prof life.

upvoted 1 times

□ **anks84** 9 months ago

Given answer and sequence is absolutely correct !!  
upvoted 1 times

□ **Deeksha1234** 10 months ago

given answer is correct  
upvoted 1 times

 **hbad** 1 year ago

淘宝店铺：<https://shop63989109.taobao.com/>

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, 1 month ago

<https://www.atlassian.com/git/tutorials/comparing-workflows/feature-branch-workflow>

upvoted 1 times

 **Kondzio** 1 year, 3 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 1 times

 **edba** 1 year, 5 months 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** 1 year, 5 months 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.

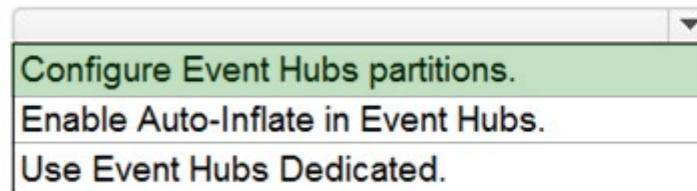
To store the Twitter feed data, use:

- An Azure Data Lake Storage Gen2 account
- An Azure Databricks high concurrency cluster
- An Azure General-purpose v2 storage account in the Premium tier

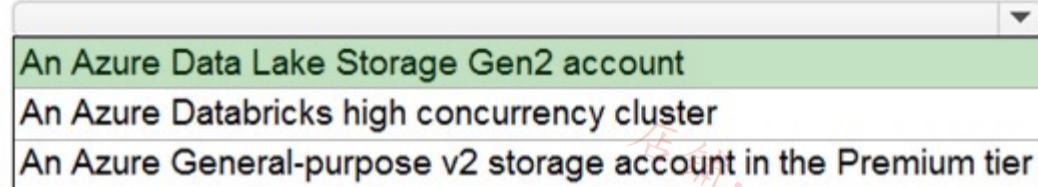
**Correct Answer:**

### Answer Area

To increase the throughput of ingesting the Twitter feeds:



To store the Twitter feed data, use:



#### 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** Highly Voted 1 year, 1 month 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 10 times

### Topic 9 - Testlet 5

upvoted 5 times

 **Lotusss** Most Recent 1 year, 1 month ago

Box one is Enable auto-inflate

Box two Data Lake account

## 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 1 year, 8 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 32 times

□ **Marcus1612** Highly Voted 1 year, 8 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/azure-sql/database/network-access-controls-overview>

upvoted 11 times

□ **AzureJobsTillRetire** Most Recent 6 months, 1 week 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 3 times

□ **Deeksha1234** 10 months ago

**Selected Answer: C**

correct

upvoted 1 times

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

**Selected Answer: C**

Answer is correct

upvoted 1 times

□ **MvanG** 1 year 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, 1 month 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 ago

淘宝店铺 : <https://shop63989109.taobao.com/>

This is a tough one! Their on-prem data is in a private network, so in hindsite 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, 4 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, 4 months ago

**Selected Answer: C**

C looks correct.

upvoted 1 times

 **SabaJamal2010AtGmail** 1 year, 5 months ago

A server-level firewall IP rule is correct

upvoted 1 times

 **Canary\_2021** 1 year, 5 months 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, 2 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** 1 year, 6 months ago

**Selected Answer: C**

Correcta

upvoted 1 times

 **FredNo** 1 year, 6 months ago

**Selected Answer: C**

Answer is C because the company doesn't want to use any virtual network tools.

upvoted 3 times

 **jefvaen** 1 year, 8 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 2 times

 **YipingRuan** 1 year, 7 months ago

How to add rule at database level?

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 1 year, 9 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 36 times

✉  **FredNo** Highly Voted 1 year, 6 months ago

**Selected Answer: C**

Answer is C

upvoted 7 times

✉  **XiltroX** Most Recent 6 months, 1 week 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 2 times

✉  **anks84** 9 months ago

**Selected Answer: C**

Correct Answer is C i.e. Column-level security !!

upvoted 2 times

✉  **yyhhh** 9 months, 2 weeks 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** 10 months ago

**Selected Answer: C**

C - column level security is correct

upvoted 1 times

✉  **dmgArtyco** 10 months ago

La verdad que no queda nada claro

upvoted 1 times

 **Remedios79** 11 months, 1 week ago

淘宝店铺 : <https://shop63989109.taobao.com/>

**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 2 times

 **Davico93** 11 months, 3 weeks ago

It's tricky, because it says "secure" and not "not to access"

upvoted 1 times

 **Arunava05** 1 year ago

Even in udemy also the answer is same as ' Data sensitivity labels'

upvoted 2 times

 **AIcubeHead** 1 year, 2 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, 4 months ago

It should be C, because only analysts are might not see those columns but others yes

upvoted 3 times

 **Remedios79** 11 months, 1 week ago

I agree with you!

upvoted 1 times

 **PallaviPatel** 1 year, 4 months ago

**Selected Answer: C**

column level security is correct answer

upvoted 2 times

 **Raghul108** 1 year, 4 months ago

**Selected Answer: C**

I fee it's C as we can limit access via CLS.

upvoted 1 times

 **datnguye** 1 year, 5 months ago

"Limit access" confusing the selection between C and D.

In this case, I would choose C because it doesn't say to elimitate column(s) e.g. Phone from querying data

upvoted 2 times

 **datnguye** 1 year, 5 months ago

Oh I mean to choose D (can't edit the comment)

upvoted 1 times

 **alexleonvalencia** 1 year, 6 months ago

Correcta [C]

upvoted 1 times

 **Marcus1612** 1 year, 8 months ago

D is wrong because "this type of data is not analytically relevant." Classification enable users to analyse sensitive data.

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 Highly Voted 1 year, 9 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 Reional redundancy for DL would be needed.

upvoted 9 times

sachabess79 1 year, 8 months ago

NB : it's an asynchronous copy.

upvoted 1 times

ian\_viana 1 year, 8 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 1 year, 8 months ago

Please desconsidere my answer!

Event Hub can capture data to Data Lake and Blob. But I think the key word in the question is: eal-time data PROCESSING solution azure.

Event hub is just for capture. Stream Analytics do the processing so I'm going with answer D

upvoted 8 times

Marcus1612 1 year, 8 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, 4 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

Deeksha1234 Most Recent 10 months ago

Selected Answer: D

answer D is correct

upvoted 2 times

StudentFromAus 11 months, 2 weeks ago

Selected Answer: D

Answer is correct  
upvoted 3 times

淘宝店铺：<https://shop63989109.taobao.com/>

 **GDJ2022** 1 year, 4 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, 4 months ago

**Selected Answer: D**  
I go with D and info provided by Canary\_2021 is correct.  
upvoted 2 times

 **HaBroNounen** 1 year, 5 months 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** 11 months, 2 weeks ago

You are right, but HC doesn't improve one shot processing, this would work better with multiple users  
upvoted 1 times

 **edba** 1 year, 5 months ago

I think the answer is correct!  
upvoted 2 times

 **Canary\_2021** 1 year, 5 months 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** 1 year, 5 months 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** 1 year, 5 months ago

sorry, typo, right answer is D  
upvoted 3 times

 **FredNo** 1 year, 6 months ago

What is the correct answer?  
upvoted 2 times