[Q116-Q140 DP-203 Practice Test Give You First Time Success with 100% Money Back Guarantee!

NEWS
D
G
[Q116-Q140] DP-203 Practice Test Give You First Time Success with 100% Money Back Guarantee!

DP-203 Practice Test Give You First Time Success with 100% Money Back Guarantee! All Obstacles During DP-203 Exam Preparation with DP-203 Real Test Questions

What should I know before taking the Microsoft DP-203 exam?

Microsoft offers Data Engineering on Microsoft Azure certification to those who wish to demonstrate their knowledge of data engineering on the Microsoft Cloud. The exam comprises multiple-choice questions, and each question is worth one mark. The candidates are required to attempt 60 questions in total, i.e., within 130 minutes. In order to take this exam, the candidates are required to have knowledge of the fundamental concepts related to the subject matter. Knowledge of basic cloud computing concepts (such as virtual machines, virtual networks, etc.) would be beneficial for the candidates. **Microsoft DP-203 exam dumps** contains an online study guide that explains all the concepts and answers to practice questions. Candidates should try to understand the concepts completely to gain good marks on the test.

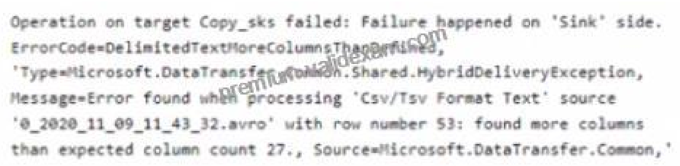
Learn about the benefits of Microsoft DP-203 Certification

Microsoft DP-203 certification is a professional certification given to the candidates who successfully complete the DP-203 exam.

Microsoft Data Platform with Hadoop Developer 203: Administration certification is an international standard for demonstrating competence in data platform administration. The exam validates the candidate's ability to administer and develop data platforms on the cloud-based environment of Microsoft Azure. The DP-203 certification is a globally recognized credential that can enable you to stand out from your peers and make your career more rewarding. The DP-203 course will help you to become a specialist who is able to manage, maintain and develop applications running on Hadoop frameworks on the Azure cloud platform. **Microsoft DP-203 Dumps** is designed to achieve your goal. The DP-203 training course covers the fundamental concepts of cloud computing, creating and managing virtual machines, storage accounts, load balancers, web and worker roles, databases, HDInsight, etc. It also covers how to implement security infrastructure and management of virtual networks using PowerShell commands. You will receive lifetime access to the content along with practice exam questions from real exams after each module. The DP-203 course provides an opportunity for career advancement as it enables you to enhance your expertise in developing solutions with the Hadoop framework and other data sources using the Microsoft Azure cloud platform. It will also help you boost your proficiency in implementing. Correct mapping and auditing exception testing for data.

NO.116 You have an Azure Data Lake Storage Gen2 account that contains two folders named Folder and Folder2.

You use Azure Data Factory to copy multiple files from Folder1 to Folder2.

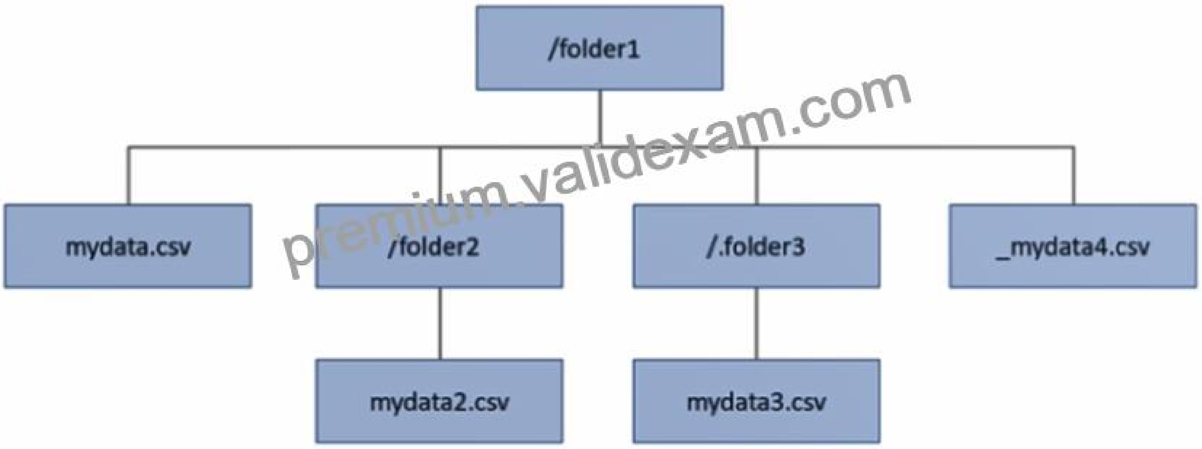


```
Operation on target Copy_sks failed: Failure happened on 'Sink' side.
ErrorCode=DelimitedTextMoreColumnsThanDefined,
'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,
Message=Error found when processing 'Csv/Tsv Format Text' source
'0_2020_11_09_11_43_32.avro' with row number 53: found more columns
than expected column count 27., Source=Microsoft.DataTransfer.Common,'
```

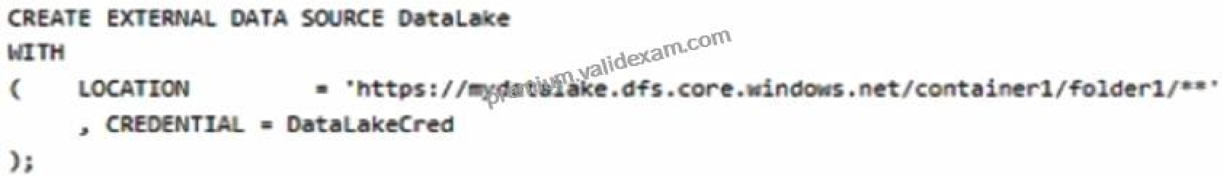
You receive the following error.

What should you do to resolve the error.
* Add an explicit mapping.
* Enable fault tolerance to skip incompatible rows.
* Lower the degree of copy parallelism
* Change the Copy activity setting to Binary Copy

NO.117 You have an Azure Data Lake Storage Gen2 account that contains a container named container1. You have an Azure Synapse Analytics serverless SQL pool that contains a native external table named dbo.Table1. The source data for dbo.Table1 is stored in container1. The folder structure of container1 is shown in the following exhibit.



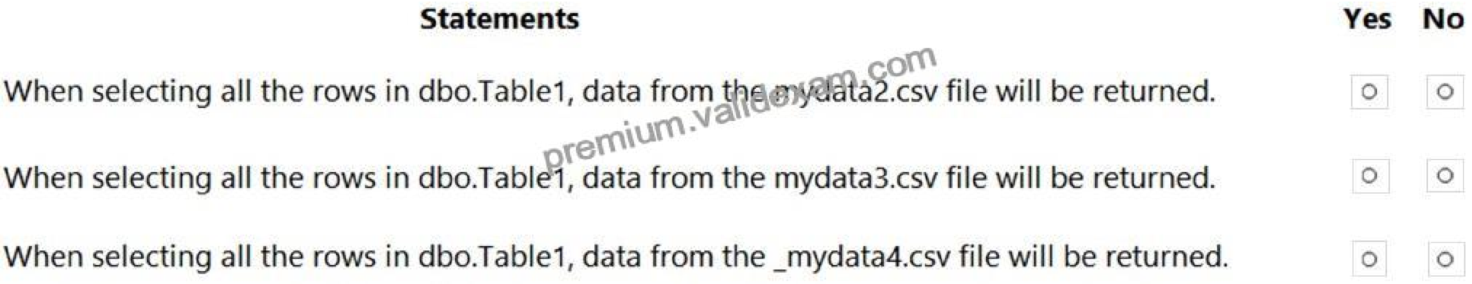
The external data source is defined by using the following statement.



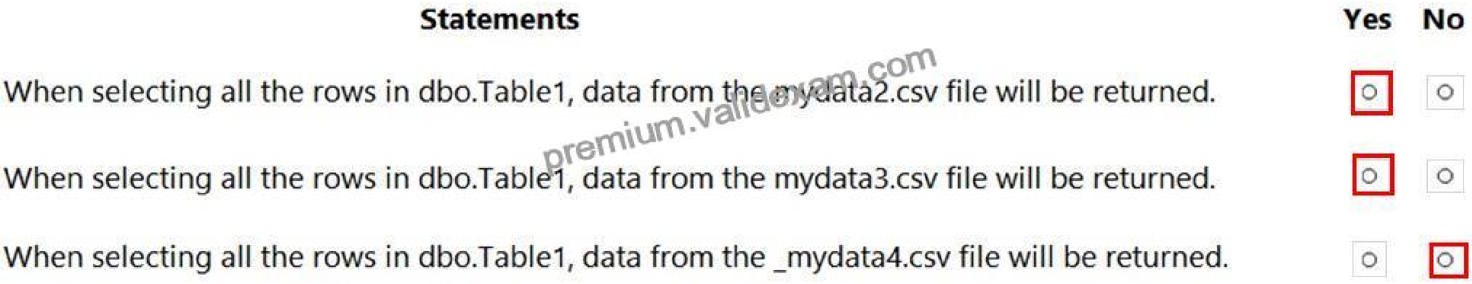
```
CREATE EXTERNAL DATA SOURCE DataLake
WITH
(    LOCATION        = 'https://mydatalake.dfs.core.windows.net/container1/folder1/**'
     , CREDENTIAL = DataLakeCred
);
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.



Statements	Yes	No
When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.	○	○
When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	○	○
When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	○	○



Statements	Yes	No
When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.	[✓ ○]	○
When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	[✓ ○]	○
When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	○	[✓ ○]

NO.118 You have a self-hosted integration runtime in Azure Data Factory.

The current status of the integration runtime has the following configurations:

Status: Running

Type: Self-Hosted

Running / Registered Node(s): 1/1

High Availability Enabled: False

Linked Count: 0

Queue Length: 0

Average Queue Duration. 0.00s

The integration runtime has the following node details:

Name: X-M

Status: Running

Available Memory: 7697MB

CPU Utilization: 6%

Network (In/Out): 1.21KBps/0.83KBps

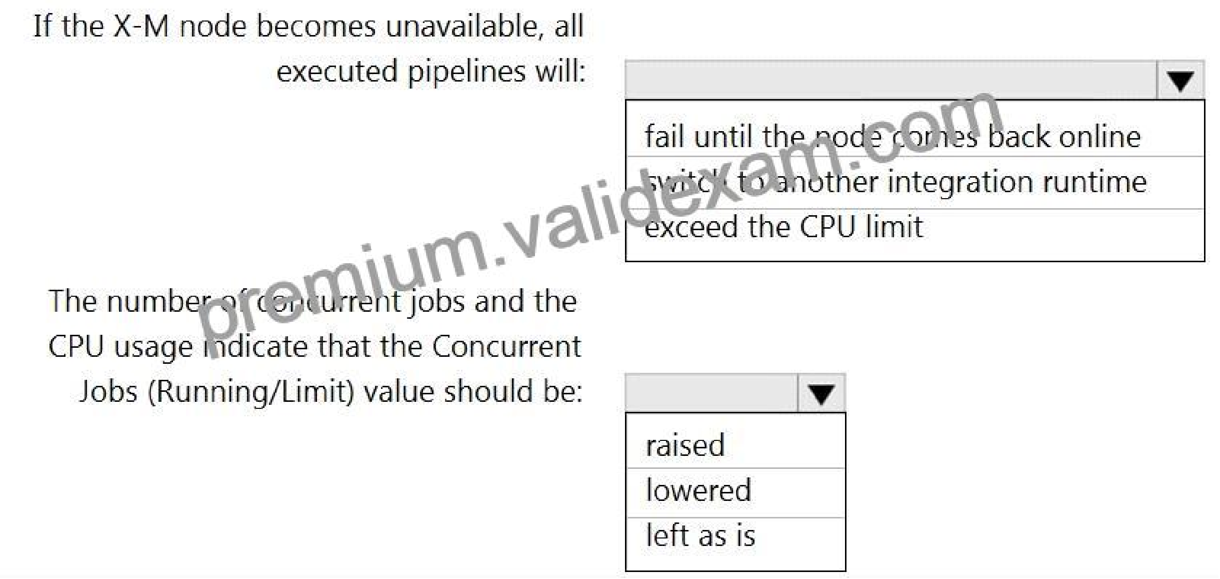
Concurrent Jobs (Running/Limit): 2/14

Role: Dispatcher/Worker

Credential Status: In Sync

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

NOTE: Each correct selection is worth one point.

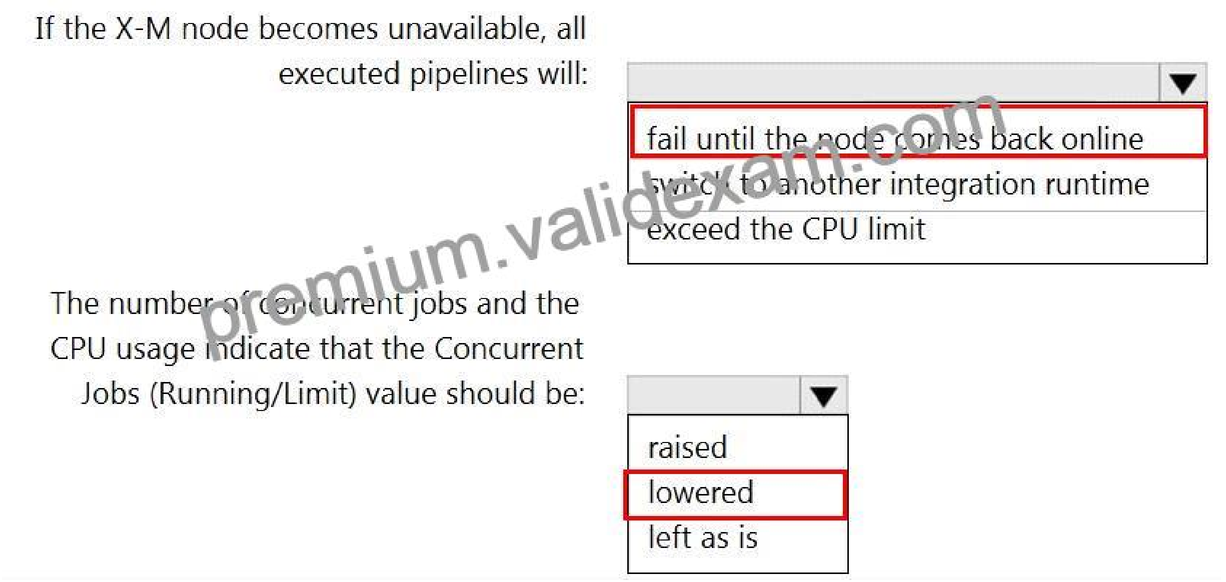


If the X-M node becomes unavailable, all executed pipelines will:

- fail until the node comes back online
- switch to another integration runtime
- exceed the CPU limit

The number of concurrent jobs and the CPU usage indicate that the Concurrent Jobs (Running/Limit) value should be:

- raised
- lowered
- left as is



If the X-M node becomes unavailable, all executed pipelines will:

- [✓ fail until the node comes back online]
- switch to another integration runtime
- exceed the CPU limit

The number of concurrent jobs and the CPU usage indicate that the Concurrent Jobs (Running/Limit) value should be:

- raised
- [✓ lowered]
- left as is

Reference:

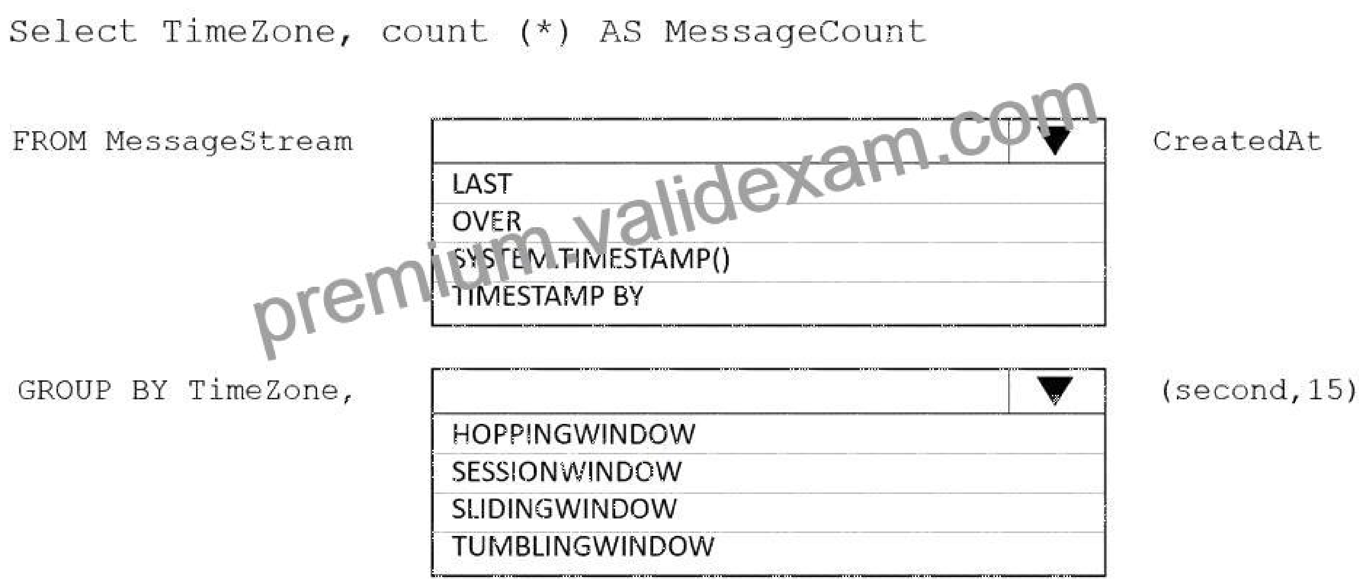
https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime

NO.119 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.

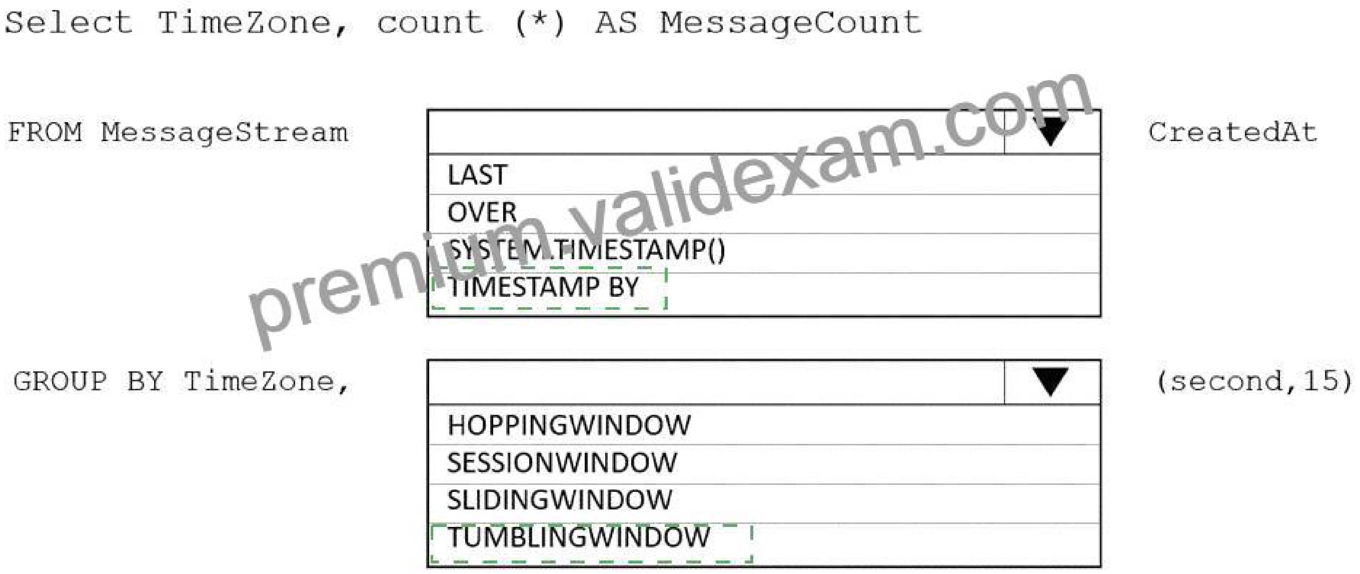
NOTE: Each correct selection is worth one point.



```
Select TimeZone, count (*) AS MessageCount

FROM MessageStream  [ ▼ ]  CreatedAt

GROUP BY TimeZone,  [ ▼ ]  (second,15)
```

- LAST
- OVER
- SYSTEM.TIMESTAMP()
- TIMESTAMP BY

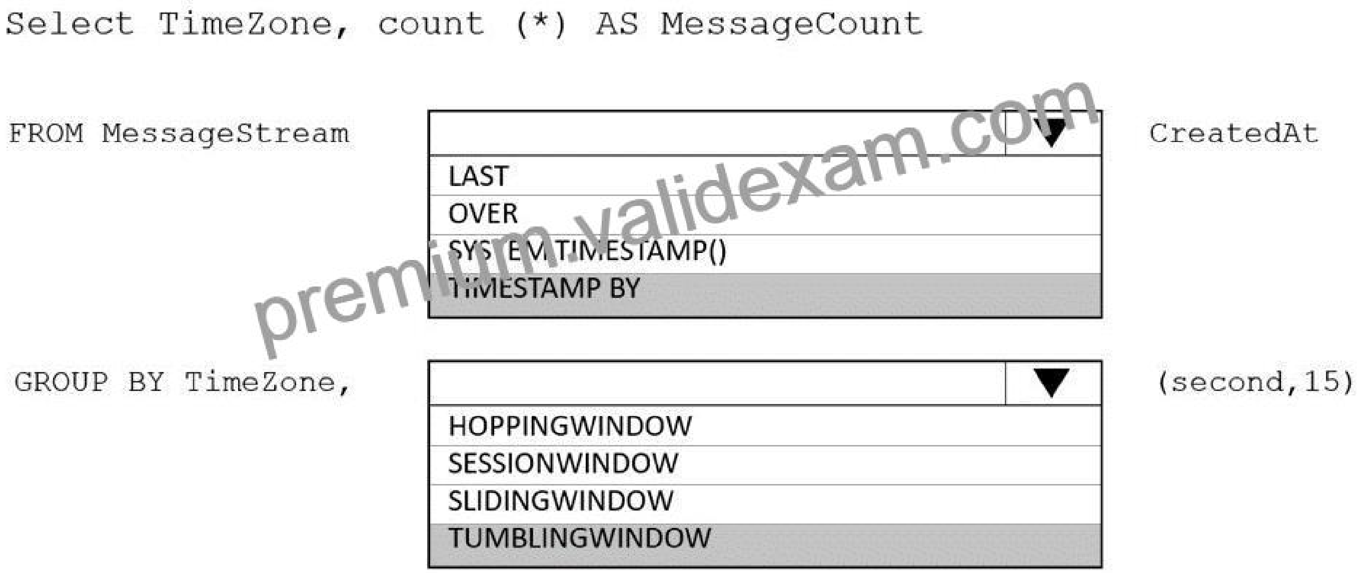
- HOPPINGWINDOW
- SESSIONWINDOW
- SLIDINGWINDOW
- TUMBLINGWINDOW



```
Select TimeZone, count (*) AS MessageCount

FROM MessageStream [          ▼] CreatedAt

GROUP BY TimeZone, [          ▼] (second,15)
```

Explanation

Table Description automatically generated



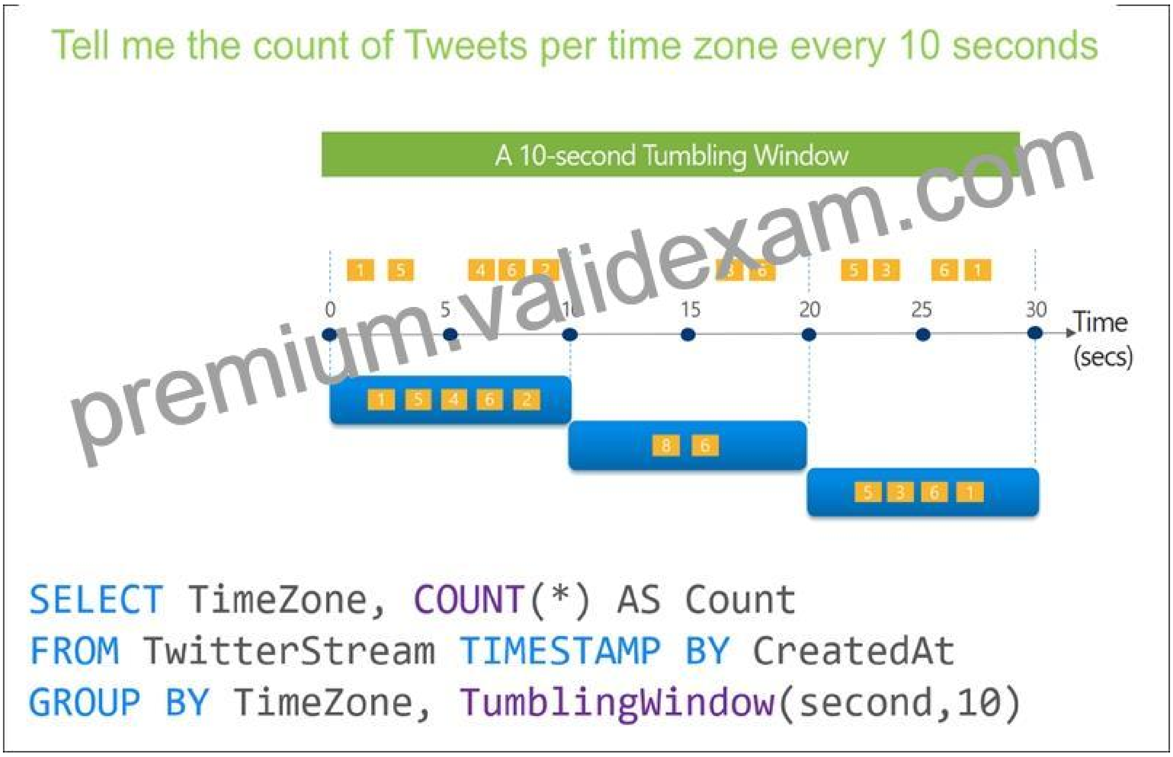
```
Select TimeZone, count (*) AS MessageCount

FROM MessageStream [          ▼] CreatedAt

GROUP BY TimeZone, [          ▼] (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.

Timeline Description automatically generated



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

NO.120 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?
* Change the compatibility level of the Stream Analytics job.
* Increase the number of streaming units (SUs).
* Remove any named consumer groups from the connection and use $default.
* Create an additional output stream for the existing input stream.

Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn&#8217;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

NO.121 You develop a dataset named DBTBL1 by using Azure Databricks.

DBTBL1 contains the following columns:

SensorTypeID

GeographyRegionID

Year

Month

Day

Hour

Minute

Temperature

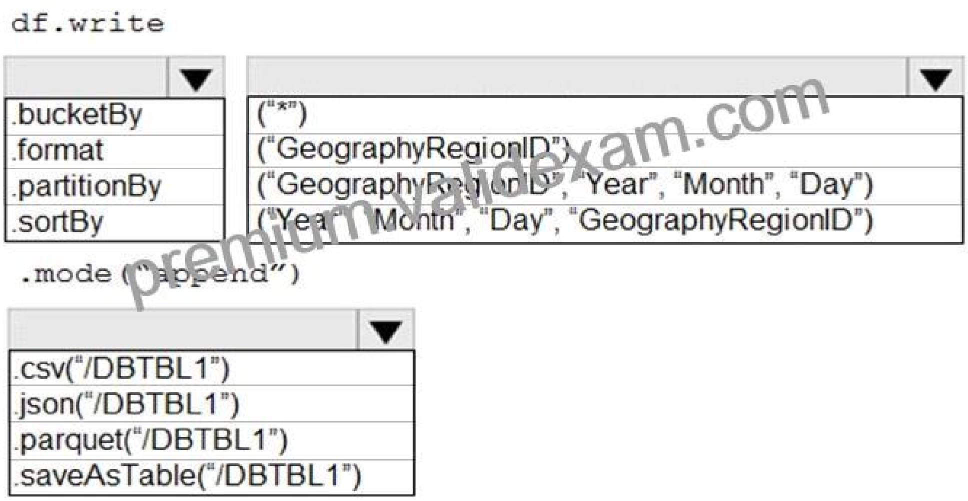
WindSpeed

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.



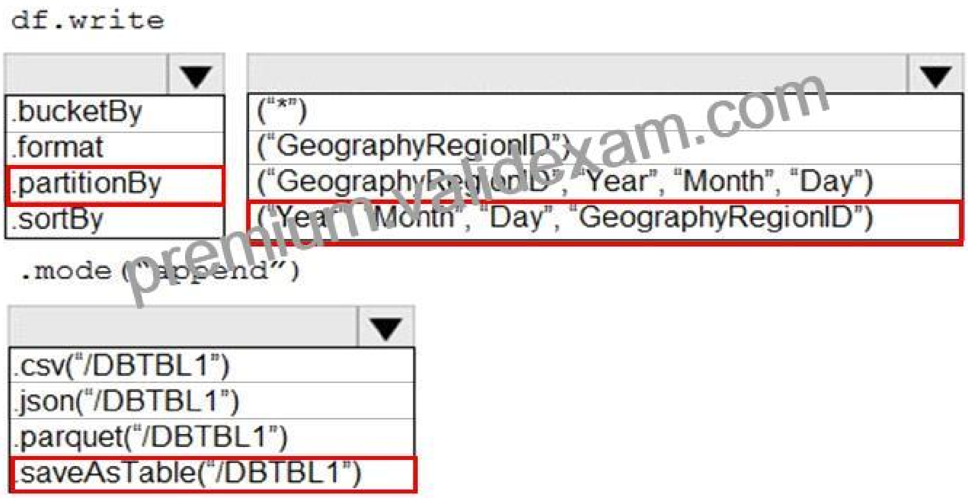
```
df.write
```

.bucketBy
.format
.partitionBy
.sortBy

("*")
("GeographyRegionID")
("GeographyRegionID", "Year", "Month", "Day")
("Year", "Month", "Day", "GeographyRegionID")

```
.mode("append")
```

.csv("/DBTBL1")
.json("/DBTBL1")
.parquet("/DBTBL1")
.saveAsTable("/DBTBL1")



df.write

- .bucketBy
- .format
- [✓ .partitionBy]
- .sortBy

- ("*")
- ("GeographyRegionID")
- ("GeographyRegionID", "Year", "Month", "Day")
- [✓ ("Year", "Month", "Day", "GeographyRegionID")]

.mode ("append")

- .csv("/DBTBL1")
- .json("/DBTBL1")
- .parquet("/DBTBL1")
- [✓ .saveAsTable("/DBTBL1")]

NO.122 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.
* Backlogged Input Events
* Watermark Delay
* Function Events
* Out of order Events
* Late Input Events
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&#8217;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

NO.123 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.

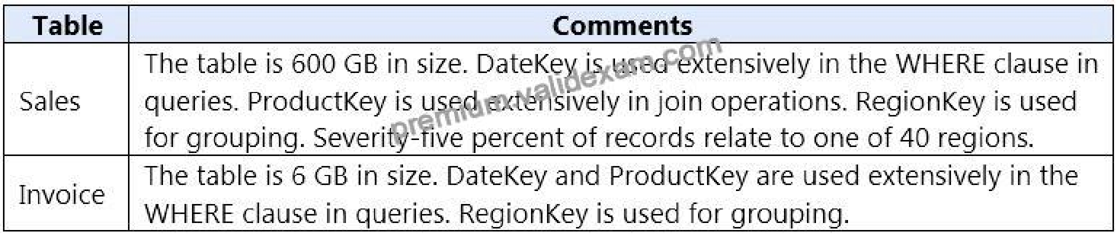


Table	Comments
Sales	The table is 600 GB in size. DateKey is used extensively in the WHERE clause in queries. ProductKey is used extensively in join operations. RegionKey is used for grouping. Seventy-five percent of records relate to one of 40 regions.
Invoice	The table is 6 GB in size. DateKey and 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

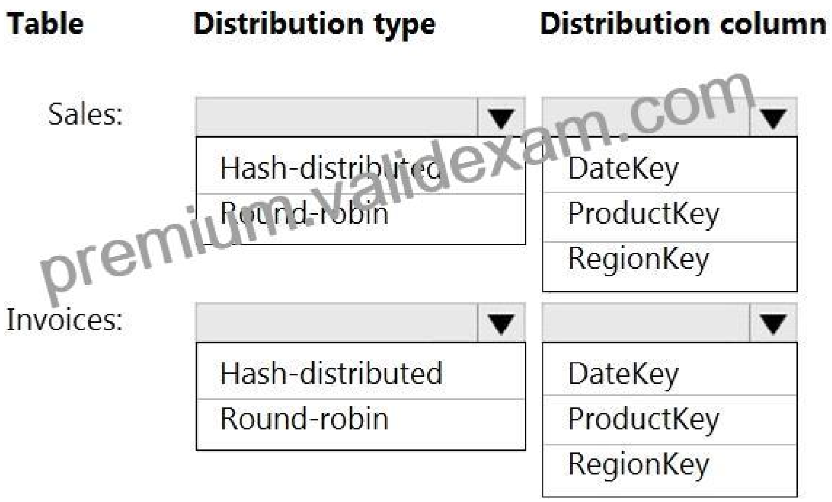


Table	Distribution type	Distribution column
Sales:	Hash-distributed Round-robin	DateKey ProductKey RegionKey
Invoices:	Hash-distributed Round-robin	DateKey ProductKey RegionKey

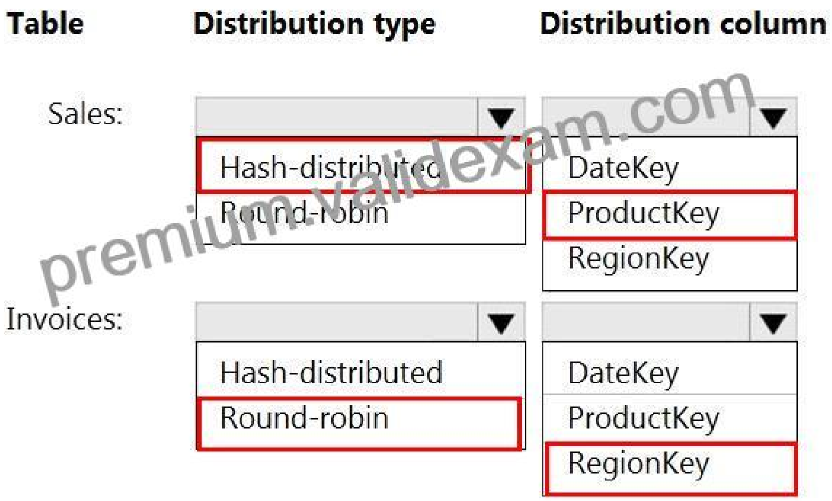
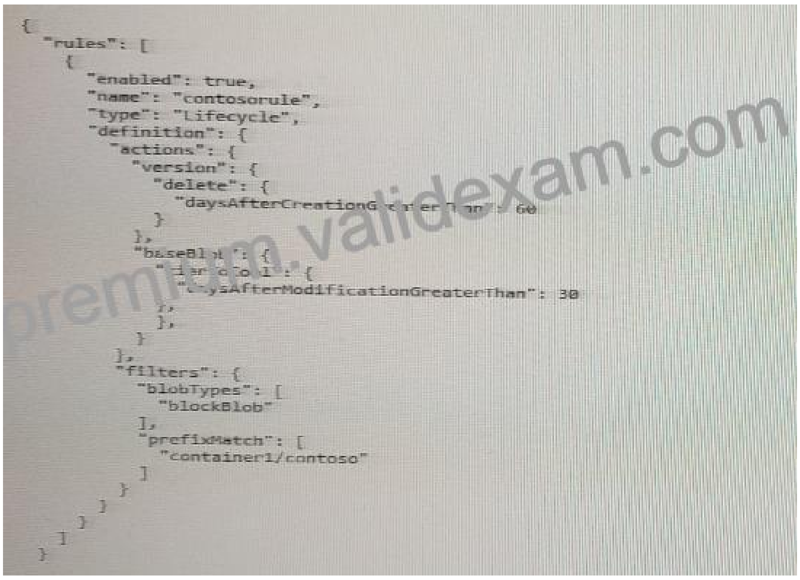


Table	Distribution type	Distribution column
Sales:	[✓ Hash-distributed] Round-robin	DateKey [✓ ProductKey] RegionKey
Invoices:	Hash-distributed [✓ Round-robin]	DateKey ProductKey [✓ RegionKey]

Reference:

https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute

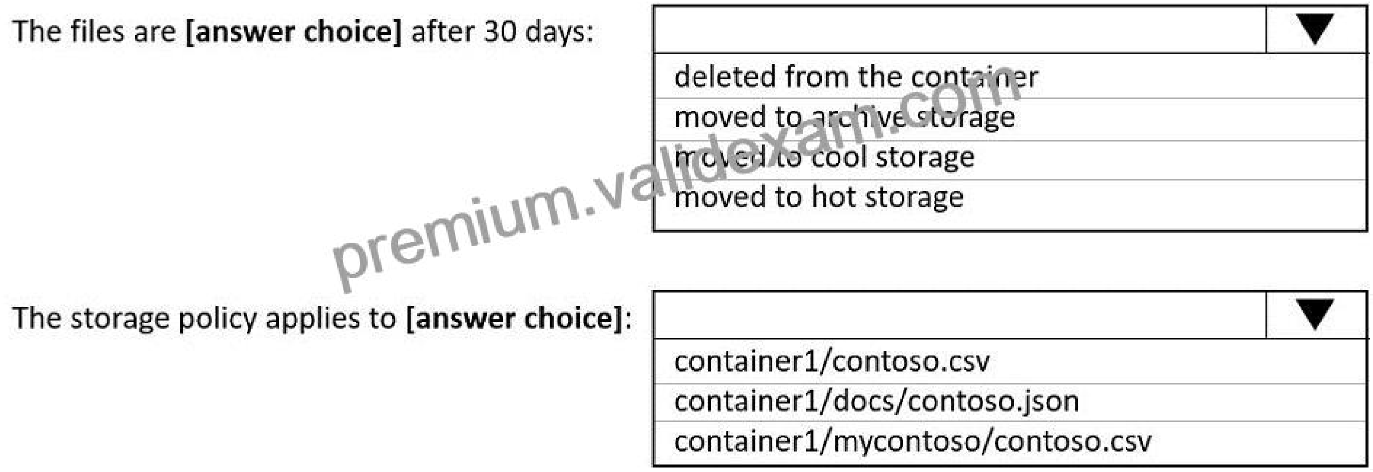
NO.124 You store files in an Azure Data Lake Storage Gen2 container. The container has the storage policy shown in the following exhibit.



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

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

NOTE: Each correct selection Is worth one point.

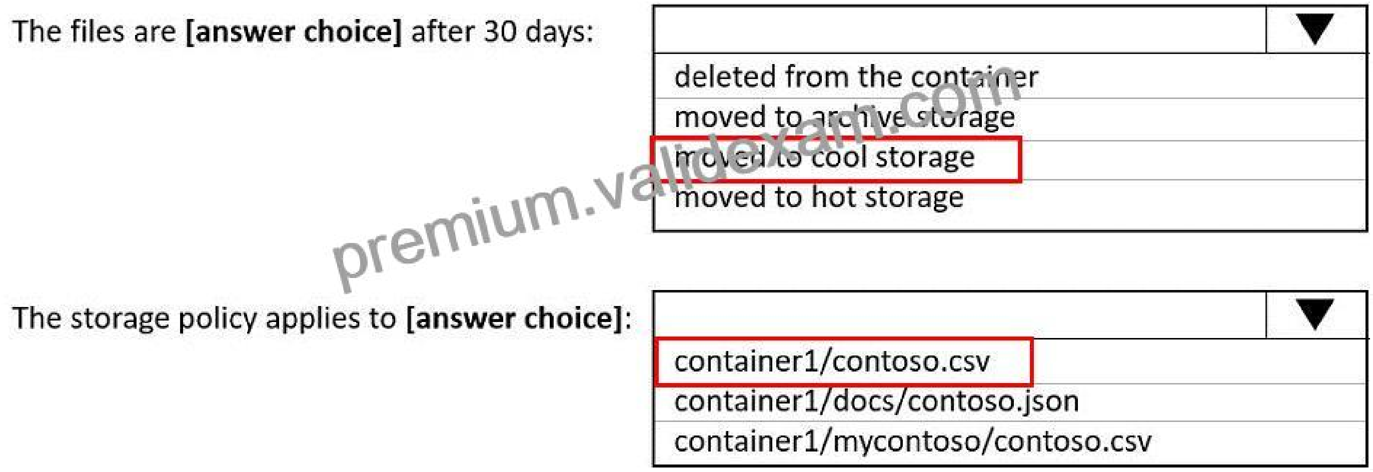


The files are **[answer choice]** after 30 days: ▼

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

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

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



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

- deleted from the container
- moved to archive storage
- [✓ moved to cool storage]
- moved to hot storage

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

- [✓ container1/contoso.csv]
- container1/docs/contoso.json
- container1/mycontoso/contoso.csv

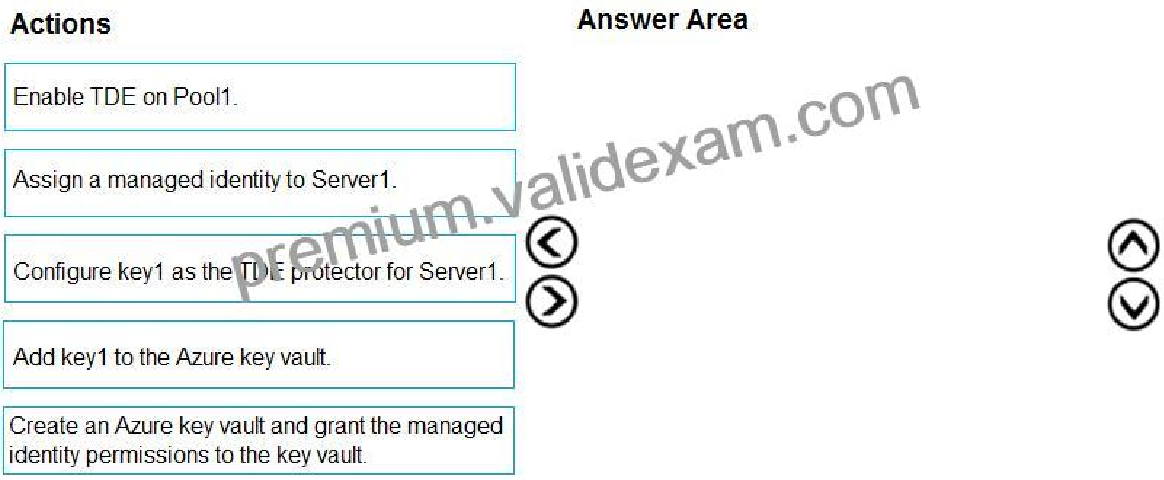
Reference:

https://docs.microsoft.com/en-us/dotnet/api/microsoft.azure.management.storage.fluent.models.managementpolicybaseblob.tiertocool

NO.125 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.



Actions

Enable TDE on Pool1.

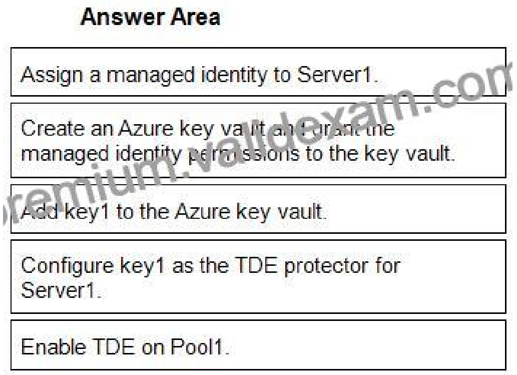
Assign a managed identity to Server1.

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.

Answer Area



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.

1 &#8211; Assign a managed identity to Server1.

2 &#8211; Create an Azure key vault and grant the managed identity permissions to the key vault.

3 &#8211; Add key1 to the Azure key vault.

4 &#8211; Configure key1 as the TDE protector for Server1.

5 &#8211; Enable TDE on Pool1.

Reference:

https://docs.microsoft.com/en-us/azure/azure-sql/managed-instance/scripts/transparent-data-encryption-byok-powershell

NO.126 You plan to implement an Azure Data Lake Storage Gen2 container that will contain CSV files. The size of the files will vary based on the number of events that occur per hour.

File sizes range from 4.KB to 5 GB.

You need to ensure that the files stored in the container are optimized for batch processing.

What should you do?
* Compress the files.
* Merge the files.
* Convert the files to JSON
* Convert the files to Avro.
Explanation

Avro supports batch and is very relevant for streaming.

Note: Avro is framework developed within Apache&#8217;s Hadoop project. It is a row-based storage format which is widely used as a serialization process. AVRO stores its schema in JSON format making it easy to read and interpret by any program. The data itself is stored in binary format by doing it compact and efficient.

Reference:

https://www.adaltas.com/en/2020/07/23/benchmark-study-of-different-file-format/

NO.127 You need to design a data retention solution for the Twitter teed data records. The solution must meet the customer sentiment analytics requirements.

Which Azure Storage functionality should you include in the solution?
* time-based retention
* change feed
* soft delete
* Iifecycle management

NO.128 You are designing a real-time dashboard solution that will visualize streaming data from remote sensors that connect to the internet. The streaming data must be aggregated to show the average value of each 10-second interval. The data will be discarded after being displayed in the dashboard.

The solution will use Azure Stream Analytics and must meet the following requirements:

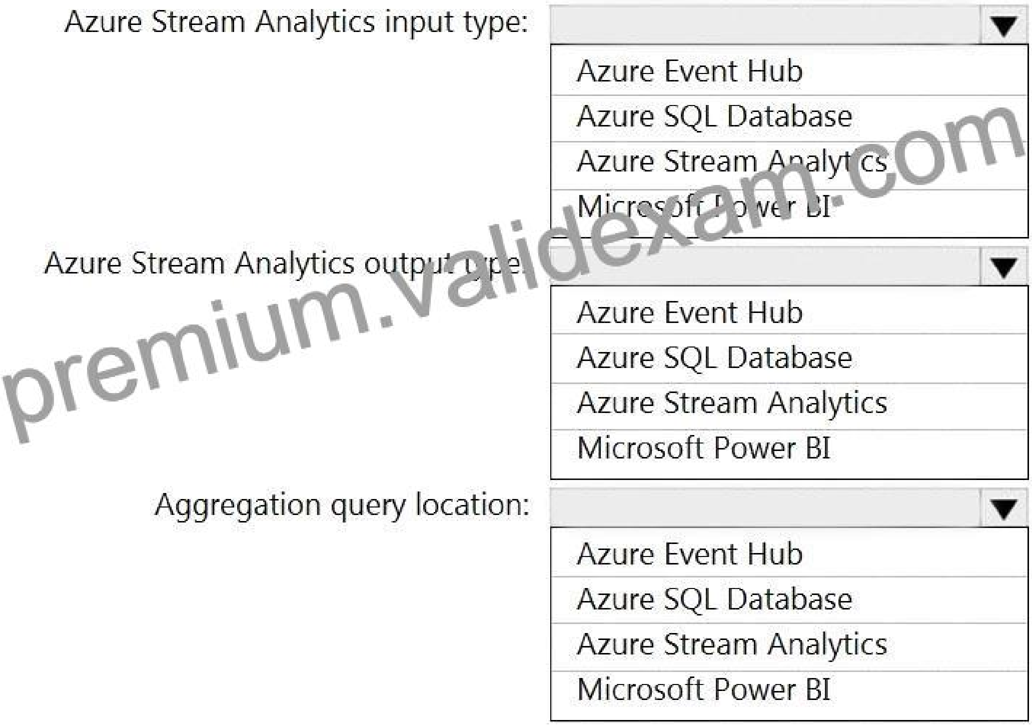
* Minimize latency from an Azure Event hub to the dashboard.

* Minimize the required storage.

* Minimize development effort.

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

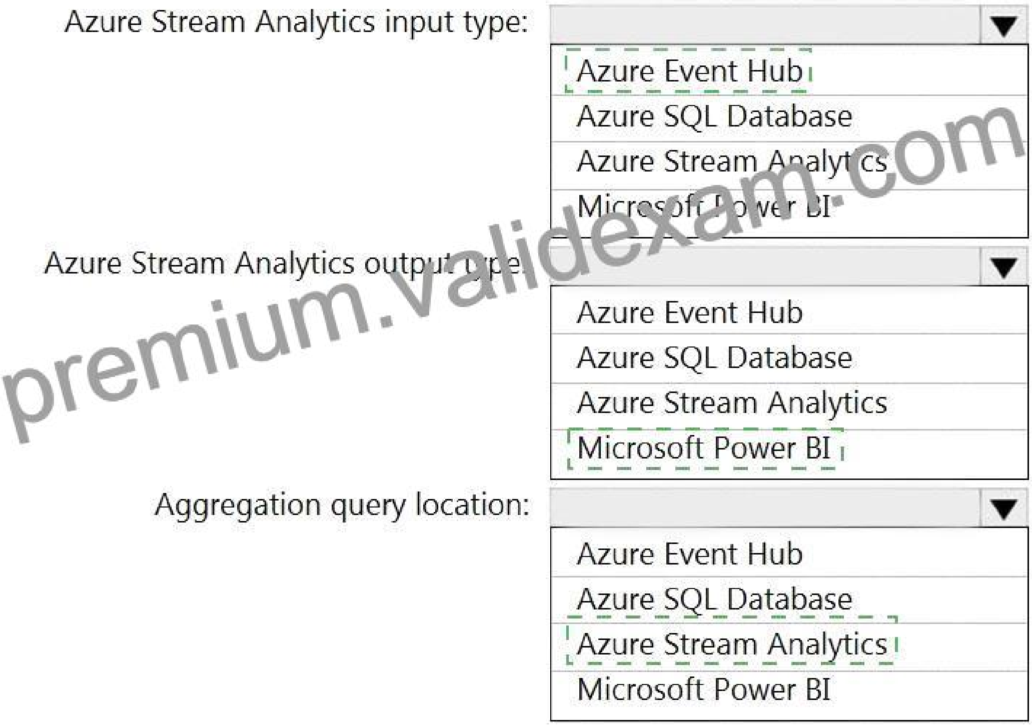
NOTE: Each correct selection is worth one point



Azure Stream Analytics input type:
- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Azure Stream Analytics output type:
- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Aggregation query location:
- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

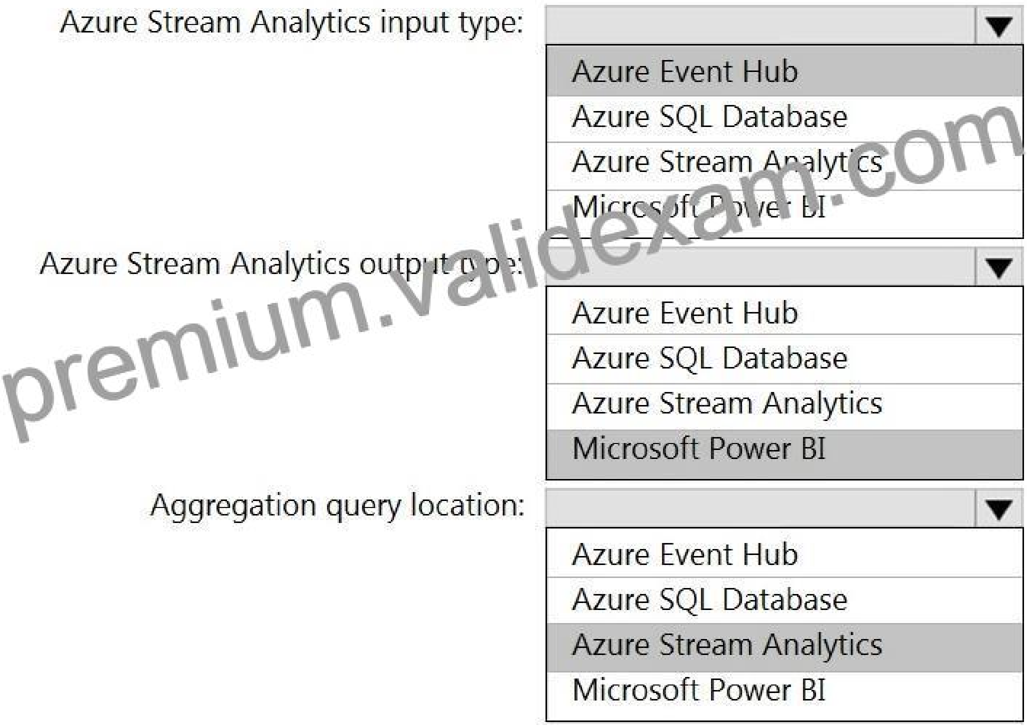


Azure Stream Analytics input type:
- [✓ Azure Event Hub]
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Azure Stream Analytics output type:
- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- [✓ Microsoft Power BI]

Aggregation query location:
- Azure Event Hub
- Azure SQL Database
- [✓ Azure Stream Analytics]
- Microsoft Power BI

Explanation



Azure Stream Analytics input type:
- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Azure Stream Analytics output type:
- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Aggregation query location:
- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Reference:

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-power-bi-dashboard

NO.129 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.

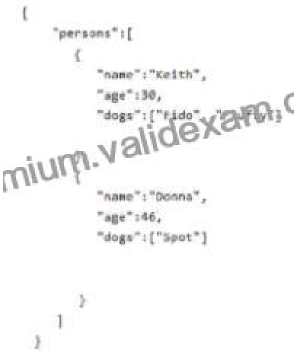
Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data Flow, and then inserts the data info the data warehouse.

Does this meet the goal?
* Yes
* No
If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity, not a mapping flow,5 with your own data processing logic and use the activity in the pipeline. You can create a custom activity to run R scripts on your HDInsight cluster with R installed.

Reference:

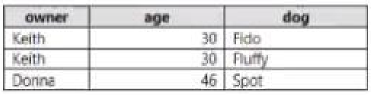
https://docs.microsoft.com/en-US/azure/data-factory/transform-data

NO.130 You use PySpark in Azure Databricks to parse the following JSON input.



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

        }
    ]
}
```

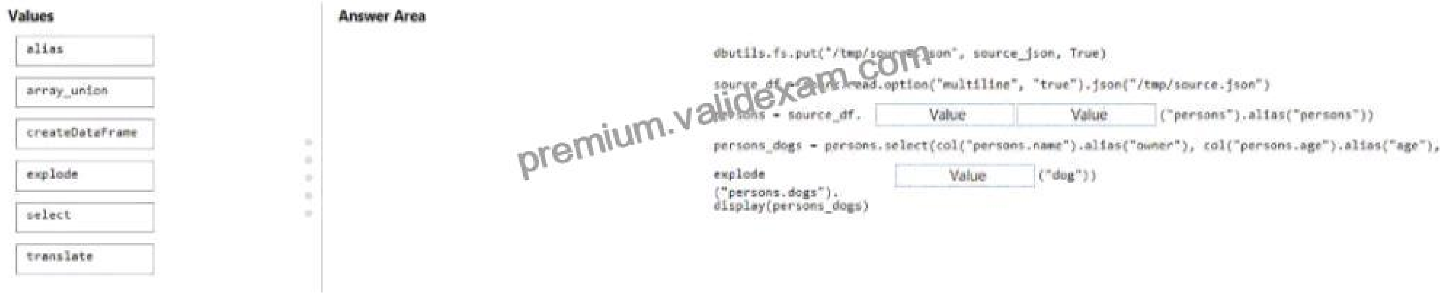
You need to output the data in the following tabular format.



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

How should you complete the PySpark code? To answer, drag the appropriate values to he 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.

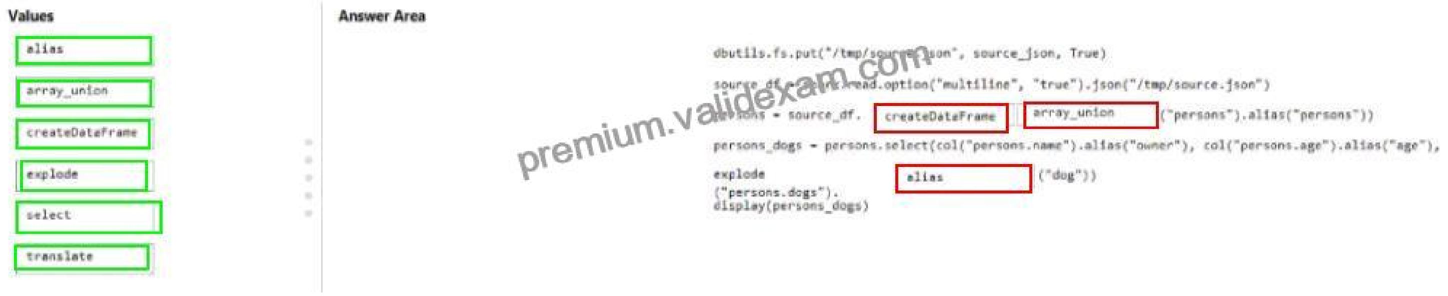


Values

- alias
- array_union
- createDataFrame
- explode
- select
- translate

Answer Area

```
dbutils.fs.put("/tmp/source.json", source_json, True)

source_df = spark.read.option("multiline", "true").json("/tmp/source.json")

persons = source_df. [Value] [Value] ("persons").alias("persons"))

persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"),

explode                [Value] ("dog"))
("persons.dogs").
display(persons_dogs)
```



Values

- alias
- array_union
- createDataFrame
- explode
- select
- translate

Answer Area

```
dbutils.fs.put("/tmp/source.json", source_json, True)

source_df = spark.read.option("multiline", "true").json("/tmp/source.json")

persons = source_df. [createDataFrame] [array_union] ("persons").alias("persons"))

persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"),

explode                [alias] ("dog"))
("persons.dogs").
display(persons_dogs)
```

NO.131 You configure monitoring for a Microsoft Azure SQL Data Warehouse implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Gen 2 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?

* EXTERNAL TABLE access failed due to internal error: &#8216;Java exception raised on call to HdfsBridge_Connect: Error

[com.microsoft.polybase.client.KerberosSecureLogin] occurred while accessing external files.&#8217;
* EXTERNAL TABLE access failed due to internal error: &#8216;Java exception raised on call to HdfsBridge_Connect: Error [No FileSystem for scheme: wasbs] occurred while accessing external file.&#8217;
* Cannot execute the query &#8220;Remote Query&#8221; against OLE DB provider &#8220;SQLNCLI11&#8221;: for linked server

&#8220;(null)&#8221;, Query aborted- the maximum reject threshold (o

rows) was reached while regarding from an external source: 1 rows rejected out of total 1 rows processed.
* EXTERNAL TABLE access failed due to internal error: &#8216;Java exception raised on call to HdfsBridge_Connect: Error [Unable to instantiate LoginClass] occurred while accessing external files.&#8217;

Explanation

Customer Scenario:

SQL Server 2016 or SQL DW connected to Azure blob storage. The CREATE EXTERNAL TABLE DDL points to a directory (and not a specific file) and the directory contains files with different schemas.

SSMS Error:

Select query on the external table gives the following error:

Msg 7320, Level 16, State 110, Line 14

Cannot execute the query &#8220;Remote Query&#8221; against OLE DB provider &#8220;SQLNCLI11&#8221; for linked server &#8220;(null)&#8221;.

Query aborted&#8211; the maximum reject threshold (0 rows) was reached while reading from an external source: 1 rows rejected out of total 1 rows processed.

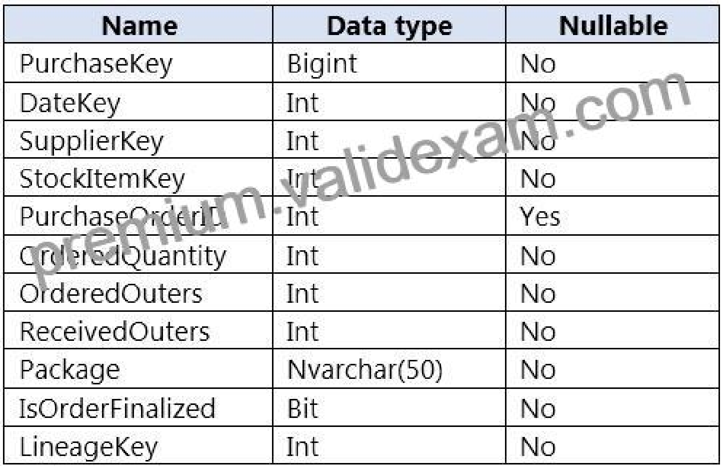
Possible Reason:

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.

Possible Solution:

If the data for each table consists of one file, then use the filename in the LOCATION section prepended by the directory of the external files. If there are multiple files per table, put each set of files into different directories in Azure Blob Storage and then you can point LOCATION to the directory instead of a particular file. The latter suggestion is the best practices recommended by SQLCAT even if you have one file per table.

NO.132 You are designing a fact table named FactPurchase in an Azure Synapse Analytics dedicated SQL pool. The table contains purchases from suppliers for a retail store. FactPurchase will contain the following columns.



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

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

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

SELECT

SupplierKey, StockItemKey, COUNT(*)

FROM FactPurchase

WHERE DateKey >= 20210101

AND DateKey <= 20210131

GROUP By SupplierKey, StockItemKey

Which table distribution will minimize query times?
* round-robin
* replicated
* hash-distributed on DateKey
* hash-distributed on PurchaseKey

Explanation

Hash-distributed tables improve query performance on large fact tables, and are the focus of this article.

Round-robin tables are useful for improving loading speed.

Reference:

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

NO.133 You have an Azure subscription.

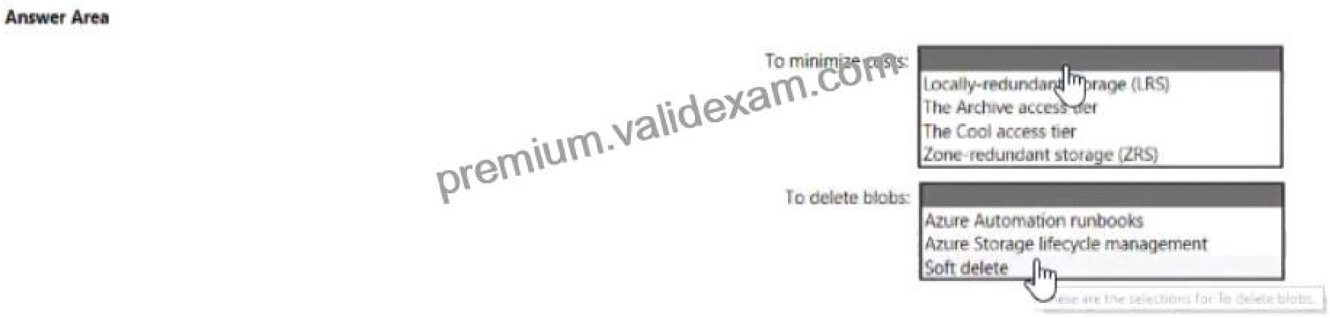
You need to deploy an Azure Data Lake Storage Gen2 Premium account. The solution must meet the following requirements:

* Blobs that are older than 365 days must be deleted.

* Administrator efforts must be minimized.

* Costs must be minimized

What should you use? To answer, select the appropriate options in the answer are a. NOTE Each correct selection is worth one point.

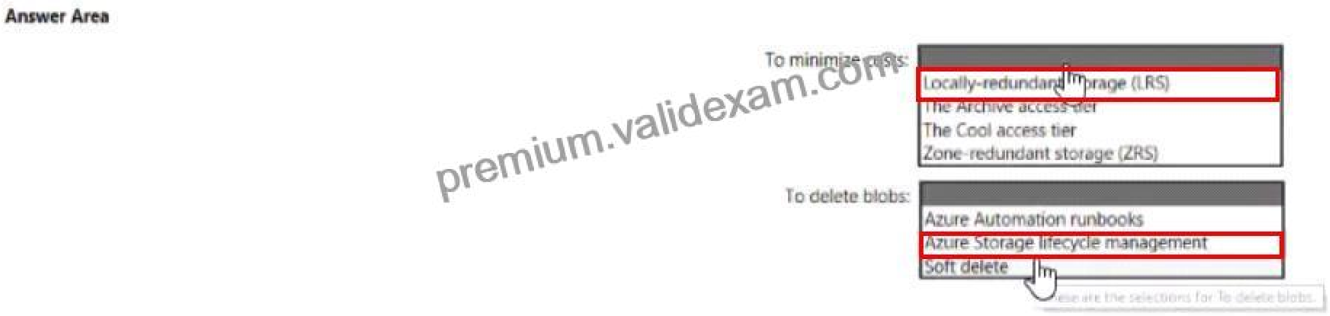


Answer Area

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

These are the selections for To delete blobs.



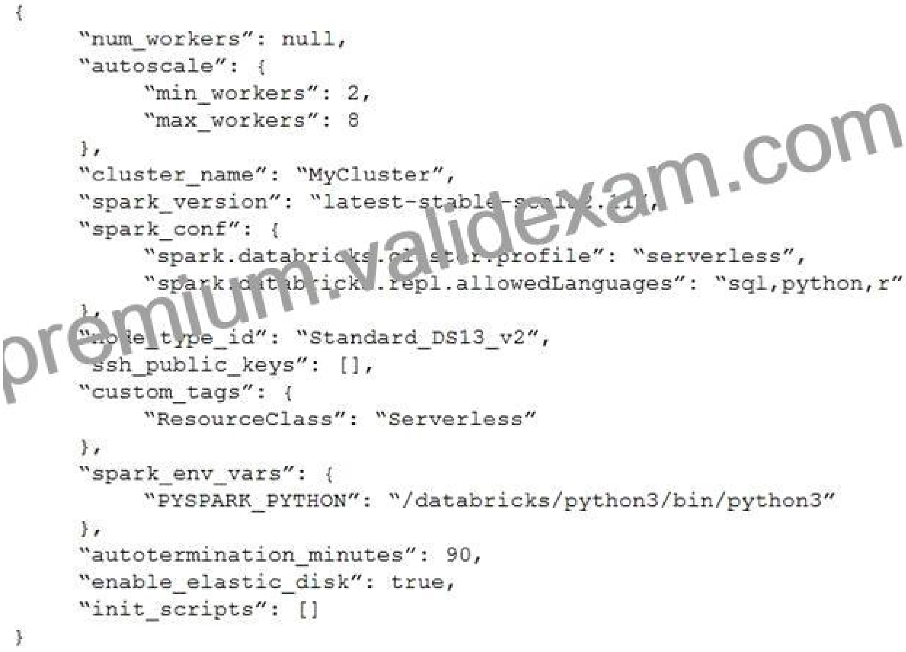
Answer Area

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

These are the selections for To delete blobs.

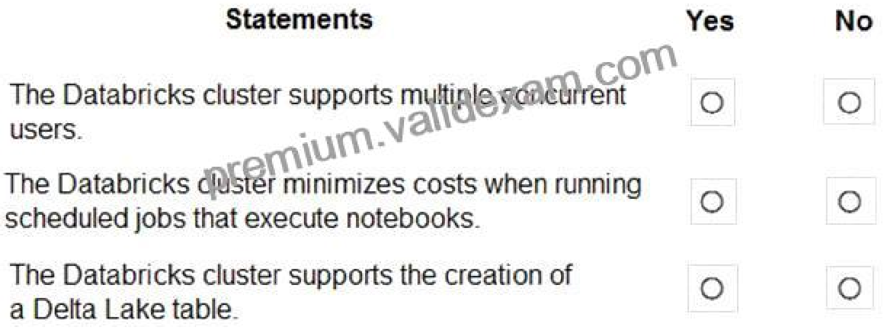
NO.134 The following code segment is used to create an Azure Databricks cluster.



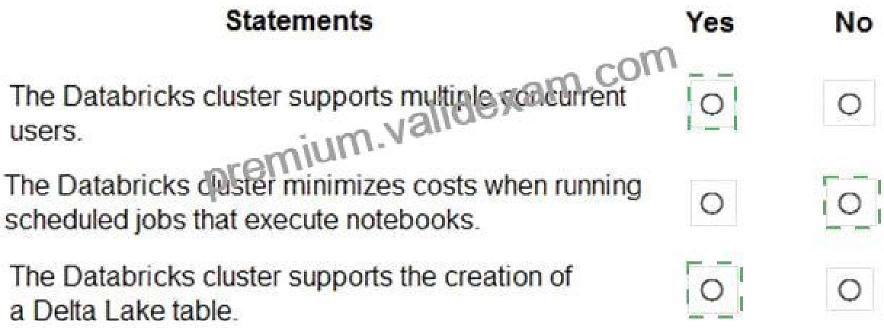
```
{
     "num_workers": null,
     "autoscale": {
          "min_workers": 2,
          "max_workers": 8
     },
     "cluster_name": "MyCluster",
     "spark_version": "latest-stable-scala2.11",
     "spark_conf": {
          "spark.databricks.cluster.profile": "serverless",
          "spark.databricks.repl.allowedLanguages": "sql,python,r"
     },
     "node_type_id": "Standard_DS13_v2",
     "ssh_public_keys": [],
     "custom_tags": {
          "ResourceClass": "Serverless"
     },
     "spark_env_vars": {
          "PYSPARK_PYTHON": "/databricks/python3/bin/python3"
     },
     "autotermination_minutes": 90,
     "enable_elastic_disk": true,
     "init_scripts": []
}
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.



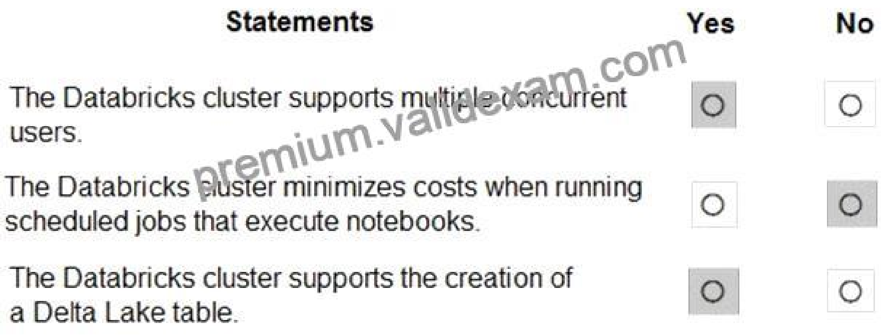
Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	○	○
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	○	○
The Databricks cluster supports the creation of a Delta Lake table.	○	○



Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	[✓ ○]	○
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	○	[✓ ○]
The Databricks cluster supports the creation of a Delta Lake table.	[✓ ○]	○

Explanation

Graphical user interface, text, application Description automatically generated



Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	[✓ ○]	○
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	○	[✓ ○]
The Databricks cluster supports the creation of a Delta Lake table.	[✓ ○]	○

Box 1: Yes

A cluster mode of &#8216;High Concurrency&#8217; is selected, unlike all the others which are &#8216;Standard&#8217;. This results in a worker type of Standard_DS13_v2.

Box 2: No

When you run a job on a new cluster, the job is treated as a data engineering (job) workload subject to the job workload pricing. When you run a job on an existing cluster, the job is treated as a data analytics (all-purpose) workload subject to all-purpose workload pricing.

Box 3: Yes

Delta Lake on Databricks allows you to configure Delta Lake based on your workload patterns.

Reference:

https://adatis.co.uk/databricks-cluster-sizing/

https://docs.microsoft.com/en-us/azure/databricks/jobs

https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html

https://docs.databricks.com/delta/index.html

NO.135 You develop a dataset named DBTBL1 by using Azure Databricks.

DBTBL1 contains the following columns:

SensorTypeID

GeographyRegionID

Year

Month

Day

Hour

Minute

Temperature

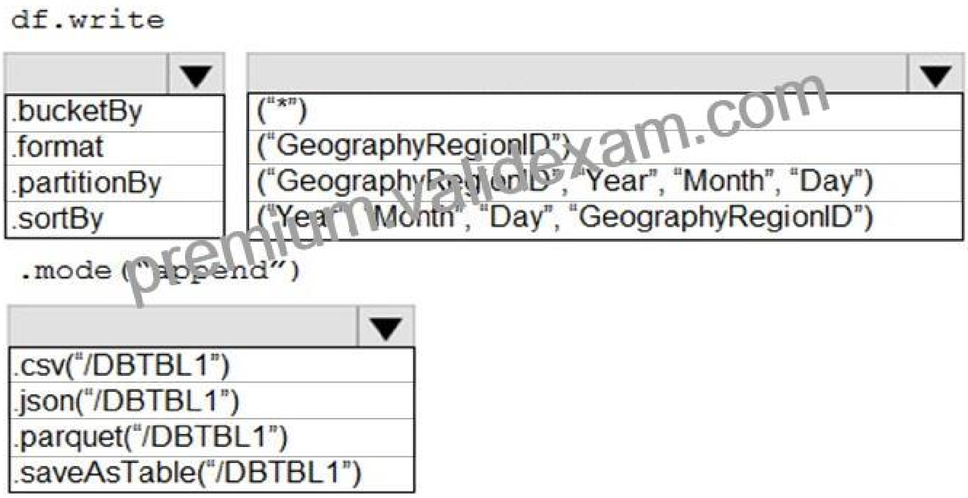
WindSpeed

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.



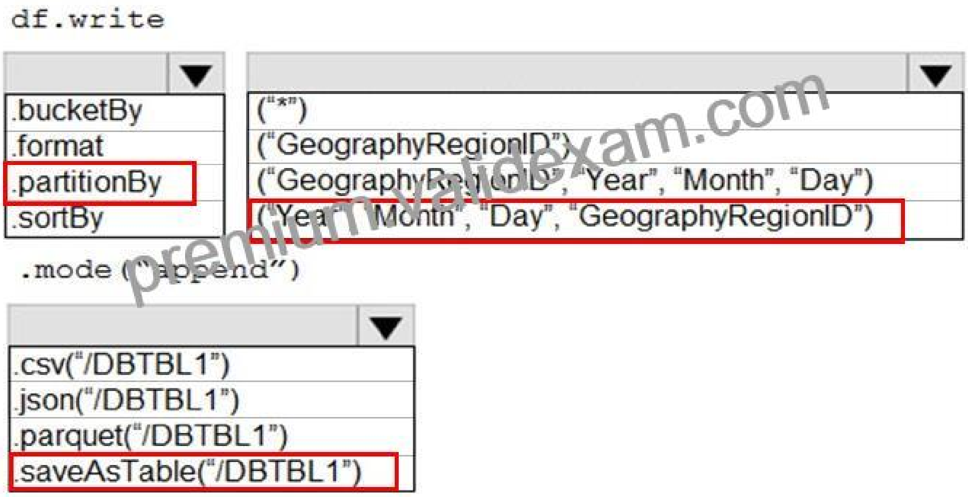
```
df.write
```

- .bucketBy
- .format
- .partitionBy
- .sortBy

- (“*”)
- (“GeographyRegionID”)
- (“GeographyRegionID”, “Year”, “Month”, “Day”)
- (“Year”, “Month”, “Day”, “GeographyRegionID”)

```
.mode (“append”)
```

- .csv(“/DBTBL1”)
- .json(“/DBTBL1”)
- .parquet(“/DBTBL1”)
- .saveAsTable(“/DBTBL1”)



```
df.write
```

- .bucketBy
- .format
- [✓ .partitionBy]
- .sortBy

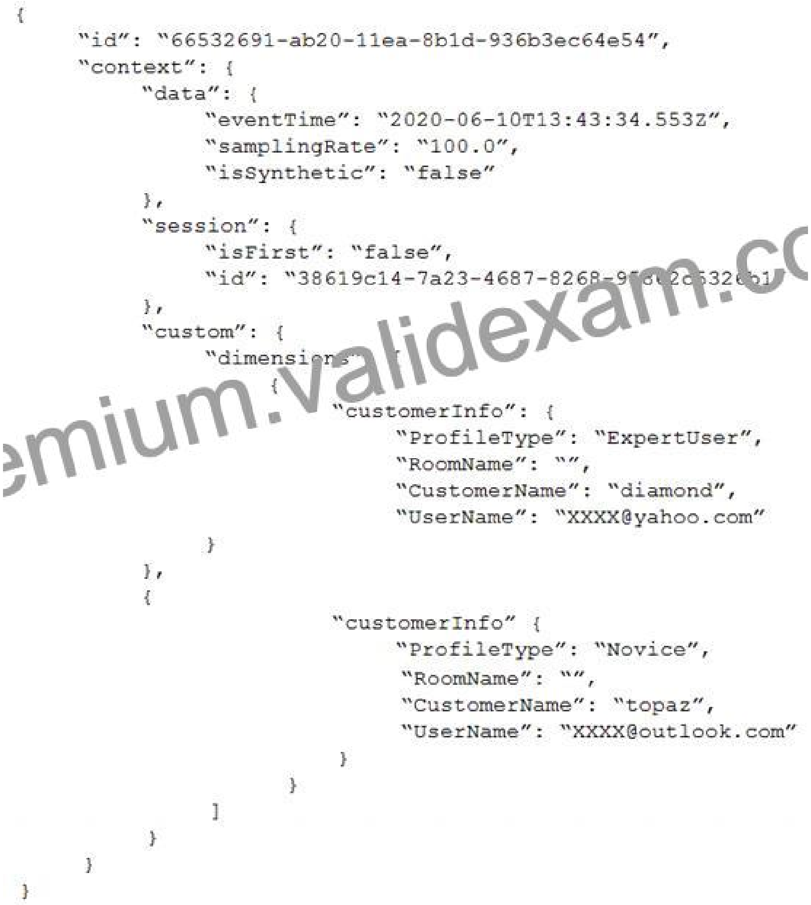
- (“*”)
- (“GeographyRegionID”)
- (“GeographyRegionID”, “Year”, “Month”, “Day”)
- [✓ (“Year”, “Month”, “Day”, “GeographyRegionID”)]

```
.mode (“append”)
```

- .csv(“/DBTBL1”)
- .json(“/DBTBL1”)
- .parquet(“/DBTBL1”)
- [✓ .saveAsTable(“/DBTBL1”)]

NO.136 You have an Azure Synapse Analytics workspace named WS1.

You have an Azure Data Lake Storage Gen2 container that contains JSON-formatted files in the following format.

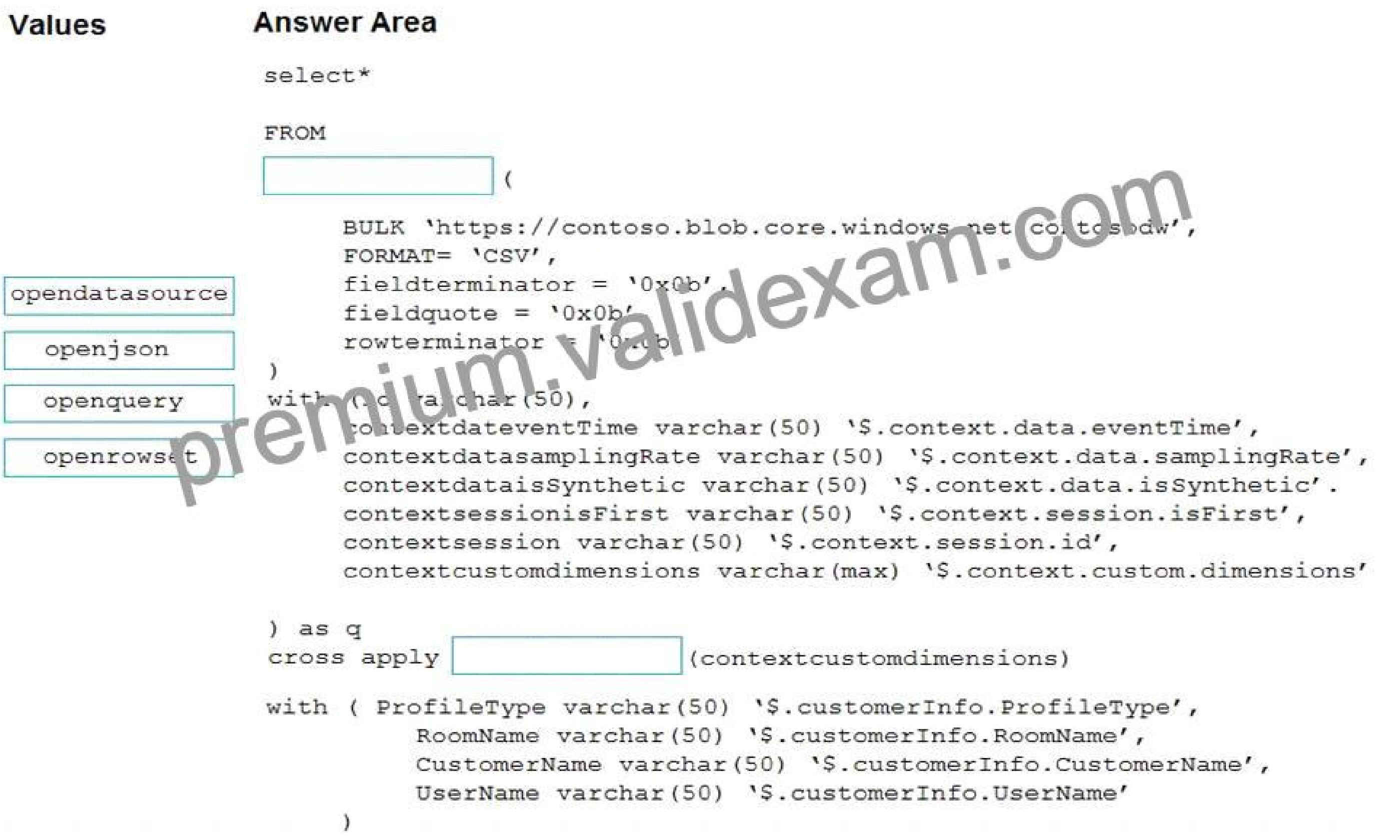


```
{
      “id”: “66532691-ab20-11ea-8b1d-936b3ec64e54”,
      “context”: {
            “data”: {
                  “eventTime”: “2020-06-10T13:43:34.553Z”,
                  “samplingRate”: “100.0”,
                  “isSynthetic”: “false”
            },
            “session”: {
                  “isFirst”: “false”,
                  “id”: “38619c14-7a23-4687-8268-9f8e2c6326b1
            },
            “custom”: {
                  “dimensions”: [
                        {
                              “customerInfo”: {
                                    “ProfileType”: “ExpertUser”,
                                    “RoomName”: “”,
                                    “CustomerName”: “diamond”,
                                    “UserName”: “XXXX@yahoo.com”
                  }
            },
            {
                              “customerInfo” {
                                    “ProfileType”: “Novice”,
                                    “RoomName”: “”,
                                    “CustomerName”: “topaz”,
                                    “UserName”: “XXXX@outlook.com”
                               }
                        }
                  ]
            }
      }
}
```

You need to use the serverless SQL pool in WS1 to read the files.

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

NOTE: Each correct selection is worth one point.



Values

opendatasource

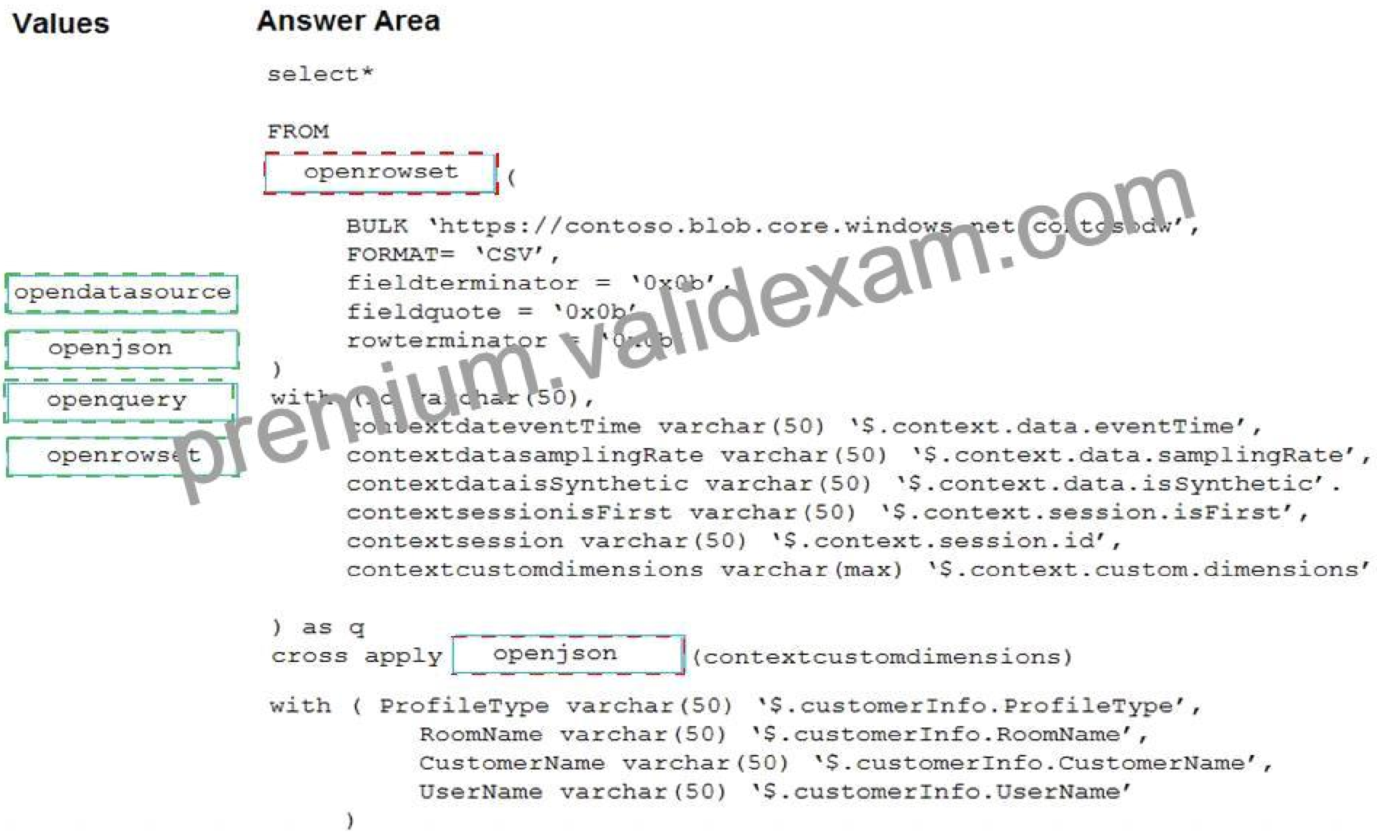
openjson

openquery

openrowset

Answer Area

```
select*

FROM
[          ] (

      BULK ‘https://contoso.blob.core.windows.net/contosodw’,
      FORMAT= ‘CSV’,
      fieldterminator = ‘0x0b’,
      fieldquote = ‘0x0b’,
      rowterminator = ‘0x0b’
)
with (id varchar(50),
      contextdateventTime varchar(50) ‘$.context.data.eventTime’,
      contextdatasamplingRate varchar(50) ‘$.context.data.samplingRate’,
      contextdataisSynthetic varchar(50) ‘$.context.data.isSynthetic’.
      contextsessionisFirst varchar(50) ‘$.context.session.isFirst’,
      contextsession varchar(50) ‘$.context.session.id’,
      contextcustomdimensions varchar(max) ‘$.context.custom.dimensions’

) as q
cross apply [          ] (contextcustomdimensions)

with ( ProfileType varchar(50) ‘$.customerInfo.ProfileType’,
         RoomName varchar(50) ‘$.customerInfo.RoomName’,
         CustomerName varchar(50) ‘$.customerInfo.CustomerName’,
         UserName varchar(50) ‘$.customerInfo.UserName’
     )
```



Values

opendatasource

openjson

openquery

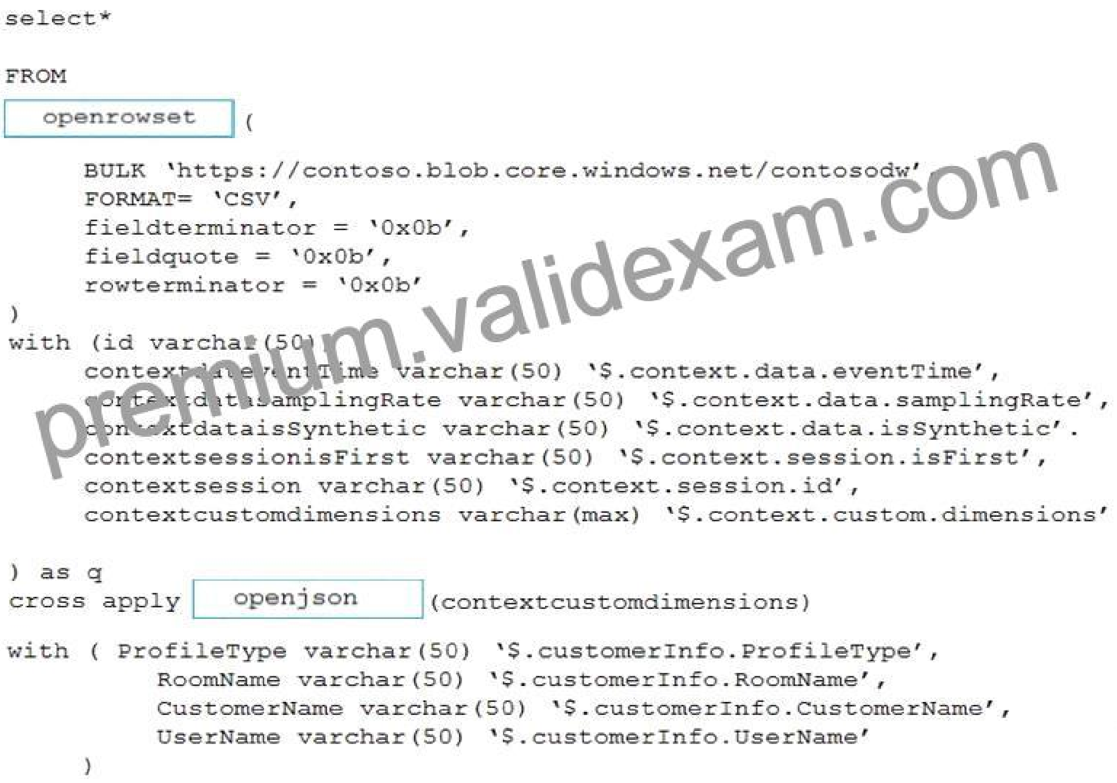
openrowset

Answer Area

```
select*

FROM
[ openrowset ] (

      BULK ‘https://contoso.blob.core.windows.net/contosodw’,
      FORMAT= ‘CSV’,
      fieldterminator = ‘0x0b’,
      fieldquote = ‘0x0b’,
      rowterminator = ‘0x0b’
)
with (id varchar(50),
      contextdateventTime varchar(50) ‘$.context.data.eventTime’,
      contextdatasamplingRate varchar(50) ‘$.context.data.samplingRate’,
      contextdataisSynthetic varchar(50) ‘$.context.data.isSynthetic’.
      contextsessionisFirst varchar(50) ‘$.context.session.isFirst’,
      contextsession varchar(50) ‘$.context.session.id’,
      contextcustomdimensions varchar(max) ‘$.context.custom.dimensions’

) as q
cross apply [ openjson ] (contextcustomdimensions)

with ( ProfileType varchar(50) ‘$.customerInfo.ProfileType’,
         RoomName varchar(50) ‘$.customerInfo.RoomName’,
         CustomerName varchar(50) ‘$.customerInfo.CustomerName’,
         UserName varchar(50) ‘$.customerInfo.UserName’
     )
```

Explanation

Graphical user interface, text, application, email Description automatically generated



```
select*

FROM
 openrowset (

     BULK 'https://contoso.blob.core.windows.net/contosodw',
     FORMAT= 'CSV',
     fieldterminator = '0x0b',
     fieldquote = '0x0b',
     rowterminator = '0x0b'
)
with (id varchar(50),
     contextdataeventTime varchar(50) '$.context.data.eventTime',
     contextdatasamplingRate varchar(50) '$.context.data.samplingRate',
     contextdataisSynthetic varchar(50) '$.context.data.isSynthetic'.
     contextsessionisFirst varchar(50) '$.context.session.isFirst',
     contextsession varchar(50) '$.context.session.id',
     contextcustomdimensions varchar(max) '$.context.custom.dimensions'

) as q
cross apply  openjson  (contextcustomdimensions)

with ( ProfileType varchar(50) '$.customerInfo.ProfileType',
         RoomName varchar(50) '$.customerInfo.RoomName',
         CustomerName varchar(50) '$.customerInfo.CustomerName',
         UserName varchar(50) '$.customerInfo.UserName'
     )
```

Box 1: openrowset

The easiest way to see to the content of your CSV file is to provide file URL to OPENROWSET function, specify csv FORMAT.

Example:

SELECT *

FROM OPENROWSET(

BULK &#8216;csv/population/population.csv&#8217;,

DATA_SOURCE = &#8216;SqlOnDemandDemo&#8217;,

FORMAT = &#8216;CSV&#8217;, PARSER_VERSION = &#8216;2.0&#8217;,

FIELDTERMINATOR =&#8217;,&#8217;,

ROWTERMINATOR = &#8216;n&#8217;

Box 2: openjson

You can access your JSON files from the Azure File Storage share by using the mapped drive, as shown in the following example:

SELECT book.* FROM

OPENROWSET(BULK N&#8217;t:booksbooks.json&#8217;, SINGLE_CLOB) AS json

CROSS APPLY OPENJSON(BulkColumn)

WITH(id nvarchar(100), name nvarchar(100), price float,

pages_i int, author nvarchar(100)) AS book

Reference:

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-single-csv-file

https://docs.microsoft.com/en-us/sql/relational-databases/json/import-json-documents-into-sql-server

NO.137 You plan to implement an Azure Data Lake Storage Gen2 container that will contain CSV files. The size of the files will vary based on the number of events that occur per hour.

File sizes range from 4.KB to 5 GB.

You need to ensure that the files stored in the container are optimized for batch processing.

What should you do?
* Compress the files.
* Merge the files.
* Convert the files to JSON
* Convert the files to Avro.

Avro supports batch and is very relevant for streaming.

Note: Avro is framework developed within Apache&#8217;s Hadoop project. It is a row-based storage format which is widely used as a serialization process. AVRO stores its schema in JSON format making it easy to read and interpret by any program. The data itself is stored in binary format by doing it compact and efficient.

Reference:

https://www.adaltas.com/en/2020/07/23/benchmark-study-of-different-file-format/

NO.138 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.

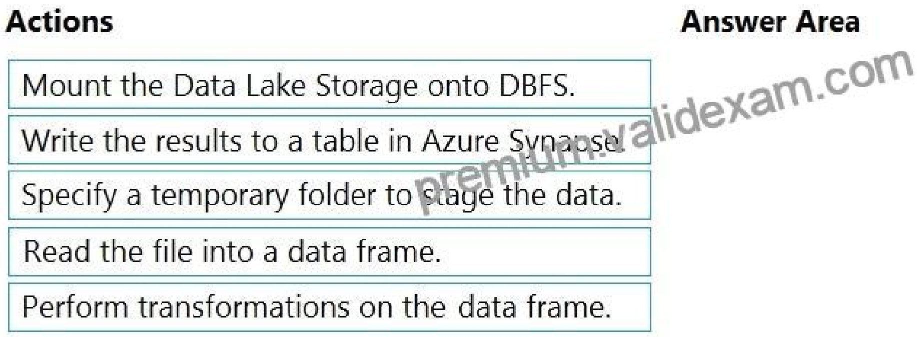
You create the following components:

A destination table in Azure Synapse

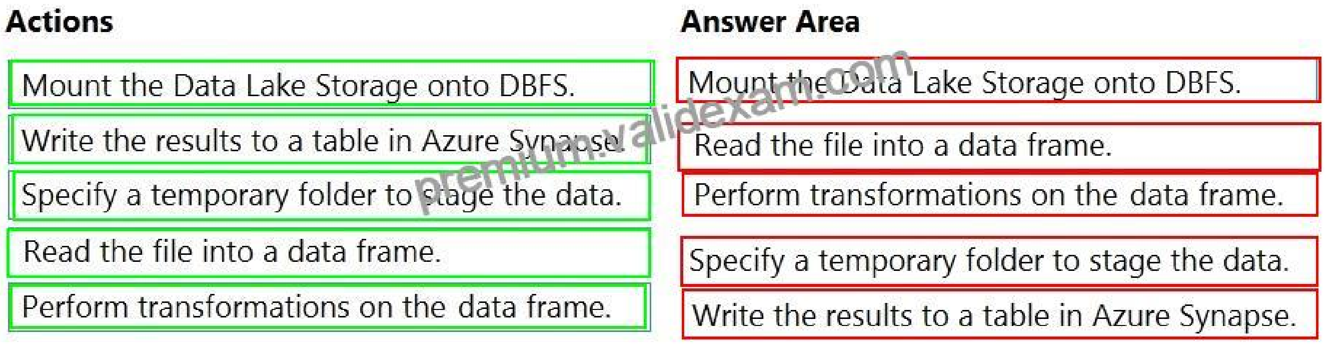
An Azure Blob storage container

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.



Actions	Answer Area
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.	



Actions	Answer Area
Mount the Data Lake Storage onto DBFS.	Mount the Data Lake Storage onto DBFS.
Write the results to a table in Azure Synapse.	Read the file into a data frame.
Specify a temporary folder to stage the data.	Perform transformations on the data frame.
Read the file into a data frame.	Specify a temporary folder to stage the data.
Perform transformations on the data frame.	Write the results to a table in Azure Synapse.

Reference:

https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-extract-load-sql-data-warehouse

NO.139 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 scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You convert the files to compressed delimited text files.

Does this meet the goal?
* Yes
* No
Explanation

All file formats have different performance characteristics. For the fastest load, use compressed delimited text files.

Reference:

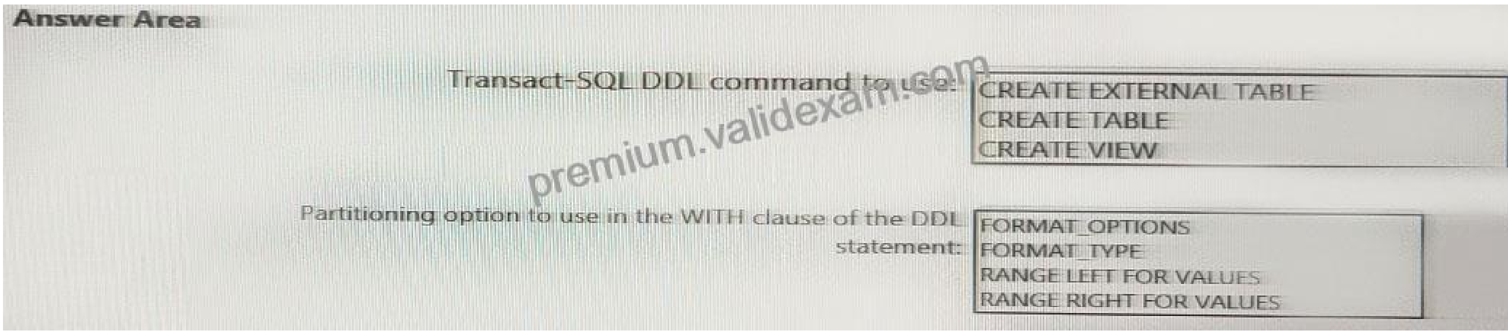
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

NO.140 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 solution must meet the sales transaction dataset requirements.

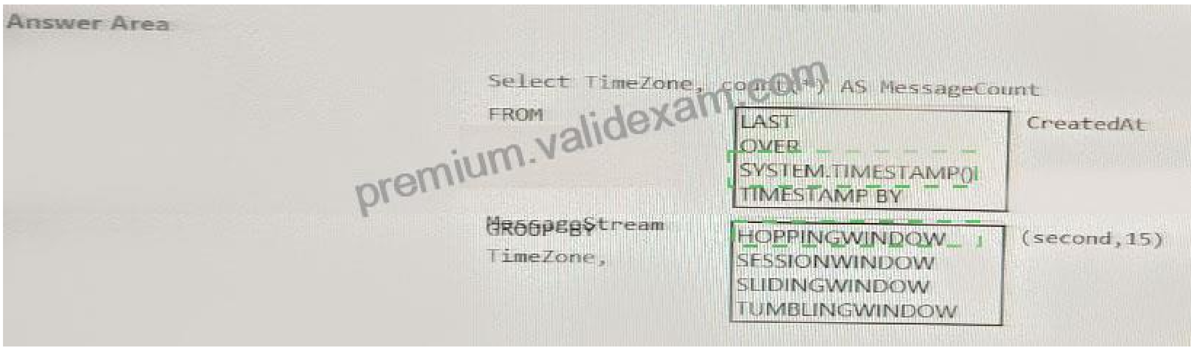
What should you do? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.


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


Answer Area

```
Select TimeZone, count(*) AS MessageCount
FROM                         [LAST | OVER | SYSTEM.TIMESTAMP() | TIMESTAMP BY]   CreatedAt
MessageStream
GROUP BY
TimeZone,                    [HOPPINGWINDOW | SESSIONWINDOW | SLIDINGWINDOW | TUMBLINGWINDOW]   (second,15)
```

Fully Updated Free Actual Microsoft DP-203 Exam Questions: https://www.validexam.com/DP-203-latest-dumps.html]