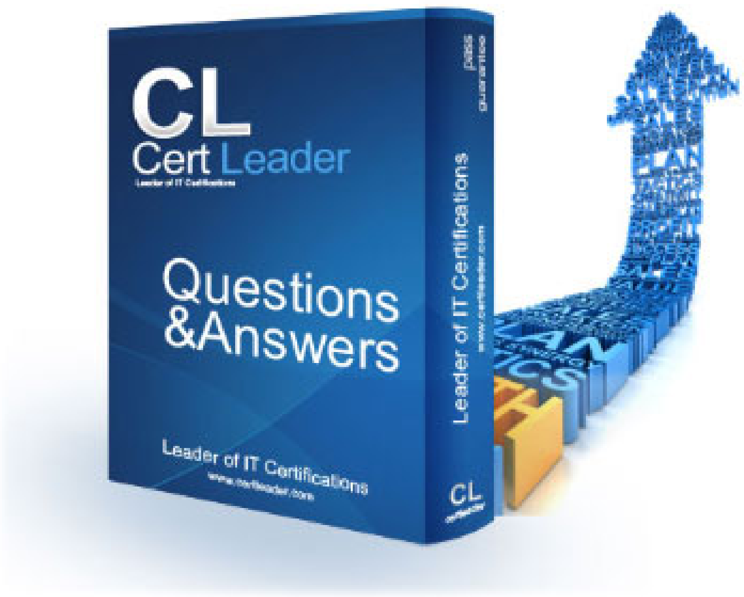
DP-203 Dumps

Data Engineering on Microsoft Azure

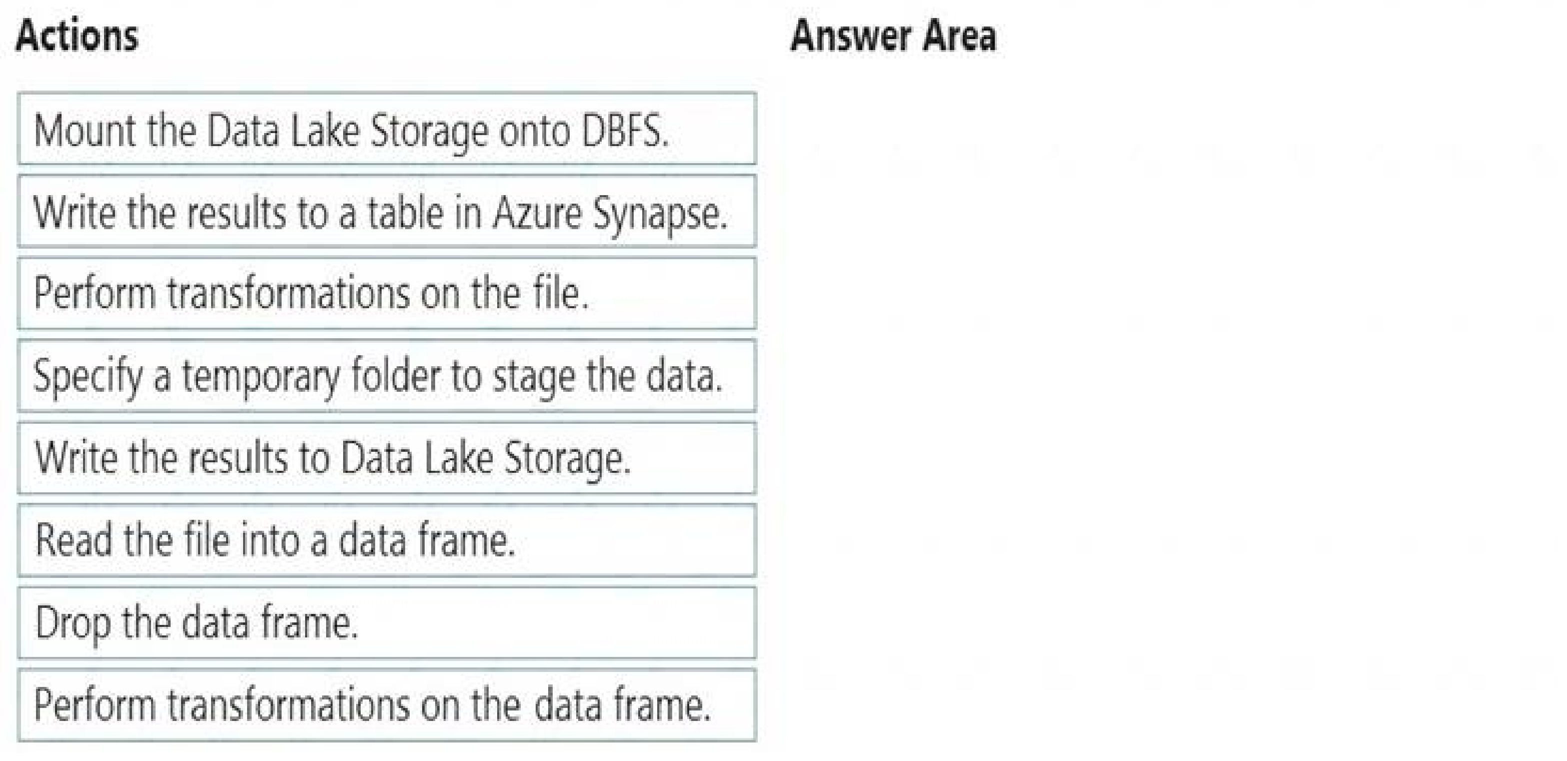
https://www.certleader.com/DP-203-dumps.html



NEW QUESTION 1
- (Exam Topic 3)
You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers. The file contains two attributes named FirstName and LastName.
You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the FirstName and LastName values.
You create the following components:

- A destination table in Azure Synapse
- An Azure Blob storage container
- A service principal

Which five actions should you perform in sequence next in is Databricks notebook? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



Actions

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

Answer Area

A. Mastered
B. Not Mastered

Answer: A

Explanation:
1) mount onto DBFS
2) read into data frame
3) transform data frame
4) specify temporary folder
5) write the results to table in in Azure Synapse https://docs.databricks.com/data/data-sources/azure/azure-datalake-gen2.html
https://docs.microsoft.com/en-us/azure/databricks/scenarios/databricks-extract-load-sql-data-warehouse

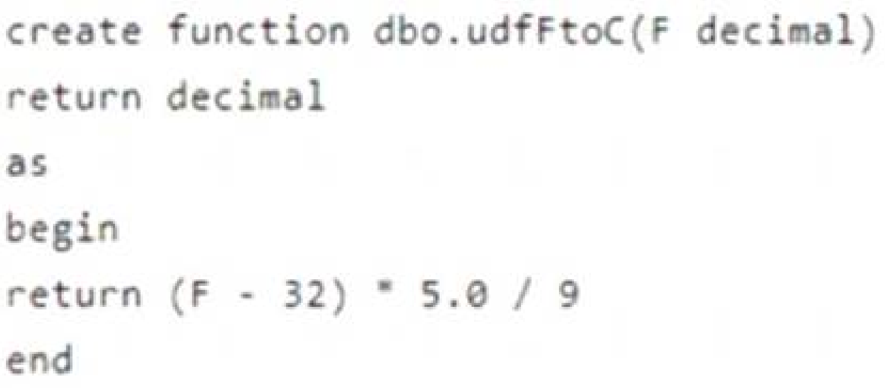
NEW QUESTION 2
- (Exam Topic 3)
A company plans to use Apache Spark analytics to analyze intrusion detection data.
You need to recommend a solution to analyze network and system activity data for malicious activities and policy violations. The solution must minimize administrative efforts.
What should you recommend?

A. Azure Data Lake Storage
B. Azure Databricks
C. Azure HDInsight
D. Azure Data Factory

Answer: B

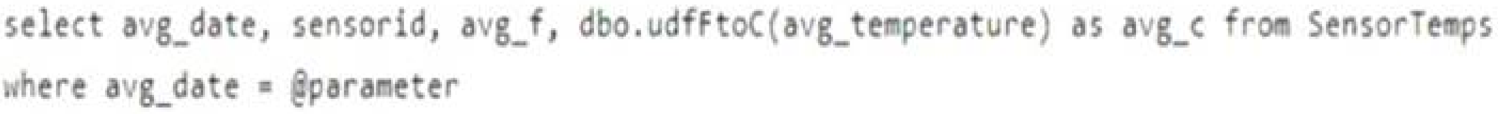
Explanation:
Three common analytics use cases with Microsoft Azure Databricks
Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.
Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.
Note: Recommendation engines, churn analysis, and intrusion detection are common scenarios that many
organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.
Reference:
https://azure.microsoft.com/es-es/blog/three-critical-analytics-use-cases-with-microsoft-azure-databricks/

NEW QUESTION 3
- (Exam Topic 3)
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.

Answer: BD

Explanation:
https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-cac

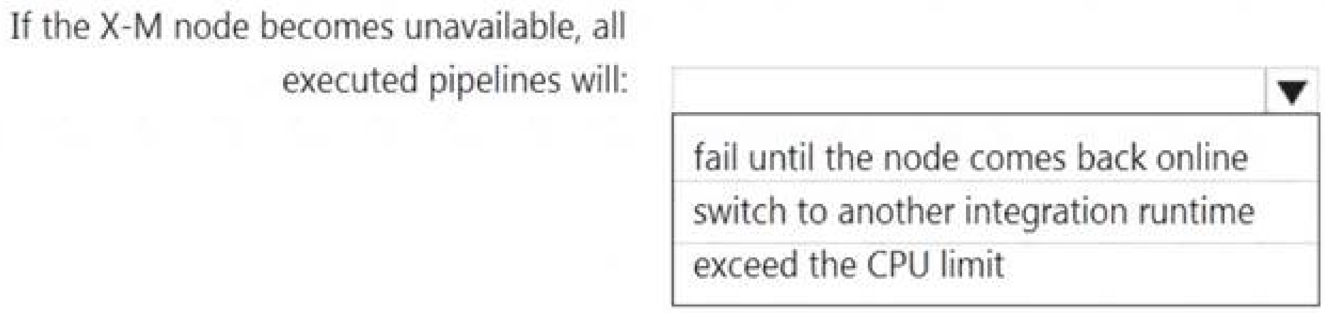
NEW QUESTION 4
- (Exam Topic 3)
You have a self-hosted integration runtime in Azure Data Factory.
The current status of the integration runtime has the following configurations:

- Status: Running
- Type: Self-Hosted
- Version: 4.4.7292.1
- Running / Registered Node(s): 1/1
- High Availability Enabled: False
- Linked Count: 0
- Queue Length: 0
- Average Queue Duration. 0.00s

The integration runtime has the following node details:

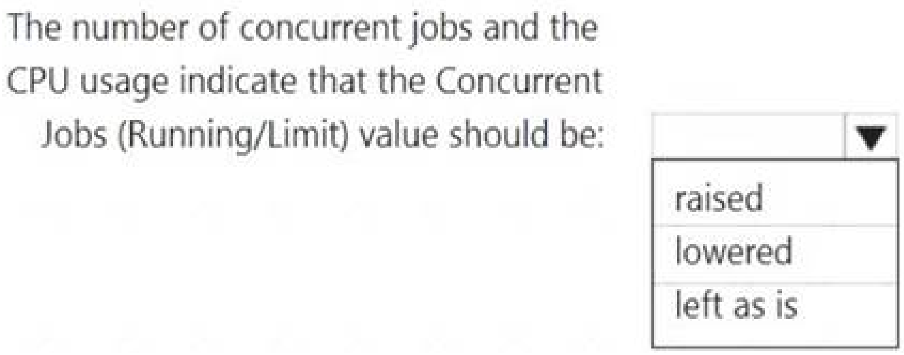
- Name: X-M
- Status: Running
- Version: 4.4.7292.1
- Available Memory: 7697MB
- CPU Utilization: 6%
- Network (In/Out): 1.21KBps/0.83KBps
- Concurrent Jobs (Running/Limit): 2/14
- Role: Dispatcher/Worker
- Credential Status: In Sync

Use the drop-down menus to select the answer choice that completes each statement based on the information presented.
NOTE: Each correct selection is worth one point.



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



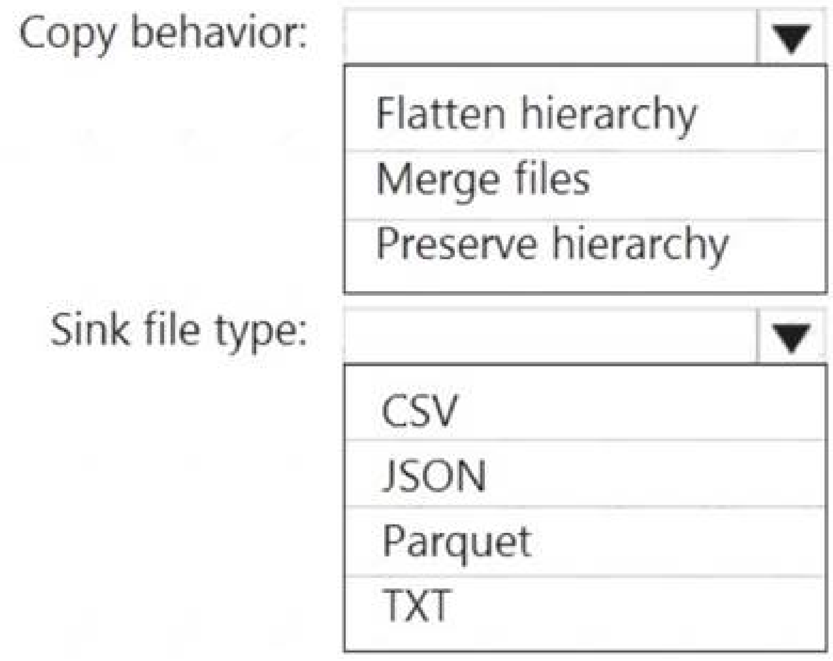
The number of concurrent jobs and the CPU usage indicate that the Concurrent Jobs (Running/Limit) value should be:

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: fail until the node comes back online
We see: High Availability Enabled: False
Note: Higher availability of the self-hosted integration runtime so that it's no longer the single point of failure in your big data solution or cloud data integration with Data Factory.
Box 2: lowered We see:
Concurrent Jobs (Running/Limit): 2/14 CPU Utilization: 6%
Note: When the processor and available RAM aren't well utilized, but the execution of concurrent jobs reaches a node's limits, scale up by increasing the number of concurrent jobs that a node can run
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime

NEW QUESTION 5
- (Exam Topic 3)
You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools. Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company.
You need to move the files to a different folder and transform the data to meet the following requirements: ➢ Provide the fastest possible query times.
➢ Automatically infer the schema from the underlying files.
How should you configure the Data Factory copy activity? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Copy behavior: ▼
Flatten hierarchy
Merge files
Preserve hierarchy

Sink file type: ▼
CSV
JSON
Parquet
TXT

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Preserver herarchy
Compared to the flat namespace on Blob storage, the hierarchical namespace greatly improves the performance of directory management operations, which improves overall job performance.
Box 2: Parquet
Azure Data Factory parquet format is supported for Azure Data Lake Storage Gen2. Parquet supports the schema property.
Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction https://docs.microsoft.com/en-us/azure/data-factory/format-parquet

NEW QUESTION 6
- (Exam Topic 3)
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?

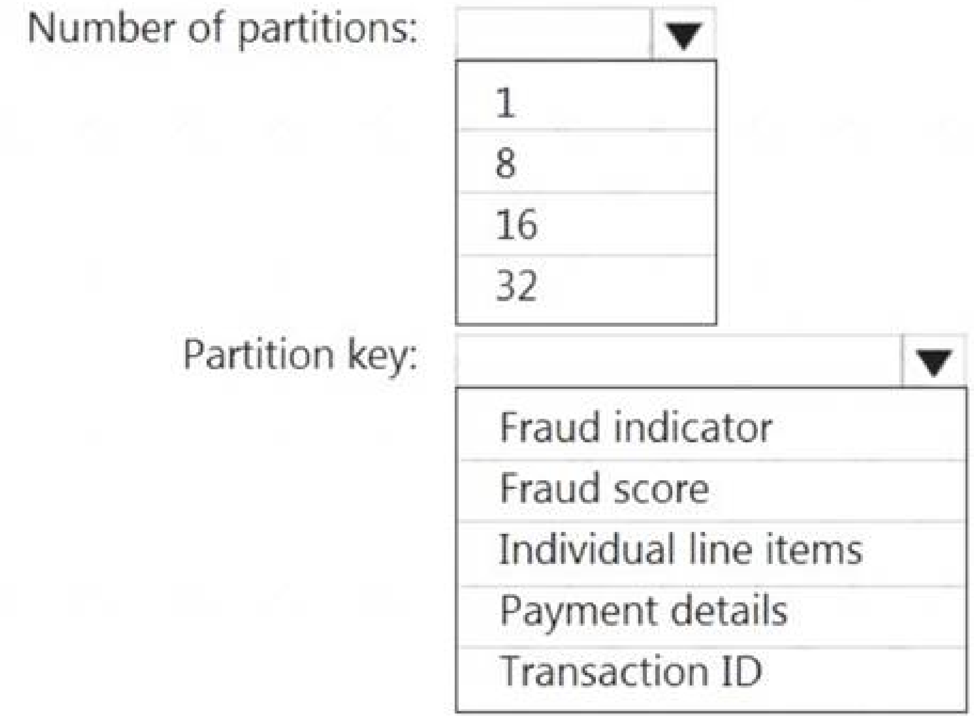
A. Yes
B. No

Answer: B

Explanation:

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

NEW QUESTION 7
- (Exam Topic 3)
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.



A. Mastered
B. Not Mastered

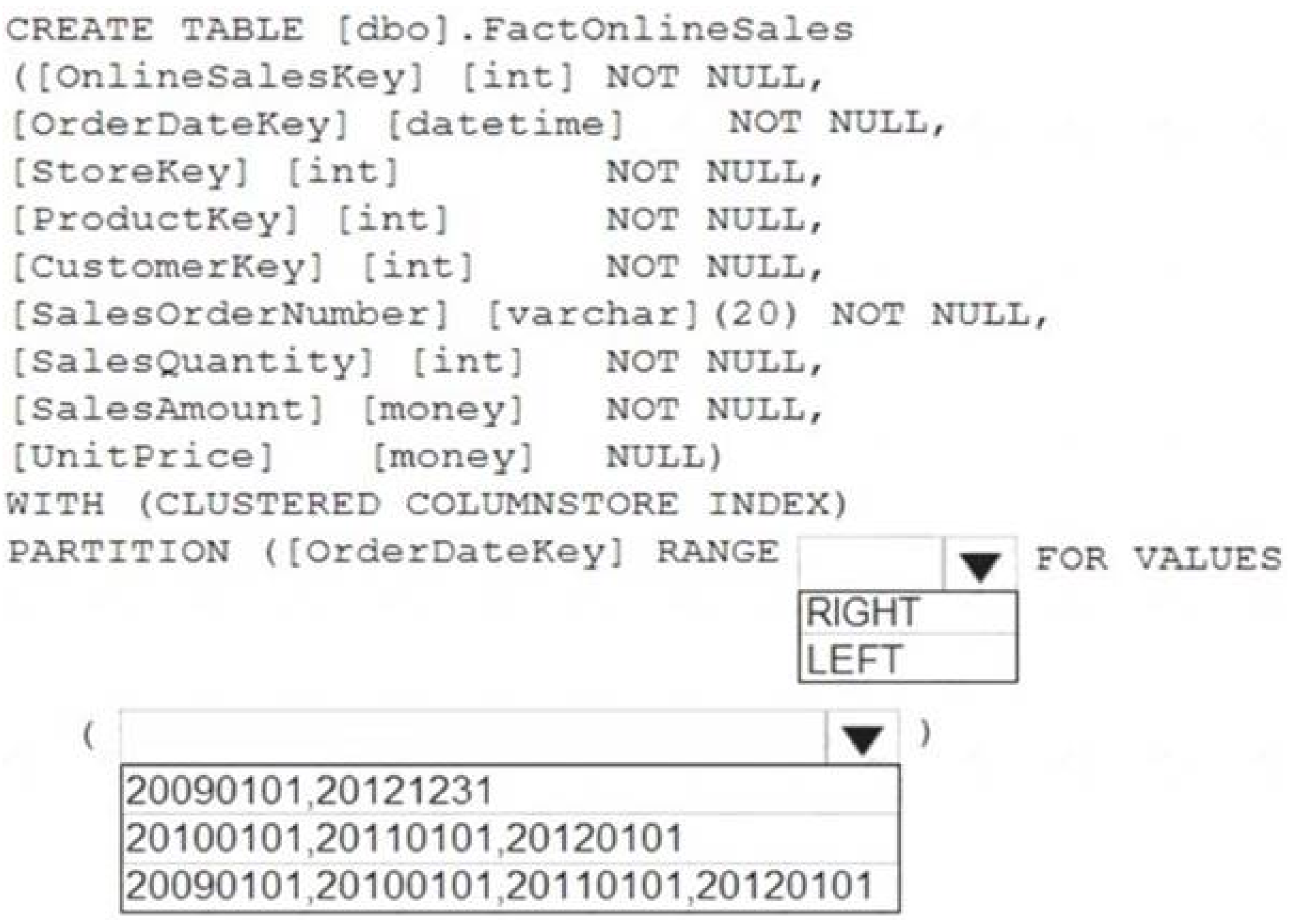
Answer: A

Explanation:
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

NEW QUESTION 8
- (Exam Topic 3)
You have an enterprise data warehouse in Azure Synapse Analytics that contains a table named FactOnlineSales. The table contains data from the start of 2009 to the end of 2012.
You need to improve the performance of queries against FactOnlineSales by using table partitions. The solution must meet the following requirements:

- Create four partitions based on the order date.
- Ensure that each partition contains all the orders places during a given calendar year.

How should you complete the T-SQL command? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



```
CREATE TABLE [dbo].FactOnlineSales
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime]    NOT NULL,
[StoreKey] [int]          NOT NULL,
[ProductKey] [int]        NOT NULL,
[CustomerKey] [int]       NOT NULL,
[SalesOrderNumber] [varchar](20) NOT NULL,
[SalesQuantity] [int]     NOT NULL,
[SalesAmount] [money]     NOT NULL,
[UnitPrice]    [money]    NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey] RANGE [ ▼ ] FOR VALUES
                                RIGHT
                                LEFT
   ( [ ▼ ] )
     20090101,20121231
     20100101,20110101,20120101
     20090101,20100101,20110101,20120101
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Text Description automatically generated
Range Left or Right, both are creating similar partition but there is difference in comparison For example: in this scenario, when you use LEFT and 20100101,20110101,20120101
Partition will be, datecol<=20100101, datecol>20100101 and datecol<=20110101, datecol>20110101 and datecol<=20120101, datecol>20120101
But if you use range RIGHT and 20100101,20110101,20120101
Partition will be, datecol<20100101, datecol>=20100101 and datecol<20110101, datecol>=20110101 and datecol<20120101, datecol>=20120101
In this example, Range RIGHT will be suitable for calendar comparison Jan 1st to Dec 31st Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql?view=sql-server-ver1

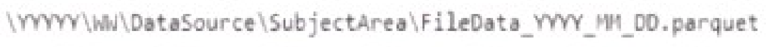
NEW QUESTION 9
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.
You need to identify the extent of the data skew in Table1. What should you do in Synapse Studio?

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_scacus.
D. Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_scacs.

Answer: A

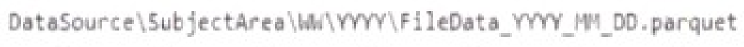
Explanation:
A quick way to check for data skew is to use DBCC PDW_SHOWSPACEUSED. The following SQL code returns the number of table rows that are stored in each of the 60 distributions. For balanced performance, the rows in your distributed table should be spread evenly across all the distributions.
DBCC PDW_SHOWSPACEUSED('dbo.FactInternetSales'); Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

NEW QUESTION 10
- (Exam Topic 3)
You are designing the folder structure for an Azure Data Lake Storage Gen2 account. You identify the following usage patterns:
• Users will query data by using Azure Synapse Analytics serverless SQL pools and Azure Synapse Analytics serverless Apache Spark pods.
• Most queries will include a filter on the current year or week.
• Data will be secured by data source.
You need to recommend a folder structure that meets the following requirements:
• Supports the usage patterns
• Simplifies folder security
• Minimizes query times
Which folder structure should you recommend?
A)



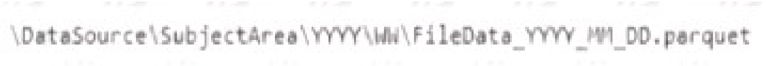
```
\YYYY\WW\DataSource\SubjectArea\FileData_YYYY_MM_DD.parquet
```

B)



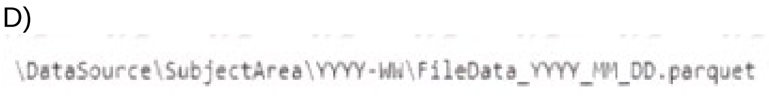
```
DataSource\SubjectArea\WW\YYYY\FileData_YYYY_MM_DD.parquet
```

C)



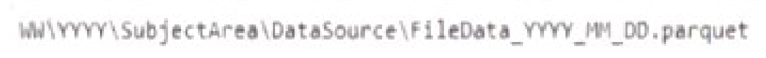
```
\DataSource\SubjectArea\YYYY\WW\FileData_YYYY_MM_DD.parquet
```

D)



```
\DataSource\SubjectArea\YYYY-WW\FileData_YYYY_MM_DD.parquet
```

E)



```
WW\YYYY\SubjectArea\DataSource\FileData_YYYY_MM_DD.parquet
```

A. Option A
B. Option B
C. Option C
D. Option D
E. Option E

Answer: C

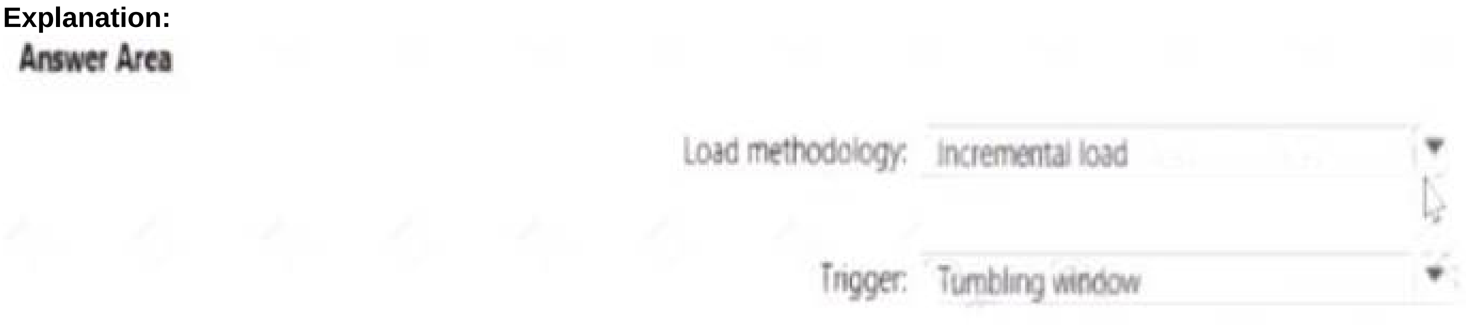
Explanation:
Data will be secured by data source. -> Use DataSource as top folder.
Most queries will include a filter on the current year or week -> Use \YYYY\WW\ as subfolders. Common Use Cases
A common use case is to filter data stored in a date (and possibly time) folder structure such as
/YYYY/MM/DD/ or /YYYY/MM/YYYY-MM-DD/. As new data is generated/sent/copied/moved to the storage account, a new folder is created for each specific time period. This strategy organises data into a maintainable folder structure.
Reference: https://www.serverlesssql.com/optimisation/azurestoragefilteringusingfilepath/

NEW QUESTION 10
- (Exam Topic 3)
You have an Azure Storage account that generates 200.000 new files daily. The file names have a format of (YYY)/(MM)/(DD)/|HH])/(CustornerID).csv.
You need to design an Azure Data Factory solution that will toad 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.

A. Mastered
B. Not Mastered

Answer: A

Explanation:



Answer Area

Load methodology: Incremental load

Trigger: Tumbling window

NEW QUESTION 11
- (Exam Topic 3)
You plan to create a dimension table in Azure Synapse Analytics that will be less than 1 GB. You need to create the table to meet the following requirements:
• Provide the fastest Query time.
• Minimize data movement during queries. Which type of table should you use?

A. hash distributed
B. heap
C. replicated
D. round-robin

Answer: C

Explanation:
A replicated table has a full copy of the table accessible on each Compute node. Replicating a table removes the need to transfer data among Compute nodes before a join or aggregation. Since the table has multiple copies, replicated tables work best when the table size is less than 2 GB compressed. 2 GB is not a hard limit.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-guidance-for-replicated-tab

NEW QUESTION 15
- (Exam Topic 3)
You have an Azure data factory.
You need to examine the pipeline failures from the last 180 flays. What should you use?

A. the Activity tog blade for the Data Factory resource
B. Azure Data Factory activity runs in Azure Monitor
C. Pipeline runs in the Azure Data Factory user experience
D. the Resource health blade for the Data Factory resource

Answer: B

Explanation:
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

NEW QUESTION 20
- (Exam Topic 3)
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 SOL.
- 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

Answer: B

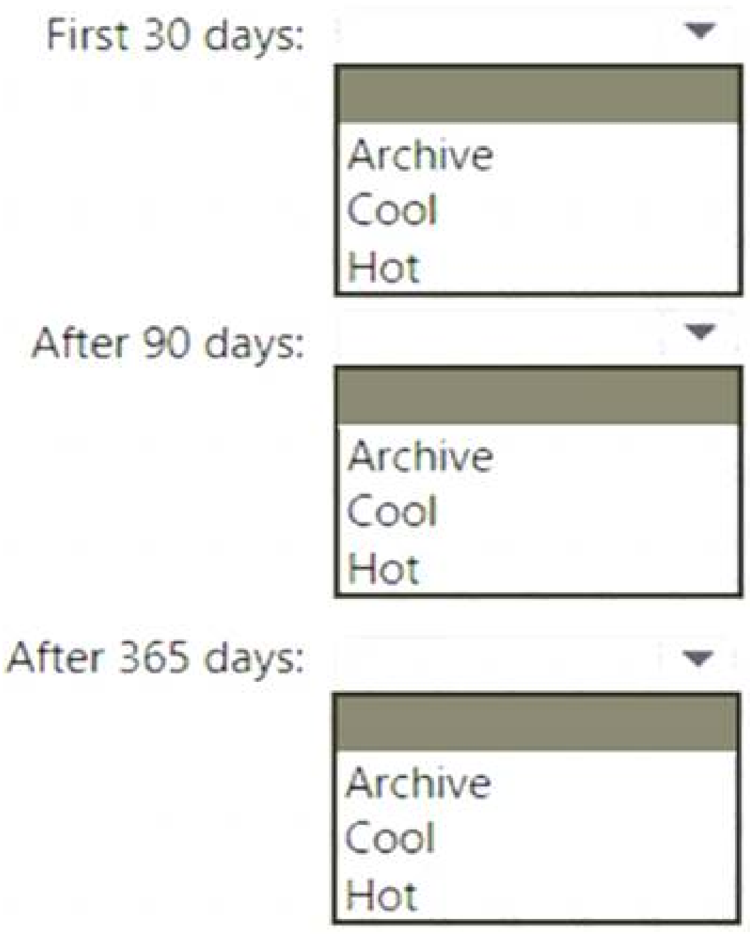
Explanation:
We need a High Concurrency cluster for the data engineers and the jobs.
Note: Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.
A high concurrency cluster is a managed cloud resource. The key benefits of 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

NEW QUESTION 25
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool mat 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. row-level security
B. column-level security
C. Dynamic data masking
D. Transparent Data Encryption (TDD

Answer: B

NEW QUESTION 29
- (Exam Topic 3)
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 SU 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.



First 30 days:
Archive
Cool
Hot

After 90 days:
Archive
Cool
Hot

After 365 days:
Archive
Cool
Hot

A. Mastered
B. Not Mastered

Answer: A

Explanation:
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. 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

NEW QUESTION 33
- (Exam Topic 3)
You are implementing a batch dataset in the Parquet format.
Data tiles will be produced by using Azure Data Factory and stored in Azure Data Lake Storage Gen2. The files will be consumed by an Azure Synapse Analytics serverless SQL pool.
You need to minimize storage costs for the solution. What should you do?

A. Store all the data as strings in the Parquet tiles.
B. Use OPENROWEST to query the Parquet files.
C. Create an external table mat contains a subset of columns from the Parquet files.
D. Use Snappy compression for the files.

Answer: C

Explanation:
An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

NEW QUESTION 34
- (Exam Topic 3)
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.

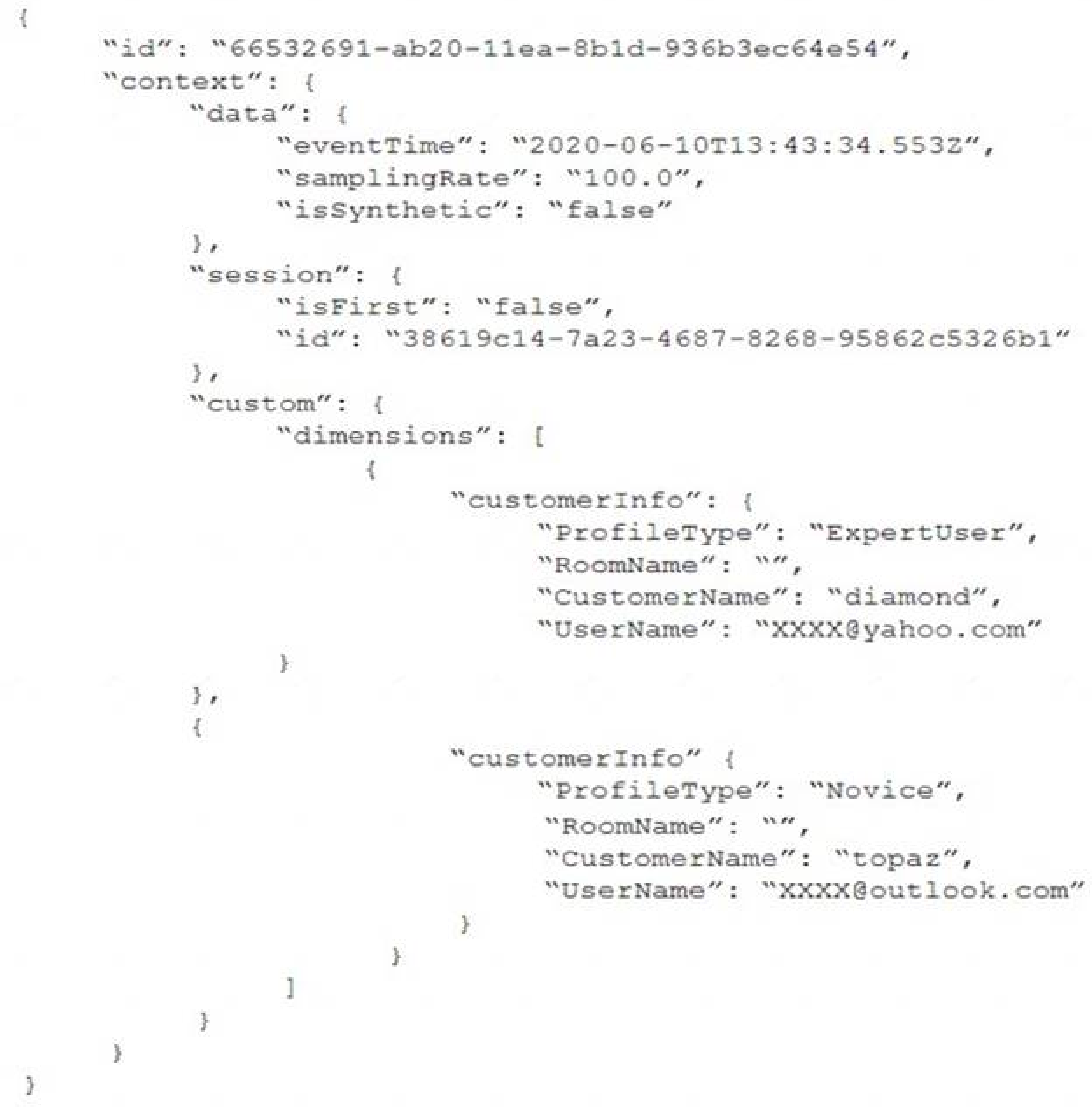
What should you include in the recommendation?

A. Transparent Data Encryption (TDE)
B. row-level security (RLS)
C. column-level encryption
D. Azure Active Directory (Azure AD) pass-through authentication

Answer: C

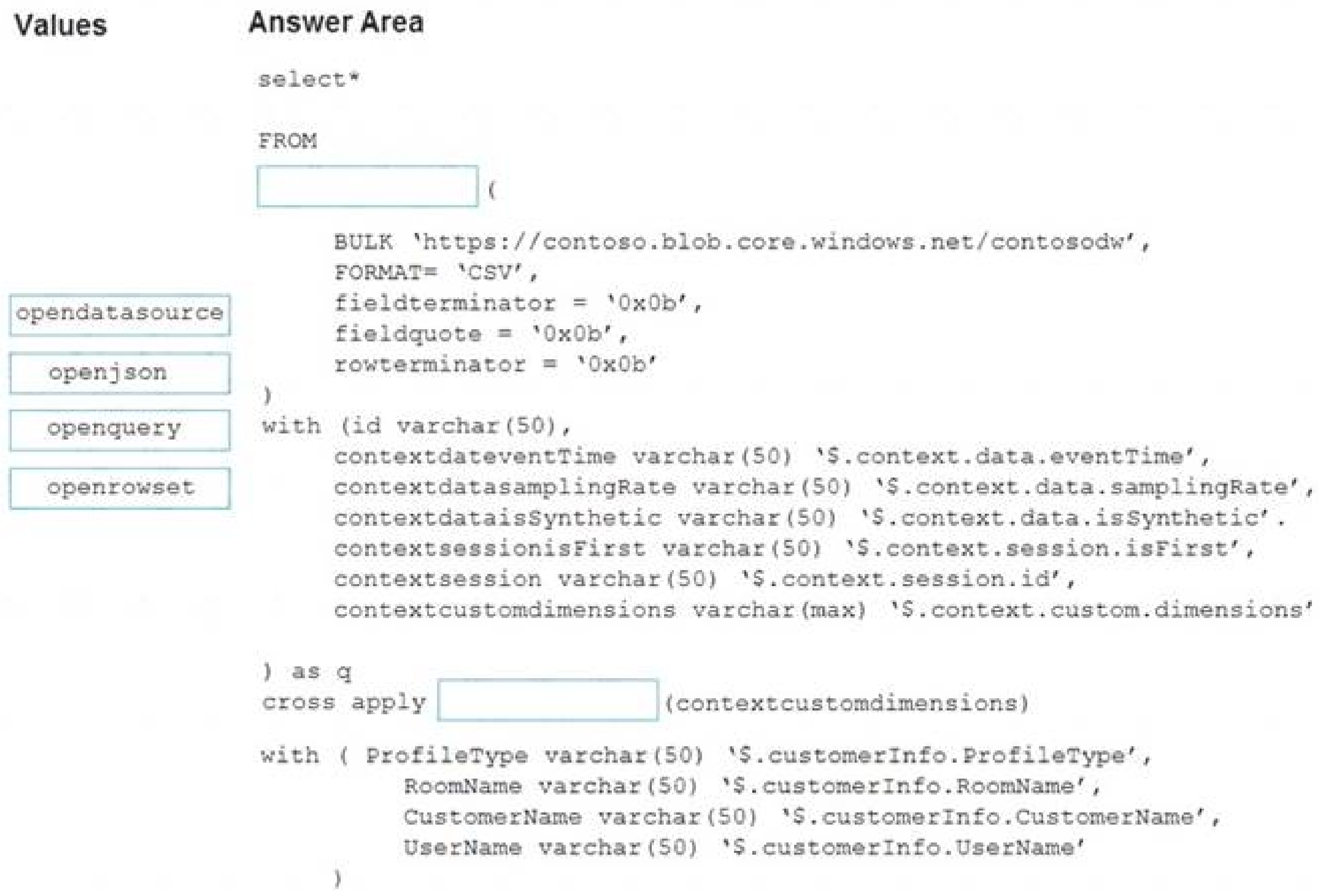
Explanation:
Use Always Encrypted to secure the required columns. You can configure Always Encrypted for individual database columns containing your sensitive data. Always Encrypted is a feature designed to protect sensitive data, such as credit card numbers or national identification numbers (for example, 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

NEW QUESTION 35
- (Exam Topic 3)
You have an Azure Synapse Analytics workspace named WS1.
You have an Azure Data Lake Storage Gen2 container that contains JSON-formatted files in the following format.



```
{
      “id”: “66532691-ab20-11ea-8b1d-936b3ec64e54”,
      “context”: {
            “data”: {
                  “eventTime”: “2020-06-10T13:43:34.553Z”,
                  “samplingRate”: “100.0”,
                  “isSynthetic”: “false”
            },
            “session”: {
                  “isFirst”: “false”,
                  “id”: “38619c14-7a23-4687-8268-95862c5326b1”
            },
            “custom”: {
                  “dimensions”: [
                        {
                              “customerInfo”: {
                                    “ProfileType”: “ExpertUser”,
                                    “RoomName”: “”,
                                    “CustomerName”: “diamond”,
                                    “UserName”: “XXXX@yahoo.com”
                  }
            },
            {
                              “customerInfo” {
                                    “ProfileType”: “Novice”,
                                    “RoomName”: “”,
                                    “CustomerName”: “topaz”,
                                    “UserName”: “XXXX@outlook.com”
                              }
                          }
                   ]
             }
       }
 }
```

You need to use the serverless SQL pool in WS1 to read the files.
How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.



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'
     )
```

A. Mastered
B. Not Mastered

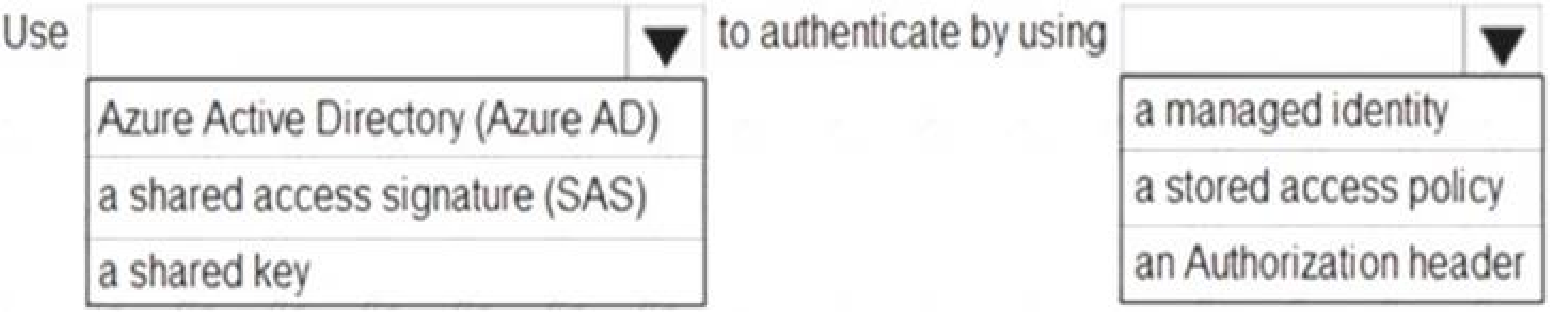
Answer: A

Explanation:
Graphical user interface, text, application, email Description automatically generated
Box 1: openrowset
The easiest way to see to the content of your CSV file is to provide file URL to OPENROWSET function, specify csv FORMAT.
Example: SELECT *
FROM OPENROWSET(
BULK 'csv/population/population.csv', DATA_SOURCE = 'SqlOnDemandDemo', FORMAT = 'CSV', PARSER_VERSION = '2.0', FIELDTERMINATOR =',', ROWTERMINATOR = '\n'
Box 2: openjson
You can access your JSON files from the Azure File Storage share by using the mapped drive, as shown in the following example:
SELECT book.* FROM
OPENROWSET(BULK N't:\books\books.json', SINGLE_CLOB) AS json CROSS APPLY OPENJSON(BulkColumn)
WITH(id nvarchar(100), name nvarchar(100), price float, pages_i int, author nvarchar(100)) AS book
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-single-csv-file https://docs.microsoft.com/en-us/sql/relational-databases/json/import-json-documents-into-sql-server

NEW QUESTION 36
- (Exam Topic 3)
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.



Use [▼] to authenticate by using [▼]

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

a managed identity
a stored access policy
an Authorization header

A. Mastered
B. Not Mastered

Answer: A

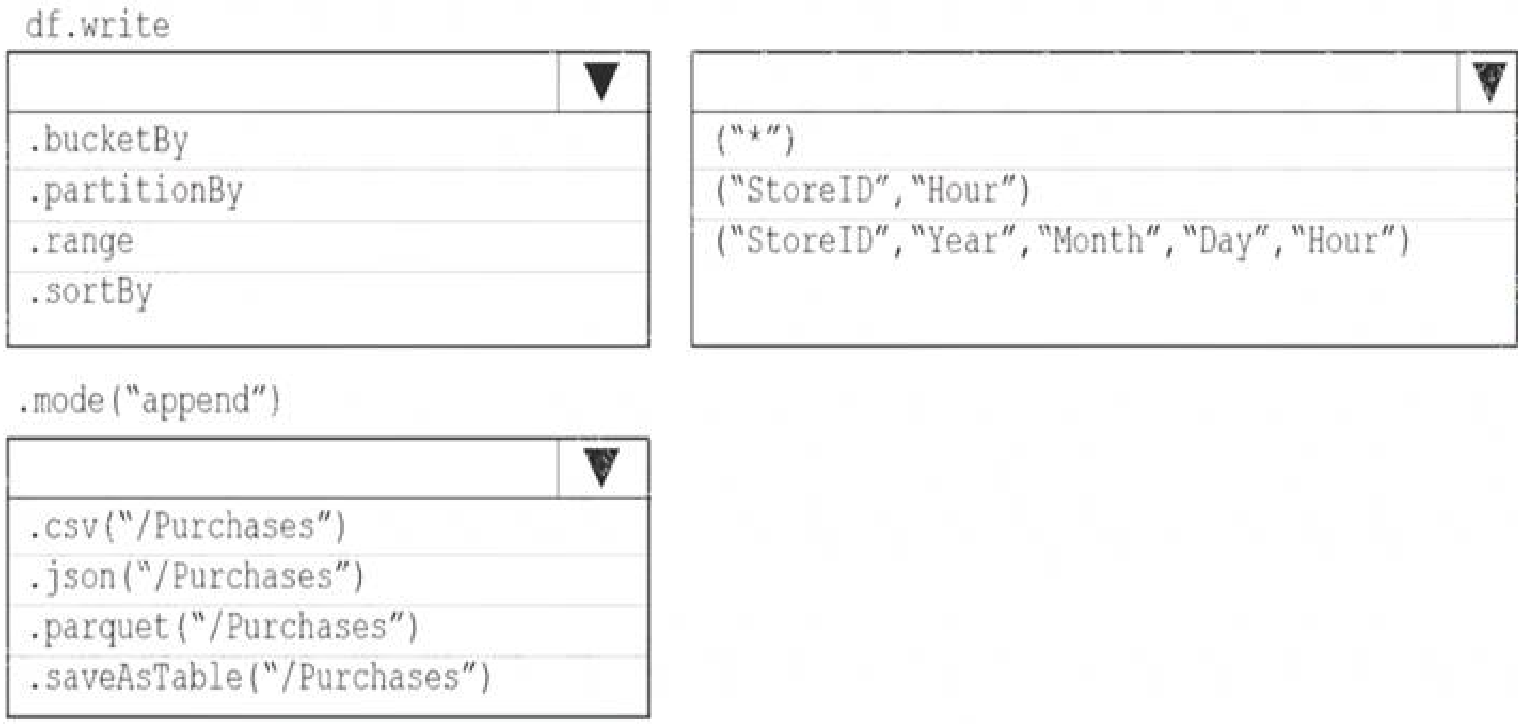
Explanation:
Text Description automatically generated
Box 1: Azure Active Directory (Azure AD)

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

NEW QUESTION 41
- (Exam Topic 3)
You plan to develop a dataset named Purchases by using Azure databricks Purchases will contain the following columns:
• ProductID
• ItemPrice
• lineTotal
• Quantity
• StoreID
• Minute
• Month
• Hour
• Year
• Day
You need to store the data to support hourly incremental load pipelines that will vary for each StoreID. the solution must minimize storage costs. How should you complete the rode? To answer, select the appropriate options In the answer area.
NOTE: Each correct selection is worth one point.



```
df.write
```

```
.bucketBy
.partitionBy
.range
.sortBy
```

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

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

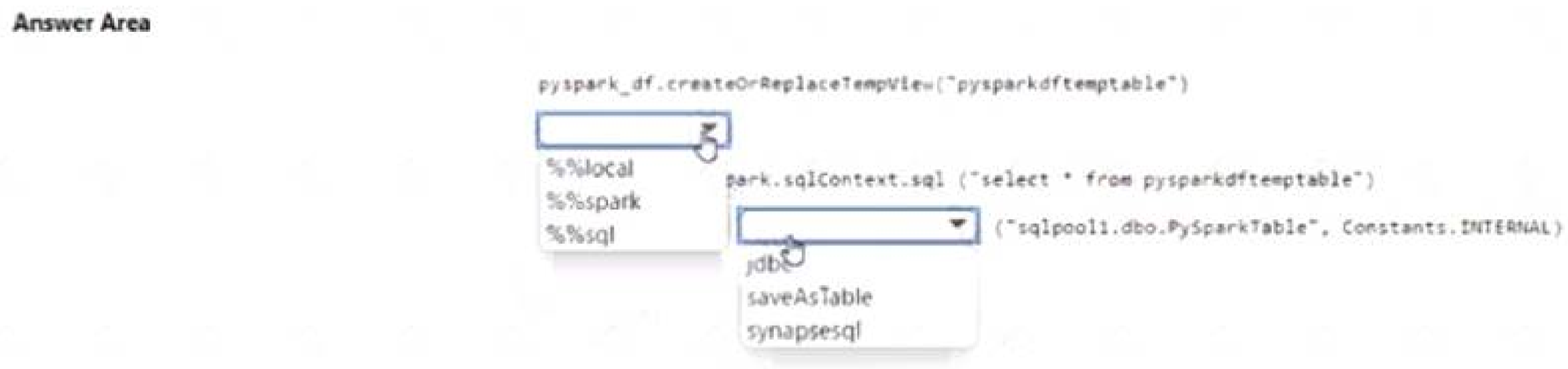
```
.csv("/Purchases")
.json("/Purchases")
.parquet("/Purchases")
.saveAsTable("/Purchases")
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: partitionBy
We should overwrite at the partition level. Example: df.write.partitionBy("y","m","d") mode(SaveMode.Append) parquet("/data/hive/warehouse/db_name.db/" + tableName) Box 2: ("StoreID", "Year", "Month", "Day", "Hour", "StoreID") Box 3: parquet("/Purchases")
Reference:
https://intellipaat.com/community/11744/how-to-partition-and-write-dataframe-in-spark-without-deleting-partiti

NEW QUESTION 44
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 contains a dedicated SQL pool named SQL Pool and an Apache Spark pool named sparkpool. Sparkpool1 contains a DataFrame named pyspark.df.
You need to write the contents of pyspark_df to a tabte in SQLPooM 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.



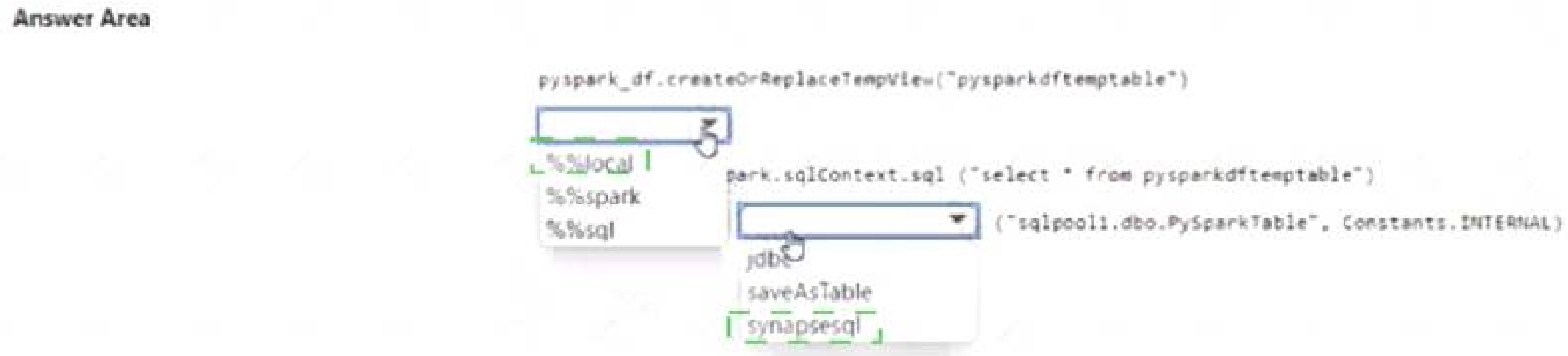
Answer Area

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")
park.sqlContext.sql ("select * from pysparkdftemptable")
("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)
```

A. Mastered
B. Not Mastered

Answer: A

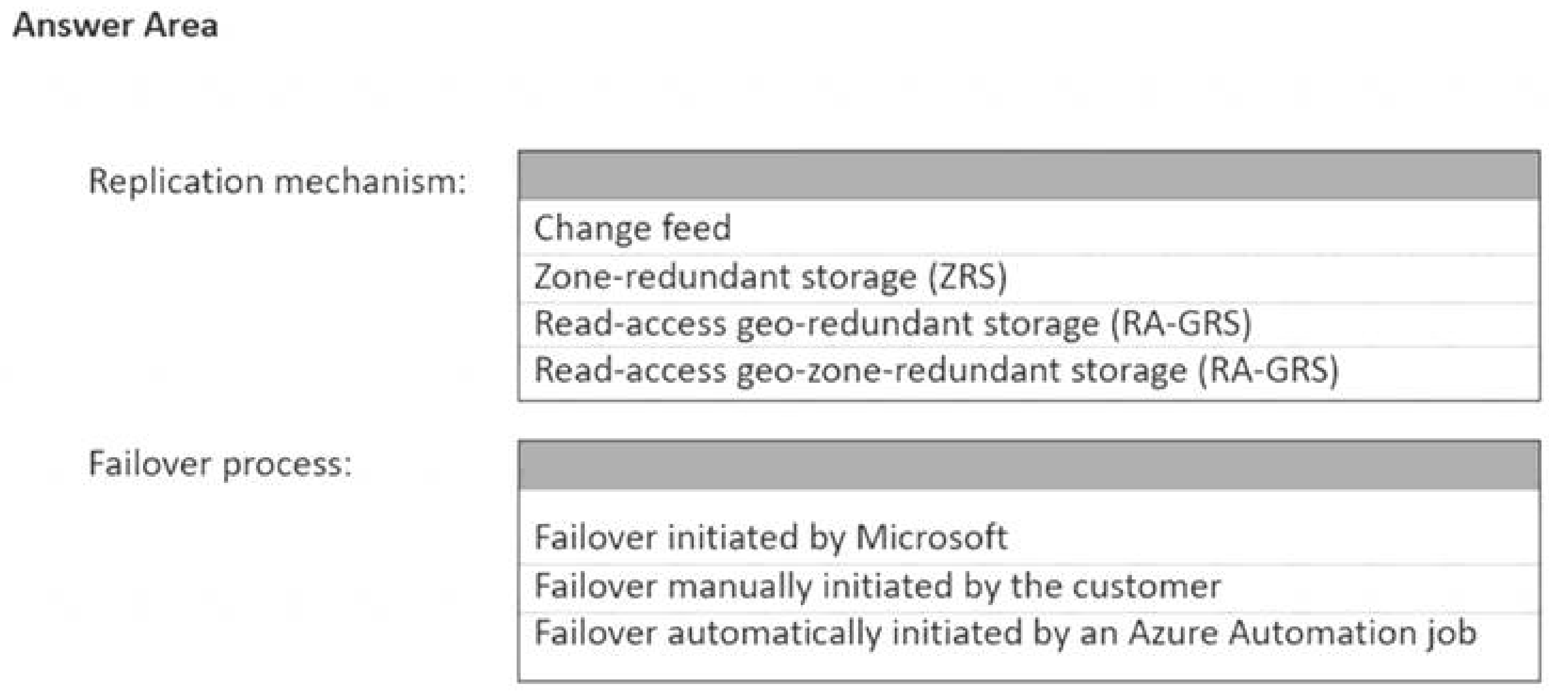
Explanation:



Answer Area

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")
park.sqlContext.sql ("select * from pysparkdftemptable")
("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)
```

NEW QUESTION 49
- (Exam Topic 3)
You plan to create an Azure Data Lake Storage Gen2 account
You need to recommend a storage solution that meets the following requirements:
• Provides the highest degree of data resiliency
• Ensures that content remains available for writes if a primary data center fails
What should you include in the recommendation? To answer, select the appropriate options in the answer area.



Answer Area

Replication mechanism:

Change feed
Zone-redundant storage (ZRS)
Read-access geo-redundant storage (RA-GRS)
Read-access geo-zone-redundant storage (RA-GRS)

Failover process:

Failover initiated by Microsoft
Failover manually initiated by the customer
Failover automatically initiated by an Azure Automation job

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated
Availability : "Microsoft recommends RA-GZRS for maximum availability and durability for your applications."
Failover: "The customer initiates the account failover to the secondary endpoint. " https://docs.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance?toc=/azure/storage/
https://docs.microsoft.com/en-us/answers/questions/32583/azure-data-lake-gen2-disaster-recoverystorage-acco.h

NEW QUESTION 53
- (Exam Topic 3)
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?

A. Yes
B. No

Answer: A

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

NEW QUESTION 56
- (Exam Topic 3)
You are developing an application that uses Azure Data Lake Storage Gen 2.
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. Azure Active Directory (Azure AD) identities
B. shared access signatures (SAS)
C. account keys
D. role assignments

Answer: B

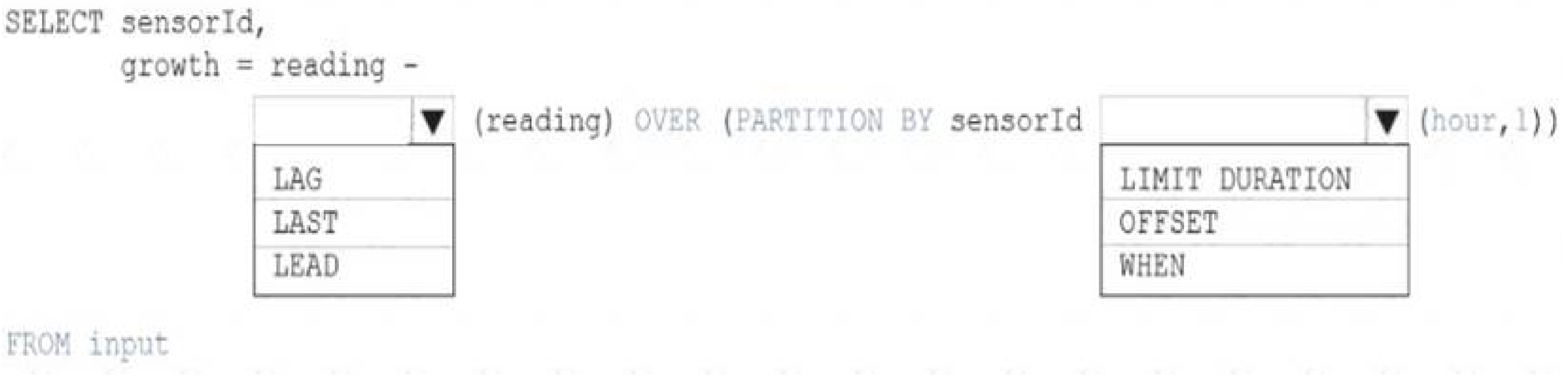
Explanation:
A shared access signature (SAS) provides secure delegated access to resources in your storage account. With a SAS, you have granular control over how a client can access your data. For example:
What resources the client may access.
What permissions they have to those resources. How long the SAS is valid.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/storage-sas-overview

NEW QUESTION 58
- (Exam Topic 3)
You have an Azure Factory instance named DF1 that contains a pipeline named PL1.PL1 includes a tumbling window trigger.
You create five clones of PL1. You configure each clone pipeline to use a different data source.
You need to ensure that the execution schedules of the clone pipeline match the execution schedule of PL1. What should you do?

A. Add a new trigger to each cloned pipeline
B. Associate each cloned pipeline to an existing trigger.
C. Create a tumbling window trigger dependency for the trigger of PL1.
D. Modify the Concurrency setting of each pipeline.

Answer: B

NEW QUESTION 59
- (Exam Topic 3)
You are building an Azure Analytics query that will receive input data from Azure IoT Hub and write the results to Azure Blob storage.
You need to calculate the difference in readings per sensor per hour.
How should you complete the query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



```
SELECT sensorId,
       growth = reading -
            [          ▼] (reading) OVER (PARTITION BY sensorId [              ▼] (hour,1))
FROM input
```

A. Mastered
B. Not Mastered

Answer: A

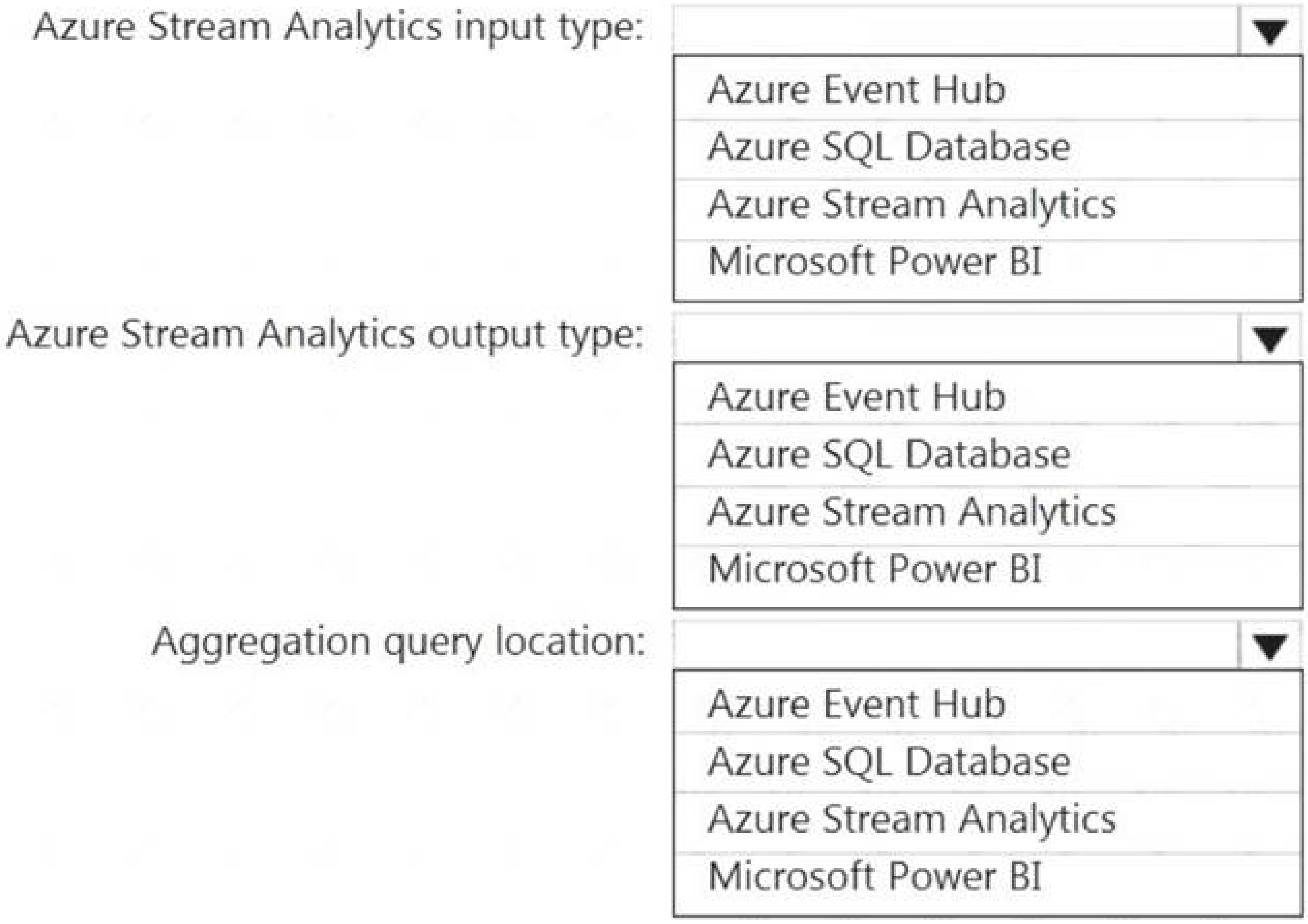
Explanation:
Box 1: LAG
The LAG analytic operator allows one to look up a “previous” event in an event stream, within certain constraints. It is very useful for computing the rate of growth of a variable, detecting when a variable crosses a threshold, or when a condition starts or stops being true.
Box 2: LIMIT DURATION
Example: Compute the rate of growth, per sensor: SELECT sensorId,
growth = reading

LAG(reading) OVER (PARTITION BY sensorId LIMIT DURATION(hour, 1)) FROM input
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/lag-azure-stream-analytics

NEW QUESTION 62
- (Exam Topic 3)
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

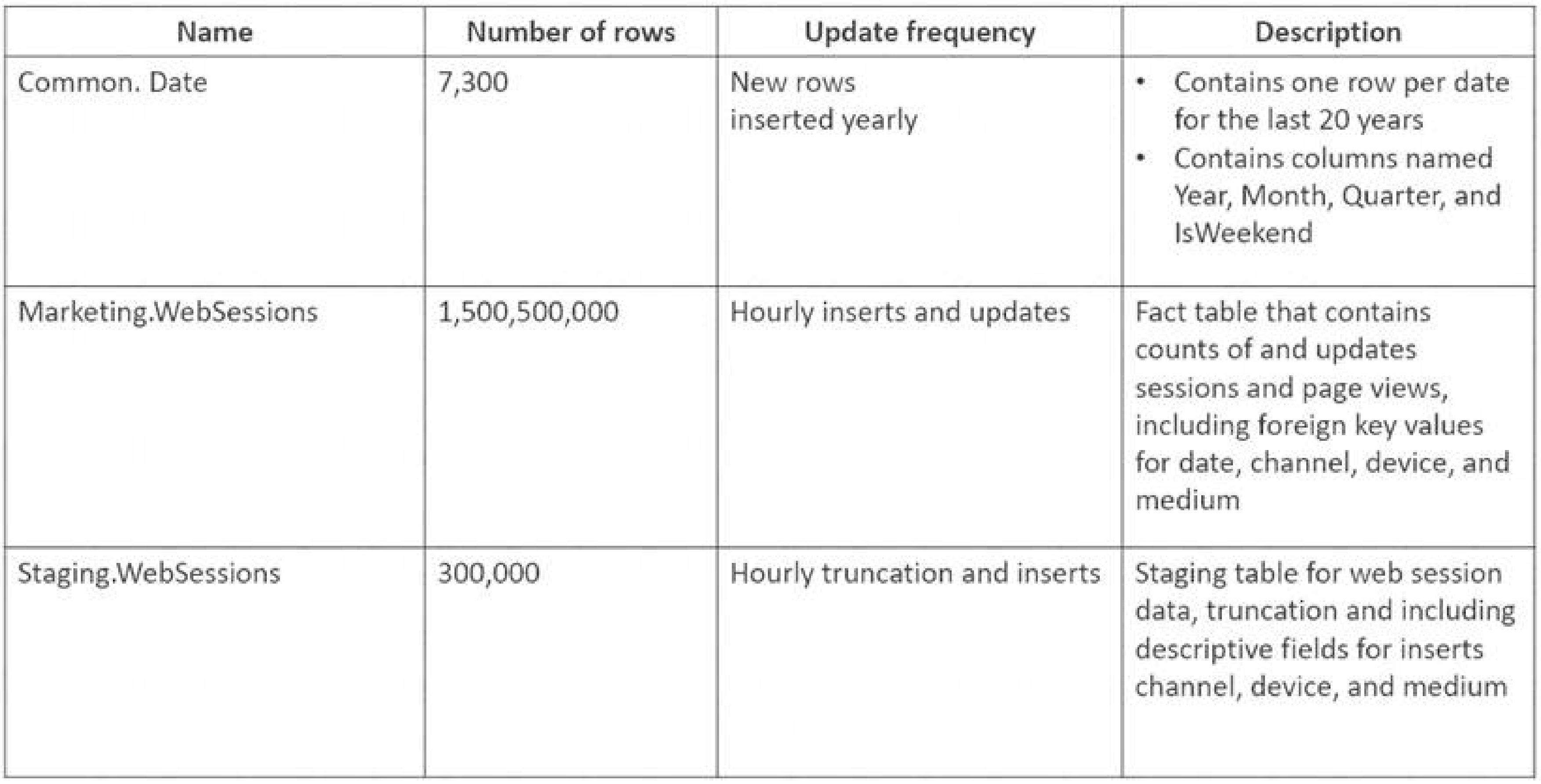
Aggregation query location:
- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 67
- (Exam Topic 3)
You have an Azure subscription.
You plan to build a data warehouse in an Azure Synapse Analytics dedicated SQL pool named pool1 that will contain staging tables and a dimensional model.
Pool1 will contain the following tables.



Name	Number of rows	Update frequency	Description
Common. Date	7,300	New rows inserted yearly	• Contains one row per date for the last 20 years • Contains columns named Year, Month, Quarter, and IsWeekend
Marketing.WebSessions	1,500,500,000	Hourly inserts and updates	Fact table that contains counts of and updates sessions and page views, including foreign key values for date, channel, device, and medium
Staging.WebSessions	300,000	Hourly truncation and inserts	Staging table for web session data, truncation and including descriptive fields for inserts channel, device, and medium

You need to design the table storage for pool1. The solution must meet the following requirements:

- Maximize the performance of data loading operations to Staging.WebSessions.
- Minimize query times for reporting queries against the dimensional model.

Which type of table distribution should you use for each table? To answer, drag the appropriate table distribution types to the correct tables. Each table distribution type may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.

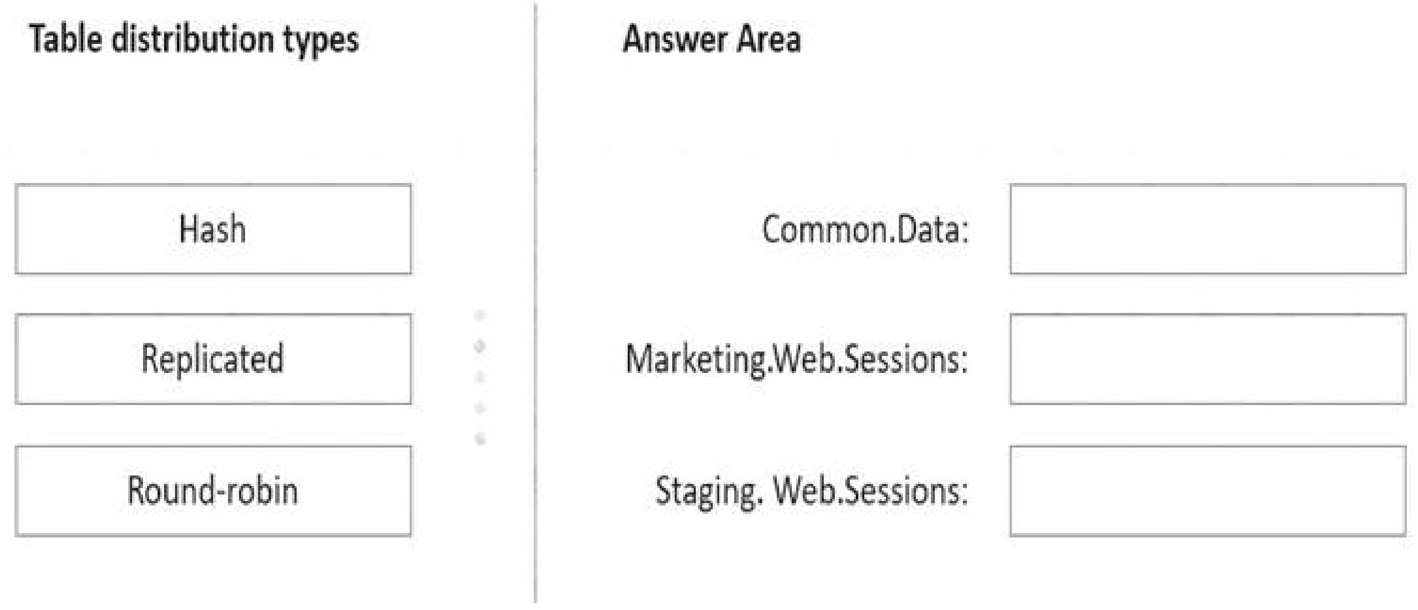


Table distribution types

Hash

Replicated

Round-robin

Answer Area

Common.Data:

Marketing.Web.Sessions:

Staging. Web.Sessions:

A. Mastered
B. Not Mastered

Answer: A

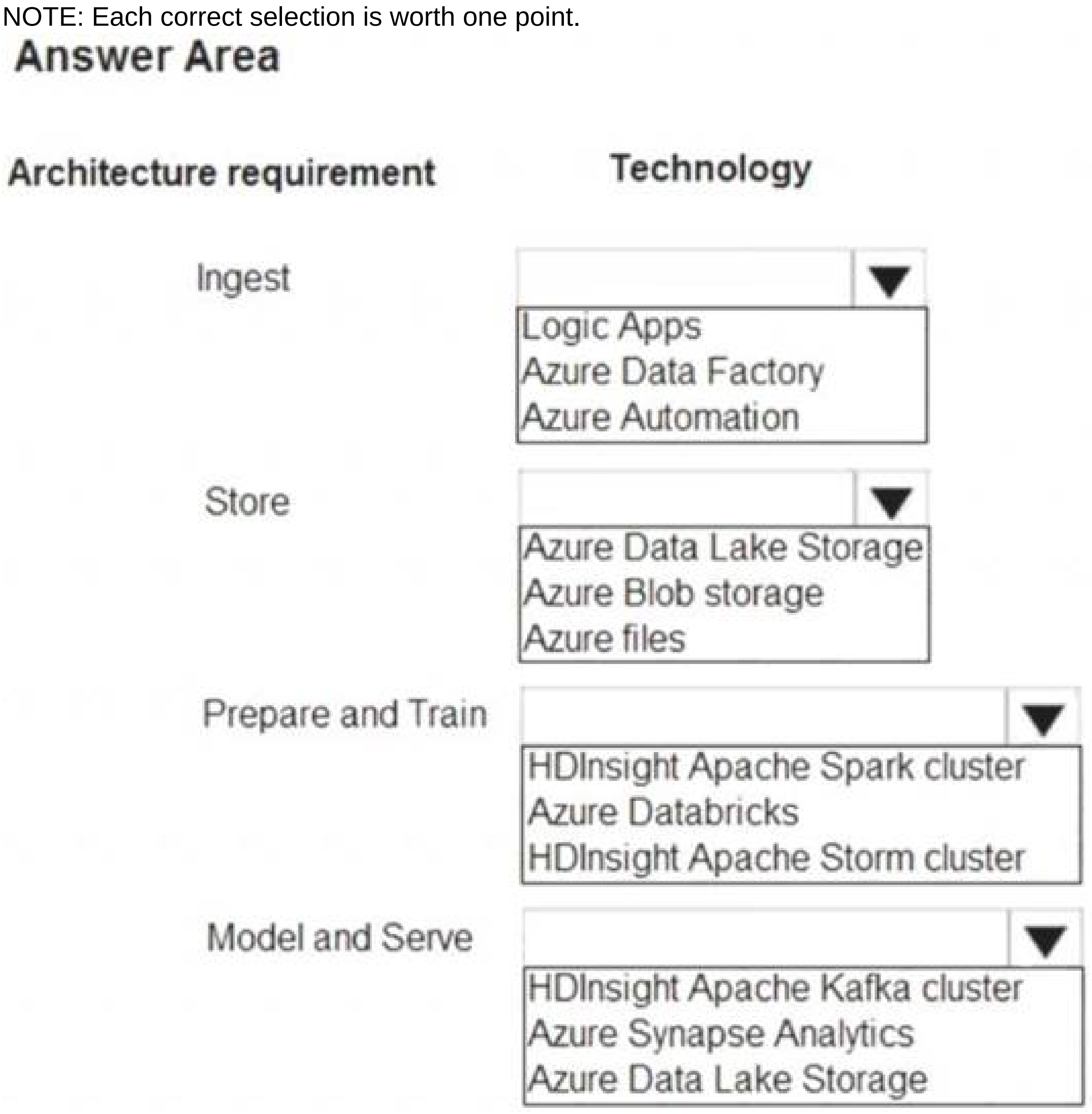
Explanation:
Box 1: Replicated
The best table storage option for a small table is to replicate it across all the Compute nodes. Box 2: Hash
Hash-distribution improves query performance on large fact tables. Box 3: Round-robin
Round-robin distribution is useful for improving loading speed. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

NEW QUESTION 71
- (Exam Topic 3)
A company plans to use Platform-as-a-Service (PaaS) to create the new data pipeline process. The process must meet the following requirements:
Ingest:

- Access multiple data sources.
- Provide the ability to orchestrate workflow.
- Provide the capability to run SQL Server Integration Services packages. Store:
- Optimize storage for big data workloads.
- Provide encryption of data at rest.
- Operate with no size limits. Prepare and Train:
- Provide a fully-managed and interactive workspace for exploration and visualization.
- Provide the ability to program in R, SQL, Python, Scala, and Java.
- Provide seamless user authentication with Azure Active Directory. Model & Serve:
- Implement native columnar storage.
- Support for the SQL language

- Provide support for structured streaming. You need to build the data integration pipeline.

Which technologies should you use? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer Area

Architecture requirement	Technology
Ingest	Logic Apps Azure Data Factory Azure Automation
Store	Azure Data Lake Storage Azure Blob storage Azure files
Prepare and Train	HDInsight Apache Spark cluster Azure Databricks HDInsight Apache Storm cluster
Model and Serve	HDInsight Apache Kafka cluster Azure Synapse Analytics Azure Data Lake Storage

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, application, table, email Description automatically generated

NEW QUESTION 75
- (Exam Topic 3)
You are designing a statistical analysis solution that will use custom proprietary1 Python functions on near real-time data from Azure Event Hubs.
You need to recommend which Azure service to use to perform the statistical analysis. The solution must minimize latency.
What should you recommend?

A. Azure Stream Analytics
B. Azure SQL Database
C. Azure Databricks
D. Azure Synapse Analytics

Answer: A

Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics

NEW QUESTION 78
- (Exam Topic 3)
You haw 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 m 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. Create an Azure Data Factory trigger
B. From the UX Authoring canvas, select Set up code repository
C. Create a GitHub action
D. From the UX Authoring canvas, run Publish All.
E. Create a Git repository
F. From the UX Authoring canvas, select Publish

Answer: DE

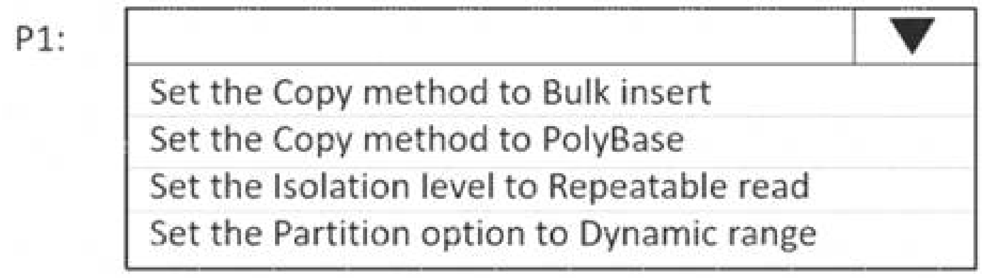
Explanation:

Reference:
https://docs.microsoft.com/en-us/azure/data-factory/source-control

NEW QUESTION 79
- (Exam Topic 3)
You have an Azure Data Factory instance named ADF1 and two Azure Synapse Analytics workspaces named WS1 and WS2.
ADF1 contains the following pipelines:

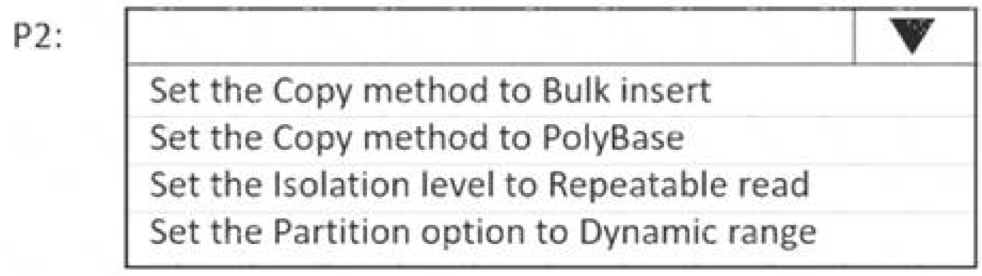
- P1: Uses a copy activity to copy data from a nonpartitioned table in a dedicated SQL pool of WS1 to an Azure Data Lake Storage Gen2 account
- P2: Uses a copy activity to copy data from text-delimited files in an Azure Data Lake Storage Gen2 account to a nonpartitioned table in a dedicated SQL pool 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.



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

A. Mastered
B. Not Mastered

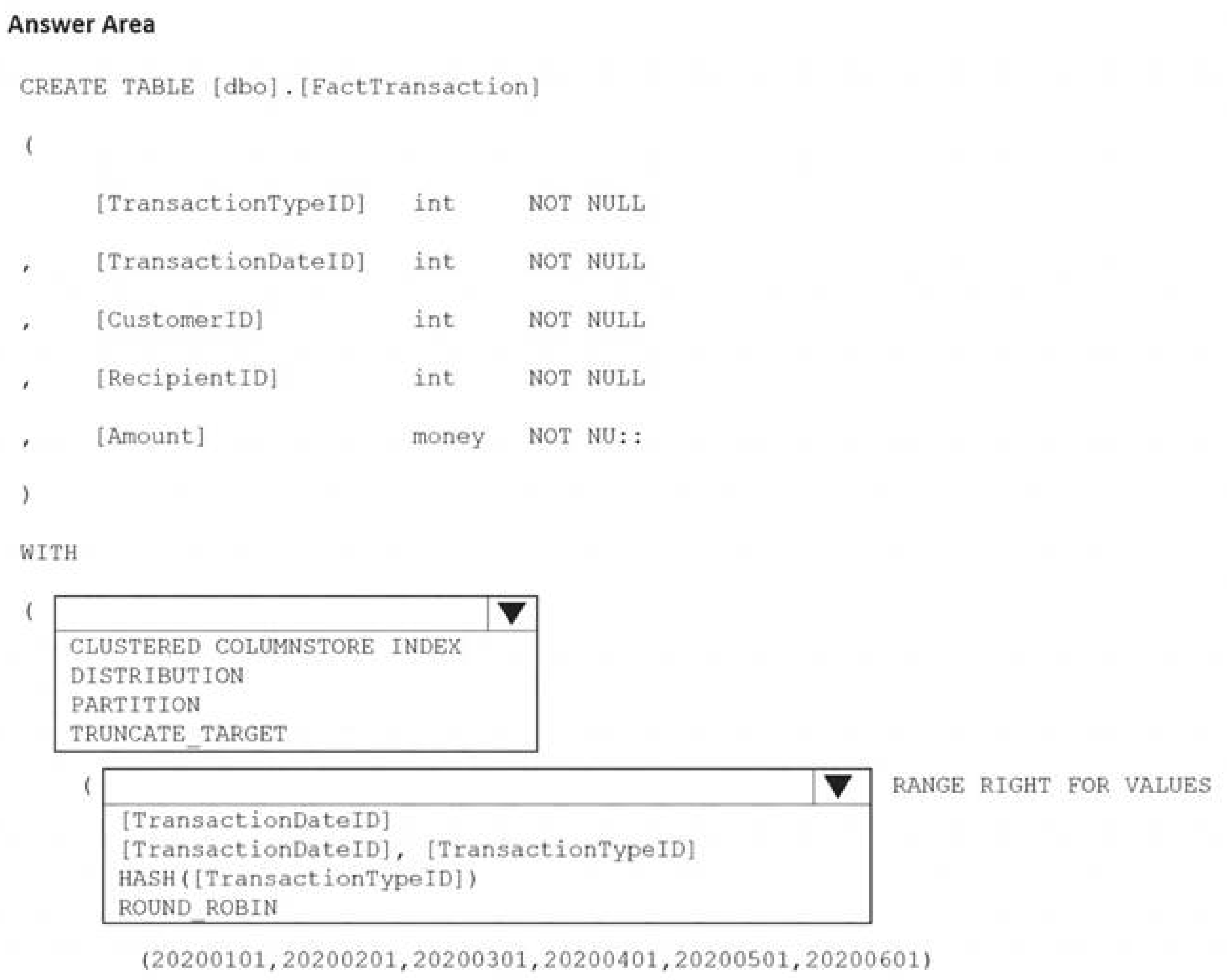
Answer: A

Explanation:
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

NEW QUESTION 83
- (Exam Topic 3)
You are building an Azure Synapse Analytics dedicated SQL pool that will contain a fact table for transactions from the first half of the year 2020.
You need to ensure that the table meets the following requirements:

- Minimizes the processing time to delete data that is older than 10 years
- Minimizes the I/O for queries that use year-to-date values

How should you complete the Transact-SQL statement? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer Area

```
CREATE TABLE [dbo].[FactTransaction]
(
      [TransactionTypeID]   int     NOT NULL
,     [TransactionDateID]   int     NOT NULL
,     [CustomerID]          int     NOT NULL
,     [RecipientID]         int     NOT NULL
,     [Amount]              money   NOT NU::
)
WITH
(  [▼ CLUSTERED COLUMNSTORE INDEX | DISTRIBUTION | PARTITION | TRUNCATE_TARGET]
     ( [▼ [TransactionDateID] | [TransactionDateID], [TransactionTypeID] | HASH([TransactionTypeID]) | ROUND_ROBIN]  RANGE RIGHT FOR VALUES
       (20200101,20200201,20200301,20200401,20200501,20200601)
```

CLUSTERED COLUMNSTORE INDEX
DISTRIBUTION
PARTITION
TRUNCATE_TARGET
[TransactionDateID]
[TransactionDateID], [TransactionTypeID]
HASH([TransactionTypeID])
ROUND_ROBIN

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Table Description automatically generated
Box 1: PARTITION
RANGE RIGHT FOR VALUES is used with PARTITION.
Part 2: [TransactionDateID] Partition on the date column.
Example: Creating a RANGE RIGHT partition function on a datetime column
The following partition function partitions a table or index into 12 partitions, one for each month of a year's worth of values in a datetime column.
CREATE PARTITION FUNCTION [myDateRangePF1] (datetime)
AS RANGE RIGHT FOR VALUES ('20030201', '20030301', '20030401',
'20030501', '20030601', '20030701', '20030801',
'20030901', '20031001', '20031101', '20031201');
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql

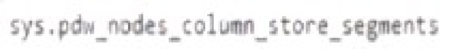
NEW QUESTION 84
- (Exam Topic 3)
You have a SQL pool in Azure Synapse that contains a table named dbo.Customers. The table contains a column name Email.
You need to prevent nonadministrative users from seeing the full email addresses in the Email column. The users must see values in a format of aXXX@XXXX.com 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.

Answer: D

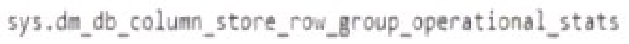
Explanation:
From Microsoft SQL Server Management Studio, set an email mask on the Email column. This is because "This feature cannot be set using portal for Azure Synapse (use PowerShell or REST API) or SQL Managed Instance." So use Create table statement with Masking e.g. CREATE TABLE Membership (MemberID int IDENTITY PRIMARY KEY, FirstName varchar(100) MASKED WITH (FUNCTION = 'partial(1,"XXXXXXX",0)') NULL, . .
https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview
upvoted 24 times

NEW QUESTION 85
- (Exam Topic 3)
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)



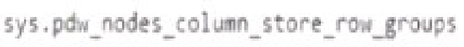
```
sys.pdw_nodes_column_store_segments
```

B)



```
sys.dm_db_column_store_row_group_operational_stats
```

C)



```
sys.pdw_nodes_column_store_row_groups
```

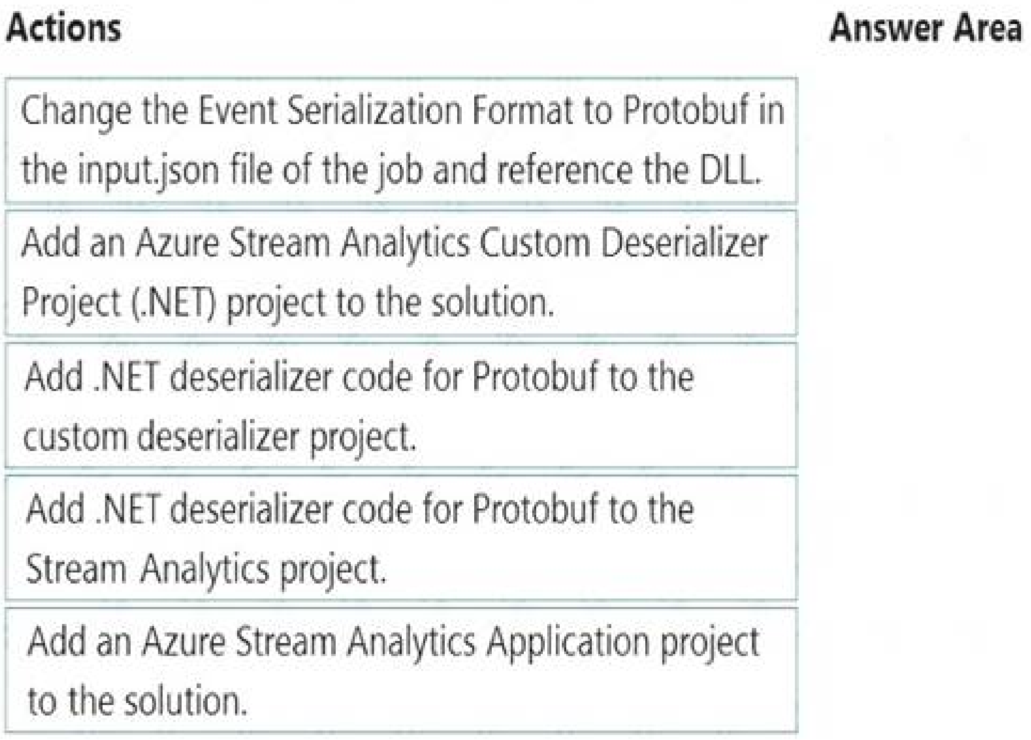
D)

```
sys.dm_db_column_store_row_group_physical_stats
```

A. Option
B. Option
C. Option
D. Option

Answer: B

NEW QUESTION 88
- (Exam Topic 3)
You have an Azure Stream Analytics job that is a Stream Analytics project solution in Microsoft Visual Studio. The job accepts data generated by IoT devices in the JSON format.
You need to modify the job to accept data generated by the IoT devices in the Protobuf format.
Which three actions should you perform from Visual Studio on sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



Actions

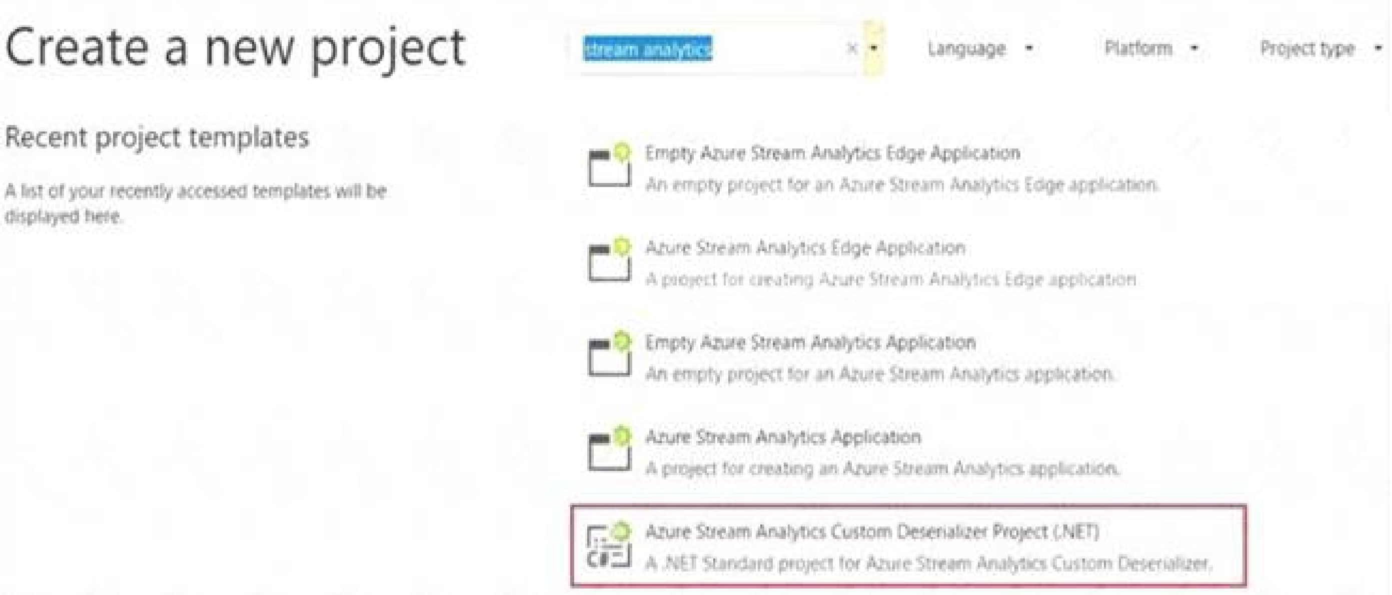
Answer Area

- Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.
- Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.
- Add .NET deserializer code for Protobuf to the custom deserializer project.
- Add .NET deserializer code for Protobuf to the Stream Analytics project.
- Add an Azure Stream Analytics Application project to the solution.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Step 1: Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution. Create a custom deserializer
* 1. Open Visual Studio and select File > New > Project. Search for Stream Analytics and select Azure Stream Analytics Custom Deserializer Project (.NET). Give the project a name, like Protobuf Deserializer.



Create a new project

stream analytics

Recent project templates

A list of your recently accessed templates will be displayed here.

Empty Azure Stream Analytics Edge Application
An empty project for an Azure Stream Analytics Edge application.

Azure Stream Analytics Edge Application
A project for creating Azure Stream Analytics Edge application

Empty Azure Stream Analytics Application
An empty project for an Azure Stream Analytics application.

Azure Stream Analytics Application
A project for creating an Azure Stream Analytics application.

Azure Stream Analytics Custom Deserializer Project (.NET)
A .NET Standard project for Azure Stream Analytics Custom Deserializer.

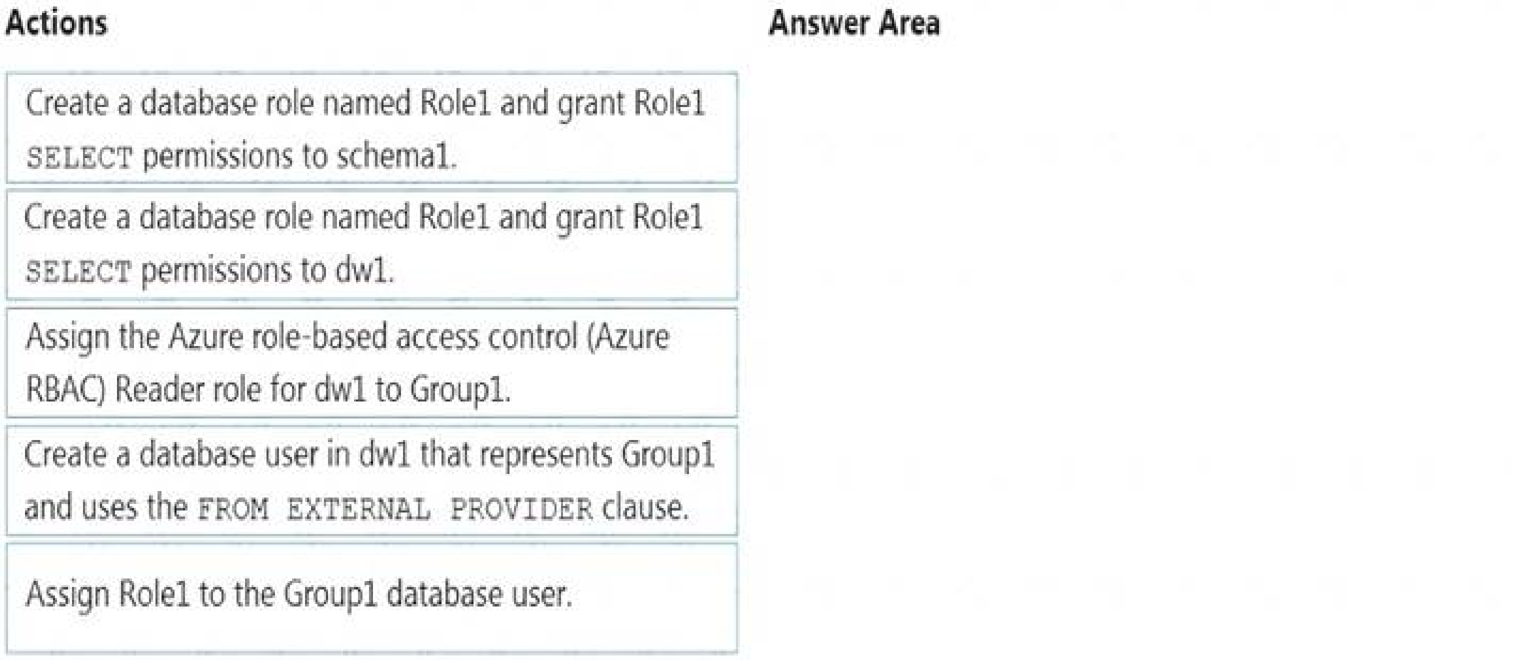
* 2. In Solution Explorer, right-click your Protobuf Deserializer project and select Manage NuGet Packages from the menu. Then install the Microsoft.Azure.StreamAnalytics and Google.Protobuf NuGet packages.
* 3. Add the MessageBodyProto class and the MessageBodyDeserializer class to your project.
* 4. Build the Protobuf Deserializer project.
Step 2: Add .NET deserializer code for Protobuf to the custom deserializer project

Azure Stream Analytics has built-in support for three data formats: JSON, CSV, and Avro. With custom .NET deserializers, you can read data from other formats such as Protocol Buffer, Bond and other user defined formats for both cloud and edge jobs.
Step 3: Add an Azure Stream Analytics Application project to the solution Add an Azure Stream Analytics project

- In Solution Explorer, right-click the Protobuf Deserializer solution and select Add > New Project. Under Azure Stream Analytics > Stream Analytics, choose Azure Stream Analytics Application. Name it ProtobufCloudDeserializer and select OK.
- Right-click References under the ProtobufCloudDeserializer Azure Stream Analytics project. Under Projects, add Protobuf Deserializer. It should be automatically populated for you.

Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/custom-deserializer

NEW QUESTION 92
- (Exam Topic 3)
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.



Actions

Answer Area

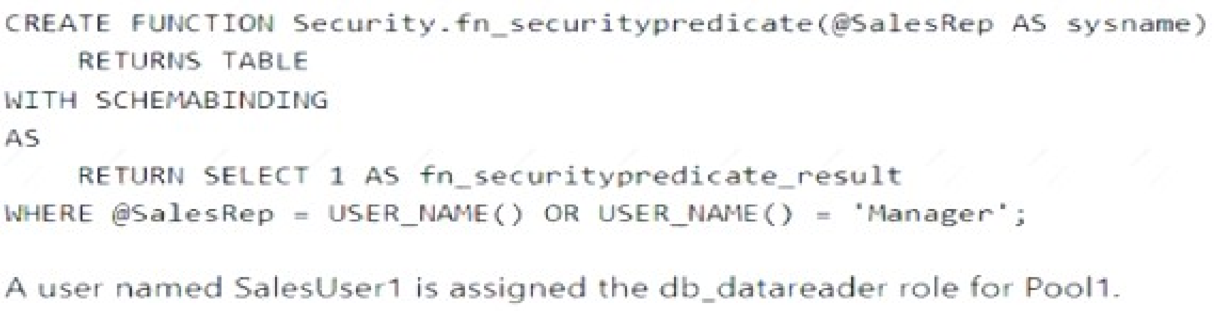
- Create a database role named Role1 and grant Role1 SELECT permissions to schema1.
- Create a database role named Role1 and grant Role1 SELECT permissions to dw1.
- Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1.
- Create a database user in dw1 that represents Group1 and uses the FROM EXTERNAL PROVIDER clause.
- Assign Role1 to the Group1 database user.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Step 1: Create a database role named Role1 and grant Role1 SELECT permissions to schema You need to grant Group1 read-only permissions to all the tables and views in schema1.
Place one or more database users into a database role and then assign permissions to the database role. Step 2: Assign Rol1 to the Group database user
Step 3: Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1 Reference:
https://docs.microsoft.com/en-us/azure/data-share/how-to-share-from-sql

NEW QUESTION 94
- (Exam Topic 3)
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.

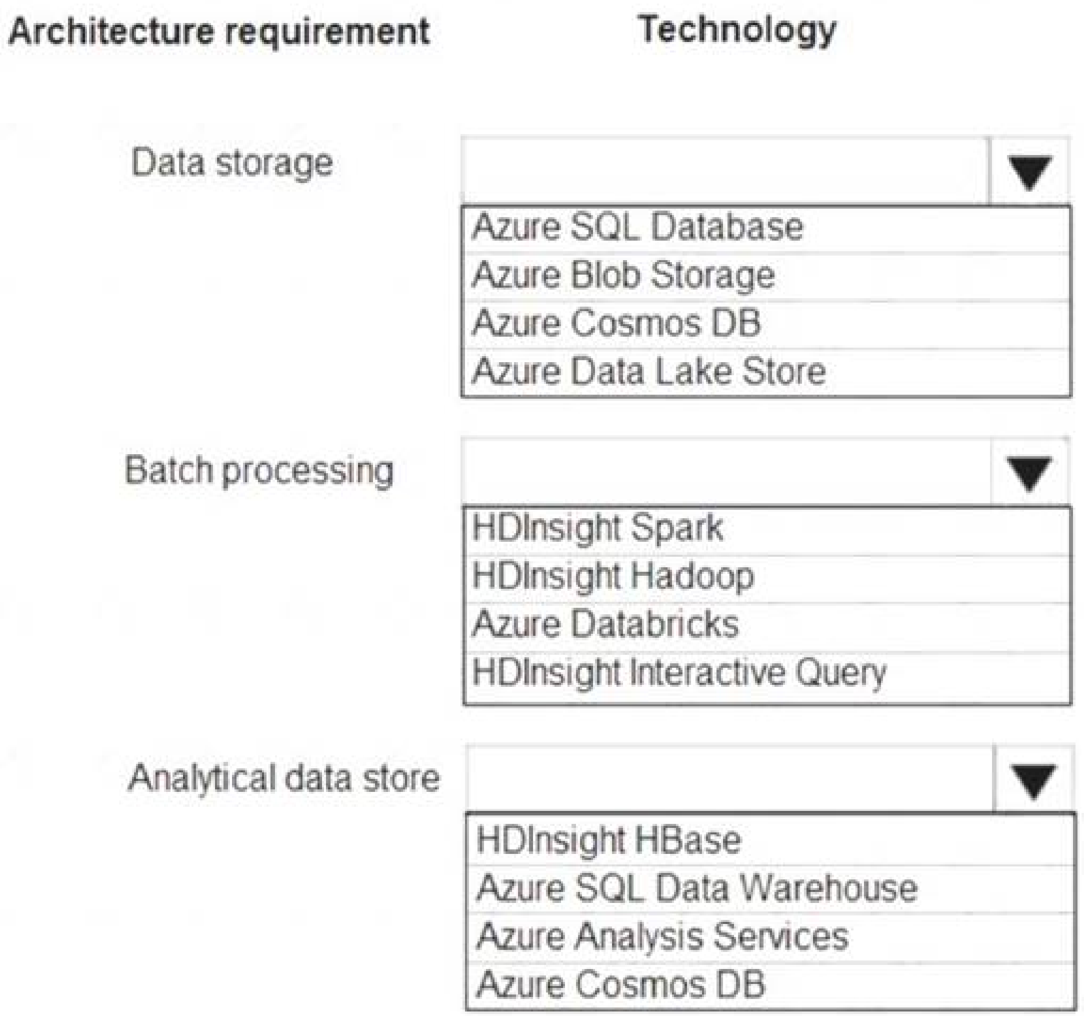
A user named SalesUser1 is assigned the db_datareader role for Pool1. Which rows in the Sales table are returned when SalesUser1 queries the table?

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

Answer: C

NEW QUESTION 99
- (Exam Topic 3)
You are developing a solution using a Lambda architecture on Microsoft Azure. The data at test layer must meet the following requirements:
Data storage:
•Serve as a repository (or high volumes of large files in various formats.

•Implement optimized storage for big data analytics workloads.
•Ensure that data can be organized using a hierarchical structure. Batch processing:
•Use a managed solution for in-memory computation processing.
•Natively support Scala, Python, and R programming languages.
•Provide the ability to resize and terminate the cluster automatically. Analytical data store:
•Support parallel processing.
•Use columnar storage.
•Support SQL-based languages.
You need to identify the correct technologies to build the Lambda architecture.
Which technologies should you use? To answer, select the appropriate options in the answer area NOTE: Each correct selection is worth one point.



Architecture requirement | Technology

Data storage
- Azure SQL Database
- Azure Blob Storage
- Azure Cosmos DB
- Azure Data Lake Store

Batch processing
- HDInsight Spark
- HDInsight Hadoop
- Azure Databricks
- HDInsight Interactive Query

Analytical data store
- HDInsight HBase
- Azure SQL Data Warehouse
- Azure Analysis Services
- Azure Cosmos DB

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Data storage: Azure Data Lake Store
A key mechanism that allows Azure Data Lake Storage Gen2 to provide file system performance at object storage scale and prices is the addition of a hierarchical namespace. This allows the collection of objects/files within an account to be organized into a hierarchy of directories and nested subdirectories in the same way that the file system on your computer is organized. With the hierarchical namespace enabled, a storage account becomes capable of providing the scalability and cost-effectiveness of object storage, with file system semantics that are familiar to analytics engines and frameworks.
Batch processing: HD Insight Spark
Aparch Spark is an open-source, parallel-processing framework that supports in-memory processing to boost the performance of big-data analysis applications.
HDInsight is a managed Hadoop service. Use it deploy and manage Hadoop clusters in Azure. For batch processing, you can use Spark, Hive, Hive LLAP, MapReduce.
Languages: R, Python, Java, Scala, SQL Analytic data store: SQL Data Warehouse
SQL Data Warehouse is a cloud-based Enterprise Data Warehouse (EDW) that uses Massively Parallel Processing (MPP).
SQL Data Warehouse stores data into relational tables with columnar storage. References:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-namespace https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/batch-processing https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-overview-what-is

NEW QUESTION 103
- (Exam Topic 3)
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

Answer: A

Explanation:
Convert nested JSON to a flattened DataFrame
You can to flatten nested JSON, using only $"column.*" and explode methods. Note: Extract and flatten
Use $"column.*" and explode methods to flatten the struct and array types before displaying the flattened DataFrame.
Scala
display(DF.select($"id" as "main_id",$"name",$"batters",$"ppu",explode($"topping")) // Exploding the topping column using explode as it is an array type withColumn("topping_id",$"col.id") // Extracting topping_id from col using DOT form withColumn("topping_type",$"col.type") // Extracting topping_tytpe from col

using DOT form drop($"col")
select($"*",$"batters.*") // Flattened the struct type batters tto array type which is batter drop($"batters")
select($"*",explode($"batter")) drop($"batter")
withColumn("batter_id",$"col.id") // Extracting batter_id from col using DOT form withColumn("battter_type",$"col.type") // Extracting battter_type from col using DOT form drop($"col")
)
Reference: https://learn.microsoft.com/en-us/azure/databricks/kb/scala/flatten-nested-columns-dynamically

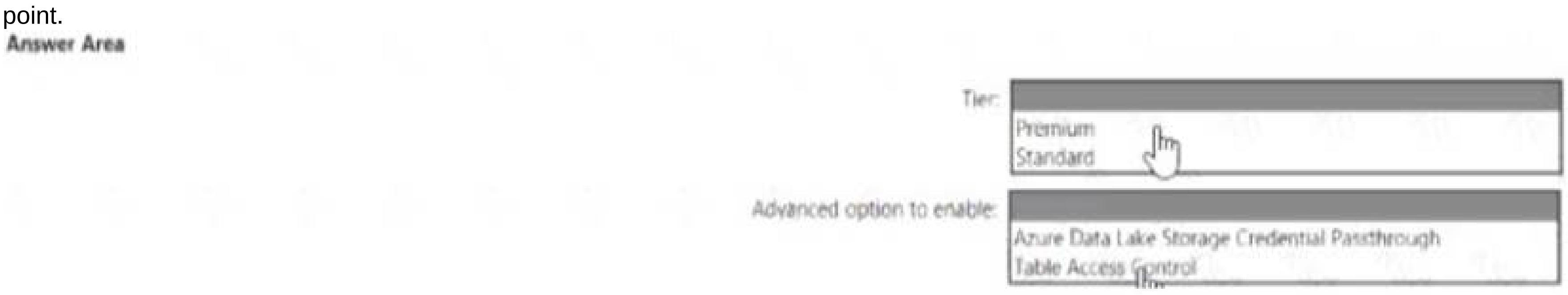
NEW QUESTION 105
- (Exam Topic 3)
You have an enterprise-wide Azure Data Lake Storage Gen2 account. The data lake is accessible only through an Azure virtual network named VNET1.
You are building a SQL pool in Azure Synapse that will use data from the data lake.
Your company has a sales team. All the members of the sales team are in an Azure Active Directory group named Sales. POSIX controls are used to assign the Sales group access to the files in the data lake.
You plan to load data to the SQL pool every hour.
You need to ensure that the SQL pool can load the sales data from the data lake.
Which three actions should you perform? Each correct answer presents part of the solution. NOTE: Each area selection is worth one point.

A. Add the managed identity to the Sales group.
B. Use the managed identity as the credentials for the data load process.
C. Create a shared access signature (SAS).
D. Add your Azure Active Directory (Azure AD) account to the Sales group.
E. Use the snared access signature (SAS) as the credentials for the data load process.
F. Create a managed identity.

Answer: ADF

Explanation:
The managed identity grants permissions to the dedicated SQL pools in the workspace.
Note: Managed identity for Azure resources is a feature of Azure Active Directory. The feature provides Azure services with an automatically managed identity in Azure AD Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-identity

NEW QUESTION 106
- (Exam Topic 3)
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 clutter? To answer, select the appropriate options in the answers area. NOTE: Each correct selection is worth one point.



Answer Area

Tier:
Premium
Standard

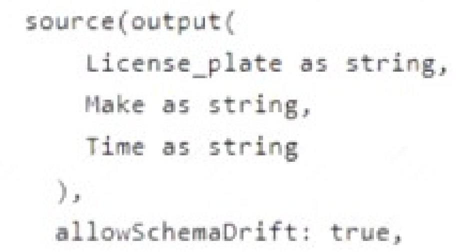
Advanced option to enable:
Azure Data Lake Storage Credential Passthrough
Table Access Control

A. Mastered
B. Not Mastered

Answer: A

Explanation:
https://docs.azuredatabricks.net/spark/latest/data-sources/azure/adls-passthrough.html

NEW QUESTION 108
- (Exam Topic 3)
You have an Azure Data Factory pipeline that contains a data flow. The data flow contains the following expression.

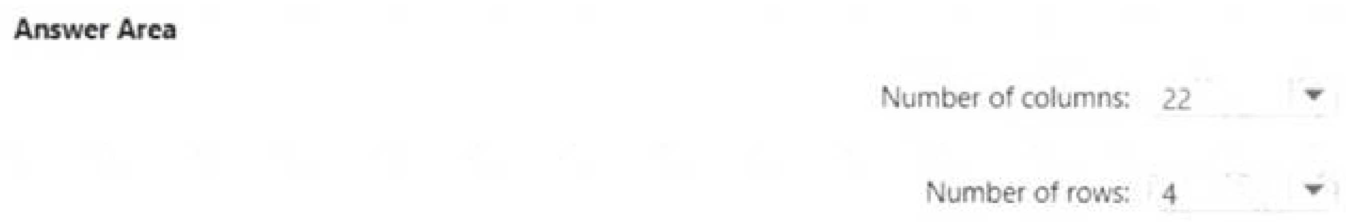


```
source(output(
    License_plate as string,
    Make as string,
    Time as string
  ),
  allowSchemaDrift: true,
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
See the answer in
See below answer.



Answer Area

Number of columns: 22

Number of rows: 4

NEW QUESTION 113
- (Exam Topic 3)
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

Answer: D

Explanation:
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

NEW QUESTION 117
- (Exam Topic 3)
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.

Answer: C

Explanation:
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

NEW QUESTION 118
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. Table1 contains the following:

- One billion rows
- A clustered columnstore index
- A hash-distributed column named Product Key
- A column named Sales Date that is of the date data type and cannot be null Thirty million rows will be added to Table1 each month.

You need to partition Table1 based on the Sales Date column. The solution must optimize query performance and data loading.
How often should you create a partition?

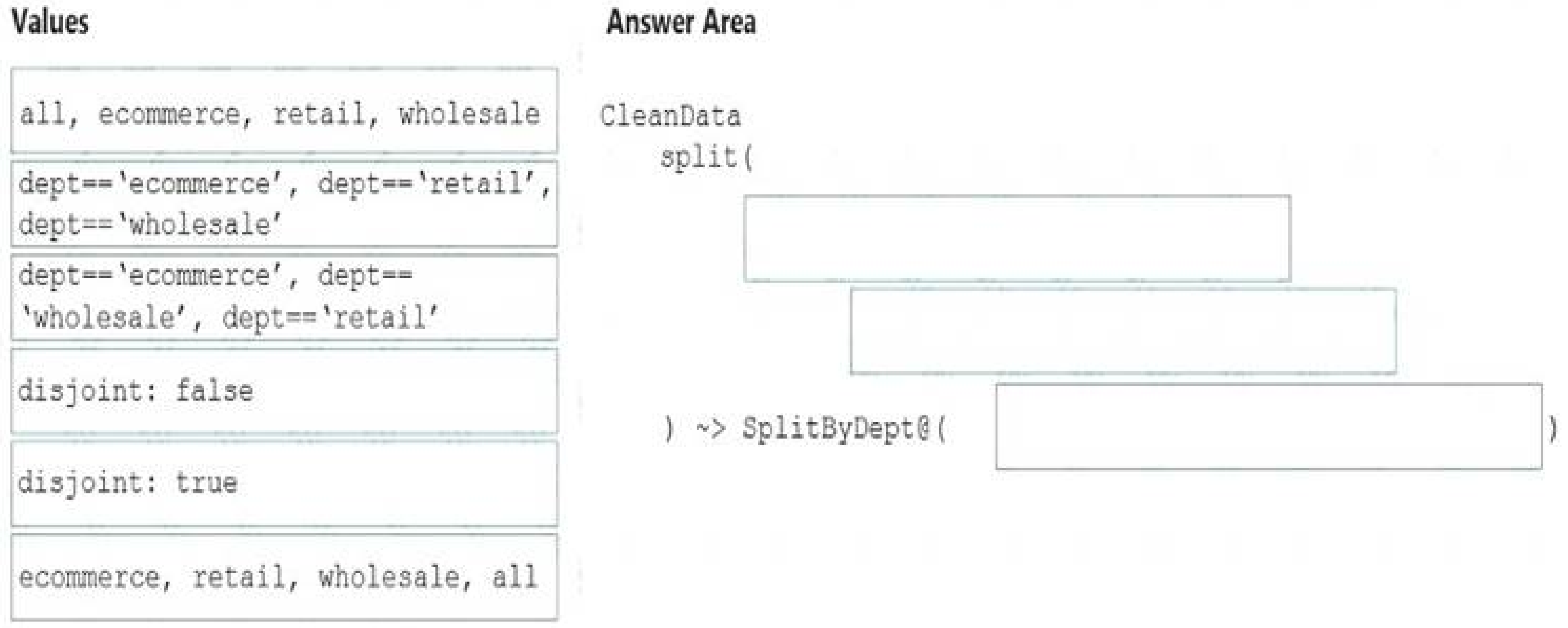
A. once per month
B. once per year
C. once per day
D. once per week

Answer: B

Explanation:
Need a minimum 1 million rows per distribution. Each table is 60 distributions. 30 millions rows is added each month. Need 2 months to get a minimum of 1 million rows per distribution in a new partition.
Note: When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributions.

Any partitioning added to a table is in addition to the distributions created behind the scenes. Using this example, if the sales fact table contained 36 monthly partitions, and given that a dedicated SQL pool has 60 distributions, then the sales fact table should contain 60 million rows per month, or 2.1 billion rows when all months are populated. If a table contains fewer than the recommended minimum number of rows per partition, consider using fewer partitions in order to increase the number of rows per partition.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partitio

NEW QUESTION 119
- (Exam Topic 3)
You need to create an Azure Data Factory pipeline to process data for the following three departments at your company: Ecommerce, retail, and wholesale. The solution must ensure that data can also be processed for the entire company.
How should you complete the Data Factory data flow script? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.



Values

- all, ecommerce, retail, wholesale
- dept=='ecommerce', dept=='retail', dept=='wholesale'
- dept=='ecommerce', dept== 'wholesale', dept=='retail'
- disjoint: false
- disjoint: true
- ecommerce, retail, wholesale, all

Answer Area

```
CleanData
    split(
        [          ]
            [          ]
    ) ~> SplitByDept@( [          ] )
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
The conditional split transformation routes data rows to different streams based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language. The transformation evaluates expressions, and based on the results, directs the data row to the specified stream.
Box 1: dept=='ecommerce', dept=='retail', dept=='wholesale'
First we put the condition. The order must match the stream labeling we define in Box 3. Syntax:
<incomingStream> split(
<conditionalExpression1>
<conditionalExpression2> disjoint: {true | false}
) ~> <splitTx>@(stream1, stream2, ..., <defaultStream>)
Box 2: discount : false
disjoint is false because the data goes to the first matching condition. All remaining rows matching the third condition go to output stream all.
Box 3: ecommerce, retail, wholesale, all Label the streams
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split

NEW QUESTION 120
- (Exam Topic 3)
You have two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.

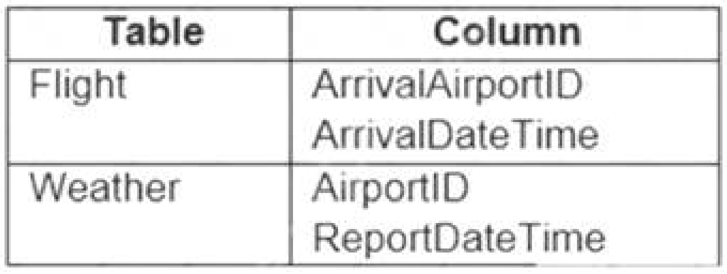


Table	Column
Flight	ArrivalAirportID ArrivalDateTime
Weather	AirportID ReportDateTime

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.

Answer: B

Explanation:
Hash-distribution improves query performance on large fact tables.

NEW QUESTION 122
- (Exam Topic 3)

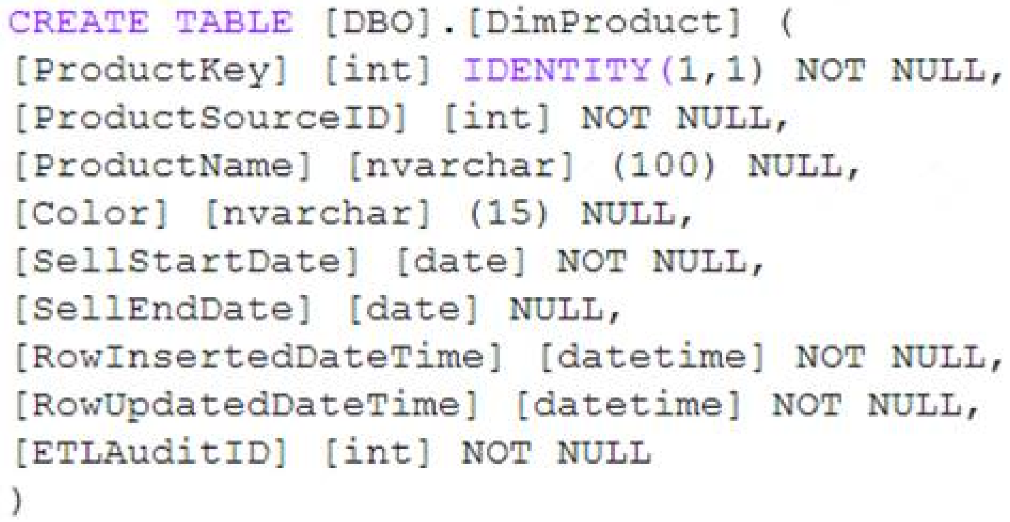
You have an Azure Data lake Storage account that contains a staging zone.
You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: 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.
Dow this meet the goal?

A. Yes
B. No

Answer: B

Explanation:
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

NEW QUESTION 124
- (Exam Topic 3)
You need to implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool.
You have a table that was created by using the following Transact-SQL statement.



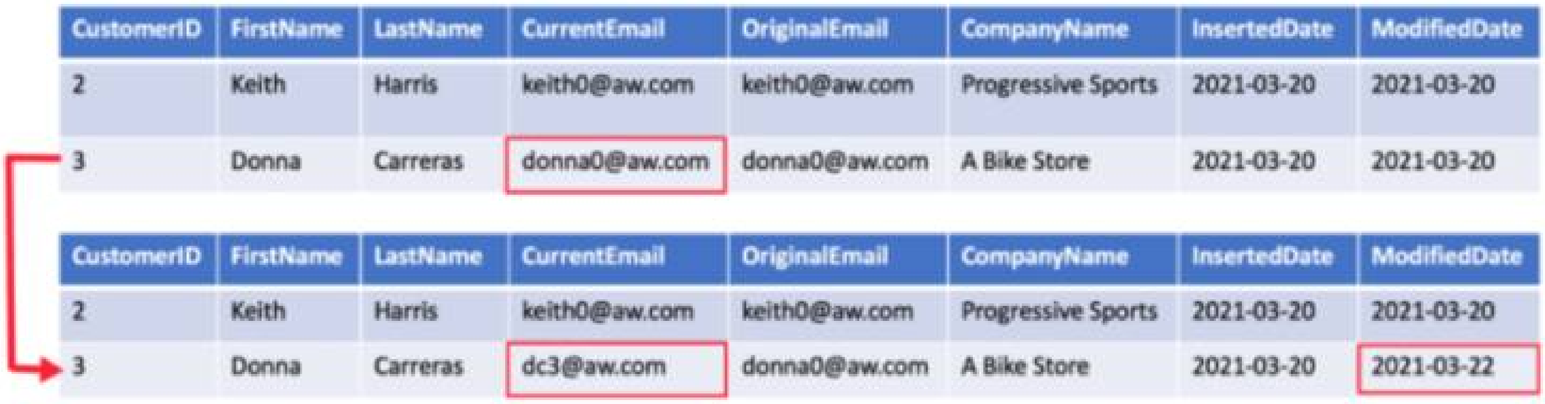
```
CREATE TABLE [DBO].[DimProduct] (
[ProductKey] [int] IDENTITY(1,1) NOT NULL,
[ProductSourceID] [int] NOT NULL,
[ProductName] [nvarchar] (100) NULL,
[Color] [nvarchar] (15) NULL,
[SellStartDate] [date] NOT NULL,
[SellEndDate] [date] NULL,
[RowInsertedDateTime] [datetime] NOT NULL,
[RowUpdatedDateTime] [datetime] NOT NULL,
[ETLAuditID] [int] NOT NULL
)
```

Which two columns should you add to the table? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. [EffectiveScarcDate] [datetime] NOT NULL,
B. [CurrentProduccCacegory] [nvarchar] (100) NOT NULL,
C. [EffectiveEndDace] [dacecime] NULL,
D. [ProductCategory] [nvarchar] (100) NOT NULL,
E. [OriginalProduccCacegory] [nvarchar] (100) NOT NULL,

Answer: BE

Explanation:
A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD.
This type of tracking may be used for one or two columns in a dimension table. It is not common to use it for many members of the same table. It is often used in combination with Type 1 or Type 2 members.
Graphical user interface, application, email Description automatically generated

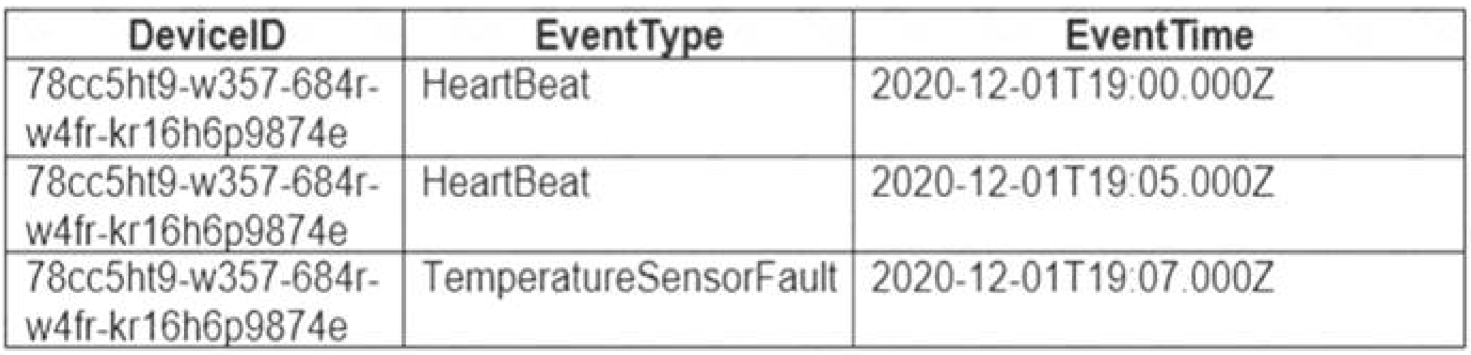


CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	donna0@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-20

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	dc3@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-22

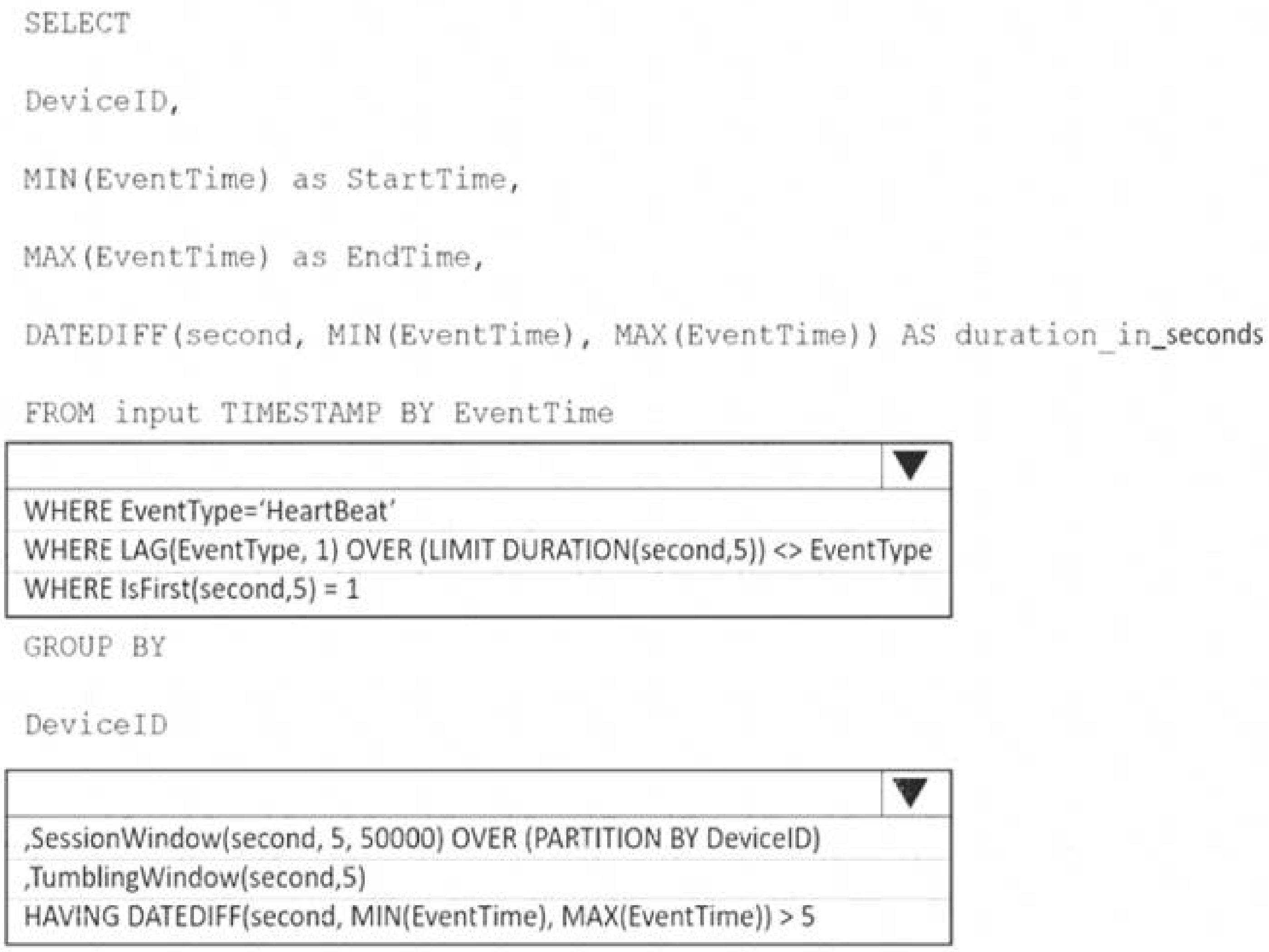
Reference:
https://k21academy.com/microsoft-azure/azure-data-engineer-dp203-q-a-day-2-live-session-review/

NEW QUESTION 126
- (Exam Topic 3)
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.



```
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

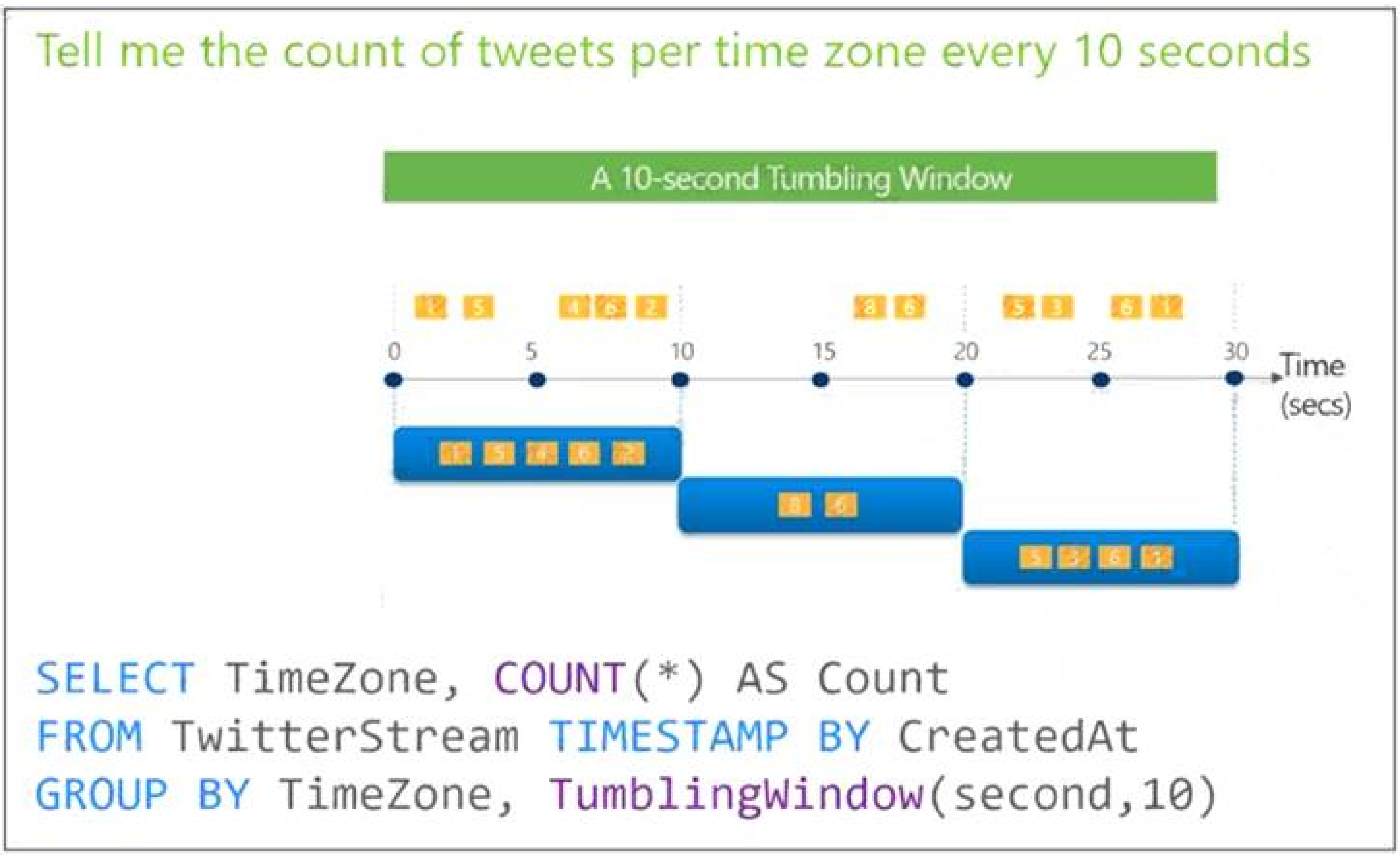
```
GROUP BY

DeviceID
```

- ,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID)
- ,TumblingWindow(second,5)
- HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated
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.
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/stream-analytics-query/session-window-azure-stream-analytics https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

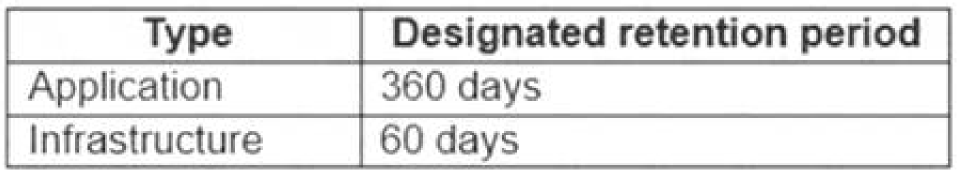
NEW QUESTION 129
- (Exam Topic 3)
You have an Azure Synapse Analytics Apache Spark pool named Pool1.
You plan to load JSON files from an Azure Data Lake Storage Gen2 container into the tables in Pool1. The structure and data types vary by file.
You need to load the files into the tables. The solution must maintain the source data types. What should you do?

A. Use a Get Metadata activity in Azure Data Factory.
B. Use a Conditional Split transformation in an Azure Synapse data flow.
C. Load the data by using the OPEHROwset Transact-SQL command in an Azure Synapse Anarytics serverless SQL pool.
D. Load the data by using PySpark.

Answer: A

Explanation:
Serverless SQL pool can automatically synchronize metadata from Apache Spark. A serverless SQL pool database will be created for each database existing in serverless Apache Spark pools.
Serverless SQL pool enables you to query data in your data lake. It offers a T-SQL query surface area that accommodates semi-structured and unstructured data queries.
To support a smooth experience for in place querying of data that's located in Azure Storage files, serverless SQL pool uses the OPENROWSET function with additional capabilities.
The easiest way to see to the content of your JSON file is to provide the file URL to the OPENROWSET function, specify csv FORMAT.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-json-files https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-data-storage

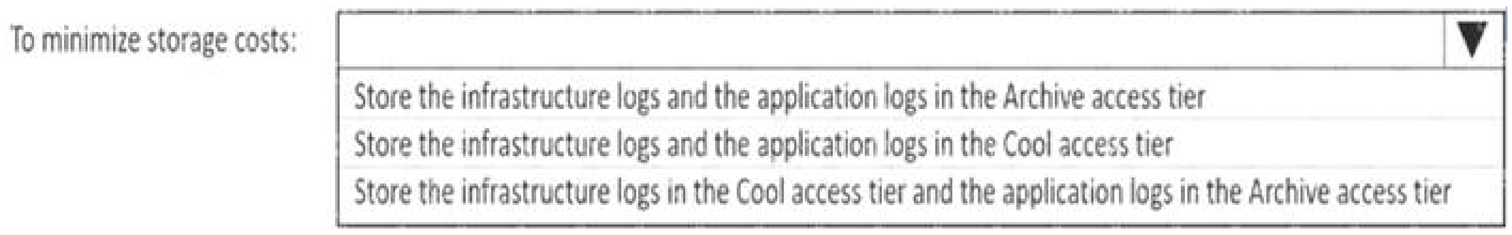
NEW QUESTION 132
- (Exam Topic 3)
You have an Azure Data Lake Storage Gen2 account named account1 that stores logs as shown in the following table.



Type	Designated retention period
Application	360 days
Infrastructure	60 days

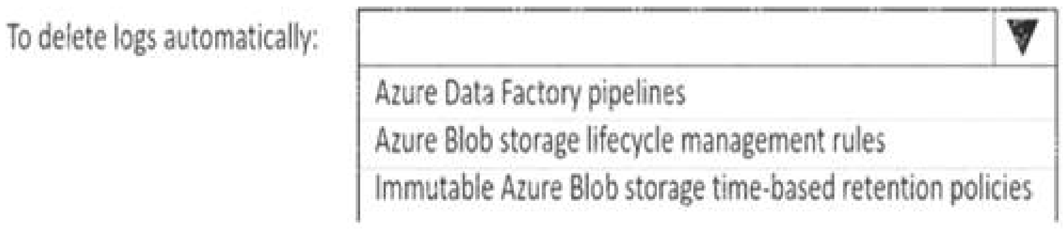
You do not expect that the logs will be accessed during the retention periods.
You need to recommend a solution for account1 that meets the following requirements:
- Automatically deletes the logs at the end of each retention period
- Minimizes storage costs

What should you include in the recommendation? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



To minimize storage costs:

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



To delete logs automatically:

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

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 134
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly. Solution: You copy the files to a table that has a columnstore index. Does this meet the goal?

A. Yes
B. No

Answer: B

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

NEW QUESTION 137
- (Exam Topic 3)
You are designing an anomaly detection solution for streaming data from an Azure IoT hub. The solution must meet the following requirements:

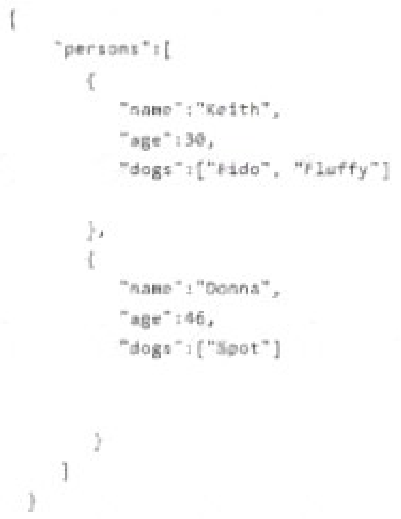
- Send the output to Azure Synapse.
- Identify spikes and dips in time series data.
- Minimize development and configuration effort. Which should you include in the solution?

A. Azure Databricks
B. Azure Stream Analytics
C. Azure SQL Database

Answer: B

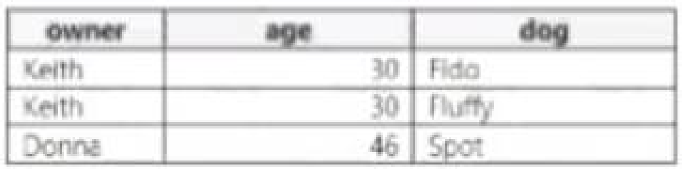
Explanation:
You can identify anomalies by routing data via IoT Hub to a built-in ML model in Azure Stream Analytics. Reference:
https://docs.microsoft.com/en-us/learn/modules/data-anomaly-detection-using-azure-iot-hub/

NEW QUESTION 140
- (Exam Topic 3)
You use PySpark in Azure Databricks to parse the following JSON input.



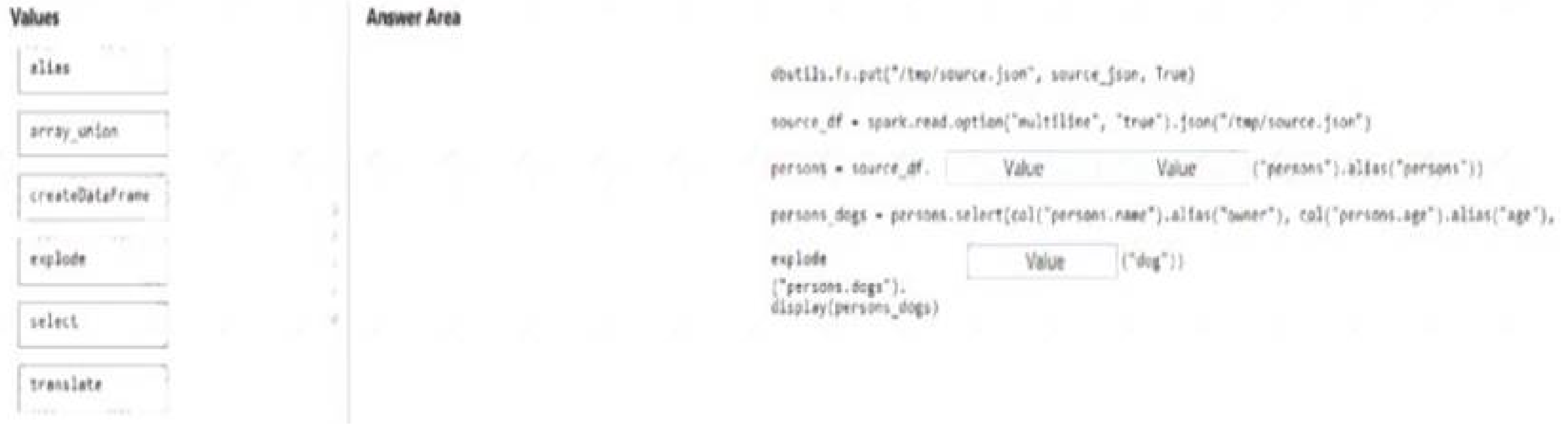
```
{
    "persons":[
      {
        "name":"Keith",
        "age":30,
        "dogs":["Fido", "Fluffy"]

      },
      {
        "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)
```

A. Mastered
B. Not Mastered

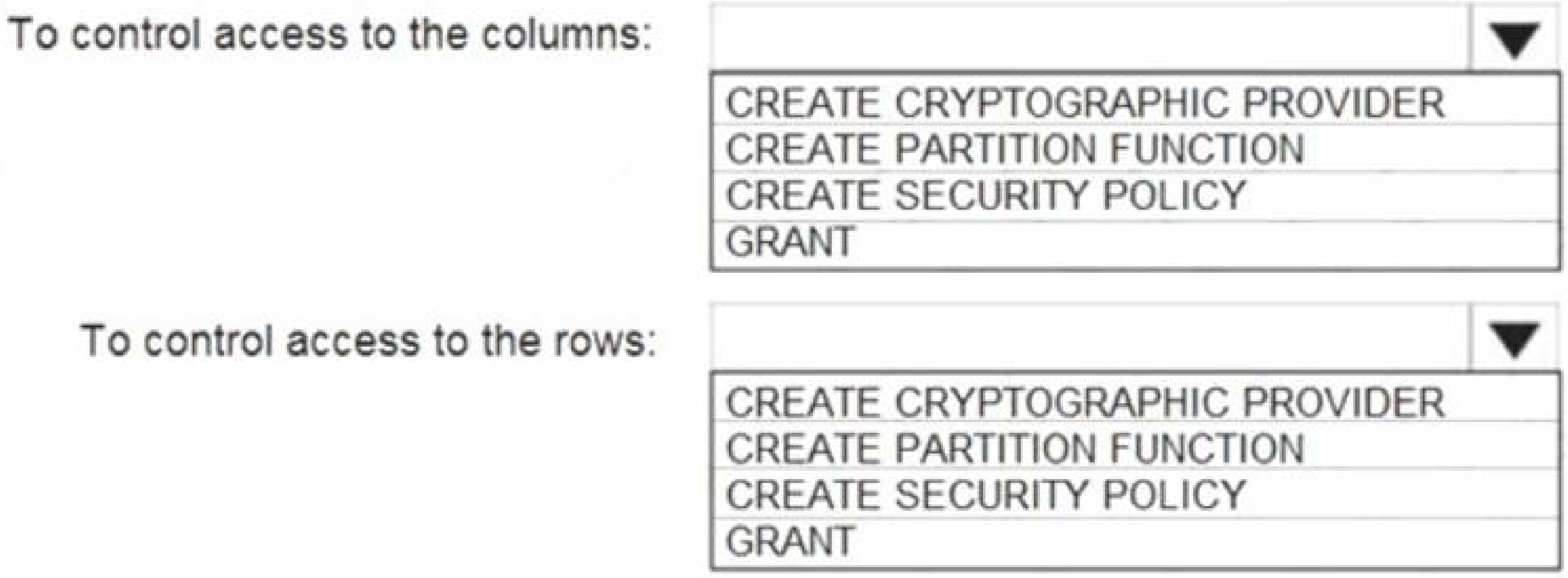
Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated
Box 1: select
Box 2: explode
Bop 3: alias
pyspark.sql.Column.alias returns this column aliased with a new name or names (in the case of expressions that return more than one column, such as explode).
Reference: https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.sql.Column.alias.html https://docs.microsoft.com/en-us/azure/databricks/sql/language-manual/functions/explode

NEW QUESTION 145
- (Exam Topic 3)
You have an Azure subscription that contains the following resources:

- An Azure Active Directory (Azure AD) tenant that contains a security group named Group1
- An Azure Synapse Analytics SQL pool named Pool1

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.



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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Text Description automatically generated
Box 1: GRANT
You can implement column-level security with the GRANT T-SQL statement. Box 2: CREATE SECURITY POLICY
Implement Row Level Security by using the CREATE SECURITY POLICY Transact-SQL statement Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security

NEW QUESTION 150
- (Exam Topic 3)
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. Assign Azure AD security groups to Azure Data Lake Storage.
B. Configure end-user authentication for the Azure Data Lake Storage account.
C. Configure service-to-service authentication for the Azure Data Lake Storage account.
D. Create security groups in Azure Active Directory (Azure AD) and add project members.
E. Configure access control lists (ACL) for the Azure Data Lake Storage account.

Answer: ADE

Explanation:
References:
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-secure-data

NEW QUESTION 155
- (Exam Topic 3)
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 SOL.
- 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

Answer: B

Explanation:
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

NEW QUESTION 158
- (Exam Topic 3)
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 theGet-AzOperationalInsightsSearchResult PowerShell cmdlet.
D. On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.

Answer: A

Explanation:
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

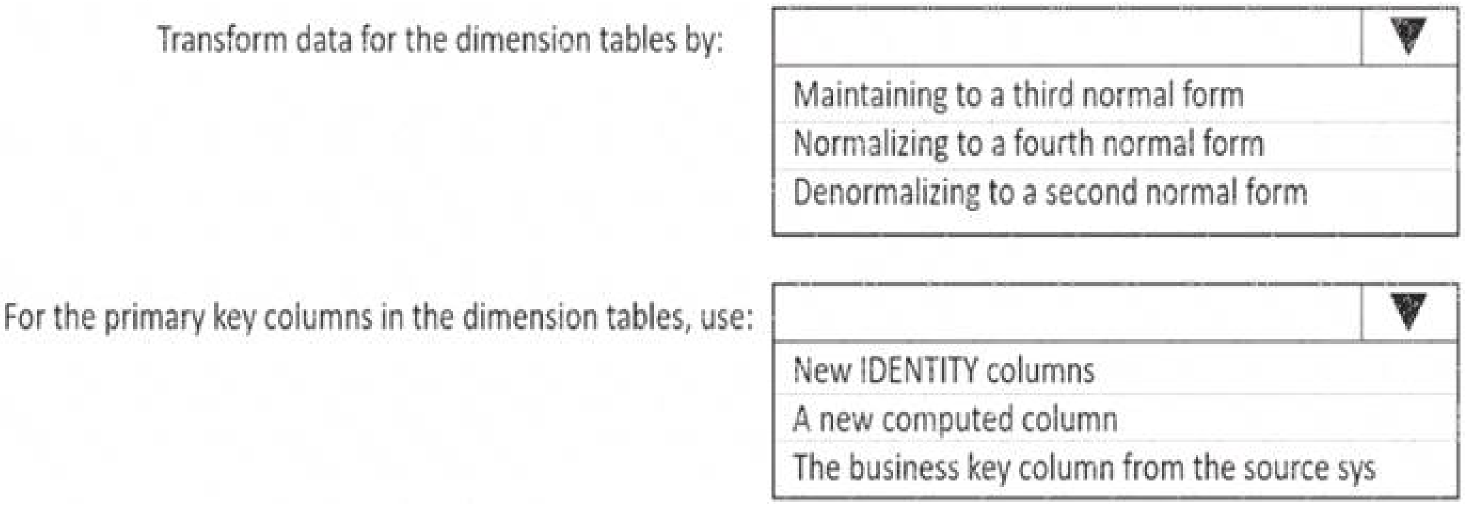
NEW QUESTION 162
- (Exam Topic 3)
You are designing a dimension table for a data warehouse. The table will track the value of the dimension attributes over time and preserve the history of the data by adding new rows as the data changes.
Which type of slowly changing dimension (SCD) should use?

A. Type 0
B. Type 1
C. Type 2
D. Type 3

Answer: C

Explanation:
Type 2 - Creating a new additional record. In this methodology all history of dimension changes is kept in the database. You capture attribute change by adding a new row with a new surrogate key to the dimension table. Both the prior and new rows contain as attributes the natural key(or other durable identifier). Also 'effective date' and 'current indicator' columns are used in this method. There could be only one record with current indicator set to 'Y'. For 'effective date' columns, i.e. start_date and end_date, the end_date for current record usually is set to value 9999-12-31. Introducing changes to the dimensional model in type 2 could be very expensive database operation so it is not recommended to use it in dimensions where a new attribute could be added in the future.
https://www.datawarehouse4u.info/SCD-Slowly-Changing-Dimensions.html

NEW QUESTION 163
- (Exam Topic 3)
You have a Microsoft SQL Server database that uses a third normal form schema.
You plan to migrate the data in the database to a star schema in an Azure Synapse Analytics dedicated SQl pool.
You need to design the dimension tables. The solution must optimize read operations.
What should you include in the solution? to answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.



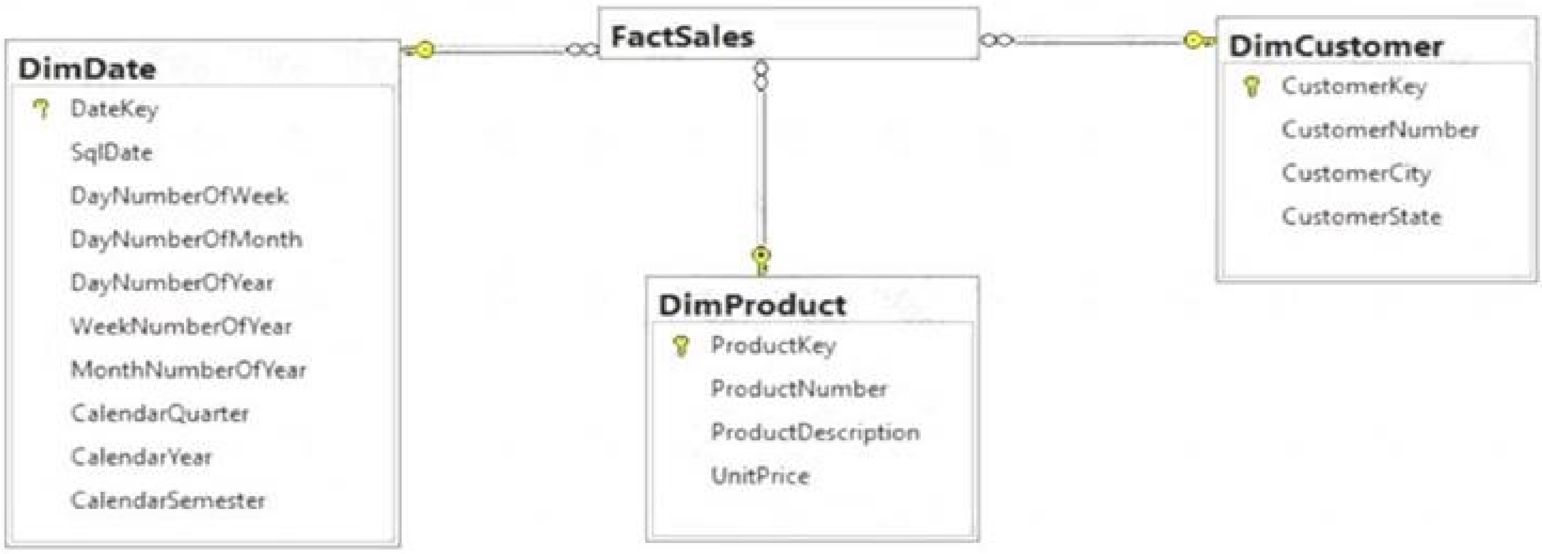
Transform data for the dimension tables by:
- Maintaining to a third normal form
- Normalizing to a fourth normal form
- Denormalizing to a second normal form

For the primary key columns in the dimension tables, use:
- New IDENTITY columns
- A new computed column
- The business key column from the source sys

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Text, table Description automatically generated
Box 1: Denormalize to a second normal form
Denormalization is the process of transforming higher normal forms to lower normal forms via storing the join of higher normal form relations as a base relation. Denormalization increases the performance in data retrieval at cost of bringing update anomalies to a database.
Box 2: New identity columns
The collapsing relations strategy can be used in this step to collapse classification entities into component entities to obtain at dimension tables with single-part keys that connect directly to the fact table. The single-part key is a surrogate key generated to ensure it remains unique over time.
Example:
Diagram Description automatically generated



Note: A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.
Reference:

https://www.mssqltips.com/sqlservertip/5614/explore-the-role-of-normal-forms-in-dimensional-modeling/ https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identity

NEW QUESTION 164
- (Exam Topic 3)
You plan to ingest streaming social media data by using Azure Stream Analytics. The data will be stored in files in Azure Data Lake Storage, and then consumed by using Azure Datiabricks and PolyBase in Azure Synapse Analytics.
You need to recommend a Stream Analytics data output format to ensure that the queries from Databricks and PolyBase against the files encounter the fewest possible errors. The solution must ensure that the tiles can be queried quickly and that the data type information is retained.
What should you recommend?

A. Parquet
B. Avro
C. CSV
D. JSON

Answer: A

Explanation:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-define-outputs

NEW QUESTION 165
- (Exam Topic 3)
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.

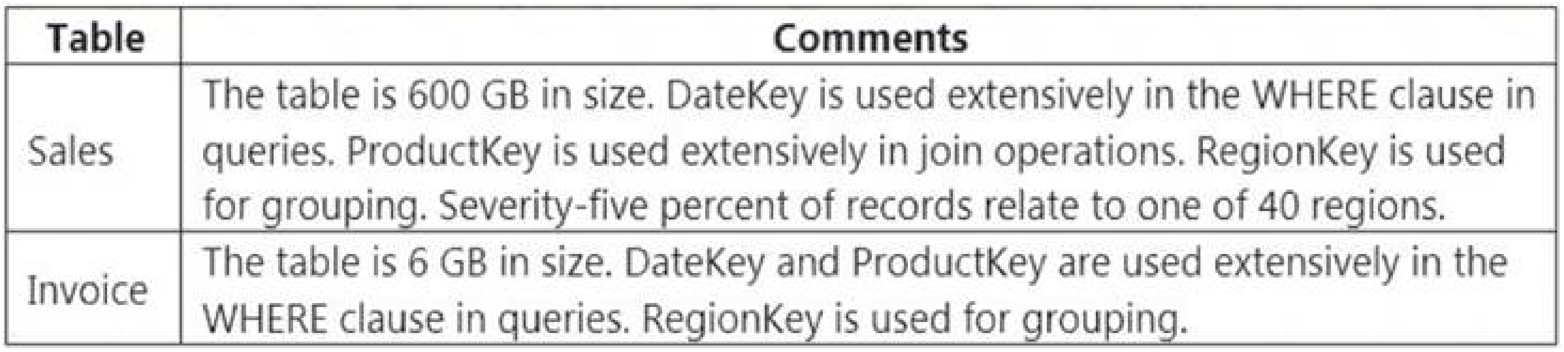
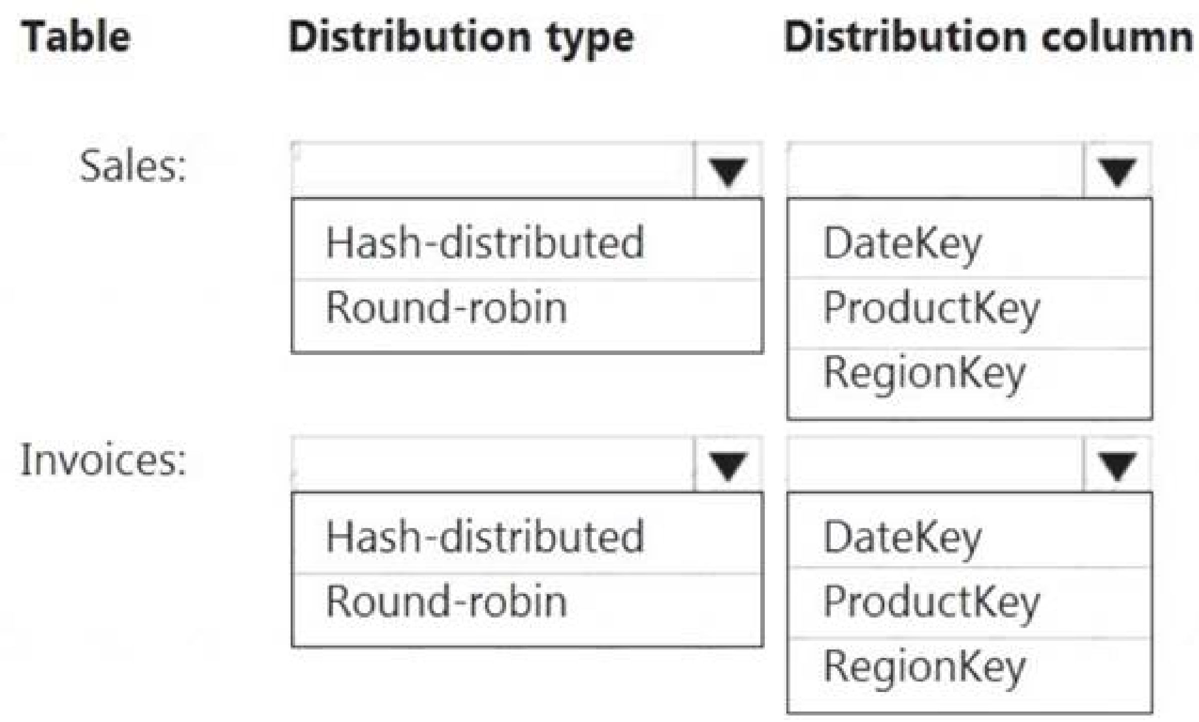


Table	Comments
Sales	The table is 600 GB in size. DateKey is used extensively in the WHERE clause in queries. ProductKey is used extensively in join operations. RegionKey is used for grouping. Severity-five percent of records relate to one of 40 regions.
Invoice	The table is 6 GB in size. DateKey and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



A. Mastered
B. Not Mastered

Answer: A

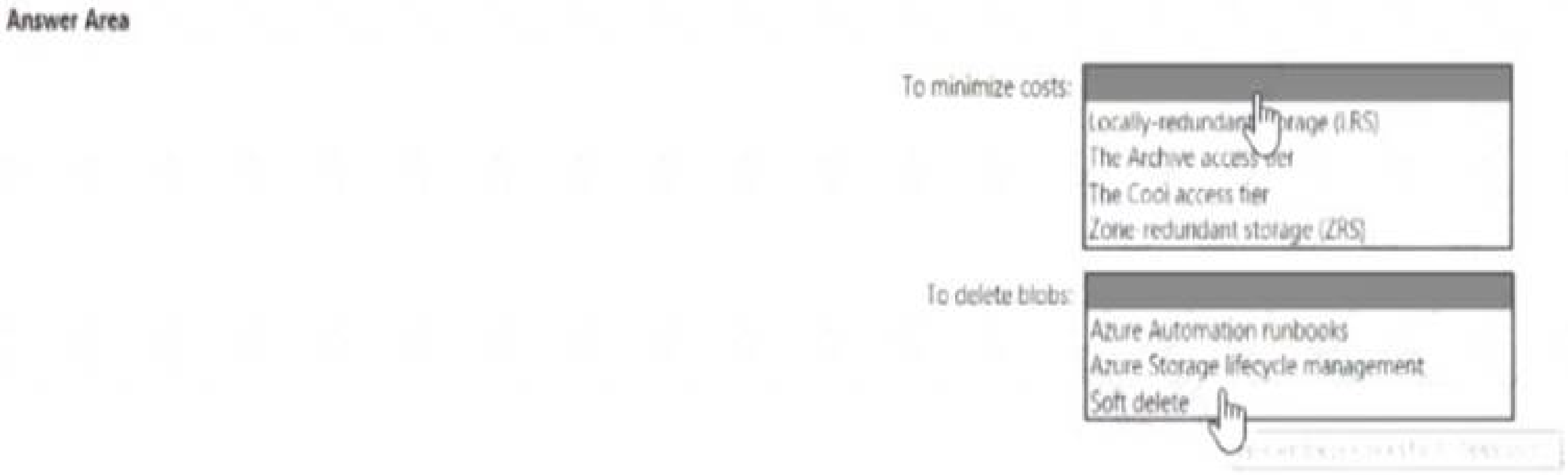
Explanation:
Box 1: Hash-distributed
Box 2: ProductKey
ProductKey is used extensively in joins.
Hash-distributed tables improve query performance on large fact tables.
Box 3: Round-robin
Box 4: RegionKey
Round-robin tables are useful for improving loading speed.
Consider using the round-robin distribution 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

NEW QUESTION 166
- (Exam Topic 3)
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 area. 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)

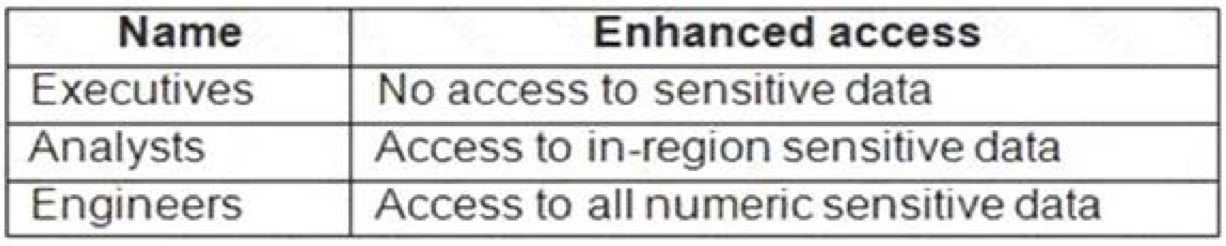
To delete blobs:
Azure Automation runbooks
Azure Storage lifecycle management
Soft delete

A. Mastered
B. Not Mastered

Answer: A

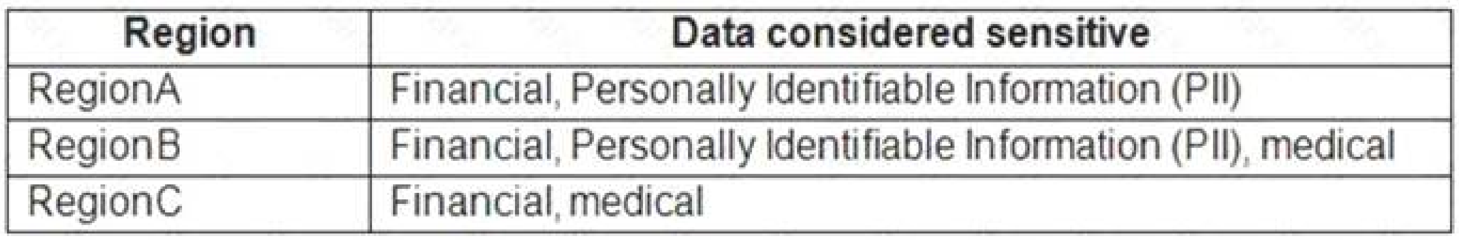
Explanation:
https://learn.microsoft.com/en-us/azure/storage/blobs/premium-tier-for-data-lake-storage

NEW QUESTION 167
- (Exam Topic 3)
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.



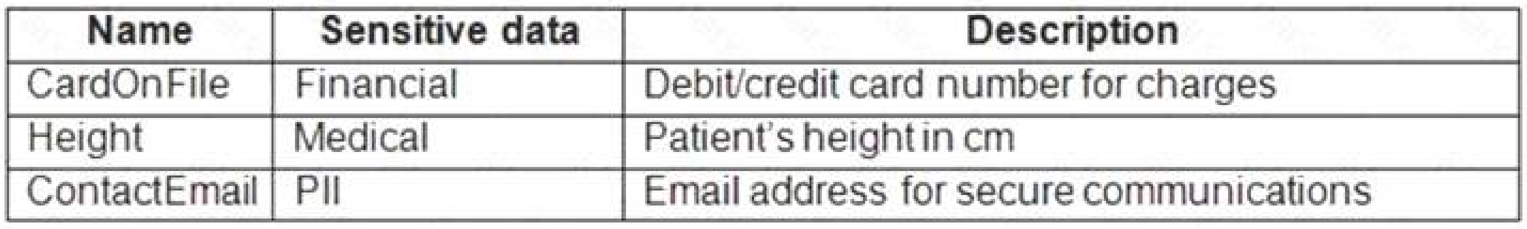
Name	Enhanced access
Executives	No access to sensitive data
Analysts	Access to in-region sensitive data
Engineers	Access to all numeric sensitive data

You have policies for the sensitive data. The policies vary be region as shown in the following table.



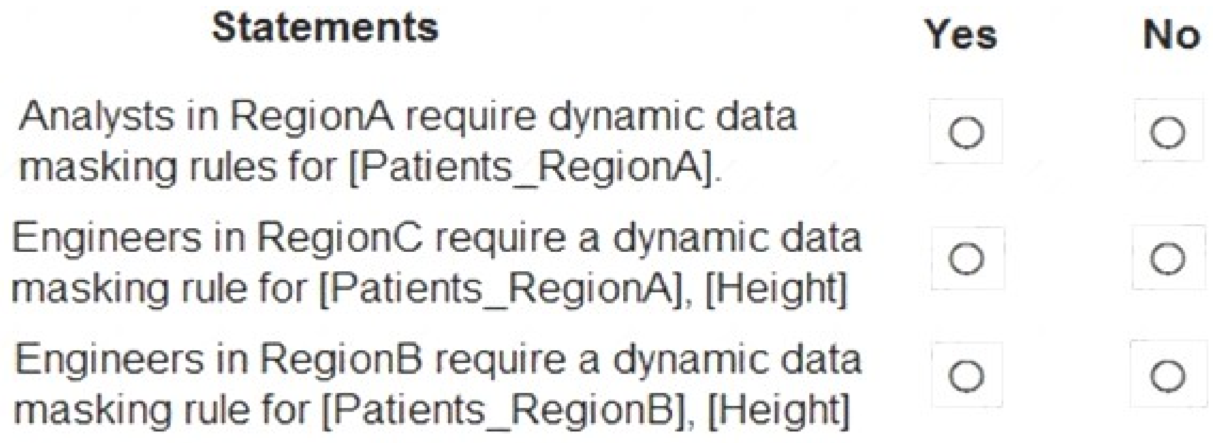
Region	Data considered sensitive
RegionA	Financial, Personally Identifiable Information (PII)
RegionB	Financial, Personally Identifiable Information (PII), medical
RegionC	Financial, medical

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.



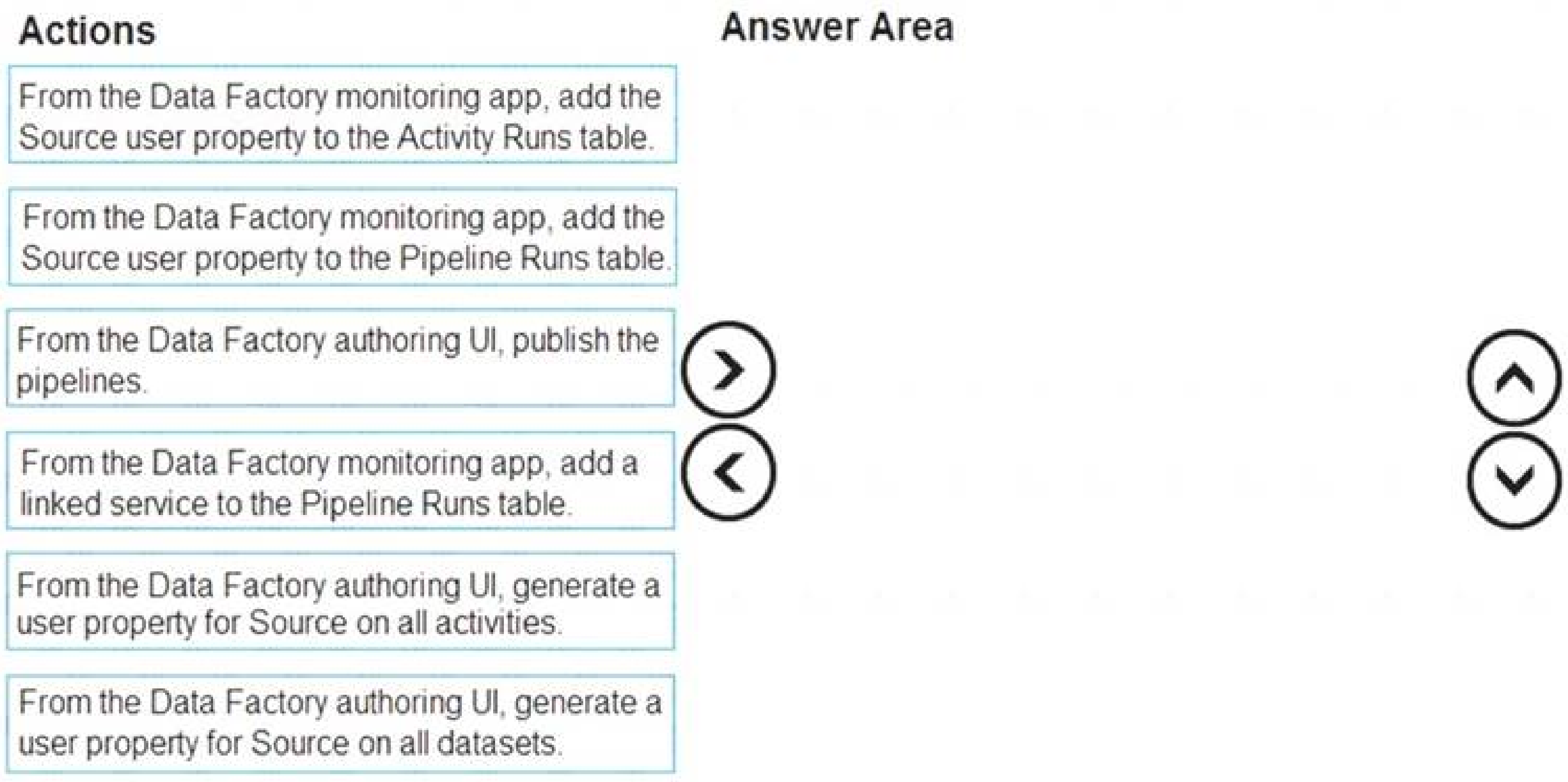
Statements	Yes	No
Analysts in RegionA require dynamic data masking rules for [Patients_RegionA].	○	○
Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height]	○	○
Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height]	○	○

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Text Description automatically generated
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

NEW QUESTION 169
- (Exam Topic 3)
You plan to monitor an Azure data factory by using the Monitor & Manage app.
You need to identify the status and duration of activities that reference a table in a source database.
Which three actions should you perform in sequence? To answer, move the actions from the list of actions to the answer are and arrange them in the correct order.



Actions

- From the Data Factory monitoring app, add the Source user property to the Activity Runs table.
- From the Data Factory monitoring app, add the Source user property to the Pipeline Runs table.
- From the Data Factory authoring UI, publish the pipelines.
- From the Data Factory monitoring app, add a linked service to the Pipeline Runs table.
- From the Data Factory authoring UI, generate a user property for Source on all activities.
- From the Data Factory authoring UI, generate a user property for Source on all datasets.

Answer Area

A. Mastered
B. Not Mastered

Answer: A

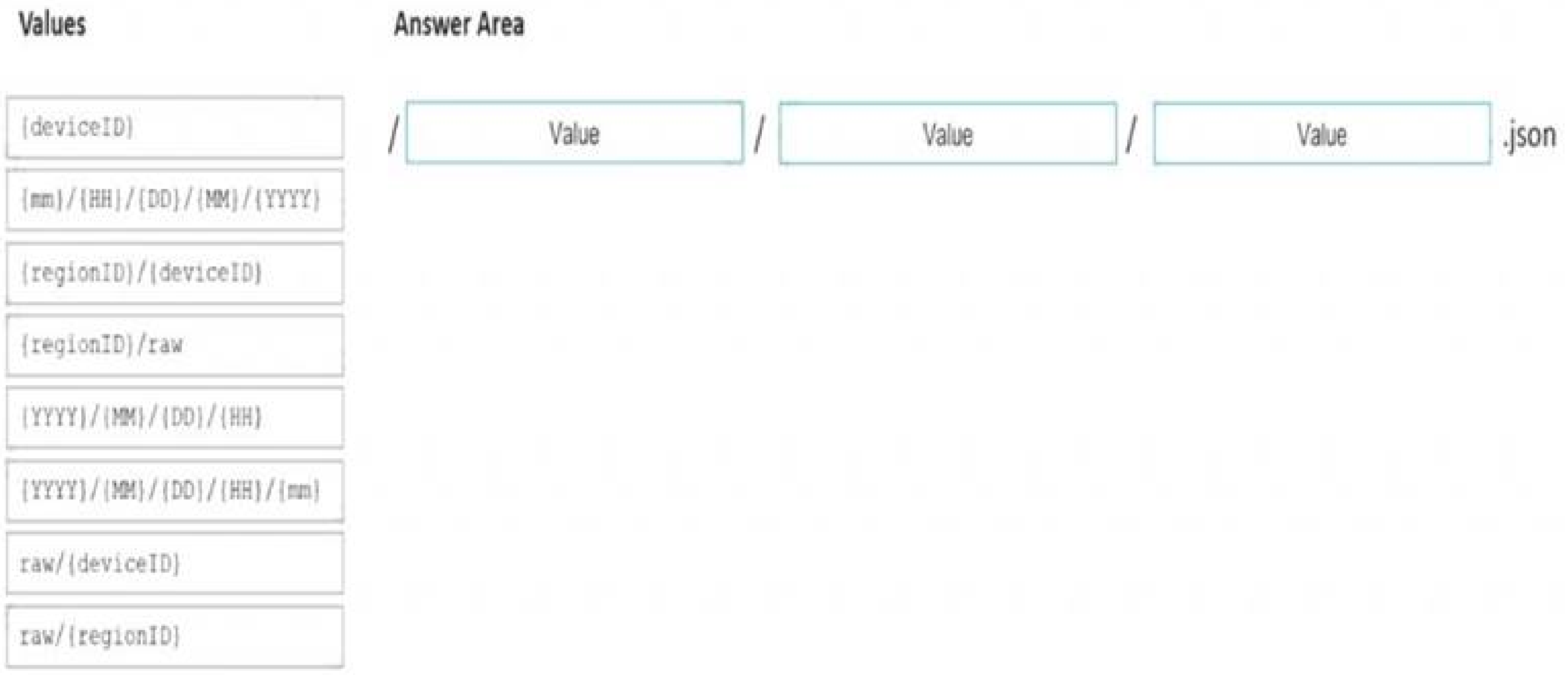
Explanation:
Step 1: From the Data Factory authoring UI, generate a user property for Source on all activities. Step 2: From the Data Factory monitoring app, add the Source user property to Activity Runs table.
You can promote any pipeline activity property as a user property so that it becomes an entity that you can
monitor. For example, you can promote the Source and Destination properties of the copy activity in your pipeline as user properties. You can also select Auto Generate to generate the Source and Destination user properties for a copy activity.
Step 3: From the Data Factory authoring UI, publish the pipelines
Publish output data to data stores such as Azure SQL Data Warehouse for business intelligence (BI) applications to consume.
References:
https://docs.microsoft.com/en-us/azure/data-factory/monitor-visually

NEW QUESTION 173
- (Exam Topic 3)
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.



Values

{deviceID}
{mm}/{HH}/{DD}/{MM}/{YYYY}
{regionID}/{deviceID}
{regionID}/raw
{YYYY}/{MM}/{DD}/{HH}
{YYYY}/{MM}/{DD}/{HH}/{mm}
raw/{deviceID}
raw/{regionID}

Answer Area

/ Value / Value / Value .json

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: {YYYY}/{MM}/{DD}/{HH}
Date Format [optional]: if the date token is used in the prefix path, you can select the date format in which your files are organized. Example: YYYY/MM/DD
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.
Box 2: {regionID}/raw
Data engineers from each region must be able to build their own pipelines for the data of their respective region only.
Box 3: {deviceID} Reference:
https://github.com/paolosalvatori/StreamAnalyticsAzureDataLakeStore/blob/master/README.md

NEW QUESTION 174
- (Exam Topic 3)
You have two Azure Data Factory instances named ADFdev and ADFprod. ADFdev connects to an Azure DevOps Git repository.
You publish changes from the main branch of the Git repository to ADFdev. You need to deploy the artifacts from ADFdev to ADFprod.
What should you do first?

A. From ADFdev, modify the Git configuration.
B. From ADFdev, create a linked service.
C. From Azure DevOps, create a release pipeline.
D. From Azure DevOps, update the main branch.

Answer: C

Explanation:
In Azure Data Factory, continuous integration and delivery (CI/CD) means moving Data Factory pipelines from one environment (development, test, production) to another.
Note:
The following is a guide for setting up an Azure Pipelines release that automates the deployment of a data factory to multiple environments.
- In Azure DevOps, open the project that's configured with your data factory.
- On the left side of the page, select Pipelines, and then select Releases.
- Select New pipeline, or, if you have existing pipelines, select New and then New release pipeline.
- In the Stage name box, enter the name of your environment.
- Select Add artifact, and then select the git repository configured with your development data factory.

Select the publish branch of the repository for the Default branch. By default, this publish branch is adf_publish.
- Select the Empty job template.

Reference:
https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment

NEW QUESTION 179
- (Exam Topic 3)
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.

Answer: B

Explanation:
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-overviewmana

NEW QUESTION 184
- (Exam Topic 3)
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. Append
B. Update
C. Complete

Answer: B

Explanation:
By default, streams run in append mode, which adds new records to the table. https://docs.databricks.com/delta/delta-streaming.html

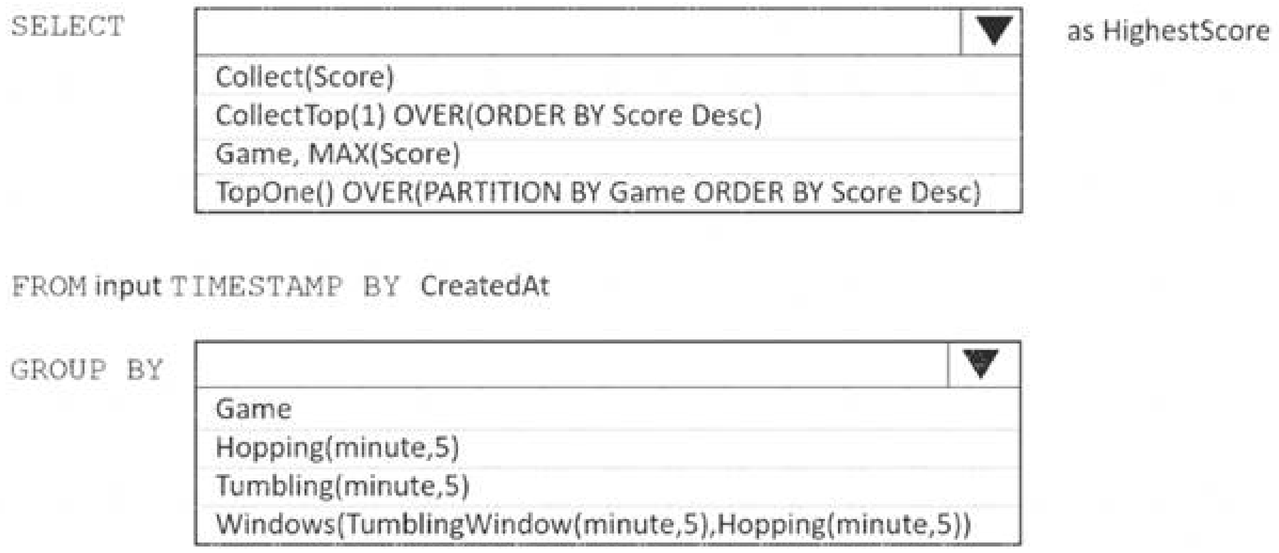
NEW QUESTION 185
- (Exam Topic 3)
You are designing a folder structure for the files m an Azure Data Lake Storage Gen2 account. The account has one container that contains three years of data.
You need to recommend a folder structure that meets the following requirements:
• Supports partition elimination for queries by Azure Synapse Analytics serverless SQL pooh
• Supports fast data retrieval for data from the current month
• Simplifies data security management by department Which folder structure should you recommend?

A. \YYY\MM\DD\Department\DataSource\DataFile_YYYMMMDD.parquet
B. \Depdftment\DataSource\YYY\MM\DataFile_YYYYMMDD.parquet
C. \DD\MM\YYYY\Department\DataSource\DataFile_DDMMYY.parquet
D. \DataSource\Department\YYYYMM\DataFile_YYYYMMDD.parquet

Answer: B

Explanation:
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.

NEW QUESTION 187
- (Exam Topic 3)
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.


SELECT [▼] as HighestScore
- Collect(Score)
- CollectTop(1) OVER(ORDER BY Score Desc)
- Game, MAX(Score)
- TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)

FROM input TIMESTAMP BY CreatedAt

GROUP BY [▼]
- Game
- Hopping(minute,5)
- Tumbling(minute,5)
- Windows(TumblingWindow(minute,5),Hopping(minute,5))

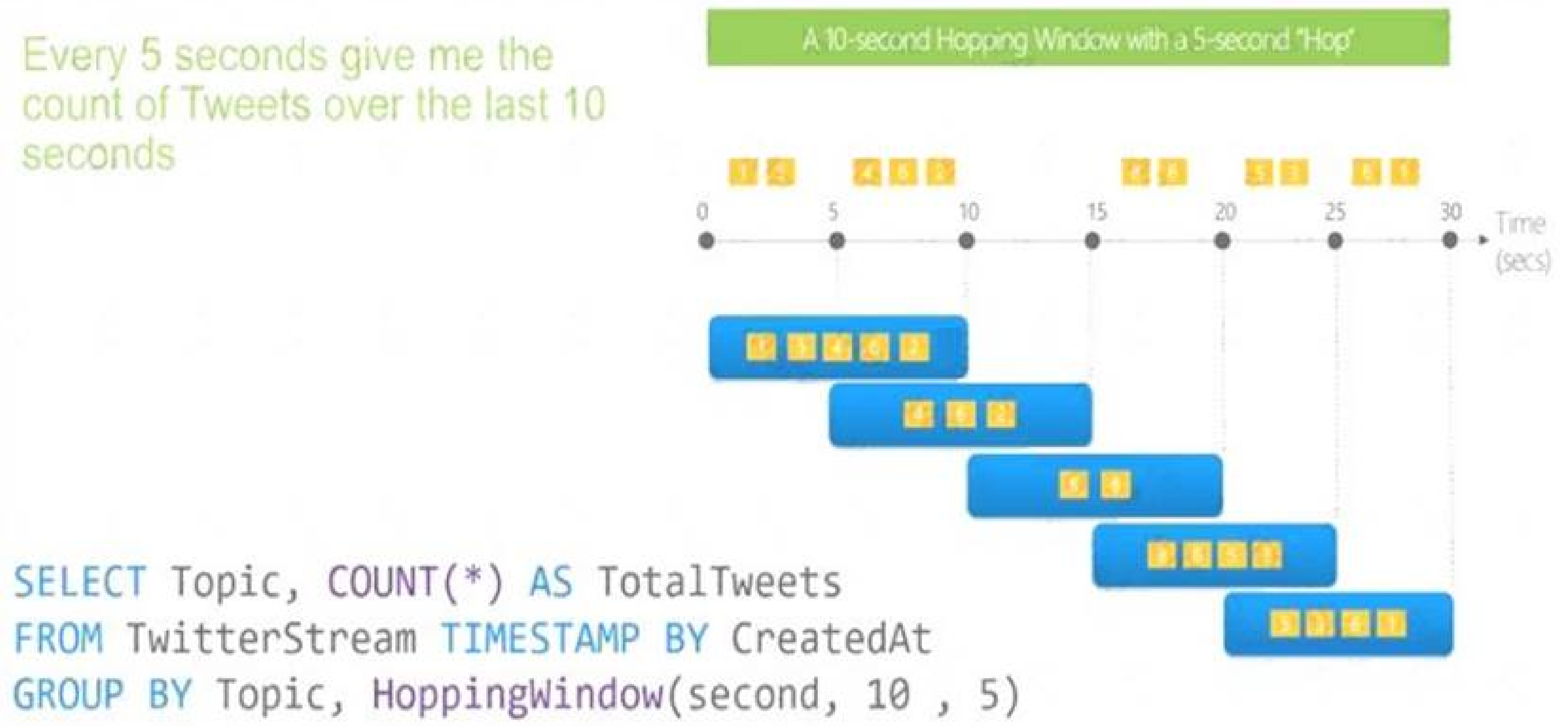
A. Mastered
B. Not Mastered

Answer: A

Explanation:
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.

A picture containing timeline Description automatically generated


Every 5 seconds give me the count of Tweets over the last 10 seconds
A 10-second Hopping Window with a 5-second "Hop"
0
5
10
15
20
25
30
Time (secs)

```
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

NEW QUESTION 190
- (Exam Topic 3)
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 into Table1 and azure Data Lake Storage Gen2 container named container1.
You plan to insert data from the files 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

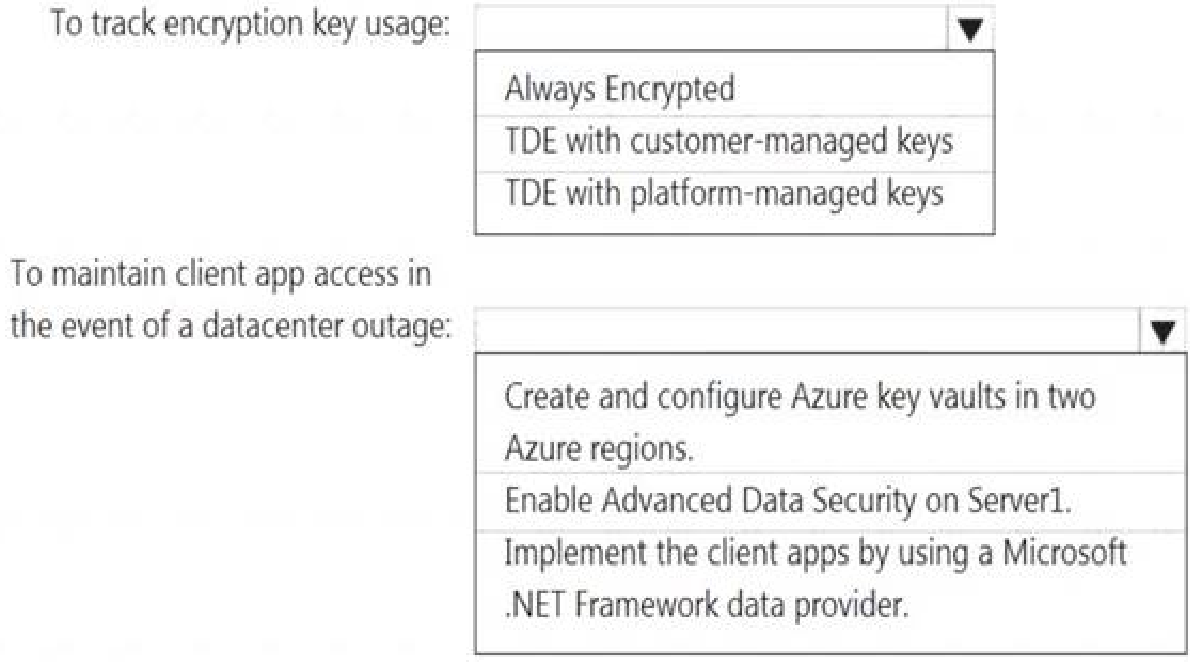
Answer: B

Explanation:
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

NEW QUESTION 192
- (Exam Topic 3)
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.


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.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: TDE with customer-managed keys
Customer-managed keys are stored in the Azure Key Vault. You can monitor how and when your key vaults are accessed, and by whom. You can do this by enabling logging for Azure Key Vault, which saves information in an Azure storage account that you provide.
Box 2: Create and configure Azure key vaults in two Azure regions
The contents of your key vault are replicated within the region and to a secondary region at least 150 miles away, but within the same geography to maintain high durability of your keys and secrets.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption https://docs.microsoft.com/en-us/azure/key-vault/general/logging

NEW QUESTION 194
- (Exam Topic 3)
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?

A. Yes
B. No

Answer: B

Explanation:
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

NEW QUESTION 199
- (Exam Topic 3)
You are designing a data mart for the human resources (MR) department at your company. The data mart will contain information and employee transactions.
From a source system you have a flat extract that has the following fields:
• EmployeeID
• FirstName
• LastName
• Recipient
• GrossArnount
• TransactionID
• GovernmentID
• NetAmountPaid
• TransactionDate
You need to design a start schema data model in an Azure Synapse analytics dedicated SQL pool for the data mart.
Which two tables should you create? Each Correct answer present part of the solution.

A. a dimension table for employee
B. a fabric for Employee
C. a dimension table far EmployeeTransaction
D. a dimension table for Transaction
E. a fact table for Transaction

Answer: AE

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

NEW QUESTION 202
- (Exam Topic 3)
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 Azure SQL data warehouse. You need to prepare the files to ensure that the data copies quickly.
Solution: You modify the files to ensure that each row is more than 1 MB. Does this meet the goal?

A. Yes
B. No

Answer: A

Explanation:
Instead modify the files to ensure that each row is less than 1 MB. References:

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

NEW QUESTION 207
- (Exam Topic 3)
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.

Answer: BD

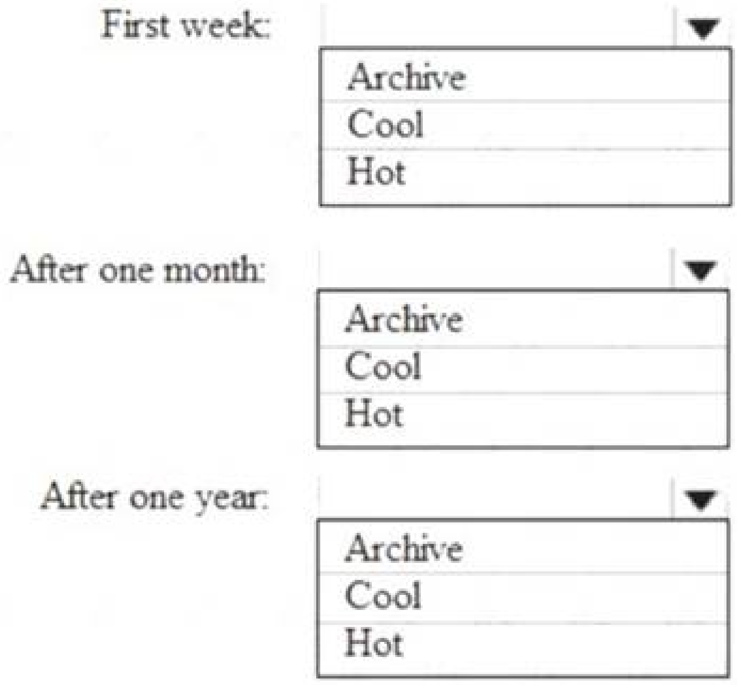
NEW QUESTION 208
- (Exam Topic 3)
You plan to implement an Azure Data Lake Gen2 storage account.
You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs.
Which type of replication should you use for the storage account?

A. geo-redundant storage (GRS)
B. zone-redundant storage (ZRS)
C. locally-redundant storage (LRS)
D. geo-zone-redundant storage (GZRS)

Answer: C

Explanation:
Locally redundant storage (LRS) copies your data synchronously three times within a single physical location in the primary region. LRS is the least expensive replication option Reference:
https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy

NEW QUESTION 210
- (Exam Topic 3)
You are designing an application that will store petabytes of medical imaging data
When the data is first created, the data will be accessed frequently during the first week. After one month, the data must be accessible within 30 seconds, but files will be accessed infrequently. After one year, the data will be accessed infrequently but must be accessible within five minutes.
You need to select a storage strategy for the data. The solution must minimize costs.
Which storage tier should you use for each time frame? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
First week: Hot
Hot - Optimized for storing data that is accessed frequently. After one month: Cool
Cool - Optimized for storing data that is infrequently accessed and stored for at least 30 days.
After one year: Cool

NEW QUESTION 214

- (Exam Topic 3)
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

Answer: A

Explanation:
SQL Database dynamic data masking limits sensitive data exposure by masking it to non-privileged users. The Credit card masking method exposes the last four digits of the designated fields and adds a constant string as a prefix in the form of a credit card.
Example: XXXX-XXXX-XXXX-1234
Reference:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-dynamic-data-masking-get-started

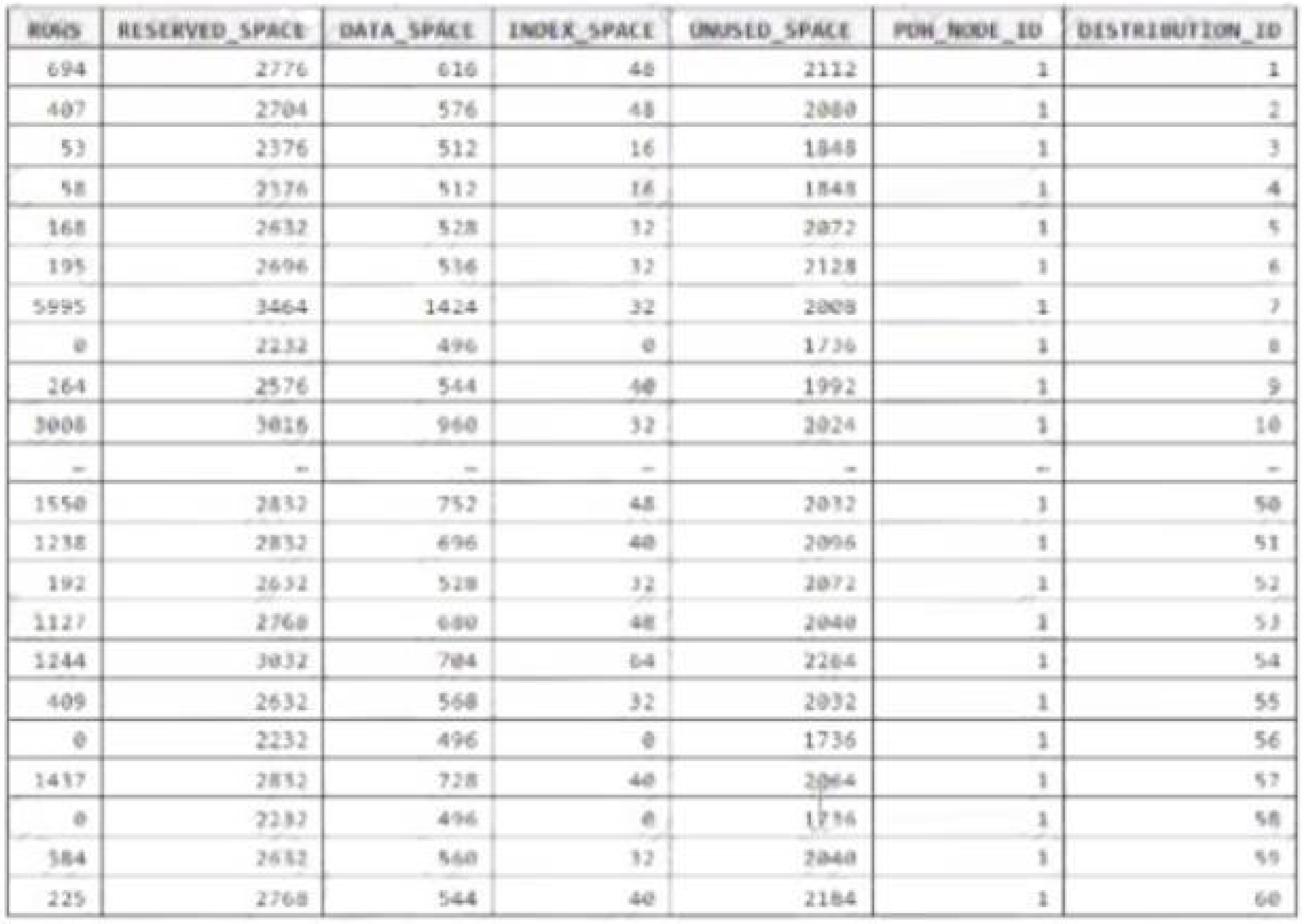
NEW QUESTION 215
- (Exam Topic 3)
You have an Azure Synapse Analytics job that uses Scala. You need to view the status of the job.
What should you do?

A. From Azure Monitor, run a Kusto query against the AzureDiagnostics table.
B. From Azure Monitor, run a Kusto query against the SparkLogying1 Event.CL table.
C. From Synapse Studio, select the workspac
D. From Monitor, select Apache Sparks applications.
E. From Synapse Studio, select the workspac
F. From Monitor, select SQL requests.

Answer: C

Explanation:
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

NEW QUESTION 217
- (Exam Topic 3)
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	784	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

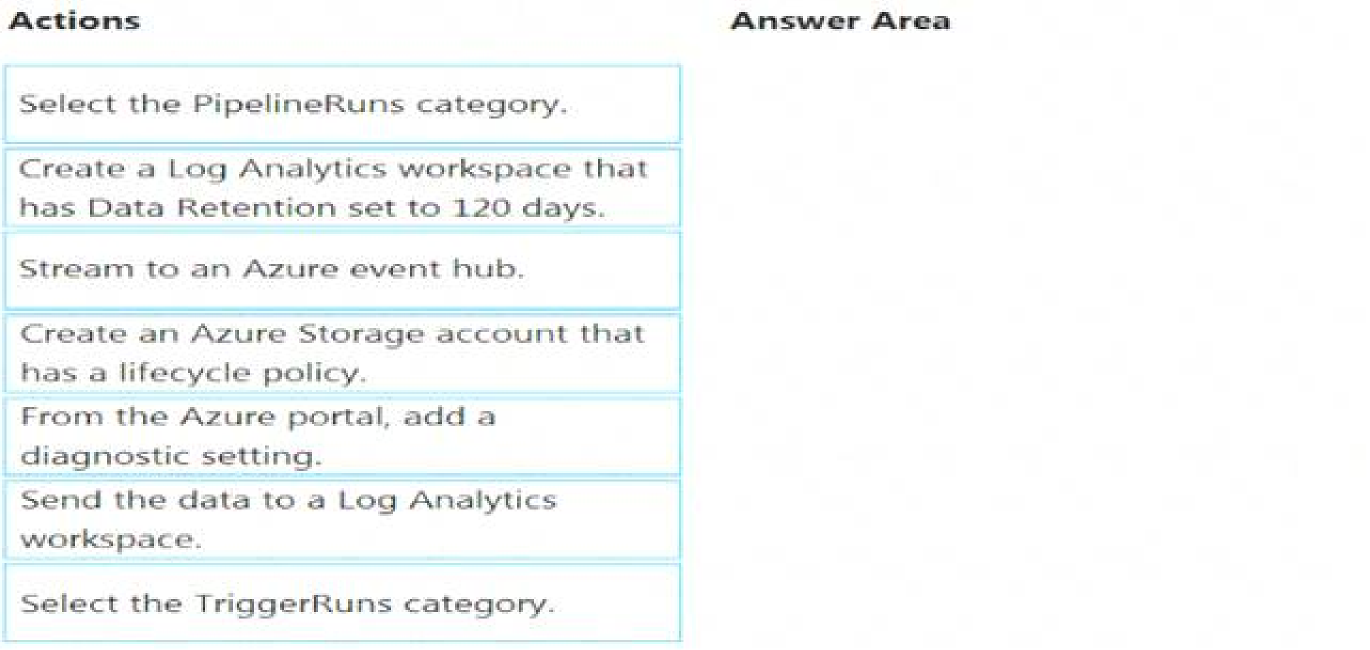
Which statement accurately describes the dbo,FactInternetSales table?

A. The table contains less than 1,000 rows.
B. All distributions contain data.
C. The table is skewed.
D. The table uses round-robin distribution.

Answer: B

Explanation:
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-distribu

NEW QUESTION 221
- (Exam Topic 3)
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.



Actions

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.

Answer Area

A. Mastered
B. Not Mastered

Answer: A

Explanation:
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.

- In the portal, go to Monitor. Select Settings > Diagnostic settings.
- Select the data factory for which you want to set a diagnostic setting.
- If no settings exist on the selected data factory, you're prompted to create a setting. Select Turn on diagnostics.
- Give your setting a name, select Send to Log Analytics, and then select a workspace from Log Analytics Workspace.
- Select Save. Reference:

https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor

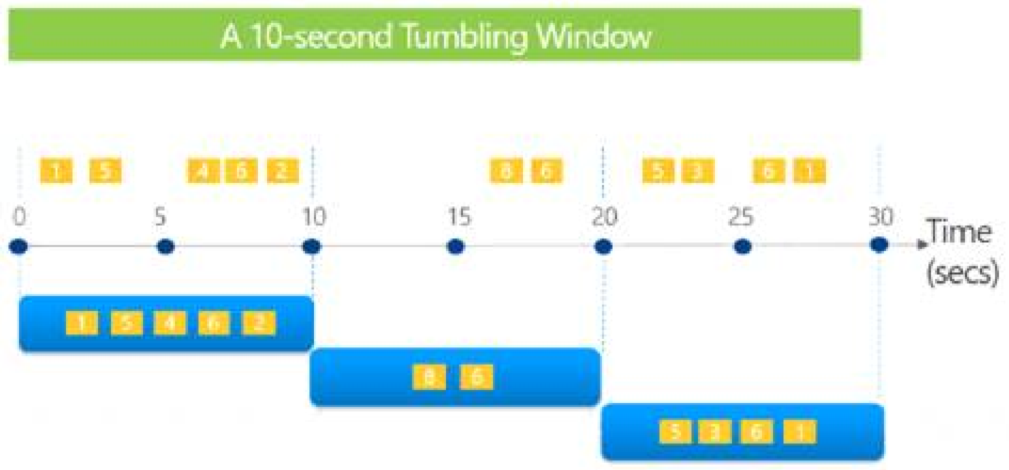
NEW QUESTION 225
- (Exam Topic 3)
You are designing an Azure Stream Analytics job to process incoming events from sensors in retail environments.
You need to process the events to produce a running average of shopper counts during the previous 15 minutes, calculated at five-minute intervals.
Which type of window should you use?

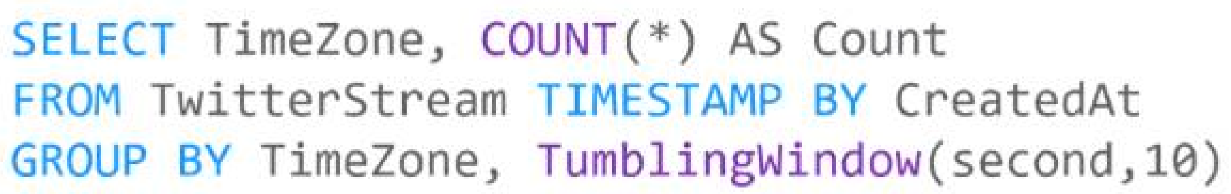
A. snapshot
B. tumbling
C. hopping
D. sliding

Answer: B

Explanation:
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

NEW QUESTION 228
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL Pool1. Pool1 contains a partitioned fact table named dbo.Sales and a staging table named stg.Sales that has the matching table and partition definitions.
You need to overwrite the content of the first partition in dbo.Sales with the content of the same partition in stg.Sales. The solution must minimize load times.
What should you do?

A. Switch the first partition from dbo.Sales to stg.Sales.
B. Switch the first partition from stg.Sales to db
C. Sales.
D. Update dbo.Sales from stg.Sales.
E. Insert the data from stg.Sales into dbo.Sales.

Answer: A

NEW QUESTION 230
- (Exam Topic 3)
You are planning a solution to aggregate streaming data that originates in Apache Kafka and is output to Azure Data Lake Storage Gen2. The developers who will implement the stream processing solution use Java,
Which service should you recommend using to process the streaming data?

A. Azure Data Factory
B. Azure Stream Analytics
C. Azure Databricks
D. Azure Event Hubs

Answer: C

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

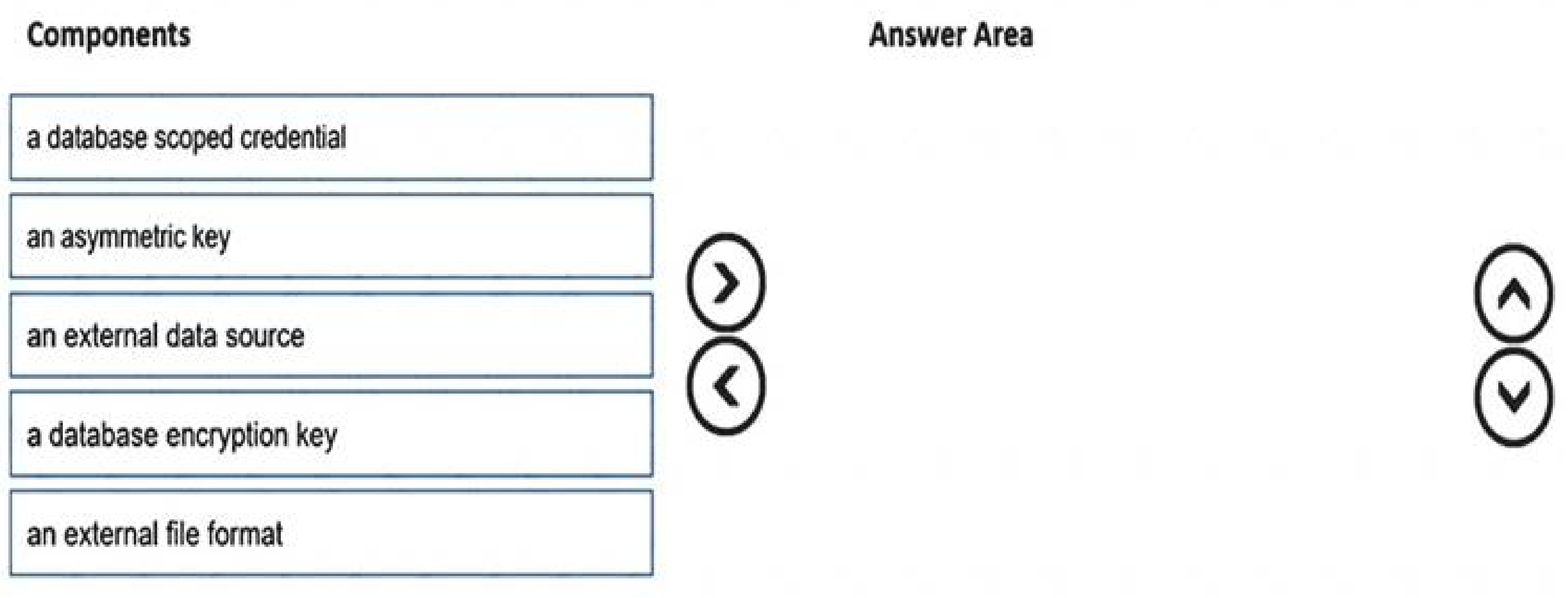
NEW QUESTION 234
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Data Lake Storage account named myaccount1. The myaccount1 account contains two containers named container1 and contained. The subscription is linked to an Azure Active Directory (Azure AD) tenant that contains a security group named Group1.
You need to grant Group1 read access to contamer1. The solution must use the principle of least privilege. Which role should you assign to Group1?

A. Storage Blob Data Reader for container1
B. Storage Table Data Reader for container1
C. Storage Blob Data Reader for myaccount1
D. Storage Table Data Reader for myaccount1

Answer: A

NEW QUESTION 237
- (Exam Topic 3)
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.



Components

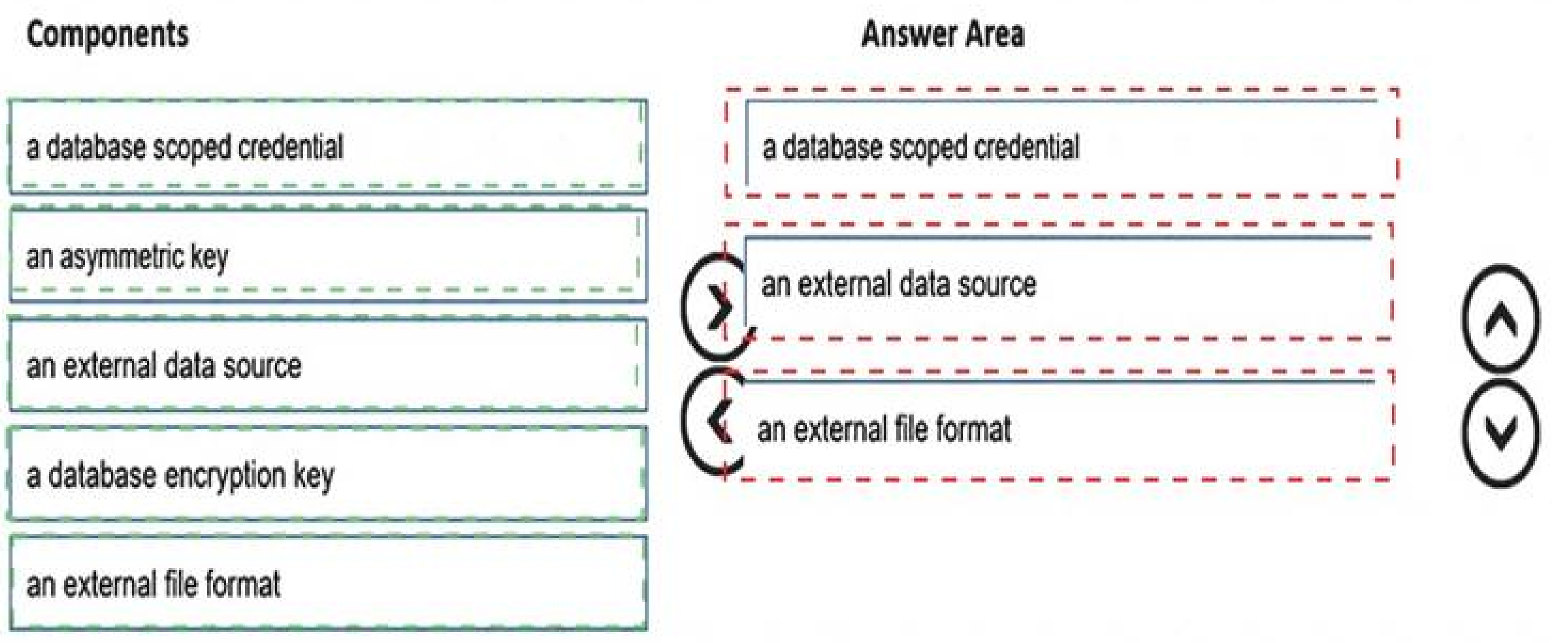
- a database scoped credential
- an asymmetric key
- an external data source
- a database encryption key
- an external file format

Answer Area

A. Mastered
B. Not Mastered

Answer: A

Explanation:



Components

- a database scoped credential
- an asymmetric key
- an external data source
- a database encryption key
- an external file format

Answer Area

- a database scoped credential
- an external data source
- an external file format

NEW QUESTION 239
- (Exam Topic 3)
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 SOL.
- 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

Answer: B

Explanation:
Need a High Concurrency cluster for the jobs.
Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.
A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.
Reference: https://docs.azuredatabricks.net/clusters/configure.html

NEW QUESTION 242

......