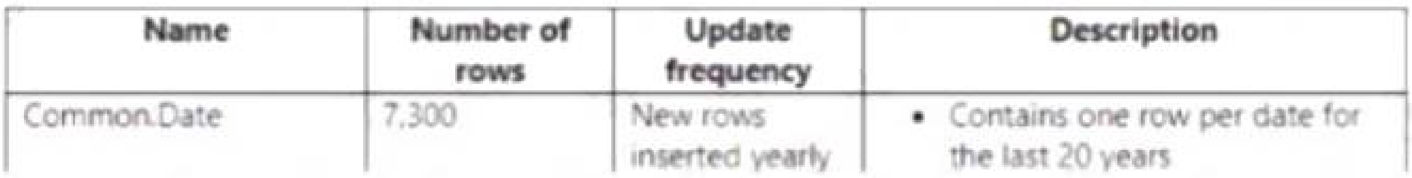
Exam Questions DP-203

Data Engineering on Microsoft Azure

https://www.2passeasy.com/dumps/DP-203/

NEW QUESTION 1
- (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

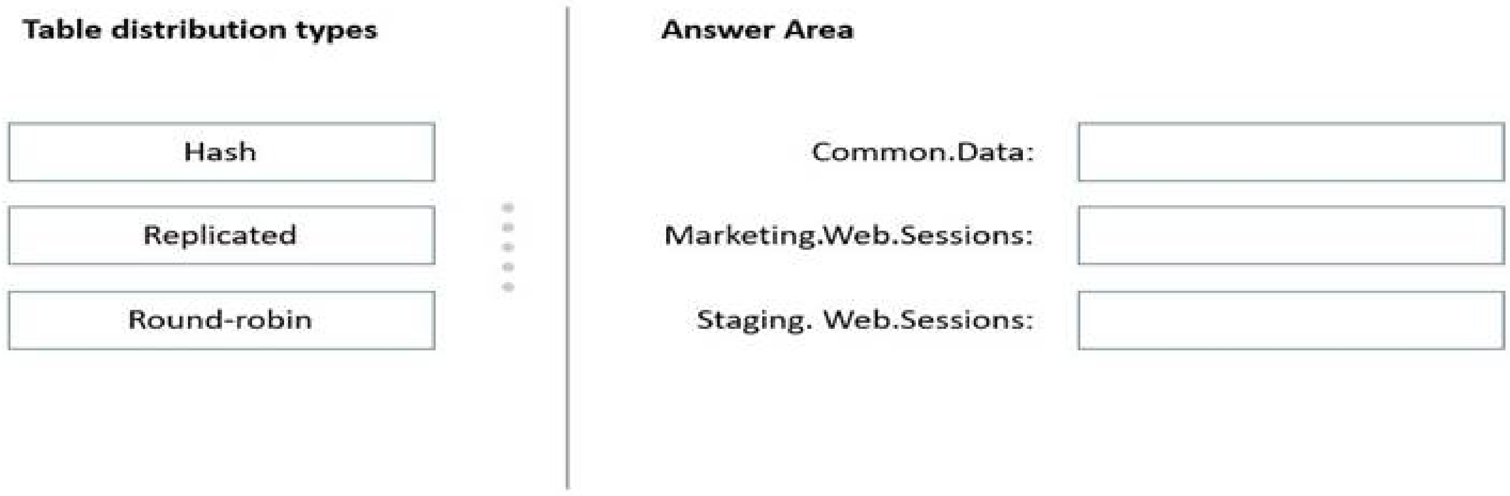


Table distribution types

Answer Area

A. Mastered
B. Not Mastered

Answer: A

Explanation:

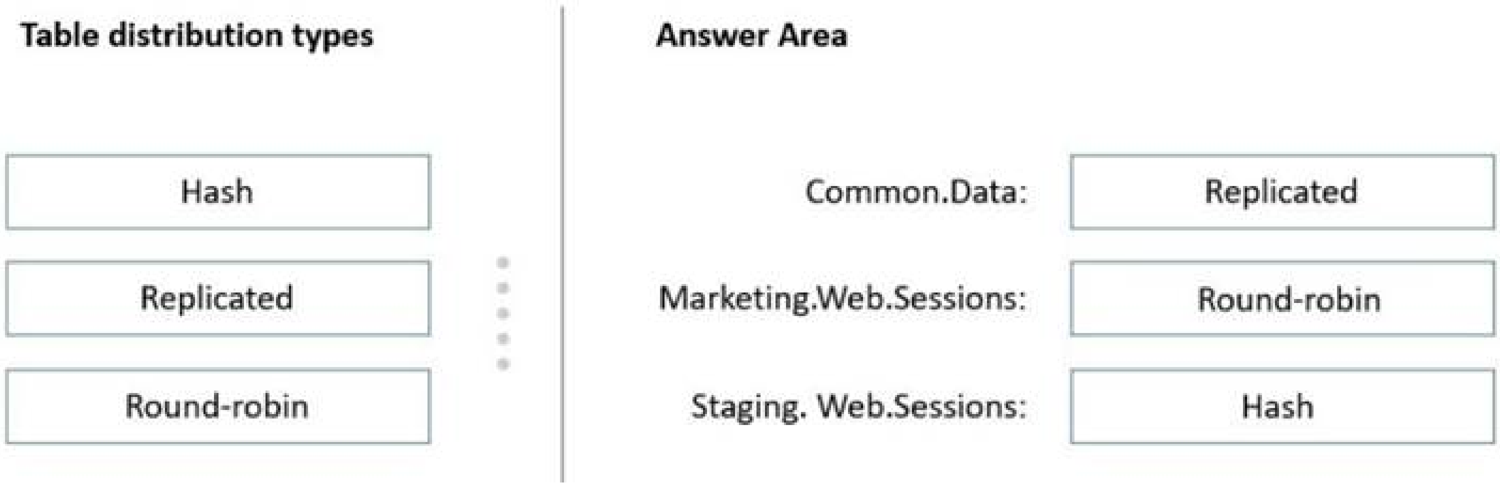


Table distribution types

Answer Area

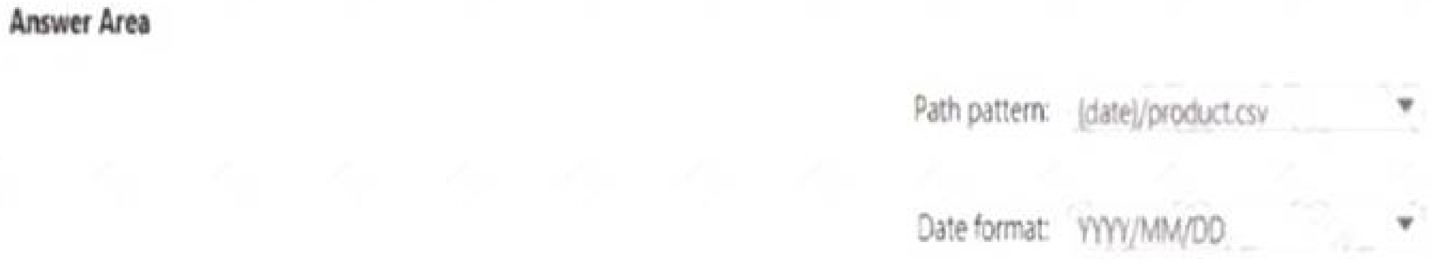
Common.Data: Replicated
Marketing.Web.Sessions: Round-robin
Staging. Web.Sessions: Hash

NEW QUESTION 2
- (Exam Topic 3)
The storage account container view is shown in the Refdata exhibit. (Click the Refdata tab.) You need to configure the Stream Analytics job to pick up the new reference data. What should you configure? To answer, select the appropriate options in the answer area NOTE: Each correct selection is worth one point.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Answer as below



Answer Area

Path pattern: {date}/product.csv
Date format: YYYY/MM/DD

NEW QUESTION 3
- (Exam Topic 3)
A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU).
You need to optimize performance for the Azure Stream Analytics job.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Implement event ordering.
B. Implement Azure Stream Analytics user-defined functions (UDF).
C. Implement query parallelization by partitioning the data output.
D. Scale the SU count for the job up.
E. Scale the SU count for the job down.
F. Implement query parallelization by partitioning the data input.

Answer: DF

Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization

NEW QUESTION 4
- (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 5
- (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 6
- (Exam Topic 3)
You plan to create an Azure Data Factory pipeline that will include a mapping data flow. You have JSON data containing objects that have nested arrays.
You need to transform the JSON-formatted data into a tabular dataset. The dataset must have one tow for each item in the arrays.
Which transformation method should you use in the mapping data flow?

A. unpivot
B. flatten
C. new branch
D. alter row

Answer: B

Explanation:
Use the flatten transformation to take array values inside hierarchical structures such as JSON and unroll them into individual rows. This process is known as denormalization.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-flatten

NEW QUESTION 7
- (Exam Topic 3)
You manage an enterprise data warehouse in Azure Synapse Analytics.
Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries.
You need to monitor resource utilization to determine the source of the performance issues. Which metric should you monitor?

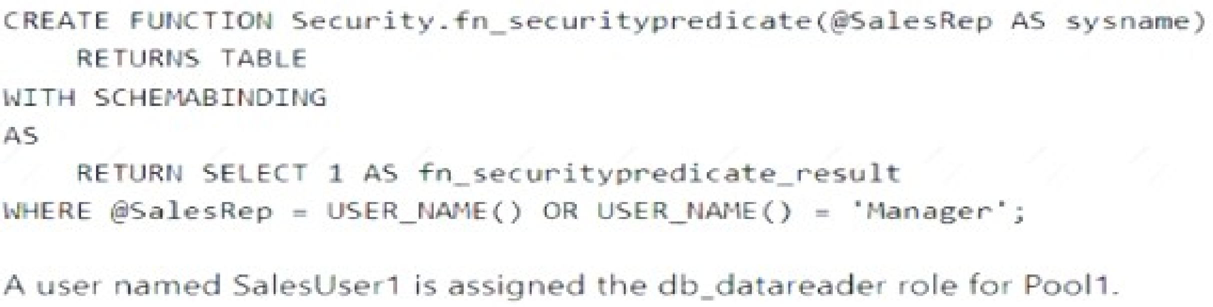
A. Data IO percentage
B. Local tempdb percentage

C. Cache used percentage
D. DWU percentage

Answer: C

Explanation:
Monitor and troubleshoot slow query performance by determining whether your workload is optimally leveraging the adaptive cache for dedicated SQL pools.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monit

NEW QUESTION 8
- (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: A

NEW QUESTION 9
- (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 10
- (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.

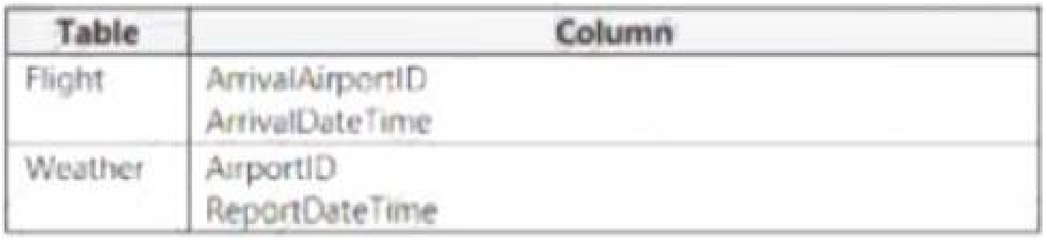


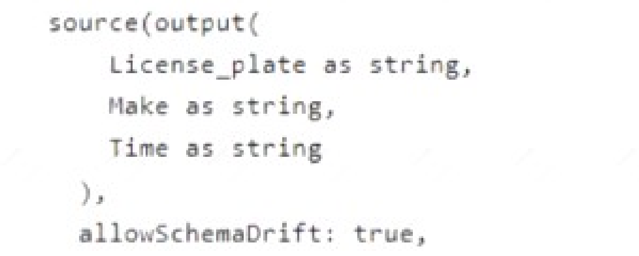
Table	Column
Flight	ArrivalAirportID ArrivalDateTime
Weather	AirportID ReportDateTime

You need to recommend a solution that maximum query performance. What should you include in the recommendation?

A. In each table, create a column as a composite of the other two columns in the table.
B. In each table, create an IDENTITY column.
C. In the tables, use a hash distribution of ArriveDateTime and ReportDateTime.
D. In the tables, use a hash distribution of ArriveAirPortID and AirportID.

Answer: D

NEW QUESTION 10
- (Exam Topic 3)
You have an Azure Data Factory pipeline that contains a data flow. The data flow contains the following expression.

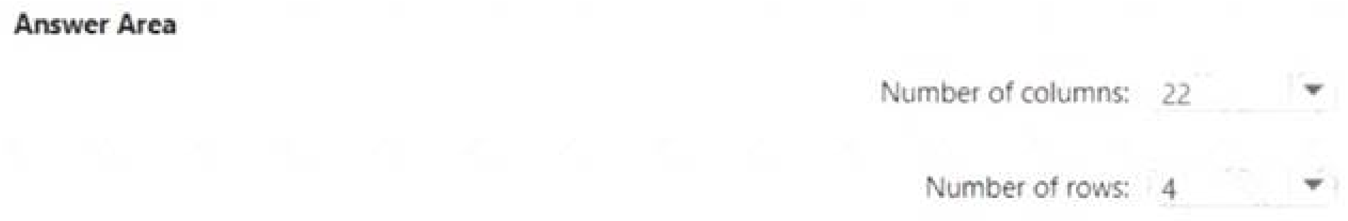


```
source(output(
    License_plate as string,
    Make as string,
    Time as string
  ),
  allowSchemaDrift: true,
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
See below answer.

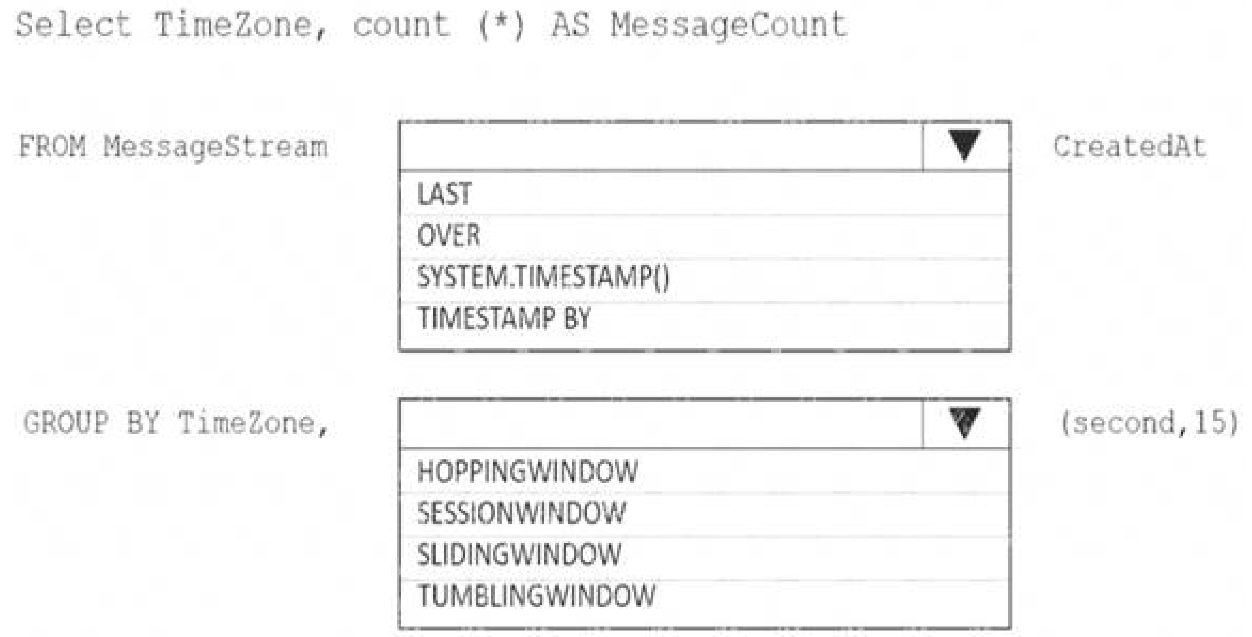


Answer Area

Number of columns: 22

Number of rows: 4

NEW QUESTION 11
- (Exam Topic 3)
You are designing an Azure Stream Analytics solution that receives instant messaging data from an Azure Event Hub.
You need to ensure that the output from the Stream Analytics job counts the number of messages per time zone every 15 seconds.
How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



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

FROM MessageStream  [            ▼]  CreatedAt
                    LAST
                    OVER
                    SYSTEM.TIMESTAMP()
                    TIMESTAMP BY

GROUP BY TimeZone,  [            ▼]  (second,15)
                    HOPPINGWINDOW
                    SESSIONWINDOW
                    SLIDINGWINDOW
                    TUMBLINGWINDOW
```

A. Mastered
B. Not Mastered

Answer: A

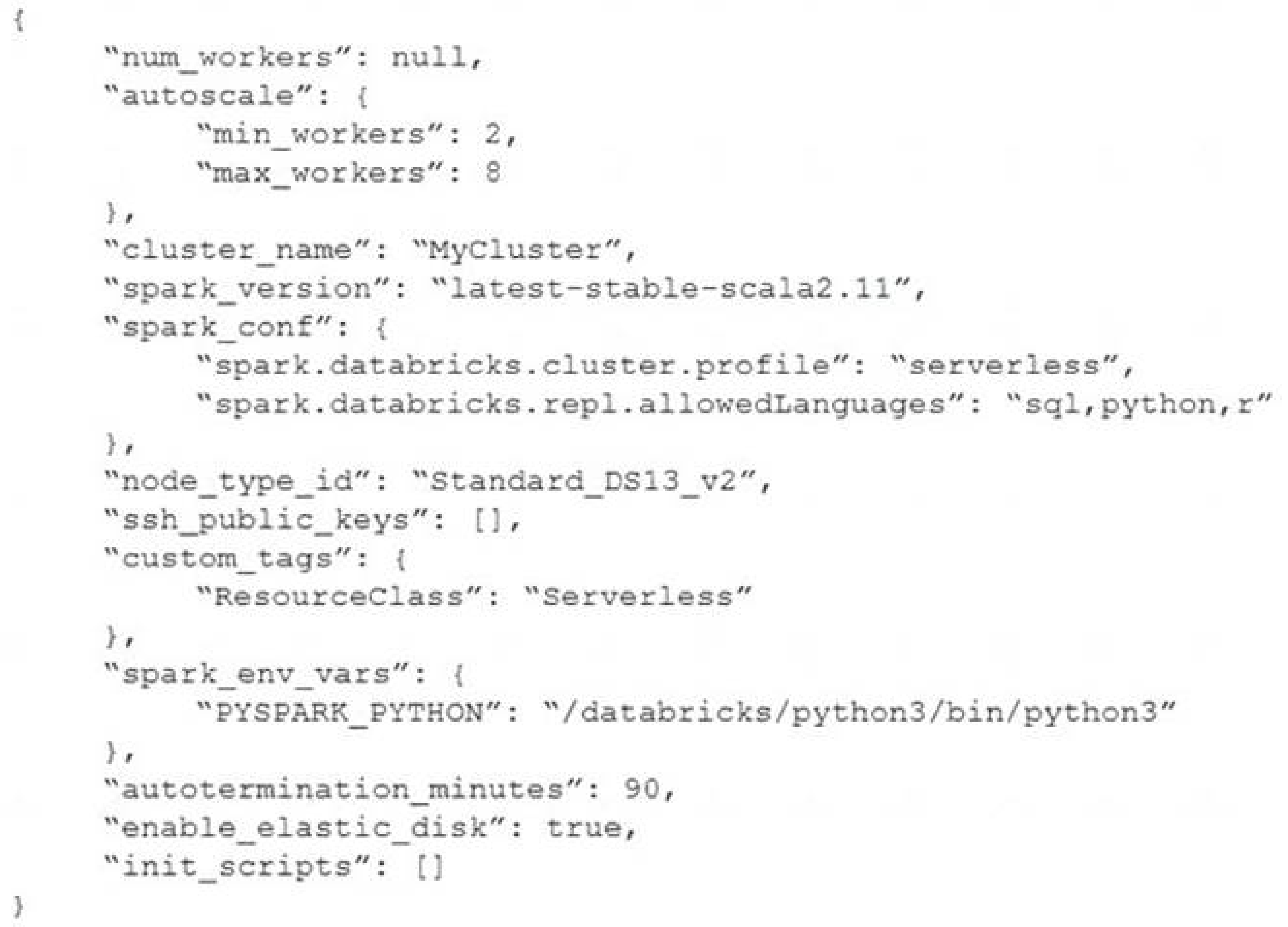
Explanation:
Table Description automatically generated
Box 1: timestamp by
Box 2: TUMBLINGWINDOW
Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.
Timeline Description automatically generated
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

NEW QUESTION 16
- (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.

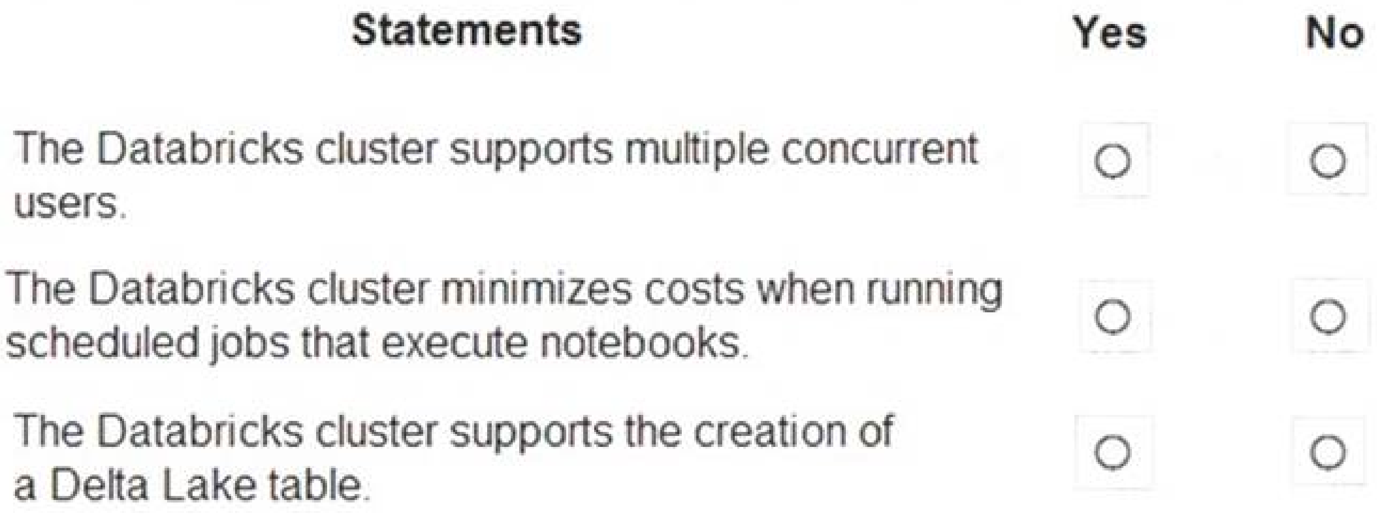
Answer: B

NEW QUESTION 21
- (Exam Topic 3)
The following code segment is used to create an Azure Databricks cluster.



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

For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.



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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated
Box 1: Yes
A cluster mode of ‘High Concurrency’ is selected, unlike all the others which are ‘Standard’. This results in a worker type of Standard_DS13_v2.
Box 2: No
When you run a job on a new cluster, the job is treated as a data engineering (job) workload subject to the job workload pricing. When you run a job on an existing cluster, the job is treated as a data analytics (all-purpose) workload subject to all-purpose workload pricing.
Box 3: Yes
Delta Lake on Databricks allows you to configure Delta Lake based on your workload patterns. Reference:
https://adatis.co.uk/databricks-cluster-sizing/ https://docs.microsoft.com/en-us/azure/databricks/jobs
https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html https://docs.databricks.com/delta/index.html

NEW QUESTION 23
- (Exam Topic 3)
You need to design a solution that will process streaming data from an Azure Event Hub and output the data to Azure Data Lake Storage. The solution must ensure that analysts can interactively query the streaming data.
What should you use?

A. event triggers in Azure Data Factory
B. Azure Stream Analytics and Azure Synapse notebooks
C. Structured Streaming in Azure Databricks
D. Azure Queue storage and read-access geo-redundant storage (RA-GRS)

Answer: C

Explanation:
Apache Spark Structured Streaming is a fast, scalable, and fault-tolerant stream processing API. You can use it to perform analytics on your streaming data in near real-time.

With Structured Streaming, you can use SQL queries to process streaming data in the same way that you would process static data.
Azure Event Hubs is a scalable real-time data ingestion service that processes millions of data in a matter of seconds. It can receive large amounts of data from multiple sources and stream the prepared data to Azure Data Lake or Azure Blob storage.
Azure Event Hubs can be integrated with Spark Structured Streaming to perform the processing of messages in near real-time. You can query and analyze the processed data as it comes by using a Structured Streaming query and Spark SQL.
Reference:
https://k21academy.com/microsoft-azure/data-engineer/structured-streaming-with-azure-event-hubs/

NEW QUESTION 28
- (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 31
- (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.

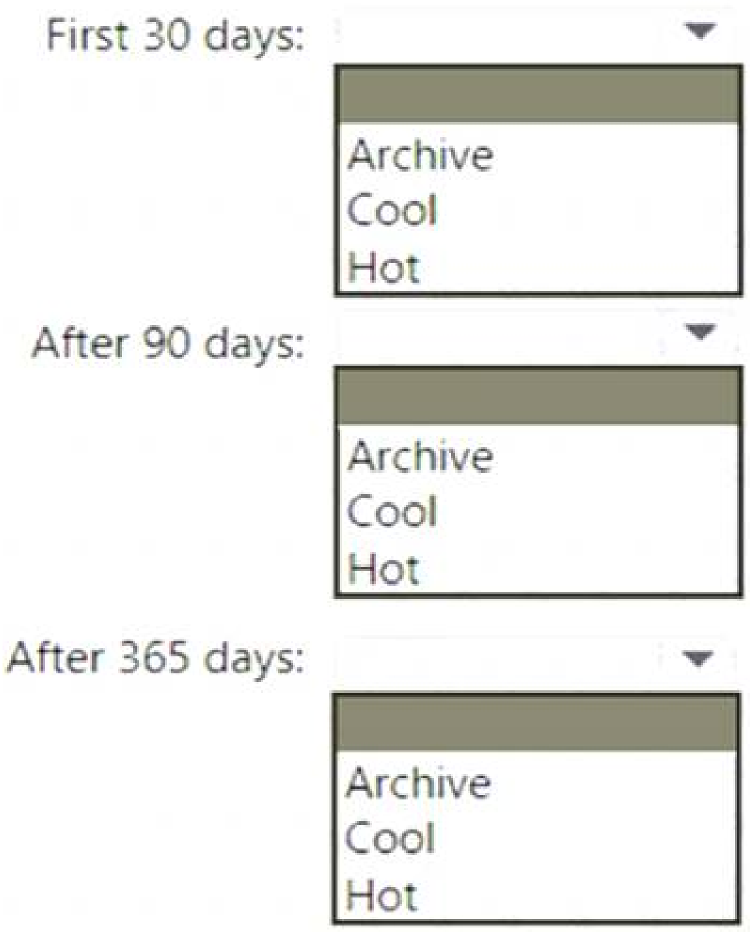
You need to create the Databricks clusters for the workloads.
Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a High Concurrency cluster for the jobs.
Does this meet the goal?

A. Yes
B. No

Answer: A

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 36
- (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

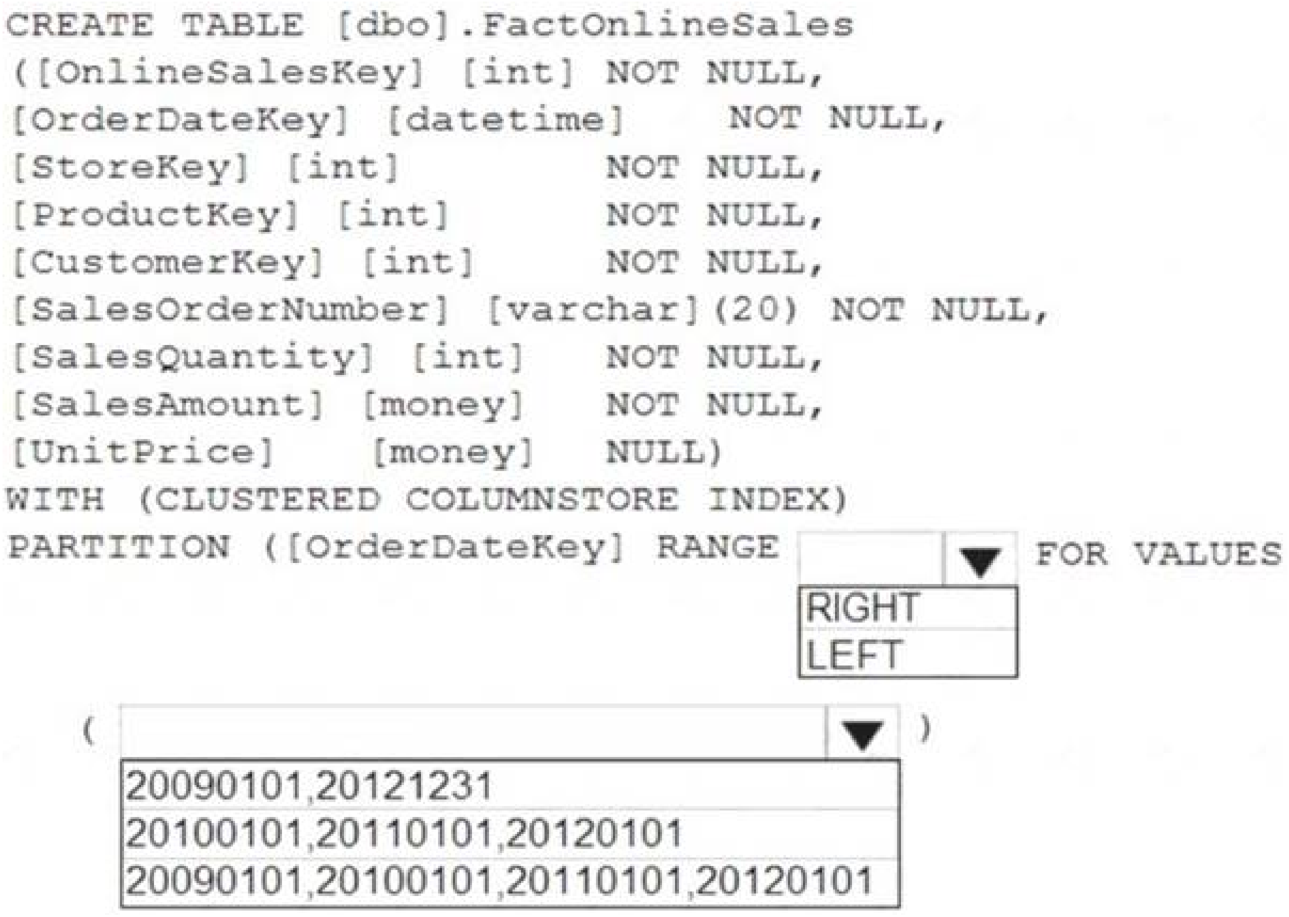
NEW QUESTION 38
- (Exam Topic 3)
You are planning a streaming data solution that will use Azure Databricks. The solution will stream sales transaction data from an online store. The solution has the following specifications:
* The output data will contain items purchased, quantity, line total sales amount, 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 39
- (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 44
- (Exam Topic 3)
You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account.
You need to output the count of tweets from the last five minutes every minute. Which windowing function should you use?

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

Answer: D

Explanation:
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.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

NEW QUESTION 46
- (Exam Topic 3)
You have an Azure Data Lake Storage Gen2 container that contains 100 TB of data.
You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region. The solution must minimize costs.
Which type of data redundancy should you use?

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

Answer: B

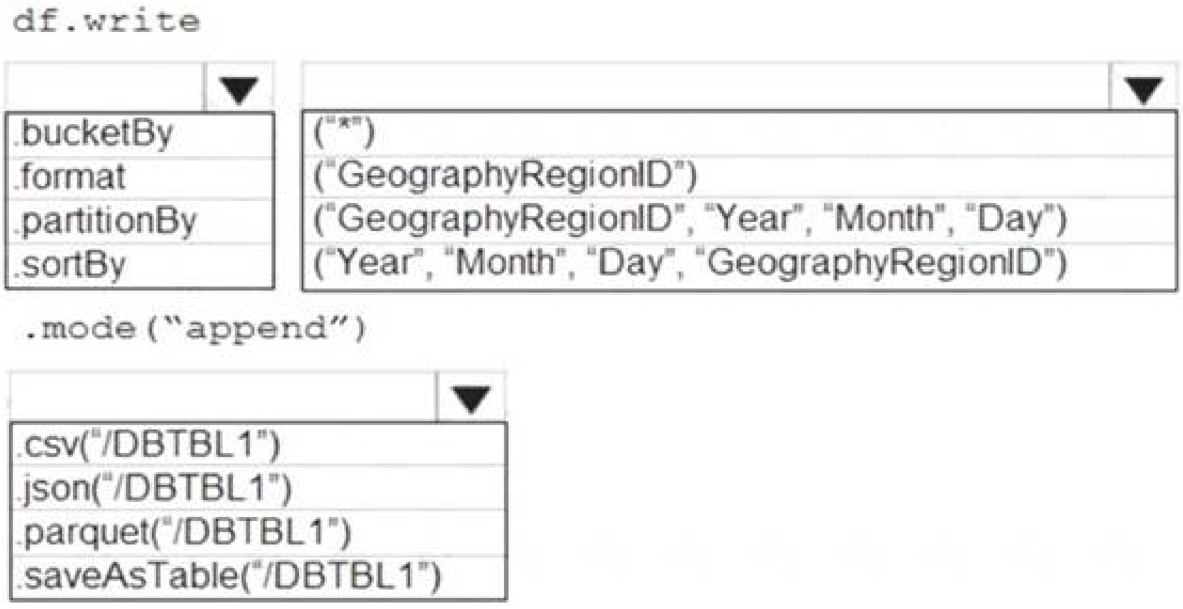
Explanation:
Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages. However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region.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

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

NEW QUESTION 50
- (Exam Topic 3)
You develop a dataset named DBTBL1 by using Azure Databricks. DBTBL1 contains the following columns:

- SensorTypeID
- GeographyRegionID
- Year
- Month
- Day
- Hour
- Minute
- Temperature
- WindSpeed
- Other

You need to store the data to support daily incremental load pipelines that vary for each GeographyRegionID. The solution must minimize storage costs.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



df.write

.bucketBy
.format
.partitionBy
.sortBy

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

.mode("append")

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated

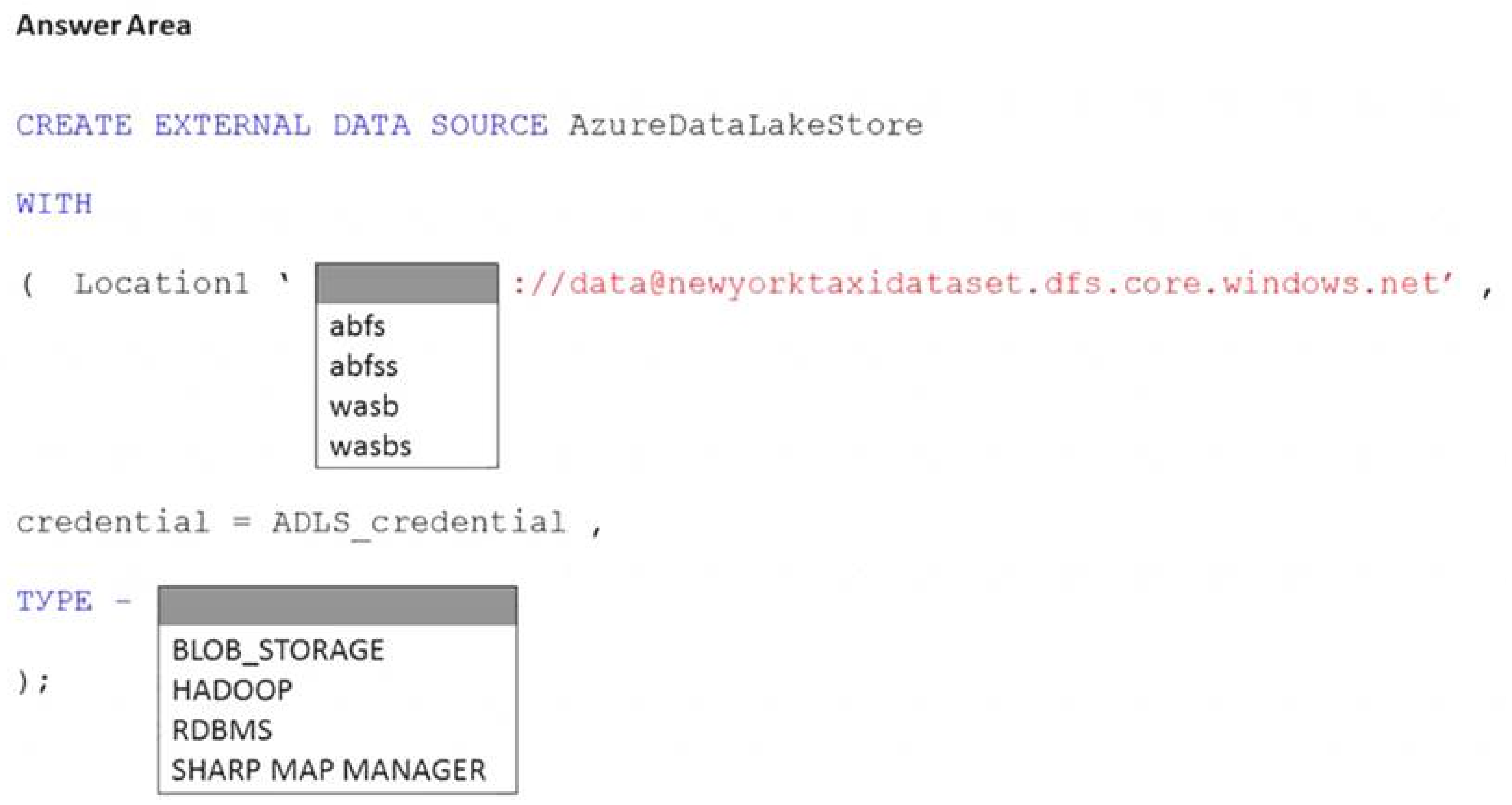
NEW QUESTION 51
- (Exam Topic 3)
You are designing a partition strategy for a fact table in an Azure Synapse Analytics dedicated SQL pool. The table has the following specifications:
• Contain sales data for 20,000 products.
• Use hash distribution on a column named ProduclID,
• Contain 2.4 billion records for the years 20l9 and 2020.
Which number of partition ranges provides optimal compression and performance of the clustered columnstore index?

A. 40
B. 240
C. 400
D. 2,400

Answer: A

Explanation:
Each partition should have around 1 millions records. Dedication SQL pools already have 60 partitions. We have the formula: Records/(Partitions*60)= 1 million
Partitions= Records/(1 million * 60)
Partitions= 2.4 x 1,000,000,000/(1,000,000 * 60) = 40
Note: Having too many partitions can reduce the effectiveness of clustered columnstore indexes if each partition has fewer than 1 million rows. Dedicated SQL pools automatically partition your data into 60 databases. So, if you create a table with 100 partitions, the result will be 6000 partitions.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

NEW QUESTION 53
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers.
You need to create an external data source in Pool1 that will be used to read .orc files in storage1. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.



Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore
WITH
(  Location1 '[ ▼ ]://data@newyorktaxidataset.dfs.core.windows.net' ,
credential = ADLS_credential ,
TYPE = [ ▼ ]
);
```

abfs
abfss
wasb
wasbs

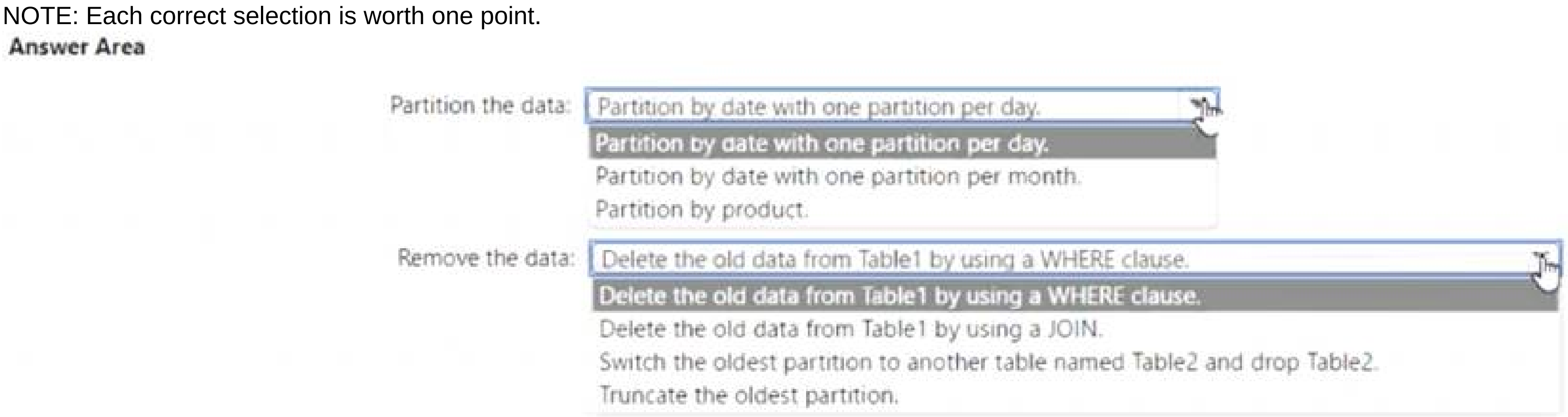
BLOB_STORAGE
HADOOP
RDBMS
SHARP MAP MANAGER

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application, email Description automatically generated
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw

NEW QUESTION 55
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Tablet. Table1 contains sales data. Sixty-five million rows of data are added to Table1 monthly.
At the end of each month, you need to remove data that is older than 36 months. The solution must minimize how long it takes to remove the data.
How should you partition Table1, and how should you remove the old data? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer Area

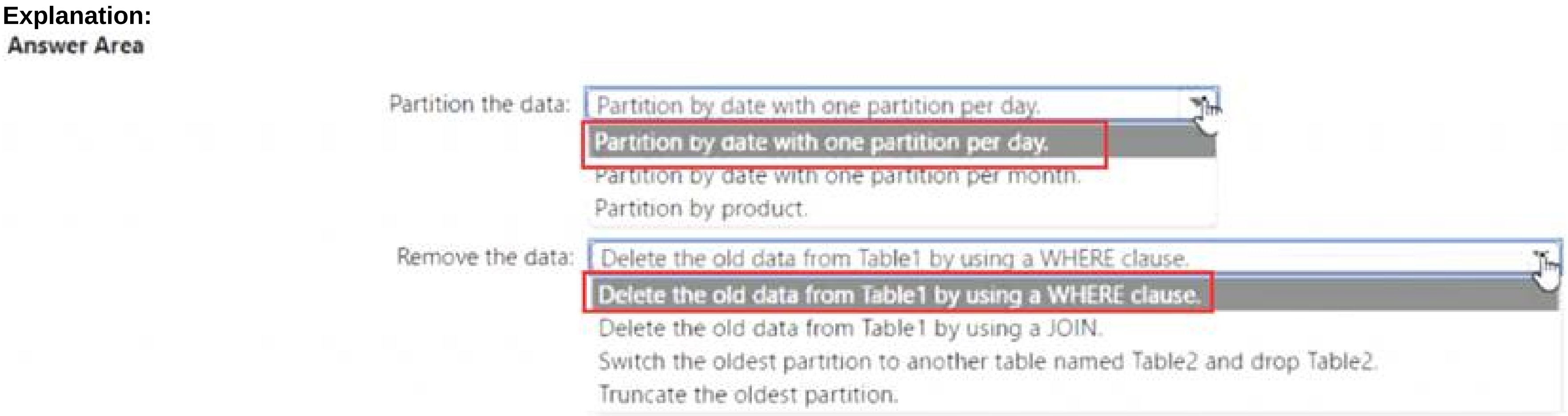
Partition the data: Partition by date with one partition per day.
- Partition by date with one partition per day.
- Partition by date with one partition per month.
- Partition by product.

Remove the data: Delete the old data from Table1 by using a WHERE clause.
- Delete the old data from Table1 by using a WHERE clause.
- Delete the old data from Table1 by using a JOIN.
- Switch the oldest partition to another table named Table2 and drop Table2.
- Truncate the oldest partition.

A. Mastered
B. Not Mastered

Answer: A

Explanation:



Answer Area

Partition the data: Partition by date with one partition per day.
- Partition by date with one partition per day.
- Partition by date with one partition per month.
- Partition by product.

Remove the data: Delete the old data from Table1 by using a WHERE clause.
- Delete the old data from Table1 by using a WHERE clause.
- Delete the old data from Table1 by using a JOIN.
- Switch the oldest partition to another table named Table2 and drop Table2.
- Truncate the oldest partition.

NEW QUESTION 58
- (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.

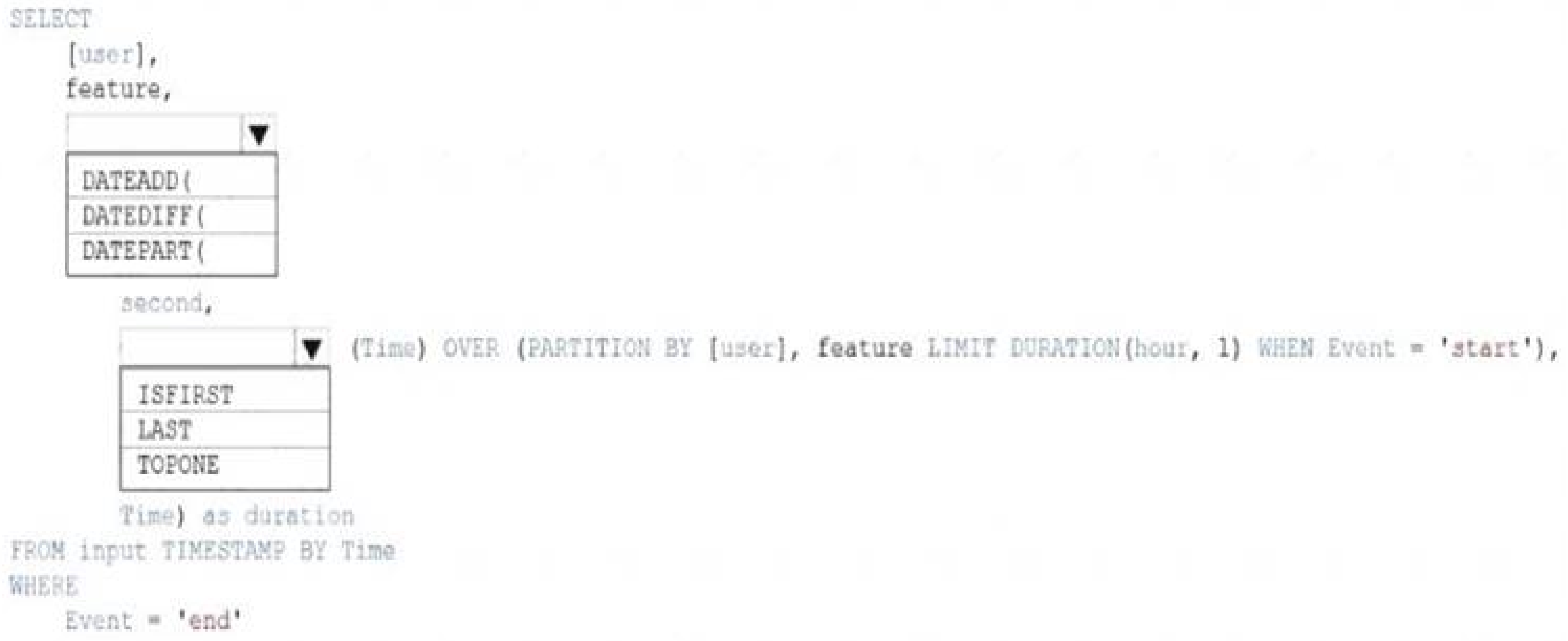
You need to create the Databricks clusters for the workloads.
Solution: You create a Standard cluster for each data scientist, a Standard cluster for the data engineers, and a High Concurrency cluster for the jobs.
Does this meet the goal?

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 62
- (Exam Topic 3)
You are building an Azure Stream Analytics job to identify how much time a user spends interacting with a feature on a webpage.
The job receives events based on user actions on the webpage. Each row of data represents an event. Each event has a type of either 'start' or 'end'.
You need to calculate the duration between start and end events.
How should you complete the query? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.



```
SELECT
    [user],
    feature,
    [            ▼]
    | DATEADD(   |
    | DATEDIFF(  |
    | DATEPART(  |
        second,
        [            ▼] (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
        | ISFIRST    |
        | LAST       |
        | TOPONE     |
        Time) as duration
FROM input TIMESTAMP BY Time
WHERE
    Event = 'end'
```

A. Mastered
B. Not Mastered

Answer: A

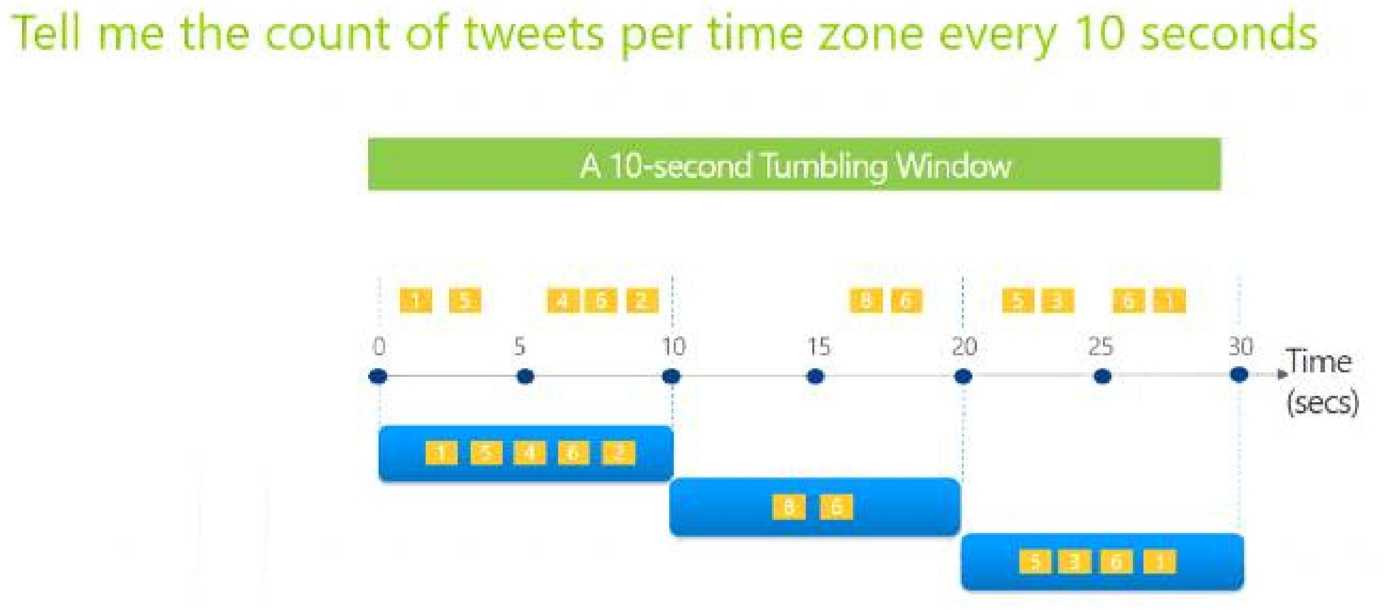
Explanation:
Box 1: DATEDIFF
DATEDIFF function returns the count (as a signed integer value) of the specified datepart boundaries crossed between the specified startdate and enddate.
Syntax: DATEDIFF (datepart , startdate, enddate) Box 2: LAST
The LAST function can be used to retrieve the last event within a specific condition. In this example, the condition is an event of type Start, partitioning the search by PARTITION BY user and feature. This way, every user and feature is treated independently when searching for the Start event. LIMIT DURATION limits the search back in time to 1 hour between the End and Start events.
Example: SELECT
[user], feature, DATEDIFF(
second,
LAST(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour,
1) WHEN Event = 'start'), Time) as duration
FROM input TIMESTAMP BY Time
WHERE
Event = 'end' Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns

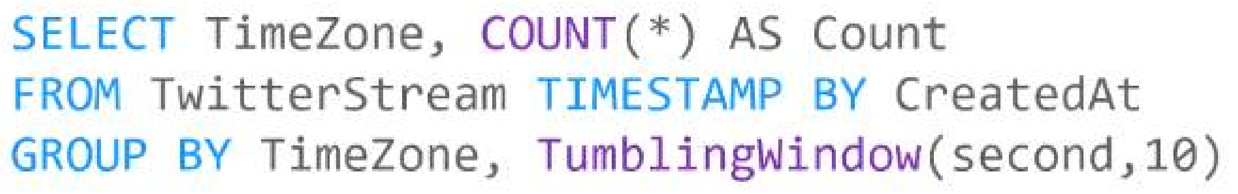
NEW QUESTION 63
- (Exam Topic 3)
You are designing an Azure Stream Analytics job to process incoming events from sensors in retail environments.
You need to process the events to produce a running average of shopper counts during the previous 15 minutes, calculated at five-minute intervals.
Which type 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
A 10-second Tumbling Window
1 5 4 6 2
8 6
5 3 6 1
0
5
10
15
20
25
30
Time (secs)
1 5 4 6 2
8 6
5 3 6 1



```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

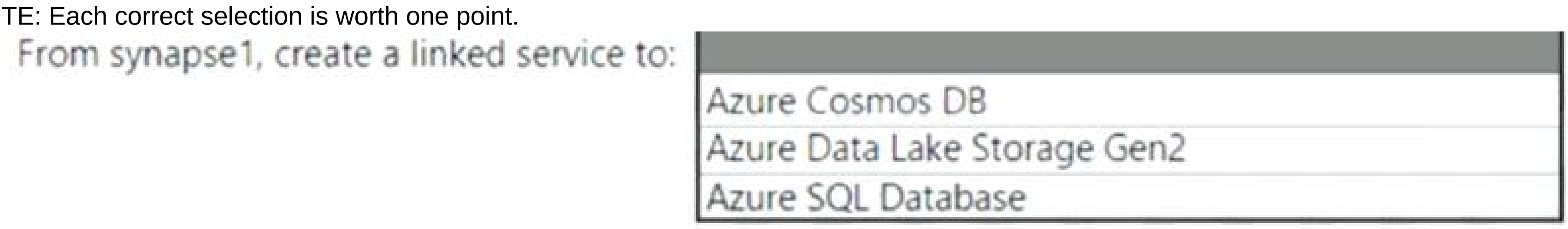
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

NEW QUESTION 67
- (Exam Topic 3)
You plan to use an Apache Spark pool in Azure Synapse Analytics to load data to an Azure Data Lake Storage Gen2 account.
You need to recommend which file format to use to store the data in the Data Lake Storage account. The solution must meet the following requirements:
• Column names and data types must be defined within the files loaded to the Data Lake Storage account.
• Data must be accessible by using queries from an Azure Synapse Analytics serverless SQL pool.
• Partition elimination must be supported without having to specify a specific partition. What should you recommend?

A. Delta Lake
B. JSON
C. CSV
D. ORC

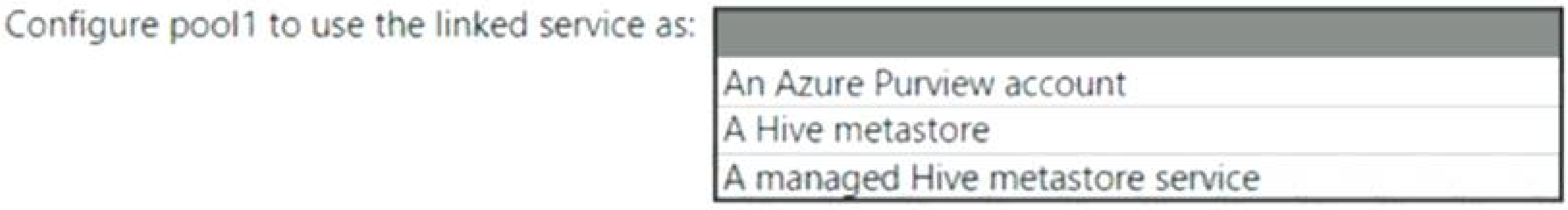
Answer: D

NEW QUESTION 72
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Databricks workspace named databricks1 and an Azure Synapse Analytics workspace named synapse1. The synapse1 workspace contains an Apache Spark pool named pool1.
You need to share an Apache Hive catalog of pool1 with databricks1.
What should you do? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



From synapse1, create a linked service to:

Azure Cosmos DB
Azure Data Lake Storage Gen2
Azure SQL Database



Configure pool1 to use the linked service as:

An Azure Purview account
A Hive metastore
A managed Hive metastore service

A. Mastered
B. Not Mastered

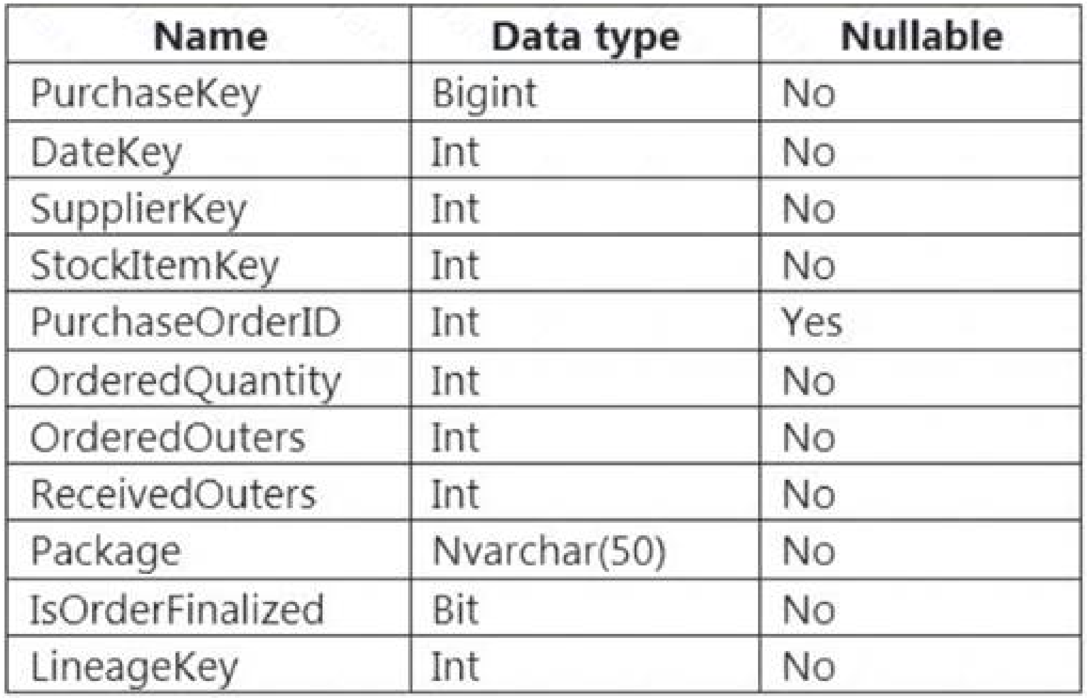
Answer: A

Explanation:
Box 1: Azure SQL Database
Use external Hive Metastore for Synapse Spark Pool
Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog.
Set up linked service to Hive Metastore
Follow below steps to set up a linked service to the external Hive Metastore in Synapse workspace.

- Set up Hive Metastore linked service
- Choose Azure SQL Database or Azure Database for MySQL based on your database type, click Continue.
- Provide Name of the linked service. Record the name of the linked service, this info will be used to configure Spark shortly.
- You can either select Azure SQL Database/Azure Database for MySQL for the external Hive Metastore from Azure subscription list, or enter the info manually.
- Provide User name and Password to set up the connection.
- Test connection to verify the username and password.
- Click Create to create the linked service.

Box 2: A Hive Metastore
nce: https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-external-metastore

NEW QUESTION 75
- (Exam Topic 3)
You are designing a fact table named FactPurchase in an Azure Synapse Analytics dedicated SQL pool. The table contains purchases from suppliers for a retail store. FactPurchase will contain the following columns.



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

FactPurchase will have 1 million rows of data added daily and will contain three years of data. Transact-SQL queries similar to the following query will be executed daily.
SELECT
SupplierKey, StockItemKey, COUNT(*) FROM FactPurchase
WHERE DateKey >= 20210101 AND DateKey <= 20210131
GROUP By SupplierKey, StockItemKey
Which table distribution will minimize query times?

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

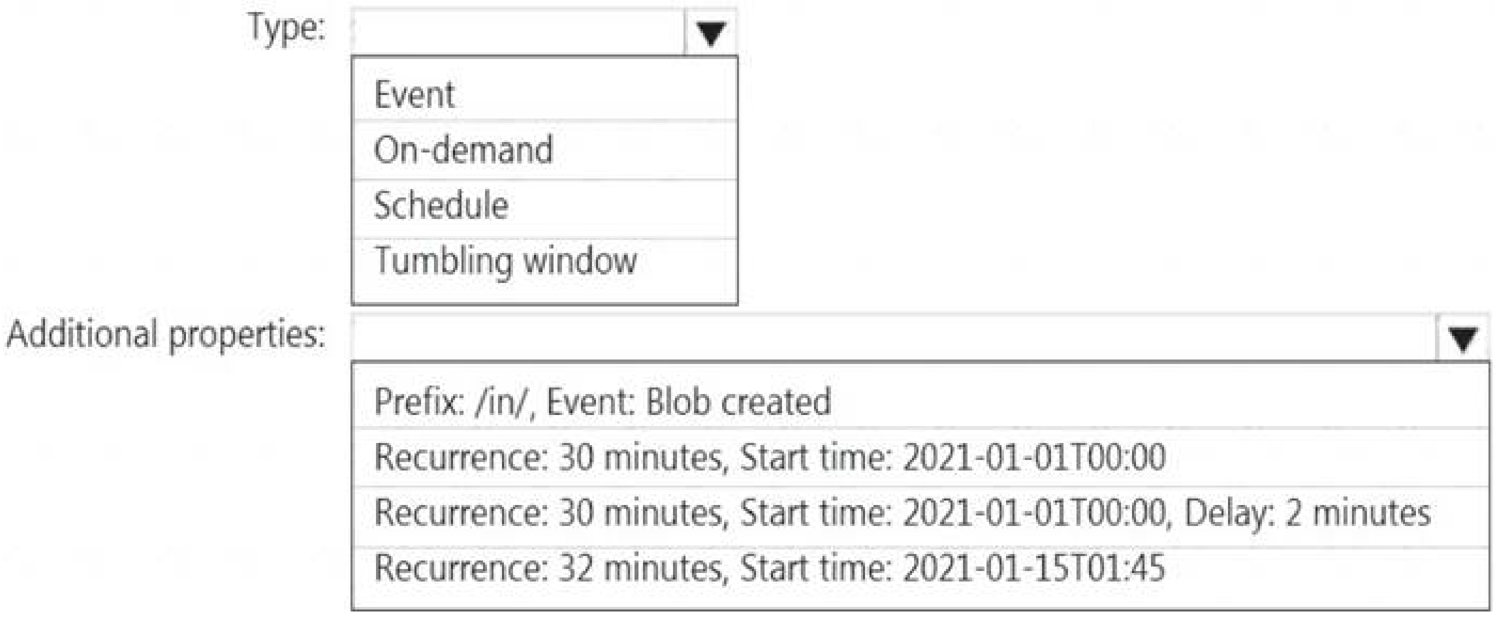
Answer: D

Explanation:
Hash-distributed tables improve query performance on large fact tables, and are the focus of this article. Round-robin tables are useful for improving loading speed.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

NEW QUESTION 79
- (Exam Topic 3)
You build an Azure Data Factory pipeline to move data from an Azure Data Lake Storage Gen2 container to a database in an Azure Synapse Analytics dedicated SQL pool.
Data in the container is stored in the following folder structure.
/in/{YYYY}/{MM}/{DD}/{HH}/{mm}
The earliest folder is /in/2021/01/01/00/00. The latest folder is /in/2021/01/15/01/45. You need to configure a pipeline trigger to meet the following requirements:

- Existing data must be loaded.
- Data must be loaded every 30 minutes.
- Late-arriving data of up to two minutes must he included in the load for the time at which the data should have arrived.

How should you configure the pipeline trigger? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.



Type:

Event

On-demand

Schedule

Tumbling window

Additional properties:

Prefix: /in/, Event: Blob created

Recurrence: 30 minutes, Start time: 2021-01-01T00:00

Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes

Recurrence: 32 minutes, Start time: 2021-01-15T01:45

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Tumbling window
To be able to use the Delay parameter we select Tumbling window. Box 2:
Recurrence: 30 minutes, not 32 minutes
Delay: 2 minutes.
The amount of time to delay the start of data processing for the window. The pipeline run is started after the expected execution time plus the amount of delay. The delay defines how long the trigger waits past the due time before triggering a new run. The delay doesn't alter the window startTime.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger

NEW QUESTION 83
- (Exam Topic 3)
You are creating a new notebook in Azure Databricks that will support R as the primary language but will also support Scale and SOL Which switch should you use to switch between languages?

A. @<Language>
B. %<Language>
C. \\(<Language>)
D. \\(<Language>)

Answer: B

Explanation:
To change the language in Databricks' cells to either Scala, SQL, Python or R, prefix the cell with '%', followed by the language.
%python //or r, scala, sql Reference:
https://www.theta.co.nz/news-blogs/tech-blog/enhancing-digital-twins-part-3-predictive-maintenance-with-azur

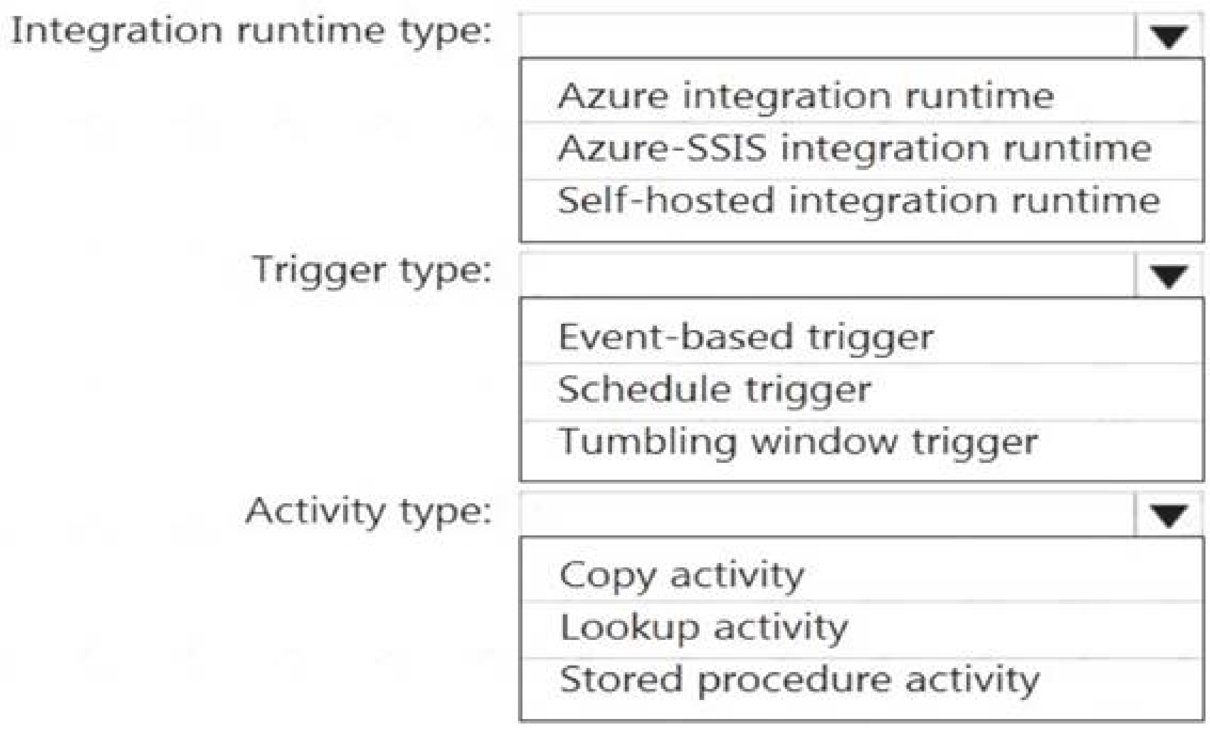
NEW QUESTION 88
- (Exam Topic 2)
What should you recommend to prevent users outside the Litware on-premises network from accessing the analytical data store?

A. a server-level virtual network rule
B. a database-level virtual network rule
C. a database-level firewall IP rule
D. a server-level firewall IP rule

Answer: A

Explanation:
Virtual network rules are one firewall security feature that controls whether the database server for your single databases and elastic pool in Azure SQL Database or for your databases in SQL Data Warehouse accepts communications that are sent from particular subnets in virtual networks.
Server-level, not database-level: Each virtual network rule applies to your whole Azure SQL Database server, not just to one particular database on the server. In other words, virtual network rule applies at the serverlevel, not at the database-level.
References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-vnet-service-endpoint-rule-overview

NEW QUESTION 93
- (Exam Topic 2)
Which Azure Data Factory components should you recommend using together to import the daily inventory data from the SQL server to Azure Data Lake Storage? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Integration runtime type:
- Azure integration runtime
- Azure-SSIS integration runtime
- Self-hosted integration runtime

Trigger type:
- Event-based trigger
- Schedule trigger
- Tumbling window trigger

Activity type:
- Copy activity
- Lookup activity
- Stored procedure activity

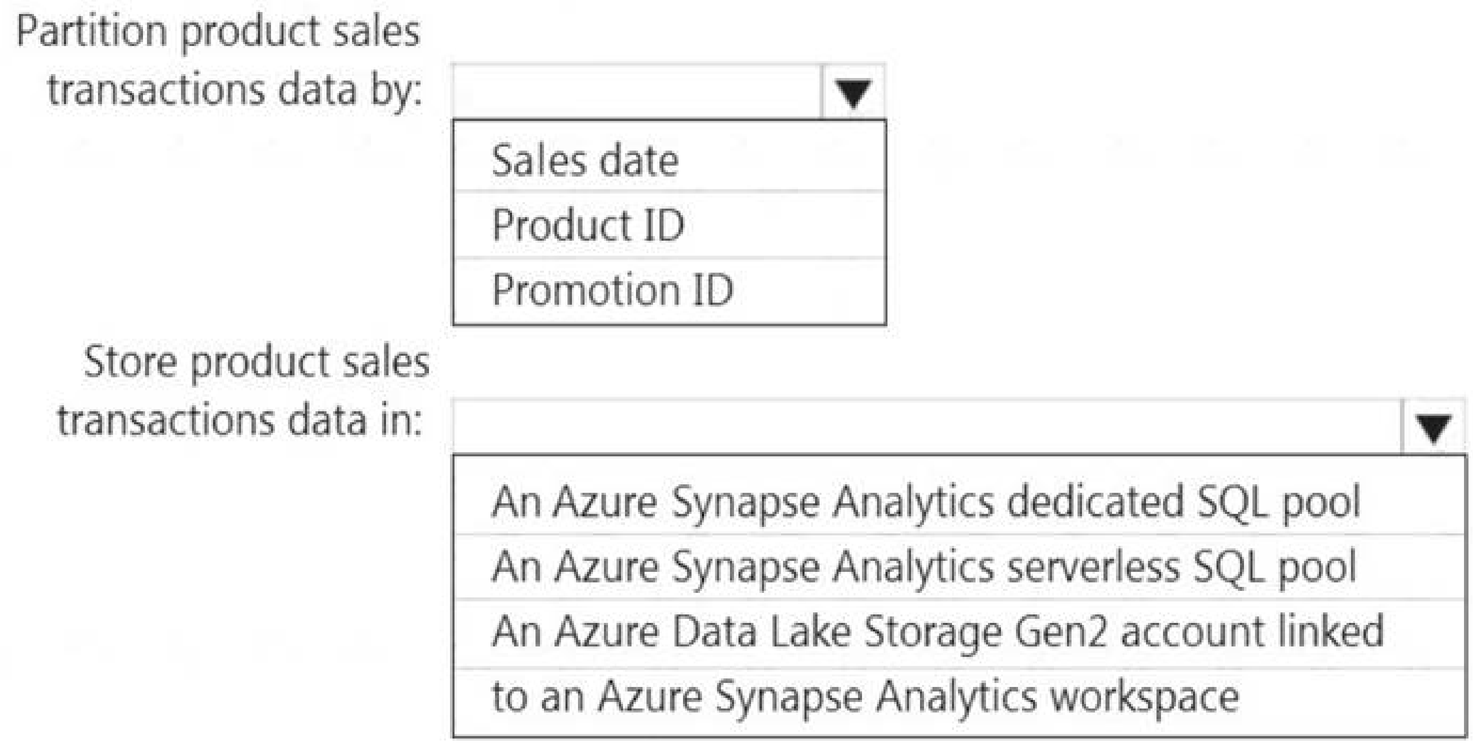
A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Self-hosted integration runtime
A self-hosted IR is capable of running copy activity between a cloud data stores and a data store in private network.
Box 2: Schedule trigger Schedule every 8 hours Box 3: Copy activity Scenario:

- Customer data, including name, contact information, and loyalty number, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
- Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.

NEW QUESTION 97
- (Exam Topic 1)
You need to design the partitions for the product sales transactions. The solution must meet the sales transaction dataset requirements.
What should you include in the solution? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.



Partition product sales transactions data by:
- Sales date
- Product ID
- Promotion ID

Store product sales transactions data in:
- An Azure Synapse Analytics dedicated SQL pool
- An Azure Synapse Analytics serverless SQL pool
- An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Sales date
Scenario: Contoso requirements for data integration include:

- Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

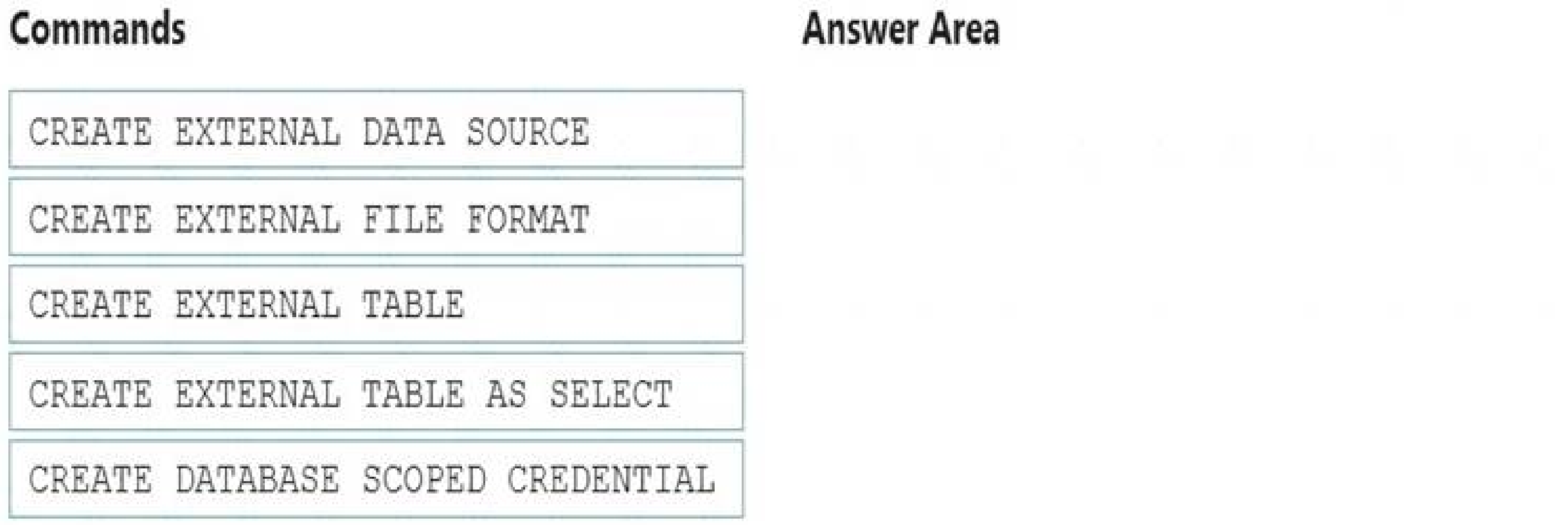
Box 2: An Azure Synapse Analytics Dedicated SQL pool Scenario: Contoso requirements for data integration include:

- Ensure that data storage costs and performance are predictable.

The size of a dedicated SQL pool (formerly SQL DW) is determined by Data Warehousing Units (DWU). Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage. This format
significantly reduces the data storage costs, and improves query performance.
Synapse analytics dedicated sql pool Reference:

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

NEW QUESTION 99
- (Exam Topic 1)
You need to ensure that the Twitter feed data can be analyzed in the dedicated SQL pool. The solution must meet the customer sentiment analytics requirements.
Which three Transaction-SQL DDL commands should you run in sequence? To answer, move the appropriate commands from the list of commands to the answer area and arrange them in the correct order.
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.



Commands

CREATE EXTERNAL DATA SOURCE

CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL TABLE

CREATE EXTERNAL TABLE AS SELECT

CREATE DATABASE SCOPED CREDENTIAL

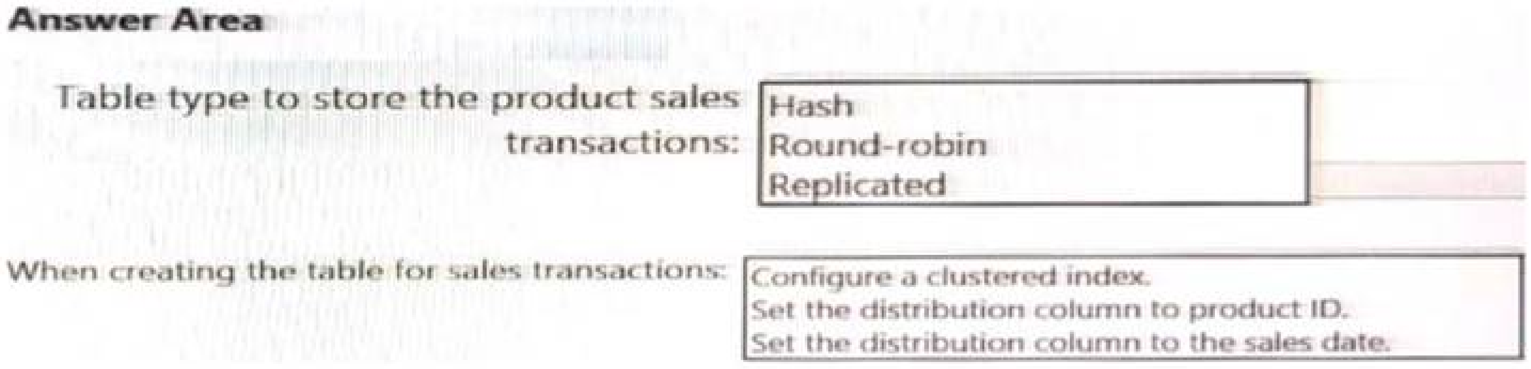
Answer Area

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Scenario: Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds. Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.
Box 1: CREATE EXTERNAL DATA SOURCE
External data sources are used to connect to storage accounts. Box 2: CREATE EXTERNAL FILE FORMAT
CREATE EXTERNAL FILE FORMAT creates an external file format object that defines external data stored in Azure Blob Storage or Azure Data Lake Storage. Creating an external file format is a prerequisite for creating an external table.
Box 3: CREATE EXTERNAL TABLE AS SELECT
When used in conjunction with the CREATE TABLE AS SELECT statement, selecting from an external table imports data into a table within the SQL pool. In addition to the COPY statement, external tables are useful for loading data.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

NEW QUESTION 103
- (Exam Topic 1)
You need to design a data storage structure for the product sales transactions. The solution must meet the sales transaction dataset requirements.
What should you include in the solution? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.



Answer Area

Table type to store the product sales transactions:
Hash
Round-robin
Replicated

When creating the table for sales transactions:
Configure a clustered index.
Set the distribution column to product ID.
Set the distribution column to the sales date.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application, chat or text message Description automatically generated
Box 1: Hash Scenario:
Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.
A hash distributed table can deliver the highest query performance for joins and aggregations on large tables. Box 2: Set the distribution column to the sales date.
Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.
Reference:
https://rajanieshkaushikk.com/2020/09/09/how-to-choose-right-data-distribution-strategy-for-azure-synapse/

NEW QUESTION 106
- (Exam Topic 1)
You need to design an analytical storage solution for the transactional data. The solution must meet the sales transaction dataset requirements.
What should you include in the solution? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

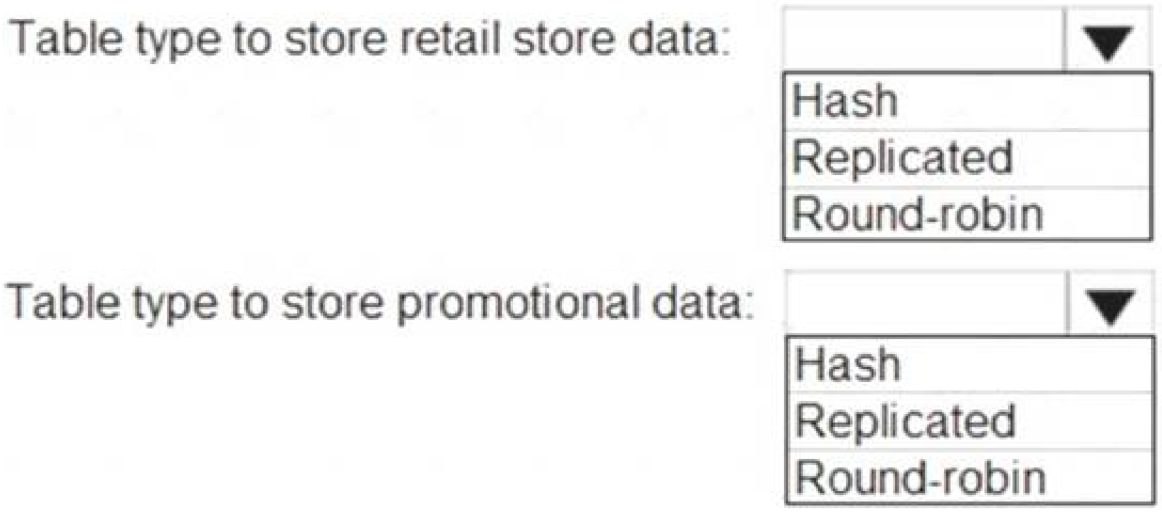


Table type to store retail store data: ▼
Hash
Replicated
Round-robin

Table type to store promotional data: ▼
Hash
Replicated
Round-robin

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application, table Description automatically generated
Box 1: Round-robin
Round-robin tables are useful for improving loading speed.
Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month.
Box 2: Hash
Hash-distributed tables improve query performance on large fact tables. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

NEW QUESTION 111
- (Exam Topic 1)
You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements.
Which Azure Storage functionality should you include in the solution?

A. change feed
B. soft delete
C. time-based retention
D. lifecycle management

Answer: D

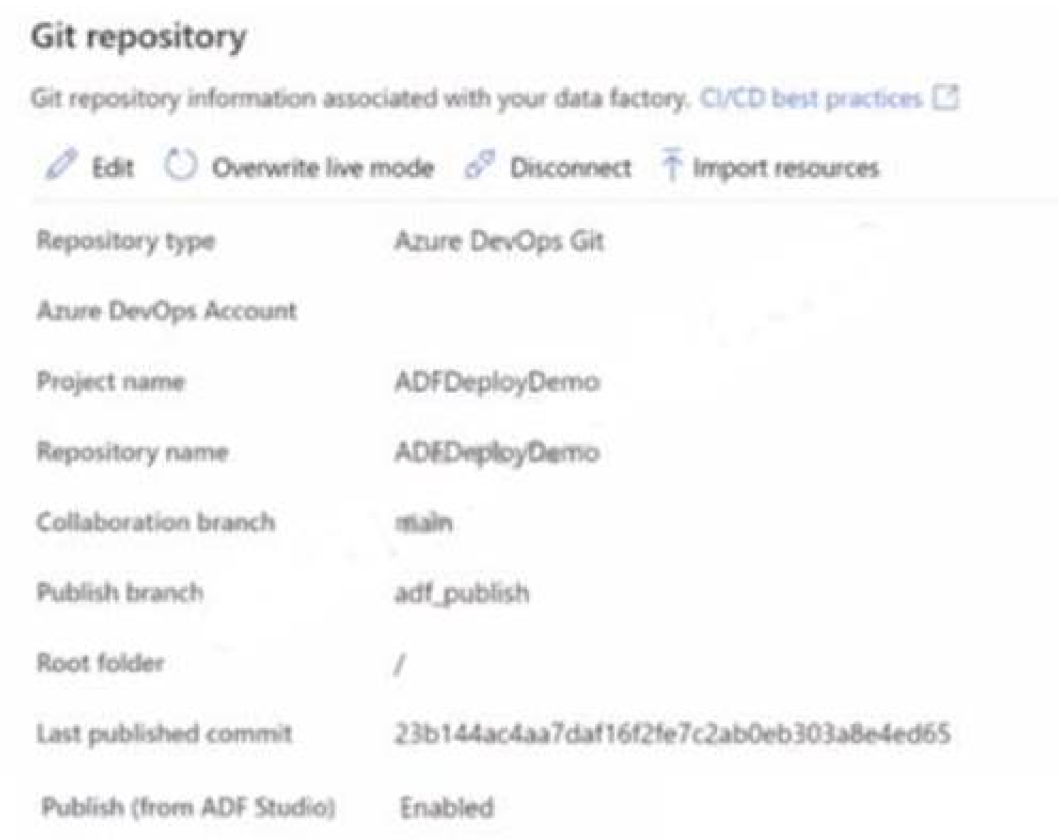
Explanation:
Scenario: Purge Twitter feed data records that are older than two years.
Data sets have unique lifecycles. Early in the lifecycle, people access some data often. But the need for access often drops drastically as the data ages. Some data remains idle in the cloud and is rarely accessed once stored. Some data sets expire days or months after creation, while other data sets are actively read and modified throughout their lifetimes. Azure Storage lifecycle management offers a rule-based policy that you can use to transition blob data to the appropriate access tiers or to expire data at the end of the data lifecycle.
Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-overview

NEW QUESTION 114
- (Exam Topic 3)
You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytics dedicated SQL pool. The CSV file contains columns named username, comment and date.
The data flow already contains the following:
• A source transformation
• A Derived Column transformation to set the appropriate types of data
• A sink transformation to land the data in the pool
You need to ensure that the data flow meets the following requirements;
• All valid rows must be written to the destination table.
• Truncation errors in the comment column must be avoided proactively.
• Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.
Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point

A. Add a select transformation that selects only the rows which will cause truncation errors.
B. Add a sink transformation that writes the rows to a file in blob storage.
C. Add a filter transformation that filters out rows which will cause truncation errors.
D. Add a Conditional Split transformation that separates the rows which will cause truncation errors.

Answer: BD

NEW QUESTION 117
- (Exam Topic 3)
You have an Azure data factory that has the Git repository settings shown in the following exhibit.

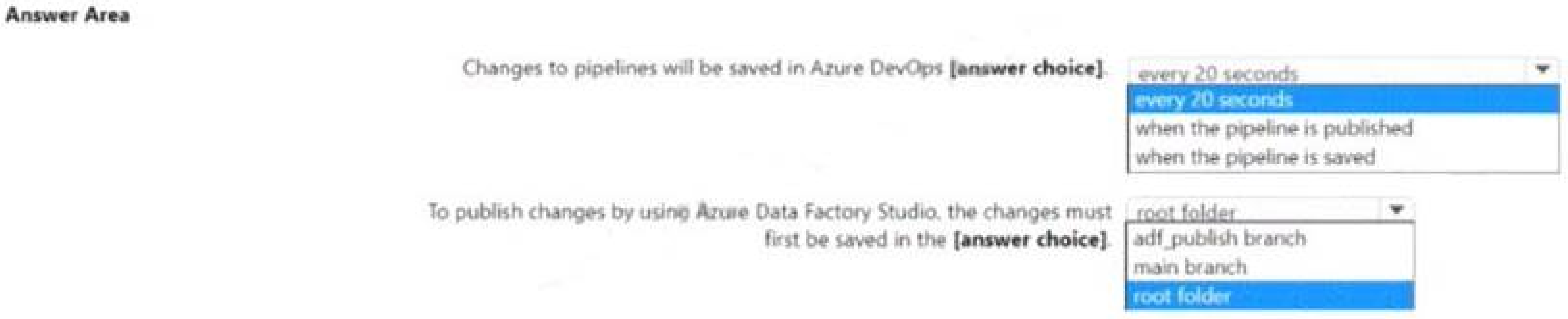


Git repository

Git repository information associated with your data factory. CI/CD best practices

Repository type	Azure DevOps Git
Azure DevOps Account	
Project name	ADFDeployDemo
Repository name	ADFDeployDemo
Collaboration branch	main
Publish branch	adf_publish
Root folder	/
Last published commit	23b144ac4aa7daf16f2fe7c2ab0eb303a8e4ed65
Publish (from ADF Studio)	Enabled

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



Answer Area

Changes to pipelines will be saved in Azure DevOps **[answer choice]**.

every 20 seconds
every 20 seconds
when the pipeline is published
when the pipeline is saved

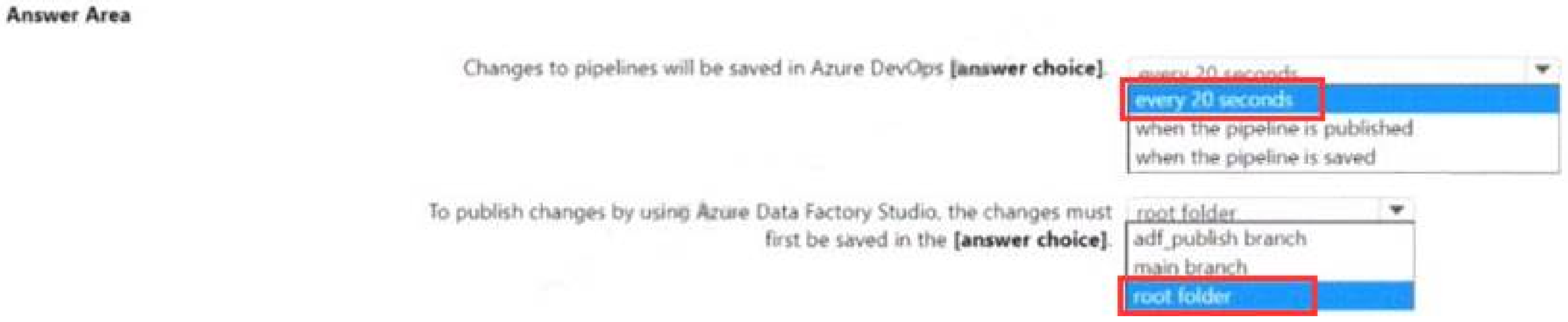
To publish changes by using Azure Data Factory Studio, the changes must first be saved in the **[answer choice]**.

root folder
adf_publish branch
main branch
root folder

A. Mastered
B. Not Mastered

Answer: A

Explanation:



Answer Area

Changes to pipelines will be saved in Azure DevOps **[answer choice]**.

every 20 seconds
every 20 seconds
when the pipeline is published
when the pipeline is saved

To publish changes by using Azure Data Factory Studio, the changes must first be saved in the **[answer choice]**.

root folder
adf_publish branch
main branch
root folder

NEW QUESTION 122
- (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:

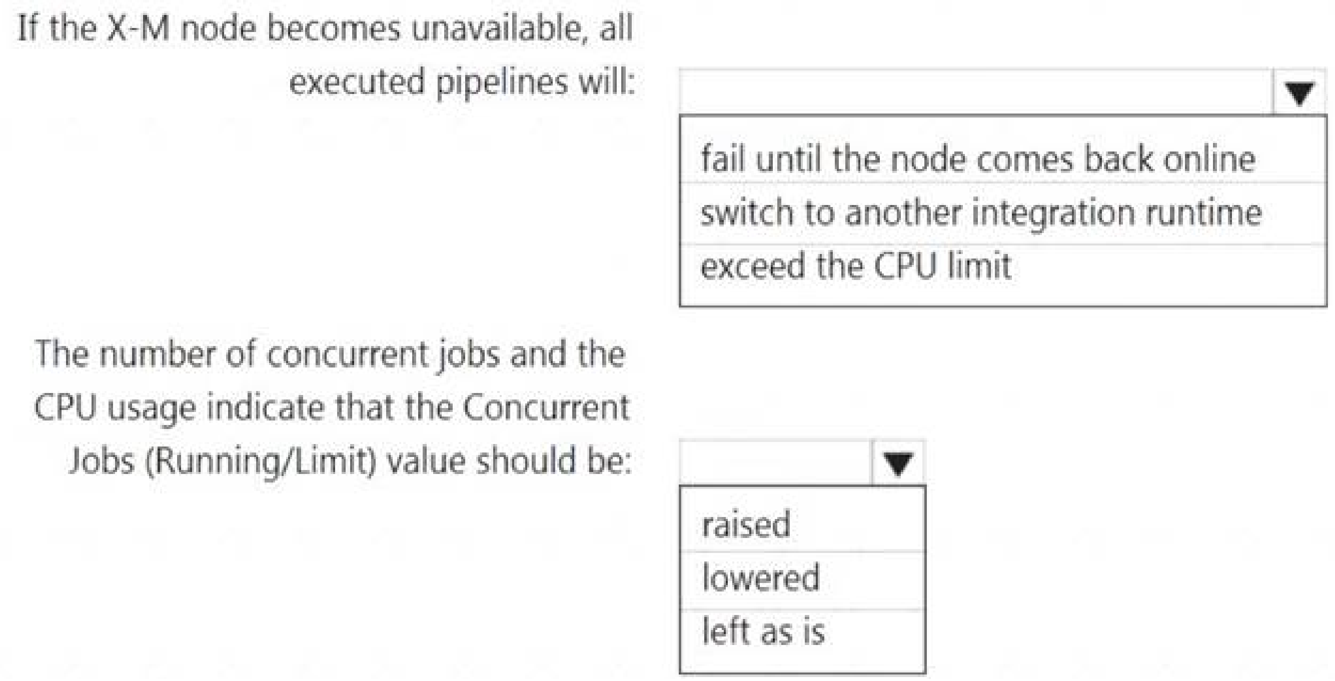
- Status: Running
- Type: Self-Hosted
- Version: 4.4.7292.1
- Running / Registered Node(s): 1/1
- High Availability Enabled: False
- Linked Count: 0
- Queue Length: 0
- Average Queue Duration. 0.00s

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: ▼

- raised
- lowered
- left as is

A. Mastered
B. Not Mastered

Answer: A

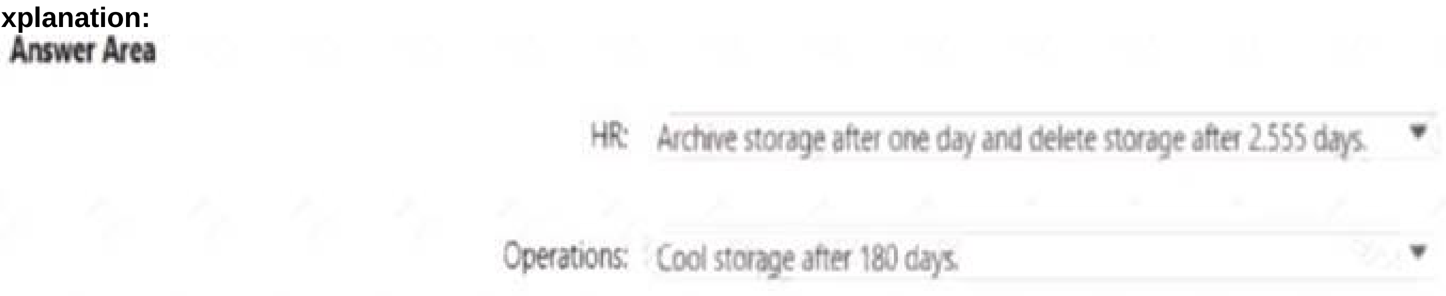
Explanation:
Box 1: fail until the node comes back online We see: High Availability Enabled: False
Note: Higher availability of the self-hosted integration runtime so that it's no longer the single point of failure in your big data solution or cloud data integration with Data Factory.
Box 2: lowered 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 126
- (Exam Topic 3)
You are designing an Azure Data Lake Storage Gen2 container to store data for the human resources (HR) department and the operations department at your company. You have the following data access requirements:
• After initial processing, the HR department data will be retained for seven years.
• The operations department data will be accessed frequently for the first six months, and then accessed once per month.
You need to design a data retention solution to meet the access requirements. The solution must minimize storage costs.

A. Mastered
B. Not Mastered

Answer: A

Explanation:

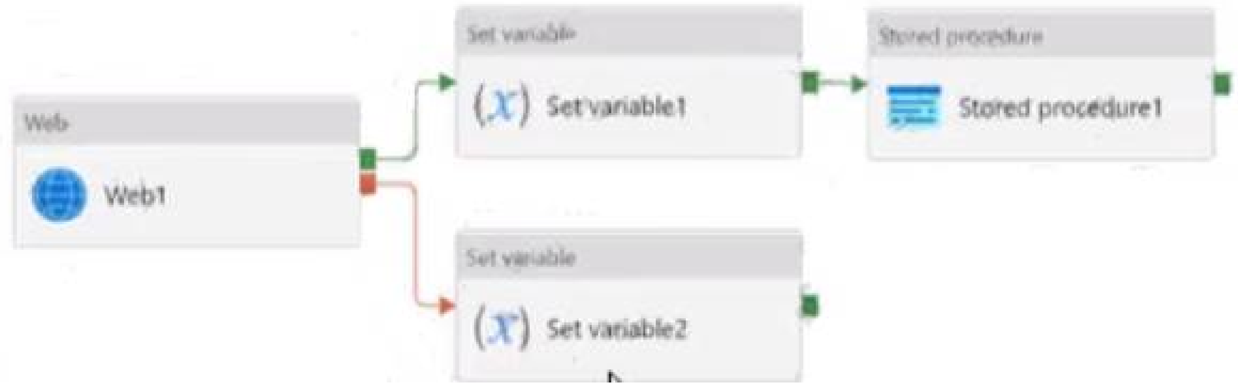


Answer Area

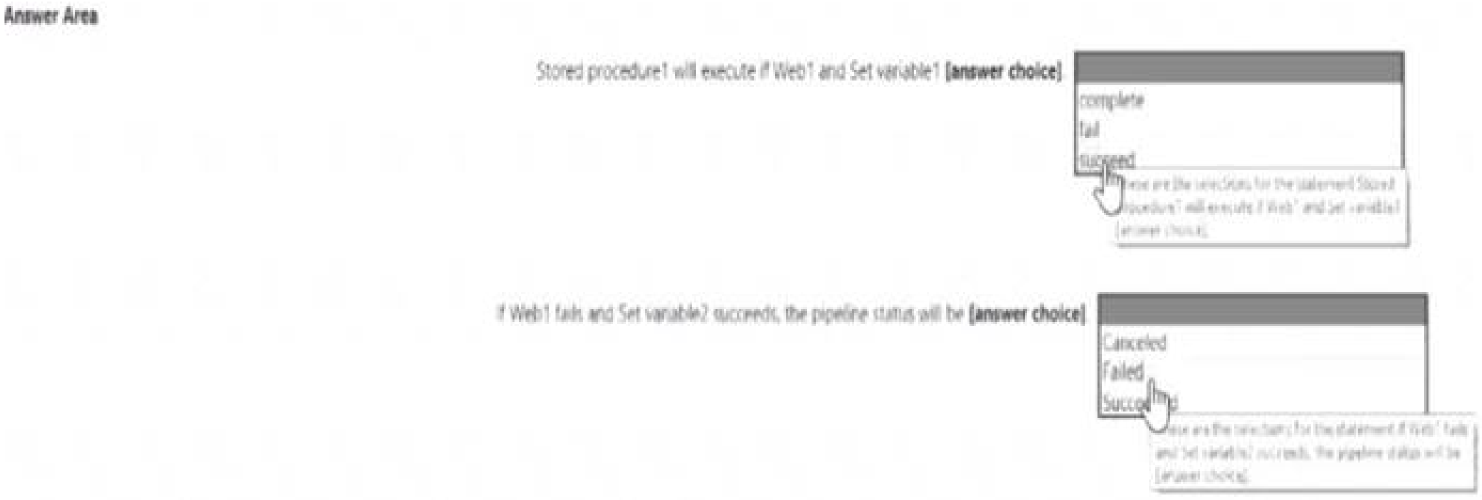
HR: Archive storage after one day and delete storage after 2,555 days. ▼

Operations: Cool storage after 180 days. ▼

NEW QUESTION 131
- (Exam Topic 3)
You have an Azure Data Factory pipeline that has the activity shown in the following exhibit.



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



Answer Area

Stored procedure1 will execute if Web1 and Set variable1 **[answer choice]**.

These are the selections for the statement Stored procedure1 will execute if Web1 and Set variable1 [answer choice].

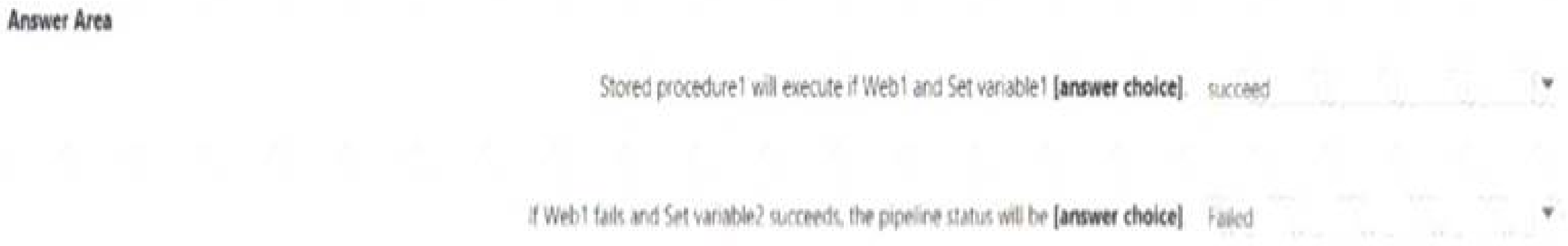
If Web1 fails and Set variable2 succeeds, the pipeline status will be **[answer choice]**

These are the selections for the statement If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice].

A. Mastered
B. Not Mastered

Answer: A

Explanation:



Answer Area

Stored procedure1 will execute if Web1 and Set variable1 **[answer choice]**. succeed

If Web1 fails and Set variable2 succeeds, the pipeline status will be **[answer choice]** Failed

NEW QUESTION 132
- (Exam Topic 3)
You are designing a sales transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will contains approximately 60 million rows per month and will be partitioned by month. The table will use a clustered column store index and round-robin distribution.
Approximately how many rows will there be for each combination of distribution and partition?

A. 1 million
B. 5 million
C. 20 million
D. 60 million

Answer: D

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

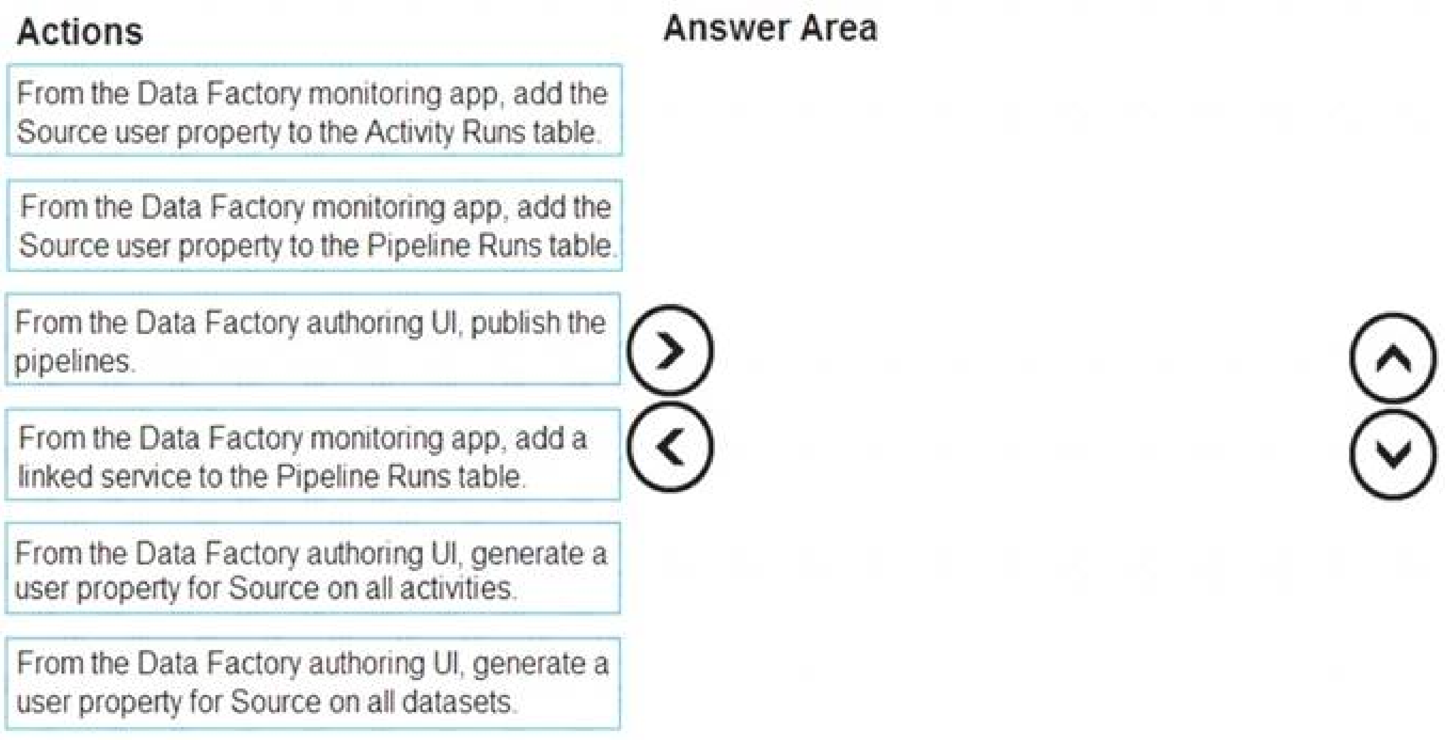
NEW QUESTION 137
- (Exam Topic 3)
You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account.
The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/. You need to design a daily Azure Data Factory data load to minimize the data transfer between the two accounts.
Which two configurations should you include in the design? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. Delete the files in the destination before loading new data.
B. Filter by the last modified date of the source files.
C. Delete the source files after they are copied.
D. Specify a file naming pattern for the destination.

Answer: BD

Explanation:
Copy data from one place to another. The requirements are : 1- need to minimize transfert and 2- need to adapte data to the destination folder structure. Filter on LastModifiedDate will copy everything that have changed since the latest load while minimizing the data transfert. Specifying the file naming pattern allows to copy data at the right place to the destination Data Lake.

NEW QUESTION 142
- (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 146
- (Exam Topic 3)
You have an Azure Data Factory pipeline named Pipeline1!. Pipelinel contains a copy activity that sends data to an Azure Data Lake Storage Gen2 account. Pipeline 1 is executed by a schedule trigger.
You change the copy activity sink to a new storage account and merge the changes into the collaboration branch.
After Pipelinel executes, you discover that data is NOT copied to the new storage account. You need to ensure that the data is copied to the new storage account.
What should you do?

A. Publish from the collaboration branch.
B. Configure the change feed of the new storage account.
C. Create a pull request.
D. Modify the schedule trigger.

Answer: A

Explanation:
CI/CD lifecycle

➢ A development data factory is created and configured with Azure Repos Git. All developers should have permission to author Data Factory resources like pipelines and datasets.

➢ A developer creates a feature branch to make a change. They debug their pipeline runs with their most recent changes

➢ After a developer is satisfied with their changes, they create a pull request from their feature branch to the main or collaboration branch to get their changes reviewed by peers.

➢ After a pull request is approved and changes are merged in the main branch, the changes get published to the development factory.

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

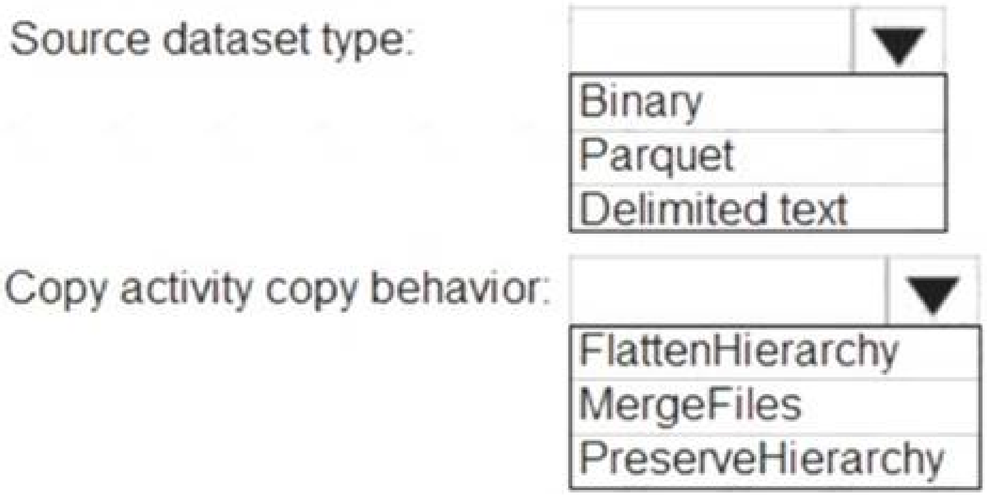
NEW QUESTION 150
- (Exam Topic 3)
You have two Azure Storage accounts named Storage1 and Storage2. Each account holds one container and has the hierarchical namespace enabled. The system has files that contain data stored in the Apache Parquet format.
You need to copy folders and files from Storage1 to Storage2 by using a Data Factory copy activity. The solution must meet the following requirements:

➢ No transformations must be performed.

➢ The original folder structure must be retained.

➢ Minimize time required to perform the copy activity.

How should you configure the copy activity? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Source dataset type:
- Binary
- Parquet
- Delimited text

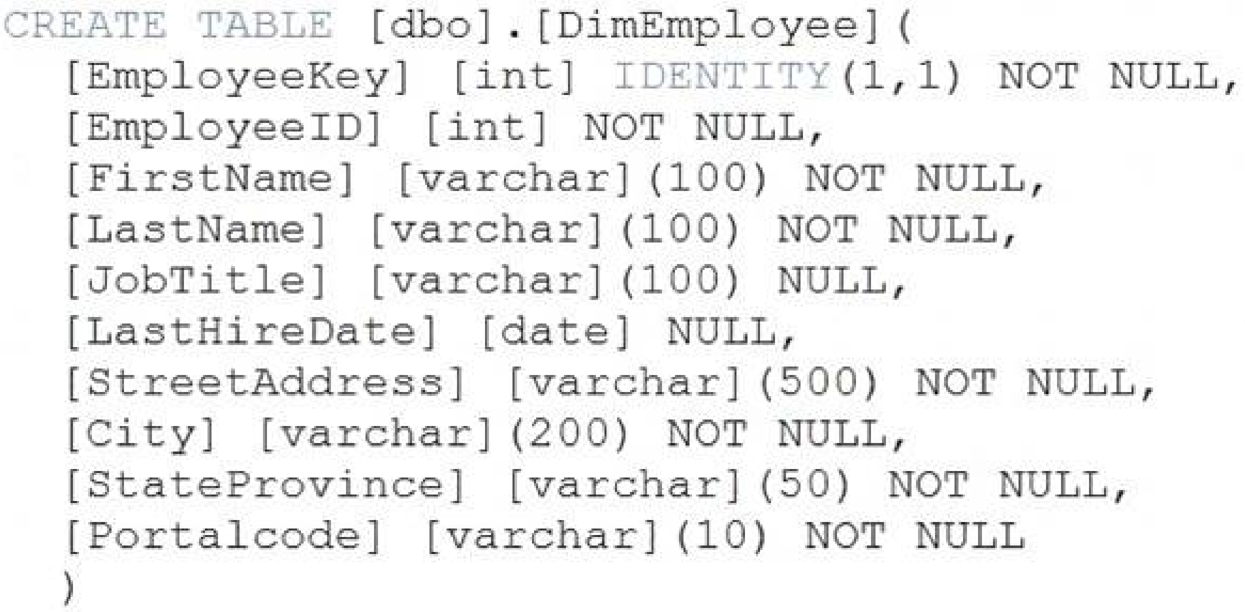
Copy activity copy behavior:
- FlattenHierarchy
- MergeFiles
- PreserveHierarchy

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application, chat or text message Description automatically generated
Box 1: Parquet
For Parquet datasets, the type property of the copy activity source must be set to ParquetSource. Box 2: PreserveHierarchy
PreserveHierarchy (default): Preserves the file hierarchy in the target folder. The relative path of the source file to the source folder is identical to the relative path of the target file to the target folder.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/format-parquet https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage

NEW QUESTION 151
- (Exam Topic 3)
You have a table in an Azure Synapse Analytics dedicated SQL pool. The table was created by using the following Transact-SQL statement.



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

You need to alter the table to meet the following requirements:
- Ensure that users can identify the current manager of employees.
- Support creating an employee reporting hierarchy for your entire company.
- Provide fast lookup of the managers' attributes such as name and job title.

Which column should you add to the table?

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

Answer: A

Explanation:
Use the same definition as the EmployeeID column. Reference:
https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular

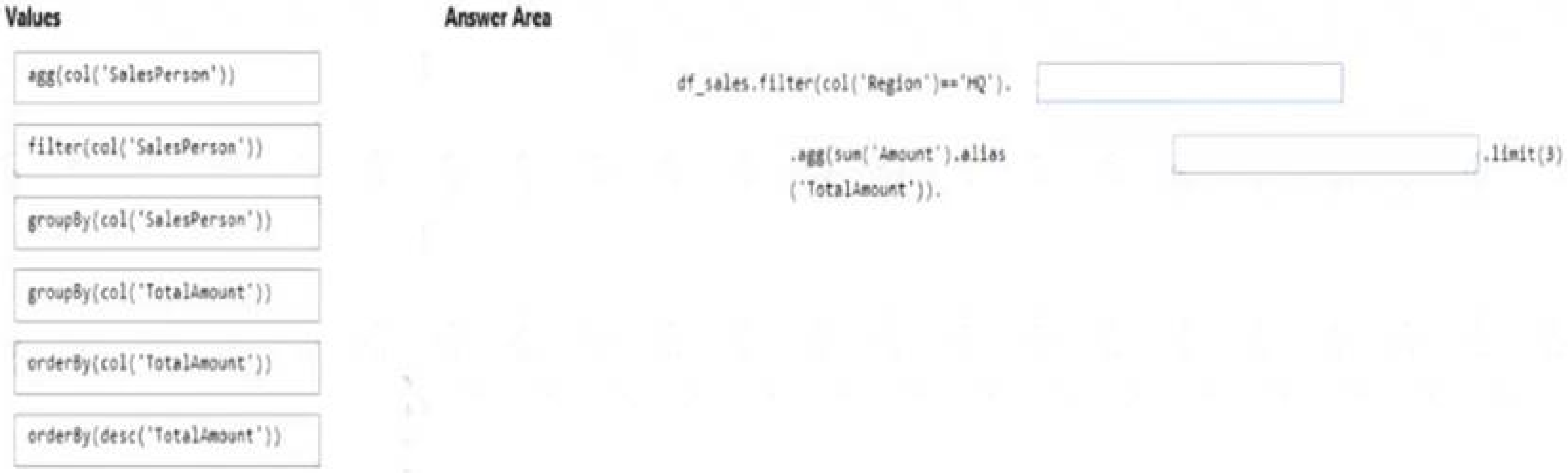
NEW QUESTION 156
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Contacts. Contacts contains a column named Phone.
You need to ensure that users in a specific role only see the last four digits of a phone number when querying the Phone column.
What should you include in the solution?

A. a default value
B. dynamic data masking
C. row-level security (RLS)
D. column encryption
E. table partitions

Answer: B

Explanation:
Dynamic data masking helps prevent unauthorized access to sensitive data by enabling customers to designate how much of the sensitive data to reveal with minimal impact on the application layer. It's a policy-based security feature that hides the sensitive data in the result set of a query over designated database fields, while the data in the database is not changed.
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

NEW QUESTION 159
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Databricks workspace. The workspace contains a notebook named Notebook1. In Notebook1, you create an Apache Spark DataFrame named df_sales that contains the following columns:
• Customer
• Salesperson
• Region
• Amount
You need to identify the three top performing salespersons by amount for a region named HQ.
How should you complete the query? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.



Values

```
agg(col('SalesPerson'))
filter(col('SalesPerson'))
groupBy(col('SalesPerson'))
groupBy(col('TotalAmount'))
orderBy(col('TotalAmount'))
orderBy(desc('TotalAmount'))
```

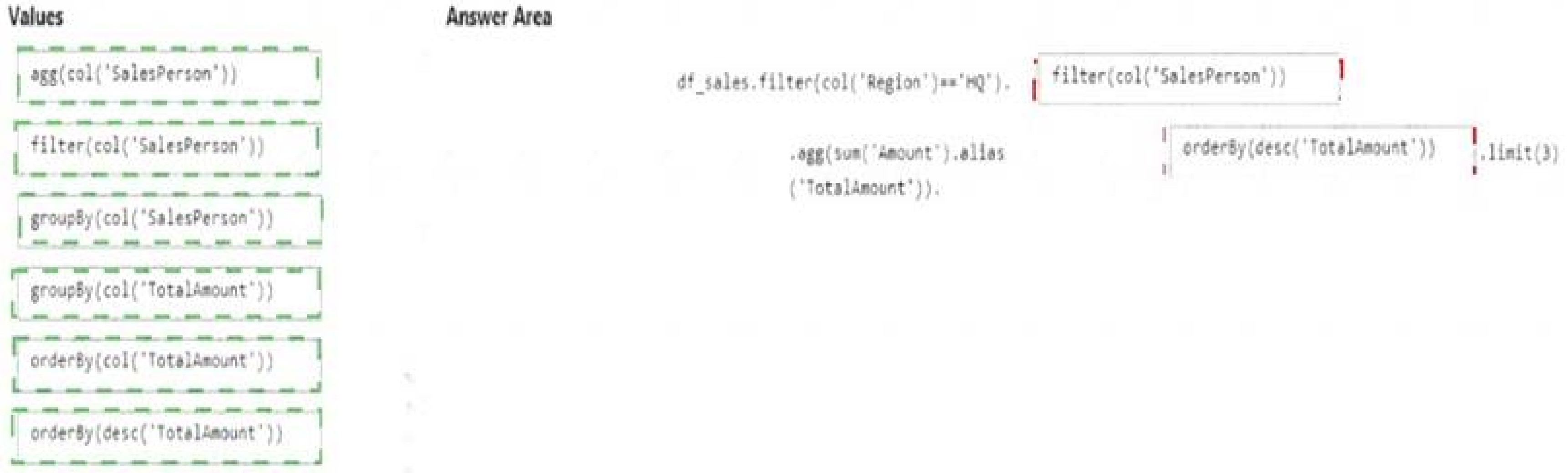
Answer Area

```
df_sales.filter(col('Region')=='HQ'). [          ]
.agg(sum('Amount').alias
('TotalAmount')).  [          ] .limit(3)
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:



Values

```
agg(col('SalesPerson'))
filter(col('SalesPerson'))
groupBy(col('SalesPerson'))
groupBy(col('TotalAmount'))
orderBy(col('TotalAmount'))
orderBy(desc('TotalAmount'))
```

Answer Area

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

NEW QUESTION 162
- (Exam Topic 3)
You are designing a highly available Azure Data Lake Storage solution that will induce geo-zone-redundant storage (GZRS).
You need to monitor for replication delays that can affect the recovery point objective (RPO). What should you include m the monitoring solution?

A. Last Sync Time
B. Average Success Latency
C. Error errors
D. availability

Answer: A

Explanation:
Because geo-replication is asynchronous, it is possible that data written to the primary region has not yet been written to the secondary region at the time an outage occurs. The Last Sync Time property indicates the last time that data from the primary region was written successfully to the secondary region. All writes made to the primary region before the last sync time are available to be read from the secondary location. Writes made to the primary region after the last sync time property may or may not be available for reads yet.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/last-sync-time-get

NEW QUESTION 165
- (Exam Topic 3)
You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination.
You need to ensure that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs.
What should you do?

A. Clone the cluster after it is terminated.
B. Terminate the cluster manually when processing completes.
C. Create an Azure runbook that starts the cluster every 90 days.
D. Pin the cluster.

Answer: D

Explanation:
To keep an interactive cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.
References:
https://docs.azuredatabricks.net/clusters/clusters-manage.html#automatic-termination

NEW QUESTION 170
- (Exam Topic 3)
You are designing a financial transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:

- TransactionType: 40 million rows per transaction type
- CustomerSegment: 4 million per customer segment
- TransactionMonth: 65 million rows per month
- AccountType: 500 million per account type

You have the following query requirements:

- Analysts will most commonly analyze transactions for a given month.
- Transactions analysis will typically summarize transactions by transaction type, customer segment, and/or account type

You need to recommend a partition strategy for the table to minimize query times. On which column should you recommend partitioning the table?

A. CustomerSegment
B. AccountType
C. TransactionType
D. TransactionMonth

Answer: C

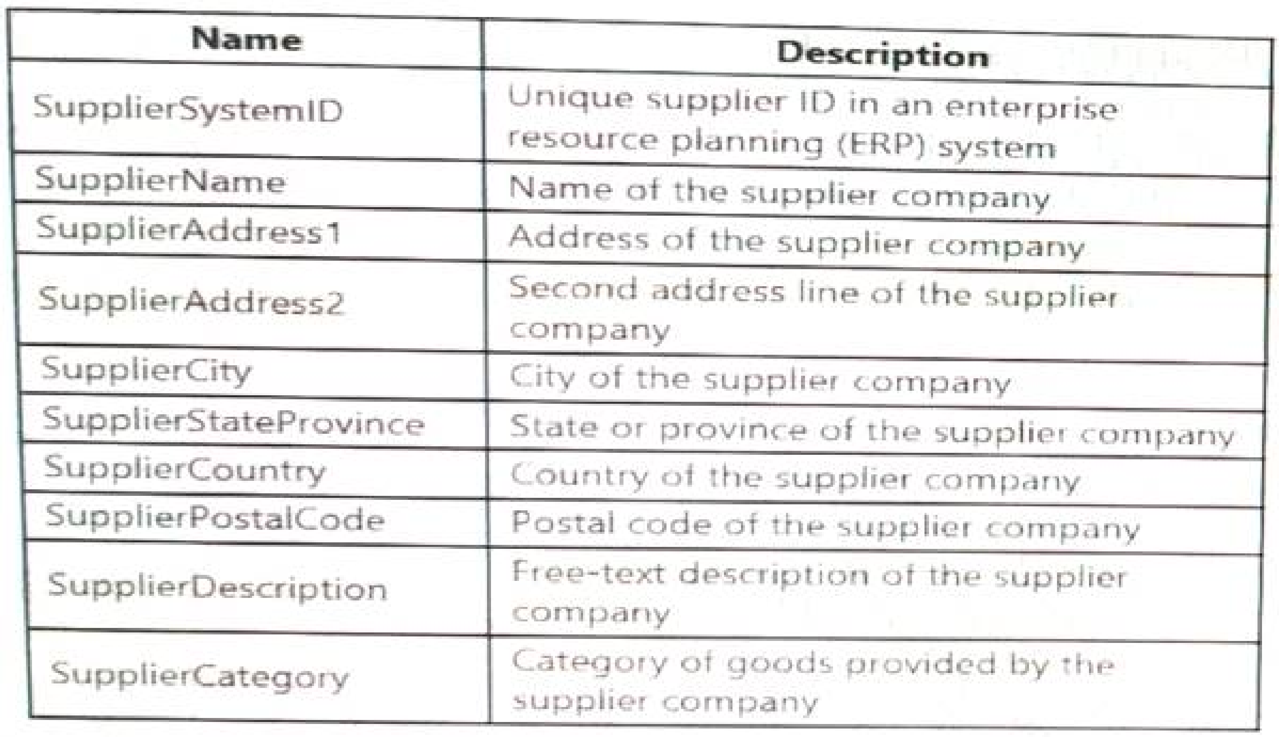
Explanation:
For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.
Example: Any partitioning added to a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

NEW QUESTION 173
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named pool1.
You need to perform a monthly audit of SQL statements that affect sensitive data. The solution must minimize administrative effort.
What should you include in the solution?

A. Microsoft Defender for SQL
B. dynamic data masking
C. sensitivity labels
D. workload management

Answer: B

NEW QUESTION 176
- (Exam Topic 3)
You are designing a slowly changing dimension (SCD) for supplier data in an Azure Synapse Analytics dedicated SQL pool.
You plan to keep a record of changes to the available fields. The supplier data contains the following columns.



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

Which three additional columns should you add to the data to create a Type 2 SCD? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

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

Answer: CDF

Explanation:
Reference:
https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/slowly-changing-dimension

NEW QUESTION 180
- (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 183
- (Exam Topic 3)
You plan to perform batch processing in Azure Databricks once daily. Which type of Databricks cluster should you use?

A. High Concurrency
B. automated
C. interactive

Answer: C

Explanation:
Azure Databricks has two types of clusters: interactive and automated. You use interactive clusters to analyze data collaboratively with interactive notebooks. You use automated clusters to run fast and robust automated jobs.
Example: Scheduled batch workloads (data engineers running ETL jobs)

This scenario involves running batch job JARs and notebooks on a regular cadence through the Databricks platform.
The suggested best practice is to launch a new cluster for each run of critical jobs. This helps avoid any issues (failures, missing SLA, and so on) due to an existing workload (noisy neighbor) on a shared cluster.
Reference:
https://docs.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-scheduled-bat

NEW QUESTION 186
- (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

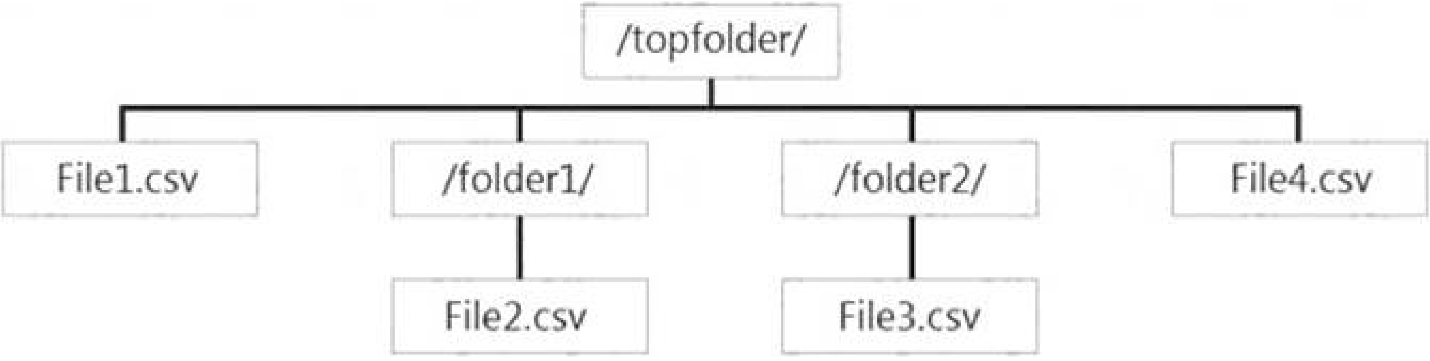
NEW QUESTION 187
- (Exam Topic 3)
You plan to implement an Azure Data Lake Gen2 storage account.
You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs.
Which type of replication should you use for the storage account?

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 189
- (Exam Topic 3)
You have files and folders in Azure Data Lake Storage Gen2 for an Azure Synapse workspace as shown in the following exhibit.



You create an external table named ExtTable that has LOCATION='/topfolder/'.
When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, which files are returned?

A. File2.csv and File3.csv only
B. File1.csv and File4.csv only
C. File1.csv, File2.csv, File3.csv, and File4.csv
D. File1.csv only

Answer: B

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

NEW QUESTION 194
- (Exam Topic 3)
You are designing a streaming data solution that will ingest variable volumes of data. You need to ensure that you can change the partition count after creation.
Which service should you use to ingest the data?

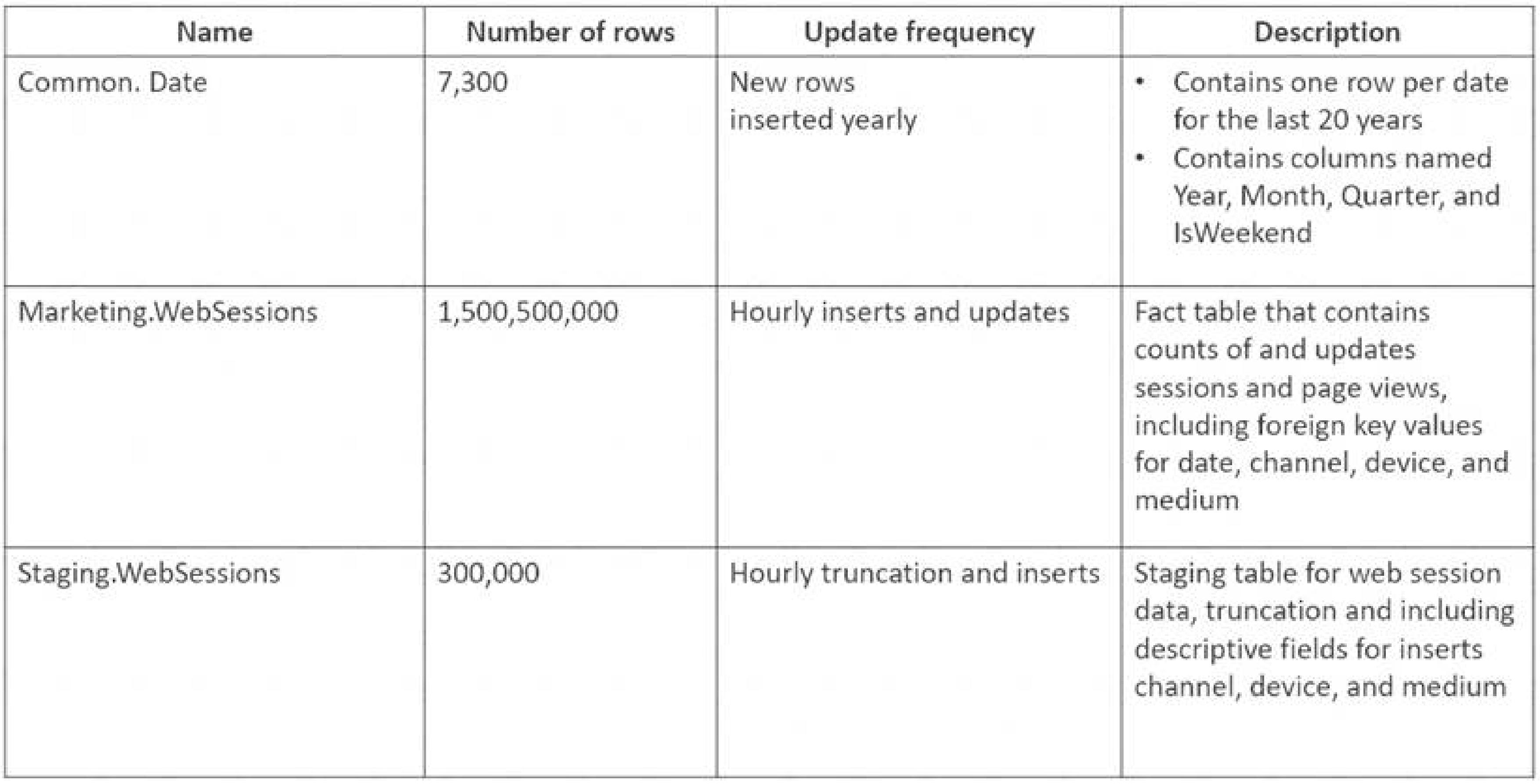
A. Azure Event Hubs Dedicated
B. Azure Stream Analytics
C. Azure Data Factory

D. Azure Synapse Analytics

Answer: B

Explanation:
You can't change the partition count for an event hub after its creation except for the event hub in a dedicated cluster.
Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features

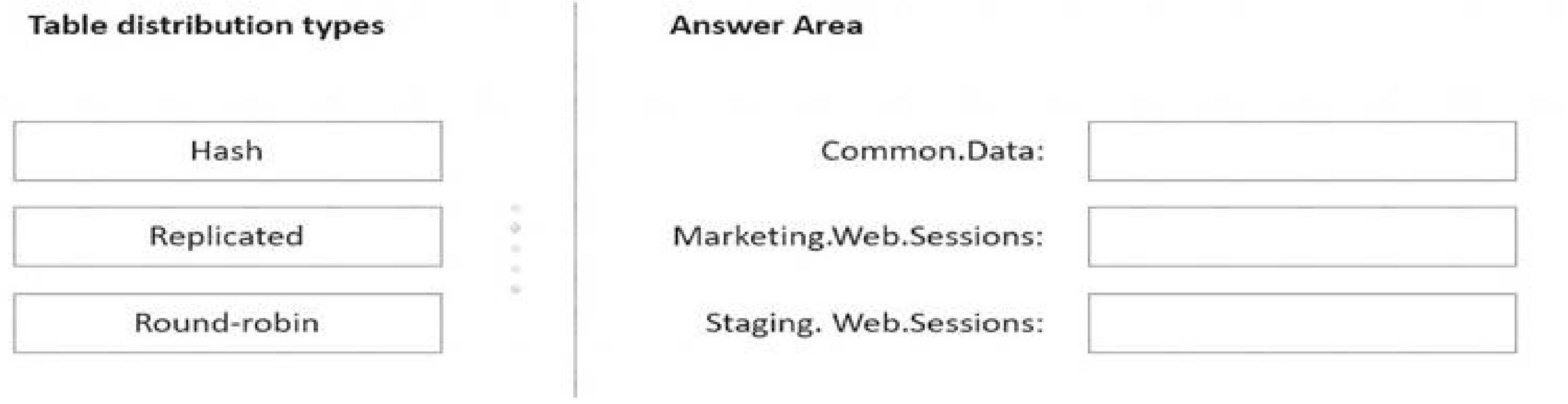
NEW QUESTION 199
- (Exam Topic 3)
You have an Azure subscription.
You plan to build a data warehouse in an Azure Synapse Analytics dedicated SQL pool named pool1 that will contain staging tables and a dimensional model.
Pool1 will contain the following tables.



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.

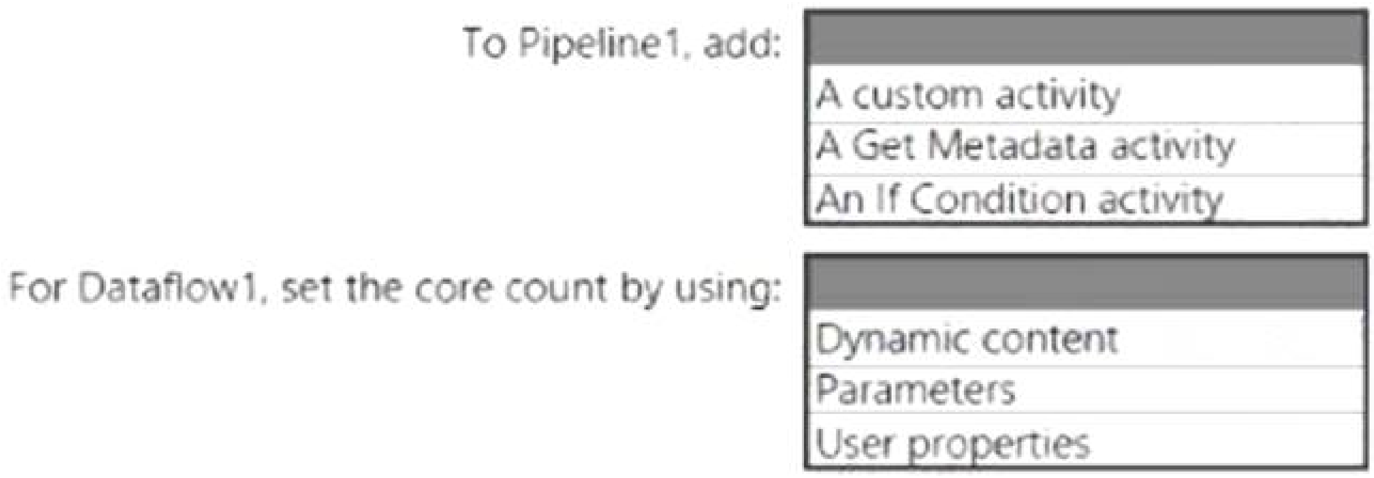


A. Mastered
B. Not Mastered

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 201
- (Exam Topic 3)
You have an Azure Synapse Analytics pipeline named Pipeline1 that contains a data flow activity named Dataflow1.
Pipeline1 retrieves files from an Azure Data Lake Storage Gen 2 account named storage1.
Dataflow1 uses the AutoResolveIntegrationRuntime integration runtime configured with a core count of 128. You need to optimize the number of cores used by Dataflow1 to accommodate the size of the files in storage1. What should you configure? To answer, select the appropriate options in the answer area.



To Pipeline1, add:
- A custom activity
- A Get Metadata activity
- An If Condition activity

For Dataflow1, set the core count by using:
- Dynamic content
- Parameters
- User properties

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: A Get Metadata activity
Dynamically size data flow compute at runtime
The Core Count and Compute Type properties can be set dynamically to adjust to the size of your incoming source data at runtime. Use pipeline activities like Lookup or Get Metadata in order to find the size of the source dataset data. Then, use Add Dynamic Content in the Data Flow activity properties.
Box 2: Dynamic content
Reference: https://docs.microsoft.com/en-us/azure/data-factory/control-flow-execute-data-flow-activity

NEW QUESTION 205
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a table named table1. You load 5 TB of data intotable1.
You need to ensure that columnstore compression is maximized for table1. Which statement should you execute?

A. ALTER INDEX ALL on table1 REORGANIZE
B. ALTER INDEX ALL on table1 REBUILD
C. DBCC DBREINOEX (table1)
D. DBCC INDEXDEFRAG (pool1,tablel)

Answer: B

Explanation:
Columnstore and columnstore archive compression
Columnstore tables and indexes are always stored with columnstore compression. You can further reduce the size of columnstore data by configuring an additional compression called archival compression. To perform archival compression, SQL Server runs the Microsoft XPRESS compression algorithm on the data. Add or remove archival compression by using the following data compression types:
Use COLUMNSTORE_ARCHIVE data compression to compress columnstore data with archival compression.
Use COLUMNSTORE data compression to decompress archival compression. The resulting data continue to be compressed with columnstore compression.
To add archival compression, use ALTER TABLE (Transact-SQL) or ALTER INDEX (Transact-SQL) with the REBUILD option and DATA COMPRESSION = COLUMNSTORE_ARCHIVE.
Reference: https://learn.microsoft.com/en-us/sql/relational-databases/data-compression/data-compression

NEW QUESTION 210
- (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?

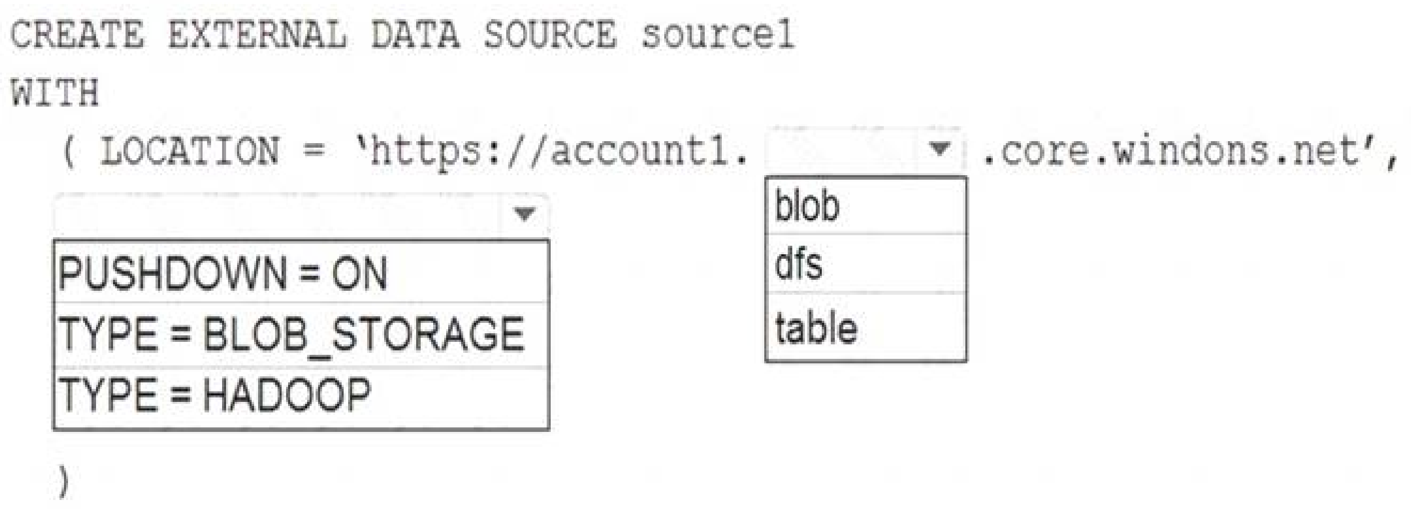
A. Yes
B. No

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 212
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named Account1.
You plan to access the files in Account1 by using an external table.

You need to create a data source in Pool1 that you can reference when you create the external table. How should you complete the Transact-SQL statement? To answer, select the appropriate options in the
answer area.
NOTE: Each correct selection is worth one point.



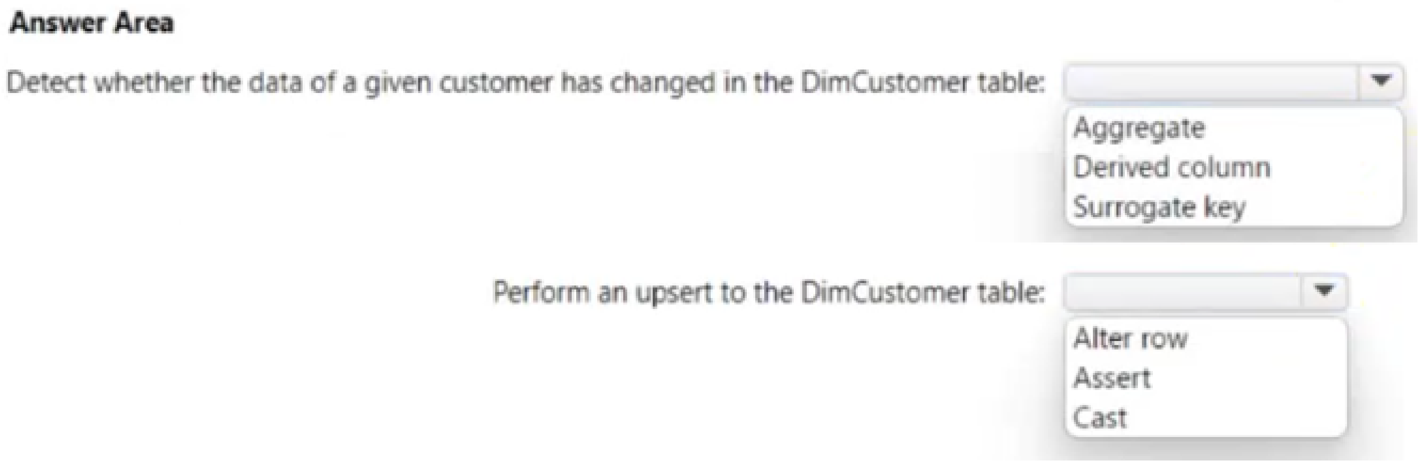
```
CREATE EXTERNAL DATA SOURCE source1
WITH
  ( LOCATION = 'https://account1. [▼] .core.windons.net',
  [▼]
  )
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, diagram Description automatically generated
Box 1: blob
The following example creates an external data source for Azure Data Lake Gen2 CREATE EXTERNAL DATA SOURCE YellowTaxi
WITH (LOCATION = 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/', TYPE = HADOOP)
Box 2: HADOOP
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

NEW QUESTION 215
- (Exam Topic 3)
You are developing an Azure Synapse Analytics pipeline that will include a mapping data flow named Dataflow1. Dataflow1 will read customer data from an external source and use a Type 1 slowly changing dimension (SCO) when loading the data into a table named DimCustomer1 in an Azure Synapse Analytics dedicated SQL pool.
You need to ensure that Dataflow1 can perform the following tasks:
* Detect whether the data of a given customer has changed in the DimCustomer table.
• Perform an upsert to the DimCustomer table.
Which type of transformation should you use for each task? To answer, select the appropriate options in the answer area
NOTE; Each correct selection is worth one point.



Answer Area

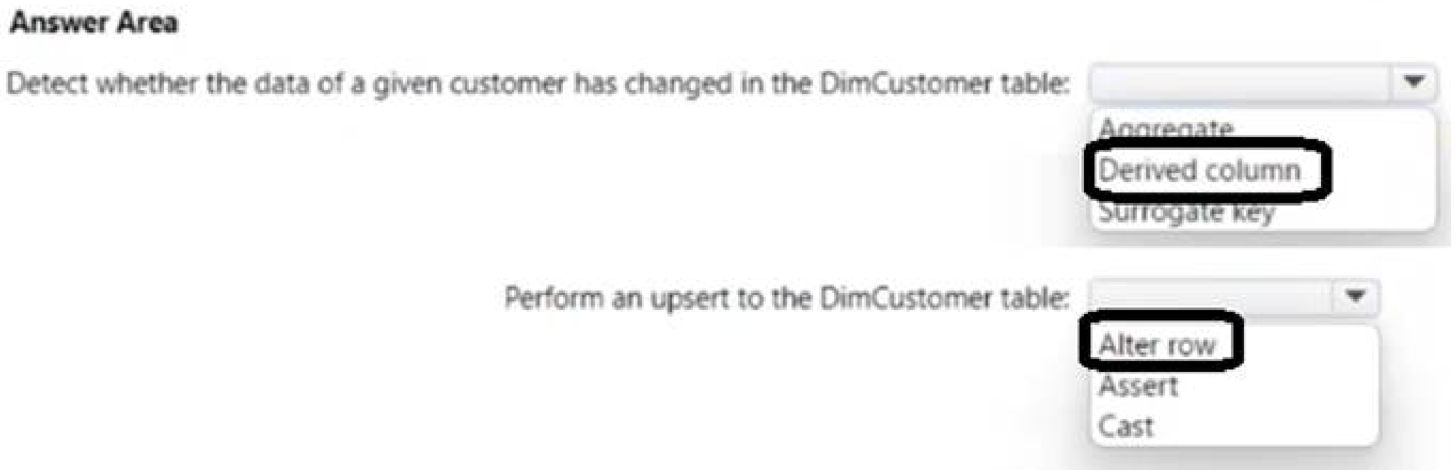
Detect whether the data of a given customer has changed in the DimCustomer table: [▼]

Perform an upsert to the DimCustomer table: [▼]

A. Mastered
B. Not Mastered

Answer: A

Explanation:

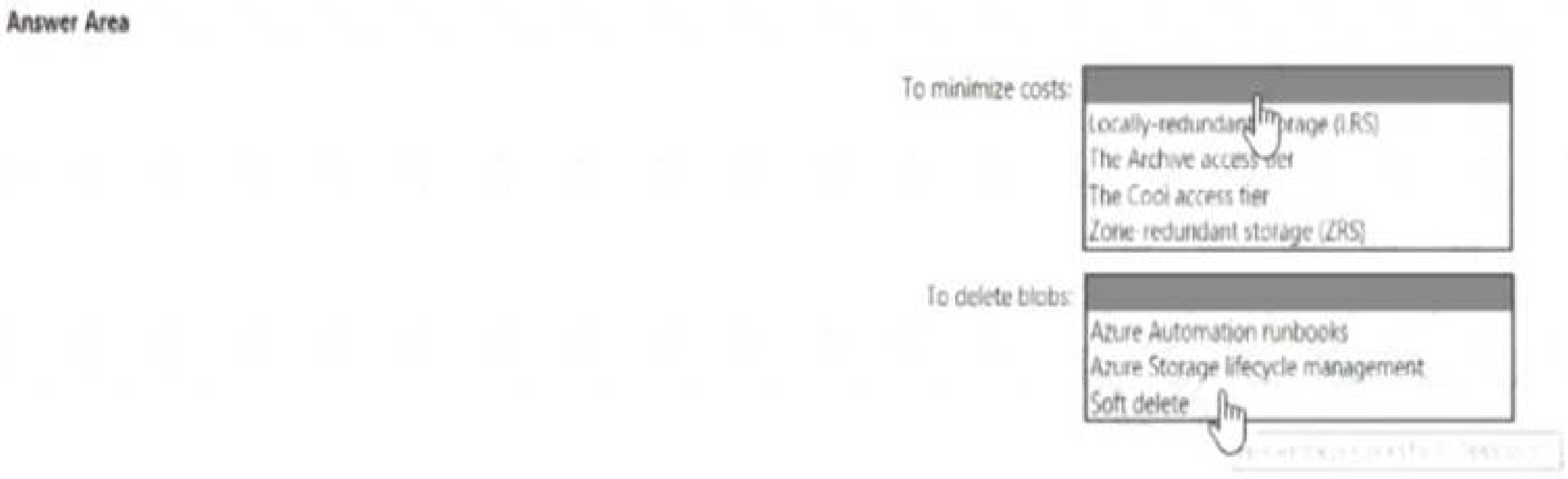


Answer Area

Detect whether the data of a given customer has changed in the DimCustomer table: [▼]

Perform an upsert to the DimCustomer table: [▼]

NEW QUESTION 218
- (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

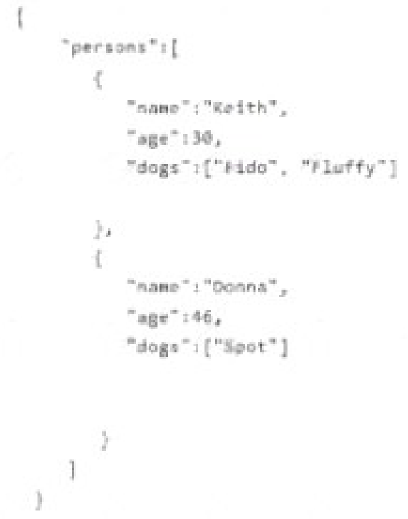
NEW QUESTION 223
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.
You need to identify the extent of the data skew in Table1. What should you do in Synapse Studio?

A. Connect to the built-in pool and query sysdm_pdw_sys_info.
B. Connect to Pool1 and run DBCC CHECKALLOC.
C. Connect to the built-in pool and run DBCC CHECKALLOC.
D. Connect to Pool! and query sys.dm_pdw_nodes_db_partition_stats.

Answer: D

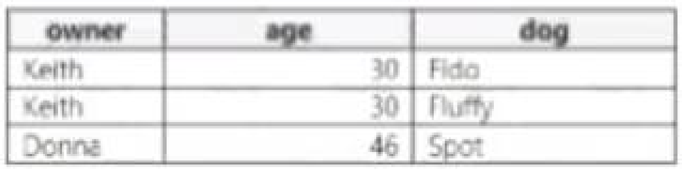
Explanation:
Microsoft recommends use of sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet

NEW QUESTION 228
- (Exam Topic 3)
You use PySpark in Azure Databricks to parse the following JSON input.



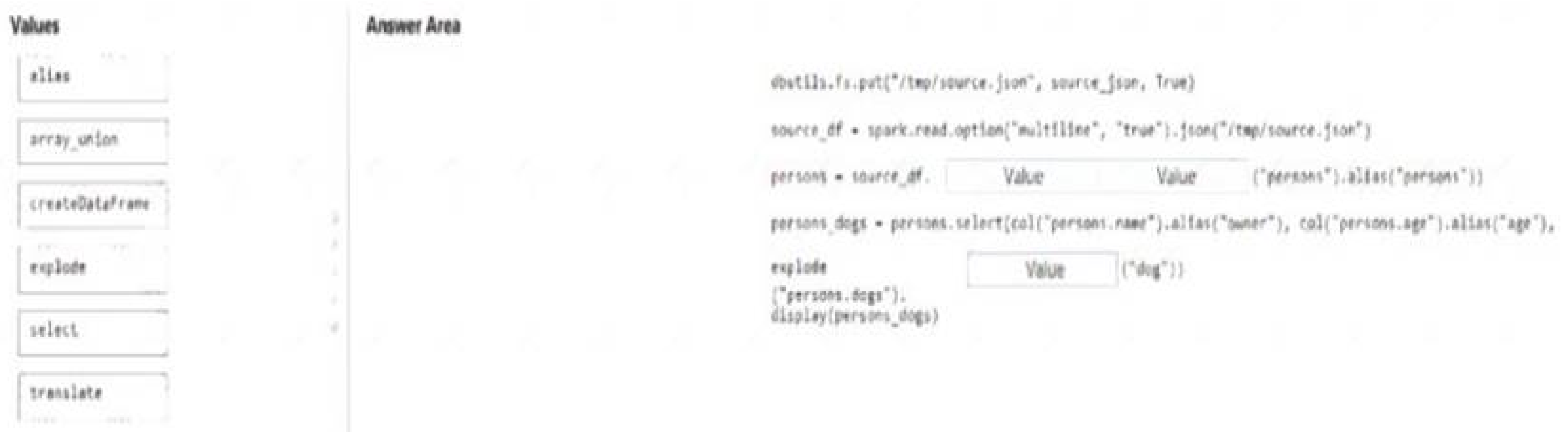
```
{
    "persons":[
       {
          "name":"Keith",
          "age":30,
          "dogs":["Fido", "Fluffy"]

       },
       {
          "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.



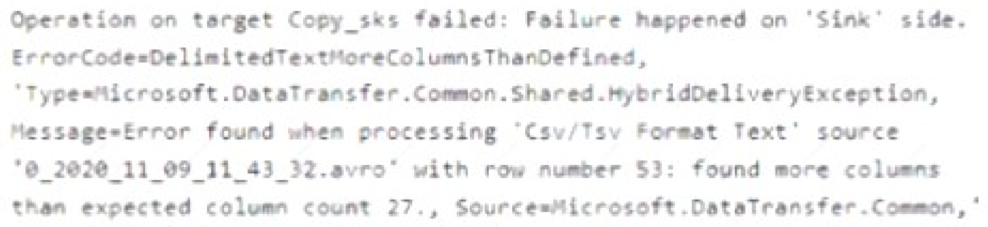
```
dbutils.fs.put("/tmp/source.json", source_json, True)

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 232
- (Exam Topic 3)
You have an Azure Data Lake Storage Gen2 account that contains two folders named Folder and Folder2. You use Azure Data Factory to copy multiple files from Folder1 to Folder2.



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

You receive the following error.
What should you do to resolve the error.

A. Add an explicit mapping.
B. Enable fault tolerance to skip incompatible rows.
C. Lower the degree of copy parallelism
D. Change the Copy activity setting to Binary Copy

Answer: A

Explanation:
Reference:
https://knowledge.informatica.com/s/article/Microsoft-Azure-Data-Lake-Store-Gen2-target-file-names-not-gene

NEW QUESTION 236
- (Exam Topic 3)
You are designing an Azure Data Lake Storage solution that will transform raw JSON files for use in an analytical workload.
You need to recommend a format for the transformed files. The solution must meet the following requirements:
- Contain information about the data types of each column in the files.
- Support querying a subset of columns in the files.
- Support read-heavy analytical workloads.
- Minimize the file size.

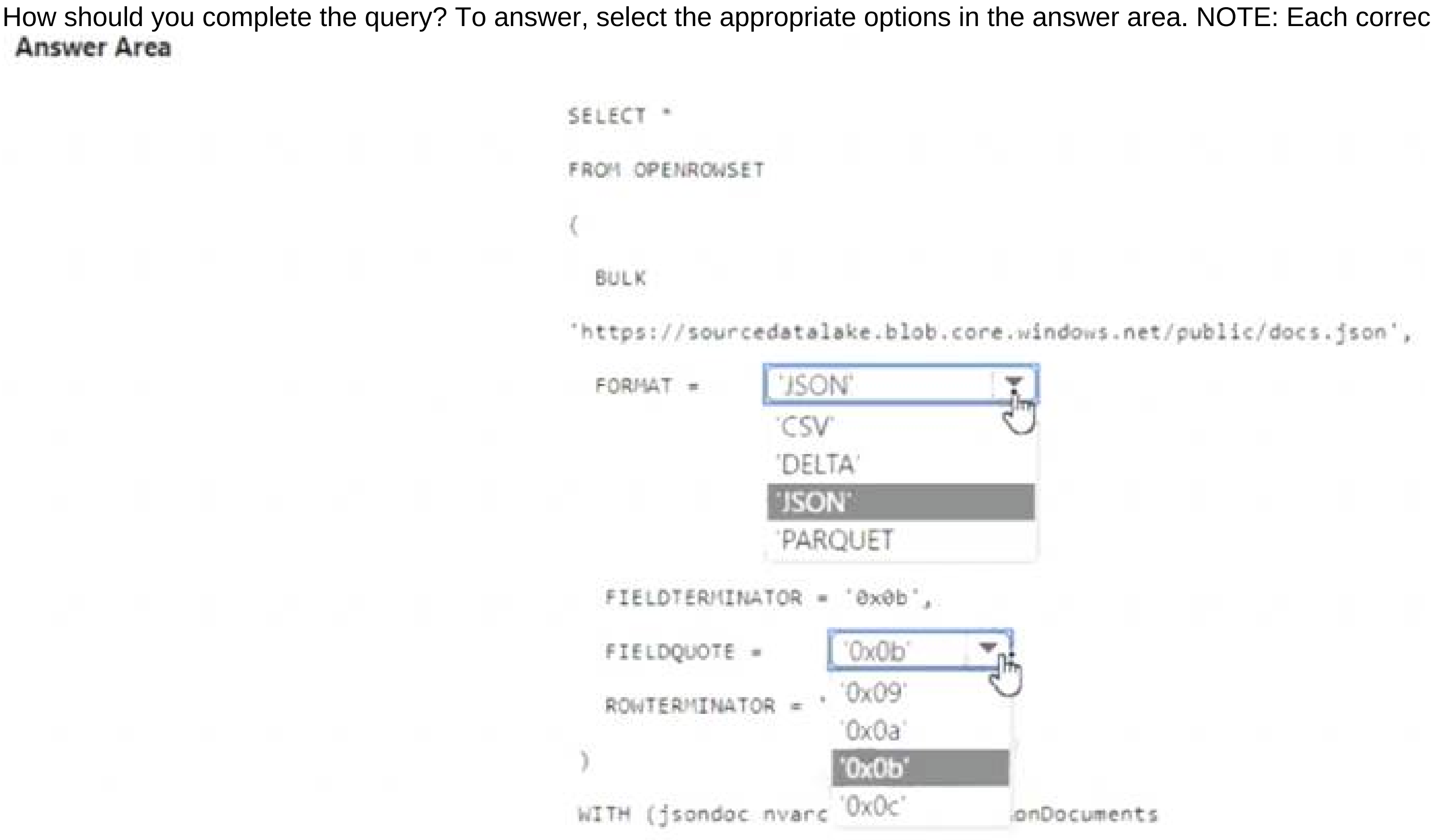
What should you recommend?

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

Answer: D

Explanation:
Parquet, an open-source file format for Hadoop, stores nested data structures in a flat columnar format. Compared to a traditional approach where data is stored in a row-oriented approach, Parquet file format is
more efficient in terms of storage and performance.
It is especially good for queries that read particular columns from a “wide” (with many columns) table since only needed columns are read, and IO is minimized.
Reference: https://www.clairvoyant.ai/blog/big-data-file-formats

NEW QUESTION 237
- (Exam Topic 3)
You have an Azure Synapse serverless SQL pool.
You need to read JSON documents from a file by using the OPENROWSET function.
How should you complete the query? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

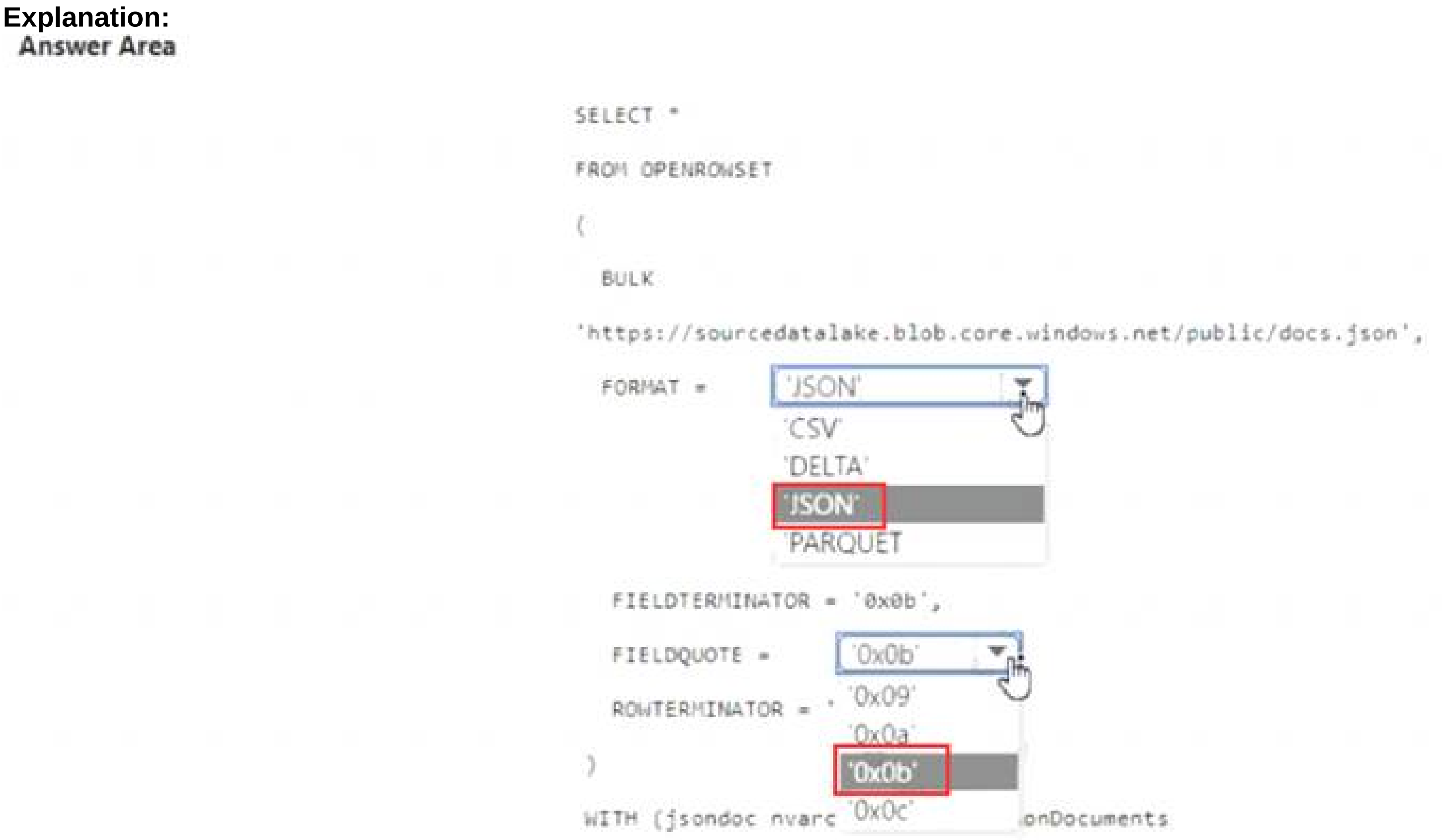

Answer Area

```
SELECT *
FROM OPENROWSET
(
  BULK
'https://sourcedatalake.blob.core.windows.net/public/docs.json',
  FORMAT =            'JSON'
                      'CSV'
                      'DELTA'
                      'JSON'
                      'PARQUET'
   FIELDTERMINATOR = '0x0b',
   FIELDQUOTE =             '0x0b'
   ROWTERMINATOR = '        '0x09'
                            '0x0a'
 )                          '0x0b'
 WITH (jsondoc nvarc        '0x0c'          onDocuments
```

A. Mastered
B. Not Mastered

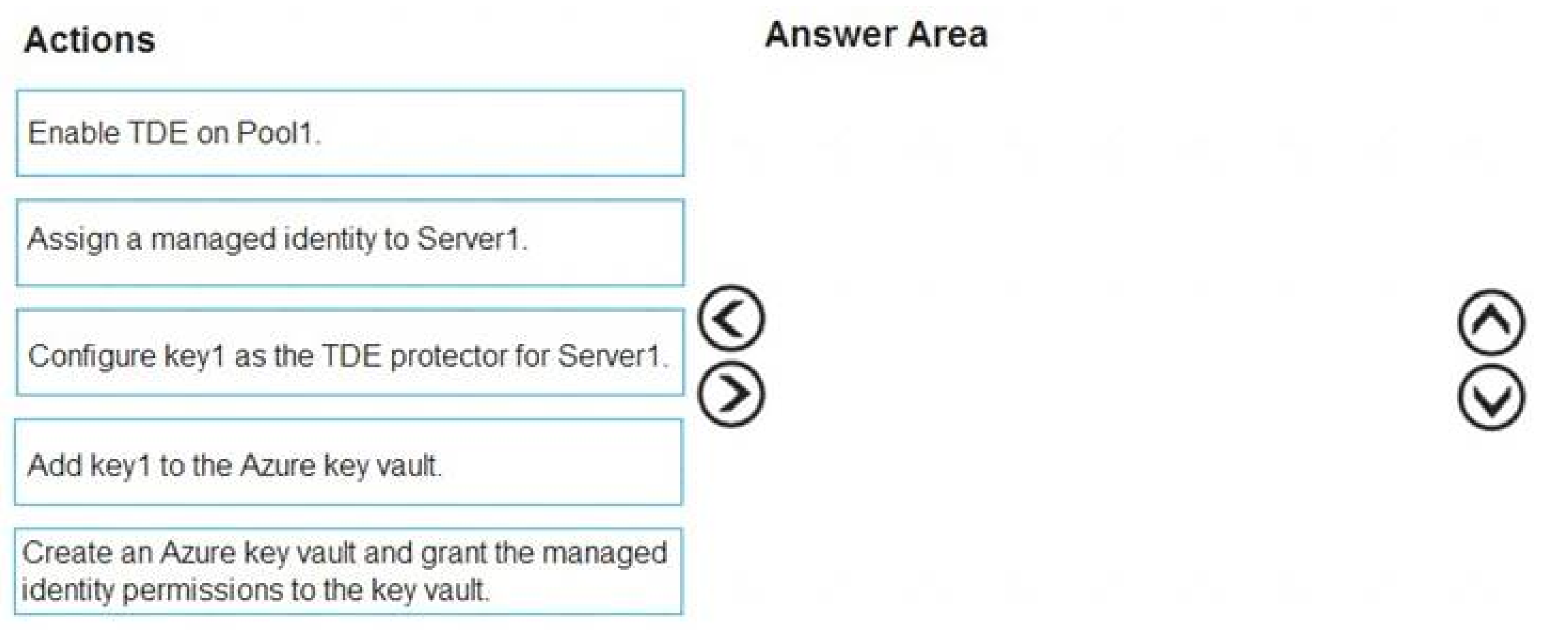
Answer: A

Explanation:


Answer Area

```
SELECT *
FROM OPENROWSET
(
  BULK
'https://sourcedatalake.blob.core.windows.net/public/docs.json',
  FORMAT =            'JSON'
                      'CSV'
                      'DELTA'
                      'JSON'
                      'PARQUET'
   FIELDTERMINATOR = '0x0b',
   FIELDQUOTE =             '0x0b'
   ROWTERMINATOR = '        '0x09'
                            '0x0a'
 )                          '0x0b'
 WITH (jsondoc nvarc        '0x0c'          onDocuments
```

NEW QUESTION 239
- (Exam Topic 3)
You have an Azure Synapse Analytics SQL pool named Pool1 on a logical Microsoft SQL server named Server1.
You need to implement Transparent Data Encryption (TDE) on Pool1 by using a custom key named key1. Which five actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



Actions

Enable TDE on Pool1.

Assign a managed identity to Server1.

Configure key1 as the TDE protector for Server1.

Add key1 to the Azure key vault.

Create an Azure key vault and grant the managed identity permissions to the key vault.

Answer Area

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated
Step 1: Assign a managed identity to Server1
You will need an existing Managed Instance as a prerequisite.
Step 2: Create an Azure key vault and grant the managed identity permissions to the vault Create Resource and setup Azure Key Vault.
Step 3: Add key1 to the Azure key vault
The recommended way is to import an existing key from a .pfx file or get an existing key from the vault. Alternatively, generate a new key directly in Azure Key Vault.
Step 4: Configure key1 as the TDE protector for Server1 Provide TDE Protector key
Step 5: Enable TDE on Pool1 Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/managed-instance/scripts/transparent-data-encryption-byok-po

NEW QUESTION 243
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pod.
You need to create a pipeline that will execute a stored procedure in the dedicated SQL pool and use the returned result set as the input (or a downstream activity. The solution must minimize development effort.
Which Type of activity should you use in the pipeline?

A. Notebook
B. U-SQL
C. Script
D. Stored Procedure

Answer: D

NEW QUESTION 244
- (Exam Topic 3)
You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container.
Which resource provider should you enable?

A. Microsoft.Sql
B. Microsoft-Automation
C. Microsoft.EventGrid
D. Microsoft.EventHub

Answer: C

Explanation:
Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure Blob Storage account. Data Factory natively integrates with Azure Event Grid, which lets you trigger pipelines on such events.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers

NEW QUESTION 249
- (Exam Topic 3)
You have an Azure Databricks workspace that contains a Delta Lake dimension table named Tablet. Table1 is a Type 2 slowly changing dimension (SCD) table. You need to apply updates from a source table to Table1. Which Apache Spark SQL operation should you use?

A. CREATE
B. UPDATE
C. MERGE
D. ALTER

Answer: C

Explanation:
The Delta provides the ability to infer the schema for data input which further reduces the effort required in managing the schema changes. The Slowly Changing Data(SCD) Type 2 records all the changes made to each key in the dimensional table. These operations require updating the existing rows to mark the previous values of the keys as old and then inserting new rows as the latest values. Also, Given a source table with the updates and the target table with dimensional data, SCD Type 2 can be expressed with the merge.
Example:

```
// Implementing SCD Type 2 operation using merge function customersTable
as("customers") merge(
stagedUpdates.as("staged_updates"), "customers.customerId = mergeKey")
whenMatched("customers.current = true AND customers.address <> staged_updates.address") updateExpr(Map(
"current" -> "false",
"endDate" -> "staged_updates.effectiveDate")) whenNotMatched()
insertExpr(Map(
"customerid" -> "staged_updates.customerId", "address" -> "staged_updates.address", "current" -> "true",
"effectiveDate" -> "staged_updates.effectiveDate",
"endDate" -> "null")) execute()
}
```

Reference:
https://www.projectpro.io/recipes/what-is-slowly-changing-data-scd-type-2-operation-delta-table-databricks

NEW QUESTION 252
- (Exam Topic 3)
You are incrementally loading data into fact tables in an Azure Synapse Analytics dedicated SQL pool. Each batch of incoming data is staged before being loaded into the fact tables. |
You need to ensure that the incoming data is staged as quickly as possible. |
How should you configure the staging tables? To answer, select the appropriate options in the answer area.

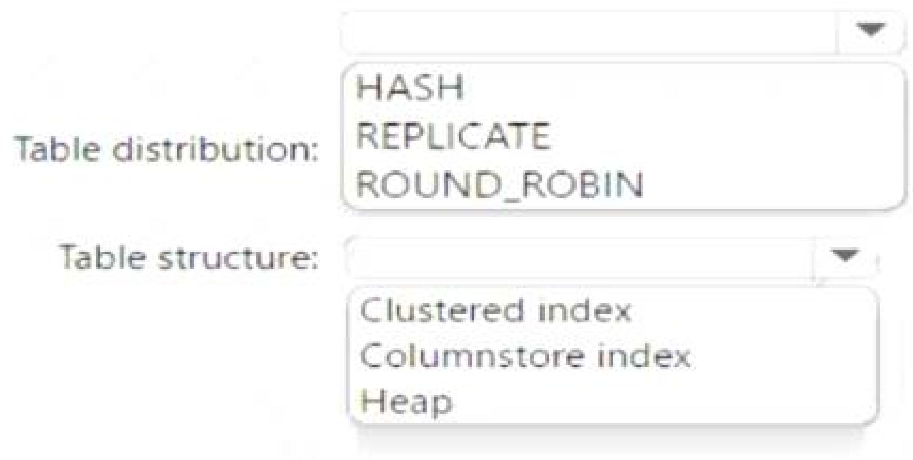


Table distribution:
HASH
REPLICATE
ROUND_ROBIN

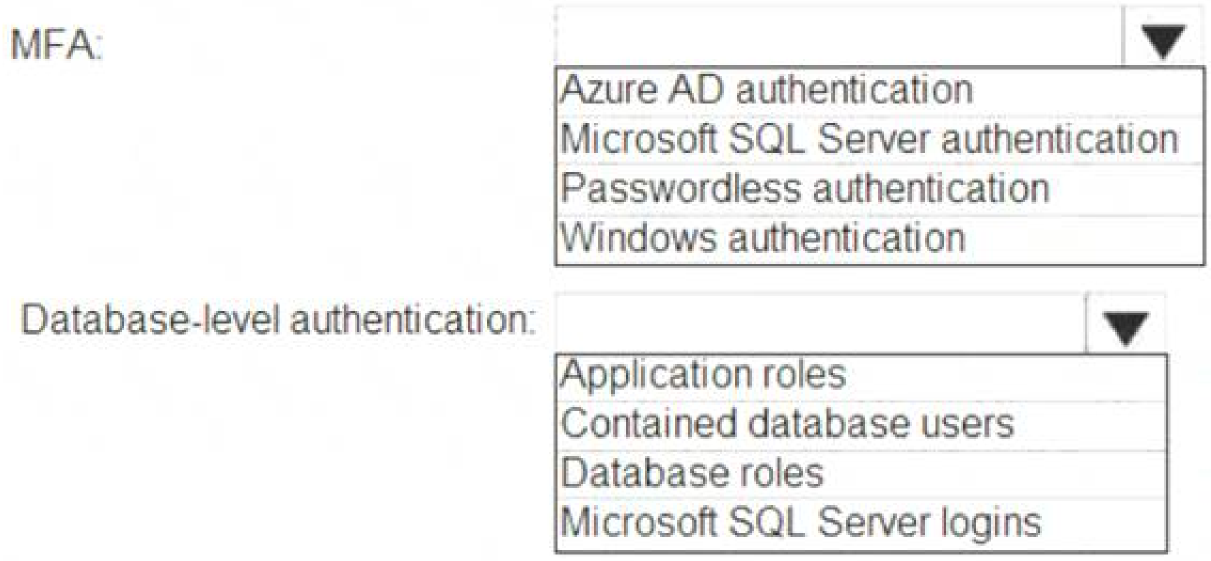
Table structure:
Clustered index
Columnstore index
Heap

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Round-robin distribution is recommended for staging tables because it distributes data evenly across all the distributions without requiring a hash column. This can improve the speed of data loading and avoid data skew. Heap tables are recommended for staging tables because they do not have any indexes or partitions that can slow down the data loading process. Heap tables are also easier to truncate and reload than clustered index or columnstore index tables.

NEW QUESTION 257
- (Exam Topic 3)
You have an Azure subscription that is linked to a hybrid Azure Active Directory (Azure AD) tenant. The subscription contains an Azure Synapse Analytics SQL pool named Pool1.
You need to recommend an authentication solution for Pool1. The solution must support multi-factor authentication (MFA) and database-level authentication.
Which authentication solution or solutions should you include in the recommendation? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



MFA:
Azure AD authentication
Microsoft SQL Server authentication
Passwordless authentication
Windows authentication

Database-level authentication:
Application roles
Contained database users
Database roles
Microsoft SQL Server logins

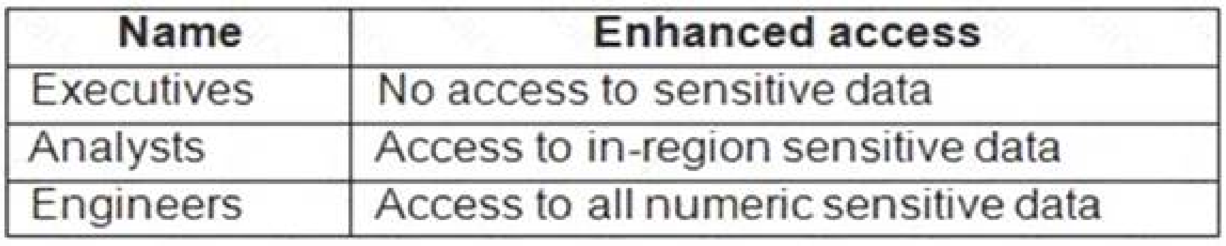
A. Mastered

B. Not Mastered

Answer: A

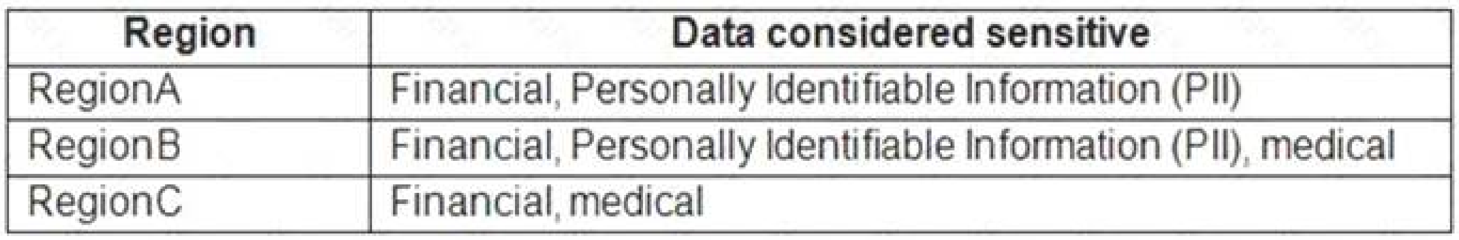
Explanation:
Graphical user interface, text, application, chat or text message Description automatically generated
Box 1: Azure AD authentication
Azure Active Directory authentication supports Multi-Factor authentication through Active Directory Universal Authentication.
Box 2: Contained database users
Azure Active Directory Uses contained database users to authenticate identities at the database level. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-authentication

NEW QUESTION 261
- (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.



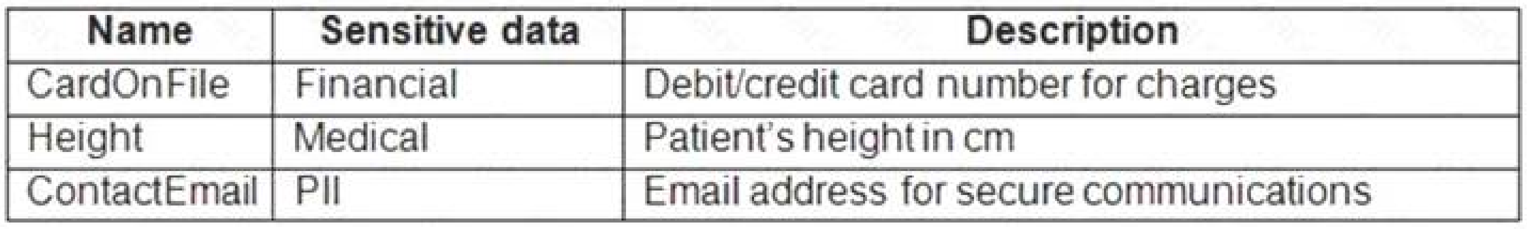
Name	Enhanced access
Executives	No access to sensitive data
Analysts	Access to in-region sensitive data
Engineers	Access to all numeric sensitive data

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



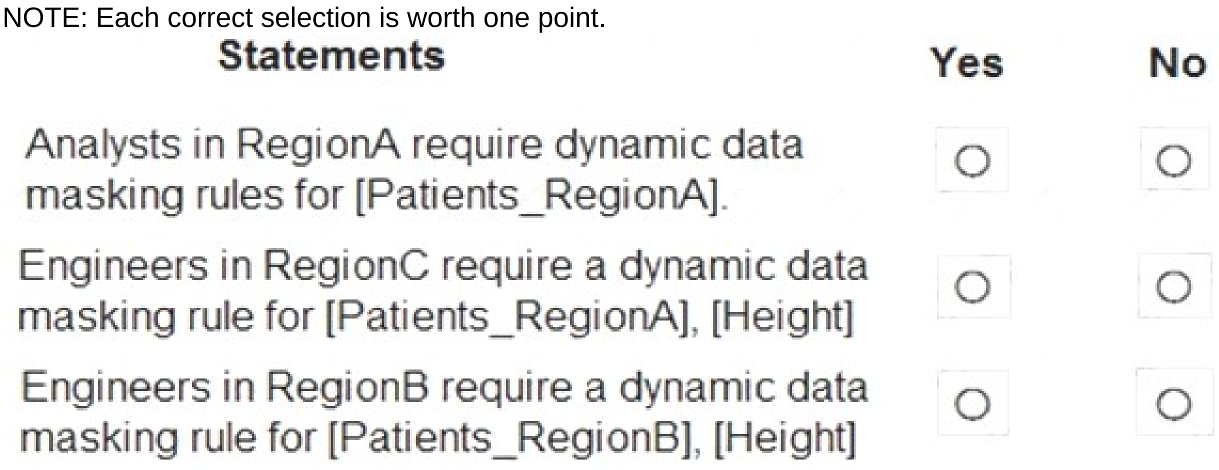
Region	Data considered sensitive
RegionA	Financial, Personally Identifiable Information (PII)
RegionB	Financial, Personally Identifiable Information (PII), medical
RegionC	Financial, medical

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.



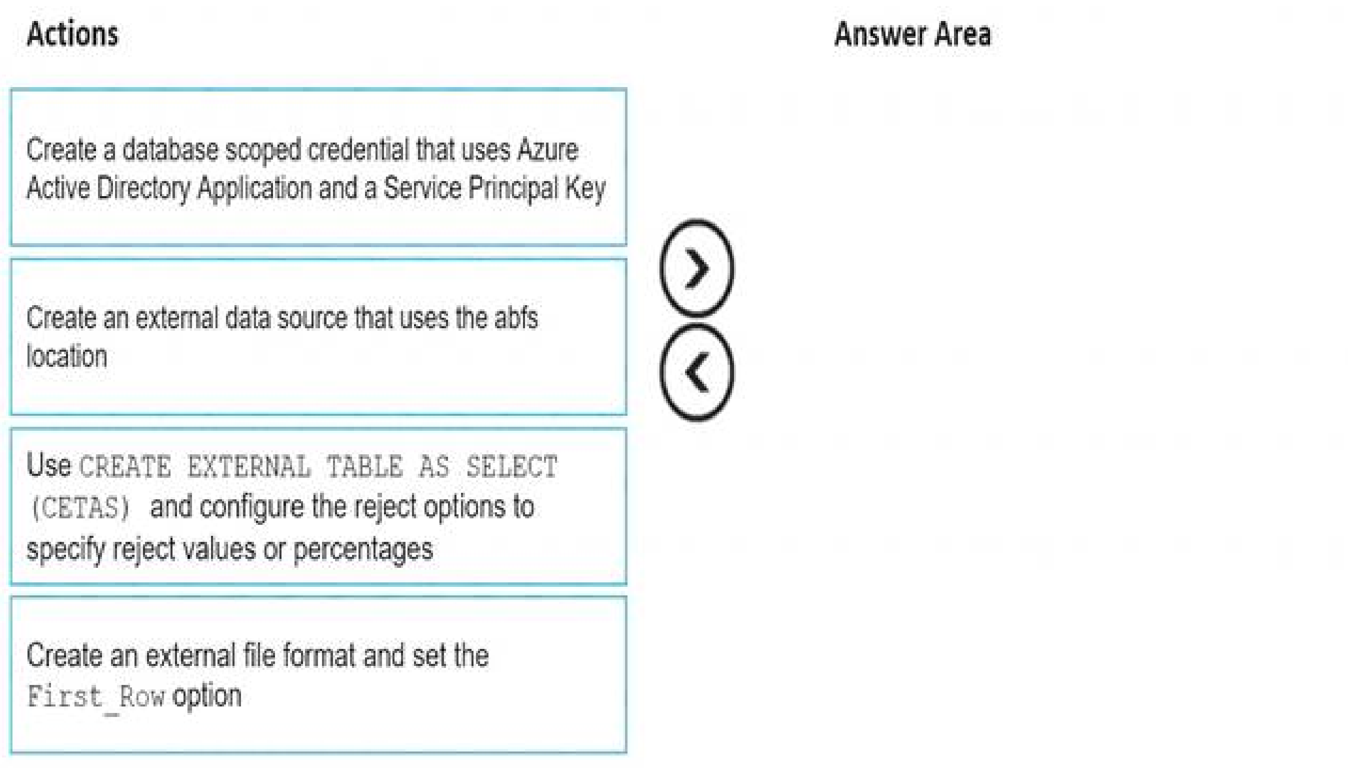
Statements	Yes	No
Analysts in RegionA require dynamic data masking rules for [Patients_RegionA].	○	○
Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height]	○	○
Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height]	○	○

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 263
- (Exam Topic 3)
You have data stored in thousands of CSV files in Azure Data Lake Storage Gen2. Each file has a header row followed by a properly formatted carriage return (/r) and line feed (/n).
You are implementing a pattern that batch loads the files daily into an enterprise data warehouse in Azure Synapse Analytics by using PolyBase.
You need to skip the header row when you import the files into the data warehouse. Before building the loading pattern, you need to prepare the required database objects in Azure Synapse Analytics.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
NOTE: Each correct selection is worth one point



Actions

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

Create an external data source that uses the abfs location

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

Create an external file format and set the First_Row option

Answer Area

A. Mastered
B. Not Mastered

Answer: A

Explanation:
A picture containing timeline Description automatically generated
Step 1: Create an external data source that uses the abfs location
Create External Data Source to reference Azure Data Lake Store Gen 1 or 2 Step 2: Create an external file format and set the First_Row option.
Create External File Format.
Step 3: Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages
To use PolyBase, you must create external tables to reference your external data. Use reject options.
Note: REJECT options don't apply at the time this CREATE EXTERNAL TABLE AS SELECT statement is run. Instead, they're specified here so that the database can use them at a later time when it imports data from the external table. Later, when the CREATE TABLE AS SELECT statement selects data from the external table, the database will use the reject options to determine the number or percentage of rows that can fail to import before it stops the import.
Reference:
https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-as-select-transact-sql

NEW QUESTION 264
- (Exam Topic 3)
You have a SQL pool in Azure Synapse.
You discover that some queries fail or take a long time to complete. You need to monitor for transactions that have rolled back.
Which dynamic management view should you query?

A. sys.dm_pdw_request_steps
B. sys.dm_pdw_nodes_tran_database_transactions
C. sys.dm_pdw_waits
D. sys.dm_pdw_exec_sessions

Answer: B

Explanation:
You can use Dynamic Management Views (DMVs) to monitor your workload including investigating query execution in SQL pool.
If your queries are failing or taking a long time to proceed, you can check and monitor if you have any transactions rolling back.
Example:
-- Monitor rollback SELECT
SUM(CASE WHEN t.database_transaction_next_undo_lsn IS NOT NULL THEN 1 ELSE 0 END), t.pdw_node_id,
nod.[type]
FROM sys.dm_pdw_nodes_tran_database_transactions t
JOIN sys.dm_pdw_nodes nod ON t.pdw_node_id = nod.pdw_node_id GROUP BY t.pdw_node_id, nod.[type]
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monit

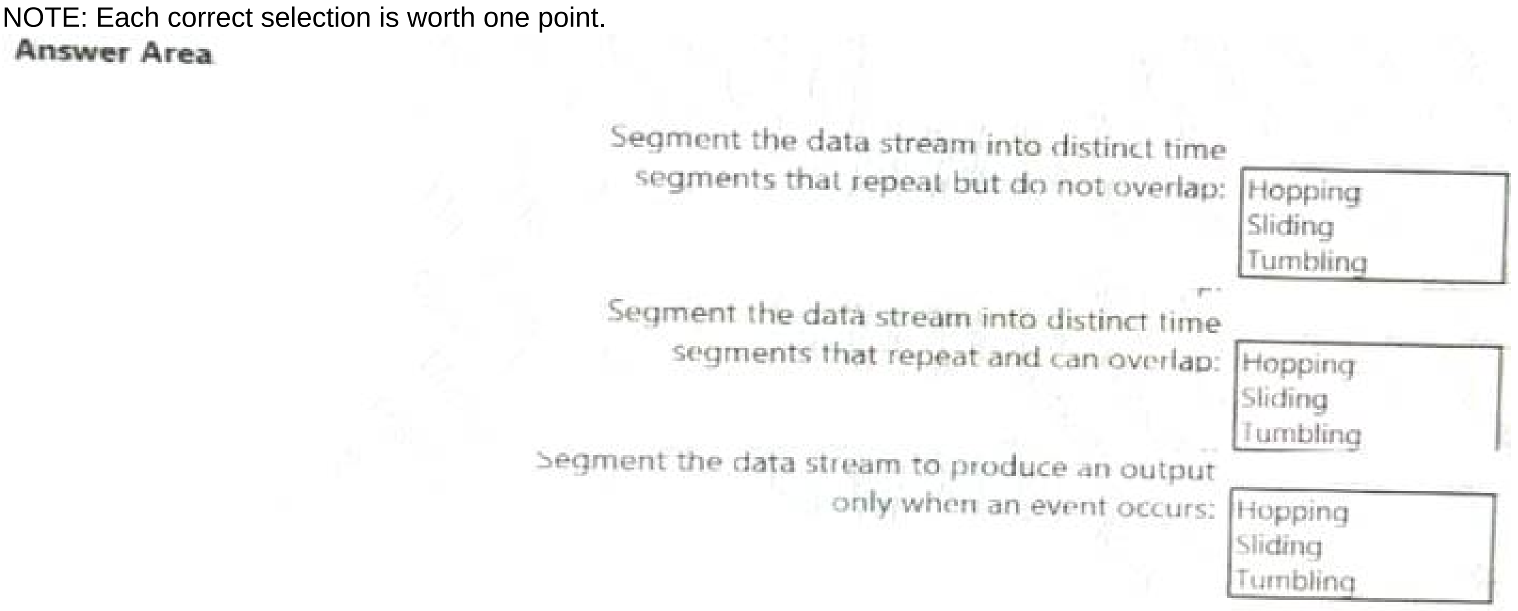
NEW QUESTION 269
- (Exam Topic 3)
You are designing an enterprise data warehouse in Azure Synapse Analytics that will contain a table named Customers. Customers will contain credit card information.
You need to recommend a solution to provide salespeople with the ability to view all the entries in Customers. 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 271
- (Exam Topic 3)
You are implementing Azure Stream Analytics windowing functions.
Which windowing function should you use for each requirement? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer Area

Segment the data stream into distinct time segments that repeat but do not overlap: Hopping
Sliding
Tumbling

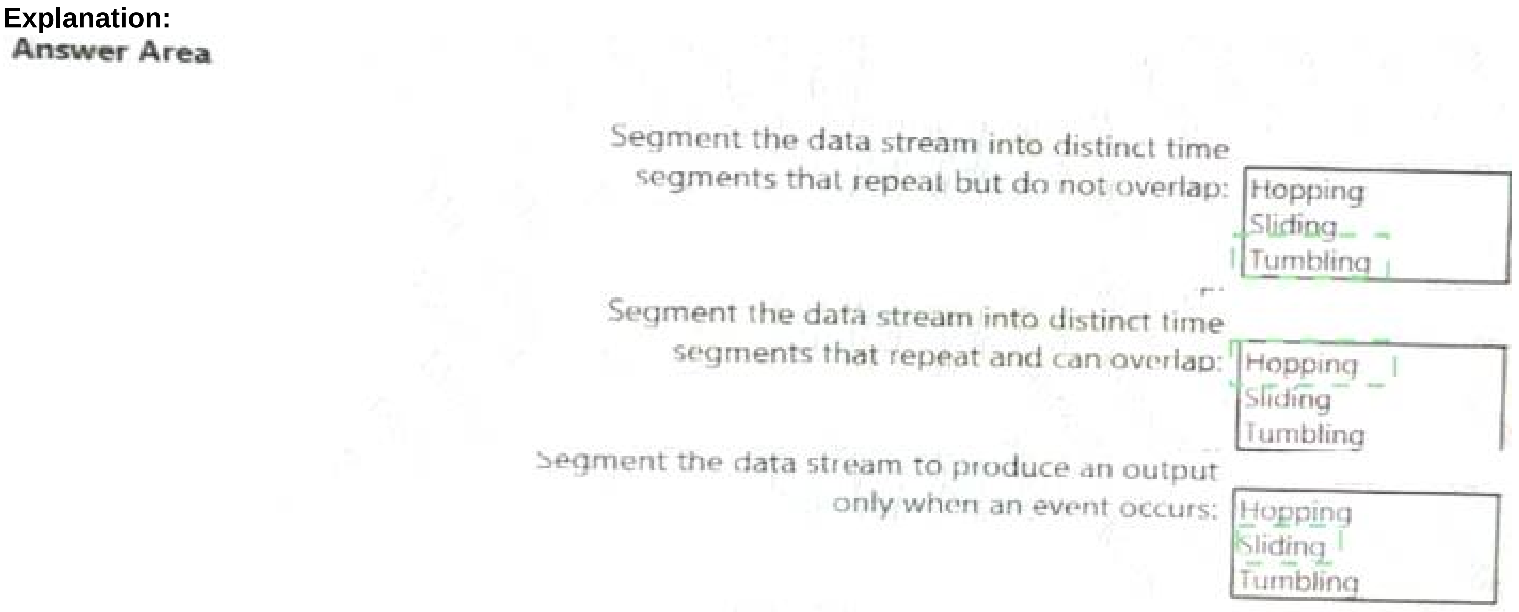
Segment the data stream into distinct time segments that repeat and can overlap: Hopping
Sliding
Tumbling

Segment the data stream to produce an output only when an event occurs: Hopping
Sliding
Tumbling

A. Mastered
B. Not Mastered

Answer: A

Explanation:



Answer Area

Segment the data stream into distinct time segments that repeat but do not overlap: Hopping
Sliding
[✓ Tumbling]

Segment the data stream into distinct time segments that repeat and can overlap: [✓ Hopping]
Sliding
Tumbling

Segment the data stream to produce an output only when an event occurs: Hopping
[✓ Sliding]
Tumbling

NEW QUESTION 275
- (Exam Topic 3)
A company uses Azure Stream Analytics to monitor devices.
The company plans to double the number of devices that are monitored.
You need to monitor a Stream Analytics job to ensure that there are enough processing resources to handle the additional load.
Which metric should you monitor?

A. Early Input Events
B. Late Input Events
C. Watermark delay
D. Input Deserialization Errors

Answer: A

Explanation:
There are a number of resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to:
- Not enough processing resources in Stream Analytics to handle the volume of input events.
- Not enough throughput within the input event brokers, so they are throttled.
- Output sinks are not provisioned with enough capacity, so they are throttled. The possible solutions vary widely based on the flavor of output service being used.

Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling

NEW QUESTION 279
- (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 Azure Data Factor Studio, run Publish All.
E. Create a Git repository
F. From the UX Authoring canvas, select Publish

Answer: DE

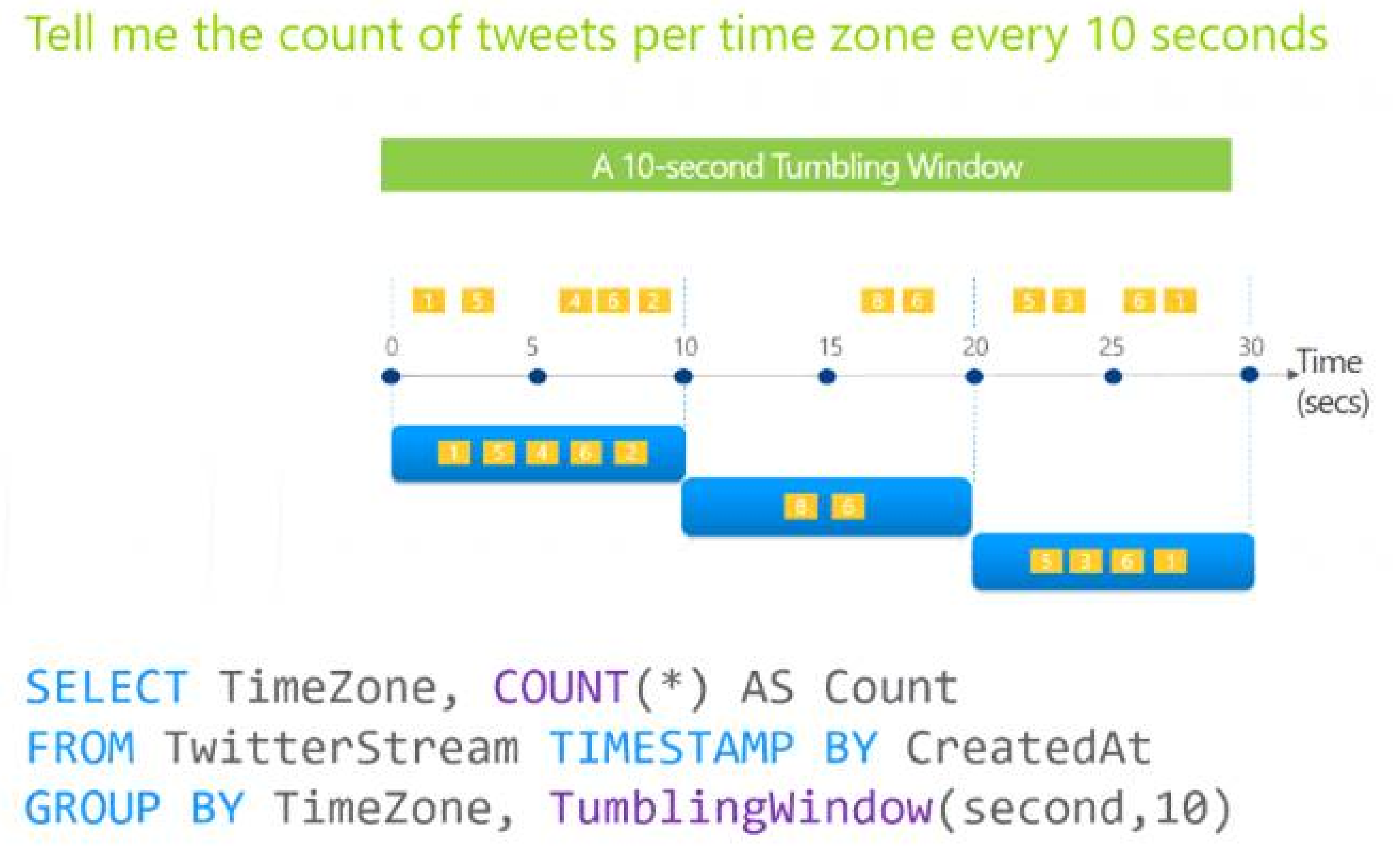
Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/source-control

NEW QUESTION 280
- (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 are designing an Azure Stream Analytics solution that will analyze Twitter data.
You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.
Solution: You use a tumbling window, and you set the window size to 10 seconds. Does this meet the goal?

A. Yes
B. No

Answer: A

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.

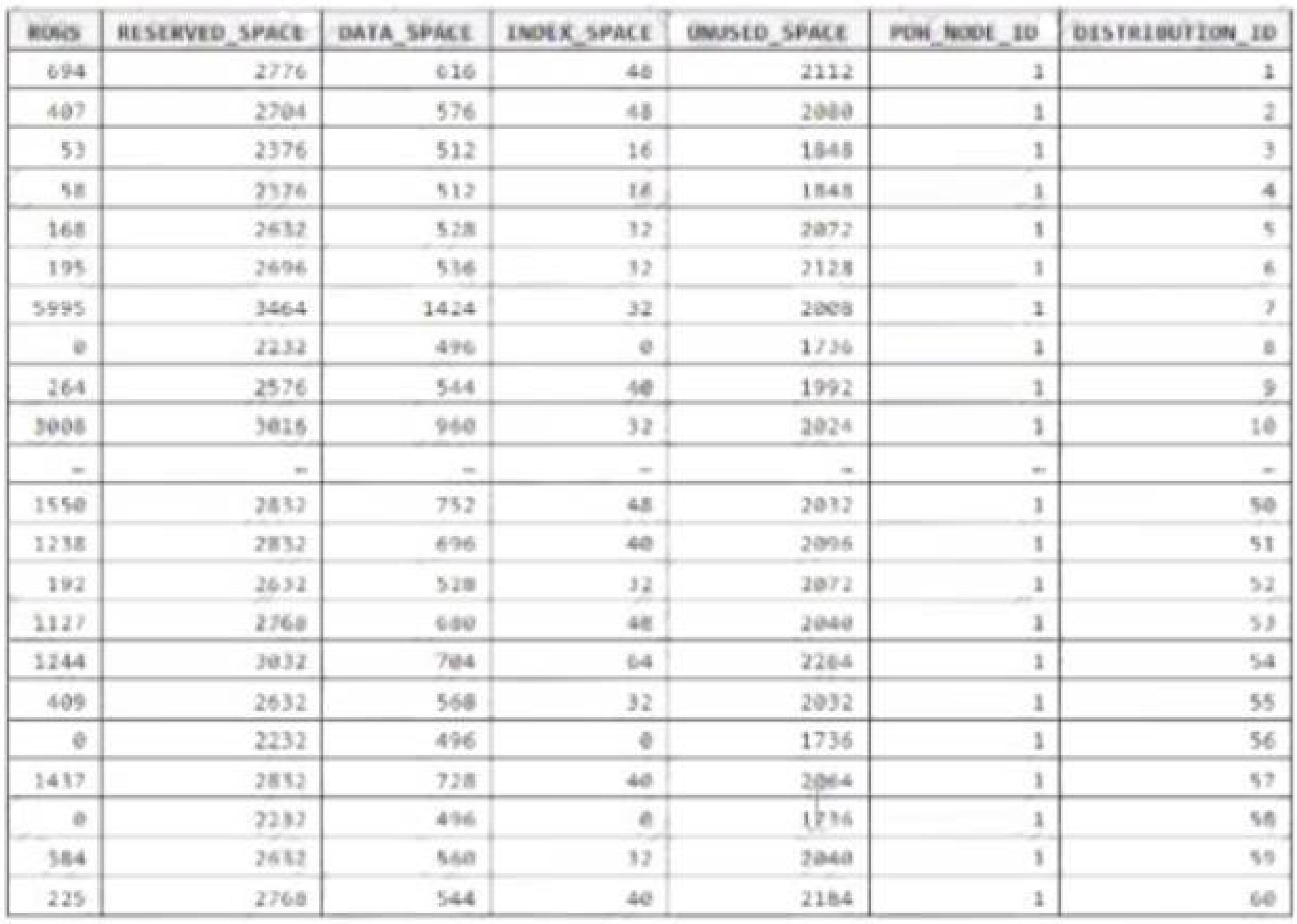


Tell me the count of tweets per time zone every 10 seconds

```
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 281
- (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	704	64	2264	1	54
409	2632	568	32	2032	1	55
0	2232	496	0	1736	1	56
1437	2832	728	40	2064	1	57
0	2232	496	0	1736	1	58
584	2632	560	32	2040	1	59
225	2768	544	40	2184	1	60

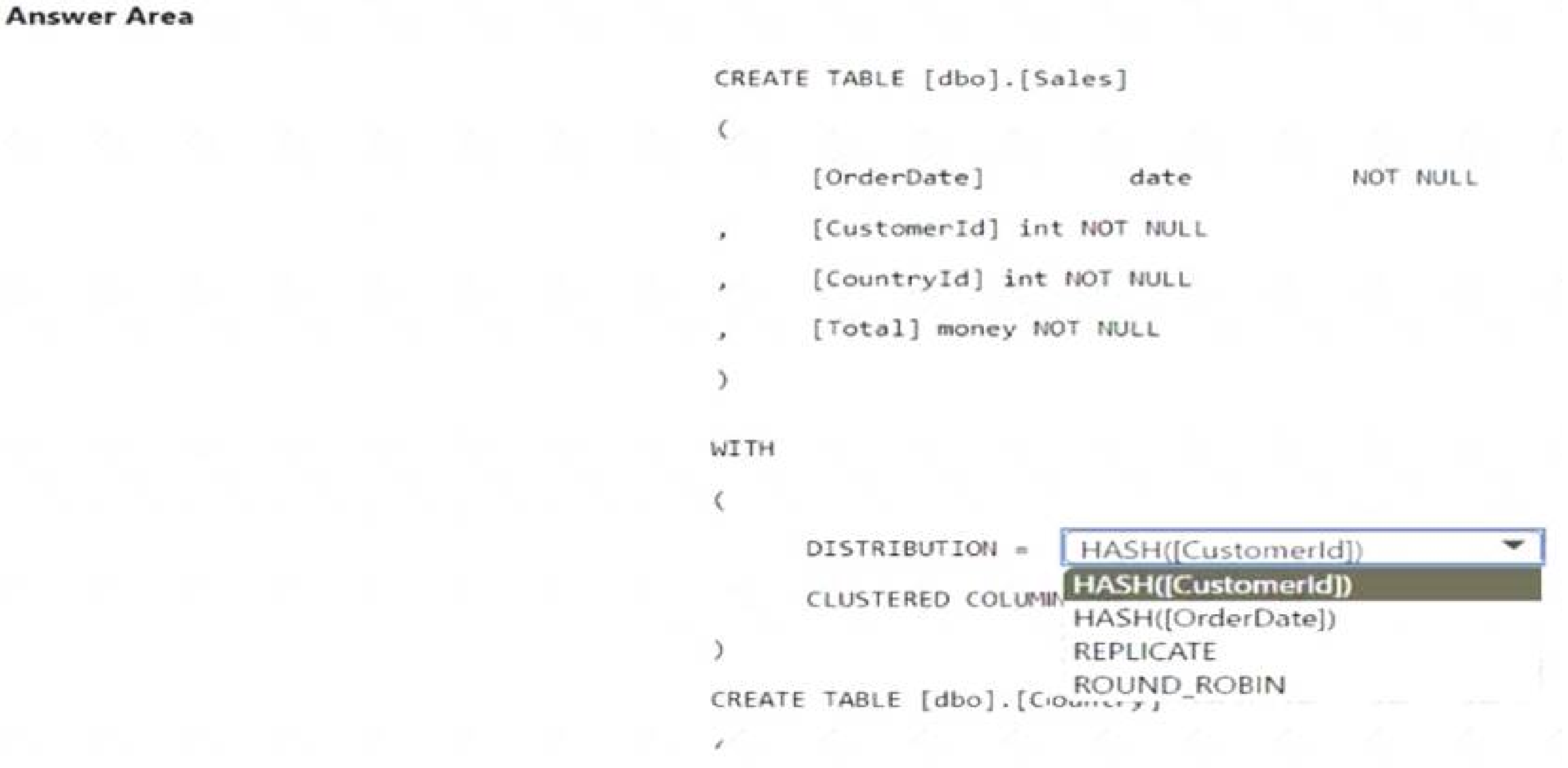
Which statement accurately describes the dbo,FactInternetSales table?

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: C

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 283
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool. You plan to deploy a solution that will analyze sales data and include the following:
• A table named Country that will contain 195 rows
• A table named Sales that will contain 100 million rows
• A query to identify total sales by country and customer from the past 30 days
You need to create the tables. The solution must maximize query performance.
How should you complete the script? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.



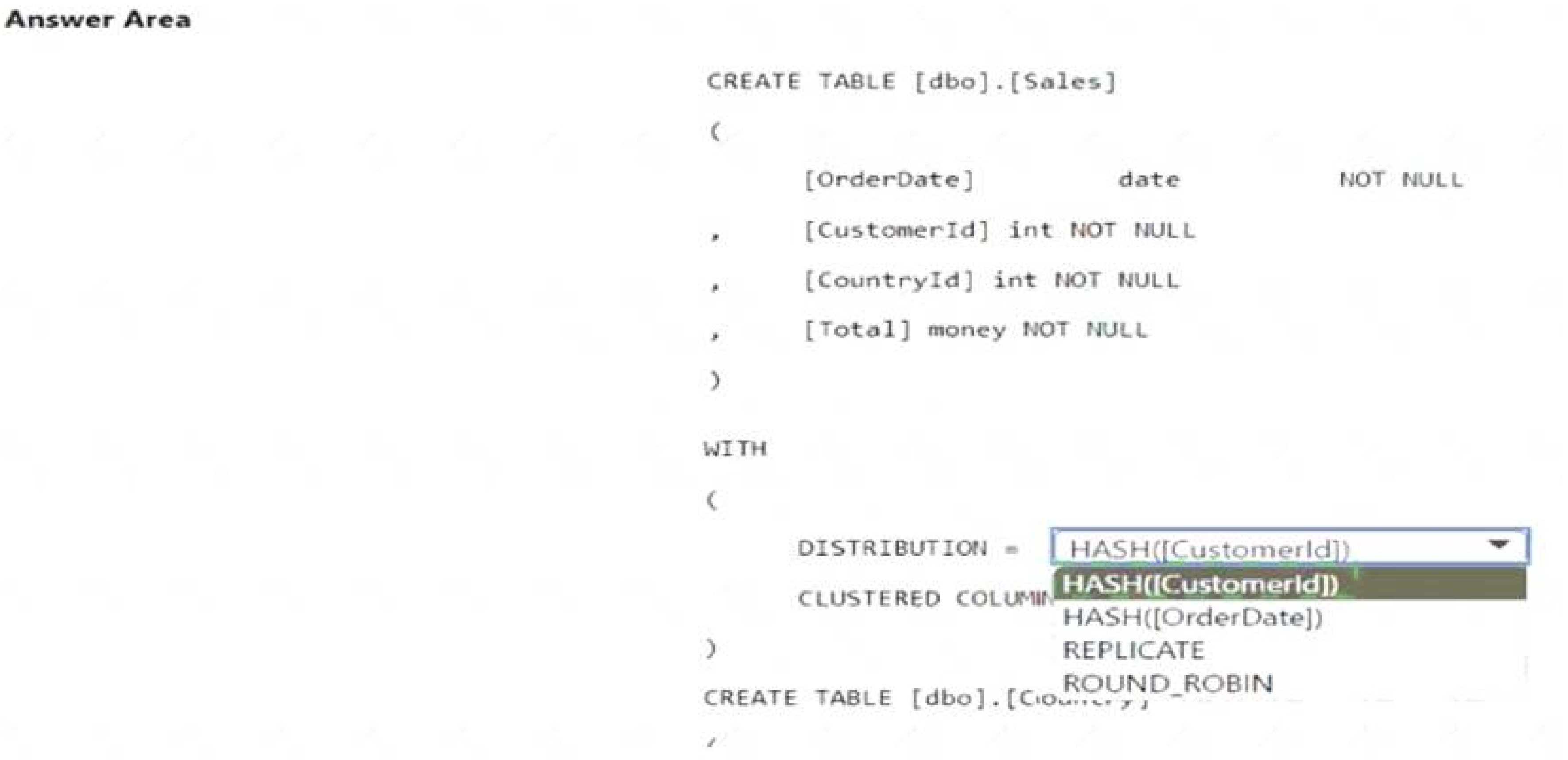
Answer Area

```
CREATE TABLE [dbo].[Sales]
(
        [OrderDate]          date            NOT NULL
,       [CustomerId] int NOT NULL
,       [CountryId] int NOT NULL
,       [Total] money NOT NULL
)
WITH
(
        DISTRIBUTION = HASH([CustomerId])
        CLUSTERED COLUMN
)
CREATE TABLE [dbo].[Co
,
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:



Answer Area

```
CREATE TABLE [dbo].[Sales]
(
        [OrderDate]          date          NOT NULL
,       [CustomerId] int NOT NULL
,       [CountryId] int NOT NULL
,       [Total] money NOT NULL
)
WITH
(
        DISTRIBUTION = HASH([CustomerId])
        CLUSTERED COLUMN
)
CREATE TABLE [dbo].[Cou
;
```

NEW QUESTION 286

- (Exam Topic 3)

You have two Azure SQL databases named DB1 and DB2.

DB1 contains a table named Table 1. Table1 contains a timestamp column named LastModifiedOn. LastModifiedOn contains the timestamp of the most recent update for each individual row.

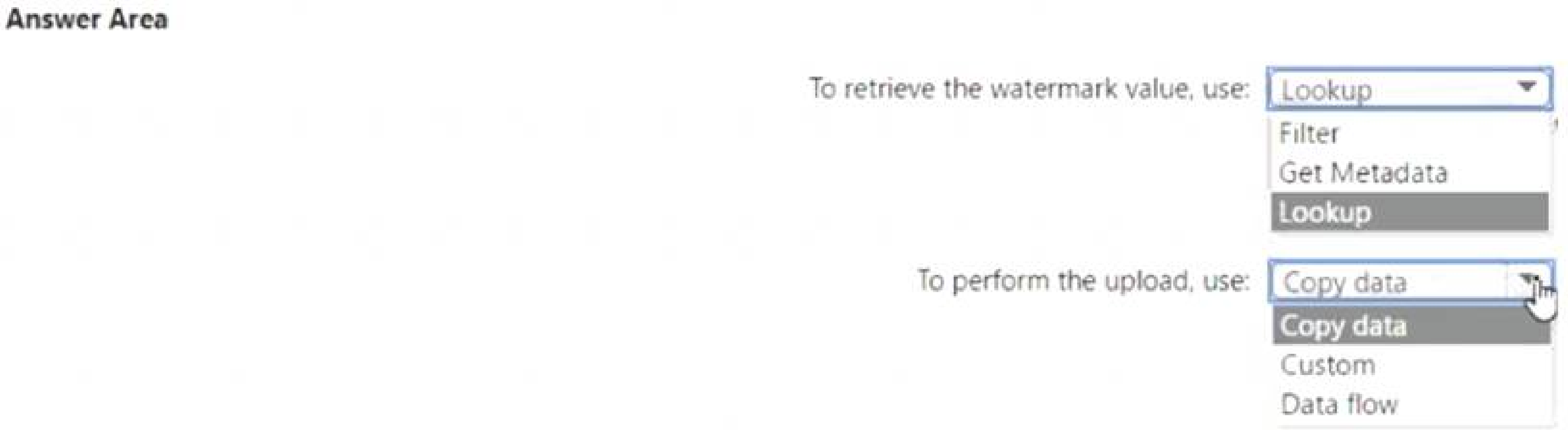
DB2 contains a table named Watermark. Watermark contains a single timestamp column named WatermarkValue.

You plan to create an Azure Data Factory pipeline that will incrementally upload into Azure Blob Storage all the rows in Table1 for which the LastModifiedOn column contains a timestamp newer than the most recent value of the WatermarkValue column in Watermark.

You need to identify which activities to include in the pipeline. The solution must meet the following requirements:

• Minimize the effort to author the pipeline.

• Ensure that the number of data integration units allocated to the upload operation can be controlled. What should you identify? To answer, select the appropriate options in the answer area.



Answer Area

To retrieve the watermark value, use: Lookup

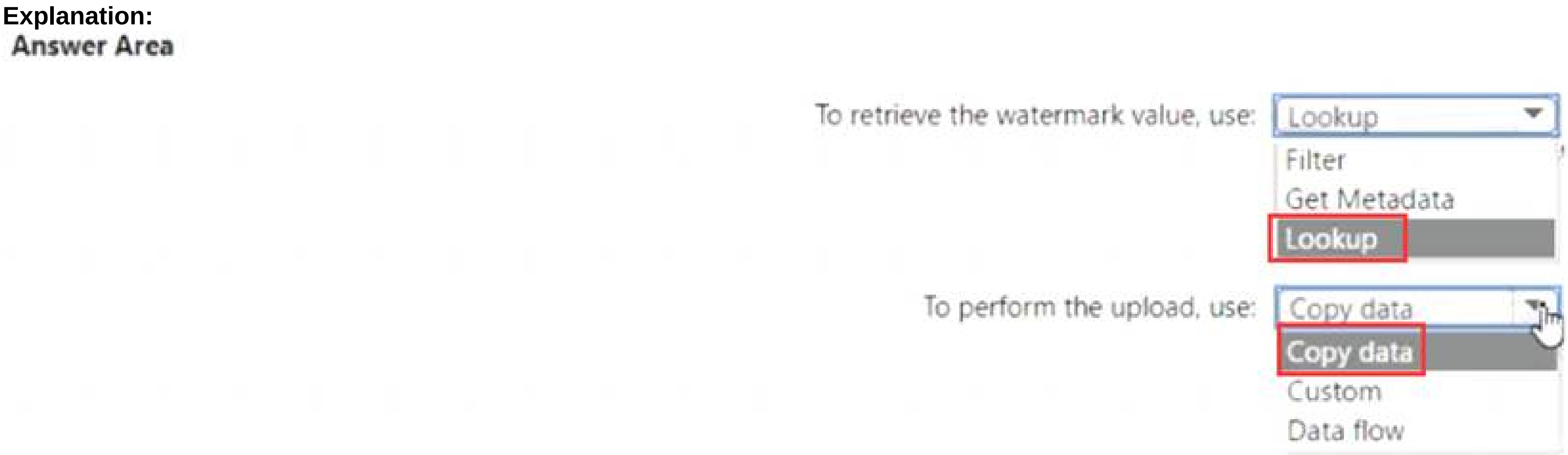
To perform the upload, use: Copy data

A. Mastered

B. Not Mastered

Answer: A

Explanation:



Answer Area

To retrieve the watermark value, use: Lookup

To perform the upload, use: Copy data

NEW QUESTION 287

- (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

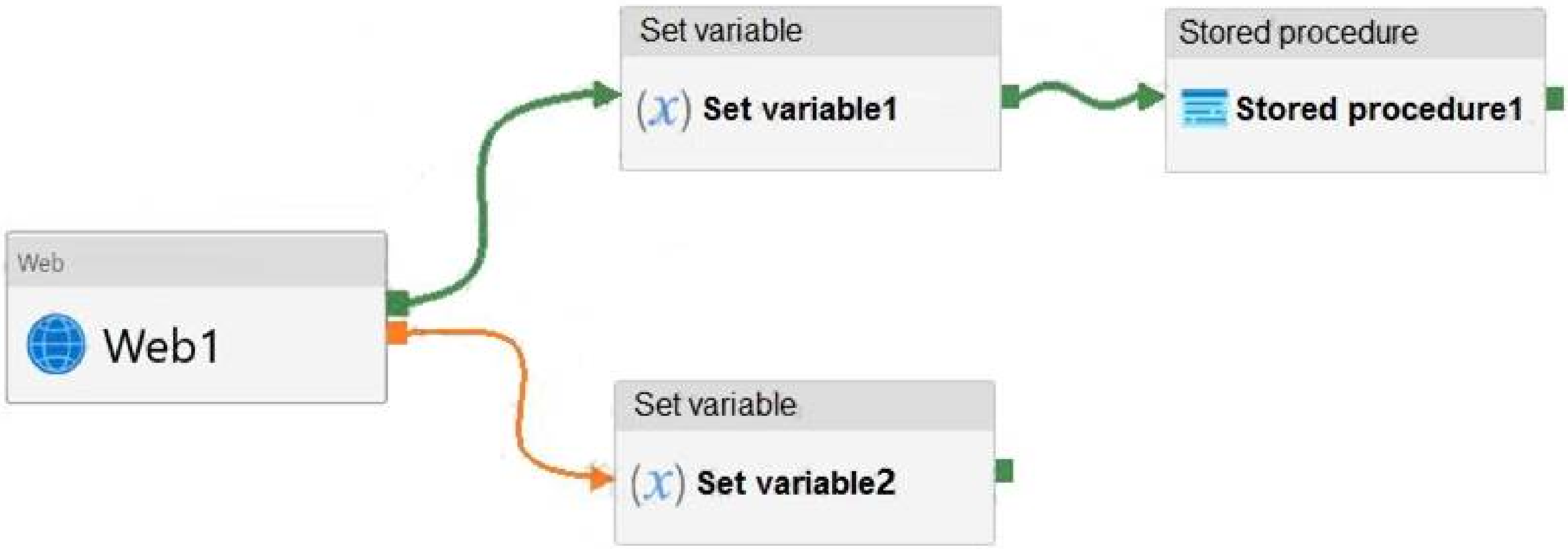
aXXX@XXXX.com instead.
What should you do?

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.

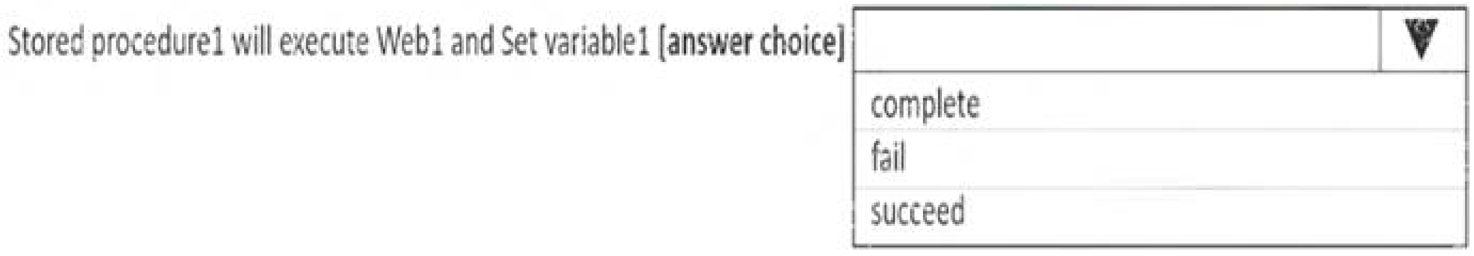
Answer: D

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

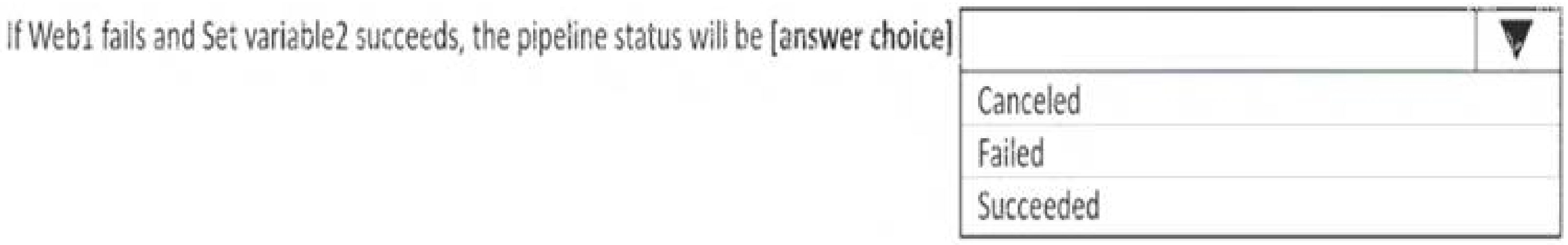
NEW QUESTION 289
- (Exam Topic 3)
You have an Azure Data Factory pipeline that has the activities shown in the following exhibit.



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


Stored procedure1 will execute Web1 and Set variable1 [answer choice]

complete
fail
succeed

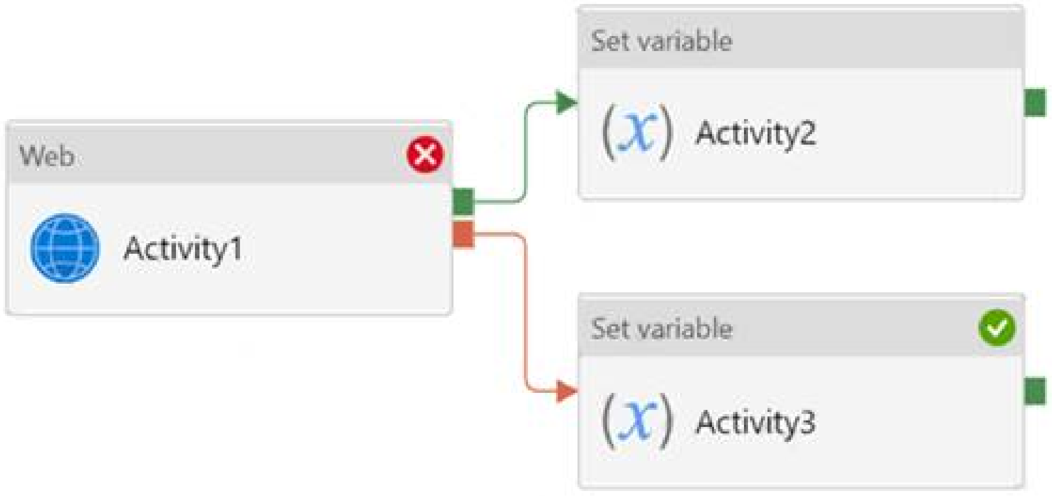

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]

Canceled
Failed
Succeeded

A. Mastered
B. Not Mastered

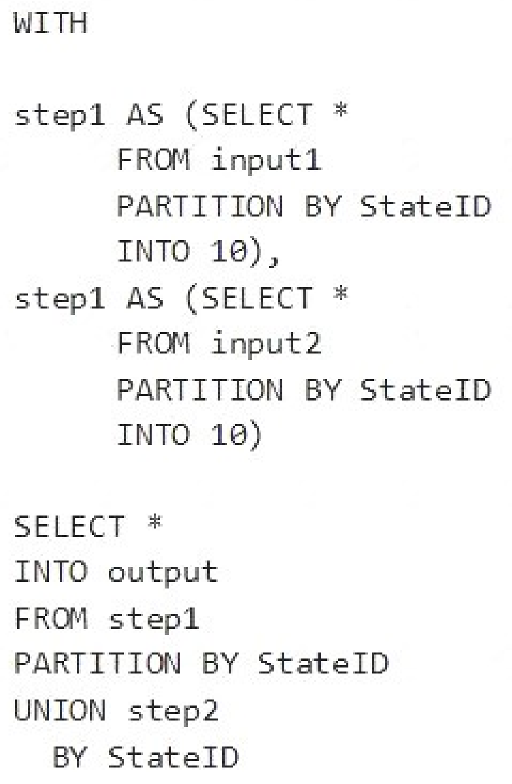
Answer: A

Explanation:
Box 1: succeed
Box 2: failed Example:
Now let's say we have a pipeline with 3 activities, where Activity1 has a success path to Activity2 and a failure path to Activity3. If Activity1 fails and Activity3 succeeds, the pipeline will fail. The presence of the success path alongside the failure path changes the outcome reported by the pipeline, even though the activity executions from the pipeline are the same as the previous scenario.



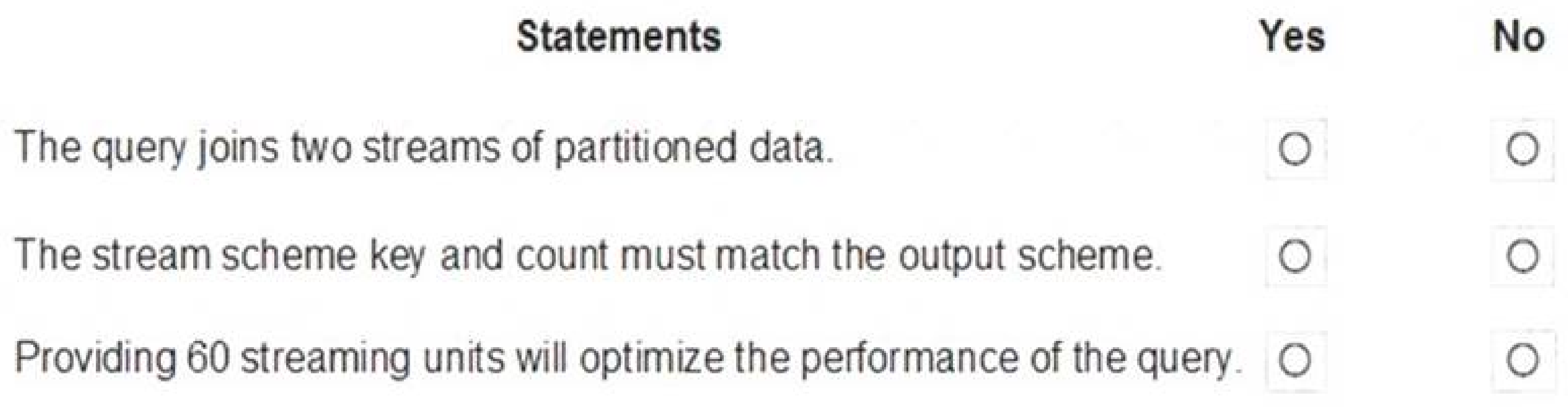
Activity1 fails, Activity2 is skipped, and Activity3 succeeds. The pipeline reports failure. Reference: https://datasavvy.me/2021/02/18/azure-data-factory-activity-failures-and-pipeline-outcomes/

NEW QUESTION 292
- (Exam Topic 3)
You have the following Azure Stream Analytics query.



```
WITH

step1 AS (SELECT *
      FROM input1
      PARTITION BY StateID
      INTO 10),
step1 AS (SELECT *
      FROM input2
      PARTITION BY StateID
      INTO 10)

SELECT *
INTO output
FROM step1
PARTITION BY StateID
UNION step2
  BY StateID
```

For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.



Statements	Yes	No
The query joins two streams of partitioned data.	○	○
The stream scheme key and count must match the output scheme.	○	○
Providing 60 streaming units will optimize the performance of the query.	○	○

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Yes
You can now use a new extension of Azure Stream Analytics SQL to specify the number of partitions of a stream when reshuffling the data.
The outcome is a stream that has the same partition scheme. Please see below for an example: WITH step1 AS (SELECT * FROM [input1] PARTITION BY DeviceID INTO 10),
step2 AS (SELECT * FROM [input2] PARTITION BY DeviceID INTO 10)
SELECT * INTO [output] FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceID Note: The new extension of Azure Stream Analytics SQL includes a keyword INTO that allows you to specify the number of partitions for a stream when performing reshuffling using a PARTITION BY statement.
Box 2: Yes
When joining two streams of data explicitly repartitioned, these streams must have the same partition key and partition count.
Box 3: Yes
10 partitions x six SUs = 60 SUs is fine.

Note: Remember, Streaming Unit (SU) count, which is the unit of scale for Azure Stream Analytics, must be adjusted so the number of physical resources available to the job can fit the partitioned flow. In general, six SUs is a good number to assign to each partition. In case there are insufficient resources assigned to the job, the system will only apply the repartition if it benefits the job.
Reference:
https://azure.microsoft.com/en-in/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/

NEW QUESTION 295
- (Exam Topic 3)
You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements:

- Can return an employee record from a given point in time.
- Maintains the latest employee information.
- Minimizes query complexity.

How should you model the employee data?

A. as a temporal table
B. as a SQL graph table
C. as a degenerate dimension table
D. as a Type 2 slowly changing dimension (SCD) table

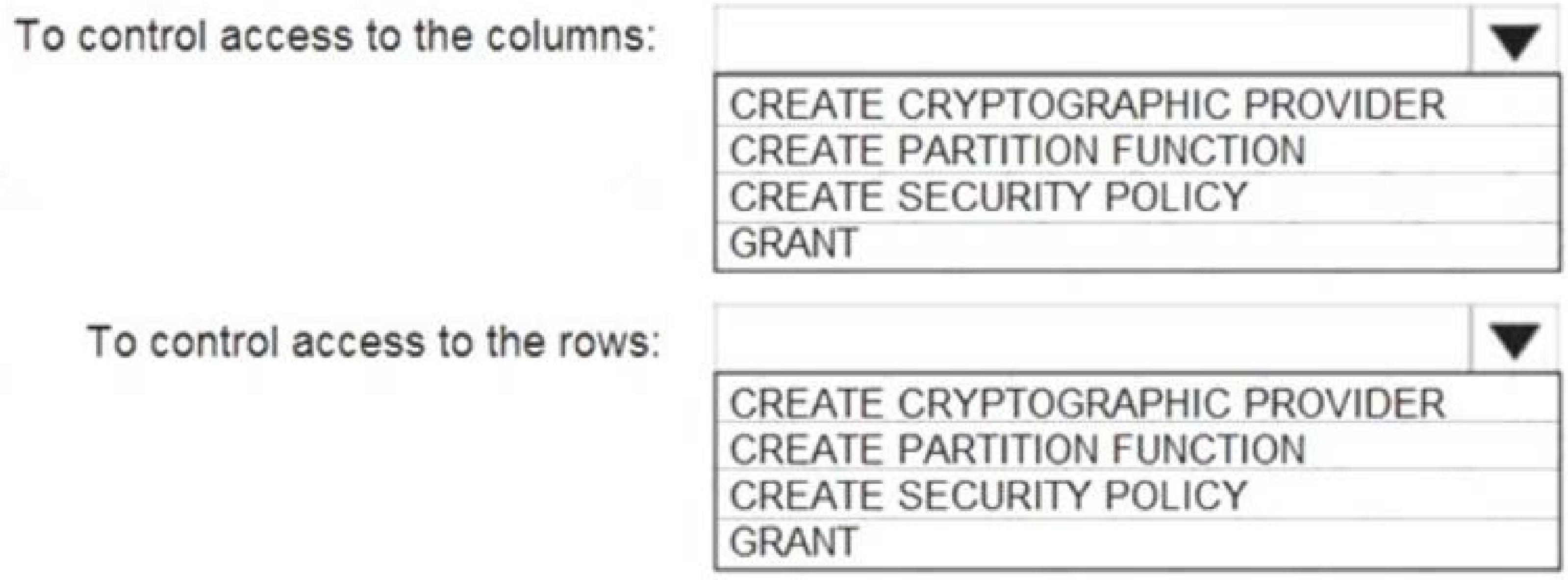
Answer: D

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

NEW QUESTION 297
- (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

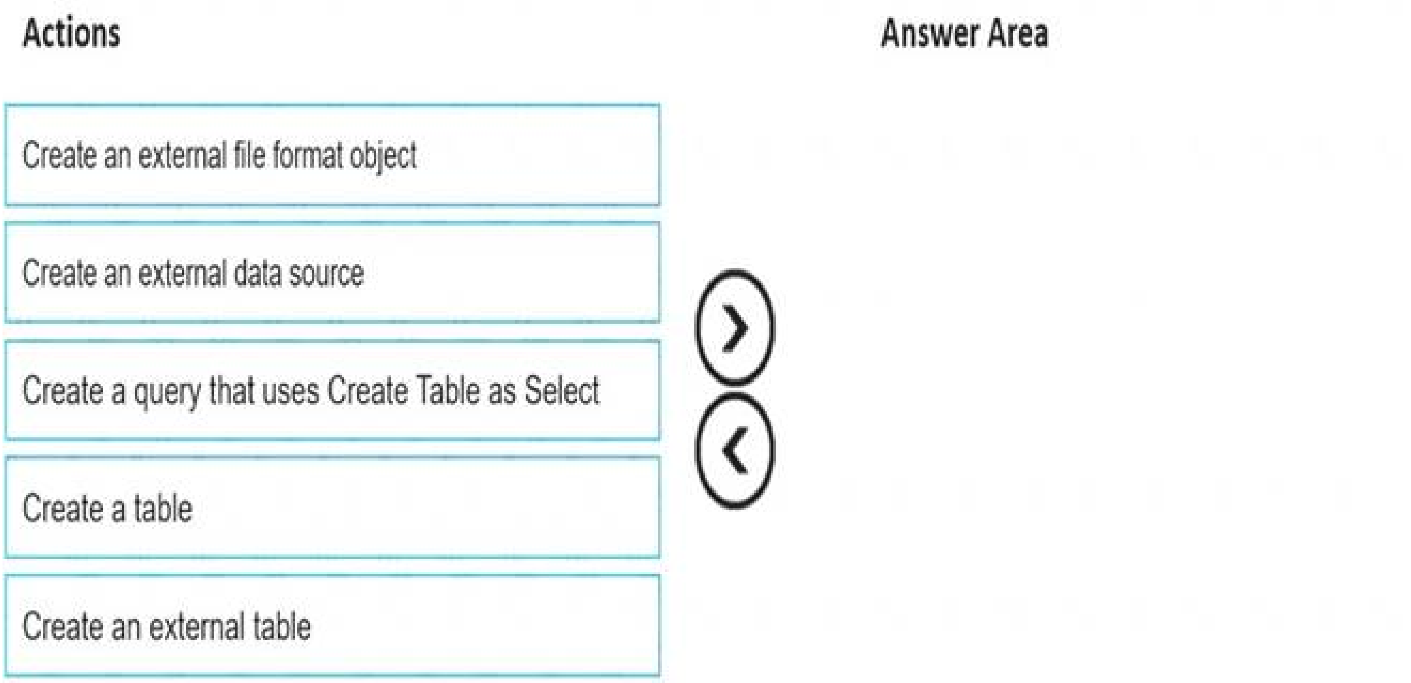
To control access to the rows:
- CREATE CRYPTOGRAPHIC PROVIDER
- CREATE PARTITION FUNCTION
- CREATE SECURITY POLICY
- GRANT

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 299
- (Exam Topic 3)
You need to build a solution to ensure that users can query specific files in an Azure Data Lake Storage Gen2 account from an Azure Synapse Analytics serverless SQL pool.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.



Actions

Create an external file format object

Create an external data source

Create a query that uses Create Table as Select

Create a table

Create an external table

Answer Area

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application, email Description automatically generated
Step 1: Create an external data source
You can create external tables in Synapse SQL pools via the following steps:

- CREATE EXTERNAL DATA SOURCE to reference an external Azure storage and specify the credential that should be used to access the storage.
- CREATE EXTERNAL FILE FORMAT to describe format of CSV or Parquet files.
- CREATE EXTERNAL TABLE on top of the files placed on the data source with the same file format. Step 2: Create an external file format object

Creating an external file format is a prerequisite for creating an external table. Step 3: Create an external table
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

NEW QUESTION 302
- (Exam Topic 3)
You are designing an Azure Databricks table. The table will ingest an average of 20 million streaming events per day.
You need to persist the events in the table for use in incremental load pipeline jobs in Azure Databricks. The solution must minimize storage costs and incremental load times.
What should you include in the solution?

A. Partition by DateTime fields.
B. Sink to Azure Queue storage.
C. Include a watermark column.
D. Use a JSON format for physical data storage.

Answer: A

Explanation:
The Databricks ABS-AQS connector uses Azure Queue Storage (AQS) to provide an optimized file source that lets you find new files written to an Azure Blob storage (ABS) container without repeatedly listing all of the files.
This provides two major advantages:

- Lower costs: no more costly LIST API requests made to ABS.

Reference:
https://docs.microsoft.com/en-us/azure/databricks/spark/latest/structured-streaming/aqs

NEW QUESTION 305
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Blob Storage account named storage1 and an Azure Synapse Analytics dedicated SQL pool named Pool1.
You need to store data in storage1. The data will be read by Pool1. The solution must meet the following requirements:

- Enable Pool1 to skip columns and rows that are unnecessary in a query.
- Automatically create column statistics.
- Minimize the size of files. Which type of file should you use?

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

Answer: B

Explanation:
Automatic creation of statistics is turned on for Parquet files. For CSV files, you need to create statistics manually until automatic creation of CSV files statistics is supported.
Reference:

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

NEW QUESTION 308

- (Exam Topic 3)

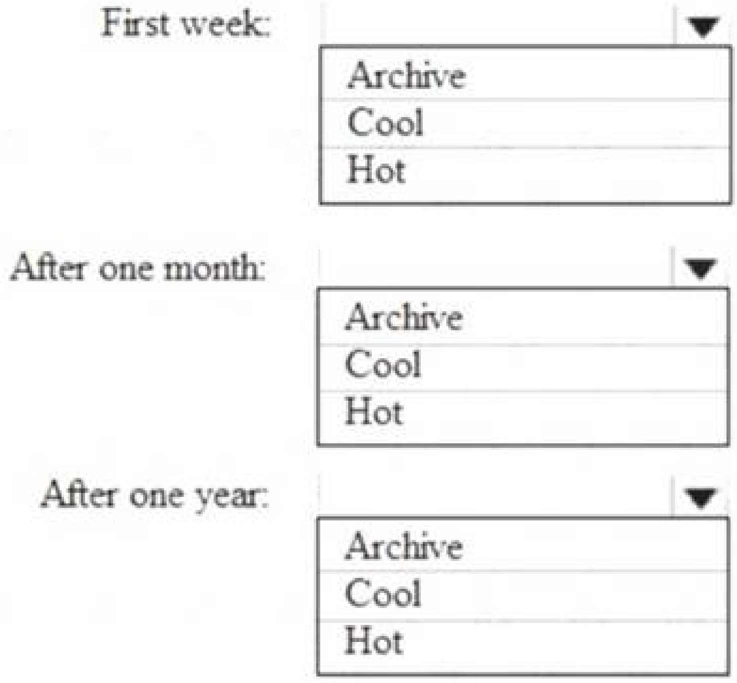
You are designing an application that will store petabytes of medical imaging data

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 312

- (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 315

......