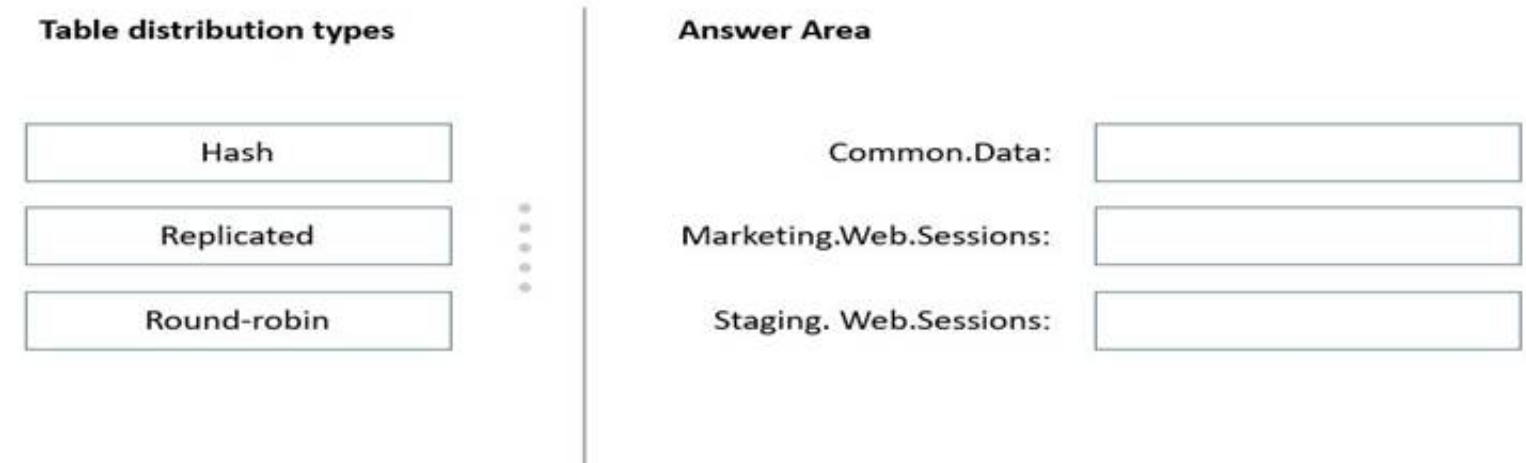
Microsoft

Exam Questions DP-203

Data Engineering on Microsoft Azure

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



A. Mastered
B. Not Mastered

Answer: A

Explanation:

NEW QUESTION 2
- (Exam Topic 3)
You have two Azure Blob Storage accounts named account1 and account2?
You plan to create an Azure Data Factory pipeline that will use scheduled intervals to replicate newly created or modified blobs from account1 to account?
You need to recommend a solution to implement the pipeline. The solution must meet the following requirements:
• Ensure that the pipeline only copies blobs that were created of modified since the most recent replication event.
• Minimize the effort to create the pipeline. What should you recommend?

A. Create a pipeline that contains a flowlet.
B. Create a pipeline that contains a Data Flow activity.
C. Run the Copy Data tool and select Metadata-driven copy task.
D. Run the Copy Data tool and select Built-in copy task.

Answer: A

NEW QUESTION 3
- (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 4
- (Exam Topic 3)
You have an Azure Synapse Analytics workspace named WS1 that contains an Apache Spark pool named Pool1.
You plan to create a database named D61 in Pool1.
You need to ensure that when tables are created in DB1, the tables are available automatically as external tables to the built-in serverless SQL pod.
Which format should you use for the tables in DB1?

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

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.
For each Spark external table based on Parquet or CSV and located in Azure Storage, an external table is created in a serverless SQL pool database.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-storage-files-spark-tables

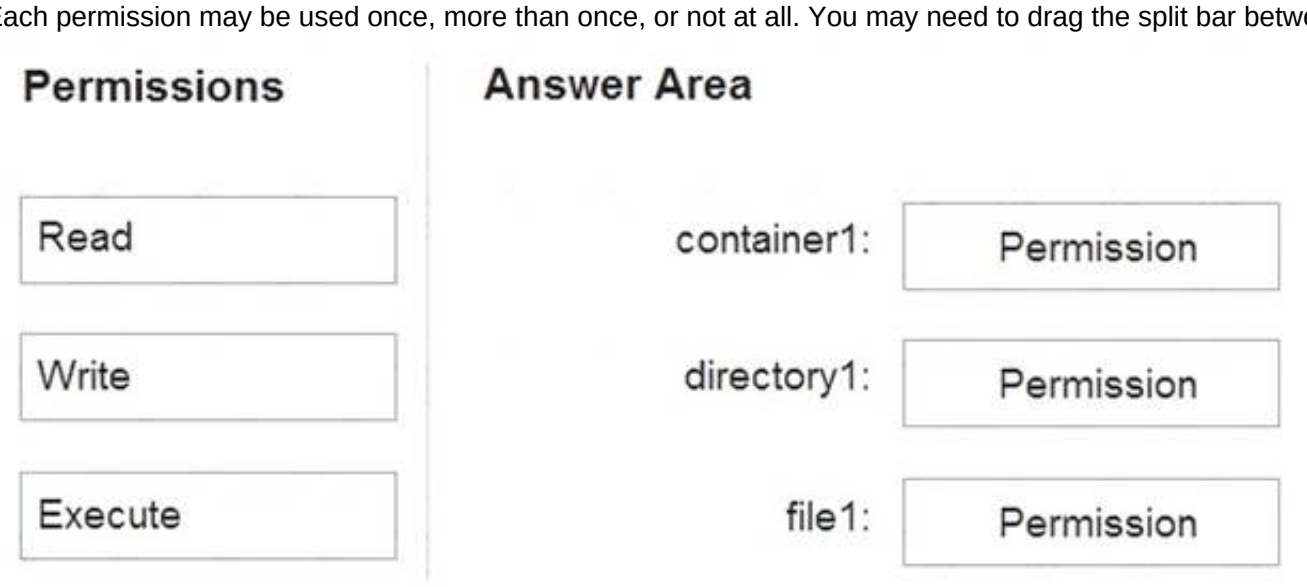
NEW QUESTION 5
- (Exam Topic 3)
You manage an enterprise data warehouse in Azure Synapse Analytics.
Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries.
You need to monitor resource utilization to determine the source 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 6
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named storage1. Storage1 contains a container named container1.
Container1 contains a directory named directory1. Directory1 contains a file named file1.
You have an Azure Active Directory (Azure AD) user named User1 that is assigned the Storage Blob Data Reader role for storage1.
You need to ensure that User1 can append data to file1. The solution must use the principle of least privilege. Which permissions should you grant? To answer, drag the appropriate permissions to the correct resources.
Each permission may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Execute
If you are granting permissions by using only ACLs (no Azure RBAC), then to grant a security principal read or write access to a file, you'll need to give the security

principal Execute permissions to the root folder of the container, and to each folder in the hierarchy of folders that lead to the file.
Box 2: Execute
On Directory: Execute (X): Required to traverse the child items of a directory Box 3: Write
On file: Write (W): Can write or append to a file. Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control

NEW QUESTION 7
- (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 8
- (Exam Topic 3)
You are designing an Azure Databricks cluster that runs user-defined local processes. You need to recommend a cluster configuration that meets the following requirements:
• Minimize query latency.
• Maximize the number of users that can run queues on the cluster at the same time « Reduce overall costs without compromising other requirements
Which cluster type should you recommend?

A. Standard with Auto termination
B. Standard with Autoscaling
C. High Concurrency with Autoscaling
D. High Concurrency with Auto Termination

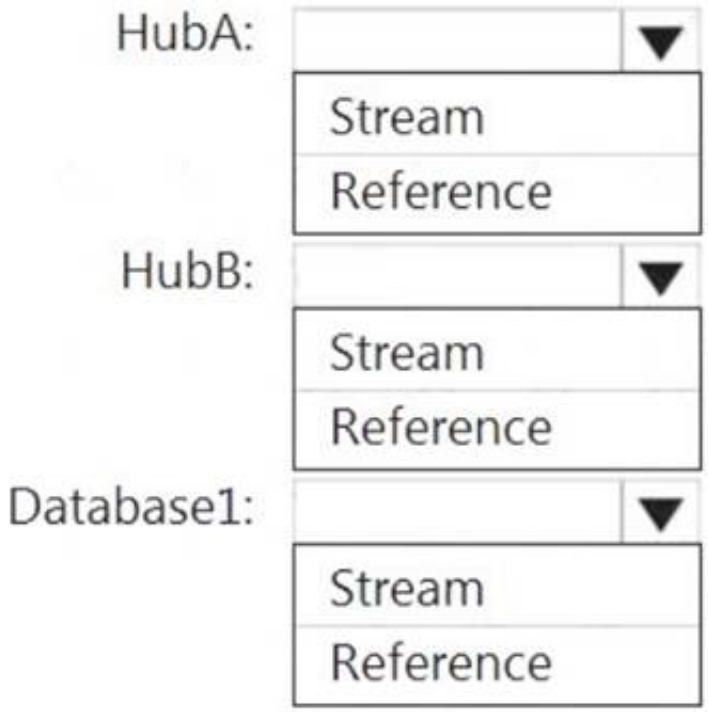
Answer: C

Explanation:
A High Concurrency cluster is a managed cloud resource. The key benefits of High Concurrency clusters are that they provide fine-grained sharing for maximum resource utilization and minimum query latencies.
Databricks chooses the appropriate number of workers required to run your job. This is referred to as autoscaling. Autoscaling makes it easier to achieve high cluster utilization, because you don't need to provision the cluster to match a workload.
Reference:
https://docs.microsoft.com/en-us/azure/databricks/clusters/configure

NEW QUESTION 9
- (Exam Topic 3)
You have an Azure SQL database named Database1 and two Azure event hubs named HubA and HubB. The data consumed from each source is shown in the following table.

Source	Data
Database1	Driver's name Driver's license number
HubA	Ride route Ride distance Ride duration
HubB	Ride fare Ride payment

You need to implement Azure Stream Analytics to calculate the average fare per mile by driver.
How should you configure the Stream Analytics input for each source? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
HubA: Stream HubB: Stream
Database1: Reference
Reference data (also known as a lookup table) is a finite data set that is static or slowly changing in nature, used to perform a lookup or to augment your data streams. For example, in an IoT scenario, you could store metadata about sensors (which don't change often) in reference data and join it with real time IoT data streams. Azure Stream Analytics loads reference data in memory to achieve low latency stream processing
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data

NEW QUESTION 10
- (Exam Topic 3)
You are designing a dimension table for a data warehouse. The table will track the value of the dimension attributes over time and preserve the history of the data by adding new rows as the data changes.
Which type of slowly changing dimension (SCD) should use?

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

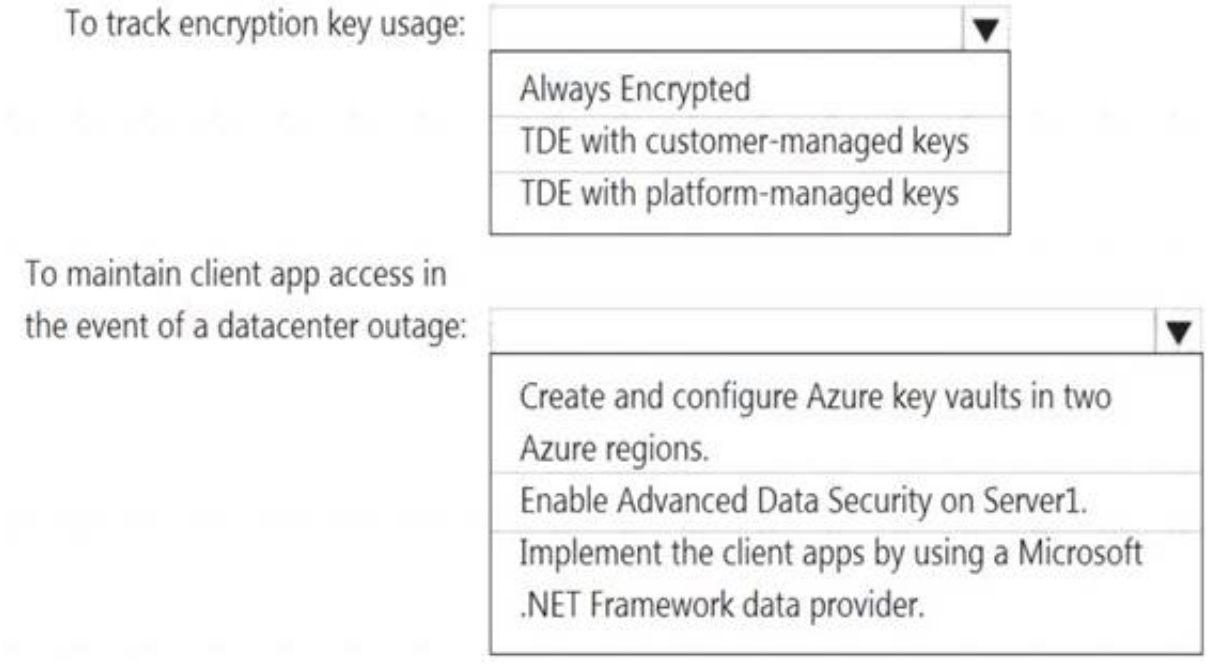
Answer: C

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

NEW QUESTION 10
- (Exam Topic 3)
You have an Azure subscription that contains a logical Microsoft SQL server named Server1. Server1 hosts an Azure Synapse Analytics SQL dedicated pool named Pool1.
You need to recommend a Transparent Data Encryption (TDE) solution for Server1. The solution must meet the following requirements:

- Track the usage of encryption keys.
- Maintain the access of client apps to Pool1 in the event of an Azure datacenter outage that affects the availability of the encryption keys.

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



A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 11
- (Exam Topic 3)
You plan to build a structured streaming solution in Azure Databricks. The solution will count new events in five-minute intervals and report only events that arrive during the interval. The output will be sent to a Delta Lake table.
Which output mode should you use?

A. complete
B. update
C. append

Answer: C

Explanation:
Append Mode: Only new rows appended in the result table since the last trigger are written to external storage. This is applicable only for the queries where existing rows in the Result Table are not expected to change.
https://docs.databricks.com/getting-started/spark/streaming.html

NEW QUESTION 16
- (Exam Topic 3)
You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date.
You need to ensure that the design provides the fastest query times of the records when querying for arbitrary date ranges and aggregating by fiscal calendar attributes.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Create a date dimension table that has a DateTime key.
B. Use built-in SQL functions to extract date attributes.
C. Create a date dimension table that has an integer key in the format of yyyymmdd.
D. In the fact table, use integer columns for the date fields.
E. Use DateTime columns for the date fields.

Answer: BD

NEW QUESTION 20
- (Exam Topic 3)
You are designing an anomaly detection solution for streaming data from an Azure IoT hub. The solution must meet the following requirements:
- Send the output to Azure Synapse.
- Identify spikes and dips in time series data.
- Minimize development and configuration effort. Which should you include in the solution?

A. Azure Databricks
B. Azure Stream Analytics

C. Azure SQL Database

Answer: B

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

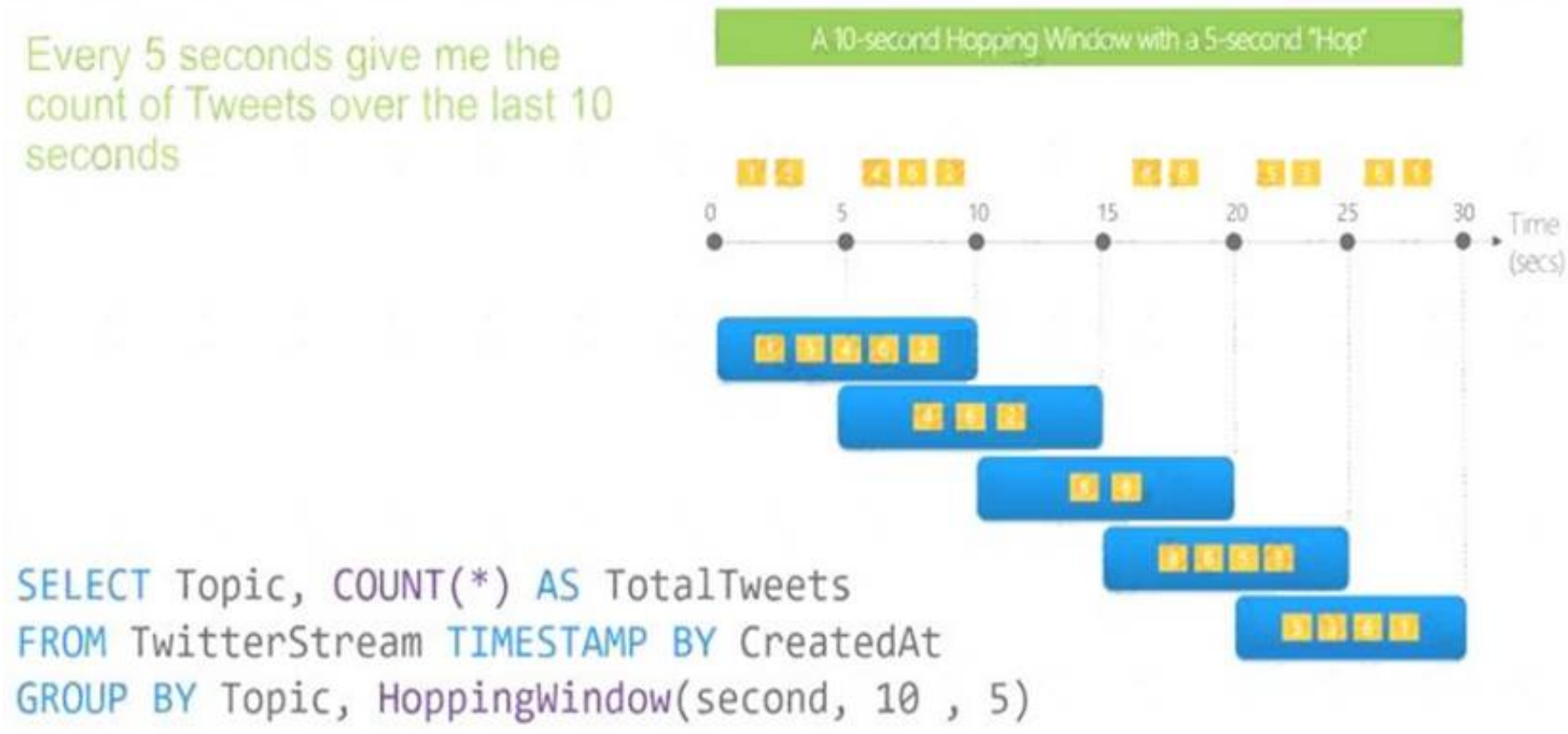
NEW QUESTION 25
- (Exam Topic 3)
You are building an Azure Stream Analytics job to retrieve game data.
You need to ensure that the job returns the highest scoring record for each five-minute time interval of each game.
How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

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

FROM input TIMESTAMP BY CreatedAt

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: TopOne OVER(PARTITION BY Game ORDER BY Score Desc)
TopOne returns the top-rank record, where rank defines the ranking position of the event in the window according to the specified ordering. Ordering/ranking is based on event columns and can be specified in ORDER BY clause.
Box 2: Hopping(minute,5)
Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.
A picture containing timeline Description automatically generated



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

NEW QUESTION 26
- (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.

A. Mastered
B. Not Mastered

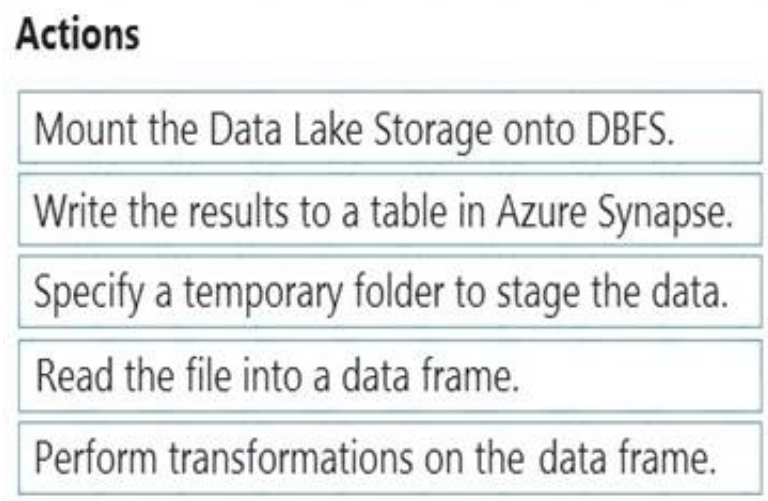
Answer: A

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 27
- (Exam Topic 3)
You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers. The file contains two attributes named FirstName and LastName.
You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the FirstName and LastName values.
You create the following components:

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

In which order should you perform the actions? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Table Description automatically generated
Step 1: Mount the Data Lake Storage onto DBFS
Begin with creating a file system in the Azure Data Lake Storage Gen2 account. Step 2: Read the file into a data frame.
You can load the json files as a data frame in Azure Databricks. Step 3: Perform transformations on the data frame.
Step 4: Specify a temporary folder to stage the data
Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse. Step 5: Write the results to a table in Azure Synapse.
You upload the transformed data frame into Azure Synapse. You use the Azure Synapse connector for Azure Databricks to directly upload a dataframe as a table in a Azure Synapse.
Reference:
https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-extract-load-sql-data-warehouse

NEW QUESTION 32
- (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 36
- (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 39
- (Exam Topic 3)
You are performing exploratory analysis of the bus fare data in an Azure Data Lake Storage Gen2 account by using an Azure Synapse Analytics serverless SQL pool.
You execute the Transact-SQL query shown in the following exhibit.

```
SELECT
   payment_type,
   SUM(fare_amount) AS fare_total
FROM OPENROWSET(
         BULK 'csv/busfare/tripdata_2020*.csv',
         DATA_SOURCE = 'BusData',
         FORMAT = 'CSV', PARSER_VERSION = '2.0',
         FIRSTROW = 2
   )
   WITH (
         payment_type INT 10,
         fare_amount FLOAT 11
   ) AS nyc
GROUP BY payment_type
ORDER BY payment_type;
```

What do the query results include?

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

Answer: D

NEW QUESTION 43
- (Exam Topic 3)
A company purchases IoT devices to monitor manufacturing machinery. The company uses an IoT appliance to communicate with the IoT devices.
The company must be able to monitor the devices in real-time. You need to design the solution.
What should you recommend?

A. Azure Stream Analytics cloud job using Azure PowerShell
B. Azure Analysis Services using Azure Portal
C. Azure Data Factory instance using Azure Portal
D. Azure Analysis Services using Azure PowerShell

Answer: C

Explanation:
Stream Analytics is a cost-effective event processing engine that helps uncover real-time insights from devices, sensors, infrastructure, applications and data quickly and easily.
Monitor and manage Stream Analytics resources with Azure PowerShell cmdlets and powershell scripting that execute basic Stream Analytics tasks.
Reference:
https://cloudblogs.microsoft.com/sqlserver/2014/10/29/microsoft-adds-iot-streaming-analytics-data-production-a

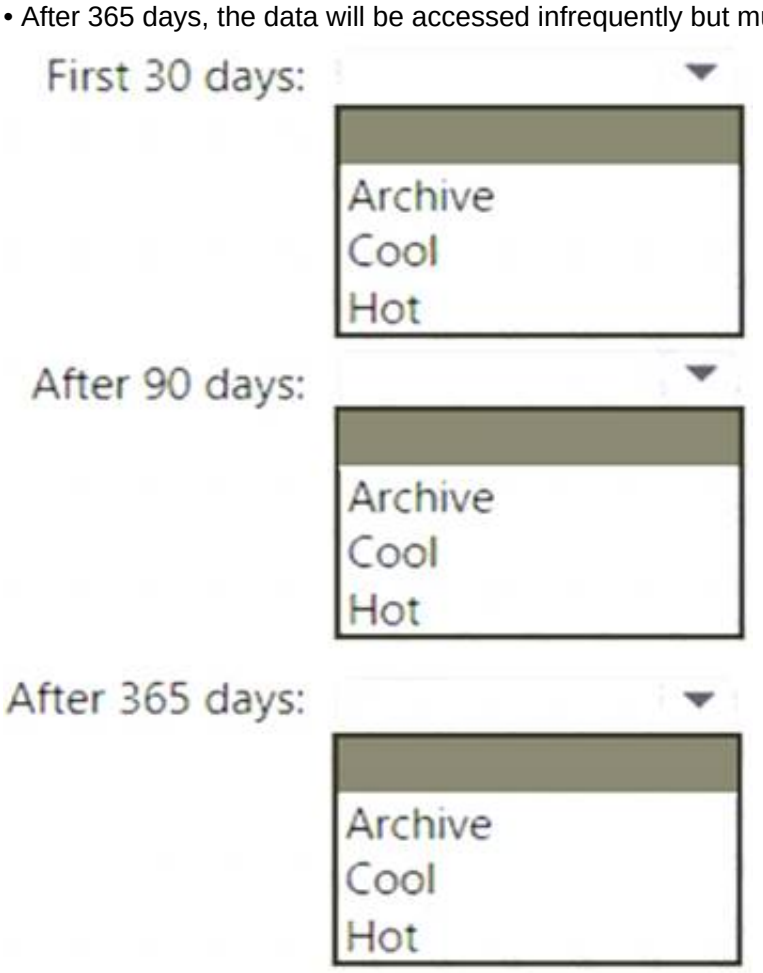
NEW QUESTION 46
- (Exam Topic 3)
You are designing an Azure Synapse solution that will provide a query interface for the data stored in an Azure Storage account. The storage account is only accessible from a virtual network.
You need to recommend an authentication mechanism to ensure that the solution can access the source data.
What should you recommend?

A. a managed identity
B. anonymous public read access
C. a shared key

Answer: A

Explanation:
Managed Identity authentication is required when your storage account is attached to a VNet. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-exa

NEW QUESTION 48
- (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.



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 49
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly. Solution: You convert the files to compressed delimited text files. Does this meet the goal?

A. Yes
B. No

Answer: A

Explanation:
All file formats have different performance characteristics. For the fastest load, use compressed delimited text files.
Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

NEW QUESTION 53
- (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 56
- (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 58
- (Exam Topic 3)
A company plans to use Platform-as-a-Service (PaaS) to create the new data pipeline process. The process must meet the following requirements:
Ingest:
- Access multiple data sources.
- Provide the ability to orchestrate workflow.
- Provide the capability to run SQL Server Integration Services packages.

Store:
Optimize storage for big data workloads. Provide encryption of data at rest. Operate with no size limits.
Prepare and Train:
- Provide a fully-managed and interactive workspace for exploration and visualization.
- Provide the ability to program in R, SQL, Python, Scala, and Java.
- Provide seamless user authentication with Azure Active Directory. Model & Serve:
- Implement native columnar storage.
- Support for the SQL language
- Provide support for structured streaming. You need to build the data integration pipeline.

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

Answer Area

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

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 63
- (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 '  [ abfs | abfss | wasb | wasbs ]  ://data@newyorktaxidataset.dfs.core.windows.net' ,
credential = ADLS_credential ,
TYPE -  [ 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 68
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 receives new data once every 24 hours.
You have the following function.

```
create function dbo.udfFtoC(F decimal)
return decimal
as
begin
return (F - 32) * 5.0 / 9
end
```

You have the following query.

```
select avg_date, sensorid, avg_f, dbo.udfFtoC(avg_temperature) as avg_c from SensorTemps
where avg_date = @parameter
```

The query is executed once every 15 minutes and the @parameter value is set to the current date. You need to minimize the time it takes for the query to return results.
Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

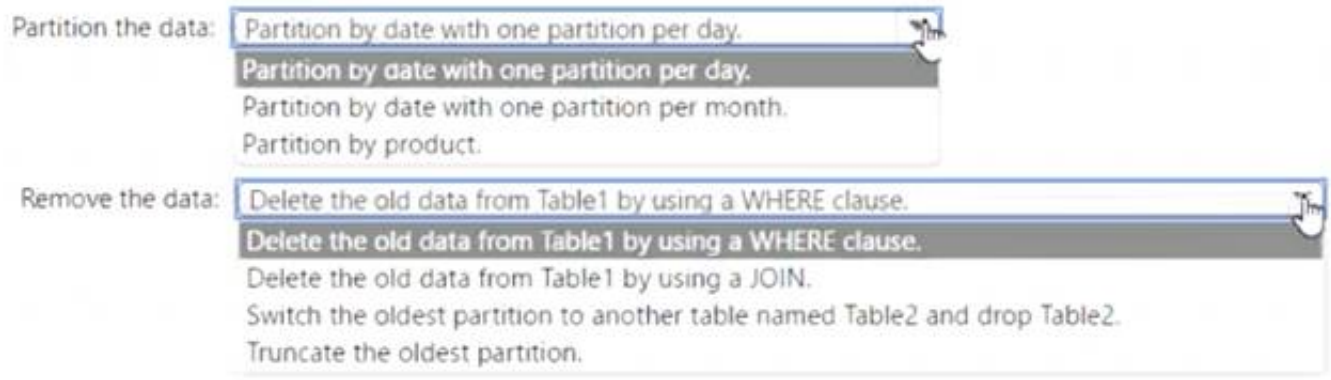
A. Create an index on the avg_f column.
B. Convert the avg_c column into a calculated column.
C. Create an index on the sensorid column.
D. Enable result set caching.

E. Change the table distribution to replicate.

Answer: BD

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

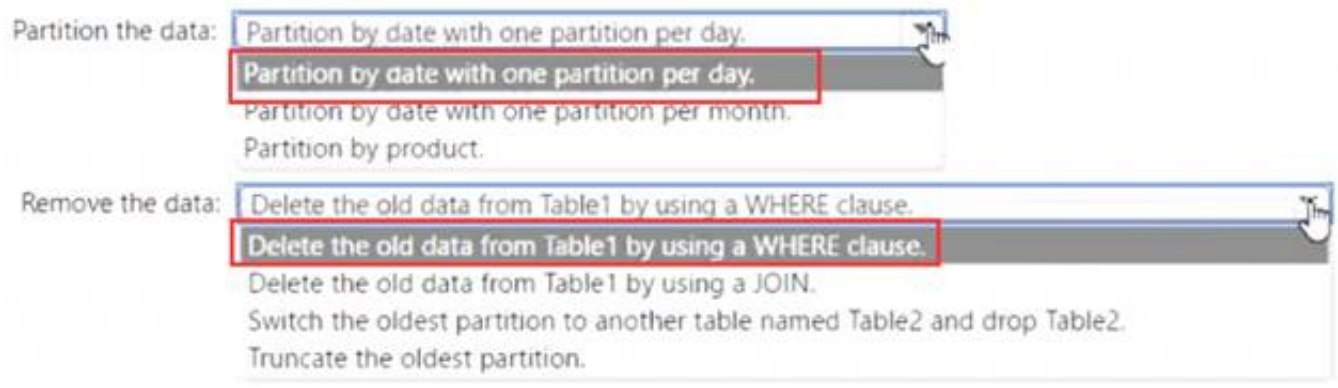
NEW QUESTION 73
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Tablet. Table1 contains sales data. Sixty-five million rows of data are added to Table1 monthly.
At the end of each month, you need to remove data that is older than 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



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Answer Area



NEW QUESTION 78
- (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 83
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- A workload for data engineers who will use Python and SQL.
- A workload for jobs that will run notebooks that use Python, Scala, and SOL.
- A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

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

A. Yes
B. No

Answer: B

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

NEW QUESTION 85
- (Exam Topic 3)
You are monitoring an Azure Stream Analytics job.
You discover that the Backlogged Input Events metric is increasing slowly and is consistently non-zero. You need to ensure that the job can handle all the events.
What should you do?

A. Change the compatibility level of the Stream Analytics job.
B. Increase the number of streaming units (SUs).
C. Remove any named consumer groups from the connection and use $default.
D. Create an additional output stream for the existing input stream.

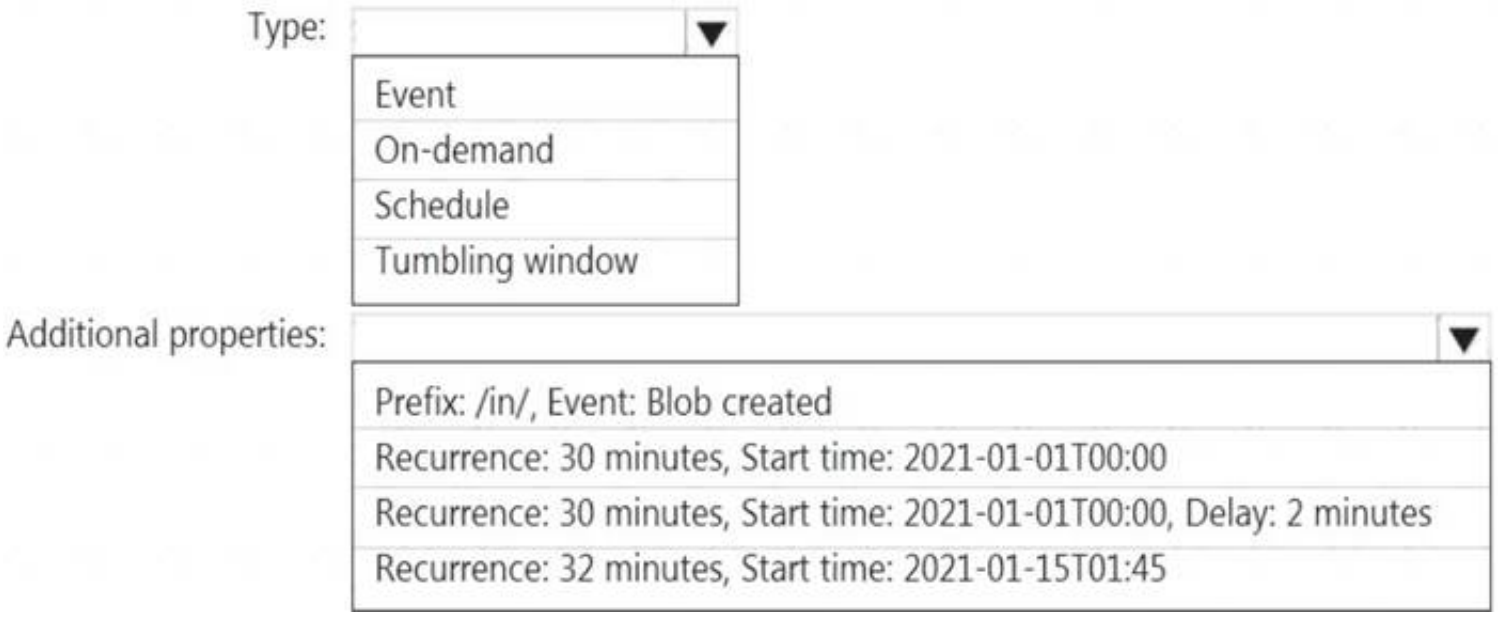
Answer: B

Explanation:
Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job. You should increase the Streaming Units.
Note: Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job.
Reference:
https://docs.microsoft.com/bs-cyrl-ba/azure/stream-analytics/stream-analytics-monitoring

NEW QUESTION 86
- (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.



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 89
- (Exam Topic 3)
You build a data warehouse in an Azure Synapse Analytics dedicated SQL pool.
Analysts write a complex SELECT query that contains multiple JOIN and CASE statements to transform data for use in inventory reports. The inventory reports will use the data and additional WHERE parameters depending on the report. The reports will be produced once daily.
You need to implement a solution to make the dataset available for the reports. The solution must minimize query times.
What should you implement?

A. a materialized view
B. a replicated table
C. in ordered clustered columnstore index
D. result set chaching

Answer: A

Explanation:
Materialized views for dedicated SQL pools in Azure Synapse provide a low maintenance method for complex analytical queries to get fast performance without any query change.
Note: When result set caching is enabled, dedicated SQL pool automatically caches query results in the user database for repetitive use. This allows subsequent query executions to get results directly from the persisted cache so recomputation is not needed. Result set caching improves query performance and reduces compute resource usage. In addition, queries using cached results set do not use any concurrency slots and thus do not count against existing concurrency limits.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-materialized- https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-cac

NEW QUESTION 90
- (Exam Topic 3)
You are monitoring an Azure Stream Analytics job by using metrics in Azure.
You discover that during the last 12 hours, the average watermark delay is consistently greater than the configured late arrival tolerance.
What is a possible cause of this behavior?

A. Events whose application timestamp is earlier than their arrival time by more than five minutes arrive as inputs.
B. There are errors in the input data.
C. The late arrival policy causes events to be dropped.
D. The job lacks the resources to process the volume of incoming data.

Answer: D

Explanation:
Watermark Delay indicates the delay of the streaming data processing job.
There are a number of resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to:
- Not enough processing resources in Stream Analytics to handle the volume of input events. To scale up resources, see Understand and adjust Streaming Units.
- Not enough throughput within the input event brokers, so they are throttled. For possible solutions, see Automatically scale up Azure Event Hubs throughput units.

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 93
- (Exam Topic 2)
What should you recommend using to secure sensitive customer contact information?

A. data labels
B. column-level security
C. row-level security
D. Transparent Data Encryption (TDE)

Answer: B

Explanation:
Scenario: All cloud data must be encrypted at rest and in transit.
Always Encrypted is a feature designed to protect sensitive data stored in specific database columns from
access (for example, credit card numbers, national identification numbers, or data on a need to know basis). This includes database administrators or other privileged users who are authorized to access the database to perform management tasks, but have no business need to access the particular data in the encrypted columns. The data is always encrypted, which means the encrypted data is decrypted only for processing by client applications with access to the encryption key.
References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-security-overview

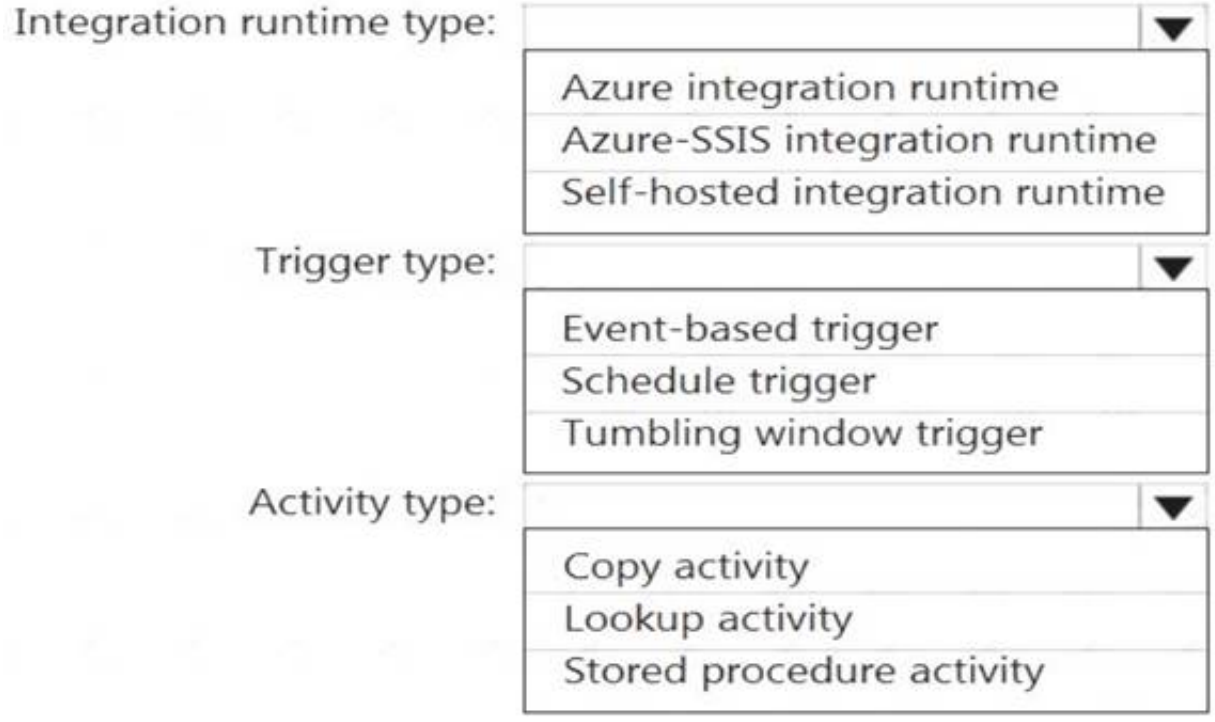
NEW QUESTION 94
- (Exam Topic 2)
What should you recommend to prevent users outside the Litware on-premises network from accessing the analytical data store?

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 96
- (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.



A. Mastered
B. Not Mastered

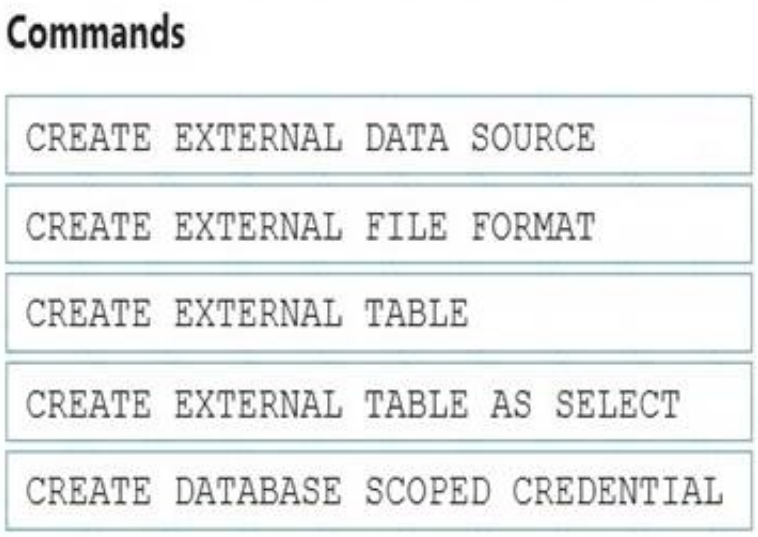
Answer: A

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 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.

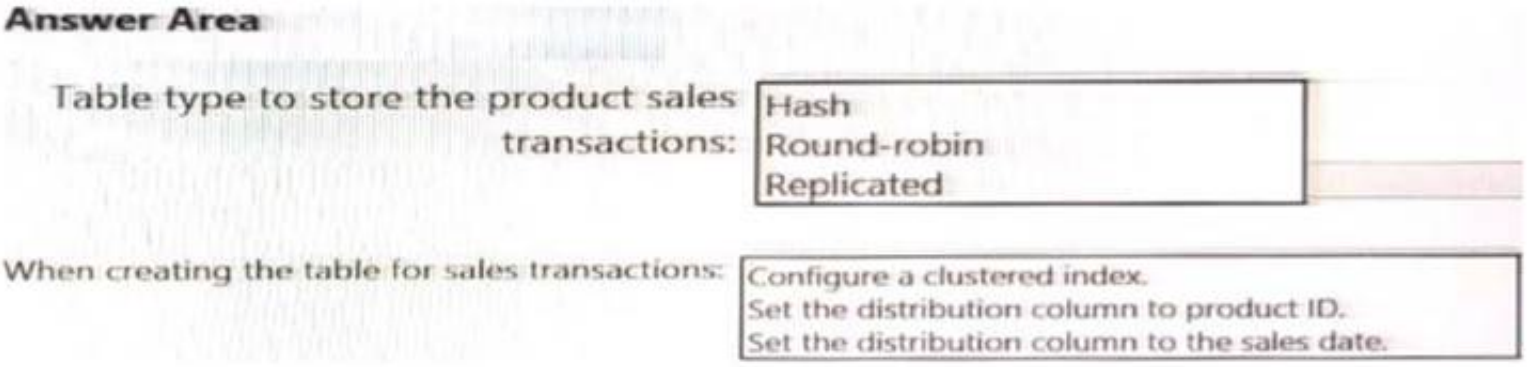


A. Mastered
B. Not Mastered

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 98
- (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.



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 101
- (Exam Topic 1)
You need to design a data retention solution for the Twitter teed data records. The solution must meet the customer sentiment analytics requirements.
Which Azure Storage functionality should you include in the solution?

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

Answer: D

NEW QUESTION 103
- (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.

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 107
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more
than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly. Solution: You copy the files to a table that has a columnstore index. Does this meet the goal?

A. Yes
B. No

Answer: B

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

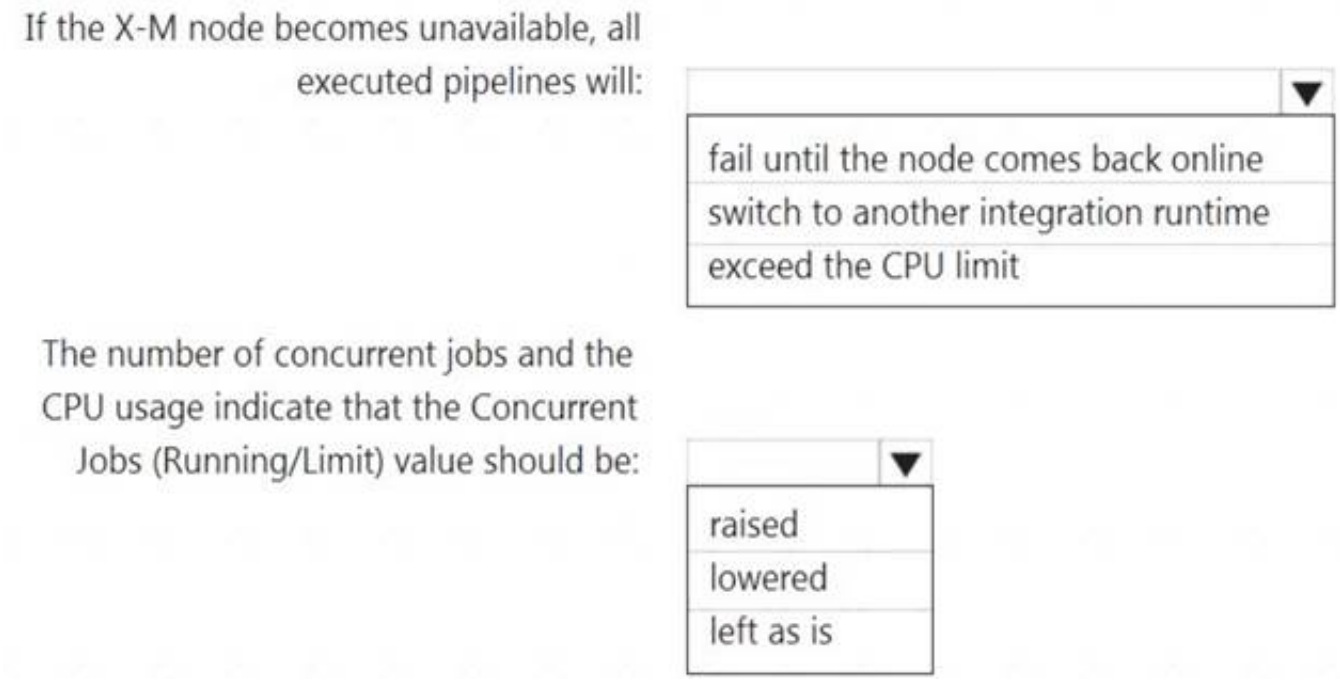
NEW QUESTION 111
- (Exam Topic 3)
You have a self-hosted integration runtime in Azure Data Factory.
The current status of the integration runtime has the following configurations:
- Status: Running
- Type: Self-Hosted
- Version: 4.4.7292.1

- 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.

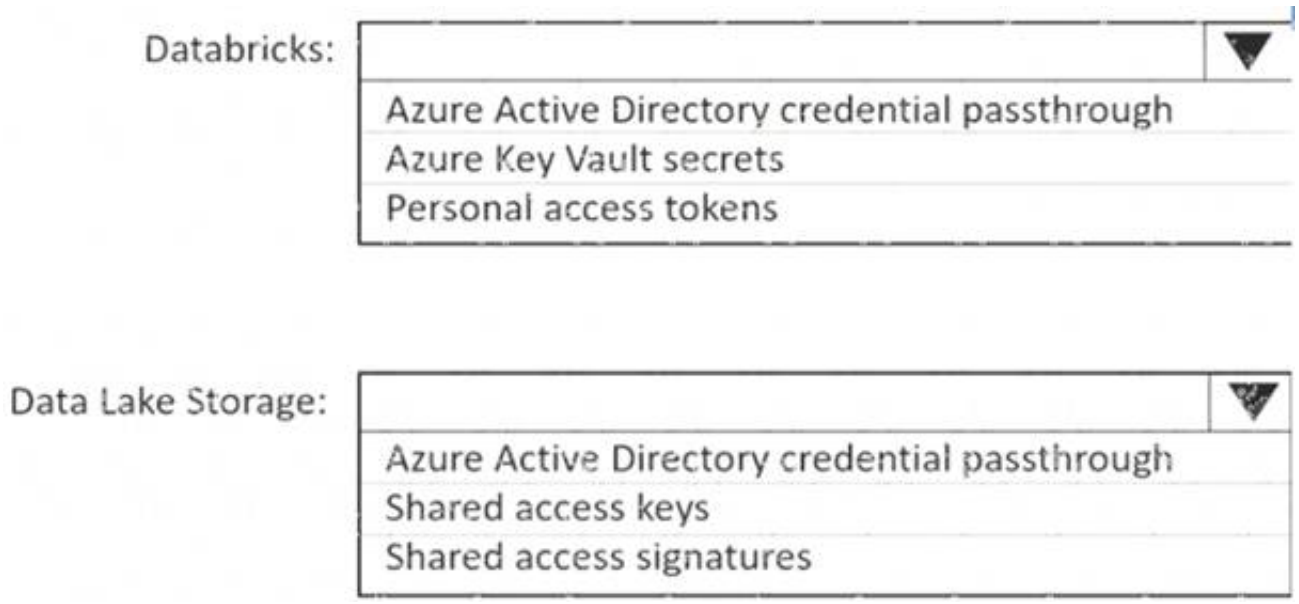


A. Mastered
B. Not Mastered

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 116
- (Exam Topic 3)
You use Azure Data Lake Storage Gen2 to store data that data scientists and data engineers will query by using Azure Databricks interactive notebooks. Users will have access only to the Data Lake Storage folders that relate to the projects on which they work.
You need to recommend which authentication methods to use for Databricks and Data Lake Storage to provide the users with the appropriate access. The solution must minimize administrative effort and development effort.
Which authentication method should you recommend for each Azure service? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Table Description automatically generated
Box 1: Personal access tokens
You can use storage shared access signatures (SAS) to access an Azure Data Lake Storage Gen2 storage account directly. With SAS, you can restrict access to a storage account using temporary tokens with fine-grained access control.
You can add multiple storage accounts and configure respective SAS token providers in the same Spark session.
Box 2: Azure Active Directory credential passthrough
You can authenticate automatically to Azure Data Lake Storage Gen1 (ADLS Gen1) and Azure Data Lake Storage Gen2 (ADLS Gen2) from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that you use to log into Azure Databricks. When you enable your cluster for Azure Data Lake Storage credential passthrough, commands that you run on that cluster can read and write data in Azure Data Lake Storage without requiring you to configure service principal credentials for access to storage.
After configuring Azure Data Lake Storage credential passthrough and creating storage containers, you can access data directly in Azure Data Lake Storage Gen1 using an adl:// path and Azure Data Lake Storage Gen2 using an abfss:// path:
Reference:
https://docs.microsoft.com/en-us/azure/databricks/data/data-sources/azure/adls-gen2/azure-datalake-gen2-sas-ac https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough

NEW QUESTION 120
- (Exam Topic 3)
You have an enterprise data warehouse in Azure Synapse Analytics.
You need to monitor the data warehouse to identify whether you must scale up to a higher service level to accommodate the current workloads
Which is the best metric to monitor?
More than one answer choice may achieve the goal. Select the BEST answer.

A. Data 10 percentage
B. CPU percentage
C. DWU used
D. DWU percentage

Answer: C

NEW QUESTION 123
- (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 126
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named SA1 that contains a table named Table1. You need to identify tables that have a high percentage of deleted rows. What should you run?
A)

```
sys.pdw_nodes_column_store_segments
```

B)

```
sys.dm_db_column_store_row_group_operational_stats
```

C)

```
sys.pdw_nodes_column_store_row_groups
```

D)

```
sys.dm_db_column_store_row_group_physical_stats
```

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

Answer: B

NEW QUESTION 131
- (Exam Topic 3)
You need to create an Azure Data Factory pipeline to process data for the following three departments at your company: Ecommerce, retail, and wholesale. The solution must ensure that data can also be processed for the entire company.
How should you complete the Data Factory data flow script? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.

Values

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

Answer Area

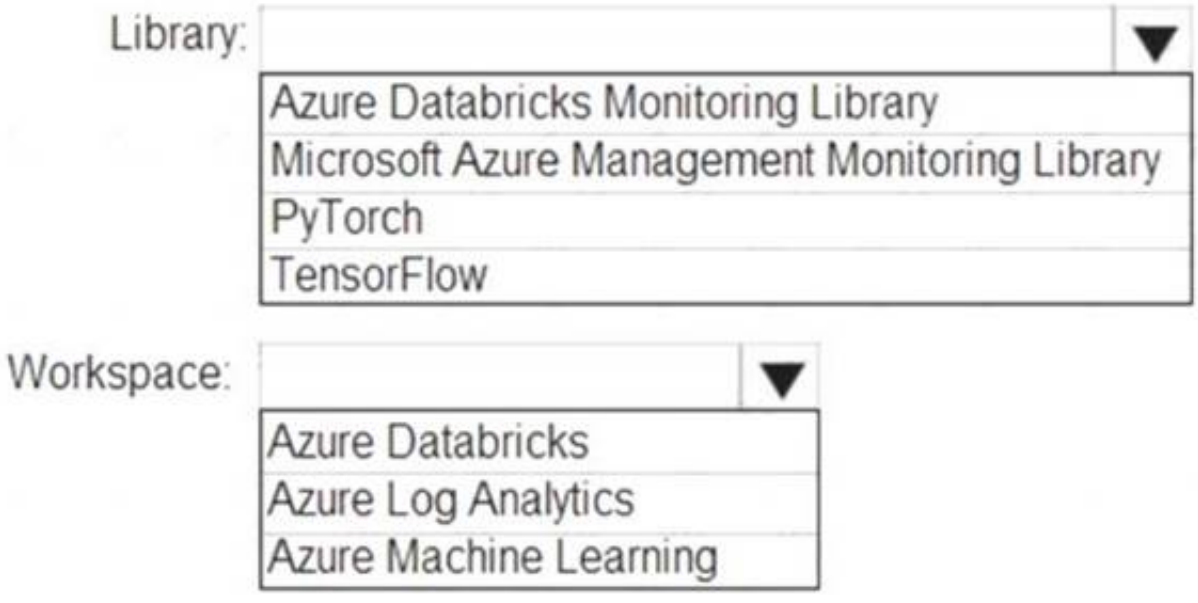
```
CleanData
    split(
        [                    ]
        [                    ]
    ) ~> SplitByDept@(  [                    ] )
```

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 133
- (Exam Topic 3)
You need to collect application metrics, streaming query events, and application log messages for an Azure Databrick cluster.
Which type of library and workspace should you implement? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
You can send application logs and metrics from Azure Databricks to a Log Analytics workspace. It uses the Azure Databricks Monitoring Library, which is available on GitHub.
References:
https://docs.microsoft.com/en-us/azure/architecture/databricks-monitoring/application-logs

NEW QUESTION 135
- (Exam Topic 3)
You have an Azure Synapse Analystics 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 136
- (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.

A. Mastered

B. Not Mastered

Answer: A

Explanation:

NEW QUESTION 138
- (Exam Topic 3)
You are designing a highly available Azure Data Lake Storage solution that will include geo-zone-redundant storage (GZRS).
You need to monitor for replication delays that can affect the recovery point objective (RPO). What should you include in the monitoring solution?

A. availability
B. Average Success E2E Latency
C. 5xx: Server Error errors
D. Last Sync Time

Answer: D

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 142
- (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 146
- (Exam Topic 3)
You are deploying a lake database by using an Azure Synapse database template.
You need to add additional tables to the database. The solution must use the same grouping method as the template tables.
'Which grouping method should you use?

A. business area
B. size
C. facts and dimensions
D. partition style

Answer: A

Explanation:
Business area: This is how the Azure Synapse database templates group tables by default. Each template consists of one or more enterprise templates that contain tables grouped by business areas. For example, the Retail template has business areas such as Customer, Product, Sales, and Store123. Using the same grouping method as the template tables can help you maintain consistency and compatibility with the industry-specific data model.

https://techcommunity.microsoft.com/t5/azure-synapse-analytics-blog/database-templates-in-azure-synapse-anal

NEW QUESTION 150
- (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 151
- (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

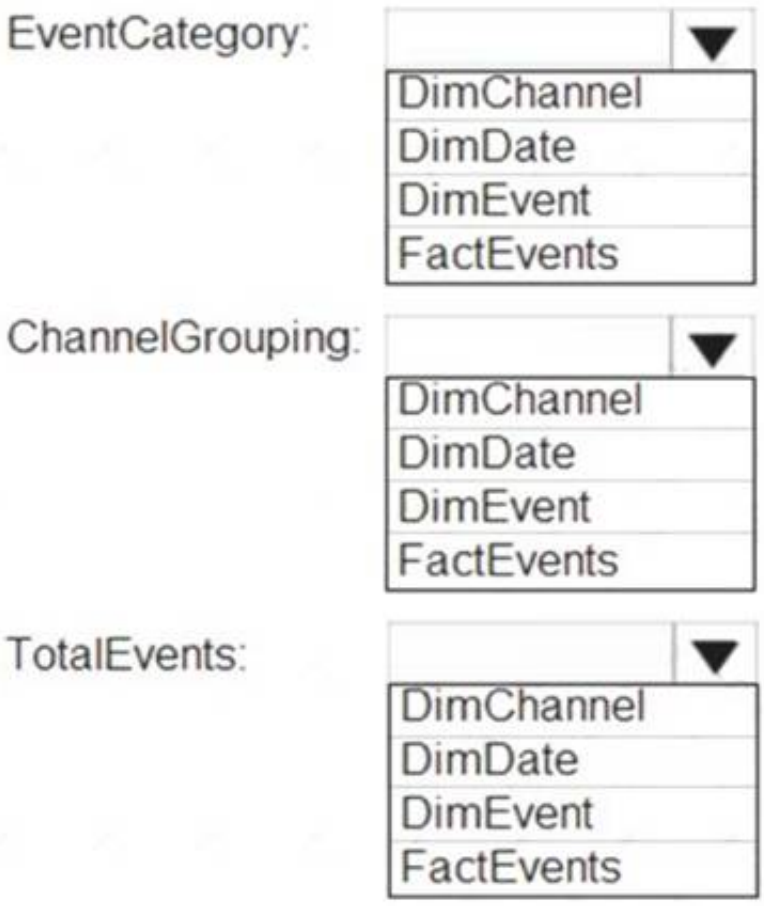
Answer: C

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 156
- (Exam Topic 3)
From a website analytics system, you receive data extracts about user interactions such as downloads, link clicks, form submissions, and video plays.
The data contains the following columns.

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

You need to design a star schema to support analytical queries of the data. The star schema will contain four tables including a date dimension.
To which table should you add each column? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

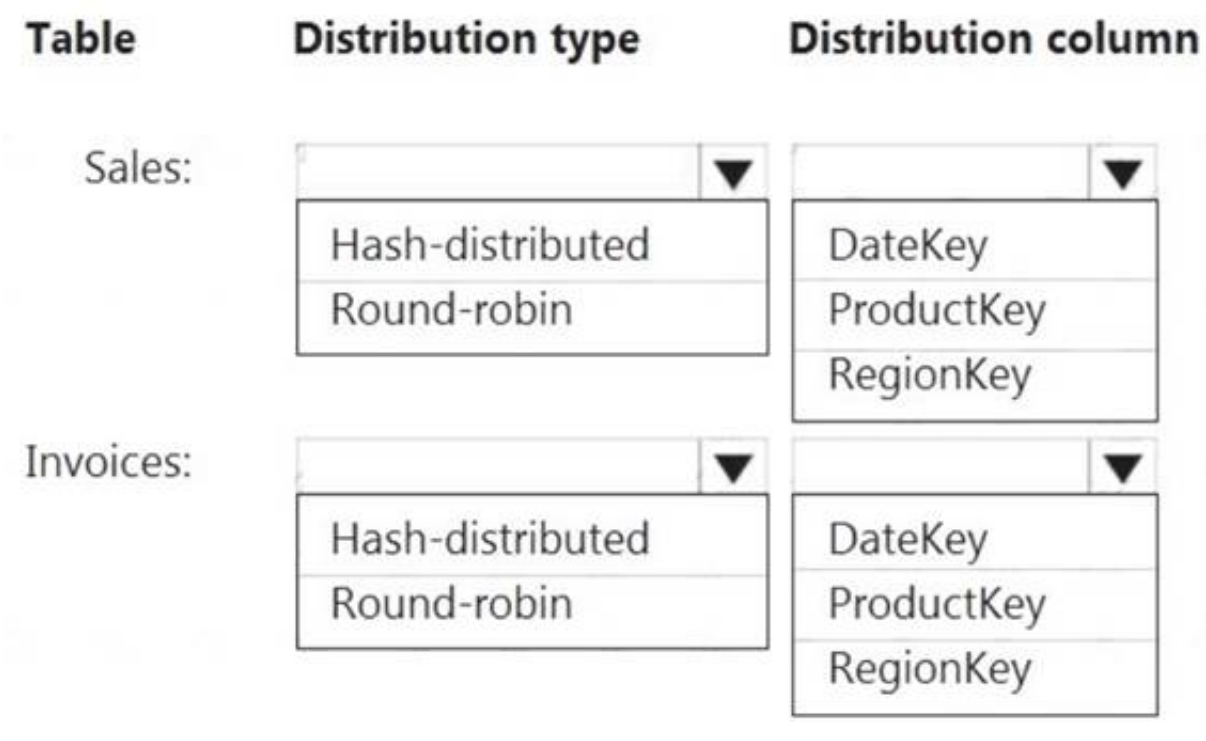
Answer: A

Explanation:
Table Description automatically generated
Box 1: DimEvent
Box 2: DimChannel
Box 3: FactEvents
Fact tables store observations or events, and can be sales orders, stock balances, exchange rates, temperatures, etc
Reference:
https://docs.microsoft.com/en-us/power-bi/guidance/star-schema

NEW QUESTION 161
- (Exam Topic 3)
You have an on-premises data warehouse that includes the following fact tables. Both tables have the following columns: DateKey, ProductKey, RegionKey. There are 120 unique product keys and 65 unique region keys.

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

Queries that use the data warehouse take a long time to complete.
You plan to migrate the solution to use Azure Synapse Analytics. You need to ensure that the Azure-based solution optimizes query performance and minimizes processing skew.
What should you recommend? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point



A. Mastered
B. Not Mastered

Answer: A

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

- When getting started as a simple starting point since it is the default
- If there is no obvious joining key
- If there is not good candidate column for hash distributing the table
- If the table does not share a common join key with other tables
- If the join is less significant than other joins in the query
- When the table is a temporary staging table

Note: A distributed table appears as a single table, but the rows are actually stored across 60 distributions. The rows are distributed with a hash or round-robin algorithm.
Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute

NEW QUESTION 162
- (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 163
- (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 167
- (Exam Topic 3)
You need to implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool.
You have a table that was created by using the following Transact-SQL statement.

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

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

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

Answer: BE

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

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

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

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

NEW QUESTION 170
- (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 session window that uses a timeout size of 10 seconds. Does this meet the goal?

A. Yes
B. No

Answer: A

Explanation:
Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

NEW QUESTION 173
- (Exam Topic 3)
You have an Azure Databricks workspace and an Azure Data Lake Storage Gen2 account named storage! New files are uploaded daily to storage1.
• Incrementally process new files as they are upkorage1 as a structured streaming source. The solution must meet the following requirements:
• Minimize implementation and maintenance effort.
• Minimize the cost of processing millions of files.
• Support schema inference and schema drift. Which should you include in the recommendation?

A. Auto Loader
B. Apache Spark FileStreamSource
C. COPY INTO
D. Azure Data Factory

Answer: D

NEW QUESTION 174
- (Exam Topic 3)
You are designing a dimension table in an Azure Synapse Analytics dedicated SQL pool.
You need to create a surrogate key for the table. The solution must provide the fastest query performance. What should you use for the surrogate key?

A. a GUID column
B. a sequence object
C. an IDENTITY column

Answer: C

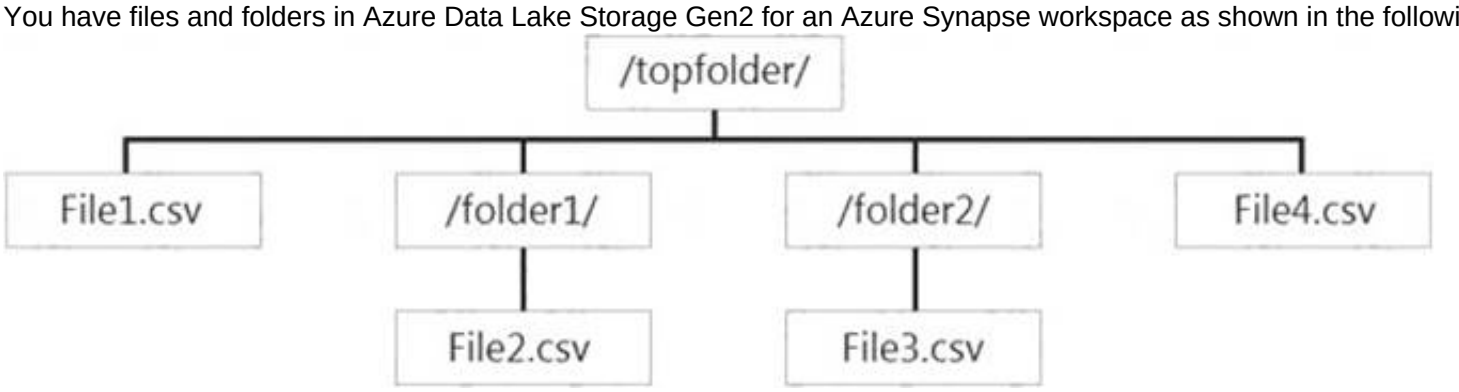
Explanation:
Use IDENTITY to create surrogate keys using dedicated SQL pool in AzureSynapse Analytics.
Note: A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identity

NEW QUESTION 175
- (Exam Topic 3)
You have a partitioned table in an Azure Synapse Analytics dedicated SQL pool. You need to design queries to maximize the benefits of partition elimination. What should you include in the Transact-SQL queries?

A. JOIN
B. WHERE
C. DISTINCT
D. GROUP BY

Answer: B

NEW QUESTION 177
- (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 182
- (Exam Topic 3)
You have an Azure Data Lake Storage Gen 2 account named storage1.
You need to recommend a solution for accessing the content in storage1. The solution must meet the following requirements:

- List and read permissions must be granted at the storage account level.
- Additional permissions can be applied to individual objects in storage1.
- Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.

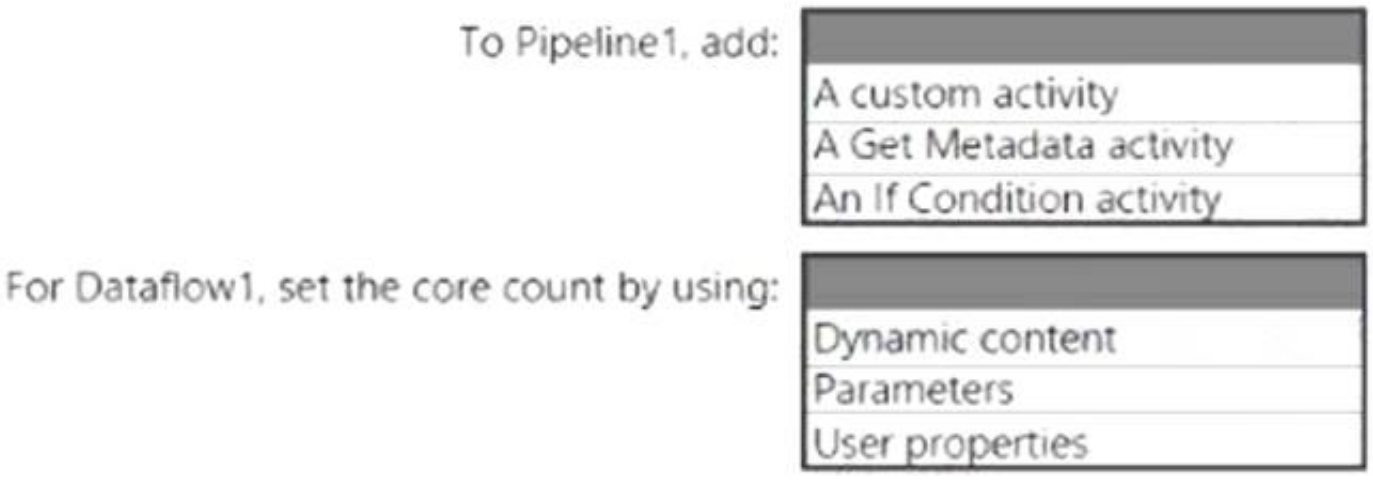
What should you use? To answer, drag the appropriate components to the correct requirements. Each component may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Role-based access control (RBAC) roles
List and read permissions must be granted at the storage account level.
Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.
Role-based access control (Azure RBAC)
Azure RBAC uses role assignments to apply sets of permissions to security principals. A security principal is an object that represents a user, group, service principal, or managed identity that is defined in Azure Active Directory (AD). A permission set can give a security principal a "coarse-grain" level of access such as read or write access to all of the data in a storage account or all of the data in a container.
Box 2: Access control lists (ACLs)
Additional permissions can be applied to individual objects in storage1. Access control lists (ACLs)
ACLs give you the ability to apply "finer grain" level of access to directories and files. An ACL is a permission construct that contains a series of ACL entries. Each ACL entry associates security principal with an access level.
Reference: https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model

NEW QUESTION 186
- (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.

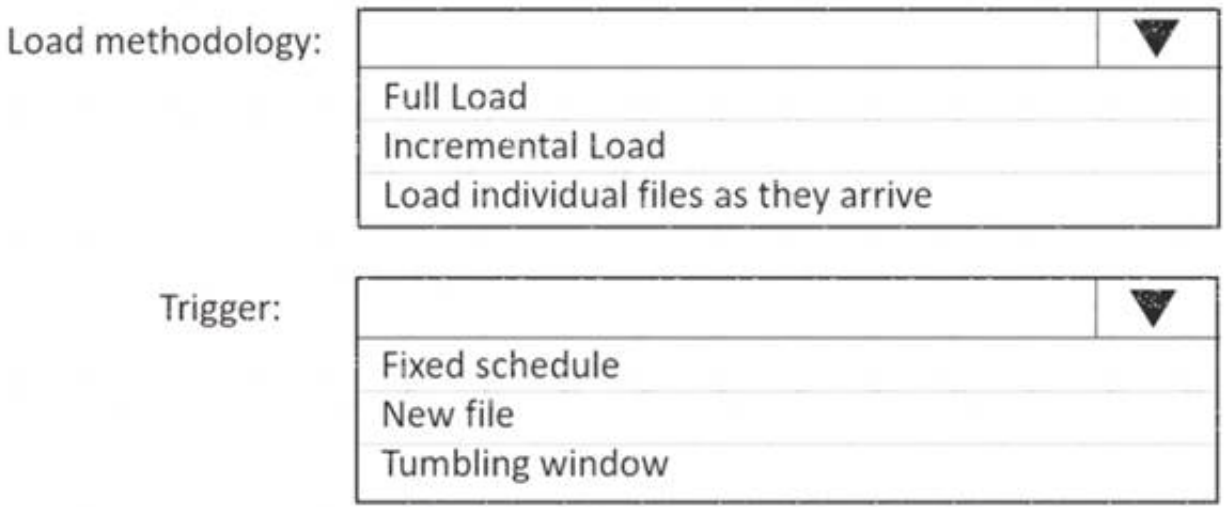


A. Mastered
B. Not Mastered

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

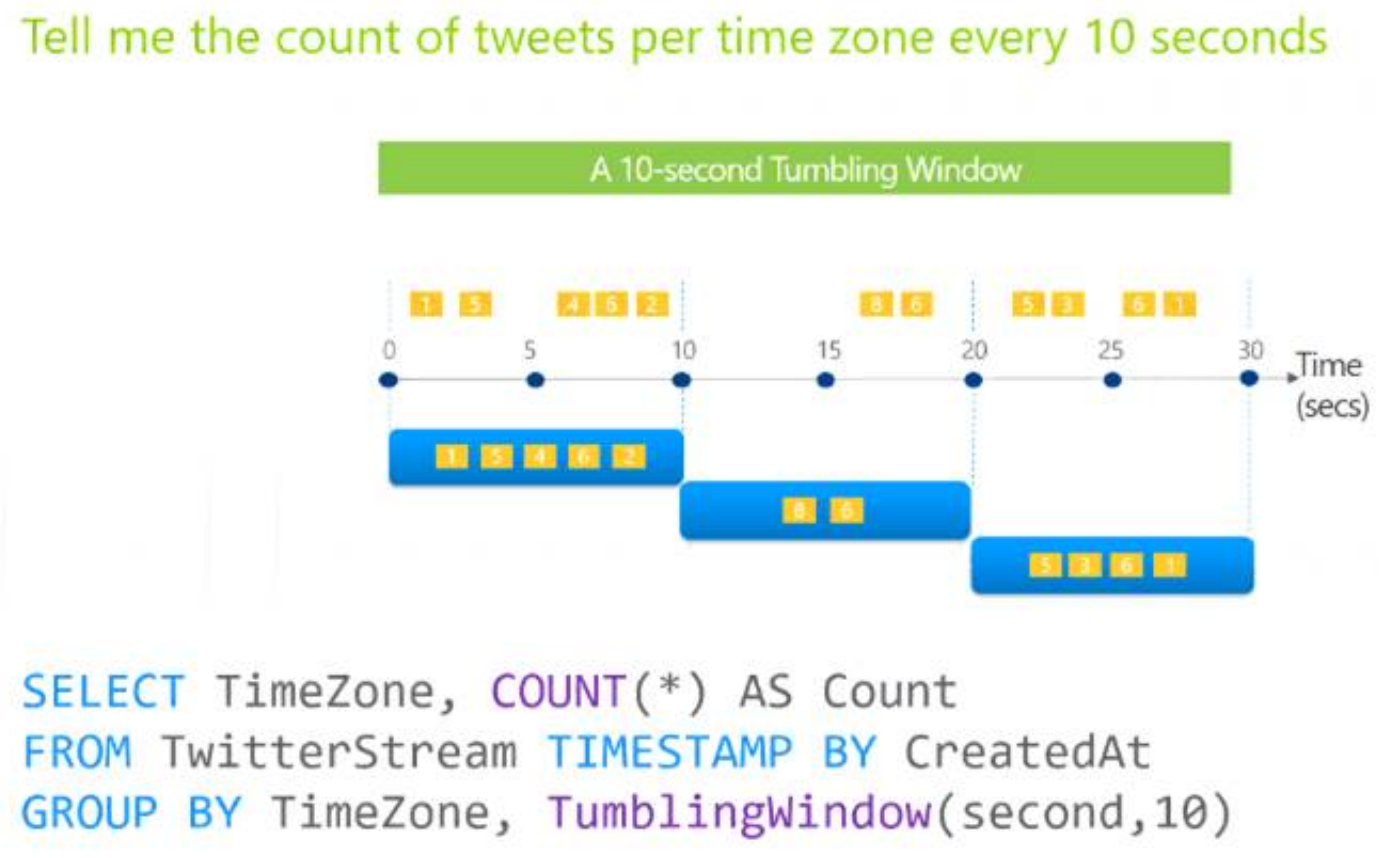
NEW QUESTION 190
- (Exam Topic 3)
You have an Azure Storage account that generates 200,000 new files daily. The file names have a format of {YYYY}/{MM}/{DD}/{HH}/{CustomerID}.csv.
You need to design an Azure Data Factory solution that will load new data from the storage account to an
Azure Data Lake once hourly. The solution must minimize load times and costs.
How should you configure the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

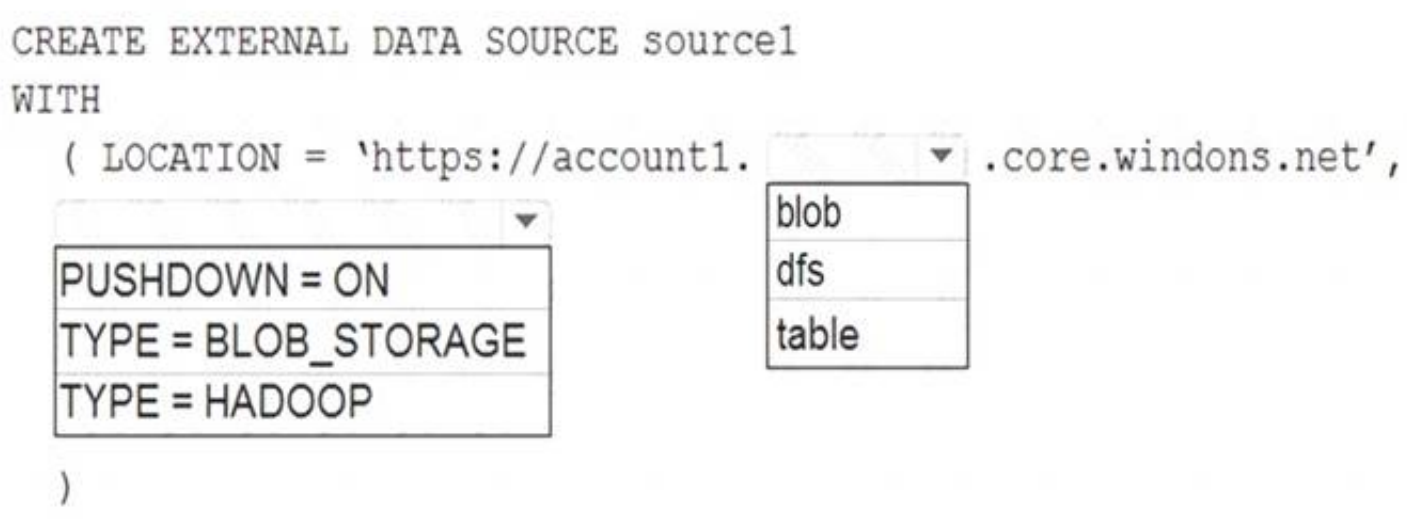
Answer: A

Explanation:
Table Description automatically generated
Box 1: Incremental load Box 2: Tumbling window
Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.
Timeline Description automatically generated



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

NEW QUESTION 191
- (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.



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

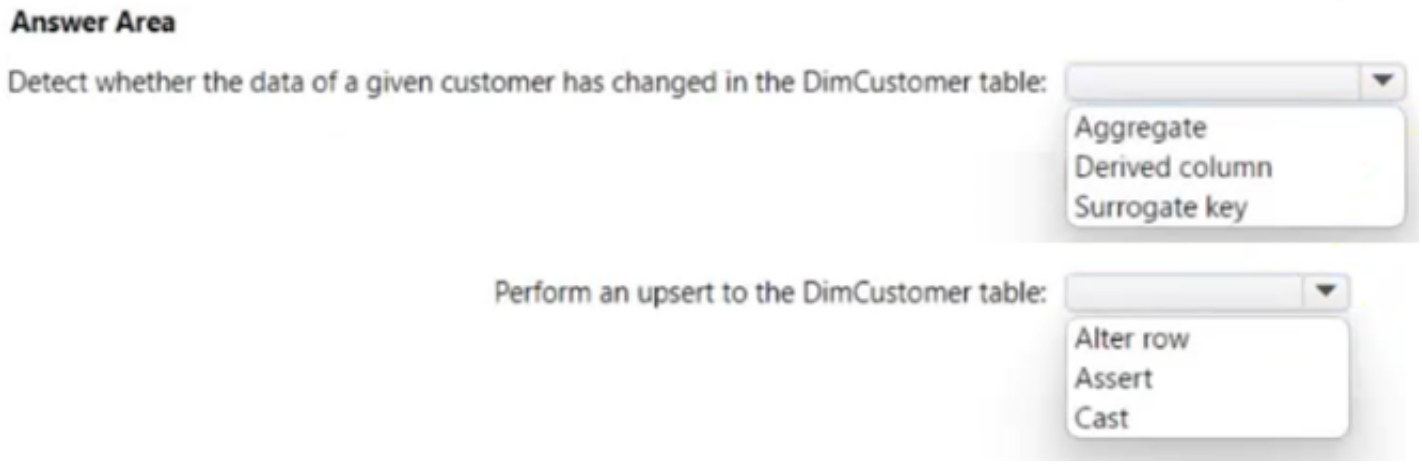
NEW QUESTION 196
- (Exam Topic 3)
You are developing an application that uses Azure Data Lake Storage Gen 2.
You need to recommend a solution to grant permissions to a specific application for a limited time period. What should you include in the recommendation?

A. Azure Active Directory (Azure AD) identities
B. shared access signatures (SAS)
C. account keys
D. role assignments

Answer: B

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

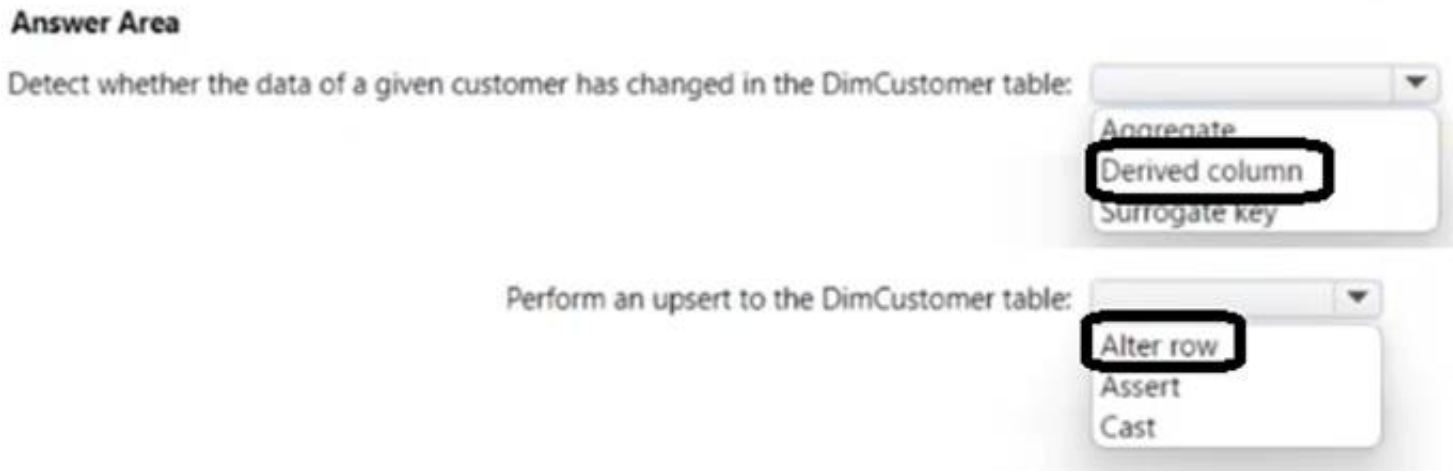
NEW QUESTION 199
- (Exam Topic 3)
You are developing an Azure Synapse Analytics pipeline that will include a mapping data flow named Dataflow1. Dataflow1 will read customer data from an external source and use a Type 1 slowly changing dimension (SCO) when loading the data into a table named DimCustomer1 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.



A. Mastered
B. Not Mastered

Answer: A

Explanation:



NEW QUESTION 202
- (Exam Topic 3)
You plan to ingest streaming social media data by using Azure Stream Analytics. The data will be stored in files in Azure Data Lake Storage, and then consumed by using Azure Datiabricks and PolyBase in Azure Synapse Analytics.
You need to recommend a Stream Analytics data output format to ensure that the queries from Databricks and PolyBase against the files encounter the fewest possible errors. The solution must ensure that the tiles can be queried quickly and that the data type information is retained.
What should you recommend?

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

Answer: A

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

NEW QUESTION 204
- (Exam Topic 3)
You have an Azure Data Factory version 2 (V2) resource named Df1. Df1 contains a linked service. You have an Azure Key vault named vault1 that contains an encryption key named key1.
You need to encrypt Df1 by using key1. What should you do first?

A. Add a private endpoint connection to vaul 1.
B. Enable Azure role-based access control on vault 1.
C. Remove the linked service from Df1.
D. Create a self-hosted integration runtime.

Answer: C

Explanation:
Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key https://docs.microsoft.com/en-us/azure/data-factory/concepts-linked-services
https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime

NEW QUESTION 209
- (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 211
- (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.

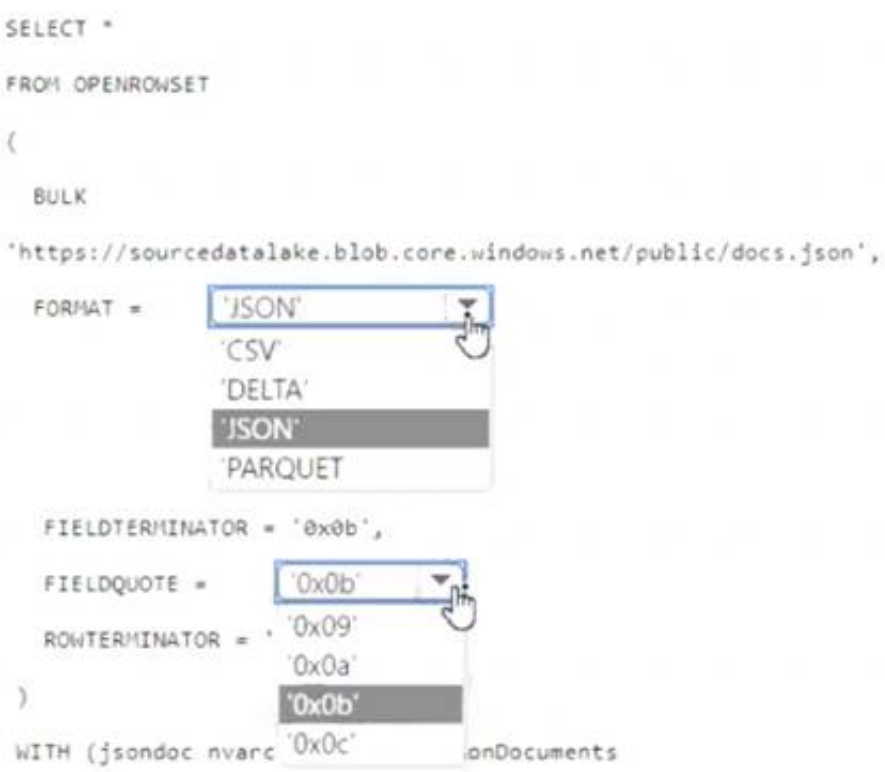
What should you recommend?

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

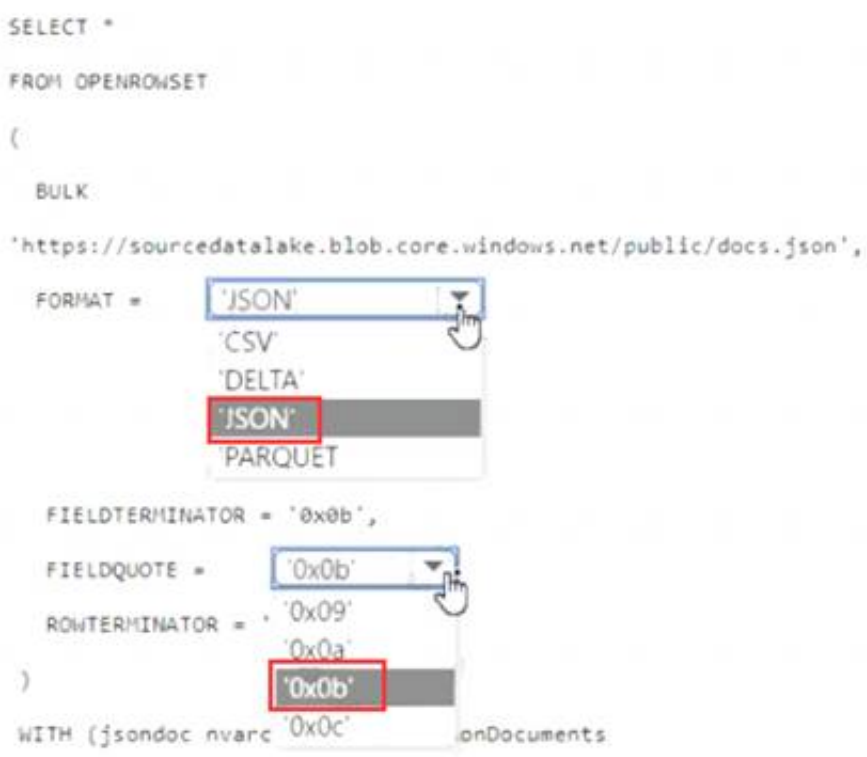
NEW QUESTION 216
- (Exam Topic 3)
You have an Azure Synapse serverless SQL pool.
You need to read JSON documents from a file by using the OPENROWSET function.
How should you complete the query? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.
Answer Area



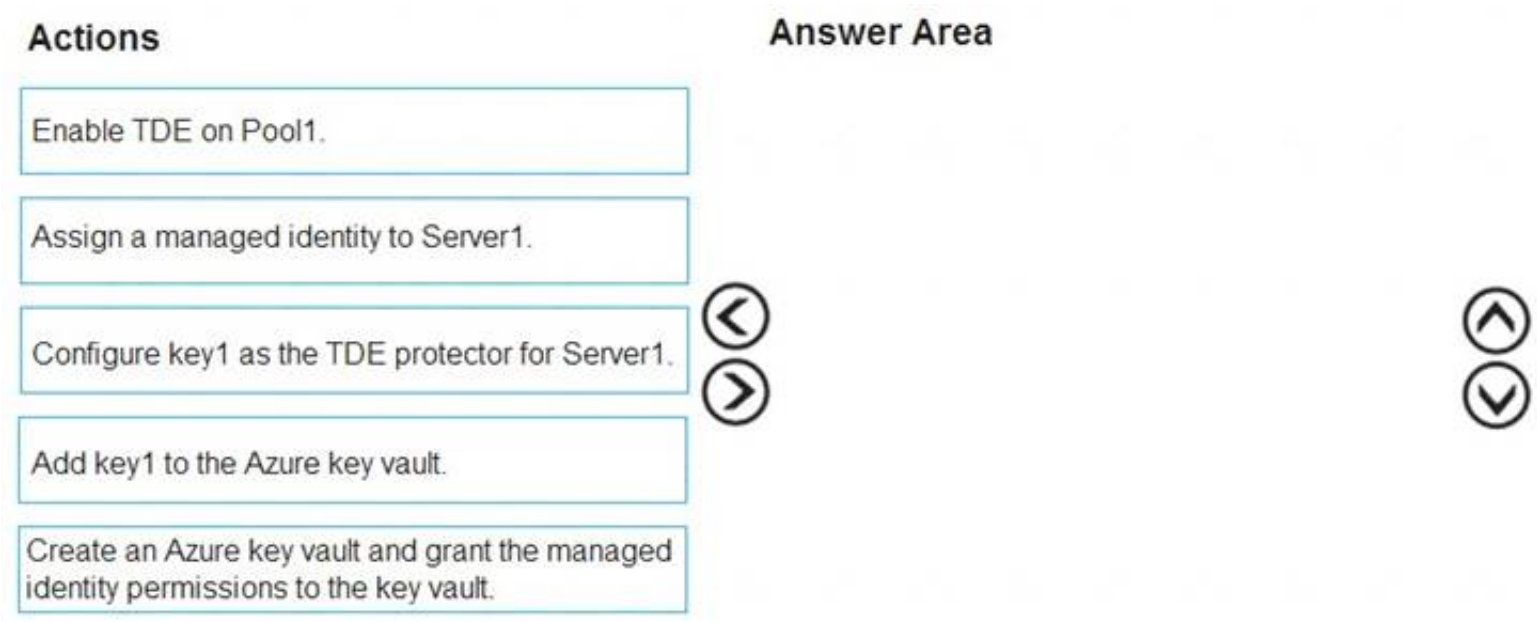
A. Mastered
B. Not Mastered

Answer: A

Explanation:
Answer Area



NEW QUESTION 219
- (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.



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 224
- (Exam Topic 3)
You have an Azure Storage account and a data warehouse in Azure Synapse Analytics in the UK South region. You need to copy blob data from the storage account to the data warehouse by using Azure Data Factory. The solution must meet the following requirements:

- Ensure that the data remains in the UK South region at all times.
- Minimize administrative effort.

Which type of integration runtime should you use?

A. Azure integration runtime
B. Azure-SSIS integration runtime
C. Self-hosted integration runtime

Answer: A

Explanation:

IR type	Public network	Private network
Azure	Data Flow Data movement Activity dispatch	
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

Reference:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

NEW QUESTION 229
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL Pool1. Pool1 contains a partitioned fact table named dbo.Sales and a staging table named stg.Sales that has the matching table and partition definitions.
You need to overwrite the content of the first partition in dbo.Sales with the content of the same partition in stg.Sales. The solution must minimize load times.
What should you do?

A. Switch the first partition from dbo.Sales to stg.Sales.
B. Switch the first partition from stg.Sales to db
C. Sales.

D. Update dbo.Sales from stg.Sales.
E. Insert the data from stg.Sales into dbo.Sales.

Answer: A

NEW QUESTION 233
- (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 234
- (Exam Topic 3)
You have an Azure subscription linked to an Azure Active Directory (Azure AD) tenant that contains a service principal named ServicePrincipal1. The subscription contains an Azure Data Lake Storage account named adls1. Adls1 contains a folder named Folder2 that has a URI of https://adls1.dfs.core.windows.net/container1/Folder1/Folder2/.
ServicePrincipal1 has the access control list (ACL) permissions shown in the following table.

Resource	Permission
container1	Access – Execute
Folder1	Access – Execute
Folder2	Access – Read

You need to ensure that ServicePrincipal1 can perform the following actions:
- Traverse child items that are created in Folder2.
- Read files that are created in Folder2.

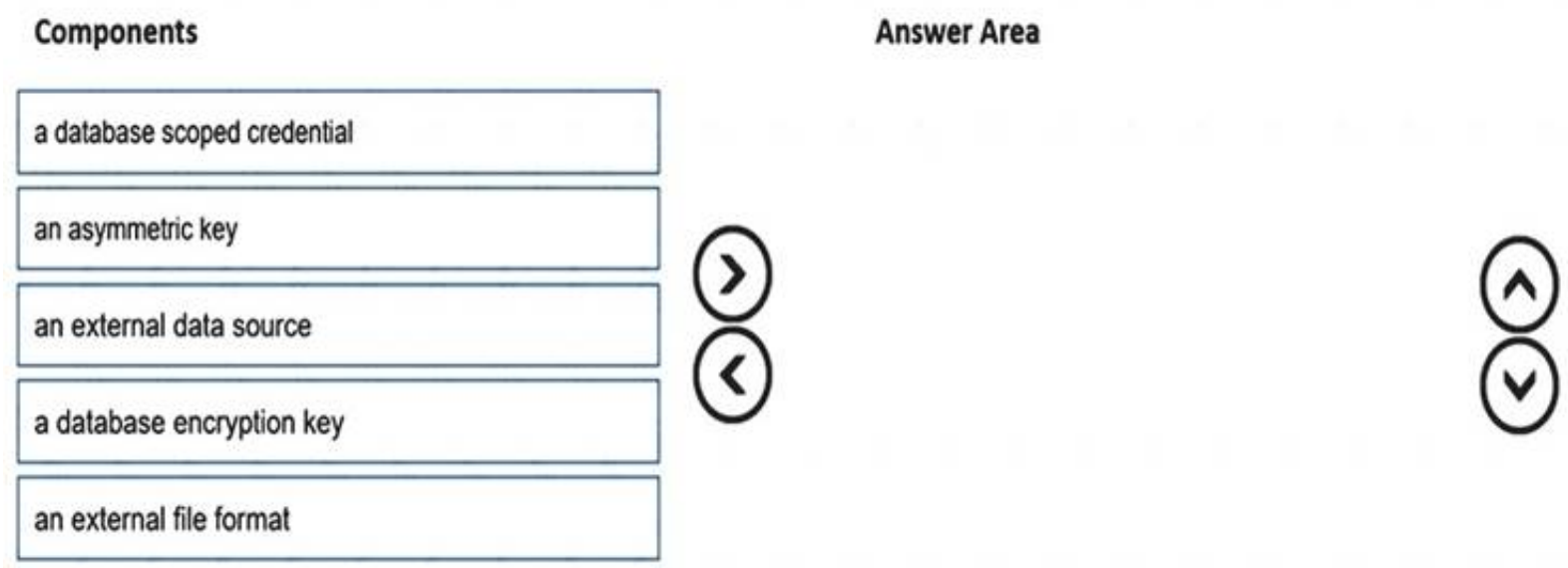
The solution must use the principle of least privilege.
Which two permissions should you grant to ServicePrincipal1 for Folder2? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Access - Read
B. Access - Write
C. Access - Execute
D. Default-Read
E. Default - Write
F. Default - Execute

Answer: DF

Explanation:
Execute (X) permission is required to traverse the child items of a folder.
There are two kinds of access control lists (ACLs), Access ACLs and Default ACLs. Access ACLs: These control access to an object. Files and folders both have Access ACLs.
Default ACLs: A "template" of ACLs associated with a folder that determine the Access ACLs for any child items that are created under that folder. Files do not have Default ACLs.
Reference:
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control

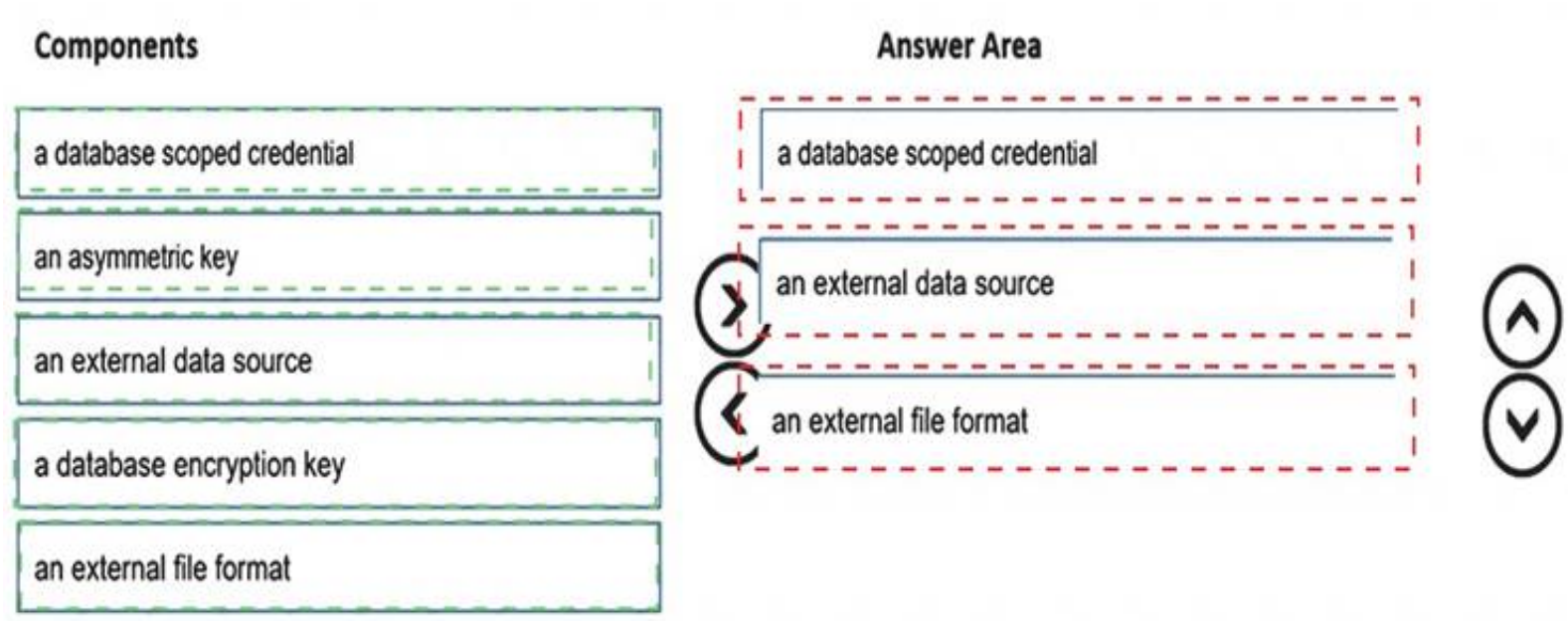
NEW QUESTION 237
- (Exam Topic 3)
You are responsible for providing access to an Azure Data Lake Storage Gen2 account.
Your user account has contributor access to the storage account, and you have the application ID and access key.
You plan to use PolyBase to load data into an enterprise data warehouse in Azure Synapse Analytics. You need to configure PolyBase to connect the data warehouse to storage account.
Which three components should you create in sequence? To answer, move the appropriate components from the list of components to the answer area and arrange them in the correct order.



A. Mastered
B. Not Mastered

Answer: A

Explanation:



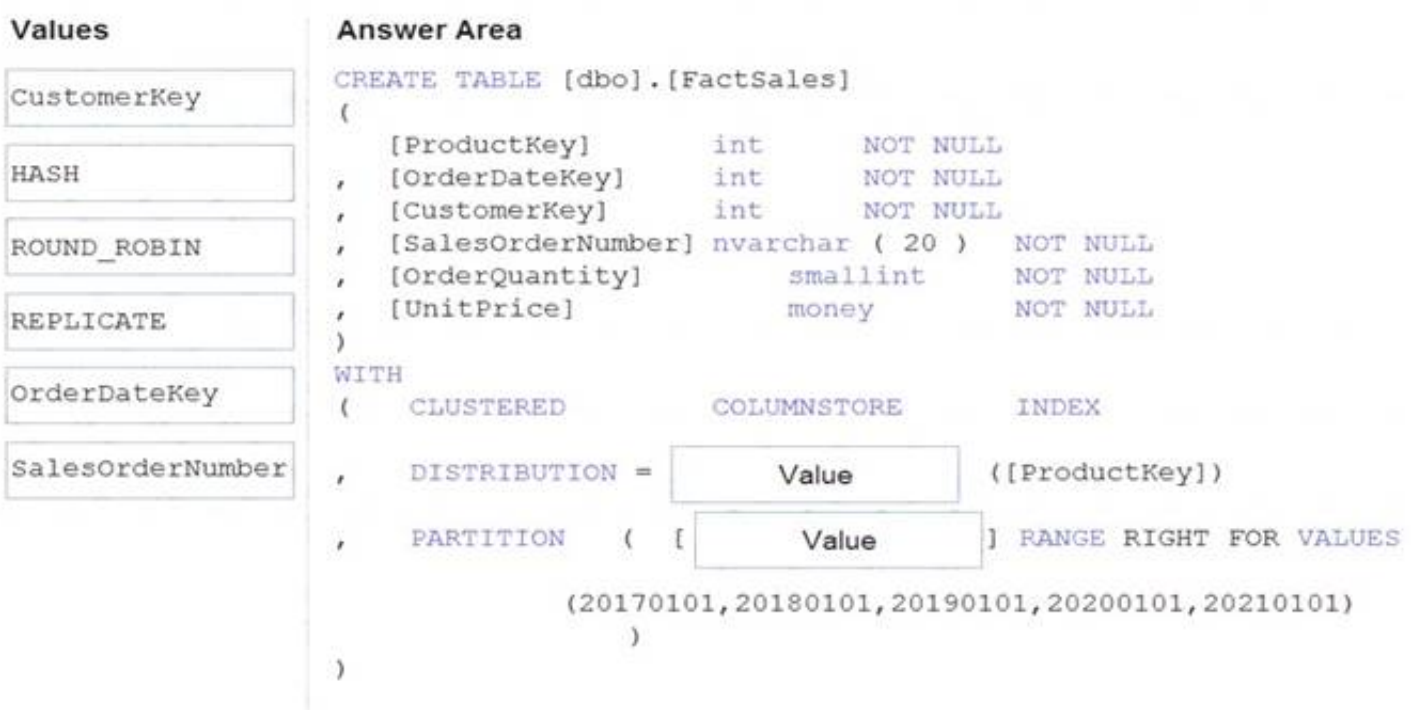
NEW QUESTION 238
- (Exam Topic 3)
You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container.
Which resource provider should you enable?

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 243
- (Exam Topic 3)
You plan to create a table in an Azure Synapse Analytics dedicated SQL pool.
Data in the table will be retained for five years. Once a year, data that is older than five years will be deleted. You need to ensure that the data is distributed evenly across partitions. The solution must minimize the
amount of time required to delete old data.
How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: HASH
Box 2: OrderDateKey
In most cases, table partitions are created on a date column.
A way to eliminate rollbacks is to use Metadata Only operations like partition switching for data management. For example, rather than execute a DELETE statement to delete all rows in a table where the order_date was in October of 2001, you could partition your data early. Then you can switch out the partition with data for an empty partition from another table.
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

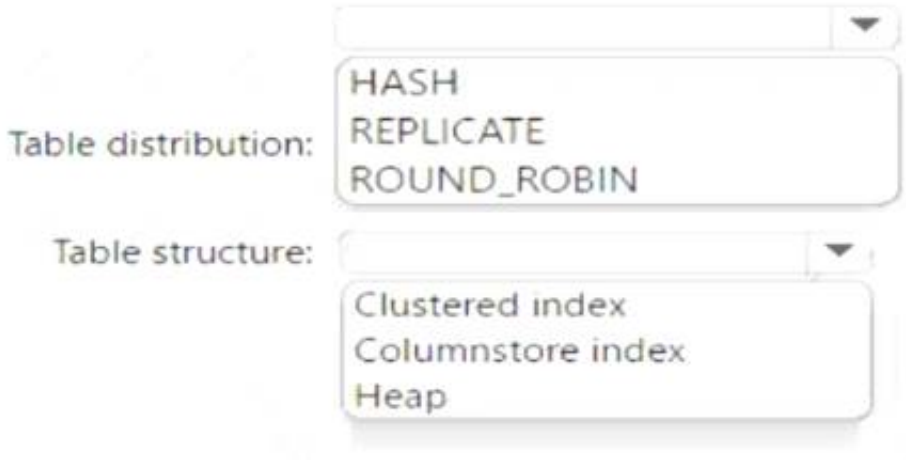
NEW QUESTION 248
- (Exam Topic 3)
You have an Azure Databricks workspace that contains a Delta Lake dimension table named Tablet. Table1 is a Type 2 slowly changing dimension (SCD) table. You need to apply updates from a source table to Table1. Which Apache Spark SQL operation should you use?

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 253
- (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.

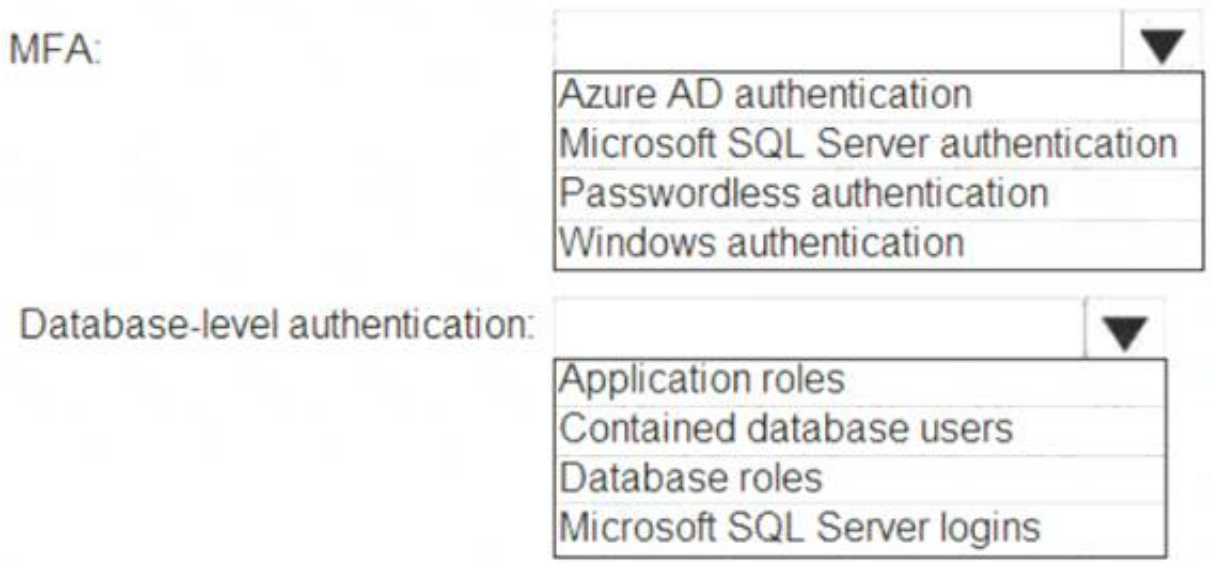


A. Mastered
B. Not Mastered

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.



A. Mastered
B. Not Mastered

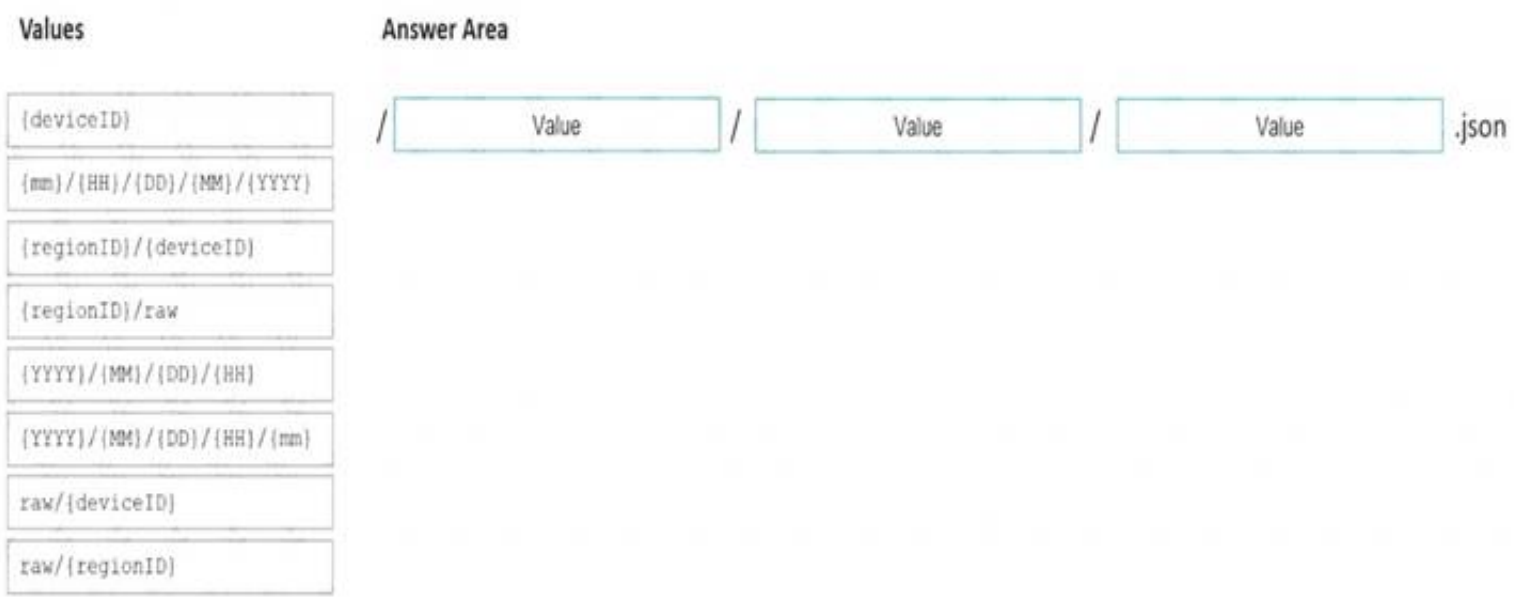
Answer: A

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 Data Lake Storage Gen2 structure for telemetry data from 25 million devices distributed across seven key geographical regions. Each minute, the devices will send a JSON payload of metrics to Azure Event Hubs.
You need to recommend a folder structure for the data. The solution must meet the following requirements:

- Data engineers from each region must be able to build their own pipelines for the data of their respective region only.
- The data must be processed at least once every 15 minutes for inclusion in Azure Synapse Analytics serverless SQL pools.

How should you recommend completing the structure? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 264
- (Exam Topic 3)
You have the following Azure Data Factory pipelines
• ingest Data from System 1
• Ingest Data from System2
• Populate Dimensions
• Populate facts
ingest Data from System1 and Ingest Data from System1 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest Data from System* Populate Facts must execute after the Populate Dimensions pipeline. All the pipelines must execute every eight hours.
What should you do to schedule the pipelines for execution?

A. Add an event trigger to all four pipelines.
B. Create a parent pipeline that contains the four pipelines and use an event trigger.
C. Create a parent pipeline that contains the four pipelines and use a schedule trigger.
D. Add a schedule trigger to all four pipelines.

Answer: C

Explanation:
Schedule trigger: A trigger that invokes a pipeline on a wall-clock schedule. Reference:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers

NEW QUESTION 266
- (Exam Topic 3)
You are designing an Azure Synapse Analytics workspace.
You need to recommend a solution to provide double encryption of all the data at rest.
Which two components should you include in the recommendation? Each coned answer presents part of the solution
NOTE: Each correct selection is worth one point.

A. an X509 certificate
B. an RSA key
C. an Azure key vault that has purge protection enabled
D. an Azure virtual network that has a network security group (NSG)
E. an Azure Policy initiative

Answer: BC

Explanation:
Synapse workspaces encryption uses existing keys or new keys generated in Azure Key Vault. A single key is used to encrypt all the data in a workspace.
Synapse workspaces support RSA 2048 and 3072 byte-sized keys, and RSA-HSM keys.
The Key Vault itself needs to have purge protection enabled. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption

NEW QUESTION 271
- (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 272
- (Exam Topic 3)
You plan to create a dimension table in Azure Synapse Analytics that will be less than 1 GB. You need to create the table to meet the following requirements:
• Provide the fastest Query time.
• Minimize data movement during queries. Which type of table should you use?

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

Answer: C

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

NEW QUESTION 276
- (Exam Topic 3)
You have an Apache Spark DataFrame named temperatures. A sample of the data is shown in the following table.

Date	Temp
...	...
18-01-2021	3
19-01-2021	4
20-01-2021	2
21-01-2021	2
...	...

You need to produce the following table by using a Spark SQL query.

Year	JAN	FEB	MAR	APR	MAY
2019	2.3	4.1	5.2	7.6	9.2
2020	2.4	4.2	4.9	7.8	9.1
2021	2.6	5.3	3.4	7.9	9.5

How should you complete the query? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all.
You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.

Values

CAST
COLLATE
CONVERT
FLATTEN
PIVOT
UNPIVOT

Answer Area

```
SELECT * FROM (
 SELECT YEAR(Date) Year, MONTH(Date) Month, Temp
 FROM temperatures
 WHERE date BETWEEN DATE '2019-01-01' AND DATE '2021-08-31'
)
[          ](
 AVG ([          ](Temp AS DECIMAL(4, 1)))
 FOR Month in (
  1 JAN, 2 FEB, 3 MAR, 4 APR, 5 MAY, 6 JUN,
  7 JUL, 8 AUG, 9 SEP, 10 OCT, 11 NOV, 12 DEC
            )
 )
 ORDER BY Year ASC
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Text Description automatically generated
Box 1: PIVOT
PIVOT rotates a table-valued expression by turning the unique values from one column in the expression into multiple columns in the output. And PIVOT runs aggregations where they're required on any remaining column values that are wanted in the final output.
Reference:
https://learnsql.com/cookbook/how-to-convert-an-integer-to-a-decimal-in-sql-server/ https://docs.microsoft.com/en-us/sql/t-sql/queries/from-using-pivot-and-unpivot

NEW QUESTION 277
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Data Lake Storage account named myaccount1. The myaccount1 account contains two containers named container1 and contained. The subscription is linked to an Azure Active Directory (Azure AD) tenant that contains a security group named Group1.
You need to grant Group1 read access to contamer1. The solution must use the principle of least privilege. Which role should you assign to Group1?

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

Answer: A

NEW QUESTION 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	2760	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
1417	2832	728	40	2064	1	57
0	2232	496	0	1736	1	58
384	2632	560	32	2040	1	59
225	2760	544	40	2184	1	60

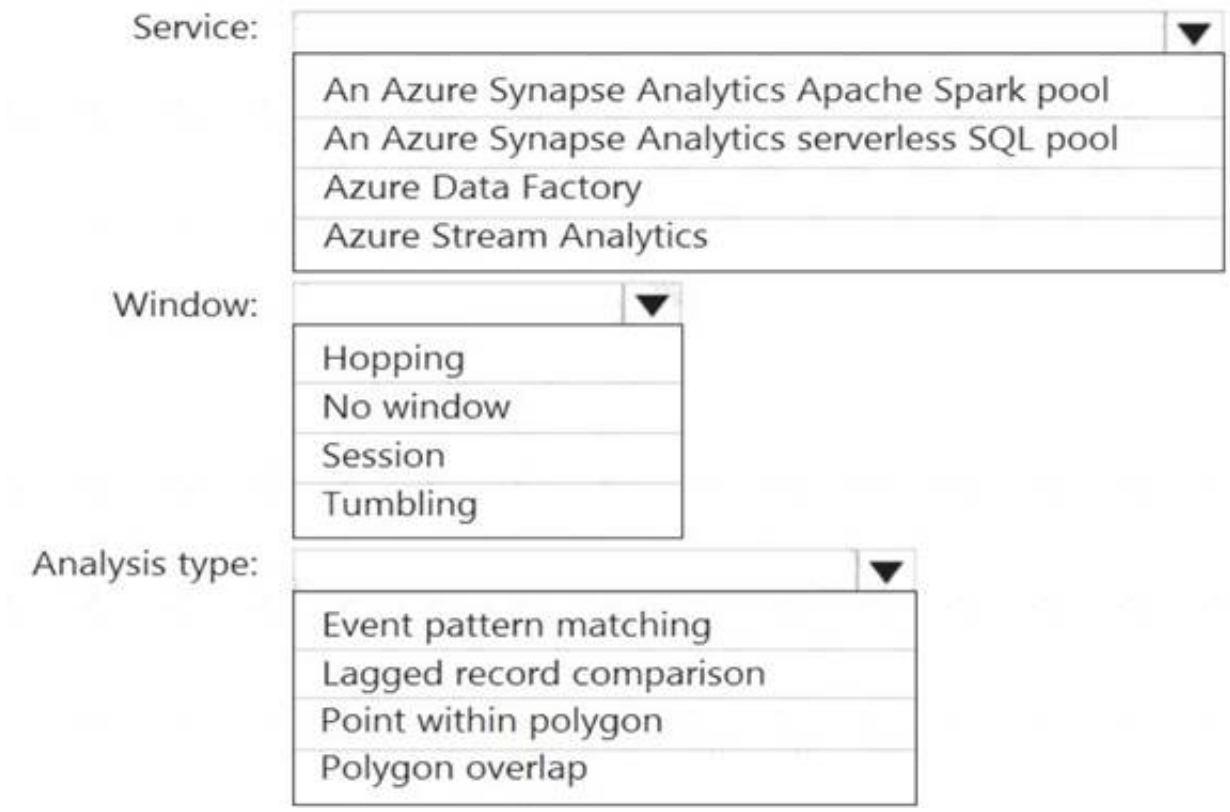
Which statement accurately describes the dbo,FactInternetSales table?

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 282
- (Exam Topic 3)
You are designing a monitoring solution for a fleet of 500 vehicles. Each vehicle has a GPS tracking device that sends data to an Azure event hub once per minute.
You have a CSV file in an Azure Data Lake Storage Gen2 container. The file maintains the expected geographical area in which each vehicle should be.
You need to ensure that when a GPS position is outside the expected area, a message is added to another event hub for processing within 30 seconds. The solution must minimize cost.
What should you include in the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



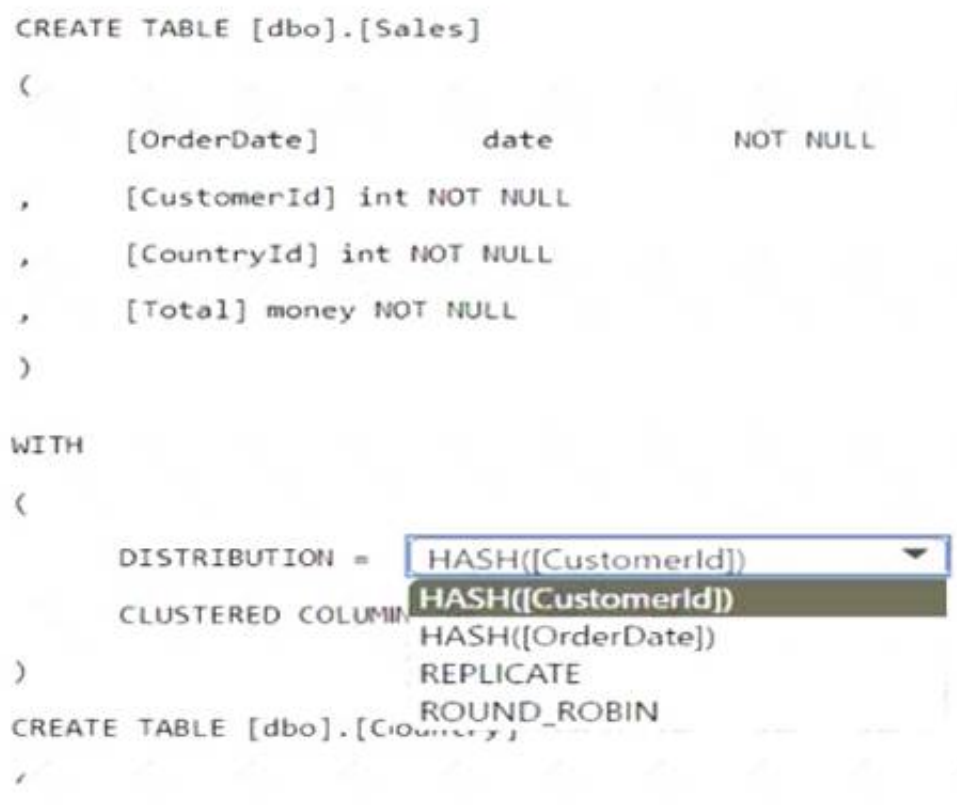
A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Azure Stream Analytics Box 2: Hopping
Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.
Box 3: Point within polygon Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

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



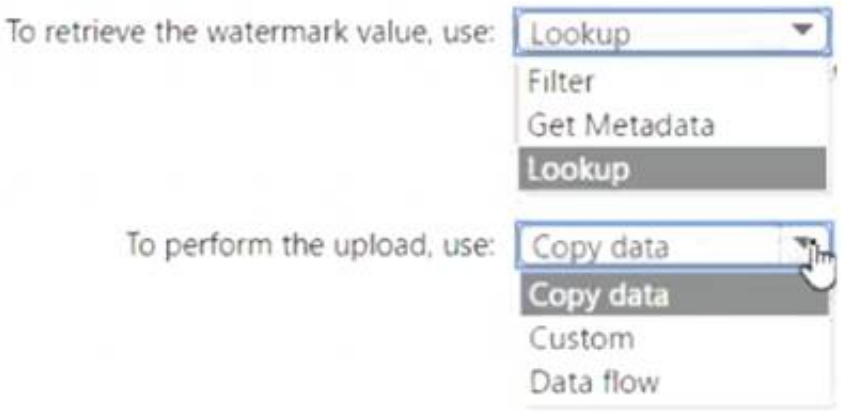
A. Mastered
B. Not Mastered

Answer: A

Explanation:
Answer Area

NEW QUESTION 288
- (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.

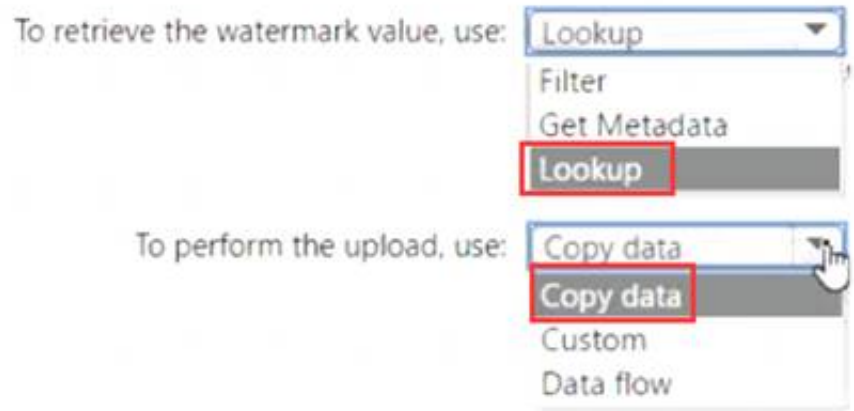
Answer Area



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Answer Area



NEW QUESTION 291
.....

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questons and Answers in PDF Format

DP-203 Practice Exam Features:

* DP-203 Questions and Answers Updated Frequently

* DP-203 Practice Questions Verified by Expert Senior Certified Staff

* DP-203 Most Realistic Questions that Guarantee you a Pass on Your FirstTry

* DP-203 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

100% Actual & Verified — Instant Download, Please Click
Order The DP-203 Practice Test Here