Microsoft

Exam Questions DP-203

Data Engineering on Microsoft Azure

NEW QUESTION 1
- (Exam Topic 3)
You are designing a data mart for the human resources (MR) department at your company. The data mart will contain information and employee transactions.
From a source system you have a flat extract that has the following fields:
• EmployeeID
• FirstName
• LastName
• Recipient
• GrossArnount
• TransactionID
• GovernmentID
• NetAmountPaid
• TransactionDate
You need to design a start schema data model in an Azure Synapse analytics dedicated SQL pool for the data mart.
Which two tables should you create? Each Correct answer present part of the solution.

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

Answer: AE

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

NEW QUESTION 2
- (Exam Topic 3)
A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU).
You need to optimize performance for the Azure Stream Analytics job.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

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

Answer: DF

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

NEW QUESTION 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?

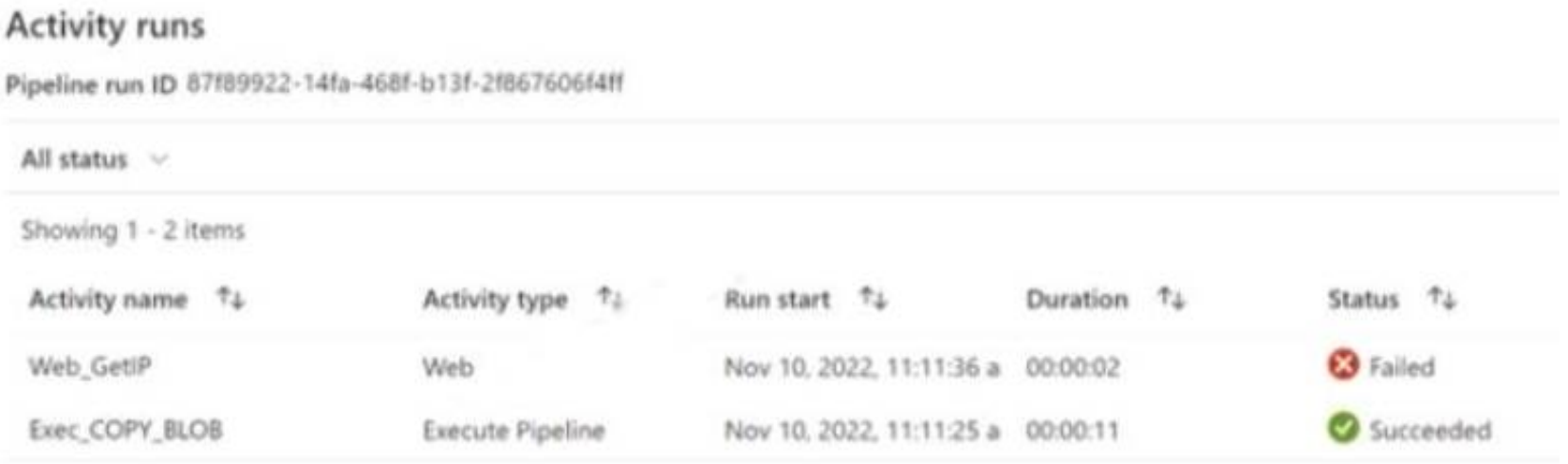
A. Store all the data as strings in the Parquet tiles.
B. Use OPENROWEST to query the Parquet files.
C. Create an external table mat contains a subset of columns from the Parquet files.
D. Use Snappy compression for the files.

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 Data Factory pipeline shown the following exhibit.

The execution log for the first pipeline run is shown in the following exhibit.



The execution log for the second pipeline run is shown in the following exhibit.

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

Answer Area

Statements	Yes	No
The `Retry` property of the Web_GetIP activity is set to `1`.	○	○
The `waitOnCompletion` property of the Exec_COPY_BLOB activity is set to `true`.	○	○
The Exec_COPY_BLOB activity was skipped during the second run due to pipeline dependencies.	○	○

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

Statements	Yes	No
The `Retry` property of the Web_GetIP activity is set to `1`.	○	○
The `waitOnCompletion` property of the Exec_COPY_BLOB activity is set to `true`.	○	○
The Exec_COPY_BLOB activity was skipped during the second run due to pipeline dependencies.	○	○

NEW QUESTION 5
- (Exam Topic 3)
You are designing a date dimension table in an Azure Synapse Analytics dedicated SQL pool. The date dimension table will be used by all the fact tables.
Which distribution type should you recommend to minimize data movement?

A. HASH
B. REPLICATE
C. ROUND ROBIN

Answer: B

Explanation:
A replicated table has a full copy of the table available on every Compute node. Queries run fast on replicated tables since joins on replicated tables don't require data movement. Replication requires extra storage, though, and isn't practical for large tables.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overvie

NEW QUESTION 6
- (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 7
- (Exam Topic 3)
You manage an enterprise data warehouse in Azure Synapse Analytics.
Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries.
You need to monitor resource utilization to determine the source of the performance issues. Which metric should you monitor?

A. Data IO percentage
B. Local tempdb percentage
C. Cache used percentage
D. DWU percentage

Answer: C

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

NEW QUESTION 8
- (Exam Topic 3)
You have several Azure Data Factory pipelines that contain a mix of the following types of activities.
* Wrangling data flow
* Notebook
* Copy
* jar
Which two Azure services should you use to debug the activities? Each correct answer presents part of the solution NOTE: Each correct selection is worth one point.

A. Azure HDInsight
B. Azure Databricks
C. Azure Machine Learning
D. Azure Data Factory
E. Azure Synapse Analytics

Answer: CE

NEW QUESTION 9
- (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 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 two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.

Table	Column
Flight	ArrivalAirportID ArrivalDateTime
Weather	AirportID ReportDateTime

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

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

Answer: D

NEW QUESTION 15
- (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

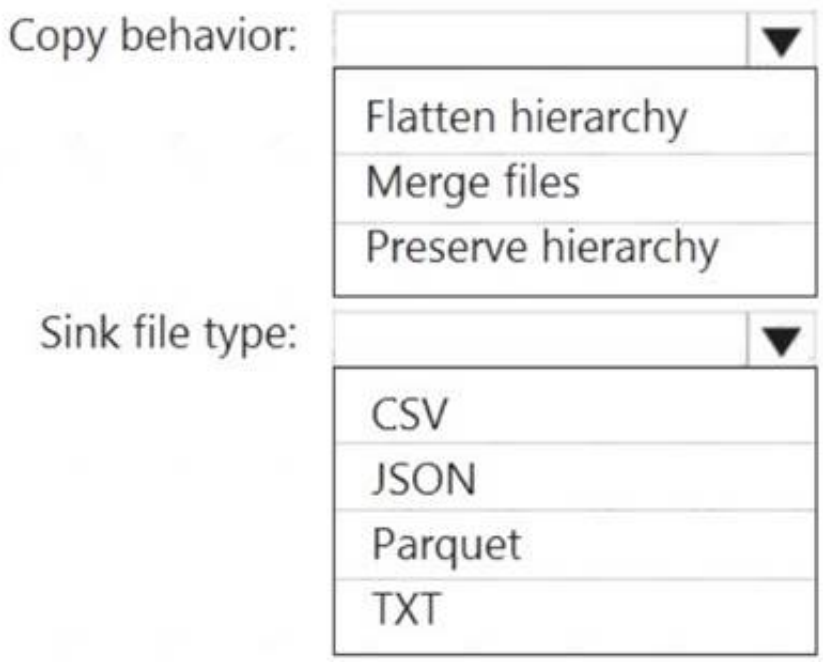
NEW QUESTION 19
- (Exam Topic 3)
You are designing an anomaly detection solution for streaming data from an Azure IoT hub. The solution must meet the following requirements:
- Send the output to Azure Synapse.
- Identify spikes and dips in time series data.
- Minimize development and configuration effort. 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 24
- (Exam Topic 3)
You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools. Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company.
You need to move the files to a different folder and transform the data to meet the following requirements: Provide the fastest possible query times.
Automatically infer the schema from the underlying files.
How should you configure the Data Factory copy activity? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

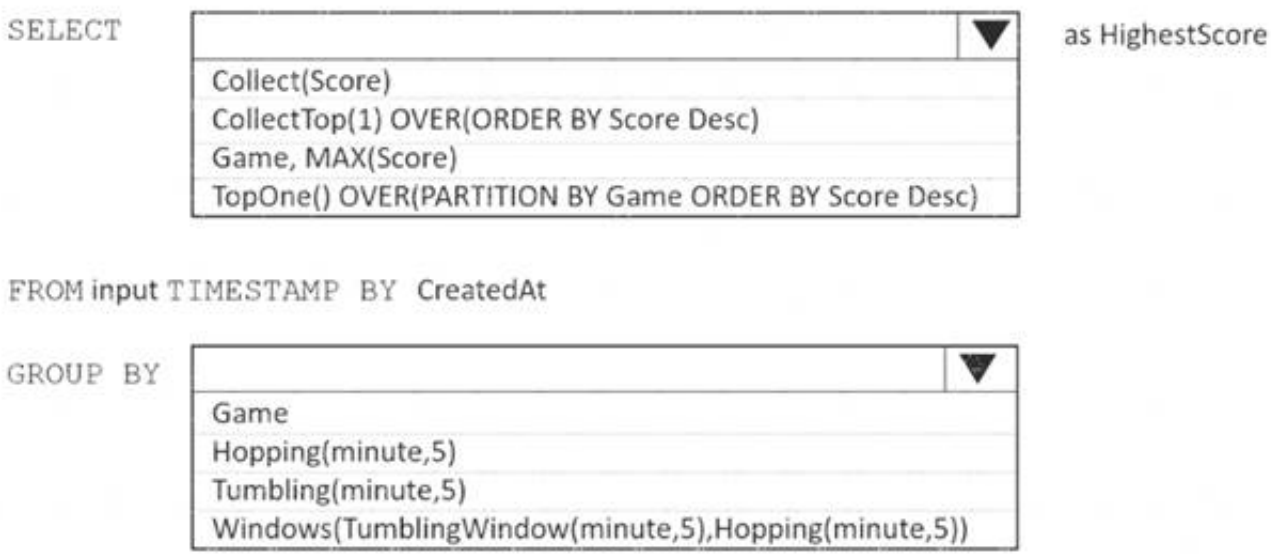


A. Mastered
B. Not Mastered

Answer: A

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

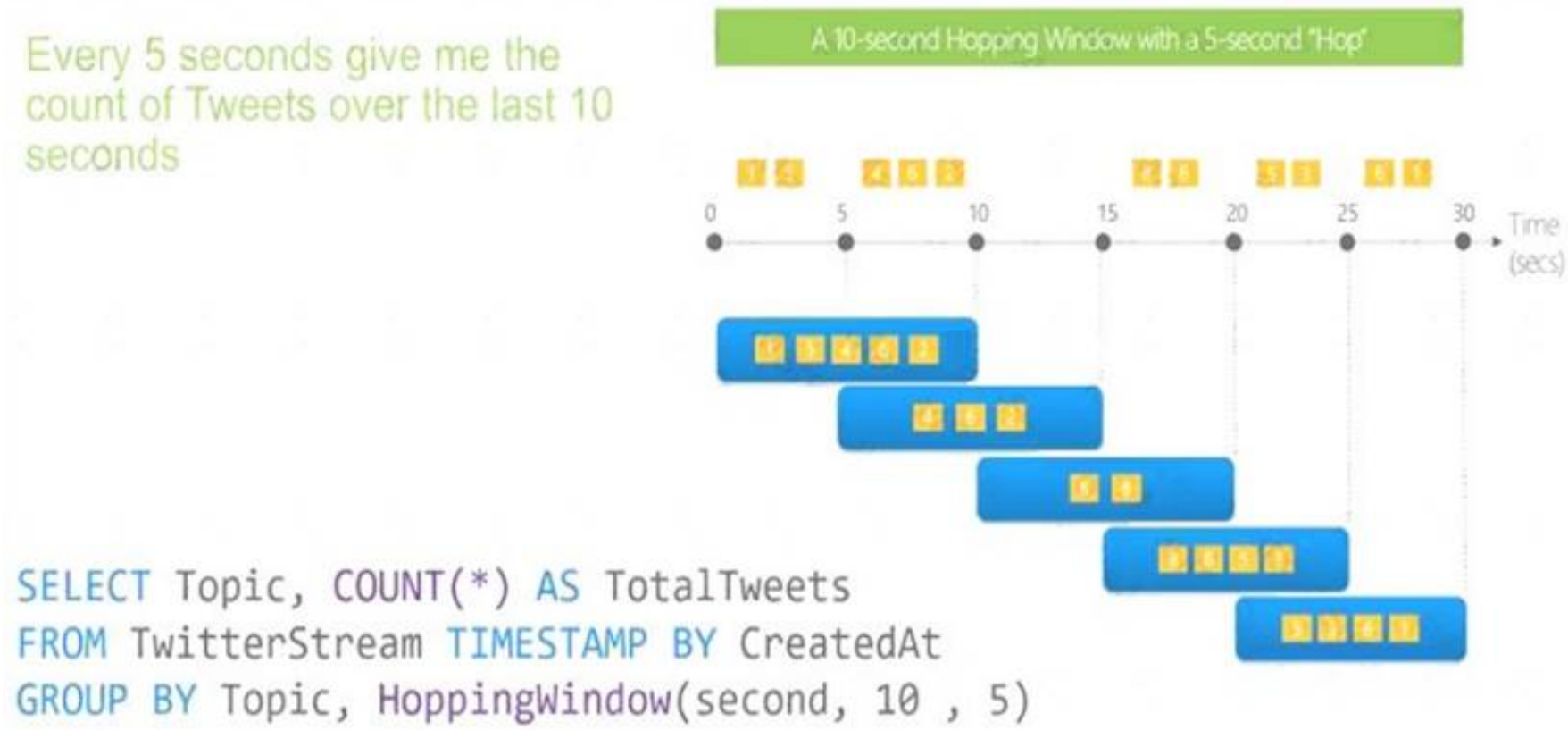
NEW QUESTION 26
- (Exam Topic 3)
You are building an Azure Stream Analytics job to retrieve game data.
You need to ensure that the job returns the highest scoring record for each five-minute time interval of each game.
How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



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 31
- (Exam Topic 3)
You have an Azure Data Factory pipeline that contains a data flow. The data flow contains the following expression.

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
See below answer.
Answer Area

Number of columns: 22

Number of rows: 4

NEW QUESTION 32
- (Exam Topic 3)
You have an Azure Stream Analytics query. The query returns a result set that contains 10,000 distinct values for a column named clusterID.
You monitor the Stream Analytics job and discover high latency. You need to reduce the latency.
Which two actions should you perform? Each correct answer presents a complete solution. NOTE: Each correct selection is worth one point.

A. Add a pass-through query.
B. Add a temporal analytic function.
C. Scale out the query by using PARTITION BY.
D. Convert the query to a reference query.
E. Increase the number of streaming units.

Answer: CE

Explanation:
C: Scaling a Stream Analytics job takes advantage of partitions in the input or output. Partitioning lets you divide data into subsets based on a partition key. A process that consumes the data (such as a Streaming Analytics job) can consume and write different partitions in parallel, which increases throughput.
E: 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. This capacity lets you focus on the query logic and abstracts the need to manage the hardware to run your Stream Analytics job in a timely manner.
References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption

NEW QUESTION 37
- (Exam Topic 3)
You have an Azure data factory that connects to a Microsoft Purview account. The data factory is registered in Microsoft Purview.
You update a Data Factory pipeline.
You need to ensure that the updated lineage is available in Microsoft Purview.
What You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1.
You need to use an Azure Synapse pipeline to copy data from the selected tables in DB1 to the files in storage1. The solution must meet the following requirements:
• The Copy activity in the pipeline must be parameterized to use the data in File1.txt to identify the source and destination of the copy.
• Copy activities must occur in parallel as often as possible.
Which two pipeline activities should you include in the pipeline? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. If Condition
B. ForEach
C. Lookup
D. Get Metadata

Answer: CD

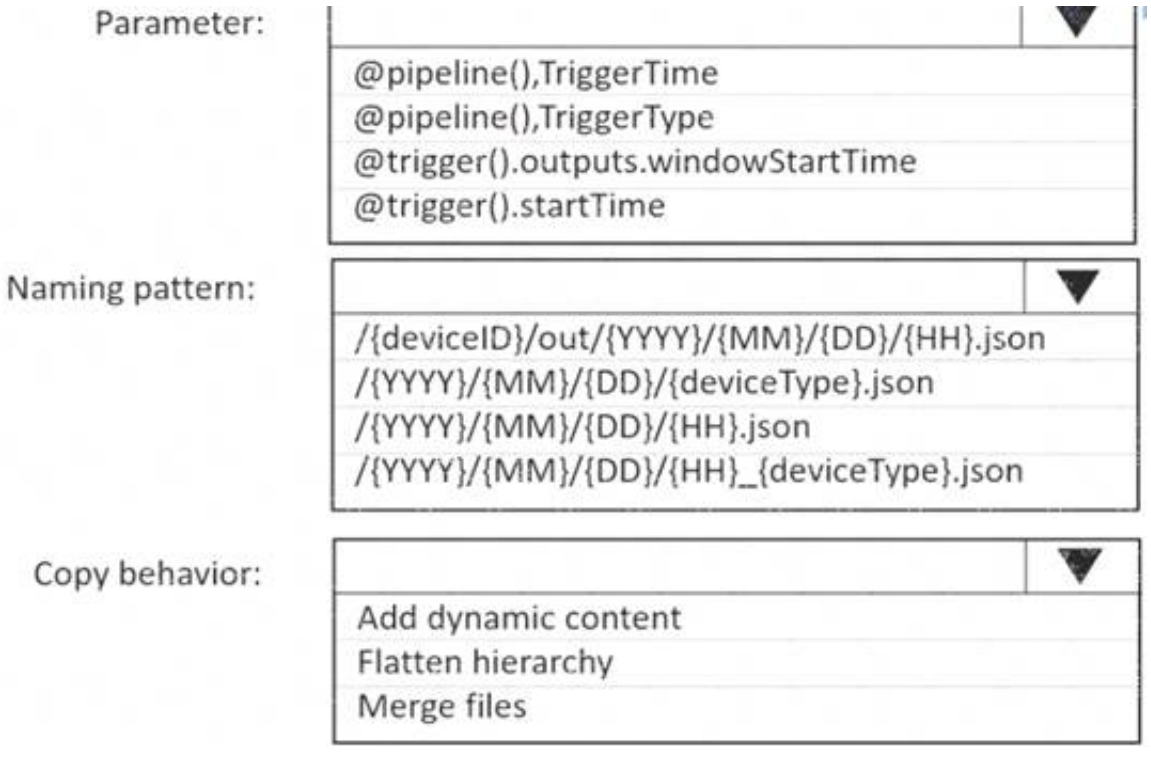
NEW QUESTION 39
- (Exam Topic 3)
You have an Azure Data Lake Storage account that contains a staging zone.
You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data Flow, and then inserts the data info the data warehouse.
Does this meet the goal?

A. Yes
B. No

Answer: B

Explanation:
If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity, not a mapping flow,5 with your own data processing logic and use the activity in the pipeline. You can create a custom activity to run R scripts on your HDInsight cluster with R installed.
Reference:
https://docs.microsoft.com/en-US/azure/data-factory/transform-data

NEW QUESTION 43
- (Exam Topic 3)
You are building an Azure Data Factory solution to process data received from Azure Event Hubs, and then ingested into an Azure Data Lake Storage Gen2 container.
The data will be ingested every five minutes from devices into JSON files. The files have the following naming pattern.
/{deviceType}/in/{YYYY}/{MM}/{DD}/{HH}/{deviceID}_{YYYY}{MM}{DD}HH}{mm}.json
You need to prepare the data for batch data processing so that there is one dataset per hour per deviceType. The solution must minimize read times.
How should you configure the sink for the copy activity? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: @trigger().startTime
startTime: A date-time value. For basic schedules, the value of the startTime property applies to the first occurrence. For complex schedules, the trigger starts no sooner than the specified startTime value.
Box 2: /{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json One dataset per hour per deviceType.
Box 3: Flatten hierarchy
- FlattenHierarchy: All files from the source folder are in the first level of the target folder. The target files have autogenerated names.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers https://docs.microsoft.com/en-us/azure/data-factory/connector-file-system

NEW QUESTION 44
- (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 48
- (Exam Topic 3)
You have the following Azure Stream Analytics query.

```
WITH

step1 AS (SELECT *
     FROM input1
     PARTITION BY StateID
     INTO 10),
step2 AS (SELECT *
     FROM input2
     PARTITION BY StateID
     INTO 10)

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

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

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: No
Note: You can now use a new extension of Azure Stream Analytics SQL to specify the number of partitions of a stream when reshuffling the data.
The outcome is a stream that has the same partition scheme. Please see below for an example: WITH step1 AS (SELECT * FROM [input1] PARTITION BY DeviceID INTO 10),
step2 AS (SELECT * FROM [input2] PARTITION BY DeviceID INTO 10)
SELECT * INTO [output] FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceID Note: The new extension of Azure Stream Analytics SQL includes a keyword INTO that allows you to specify the number of partitions for a stream when performing reshuffling using a PARTITION BY statement.
Box 2: Yes
When joining two streams of data explicitly repartitioned, these streams must have the same partition key and partition count. Box 3: Yes
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.
In general, the best practice is to start with 6 SUs for queries that don't use PARTITION BY. Here there are 10 partitions, so 6x10 = 60 SUs is good.
Note: Remember, Streaming Unit (SU) count, which is the unit of scale for Azure Stream Analytics, must be adjusted so the number of physical resources available to the job can fit the partitioned flow. In general, six SUs is a good number to assign to each partition. In case there are insufficient resources assigned to the job, the system will only apply the repartition if it benefits the job.
Reference:
https://azure.microsoft.com/en-in/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/ https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption

NEW QUESTION 52
- (Exam Topic 3)
You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1. You need to determine the size of the transaction log file for each distribution of DW1.
What should you do?

A. On DW1, execute a query against the sys.database_files dynamic management view.
B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.
C. Execute a query against the logs of DW1 by using theGet-AzOperationalInsightsSearchResult PowerShell cmdlet.
D. On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.

Answer: A

Explanation:
For information about the current log file size, its maximum size, and the autogrow option for the file, you can also use the size, max_size, and growth columns for that log file in sys.database_files.
Reference:
https://docs.microsoft.com/en-us/sql/relational-databases/logs/manage-the-size-of-the-transaction-log-file

NEW QUESTION 53
- (Exam Topic 3)
You are planning a streaming data solution that will use Azure Databricks. The solution will stream sales transaction data from an online store. The solution has the following specifications:
* The output data will contain items purchased, quantity, line total sales amount, and line total tax amount.
* Line total sales amount and line total tax amount will be aggregated in Databricks.
* Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.
You need to recommend an output mode for the dataset that will be processed by using Structured Streaming. The solution must minimize duplicate data.
What should you recommend?

A. Append
B. Update
C. Complete

Answer: B

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

NEW QUESTION 55
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 contains a dedicated SQL pool named SQL Pool and an Apache Spark pool named sparkpool. Sparkpool1 contains a DataFrame named pyspark.df.
You need to write the contents of pyspark_df to a tabte in SQLPooM by using a PySpark notebook. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.
Answer Area

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Answer Area

NEW QUESTION 58
- (Exam Topic 3)
You have an Azure Databricks workspace named workspace! in the Standard pricing tier. Workspace1 contains an all-purpose cluster named cluster). You need to reduce the time it takes for cluster 1 to start and scale up. The solution must minimize costs. What should you do first?

A. Upgrade workspace! to the Premium pricing tier.
B. Create a cluster policy in workspace1.
C. Create a pool in workspace1.
D. Configure a global init script for workspace1.

Answer: C

Explanation:
You can use Databricks Pools to Speed up your Data Pipelines and Scale Clusters Quickly.
Databricks Pools, a managed cache of virtual machine instances that enables clusters to start and scale 4 times faster.
Reference:
https://databricks.com/blog/2019/11/11/databricks-pools-speed-up-data-pipelines.html

NEW QUESTION 59
- (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

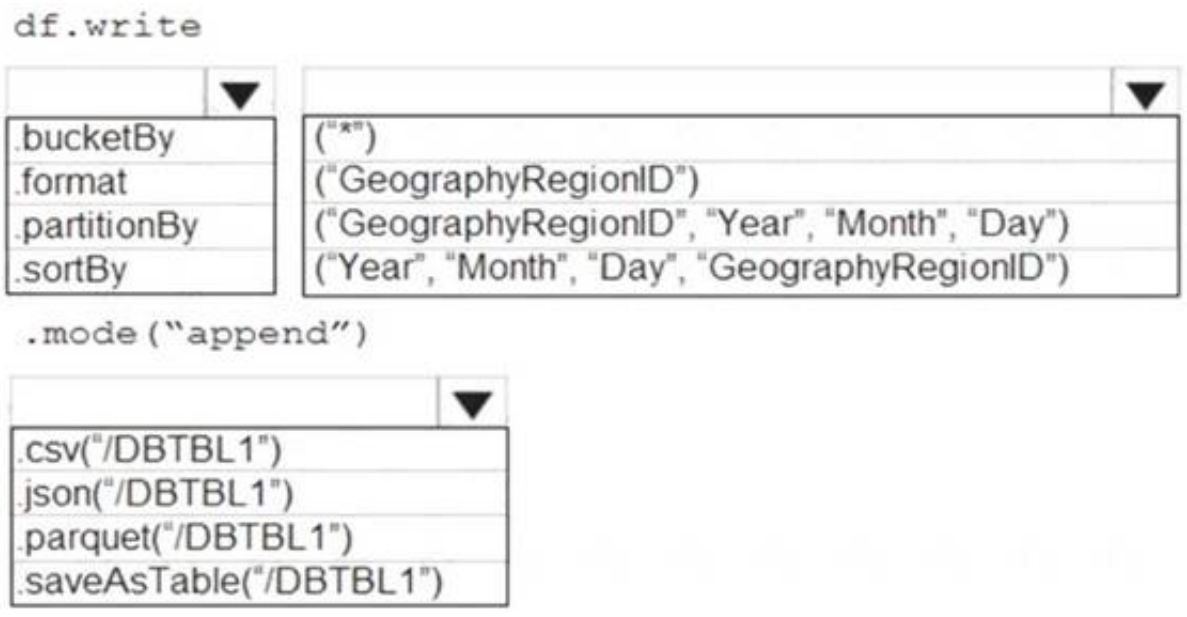
Answer: A

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

NEW QUESTION 62
- (Exam Topic 3)
You develop a dataset named DBTBL1 by using Azure Databricks. DBTBL1 contains the following columns:

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

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



A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 66
- (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 hopping window that uses a hop size of 10 seconds and a window size of 10 seconds. Does this meet the goal?

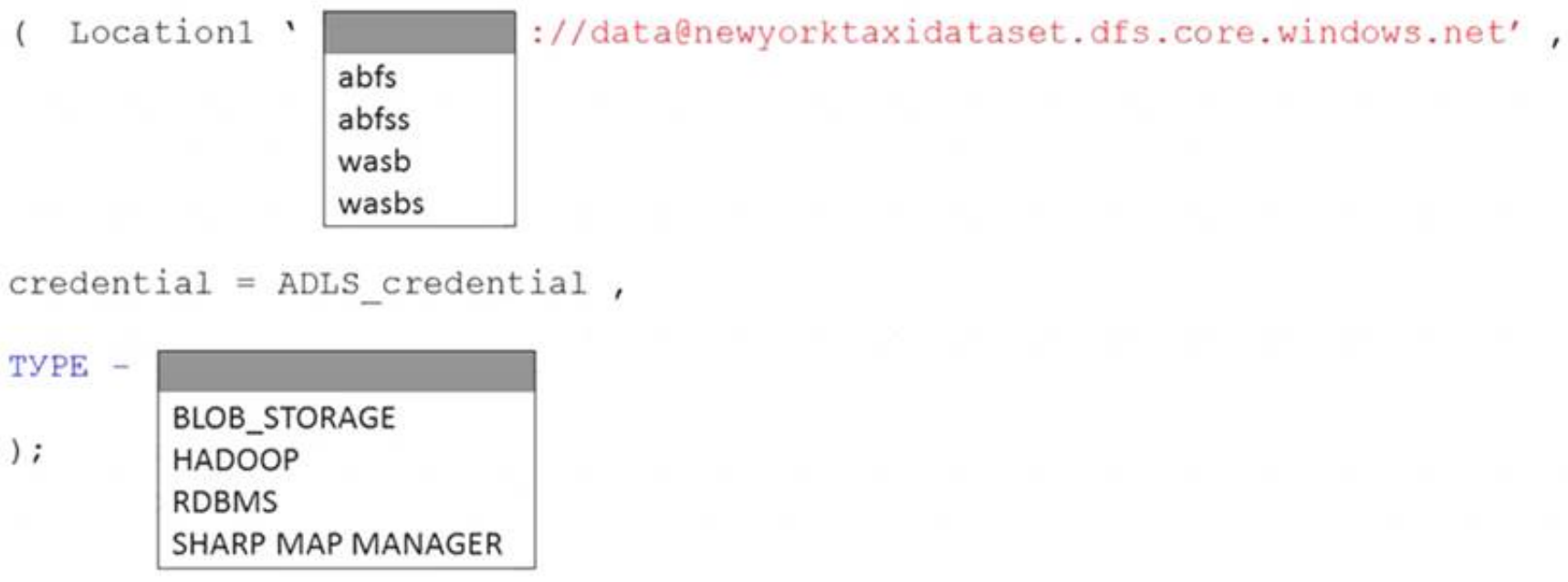
A. Yes
B. No

Answer: B

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 68
- (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



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 69
- (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 71
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

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

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

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

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

A. Yes
B. No

Answer: B

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

NEW QUESTION 72
- (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

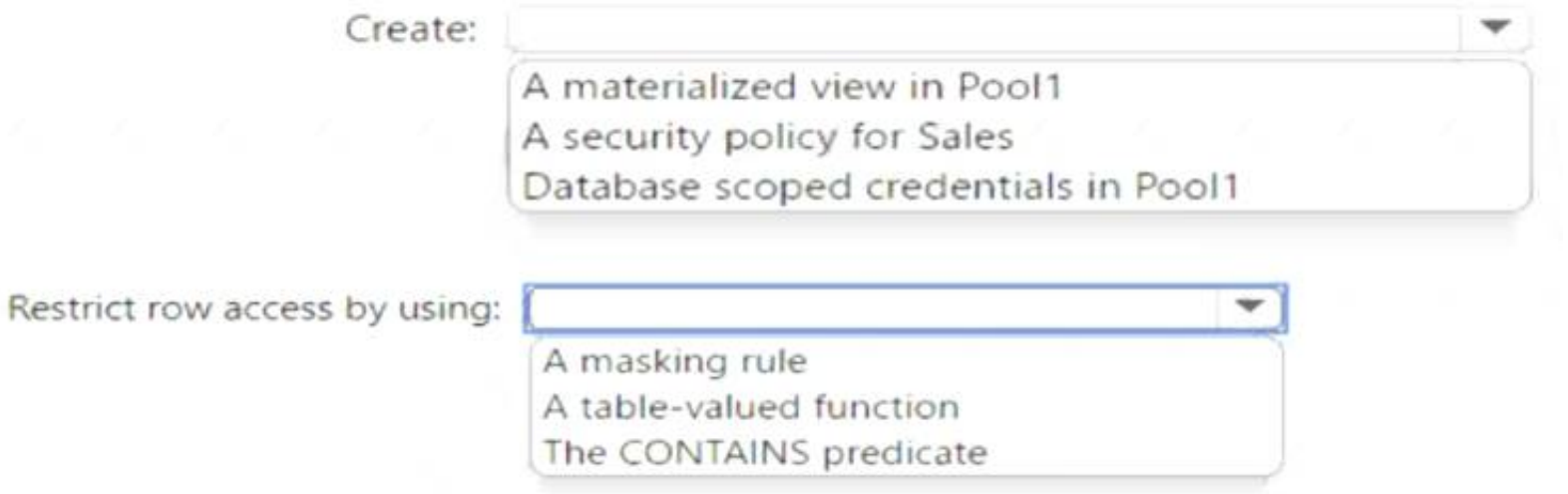
Explanation:
Need a High Concurrency cluster for the jobs.
Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.
A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that 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 75
- (Exam Topic 3)
You plan to use an Apache Spark pool in Azure Synapse Analytics to load data to an Azure Data Lake Storage Gen2 account.
You need to recommend which file format to use to store the data in the Data Lake Storage account. The solution must meet the following requirements:
• Column names and data types must be defined within the files loaded to the Data Lake Storage account.
• Data must be accessible by using queries from an Azure Synapse Analytics serverless SQL pool.
• Partition elimination must be supported without having to specify a specific partition. What should you recommend?

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

Answer: D

NEW QUESTION 79
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains an external table named Sales. Sales contains sales data. Each row in Sales
contains data on a single sale, including the name of the salesperson.
You need to implement row-level security (RLS). The solution must ensure that the salespeople can access only their respective sales.
What should you do? 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: A security policy for sale
Here are the steps to create a security policy for Sales:

- Create a user-defined function that returns the name of the current user:
- CREATE FUNCTION dbo.GetCurrentUser()
- RETURNS NVARCHAR(128)
- AS
- BEGIN
-

RETURN SUSER_SNAME();
END;
Create a security predicate function that filters the Sales table based on the current user:
CREATE FUNCTION dbo.SalesPredicate(@salesperson NVARCHAR(128))
RETURNS TABLE
WITH SCHEMABINDING
AS
RETURN SELECT 1 AS access_result
WHERE @salesperson = SalespersonName;
Create a security policy on the Sales table that uses the SalesPredicate function to filter the data:
CREATE SECURITY POLICY SalesFilter
ADD FILTER PREDICATE dbo.SalesPredicate(dbo.GetCurrentUser()) ON dbo.Sales
WITH (STATE = ON);

By creating a security policy for the Sales table, you ensure that each salesperson can only access their own sales data. The security policy uses a user-defined function to get the name of the current user and a security predicate function to filter the Sales table based on the current user.
Box 2: table-value function
to restrict row access by using row-level security, you need to create a table-valued function that returns a table of values that represent the rows that a user can access. You then use this function in a security
policy that applies a predicate on the table.

NEW QUESTION 84
- (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 85
- (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 86
- (Exam Topic 3)
You are developing a solution that will stream to Azure Stream Analytics. The solution will have both streaming data and reference data.
Which input type should you use for the reference data?

A. Azure Cosmos DB
B. Azure Blob storage
C. Azure IoT Hub
D. Azure Event Hubs

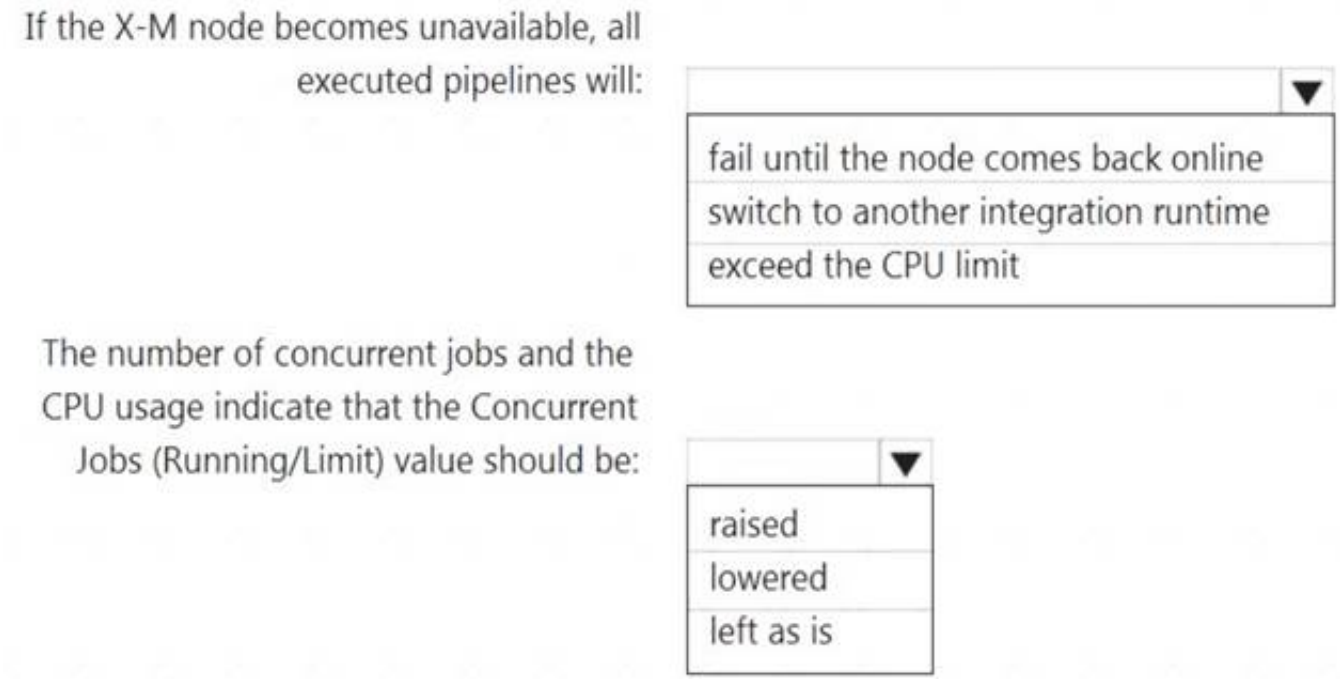
Answer: B

Explanation:
Stream Analytics supports Azure Blob storage and Azure SQL Database as the storage layer for Reference Data.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data

NEW QUESTION 87
- (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 90
- (Exam Topic 3)
You are designing an Azure Data Lake Storage Gen2 container to store data for the human resources (HR) department and the operations department at your company. You have the following data access requirements:
• After initial processing, the HR department data will be retained for seven years.
• The operations department data will be accessed frequently for the first six months, and then accessed once per month.
You need to design a data retention solution to meet the access requirements. The solution must minimize storage costs.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Answer Area

HR: Archive storage after one day and delete storage after 2.555 days.

Operations: Cool storage after 180 days.

NEW QUESTION 91
- (Exam Topic 3)
You plan to monitor an Azure data factory by using the Monitor & Manage app.
You need to identify the status and duration of activities that reference a table in a source database.
Which three actions should you perform in sequence? To answer, move the actions from the list of actions to the answer are and arrange them in the correct order.

Actions

From the Data Factory monitoring app, add the Source user property to the Activity Runs table.

From the Data Factory monitoring app, add the Source user property to the Pipeline Runs table.

From the Data Factory authoring UI, publish the pipelines.

From the Data Factory monitoring app, add a linked service to the Pipeline Runs table.

From the Data Factory authoring UI, generate a user property for Source on all activities.

From the Data Factory authoring UI, generate a user property for Source on all datasets.

Answer Area

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 95
- (Exam Topic 3)
You have an Azure Data Factory pipeline named Pipeline1!. Pipelinel contains a copy activity that sends data to an Azure Data Lake Storage Gen2 account.

Pipeline 1 is executed by a schedule trigger.
You change the copy activity sink to a new storage account and merge the changes into the collaboration branch.
After Pipelinel executes, you discover that data is NOT copied to the new storage account. You need to ensure that the data is copied to the new storage account.
What should you do?

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

Answer: A

Explanation:
CI/CD lifecycle

- A development data factory is created and configured with Azure Repos Git. All developers should have permission to author Data Factory resources like pipelines and datasets.
- A developer creates a feature branch to make a change. They debug their pipeline runs with their most recent changes
- After a developer is satisfied with their changes, they create a pull request from their feature branch to the main or collaboration branch to get their changes reviewed by peers.
- After a pull request is approved and changes are merged in the main branch, the changes get published to the development factory.

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

NEW QUESTION 96
- (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 101
- (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 104
- (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 106
- (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 111
- (Exam Topic 3)
You store files in an Azure Data Lake Storage Gen2 container. The container has the storage policy shown in the following exhibit.

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

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

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

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

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

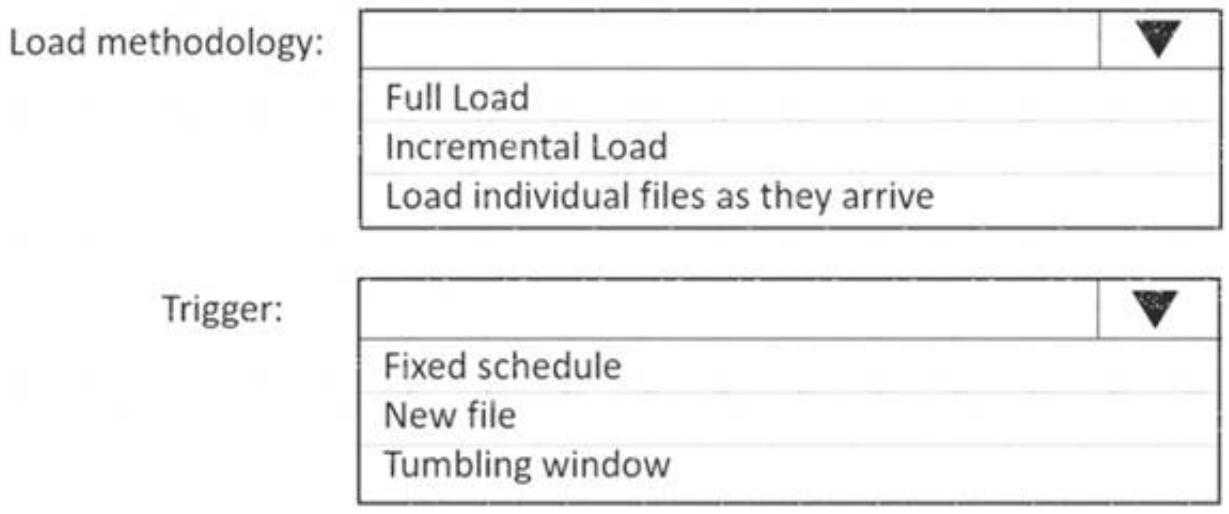
container1/contoso.csv
container1/docs/contoso.json
container1/mycontoso/contoso.csv

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated
Box 1: moved to cool storage
The ManagementPolicyBaseBlob.TierToCool property gets or sets the function to tier blobs to cool storage. Support blobs currently at Hot tier.
Box 2: container1/contoso.csv As defined by prefixMatch.
prefixMatch: An array of strings for prefixes to be matched. Each rule can define up to 10 case-senstive prefixes. A prefix string must start with a container name.
Reference:
https://docs.microsoft.com/en-us/dotnet/api/microsoft.azure.management.storage.fluent.models.managementpoli

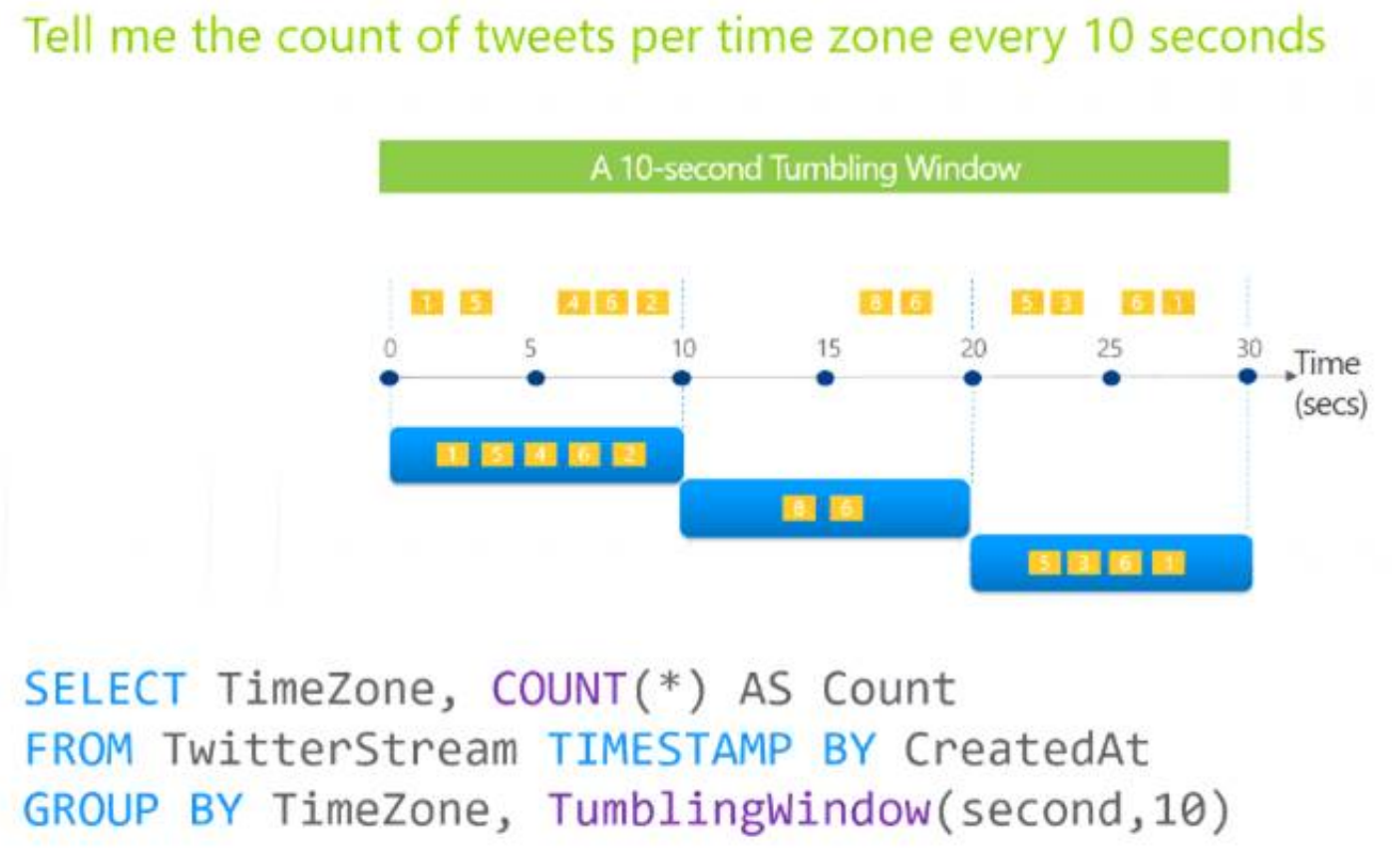
NEW QUESTION 114
- (Exam Topic 3)
You have an Azure Storage account that generates 200,000 new files daily. The file names have a format of {YYYY}/{MM}/{DD}/{HH}/{CustomerID}.csv.
You need to design an Azure Data Factory solution that will load new data from the storage account to an
Azure Data Lake 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

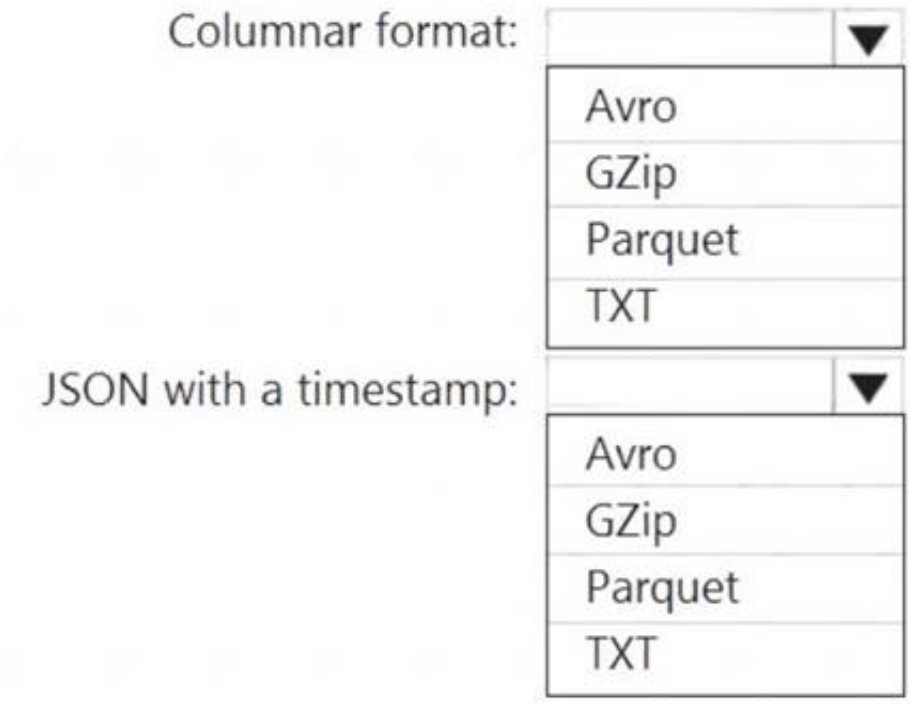
Answer: A

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 119
- (Exam Topic 3)
You need to output files from Azure Data Factory.
Which file format should you use for each type of output? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

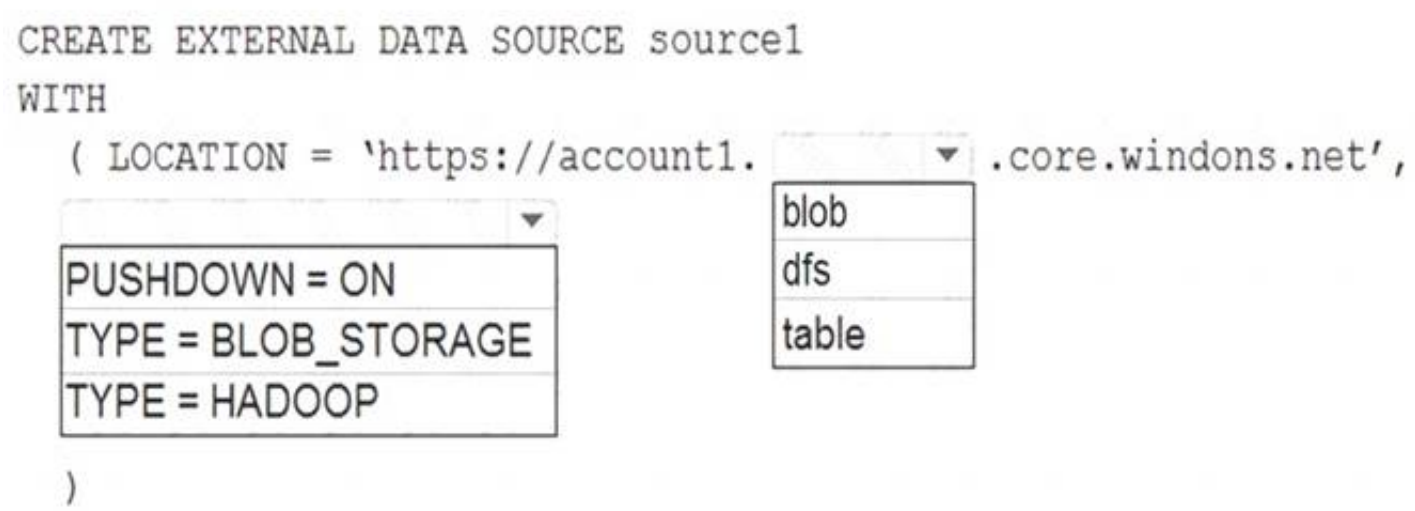
Answer: A

Explanation:
Box 1: Parquet
Parquet stores data in columns, while Avro stores data in a row-based format. By their very nature,
column-oriented data stores are optimized for read-heavy analytical workloads, while row-based databases are best for write-heavy transactional workloads.
Box 2: Avro
An Avro schema is created using JSON format. AVRO supports timestamps.
Note: Azure Data Factory supports the following file formats (not GZip or TXT).

- Avro format
- Binary format
- Delimited text format
- Excel format
- JSON format
- ORC format
- Parquet format
- XML format

Reference:
https://www.datanami.com/2018/05/16/big-data-file-formats-demystified

NEW QUESTION 121
- (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.



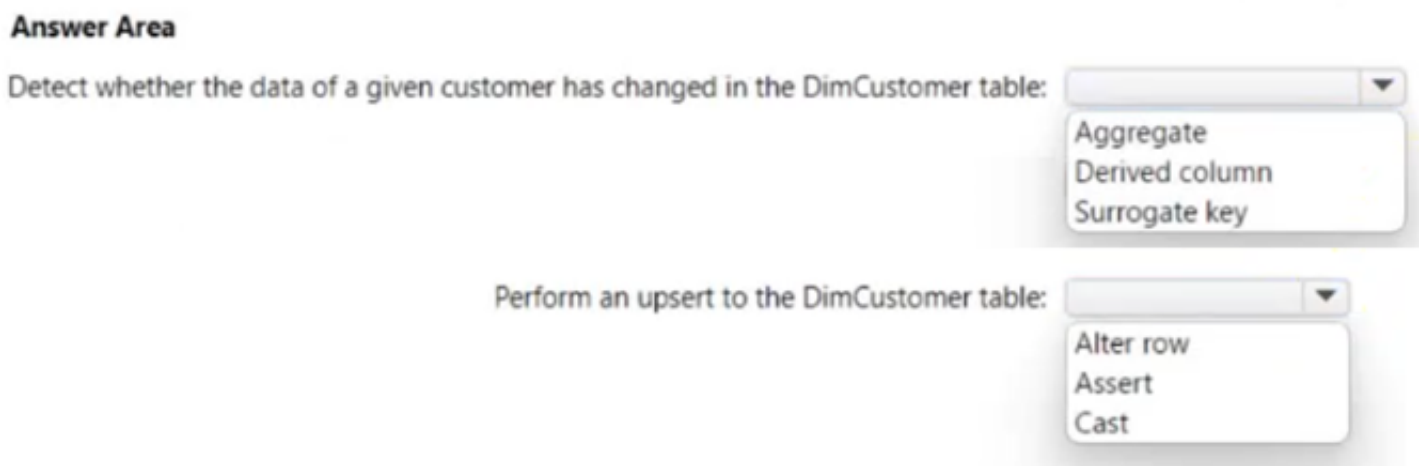
A. Mastered
B. Not Mastered

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

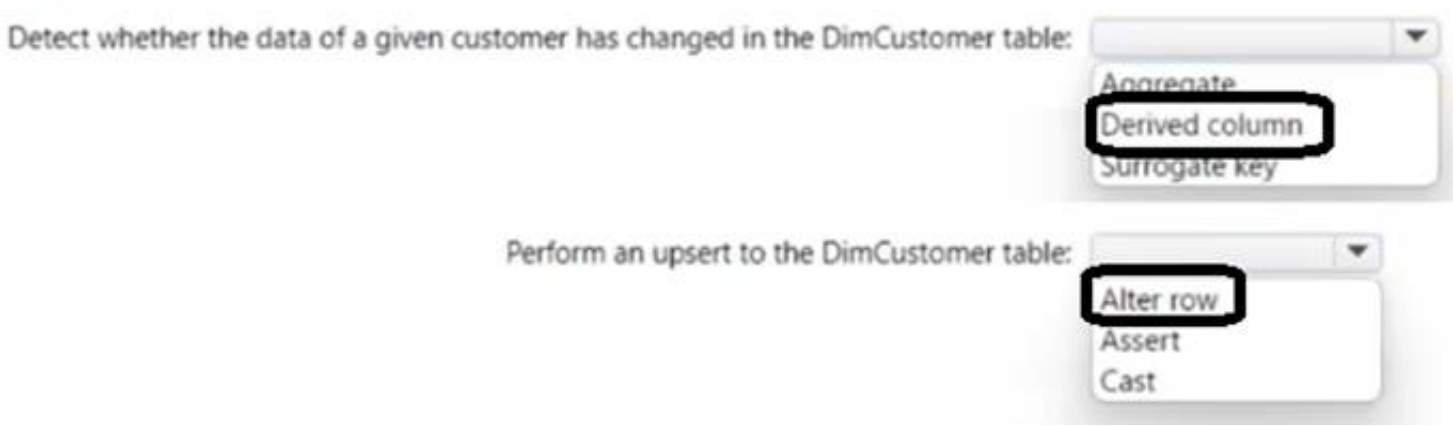
NEW QUESTION 122
- (Exam Topic 3)
You are developing an Azure Synapse Analytics pipeline that will include a mapping data flow named Dataflow1. Dataflow1 will read customer data from an external source and use a Type 1 slowly changing dimension (SCO) when loading the data into a table named DimCustomer1 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 123
- (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 124
- (Exam Topic 3)
You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytic dedicated SQL pool. The CSV file contains three columns named username, comment, and date.
The data flow already contains the following:
- A source transformation.
- A Derived Column transformation to set the appropriate types of data.
- A sink transformation to land the data in the pool.

You need to ensure that the data flow meets the following requirements:
- All valid rows must be written to the destination table.
- Truncation errors in the comment column must be avoided proactively.

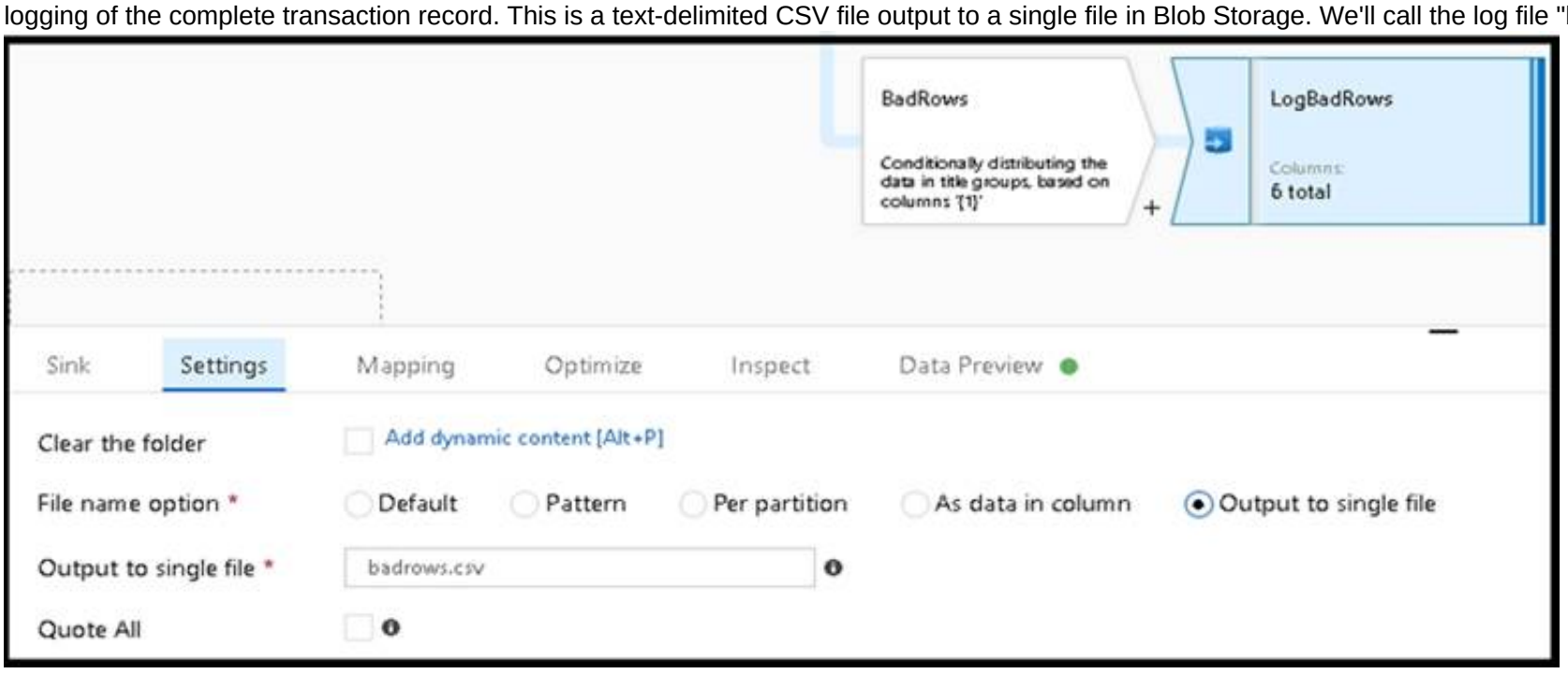
Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.
Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. To the data flow, add a sink transformation to write the rows to a file in blob storage.
B. To the data flow, add a Conditional Split transformation to separate the rows that will cause truncation errors.
C. To the data flow, add a filter transformation to filter out rows that will cause truncation errors.
D. Add a select transformation to select only the rows that will cause truncation errors.

Answer: AB

Explanation:
B: Example:
* 1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.

* 2. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the
BadRows stream. A:
* 3. Now we need to log the rows that failed. Add a sink transformation to the BadRows stream for logging. Here, we'll "auto-map" all of the fields so that we have logging of the complete transaction record. This is a text-delimited CSV file output to a single file in Blob Storage. We'll call the log file "badrows.csv".



* 4. The completed data flow is shown below. We are now able to split off error rows to avoid the SQL truncation errors and put those entries into a log file. Meanwhile, successful rows can continue to write to our target database.

Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-data-flow-error-rows

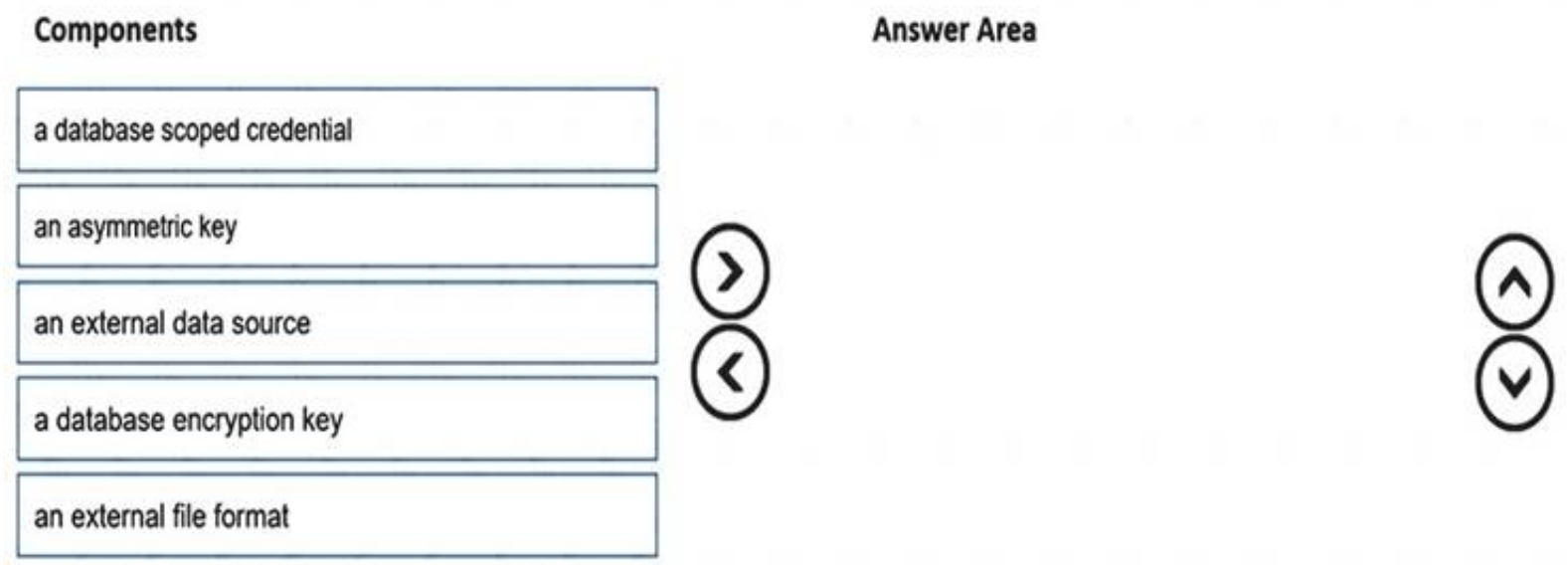
NEW QUESTION 129
- (Exam Topic 3)
You are planning a solution to aggregate streaming data that originates in Apache Kafka and is output to Azure Data Lake Storage Gen2. The developers who will implement the stream processing solution use Java, Which service should you recommend using to process the streaming data?

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

Answer: C

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

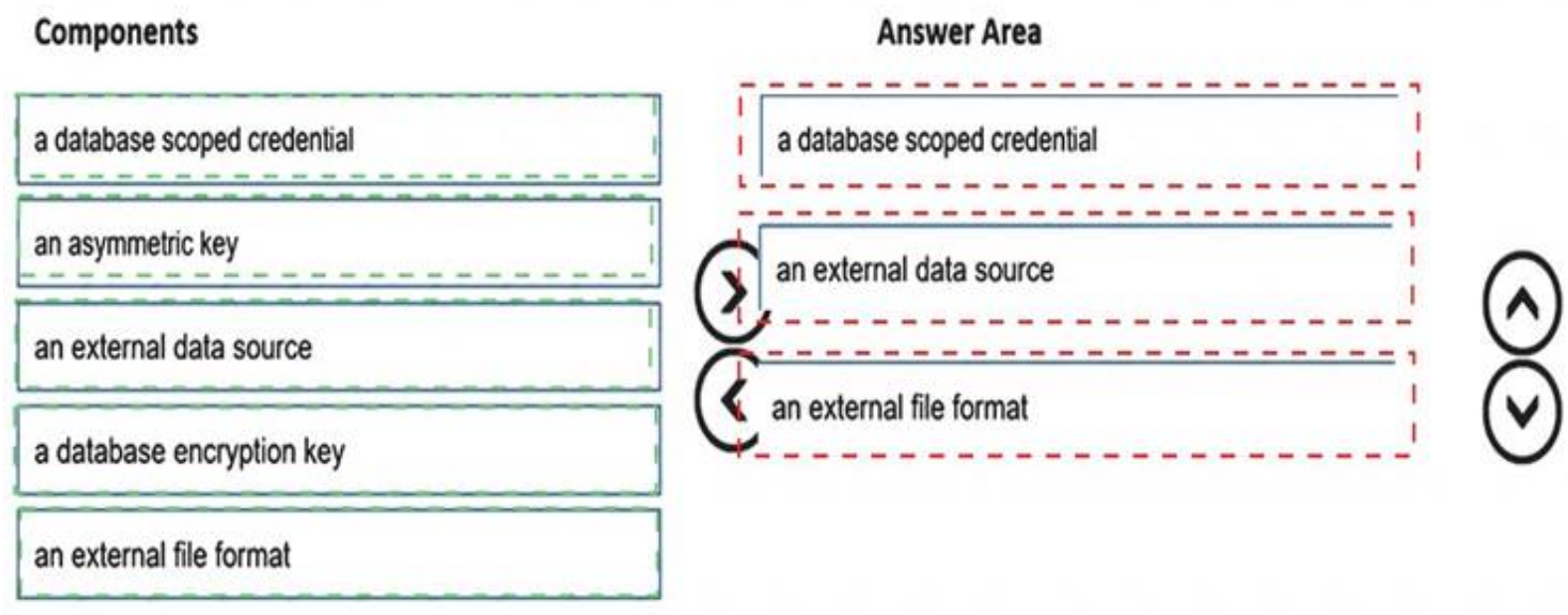
NEW QUESTION 133
- (Exam Topic 3)
You are responsible for providing access to an Azure Data Lake Storage Gen2 account.
Your user account has contributor access to the storage account, and you have the application ID and access key.
You plan to use PolyBase to load data into an enterprise data warehouse 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 135
- (Exam Topic 3)
You have an Azure Databricks workspace named workspace1 in the Standard pricing tier.
You need to configure workspace1 to support autoscaling all-purpose clusters. The solution must meet the following requirements:
- Automatically scale down workers when the cluster is underutilized for three minutes.
- Minimize the time it takes to scale to the maximum number of workers.
- Minimize costs. What should you do first?

A. Enable container services for workspace1.
B. Upgrade workspace1 to the Premium pricing tier.
C. Set Cluster Mode to High Concurrency.
D. Create a cluster policy in workspace1.

Answer: B

Explanation:
For clusters running Databricks Runtime 6.4 and above, optimized autoscaling is used by all-purpose clusters in the Premium plan
Optimized autoscaling:
Scales up from min to max in 2 steps.
Can scale down even if the cluster is not idle by looking at shuffle file state. Scales down based on a percentage of current nodes.
On job clusters, scales down if the cluster is underutilized over the last 40 seconds.
On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds.
The spark.databricks.aggressiveWindowDownS Spark configuration property specifies in seconds how often a cluster makes down-scaling decisions. Increasing the value causes a cluster to scale down more slowly. The maximum value is 600.
Note: Standard autoscaling
Starts with adding 8 nodes. Thereafter, scales up exponentially, but can take many steps to reach the max. You can customize the first step by setting the spark.databricks.autoscaling.standardFirstStepUp Spark configuration property.
Scales down only when the cluster is completely idle and it has been underutilized for the last 10 minutes. Scales down exponentially, starting with 1 node.
Reference: https://docs.databricks.com/clusters/configure.html

NEW QUESTION 138
- (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 139
- (Exam Topic 3)
You are designing an Azure Synapse Analytics dedicated SQL pool.
Groups will have access to sensitive data in the pool as shown in the following table.

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

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

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

You have a table of patients for each region. The tables contain the following potentially sensitive columns.

Name	Sensitive data	Description
CardOnFile	Financial	Debit/credit card number for charges
Height	Medical	Patient's height in cm
ContactEmail	PII	Email address for secure communications

You are designing dynamic data masking to maintain compliance.
For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.

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

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 141
- (Exam Topic 3)
You are designing an Azure Data Lake Storage Gen2 structure for telemetry data from 25 million devices distributed across seven key geographical regions. Each minute, the devices will send a JSON payload of metrics to Azure Event Hubs.
You need to recommend a folder structure for the data. The solution must meet the following requirements:

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

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

Values

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

Answer Area

/ Value / Value / Value .json

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 142
- (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 143
- (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

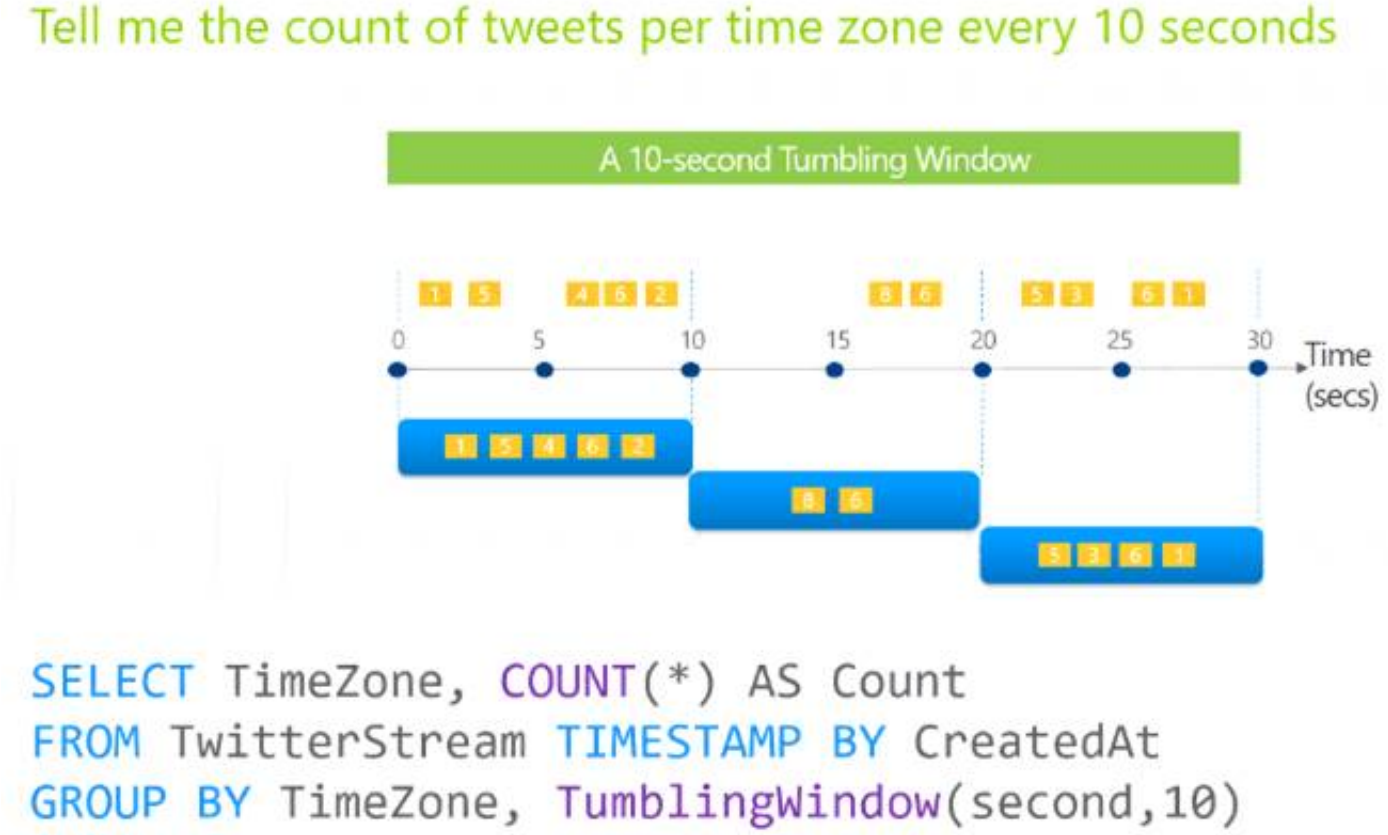
Explanation:
You can use Dynamic Management Views (DMVs) to monitor your workload including investigating query execution in SQL pool.
If your queries are failing or taking a long time to proceed, you can check and monitor if you have any transactions rolling back.
Example:
-- Monitor rollback SELECT
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 144
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an Azure Stream Analytics solution that will analyze Twitter data.
You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.
Solution: You use a tumbling window, and you set the window size to 10 seconds. Does this meet the goal?

A. Yes
B. No

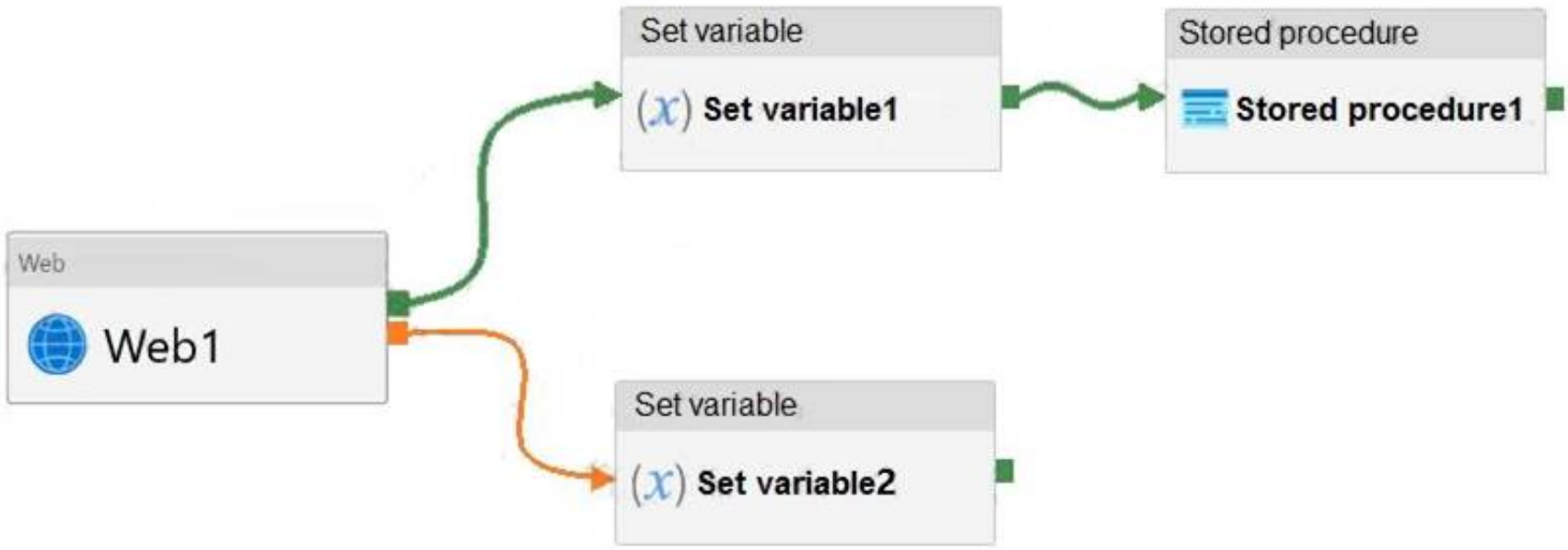
Answer: A

Explanation:
Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.

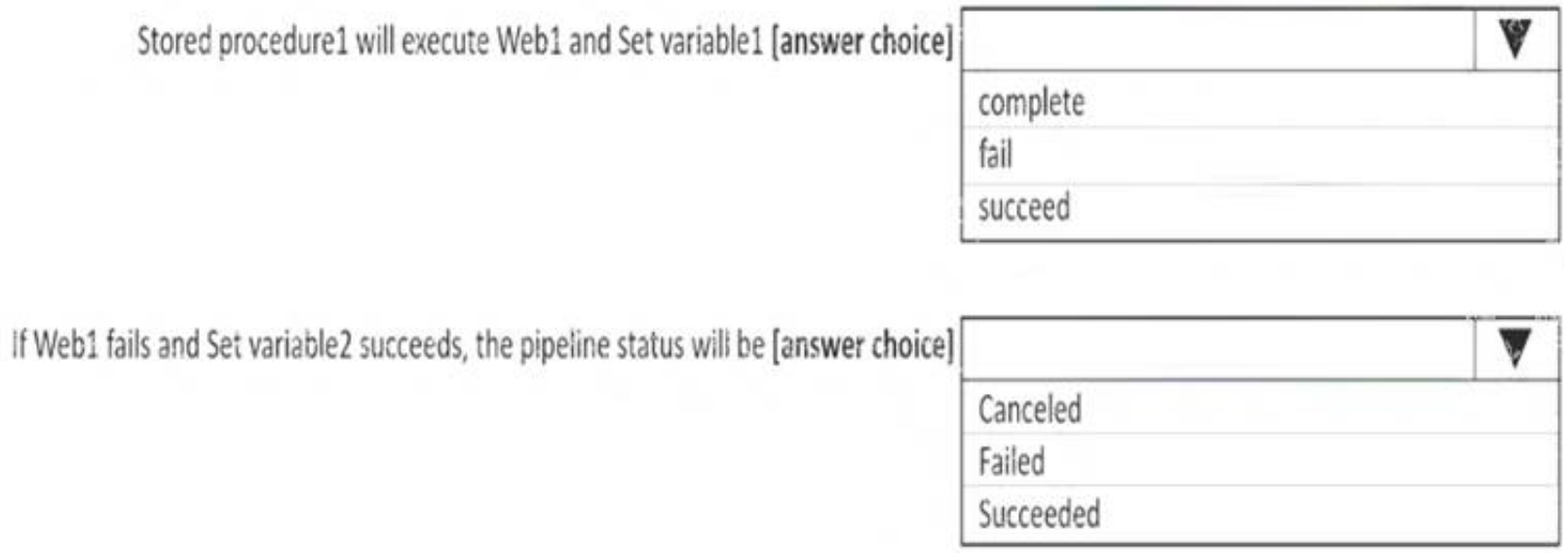


Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

NEW QUESTION 149
- (Exam Topic 3)
You have an Azure Data Factory pipeline that has the activities shown in the following exhibit.



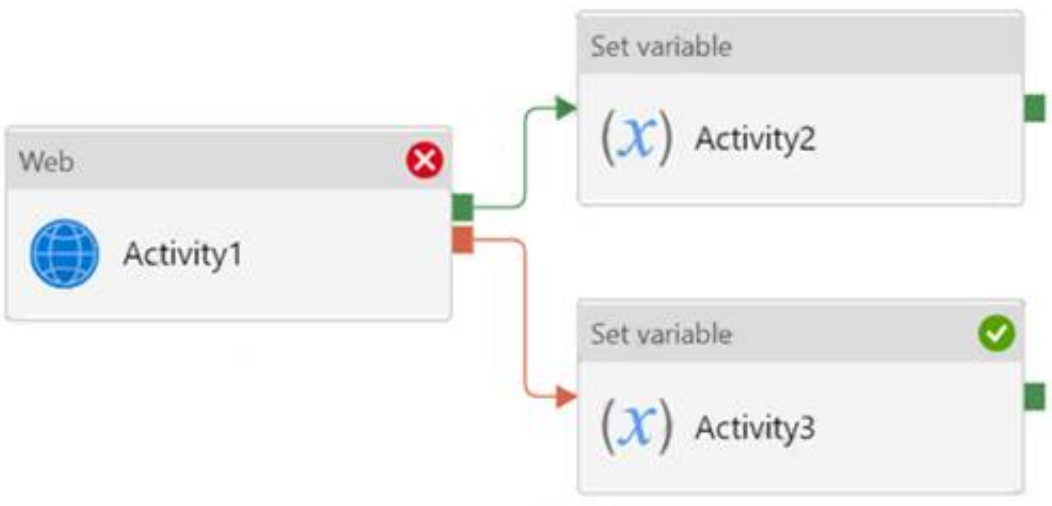
Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

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



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

NEW QUESTION 151
- (Exam Topic 3)
You have the following Azure Stream Analytics query.

```
WITH

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

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

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

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Yes
You can now use a new extension of Azure Stream Analytics SQL to specify the number of partitions of a stream when reshuffling the data.
The outcome is a stream that has the same partition scheme. Please see below for an example: WITH step1 AS (SELECT * FROM [input1] PARTITION BY DeviceID INTO 10),
step2 AS (SELECT * FROM [input2] PARTITION BY DeviceID INTO 10)
SELECT * INTO [output] FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceID Note: The new extension of Azure Stream Analytics SQL includes a keyword INTO that allows you to specify the number of partitions for a stream when performing reshuffling using a PARTITION BY statement.
Box 2: Yes
When joining two streams of data explicitly repartitioned, these streams must have the same partition key and partition count.
Box 3: Yes
10 partitions x six SUs = 60 SUs is fine.
Note: Remember, Streaming Unit (SU) count, which is the unit of scale for Azure Stream Analytics, must be adjusted so the number of physical resources available to the job can fit the partitioned flow. In general, six SUs is a good number to assign to each partition. In case there are insufficient resources assigned to the job, the system will only apply the repartition if it benefits the job.
Reference:
https://azure.microsoft.com/en-in/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/

NEW QUESTION 155
- (Exam Topic 3)
You are designing an Azure Databricks table. The table will ingest an average of 20 million streaming events per day.
You need to persist the events in the table for use in incremental load pipeline jobs in Azure Databricks. The solution must minimize storage costs and incremental load times.
What should you include in the solution?

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

Answer: A

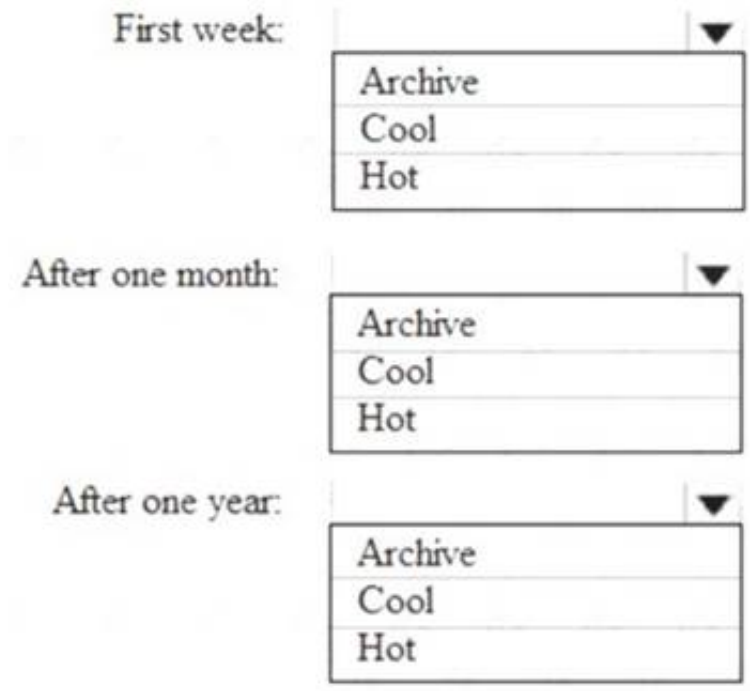
Explanation:
The Databricks ABS-AQS connector uses Azure Queue Storage (AQS) to provide an optimized file source that lets you find new files written to an Azure Blob

storage (ABS) container without repeatedly listing all of the files.
This provides two major advantages:

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

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

NEW QUESTION 159
- (Exam Topic 3)
You are designing an application that will store petabytes of medical imaging data
When the data is first created, the data will be accessed frequently during the first week. After one month, the data must be accessible within 30 seconds, but files will be accessed infrequently. After one year, the data will be accessed infrequently but must be accessible within five minutes.
You need to select a storage strategy for the data. The solution must minimize costs.
Which storage tier should you use for each time frame? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 162
- (Exam Topic 3)
You are designing an inventory updates table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:

Table	Comment
EventDate	One million records are added to the table each day
EventTypeID	The table contains 10 million records for each event type.
WarehouseID	The table contains 100 million records for each warehouse.
ProductCategoryTypeID	The table contains 25 million records for each product category type.

You identify the following usage patterns:

- Analysts will most commonly analyze transactions for a warehouse.
- Queries will summarize by product category type, date, and/or inventory event type. You need to recommend a partition strategy for the table to minimize query times.

On which column should you partition the table?

A. ProductCategoryTypeID
B. EventDate
C. WarehouseID
D. EventTypeID

Answer: C

Explanation:
The number of records for each warehouse is big enough for a good partitioning.
Note: Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.
When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

NEW QUESTION 164
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool.
You need to ensure that data in the pool is encrypted at rest. The solution must NOT require modifying applications that query the data.
What should you do?

A. Enable encryption at rest for the Azure Data Lake Storage Gen2 account.
B. Enable Transparent Data Encryption (TDE) for the pool.
C. Use a customer-managed key to enable double encryption for the Azure Synapse workspace.
D. Create an Azure key vault in the Azure subscription grant access to the pool.

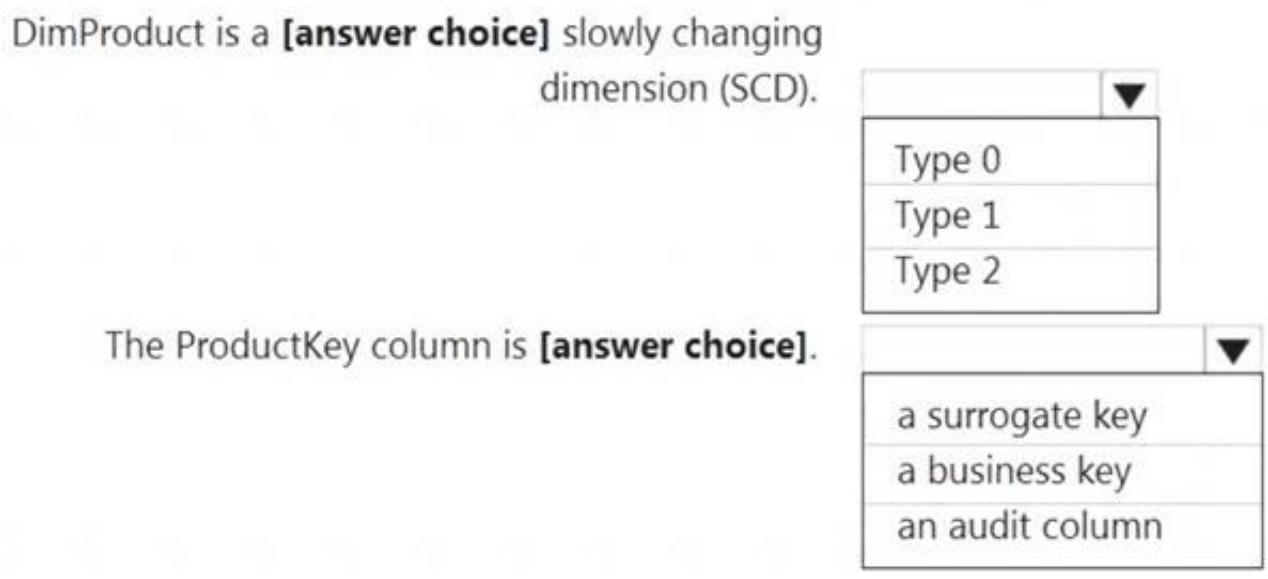
Answer: B

Explanation:
Transparent Data Encryption (TDE) helps protect against the threat of malicious activity by encrypting and decrypting your data at rest. When you encrypt your database, associated backups and transaction log files are encrypted without requiring any changes to your applications. TDE encrypts the storage of an entire database by using a symmetric key called the database encryption key.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-overviewmana

NEW QUESTION 169
- (Exam Topic 3)
You are creating dimensions for a data warehouse in an Azure Synapse Analytics dedicated SQL pool. You create a table by using the Transact-SQL statement shown in the following exhibit.

```
CREATE TABLE [DBO].[DimProduct](
      [ProductKey] [int] IDENTITY(1,1) NOT NULL,
      [ProductSourceID] [int] NOT NULL,
      [ProductName] [nvarchar](100) NOT NULL,
      [ProductNumber] [nvarchar](25) NOT NULL,
      [Color] [nvarchar](15) NULL,
      [Size] [nvarchar](5) NULL,
      [Weight] [decimal](8, 2) NULL,
      [ProductCategory] [nvarchar](100) NULL,
      [SellStartDate] [date] NOT NULL,
      [SellEndDate] [date] NULL,
      [RowInsertedDateTime] [datetime] NOT NULL,
      [RowUpdatedDateTime] [datetime] NOT NULL,
      [ETLAuditID] [int] NOT NULL
)
```

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



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Type 2
A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members.
Reference:
https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics

NEW QUESTION 174
- (Exam Topic 3)
You have two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.

Table	Column
Flight	ArrivalAirportID ArrivalDateTime
Weather	AirportID ReportDateTime

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

A. In the tables use a hash distribution of ArrivalDateTime and ReportDateTime.
B. In the tables use a hash distribution of ArrivalAirportID and AirportlD.
C. In each table, create an identity column.
D. In each table, create a column as a composite of the other two columns in the table.

Answer: B

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

NEW QUESTION 176
- (Exam Topic 3)
You have an Azure Synapse Analytics job that uses Scala. You need to view the status of the job.
What should you do?

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

Answer: C

Explanation:
Use Synapse Studio to monitor your Apache Spark applications. To monitor running Apache Spark application Open Monitor, then select Apache Spark applications. To view the details about the Apache Spark applications that are running, select the submitting Apache Spark application and view the details. If the Apache Spark application is still running, you can monitor the progress.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/monitoring/apache-spark-applications

NEW QUESTION 177
- (Exam Topic 3)
You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers. The file contains two attributes named FirstName and LastName.
You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the FirstName and LastName values.
You create the following components:
- A destination table in Azure Synapse
- An Azure Blob storage container
- A service principal

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

Actions

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

Answer Area

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 181
- (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 Data Lake Storage account that contains a staging zone.
You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You schedule an Azure Databricks job that executes an R notebook, and then inserts the data into the data warehouse.
Does this meet the goal?

A. Yes
B. No

Answer: B

Explanation:
Must use an Azure Data Factory, not an Azure Databricks job. Reference:
https://docs.microsoft.com/en-US/azure/data-factory/transform-data

NEW QUESTION 182
- (Exam Topic 3)
You configure monitoring for a Microsoft Azure SQL Data Warehouse implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Gen 2 using an external table.
Files with an invalid schema cause errors to occur. You need to monitor for an invalid schema error. For which error should you monitor?

A. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error[com.microsoft.polybase.client.KerberosSecureLogin] occurred while accessing external files.'
B. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [No FileSystem for scheme: wasbs] occurred while accessing external file.'
C. Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11": for linked server "(null)", Query aborted- the maximum reject threshold (orows) was reached while regarding from an external source: 1 rows rejected out of total 1 rows processed.
D. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [Unable to instantiate LoginClass] occurredwhile accessing external files.'

Answer: C

Explanation:
Customer Scenario:
SQL Server 2016 or SQL DW connected to Azure blob storage. The CREATE EXTERNAL TABLE DDL points to a directory (and not a specific file) and the directory contains files with different schemas.
SSMS Error:
Select query on the external table gives the following error: Msg 7320, Level 16, State 110, Line 14
Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11" for linked server "(null)". Query aborted-- the maximum reject threshold (0 rows) was reached while reading from an external source: 1 rows rejected out of total 1 rows processed.
Possible Reason:
The reason this error happens is because each file has different schema. The PolyBase external table DDL when pointed to a directory recursively reads all the files in that directory. When a column or data type mismatch happens, this error could be seen in SSMS.
Possible Solution:
If the data for each table consists of one file, then use the filename in the LOCATION section prepended by the directory of the external files. If there are multiple files per table, put each set of files into different directories in Azure Blob Storage and then you can point LOCATION to the directory instead of a particular file. The latter suggestion is the best practices recommended by SQLCAT even if you have one file per table.

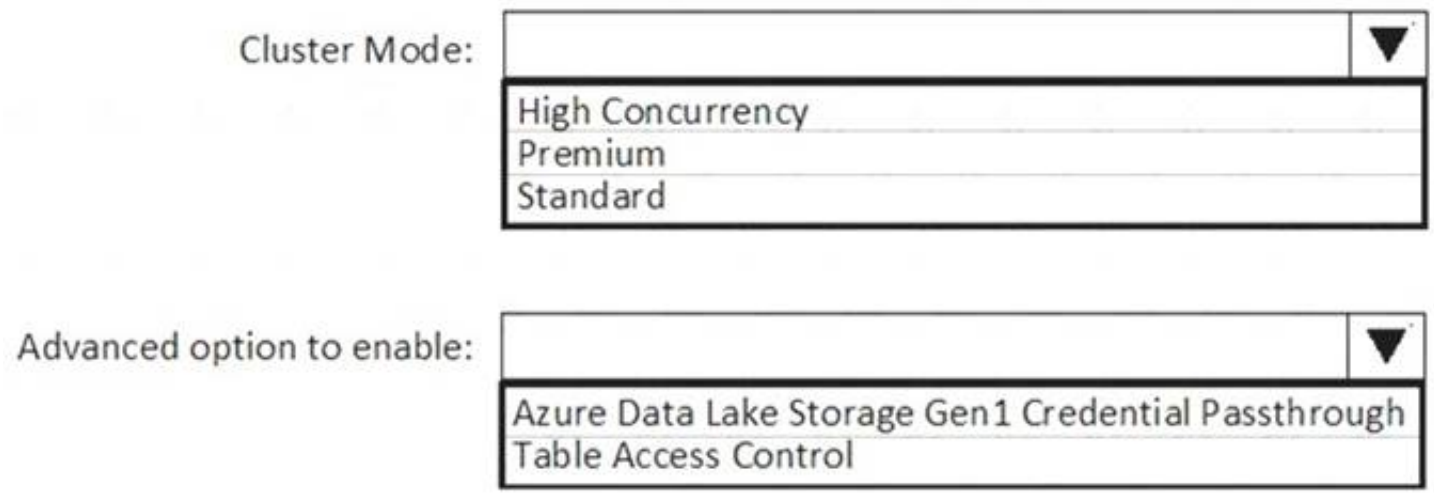
NEW QUESTION 185
- (Exam Topic 3)
You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1.
You need to use an Azure Synapse pipeline to copy data from the selected tables in DB1 to the files in storage1. The solution must meet the following requirements:
• The Copy activity in the pipeline must be parameterized to use the data in File1.txt to identify the source and destination of the copy.
• Copy activities must occur in parallel as often as possible.
Which two pipeline activities should you include in the pipeline? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. If Condition
B. ForEach
C. Lookup
D. Get Metadata

Answer: BC

Explanation:
Lookup: This is a control activity that retrieves a dataset from any of the supported data sources and makes it available for use by subsequent activities in the pipeline. You can use a Lookup activity to read File1.txt from storage1 and store its content as an array variable1.
ForEach: This is a control activity that iterates over a collection and executes specified activities in a loop. You can use a ForEach activity to loop over the array variable from the Lookup activity and pass each table name as a parameter to a Copy activity that copies data from DB1 to storage11.

NEW QUESTION 188
- (Exam Topic 3)
You need to implement an Azure Databricks cluster that automatically connects to Azure Data Lake Storage Gen2 by using Azure Active Directory (Azure AD) integration.
How should you configure the new cluster? 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: High Concurrency
Enable Azure Data Lake Storage credential passthrough for a high-concurrency cluster. Incorrect:
Support for Azure Data Lake Storage credential passthrough on standard clusters is in Public Preview.
Standard clusters with credential passthrough are supported on Databricks Runtime 5.5 and above and are limited to a single user.
Box 2: Azure Data Lake Storage Gen1 Credential Passthrough
You can authenticate automatically to Azure Data Lake Storage Gen1 and Azure Data Lake Storage 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.
References:
https://docs.azuredatabricks.net/spark/latest/data-sources/azure/adls-passthrough.html

NEW QUESTION 190
- (Exam Topic 3)
You plan to develop a dataset named Purchases by using Azure databricks Purchases will contain the following columns:
• ProductID
• ItemPrice
• lineTotal
• Quantity
• StoreID
• Minute
• Month
• Hour
• Year
• Day
You need to store the data to support hourly incremental load pipelines that will vary for each StoreID. the solution must minimize storage costs. How should you complete the rode? To answer, select the appropriate options In the answer area.
NOTE: Each correct selection is worth one point.

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 193
- (Exam Topic 3)
You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1. You need to verify whether the size of the transaction log file for each distribution of DW1 is smaller than 160 GB.
What should you do?

A. On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.
B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.
C. On DW1, execute a query against the sys.database_files dynamic management view.
D. Execute a query against the logs of DW1 by using theGet-AzOperationalInsightSearchResult PowerShell cmdlet.

Answer: A

Explanation:
The following query returns the transaction log size on each distribution. If one of the log files is reaching 160 GB, you should consider scaling up your instance or limiting your transaction size.
-- Transaction log size SELECT
instance_name as distribution_db, cntr_value*1.0/1048576 as log_file_size_used_GB, pdw_node_id
FROM sys.dm_pdw_nodes_os_performance_counters WHERE
instance_name like 'Distribution_%'
AND counter_name = 'Log File(s) Used Size (KB)'
References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-manage-monitor

NEW QUESTION 196
- (Exam Topic 3)
Note: The question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it As a result these questions will not appear in the review screen. You have an Azure Data Lake Storage account that contains a staging zone.
You need to design a dairy process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes a mapping data low. and then inserts the data into the data warehouse.
Does this meet the goal?

A. Yes
B. No

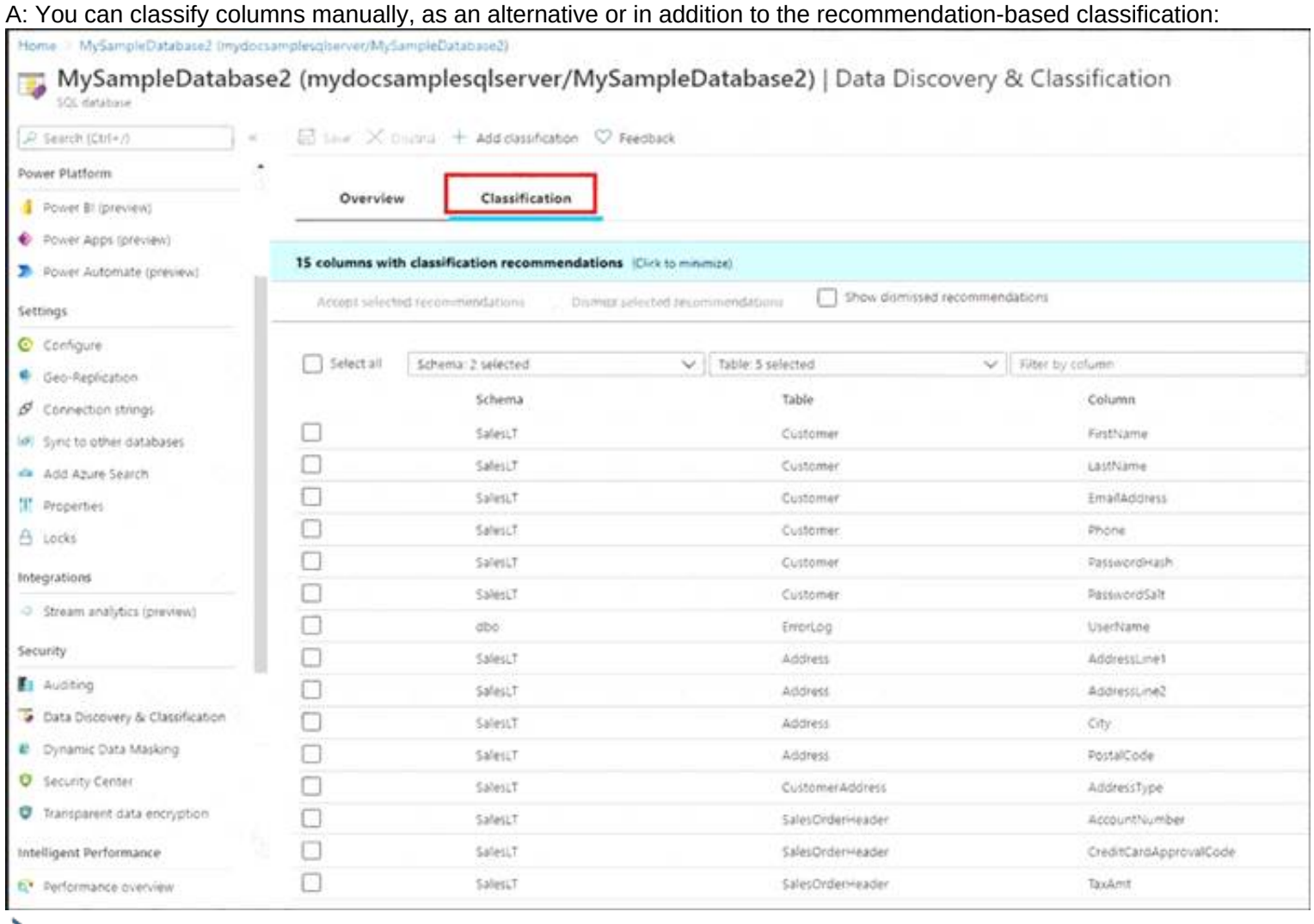
Answer: B

NEW QUESTION 201
- (Exam Topic 3)
You plan to create an Azure Synapse Analytics dedicated SQL pool.
You need to minimize the time it takes to identify queries that return confidential information as defined by the company's data privacy regulations and the users who executed the queues.
Which two components should you include in the solution? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. sensitivity-classification labels applied to columns that contain confidential information
B. resource tags for databases that contain confidential information
C. audit logs sent to a Log Analytics workspace
D. dynamic data masking for columns that contain confidential information

Answer: AC

Explanation:
A: You can classify columns manually, as an alternative or in addition to the recommendation-based classification:



- Select Add classification in the top menu of the pane.
- In the context window that opens, select the schema, table, and column that you want to classify, and the information type and sensitivity label.
- Select Add classification at the bottom of the context window.

C: An important aspect of the information-protection paradigm is the ability to monitor access to sensitive data. Azure SQL Auditing has been enhanced to include a new field in the audit log called data_sensitivity_information. This field logs the sensitivity classifications (labels) of the data that was returned by a query. Here's an example:

d	client_ip	application_name	duration_milliseconds	response_rows	affected_rows	connection_id	data_sensitivity_information
	[redacted]7.125	Microsoft SQL Server Management Studio - Query	1	847	847	C244A066-2271-...	Confidential - GDPR
	[redacted]7.125	Microsoft SQL Server Management Studio - Query	2	32	32	C244A066-2271-...	Confidential
	[redacted]7.125	Microsoft SQL Server Management Studio - Query	41	32	32	A7088FD4-759E-...	Confidential, Confidential - GDPR

Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview

NEW QUESTION 204
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool.
You need to Create a fact table named Table1 that will store sales data from the last three years. The solution must be optimized for the following query operations:
Show order counts by week.
• Calculate sales totals by region.
• Calculate sales totals by product.
• Find all the orders from a given month. Which data should you use to partition Table1?

A. region
B. product
C. week
D. month

Answer: D

Explanation:
Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.
Benefits to queries
Partitioning can also be used to improve query performance. A query that applies a filter to partitioned data can limit the scan to only the qualifying partitions. This method of filtering can avoid a full table scan and only scan a smaller subset of data. With the introduction of clustered columnstore indexes, the predicate elimination performance benefits are less beneficial, but in some cases there can be a benefit to queries.
For example, if the sales fact table is partitioned into 36 months using the sales date field, then queries that filter on the sale date can skip searching in partitions that don't match the filter.

Note: Benefits to loads
The primary benefit of partitioning in dedicated SQL pool is to improve the efficiency and performance of loading data by use of partition deletion, switching and merging. In most cases data is partitioned on a date column that is closely tied to the order in which the data is loaded into the SQL pool. One of the greatest benefits of using partitions to maintain data is the avoidance of transaction logging. While simply inserting, updating, or deleting data can be the most straightforward approach, with a little thought and effort, using partitioning during your load process can substantially improve performance.
Reference:
https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partitio

NEW QUESTION 208
.....

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