Exam Questions DP-203

Data Engineering on Microsoft Azure

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

NEW QUESTION 1
- (Exam Topic 3)
A company plans to use Apache Spark analytics to analyze intrusion detection data.
You need to recommend a solution to analyze network and system activity data for malicious activities and policy violations. The solution must minimize administrative efforts.
What should you recommend?

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

Answer: B

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

NEW QUESTION 2
- (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 3
- (Exam Topic 3)
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 copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script.
Does this meet the goal?

A. Yes
B. No

Answer: A

Explanation:
If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity with your own data processing logic and use the activity in the pipeline.
Note: You can use data transformation activities in Azure Data Factory and Synapse pipelines to transform and process your raw data into predictions and insights at scale.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/transform-data

NEW QUESTION 4
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool mat contains a table named dbo.Users.
You need to prevent a group of users from reading user email addresses from dbo.Users. What should you use?

A. row-level security
B. column-level security
C. Dynamic data masking
D. Transparent Data Encryption (TDD

Answer: B

NEW QUESTION 5
- (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

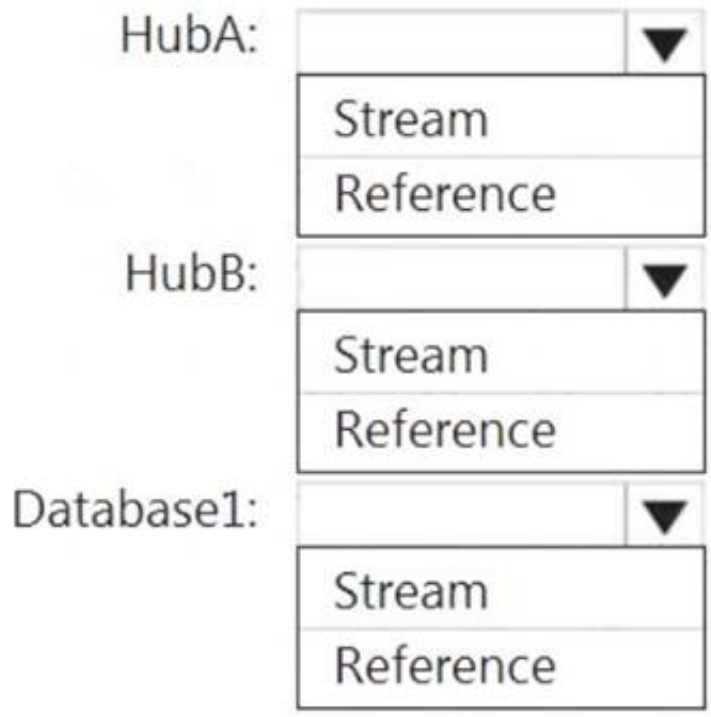
Answer: C

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 6
- (Exam Topic 3)
You have an Azure SQL database named Database1 and two Azure event hubs named HubA and HubB. The data consumed from each source is shown in the following table.

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

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



A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 7
- (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 8
- (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 9
- (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 10
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named pool1.
You plan to implement a star schema in pool1 and create a new table named DimCustomer by using the following code.

```
CREATE TABLE dbo.[DimCustomer](
    [CustomerKey] int NOT NULL,
    [CustomerSourceID] [int] NOT NULL,
    [Title] [nvarchar](8) NULL,
    [FirstName] [nvarchar](50) NOT NULL,
    [MiddleName] [nvarchar](50) NULL,
    [LastName] [nvarchar](50) NOT NULL,
    [Suffix] [nvarchar](10) NULL,
    [CompanyName] [nvarchar](128) NULL,
    [SalesPerson] [nvarchar](256) NULL,
    [EmailAddress] [nvarchar](50) NULL,
    [Phone] [nvarchar](25) NULL,
    [InsertedDate] [datetime] NOT NULL,
    [ModifiedDate] [datetime] NOT NULL,
    [HashKey] [varchar](100) NOT NULL,
    [IsCurrentRow] [bit] NOT NULL
)
WITH
(
    DISTRIBUTION = REPLICATE,
    CLUSTERED COLUMNSTORE INDEX
);
GO
```

You need to ensure that DimCustomer has the necessary columns to support a Type 2 slowly changing dimension (SCD). Which two columns should you add? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. [HistoricalSalesPerson] [nvarchar] (256) NOT NULL
B. [EffectiveEndDate] [datetime] NOT NULL
C. [PreviousModifiedDate] [datetime] NOT NULL
D. [RowID] [bigint] NOT NULL
E. [EffectiveStartDate] [datetime] NOT NULL

Answer: AB

NEW QUESTION 10
- (Exam Topic 3)
You are designing an Azure Stream Analytics solution that receives instant messaging data from an Azure Event Hub.
You need to ensure that the output from the Stream Analytics job counts the number of messages per time zone every 15 seconds.
How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

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

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

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

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 11
- (Exam Topic 3)
The following code segment is used to create an Azure Databricks cluster.

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

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

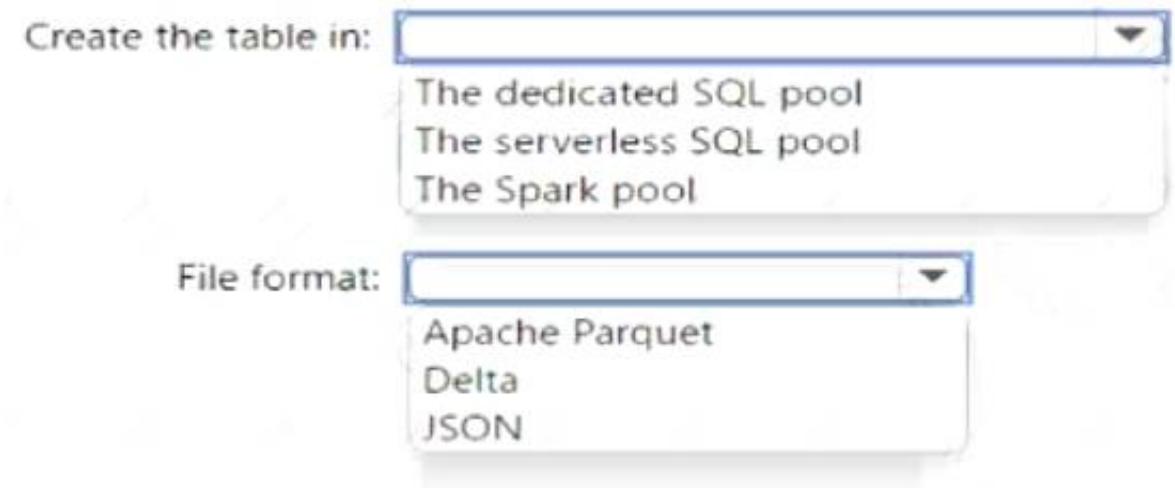
Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	○	○
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	○	○
The Databricks cluster supports the creation of a Delta Lake table.	○	○

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 14
- (Exam Topic 3)
You have an Azure Synapse Analytics serverless SQL pool, an Azure Synapse Analytics dedicated SQL pool, an Apache Spark pool, and an Azure Data Lake Storage Gen2 account.
You need to create a table in a lake database. The table must be available to both the serverless SQL pool and the Spark pool.
Where should you create the table, and Which file format should you use for data in the table? 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:
The dedicated SQL pool Apache Parquet

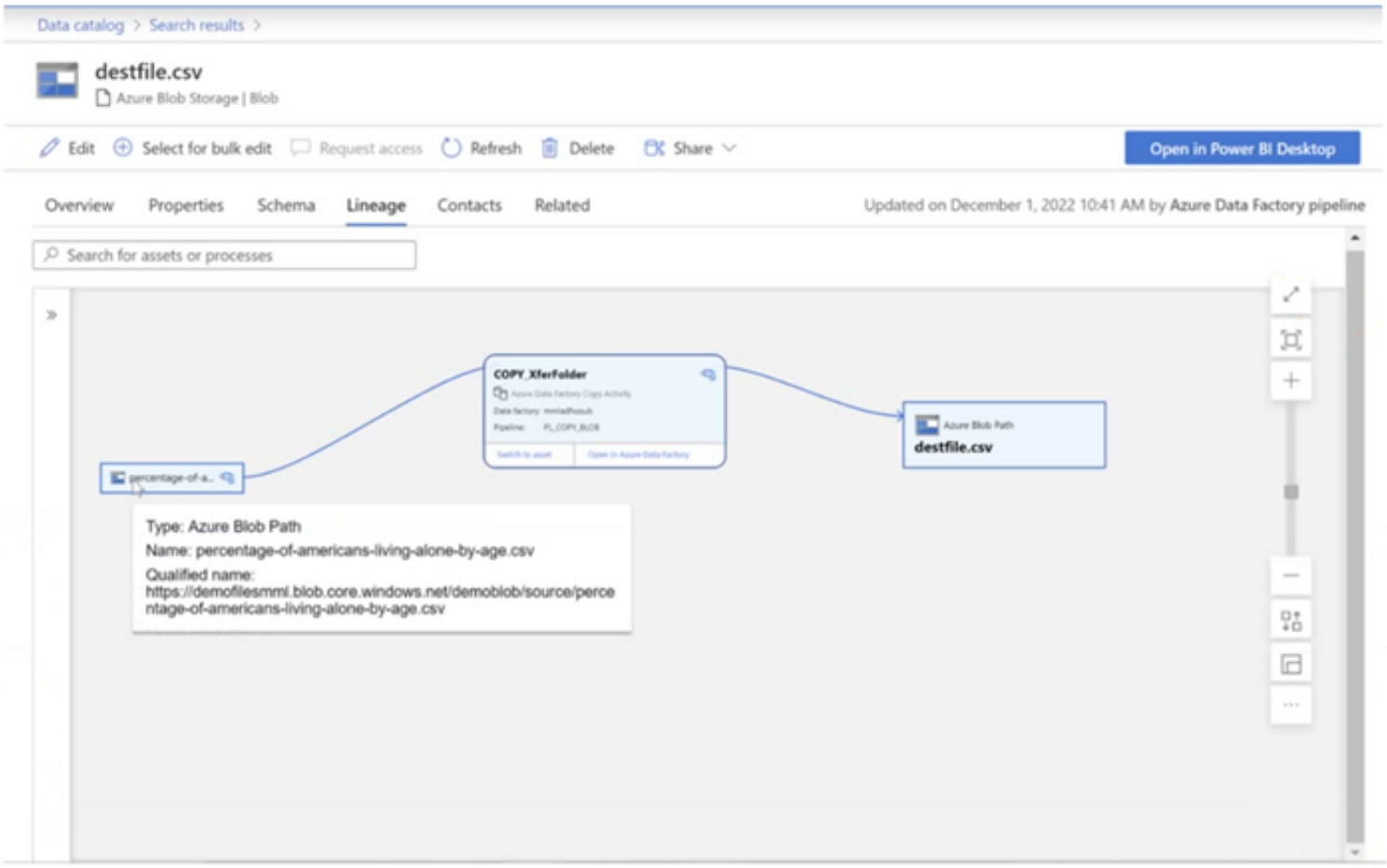
NEW QUESTION 17
- (Exam Topic 3)
You have an Azure Synapse Analytics Apache Spark pool named Pool1.
You plan to load JSON files from an Azure Data Lake Storage Gen2 container into the tables in Pool1. The structure and data types vary by file.
You need to load the files into the tables. The solution must maintain the source data types. What should you do?

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

Answer: A

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

NEW QUESTION 20
- (Exam Topic 3)
You have a Microsoft Purview account. The Lineage view of a CSV file is shown in the following exhibit.



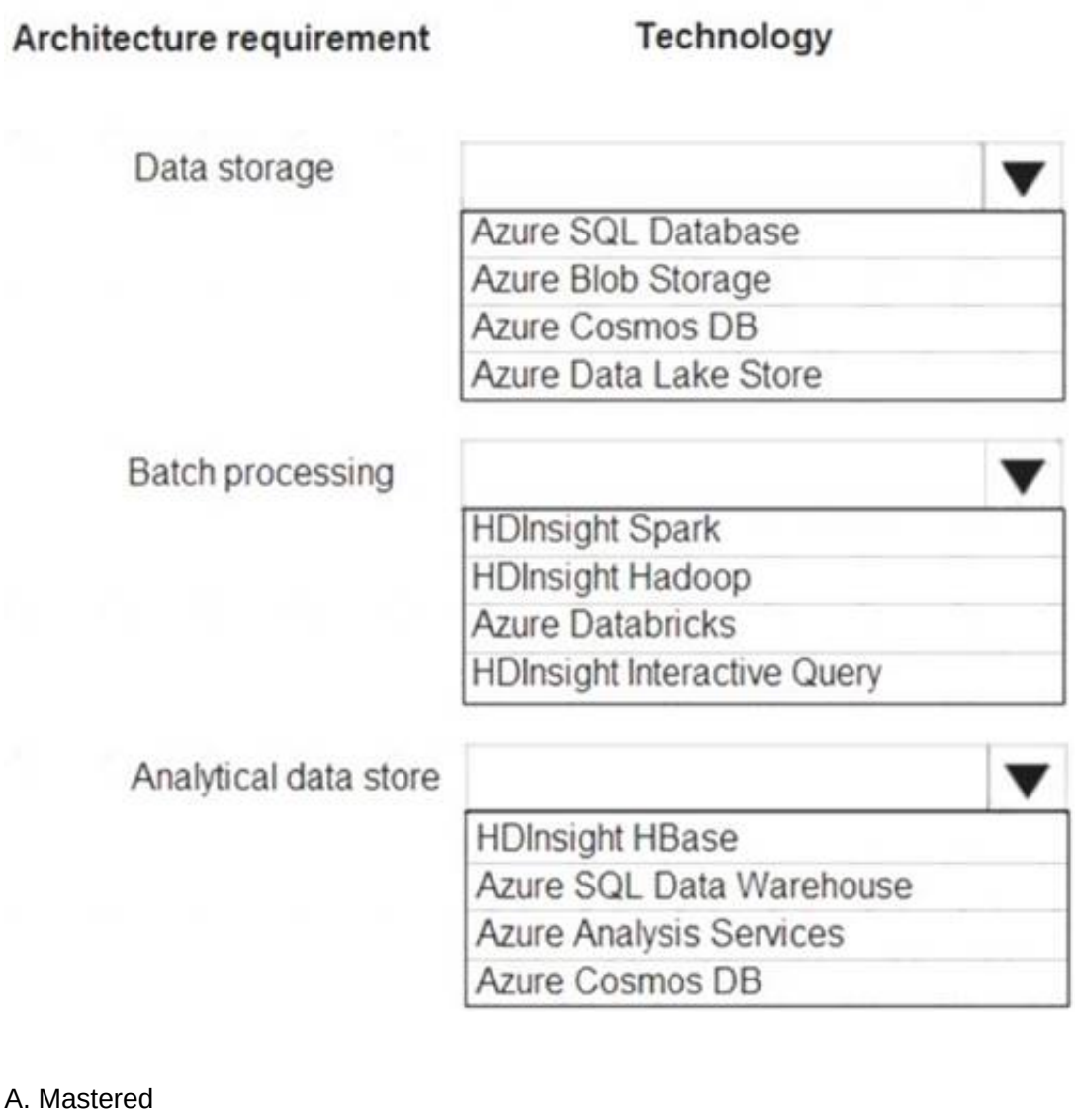
How is the data for the lineage populated?

A. manually
B. by scanning data stores
C. by executing a Data Factory pipeline

Answer: B

Explanation:
According to Microsoft Purview Data Catalog lineage user guide1, data lineage in Microsoft Purview is a core platform capability that populates the Microsoft Purview Data Map with data movement and transformations across systems2. Lineage is captured as it flows in the enterprise and stitched without gaps irrespective of its source2.

NEW QUESTION 22
- (Exam Topic 3)
You are developing a solution using a Lambda architecture on Microsoft Azure. The data at test layer must meet the following requirements:
Data storage:
•Serve as a repository (or high volumes of large files in various formats.
•Implement optimized storage for big data analytics workloads.
•Ensure that data can be organized using a hierarchical structure. Batch processing:
•Use a managed solution for in-memory computation processing.
•Natively support Scala, Python, and R programming languages.
•Provide the ability to resize and terminate the cluster automatically. Analytical data store:
•Support parallel processing.
•Use columnar storage.
•Support SQL-based languages.
You need to identify the correct technologies to build the Lambda architecture.
Which technologies should you use? To answer, select the appropriate options in the answer area NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 25
- (Exam Topic 3)

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

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

Answer: C

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

NEW QUESTION 26
- (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 27
- (Exam Topic 3)
You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account.
You need to output the count of tweets from the last five minutes every minute. Which windowing function should you use?

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

Answer: D

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

NEW QUESTION 29
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool that contains the users shown in the following table.

Name	Role
User1	Server admin
User2	db_datereader

User1 executes a query on the database, and the query returns the results shown in the following exhibit.

```
SELECT c.name,
    tbl.name as table_name,
    typ.name as datatype,
    c.is_masked,
    c.masking_function
FROM sys.masked_columns AS c
INNER JOIN sys.tables AS tbl ON c.[object_id] = tbl.[object_id]
INNER JOIN sys.types typ ON c.user_type_id = typ.user_type_id
WHERE is_masked = 1;
```

Results Messages

	name	table_name	datatype	is_masked	masking_function
1	BirthDate	DimCustomer	date	1	default()
2	Gender	DimCustomer	nvarchar	1	default()
3	EmailAddress	DimCustomer	nvarchar	1	email()
4	YearlyIncome	DimCustomer	money	1	default()

User1 is the only user who has access to the unmasked data.
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.

When User2 queries the YearlyIncome column, the values returned will be **[answer choice]**.

- a random number
- the values stored in the database
- XXXX
- 0

When User1 queries the BirthDate column, the values returned will be **[answer choice]**.

- a random date
- the values stored in the database
- XXXX
- 1900-01-01

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application, email Description automatically generated
Box 1: 0
The YearlyIncome column is of the money data type.
The Default masking function: Full masking according to the data types of the designated fields

- Use a zero value for numeric data types (bigint, bit, decimal, int, money, numeric, smallint, smallmoney, tinyint, float, real).

Box 2: the values stored in the database
Users with administrator privileges are always excluded from masking, and see the original data without any mask.
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

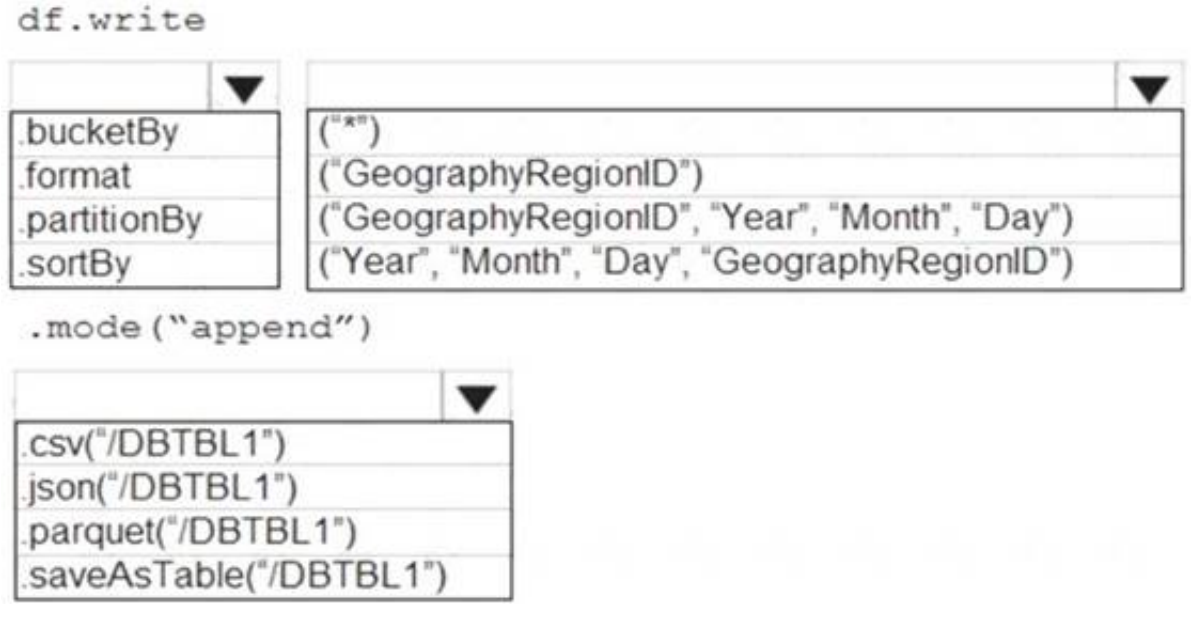
NEW QUESTION 33
- (Exam Topic 3)
You develop a dataset named DBTBL1 by using Azure Databricks. DBTBL1 contains the following columns:

- 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 34
- (Exam Topic 3)
You have an Azure Stream Analytics job.
You need to ensure that the job has enough streaming units provisioned. You configure monitoring of the SU % Utilization metric.
Which two additional metrics should you monitor? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Backlogged Input Events
B. Watermark Delay
C. Function Events
D. Out of order Events
E. Late Input Events

Answer: AB

Explanation:
To react to increased workloads and increase streaming units, consider setting an alert of 80% on the SU Utilization metric. Also, you can use watermark delay and backlogged events metrics to see if there is an impact.
Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job, by increasing the SUs.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-monitoring

NEW QUESTION 38
- (Exam Topic 3)
You are designing a partition strategy for a fact table in an Azure Synapse Analytics dedicated SQL pool. The table has the following specifications:
• Contain sales data for 20,000 products.
• Use hash distribution on a column named ProduclID,
• Contain 2.4 billion records for the years 20l9 and 2020.
Which number of partition ranges provides optimal compression and performance of the clustered columnstore index?

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

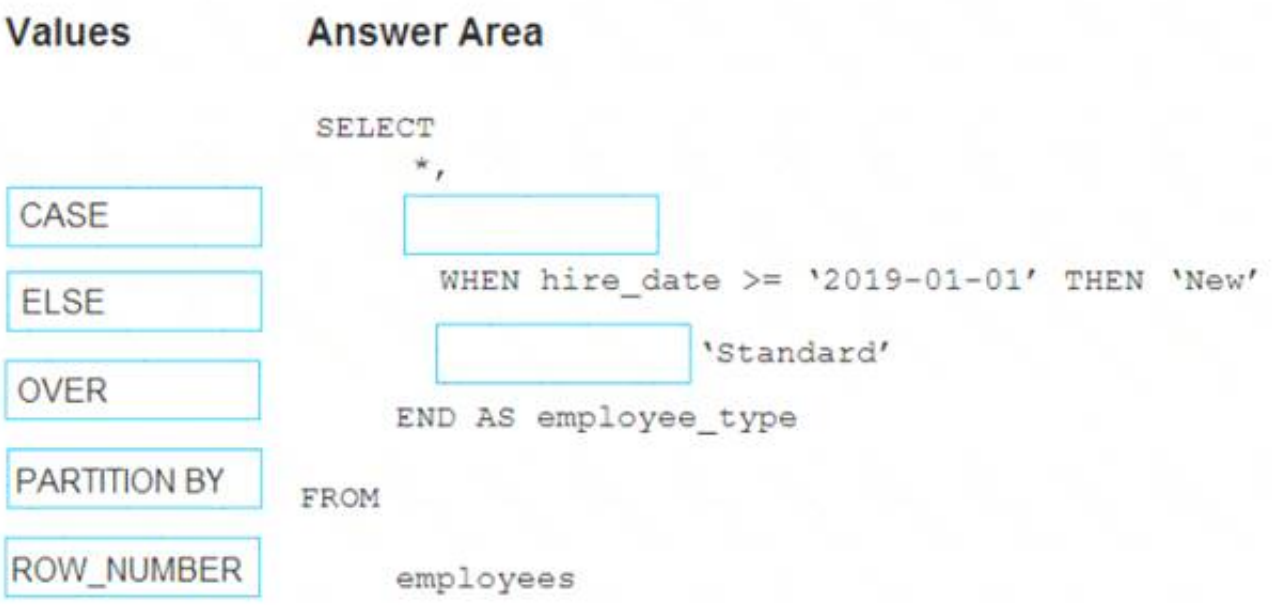
Answer: A

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

NEW QUESTION 39
- (Exam Topic 3)
You have the following table named Employees.

first_name	last_name	hire_date	employee_type
Jane	Doe	2019-08-23	new
Ben	Smith	2017-12-15	Standard

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



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, text, application Description automatically generated
Box 1: CASE
CASE evaluates a list of conditions and returns one of multiple possible result expressions.
CASE can be used in any statement or clause that allows a valid expression. For example, you can use CASE in statements such as SELECT, UPDATE, DELETE and SET, and in clauses such as select_list, IN, WHERE, ORDER BY, and HAVING.
Syntax: Simple CASE expression: CASE input_expression
WHEN when_expression THEN result_expression [...n] [ELSE else_result_expression]
END
Box 2: ELSE
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/language-elements/case-transact-sql

NEW QUESTION 42
- (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 46
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers.
You need to create an external data source in Pool1 that will be used to read .orc files in storage1. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore

WITH

(  Location1 '  [          ] ://data@newyorktaxidataset.dfs.core.windows.net' ,
                  abfs
                  abfss
                  wasb
                  wasbs

credential = ADLS_credential ,

TYPE -  [                  ]
                  BLOB_STORAGE
);        HADOOP
          RDBMS
          SHARP MAP MANAGER
```

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 51
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 receives new data once every 24 hours.
You have the following function.

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

You have the following query.

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

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

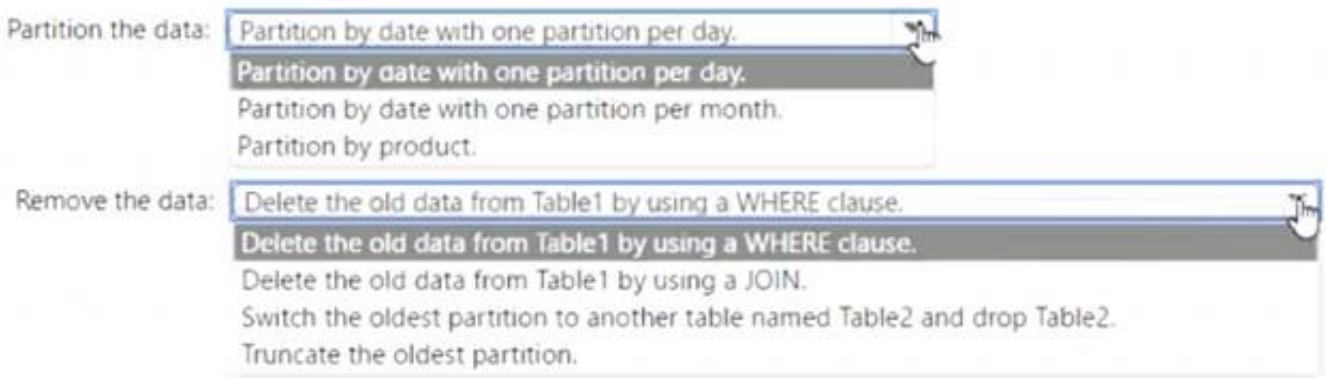
A. Create an index on the avg_f column.
B. Convert the avg_c column into a calculated column.
C. Create an index on the sensorid column.
D. Enable result set caching.
E. Change the table distribution to replicate.

Answer: BD

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

NEW QUESTION 56
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Tablet. Table1 contains sales data. Sixty-five million rows of data are added to Table1 monthly.
At the end of each month, you need to remove data that is older than 36 months. The solution must minimize how long it takes to remove the data.
How should you partition Table1, and how should you remove the old data? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

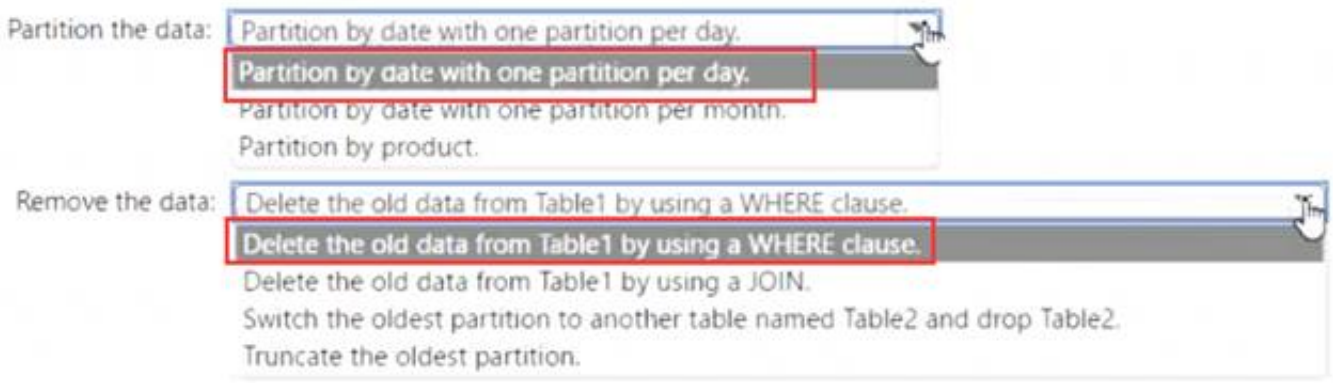
Answer Area



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Answer Area



NEW QUESTION 60
- (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:

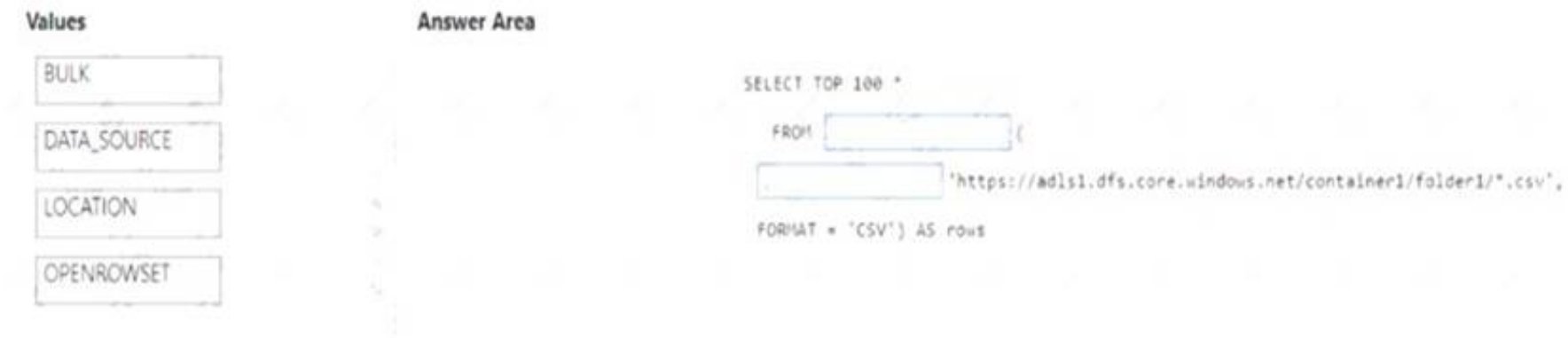
- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 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

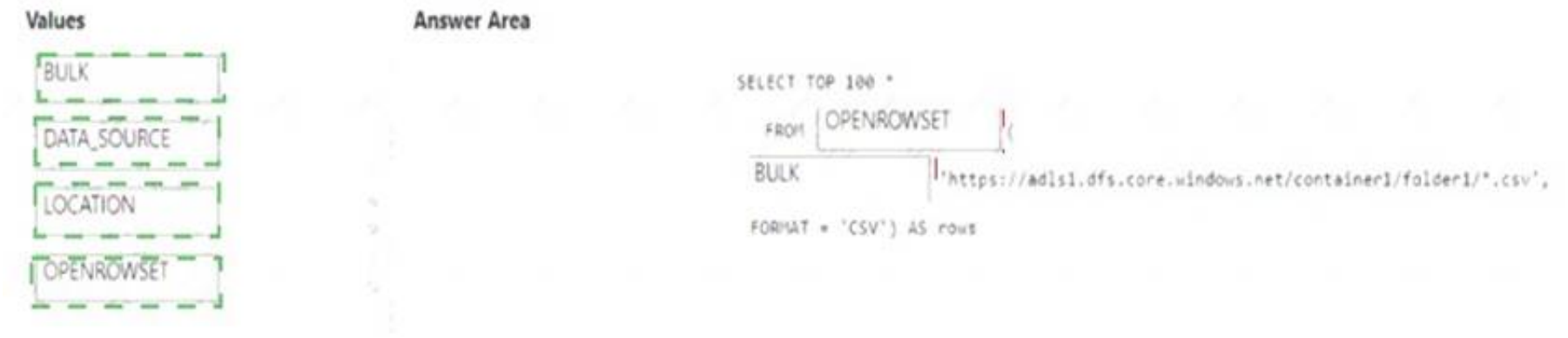
NEW QUESTION 63
- (Exam Topic 3)
You have an Azure Synapse Analytics serverless SQ1 pool.
You have an Azure Data Lake Storage account named aols1 that contains a public container named container1 The container 1 container contains a folder named folder 1.
You need to query the top 100 rows of all the CSV files in folder 1.
How shouk1 you complete the query? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all.
You may need to drag the split bar between panes or scroll to view content.
NOTE Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

Explanation:



NEW QUESTION 66
- (Exam Topic 3)
You are creating a new notebook in Azure Databricks that will support R as the primary language but will also support Scale and SOL Which switch should you use to switch between languages?

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

Answer: B

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

NEW QUESTION 71
- (Exam Topic 3)
You build a data warehouse in an Azure Synapse Analytics dedicated SQL pool.
Analysts write a complex SELECT query that contains multiple JOIN and CASE statements to transform data for use in inventory reports. The inventory reports will use the data and additional WHERE parameters depending on the report. The reports will be produced once daily.
You need to implement a solution to make the dataset available for the reports. The solution must minimize query times.
What should you implement?

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

Answer: A

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

NEW QUESTION 72
- (Exam Topic 3)

You use Azure Data Lake Storage Gen2.
You need to ensure that workloads can use filter predicates and column projections to filter data at the time the data is read from disk.
Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. Reregister the Microsoft Data Lake Store resource provider.
B. Reregister the Azure Storage resource provider.
C. Create a storage policy that is scoped to a container.
D. Register the query acceleration feature.
E. Create a storage policy that is scoped to a container prefix filter.

Answer: BD

NEW QUESTION 73
- (Exam Topic 3)
You have an Azure subscription that contains a Microsoft Purview account named MP1, an Azure data factory named DF1, and a storage account named storage.
MP1 is configured
10 scan storage1. DF1 is connected to MP1 and contains 3 dataset named DS1. DS1 references 2 file in storage.
In DF1, you plan to create a pipeline that will process data from DS1.
You need to review the schema and lineage information in MP1 for the data referenced by DS1.
Which two features can you use to locate the information? Each correct answer presents a complete solution. NOTE: Each correct answer is worth one point.

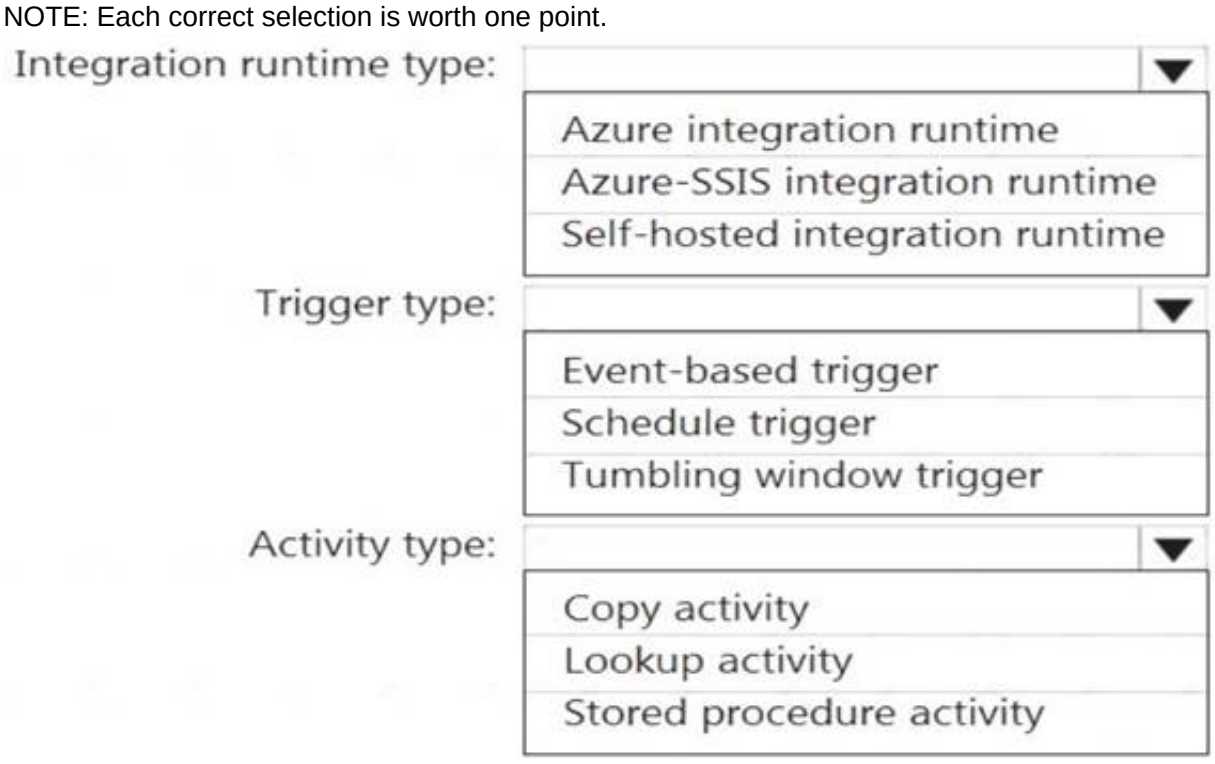
A. the Storage browser of storage1 in the Azure portal
B. the search bar in the Azure portal
C. the search bar in Azure Data Factory Studio
D. the search bar in the Microsoft Purview governance portal

Answer: CD

Explanation:
- The search bar in the Microsoft Purview governance portal: This is a feature that allows you to search for assets in your data estate using keywords, filters, and facets. You can use the search bar to find the files in storage1 that are referenced by DS1, and then view their schema and lineage information in the asset details page12.
- The search bar in Azure Data Factory Studio: This is a feature that allows you to search for datasets, linked services, pipelines, and other resources in your data factory. You can use the search bar to find DS1 in DF1, and then view its schema and lineage information in the dataset details page. You can also click on the Open in Purview button to open the corresponding asset in MP13.

The two features that can be used to locate the schema and lineage information for the data referenced by DS1 are the search bar in Azure Data Factory Studio and the search bar in the Microsoft Purview governance portal.
The search bar in Azure Data Factory Studio allows you to search for the dataset DS1 and view its properties and lineage. This can help you locate information about the source and destination data stores, as well as the transformations that were applied to the data.
The search bar in the Microsoft Purview governance portal allows you to search for the storage account and view its metadata, including schema and lineage information. This can help you understand the different data assets that are stored in the storage account and how they are related to each other.
The Storage browser of storage1 in the Azure portal may allow you to view the files that are stored in the storage account, but it does not provide lineage or schema information for those files. Similarly, the search bar in the Azure portal may allow you to search for resources in the Azure subscription, but it does not provide detailed information about the data assets themselves.
References:
- What is Azure Purview?
- Use Azure Data Factory Studio

NEW QUESTION 77
- (Exam Topic 2)
Which Azure Data Factory components should you recommend using together to import the daily inventory data from the SQL server to Azure Data Lake Storage?
To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

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

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

NEW QUESTION 78
- (Exam Topic 1)
You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements.
Which Azure Storage functionality should you include in the solution?

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

Answer: D

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

NEW QUESTION 82
- (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 86
- (Exam Topic 3)
You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytics dedicated SQL pool. The CSV file contains columns named username, comment and date.
The data flow already contains the following:
• A source transformation
• A Derived Column transformation to set the appropriate types of data
• A sink transformation to land the data in the pool
You need to ensure that the data flow meets the following requirements;
• All valid rows must be written to the destination table.
• Truncation errors in the comment column must be avoided proactively.
• Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.
Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point

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

Answer: BD

NEW QUESTION 89
- (Exam Topic 3)
You have an Azure Synapse Analytics serverless SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named storage1. The AllowedBlobpublicAccess porperty is disabled for storage1.
You need to create an external data source that can be used by Azure Active Directory (Azure AD) users to access storage1 from Pool1.
What should you create first?

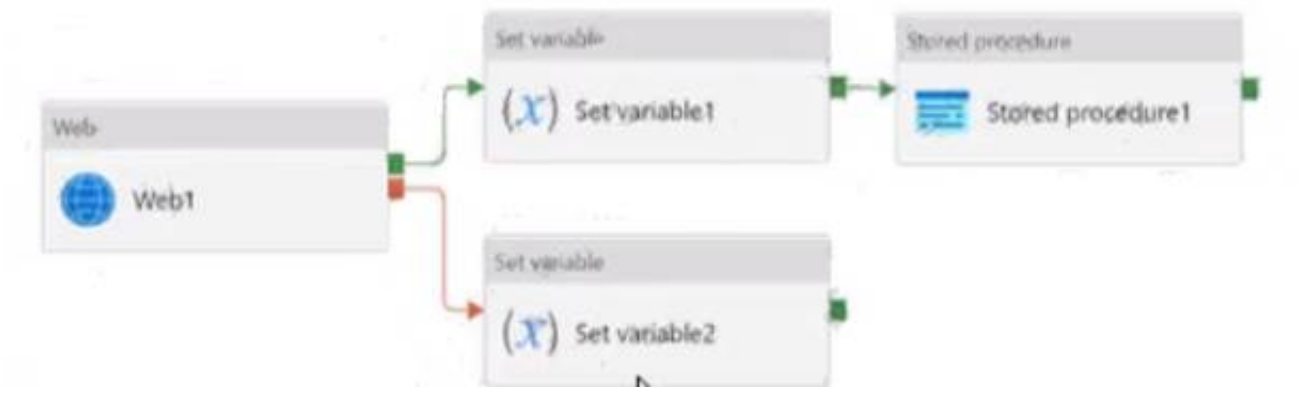
A. an external resource pool
B. a remote service binding
C. database scoped credentials

D. an external library

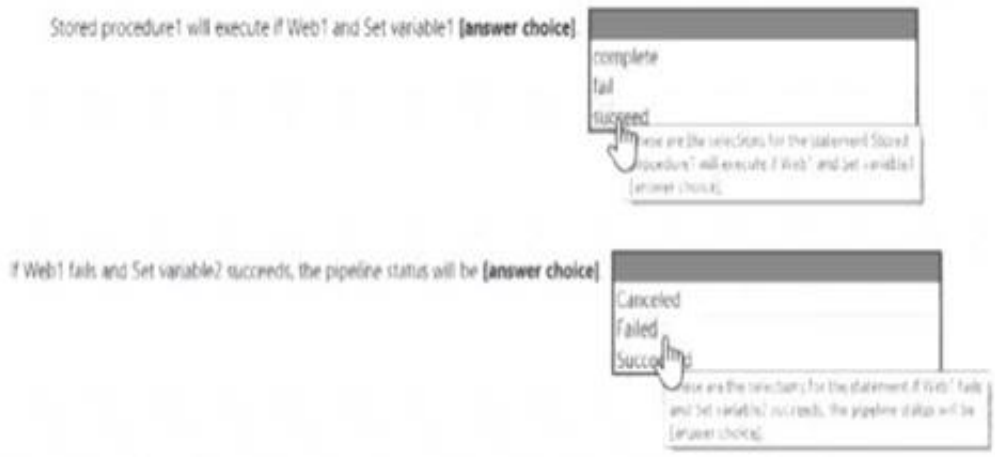
Answer: C

Explanation:
Security
User must have SELECT permission on an external table to read the data. External tables access underlying Azure storage using the database scoped credential defined in data source.
Note: A database scoped credential is a record that contains the authentication information that is required to connect to a resource outside SQL Server. Most credentials include a Windows user and password.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables https://docs.microsoft.com/en-us/sql/t-sql/statements/create-database-scoped-credential-transact-sql

NEW QUESTION 91
- (Exam Topic 3)
You have an Azure Data Factory pipeline that has the activity shown in the following exhibit.



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

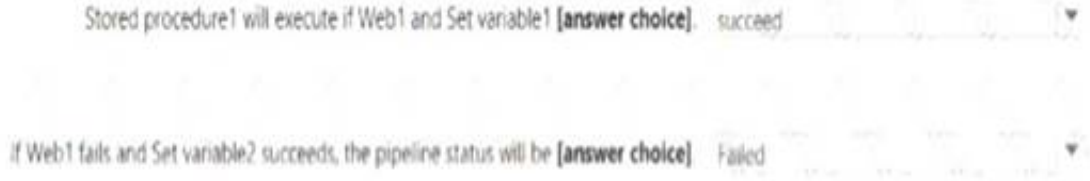


A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area



NEW QUESTION 93
- (Exam Topic 3)
You have an enterprise data warehouse in Azure Synapse Analytics.
You need to monitor the data warehouse to identify whether you must scale up to a higher service level to accommodate the current workloads
Which is the best metric to monitor?
More than one answer choice may achieve the goal. Select the BEST answer.

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

Answer: C

NEW QUESTION 94
- (Exam Topic 3)

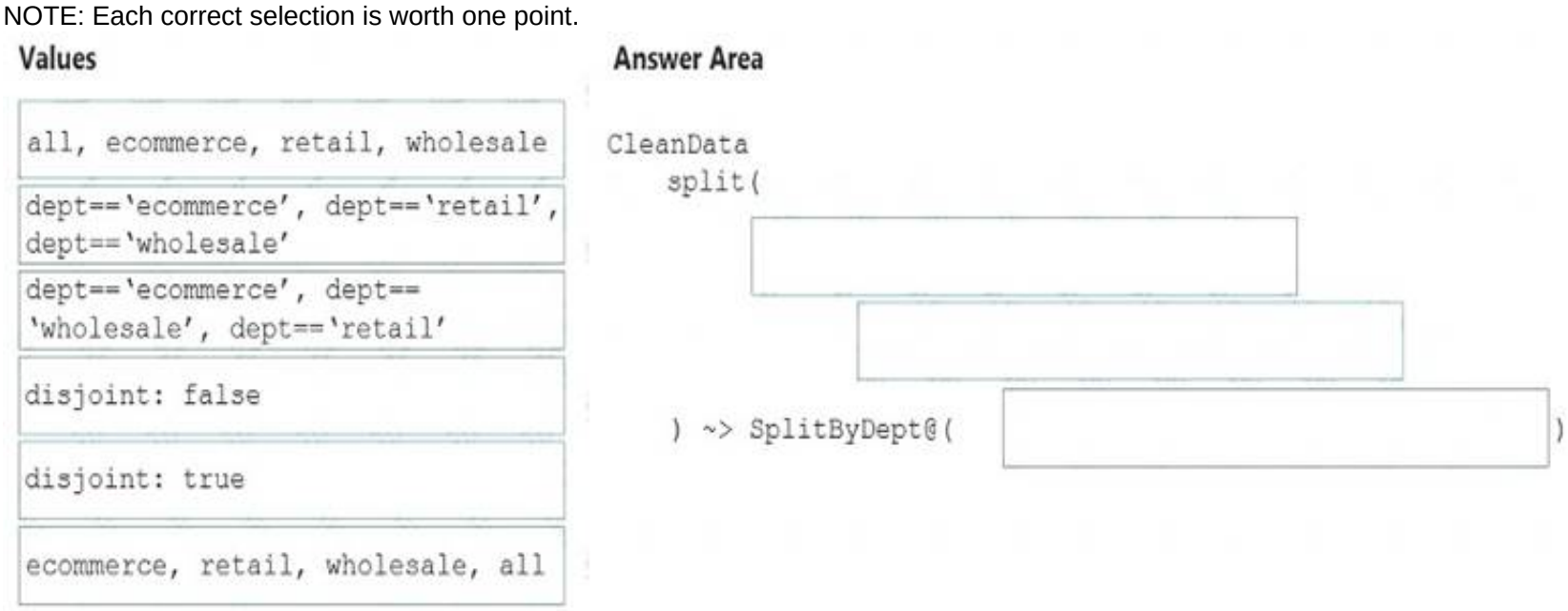
You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account.
The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/. You need to design a daily Azure Data Factory data load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

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

Answer: BD

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

NEW QUESTION 95
- (Exam Topic 3)
You need to create an Azure Data Factory pipeline to process data for the following three departments at your company: Ecommerce, retail, and wholesale. The solution must ensure that data can also be processed for the entire company.
How should you complete the Data Factory data flow script? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 100
- (Exam Topic 3)
You are designing an Azure Synapse Analytics dedicated SQL pool.
You need to ensure that you can audit access to Personally Identifiable information (PII). What should you include in the solution?

A. dynamic data masking
B. row-level security (RLS)
C. sensitivity classifications
D. column-level security

Answer: C

Explanation:
Data Discovery & Classification is built into Azure SQL Database, Azure SQL Managed Instance, and Azure Synapse Analytics. It provides basic capabilities for discovering, classifying, labeling, and reporting the sensitive data in your databases.
Your most sensitive data might include business, financial, healthcare, or personal information. Discovering and classifying this data can play a pivotal role in your organization's information-protection approach. It can serve as infrastructure for:

- Helping to meet standards for data privacy and requirements for regulatory compliance.
- Various security scenarios, such as monitoring (auditing) access to sensitive data.
- Controlling access to and hardening the security of databases that contain highly sensitive data.

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

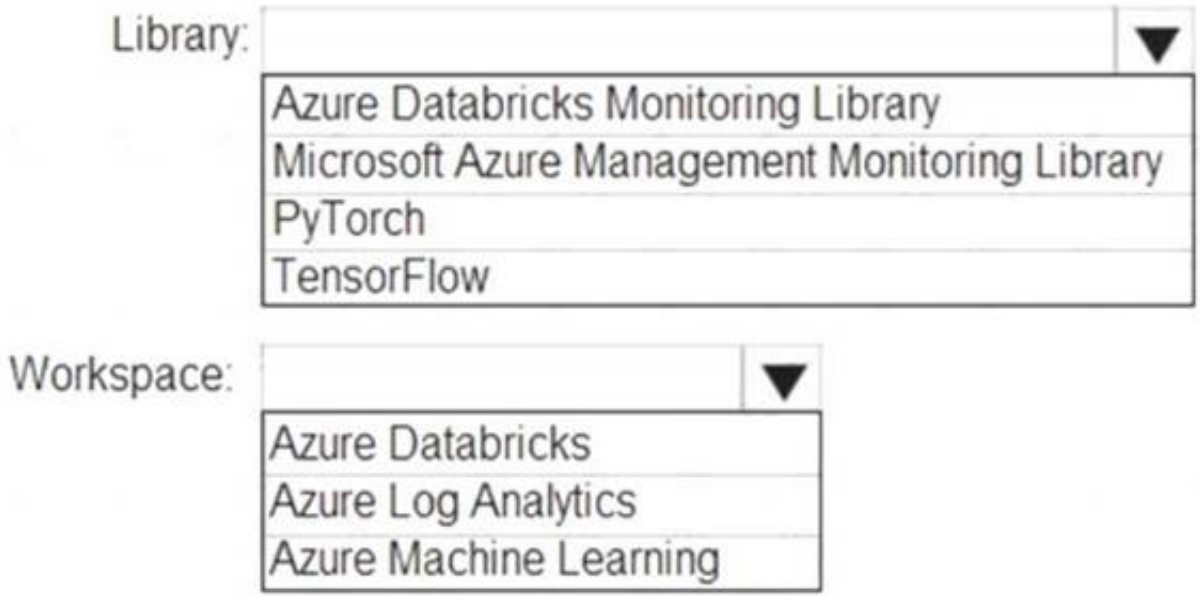
NEW QUESTION 102
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named
container1.
You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.
You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.
Solution: You use a dedicated SQL pool to create an external table that has an additional DateTime column. Does this meet the goal?

A. Yes
B. No

Answer: B

Explanation:
Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column

NEW QUESTION 107
- (Exam Topic 3)
You need to collect application metrics, streaming query events, and application log messages for an Azure Databrick cluster.
Which type of library and workspace should you implement? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



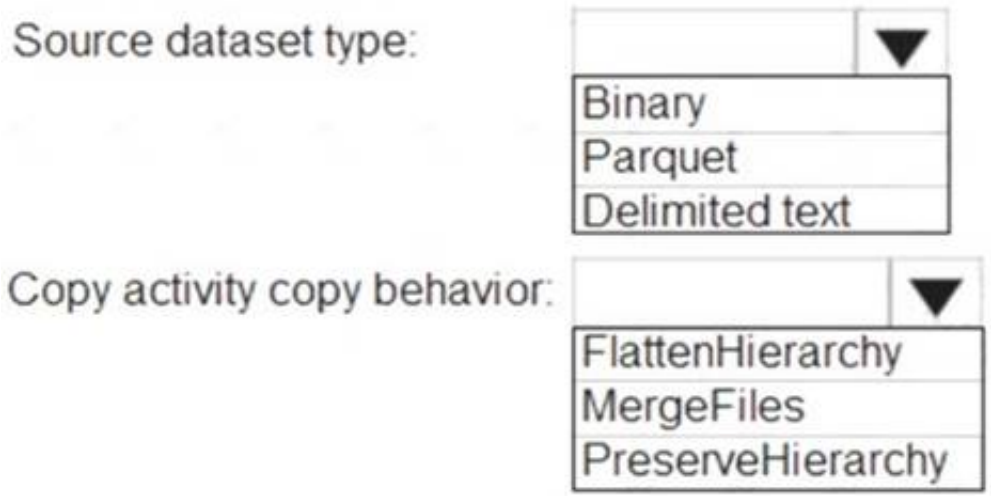
A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 112
- (Exam Topic 3)
You have two Azure Storage accounts named Storage1 and Storage2. Each account holds one container and has the hierarchical namespace enabled. The system has files that contain data stored in the Apache Parquet format.
You need to copy folders and files from Storage1 to Storage2 by using a Data Factory copy activity. The
solution must meet the following requirements:
- No transformations must be performed.
- The original folder structure must be retained.
- Minimize time required to perform the copy activity.

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



A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 114
- (Exam Topic 3)
You have a table in an Azure Synapse Analytics dedicated SQL pool. The table was created by using the following Transact-SQL statement.

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

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

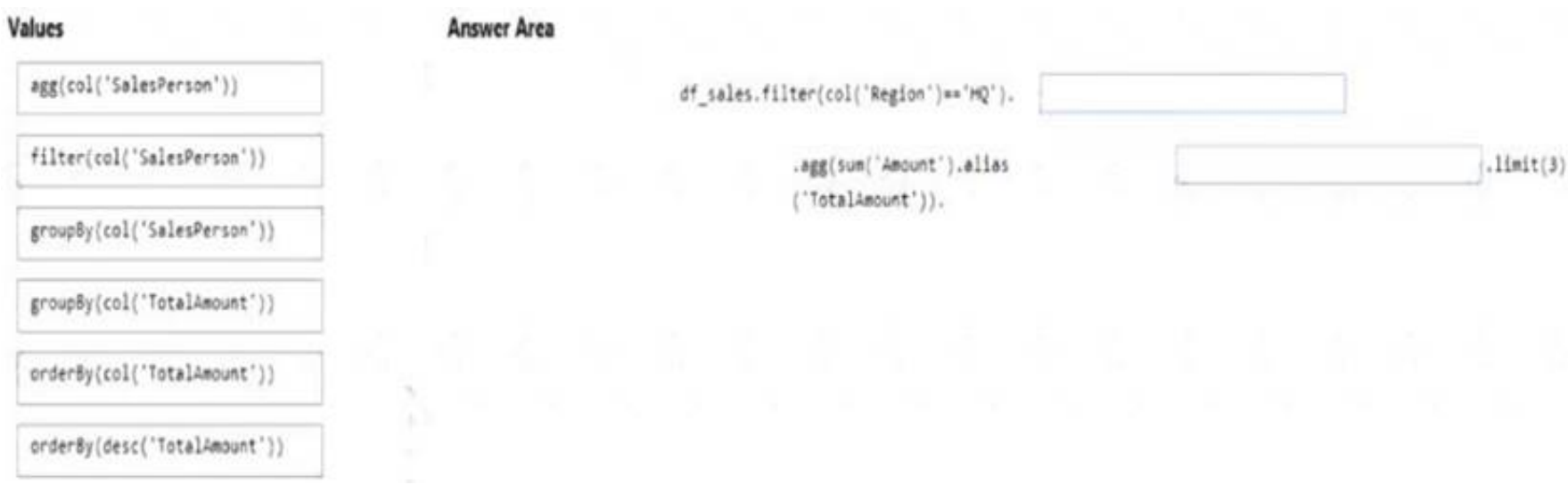
Which column should you add to the table?

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

Answer: A

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

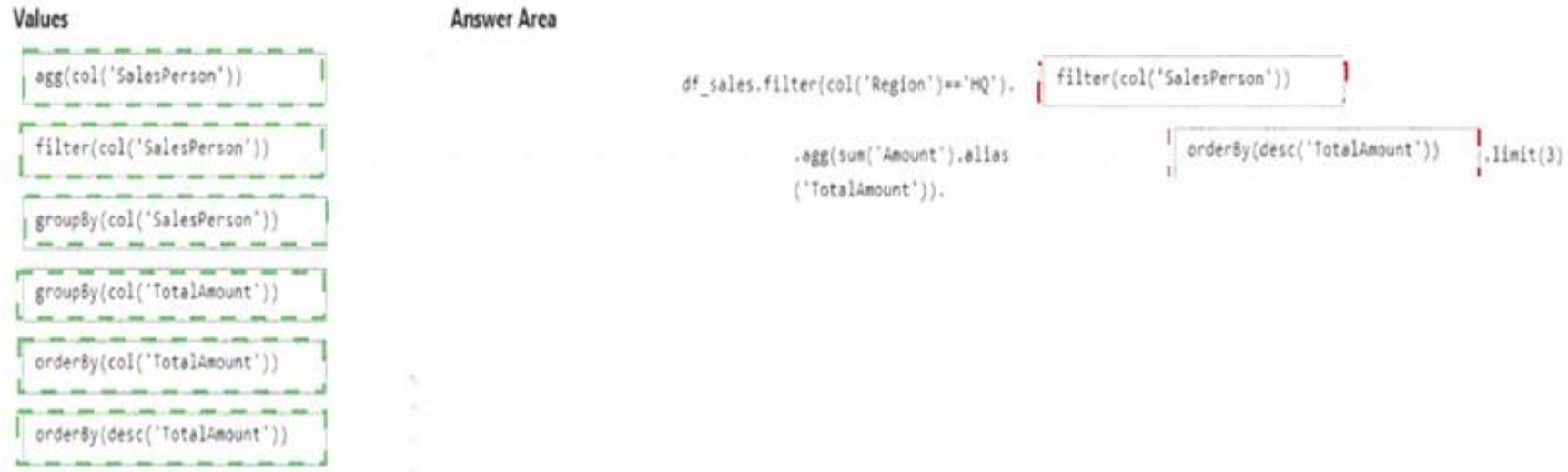
NEW QUESTION 116
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Databricks workspace. The workspace contains a notebook named Notebook1. In Notebook1, you create an Apache Spark DataFrame named df_sales that contains the following columns:
• Customer
• Salesperson
• Region
• Amount
You need to identify the three top performing salespersons by amount for a region named HQ.
How should you complete the query? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.



A. Mastered
B. Not Mastered

Answer: A

Explanation:



NEW QUESTION 118
- (Exam Topic 3)
You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination.
You need to ensure that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs.
What should you do?

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

Answer: D

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

NEW QUESTION 123
- (Exam Topic 3)
You are designing a statistical analysis solution that will use custom proprietary1 Python functions on near real-time data from Azure Event Hubs.
You need to recommend which Azure service to use to perform the statistical analysis. The solution must minimize latency.
What should you recommend?

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

Answer: A

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

NEW QUESTION 126
- (Exam Topic 3)
You plan to implement an Azure Data Lake Gen2 storage account.
You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs.

Which type of replication should you use for the storage account?

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

Answer: C

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

NEW QUESTION 130
- (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 132
- (Exam Topic 3)
You have a partitioned table in an Azure Synapse Analytics dedicated SQL pool. You need to design queries to maximize the benefits of partition elimination. What should you include in the Transact-SQL queries?

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

Answer: B

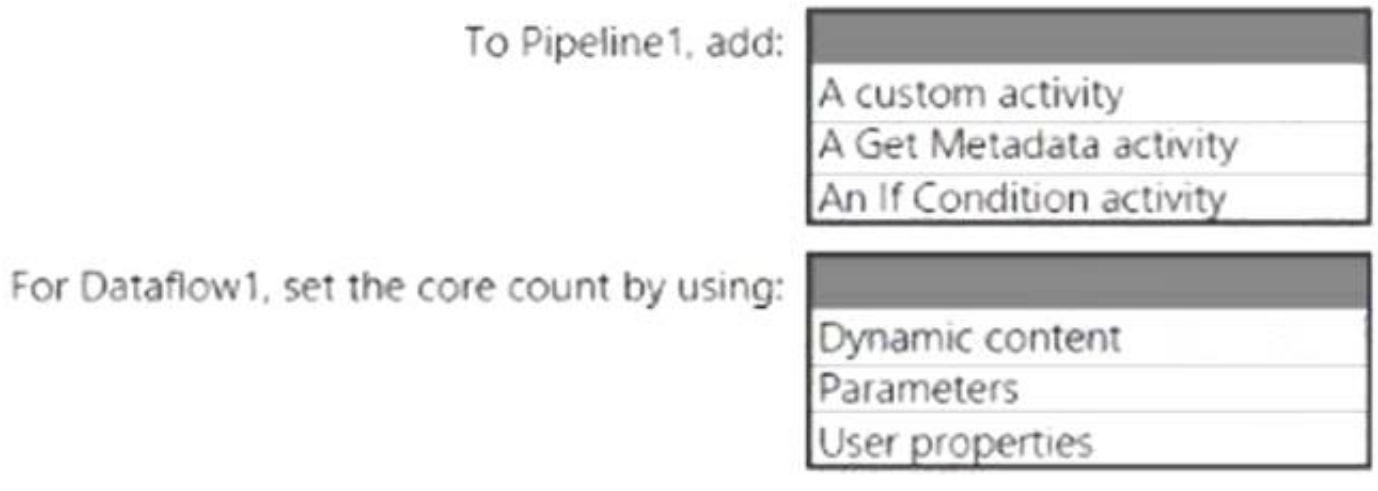
NEW QUESTION 136
- (Exam Topic 3)
You are designing a streaming data solution that will ingest variable volumes of data. You need to ensure that you can change the partition count after creation. Which service should you use to ingest the data?

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

Answer: B

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

NEW QUESTION 139
- (Exam Topic 3)
You have an Azure Synapse Analytics pipeline named Pipeline1 that contains a data flow activity named Dataflow1.
Pipeline1 retrieves files from an Azure Data Lake Storage Gen 2 account named storage1.
Dataflow1 uses the AutoResolveIntegrationRuntime integration runtime configured with a core count of 128. You need to optimize the number of cores used by Dataflow1 to accommodate the size of the files in storage1. What should you configure? To answer, select the appropriate options in the answer area.



A. Mastered
B. Not Mastered

Answer: A

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

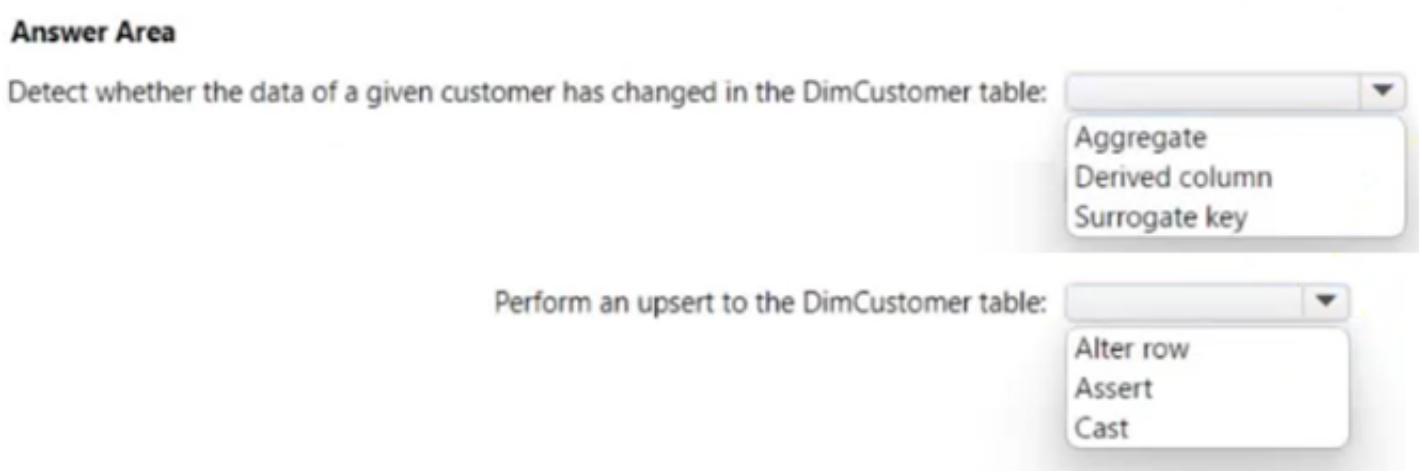
NEW QUESTION 143
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct 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 5 seconds and a window size 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

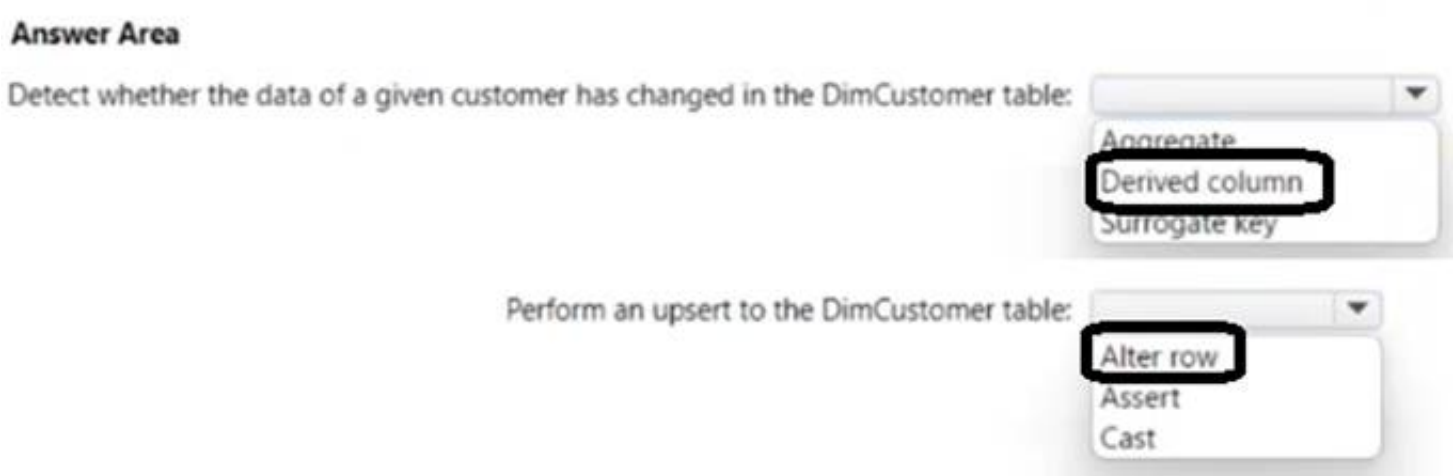
NEW QUESTION 144
- (Exam Topic 3)
You are developing an Azure Synapse Analytics pipeline that will include a mapping data flow named Dataflow1. Dataflow1 will read customer data from an external source and use a Type 1 slowly changing dimension (SCO) when loading the data into a table named DimCustomer1 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 148
- (Exam Topic 3)
You plan to ingest streaming social media data by using Azure Stream Analytics. The data will be stored in files in Azure Data Lake Storage, and then consumed by using Azure Datiabricks and PolyBase in Azure Synapse Analytics.
You need to recommend a Stream Analytics data output format to ensure that the queries from Databricks and PolyBase against the files encounter the fewest possible errors. The solution must ensure that the tiles can be queried quickly and that the data type information is retained.
What should you recommend?

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

Answer: A

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

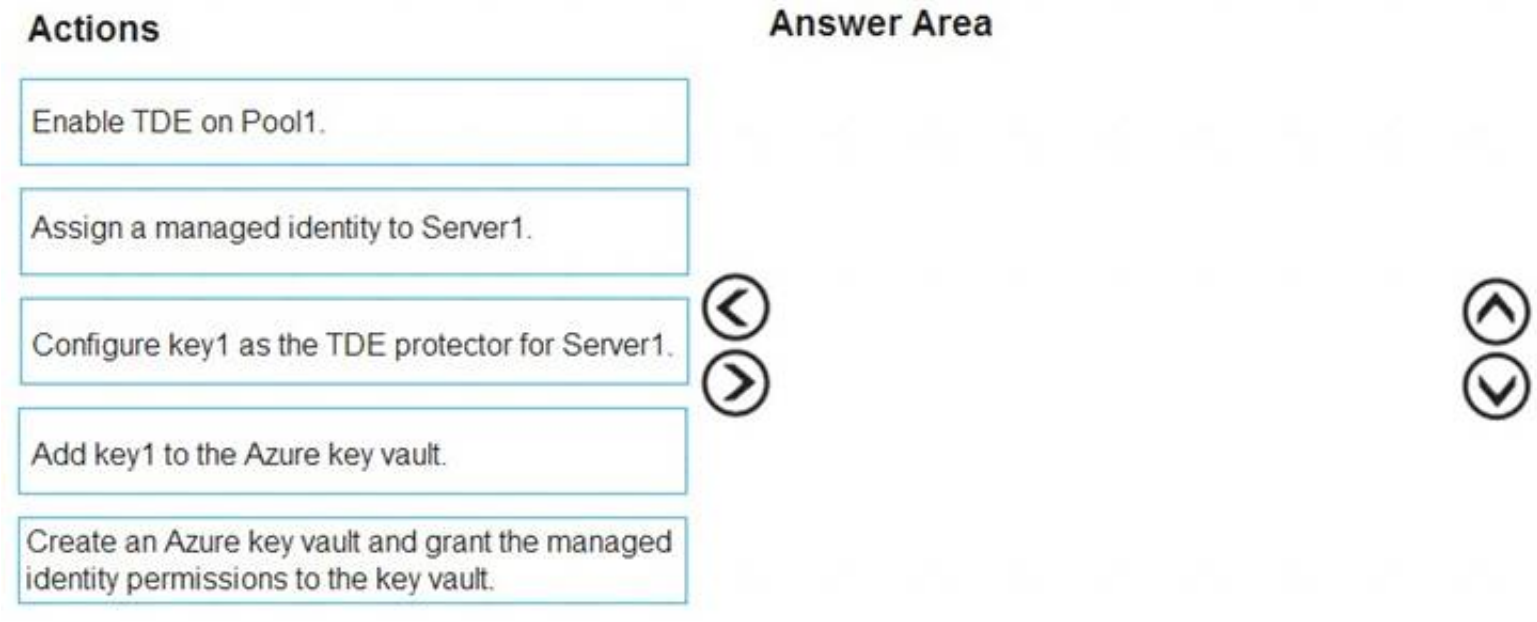
NEW QUESTION 150
- (Exam Topic 3)
You are building an Azure Analytics query that will receive input data from Azure IoT Hub and write the results to Azure Blob storage.
You need to calculate the difference in readings per sensor per hour.
How should you complete the query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: LAG
The LAG analytic operator allows one to look up a “previous” event in an event stream, within certain constraints. It is very useful for computing the rate of growth of a variable, detecting when a variable crosses a threshold, or when a condition starts or stops being true.
Box 2: LIMIT DURATION
Example: Compute the rate of growth, per sensor: SELECT sensorId,
growth = reading
LAG(reading) OVER (PARTITION BY sensorId LIMIT DURATION(hour, 1)) FROM input
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/lag-azure-stream-analytics

NEW QUESTION 153
- (Exam Topic 3)
You have an Azure Synapse Analytics SQL pool named Pool1 on a logical Microsoft SQL server named Server1.
You need to implement Transparent Data Encryption (TDE) on Pool1 by using a custom key named key1. Which five actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.



A. Mastered

B. Not Mastered

Answer: A

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

NEW QUESTION 155
- (Exam Topic 3)
You have an Azure Storage account and a data warehouse in Azure Synapse Analytics in the UK South region. You need to copy blob data from the storage account to the data warehouse by using Azure Data Factory. The solution must meet the following requirements:

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

Which type of integration runtime should you use?
A. Azure integration runtime
B. Azure-SSIS integration runtime
C. Self-hosted integration runtime

Answer: A

Explanation:

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

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

NEW QUESTION 157
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL Pool1. Pool1 contains a partitioned fact table named dbo.Sales and a staging table named stg.Sales that has the matching table and partition definitions.
You need to overwrite the content of the first partition in dbo.Sales with the content of the same partition in stg.Sales. The solution must minimize load times.
What should you do?

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

Answer: A

NEW QUESTION 162
- (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 166
- (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

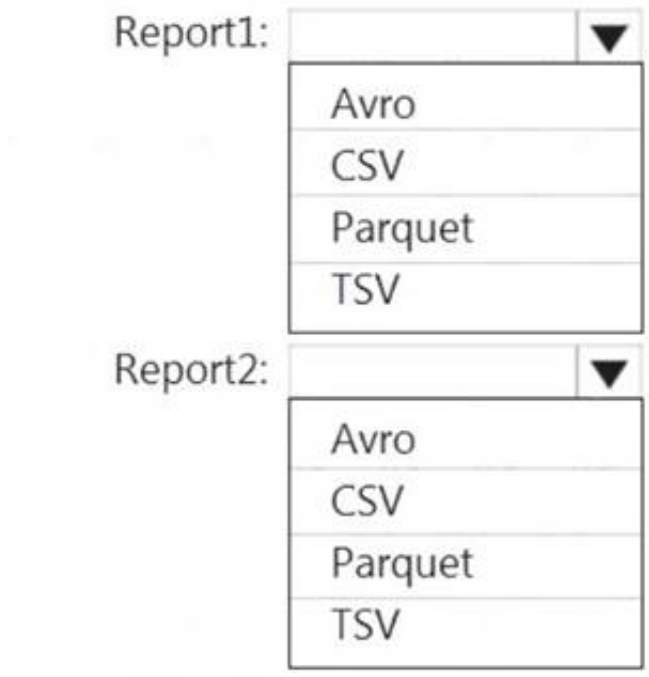
NEW QUESTION 169
- (Exam Topic 3)
You have a data warehouse in Azure Synapse Analytics.
You need to ensure that the data in the data warehouse is encrypted at rest. What should you enable?

A. Advanced Data Security for this database
B. Transparent Data Encryption (TDE)
C. Secure transfer required
D. Dynamic Data Masking

Answer: B

Explanation:
Azure SQL Database currently supports encryption at rest for Microsoft-managed service side and client-side encryption scenarios.
- Support for server encryption is currently provided through the SQL feature called Transparent Data Encryption.
- Client-side encryption of Azure SQL Database data is supported through the Always Encrypted feature. Reference:
https://docs.microsoft.com/en-us/azure/security/fundamentals/encryption-atrest

NEW QUESTION 171
- (Exam Topic 3)
You are planning the deployment of Azure Data Lake Storage Gen2. You have the following two reports that will access the data lake:
- Report1: Reads three columns from a file that contains 50 columns.
- Report2: Queries a single record based on a timestamp.
You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.
What should you recommend for each report? 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:
Report1: CSV
CSV: The destination writes records as delimited data. Report2: AVRO
AVRO supports timestamps.
Not Parquet, TSV: Not options for Azure Data Lake Storage Gen2. Reference:
https://streamsets.com/documentation/datacollector/latest/help/datacollector/UserGuide/Destinations/ADLS-G2

NEW QUESTION 172
- (Exam Topic 3)
You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container.
Which resource provider should you enable?

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

Answer: C

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

NEW QUESTION 175
- (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 179
- (Exam Topic 3)
You are incrementally loading data into fact tables in an Azure Synapse Analytics dedicated SQL pool. Each batch of incoming data is staged before being loaded into the fact tables. |
You need to ensure that the incoming data is staged as quickly as possible. |
How should you configure the staging tables? To answer, select the appropriate options in the answer area.

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 181
- (Exam Topic 3)

You have an enterprise data warehouse in Azure Synapse Analytics.
Using PolyBase, you create an external table named [Ext].[Items] to query Parquet files stored in Azure Data Lake Storage Gen2 without importing the data to the data warehouse.
The external table has three columns.
You discover that the Parquet files have a fourth column named ItemID.
Which command should you run to add the ItemID column to the external table?

A.
```
ALTER EXTERNAL TABLE [Ext].[Items]
  ADD [ItemID] int;
```

B.
```
DROP EXTERNAL FILE FORMAT parquetfile1;
CREATE EXTERNAL FILE FORMAT parquetfile1
WITH (
     FORMAT_TYPE = PARQUET,
     DATA_COMPRESSION = 'org.apache.hadoop.io.compress.SnappyCodec'
);
```

C.
```
DROP EXTERNAL TABLE [Ext].[Items]
CREATE EXTERNAL TABLE [Ext].[Items]
([ItemID] [int] NULL,
 [ItemName] nvarchar(50) NULL,
 [ItemType] nvarchar(20) NULL,
 [ItemDescription] nvarchar(250))
WITH
(
     LOCATION= '/Items/',
          DATA_SOURCE = AzureDataLakeStore,
          FILE_FORMAT = PARQUET,
          REJECT_TYPE = VALUE,
          REJECT_VALUE = 0
);
```

D.
```
ALTER TABLE [Ext].[Items]
ADD [ItemID] int;
```

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

Answer: C

Explanation:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql

NEW QUESTION 184
- (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 186
- (Exam Topic 3)
You are designing an Azure Synapse Analytics workspace.
You need to recommend a solution to provide double encryption of all the data at rest.
Which two components should you include in the recommendation? Each coned answer presents part of the solution
NOTE: Each correct selection is worth one point.

A. an X509 certificate

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

Answer: BC

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

NEW QUESTION 187
- (Exam Topic 3)
You have an Azure Synapse Analytics dedicated SQL pool.
You run PDW_SHOWSPACEUSED(dbo,FactInternetSales'); and get the results shown in the following table.

ROWS	RESERVED_SPACE	DATA_SPACE	INDEX_SPACE	UNUSED_SPACE	PDW_NODE_ID	DISTRIBUTION_ID
694	2776	616	48	2112	1	1
407	2704	576	48	2080	1	2
53	2376	512	16	1848	1	3
58	2376	512	16	1848	1	4
168	2632	528	32	2072	1	5
195	2696	536	32	2128	1	6
5995	3464	1424	32	2008	1	7
0	2232	496	0	1736	1	8
264	2576	544	40	1992	1	9
3008	3016	960	32	2024	1	10
-	-	-	-	-	-	-
1550	2832	752	48	2032	1	50
1238	2832	696	40	2096	1	51
192	2632	528	32	2072	1	52
1127	2760	680	48	2040	1	53
1244	3032	704	64	2264	1	54
409	2632	568	32	2032	1	55
0	2232	496	0	1736	1	56
1437	2832	728	40	2064	1	57
0	2232	496	0	1736	1	58
584	2632	560	32	2040	1	59
225	2760	544	40	2184	1	60

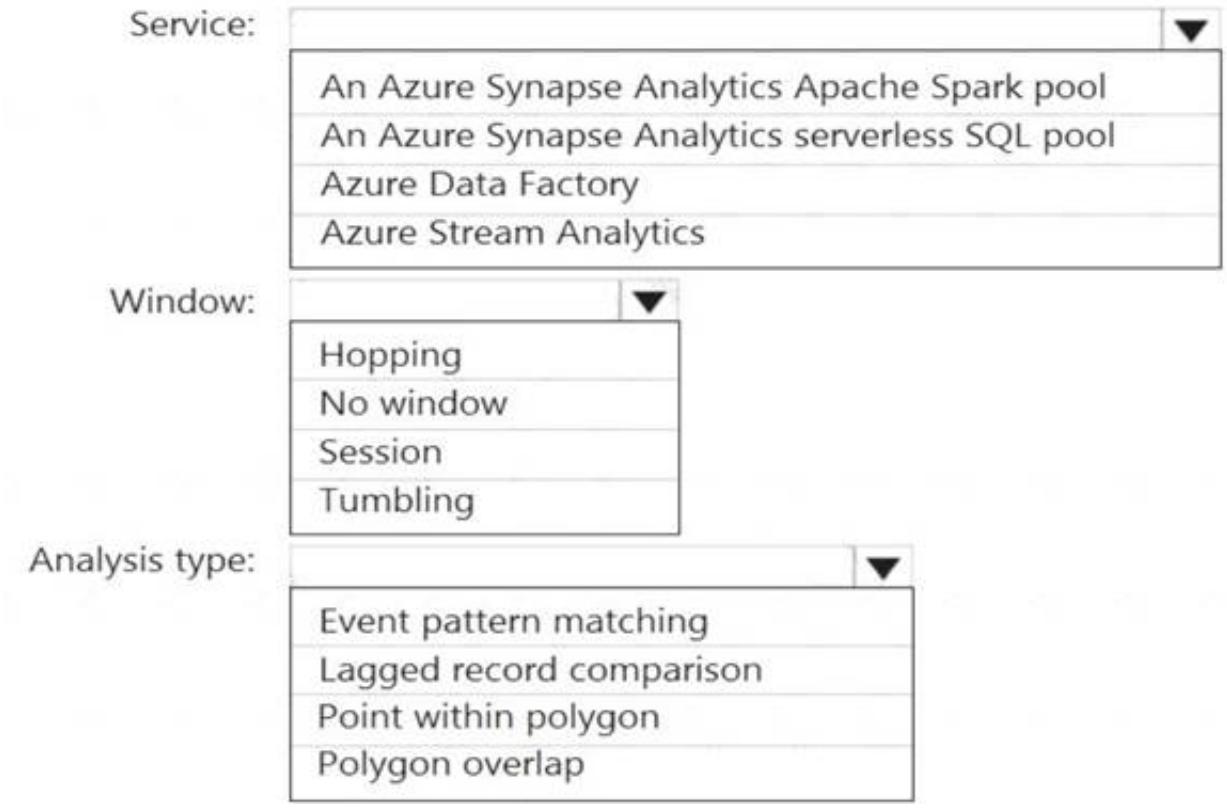
Which statement accurately describes the dbo,FactInternetSales table?

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

Answer: C

Explanation:
Data skew means the data is not distributed evenly across the distributions. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

NEW QUESTION 191
- (Exam Topic 3)
You are designing a monitoring solution for a fleet of 500 vehicles. Each vehicle has a GPS tracking device that sends data to an Azure event hub once per minute.
You have a CSV file in an Azure Data Lake Storage Gen2 container. The file maintains the expected geographical area in which each vehicle should be.
You need to ensure that when a GPS position is outside the expected area, a message is added to another event hub for processing within 30 seconds. The solution must minimize cost.
What should you include in the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

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

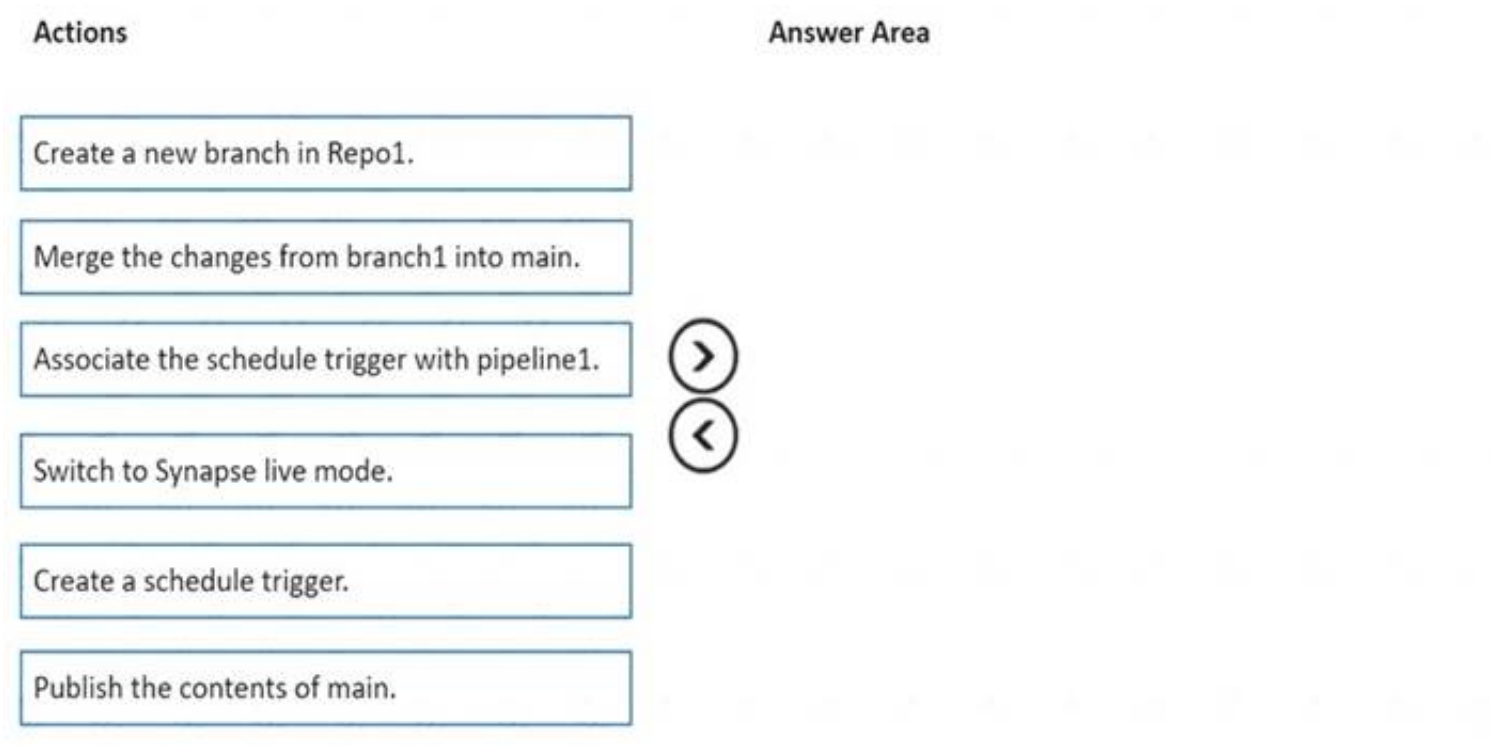
NEW QUESTION 193
- (Exam Topic 3)
You create an Azure Databricks cluster and specify an additional library to install. When you attempt to load the library to a notebook, the library in not found.
You need to identify the cause of the issue. What should you review?

A. notebook logs
B. cluster event logs
C. global init scripts logs
D. workspace logs

Answer: C

Explanation:
Cluster-scoped Init Scripts: Init scripts are shell scripts that run during the startup of each cluster node before the Spark driver or worker JVM starts. Databricks customers use init scripts for various purposes such as installing custom libraries, launching background processes, or applying enterprise security policies.
Logs for Cluster-scoped init scripts are now more consistent with Cluster Log Delivery and can be found in the same root folder as driver and executor logs for the cluster.
Reference:
https://databricks.com/blog/2018/08/30/introducing-cluster-scoped-init-scripts.html

NEW QUESTION 197
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1. Branch1 contains an Azure Synapse pipeline named pipeline1.
In workspace1, you complete testing of pipeline1. You need to schedule pipeline1 to run daily at 6 AM.
Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Timeline Description automatically generated

NEW QUESTION 202
- (Exam Topic 3)
You have a SQL pool in Azure Synapse.
A user reports that queries against the pool take longer than expected to complete. You need to add monitoring to the underlying storage to help diagnose the issue.
Which two metrics should you monitor? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. Cache used percentage
B. DWU Limit
C. Snapshot Storage Size
D. Active queries
E. Cache hit percentage

Answer: AE

Explanation:
A: Cache used is the sum of all bytes in the local SSD cache across all nodes and cache capacity is the sum of the storage capacity of the local SSD cache across all nodes.
E: Cache hits is the sum of all columnstore segments hits in the local SSD cache and cache miss is the columnstore segments misses in the local SSD cache summed across all nodes
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-resou

NEW QUESTION 207
- (Exam Topic 3)
You have a SQL pool in Azure Synapse that contains a table named dbo.Customers. The table contains a column name Email.
You need to prevent nonadministrative users from seeing the full email addresses in the Email column. The users must see values in a format of aXXX@XXXX.com instead.
What should you do?

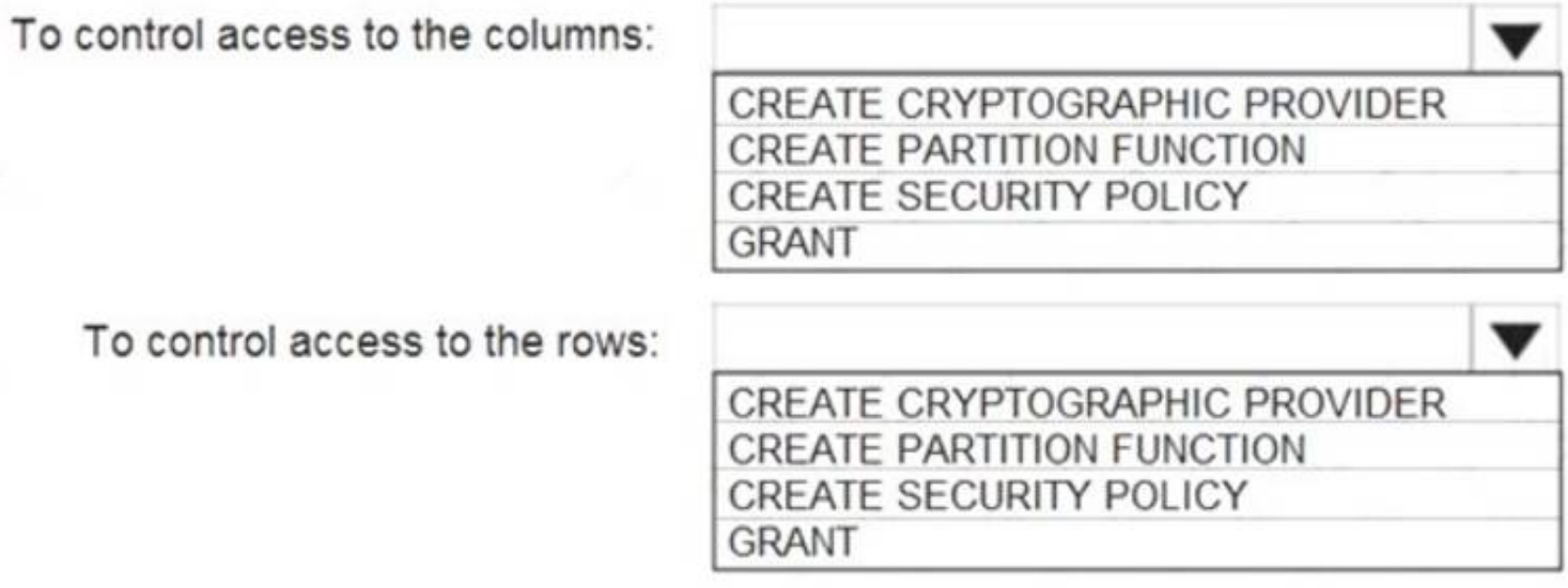
A. From Microsoft SQL Server Management Studio, set an email mask on the Email column.
B. From the Azure portal, set a mask on the Email column.
C. From Microsoft SQL Server Management studio, grant the SELECT permission to the users for all the columns in the dbo.Customers table except Email.
D. From the Azure portal, set a sensitivity classification of Confidential for the Email column.

Answer: D

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

NEW QUESTION 209
- (Exam Topic 3)
You have an Azure subscription that contains the following resources:
- An Azure Active Directory (Azure AD) tenant that contains a security group named Group1

An Azure Synapse Analytics SQL pool named Pool1
You need to control the access of Group1 to specific columns and rows in a table in Pool1.
Which Transact-SQL commands should you use? To answer, select the appropriate options in the answer area.



A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 213
- (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 216
- (Exam Topic 3)
You have an Azure subscription that contains an Azure Blob Storage account named storage1 and an Azure Synapse Analytics dedicated SQL pool named Pool1.
You need to store data in storage1. The data will be read by Pool1. The solution must meet the following requirements:
- Enable Pool1 to skip columns and rows that are unnecessary in a query.
- Automatically create column statistics.
- Minimize the size of files. Which type of file should you use?

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

Answer: B

Explanation:
Automatic creation of statistics is turned on for Parquet files. For CSV files, you need to create statistics manually until automatic creation of CSV files statistics is supported.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-statistics

NEW QUESTION 217

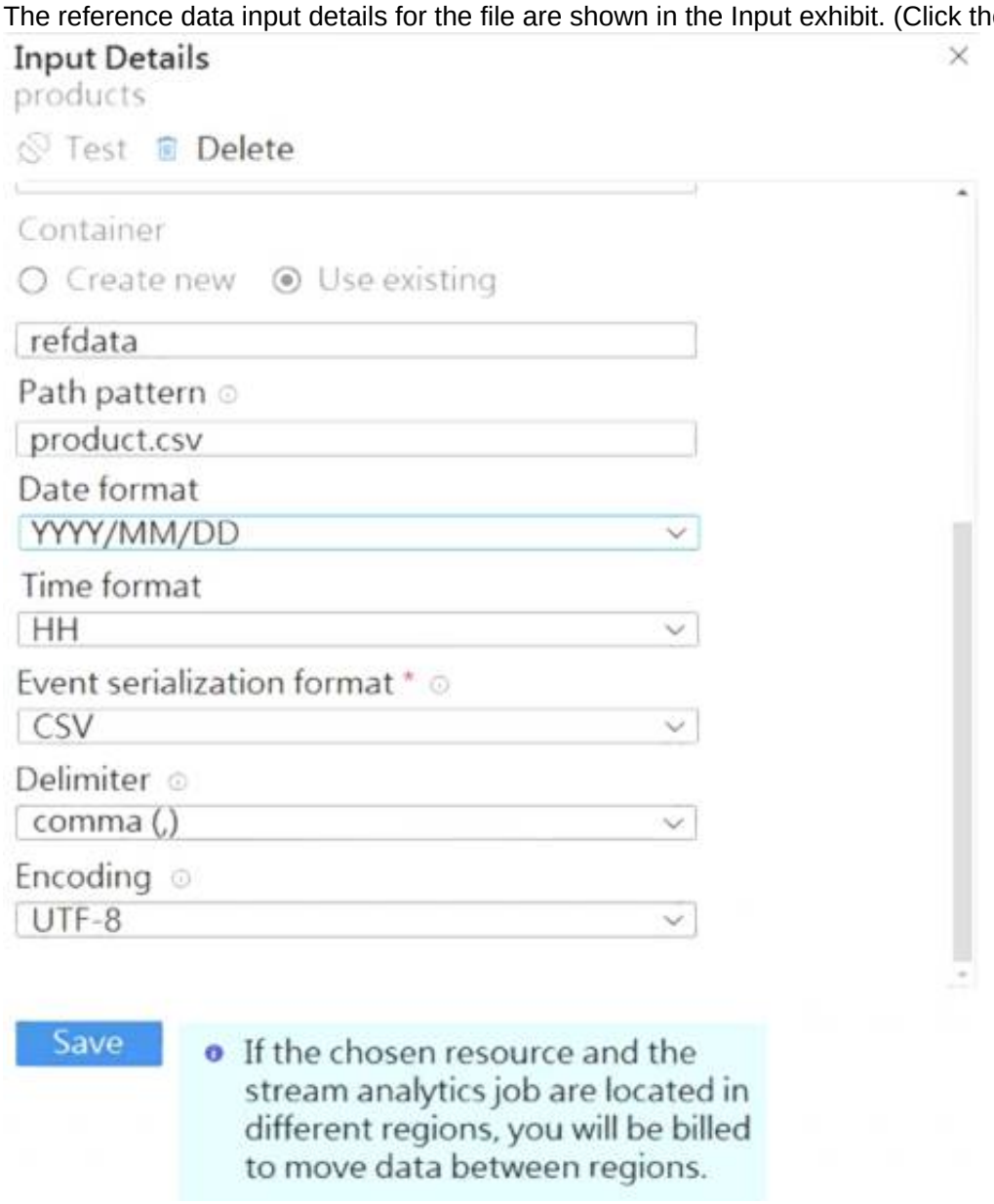
- (Exam Topic 3)
You are monitoring an Azure Stream Analytics job.
The Backlogged Input Events count has been 20 for the last hour. You need to reduce the Backlogged Input Events count.
What should you do?

A. Drop late arriving events from the job.
B. Add an Azure Storage account to the job.
C. Increase the streaming units for the job.
D. Stop the job.

Answer: C

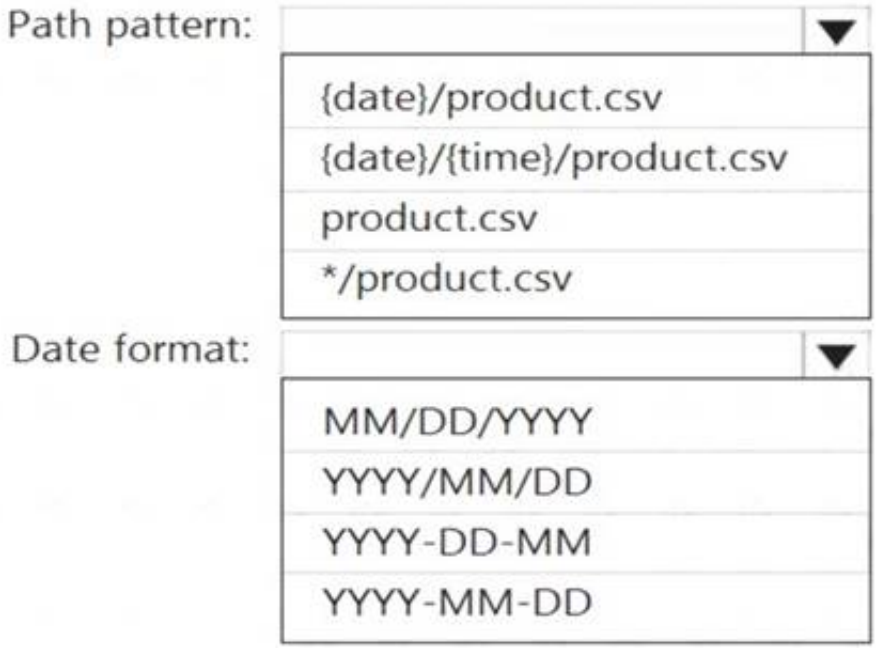
Explanation:
General symptoms of the job hitting system resource limits include:
If the backlog event metric keeps increasing, it's an indicator that the system resource is constrained (either because of output sink throttling, or high CPU).
Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job: adjust Streaming Units.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-scale-jobs https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-monitoring

NEW QUESTION 222
- (Exam Topic 3)
You are building an Azure Stream Analytics job that queries reference data from a product catalog file. The file is updated daily.
The reference data input details for the file are shown in the Input exhibit. (Click the Input tab.)



The storage account container view is shown in the Refdata exhibit. (Click the Refdata tab.)

You need to configure the Stream Analytics job to pick up the new reference data.
What should you configure? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

Explanation:
Graphical user interface, application, table Description automatically generated
Box 1: {date}/product.csv
In the 2nd exhibit we see: Location: refdata / 2020-03-20
Note: Path Pattern: This is a required property that is used to locate your blobs within the specified container. Within the path, you may choose to specify one or more instances of the following 2 variables:
{date}, {time}
Example 1: products/{date}/{time}/product-list.csv
Example 2: products/{date}/product-list.csv
Example 3: product-list.csv
Box 2: YYYY-MM-DD
Note: Date Format [optional]: If you have used {date} within the Path Pattern that you specified, then you can select the date format in which your blobs are organized from the drop-down of supported formats.
Example: YYYY/MM/DD, MM/DD/YYYY, etc. Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data

NEW QUESTION 225
- (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 226
- (Exam Topic 3)
You have an Azure subscription that contains the resources shown in the following table.

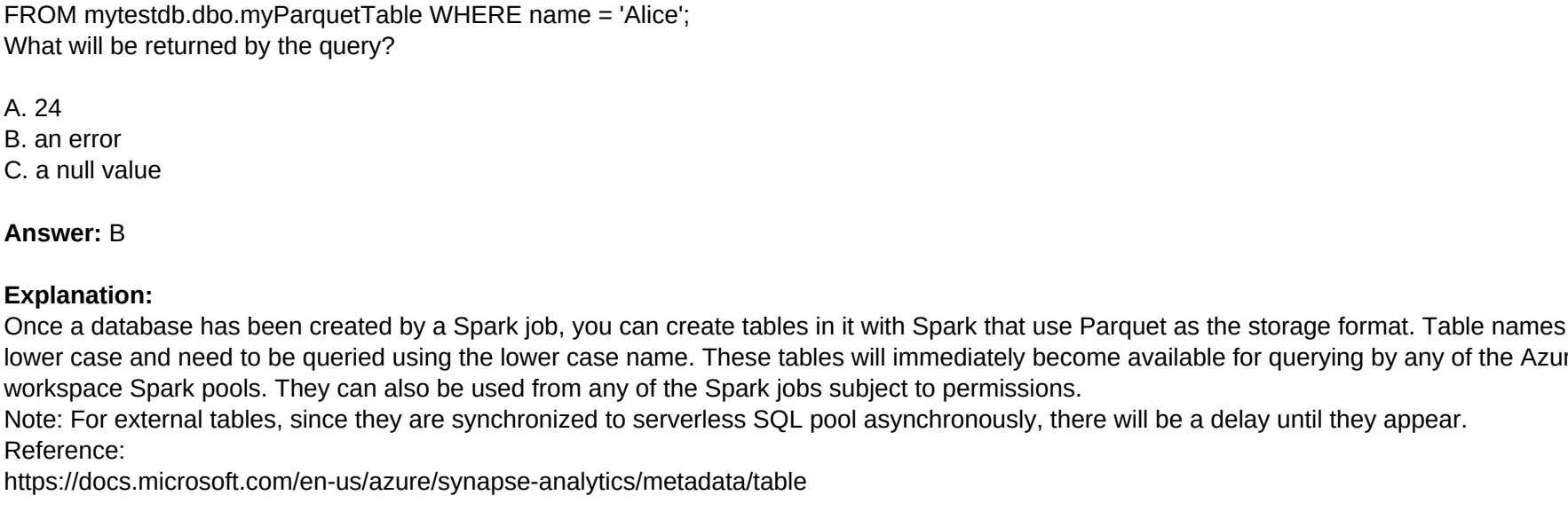
Name	Type	Description
ws1	Azure Synapse Analytics workspace	*None*
kv1	Azure Key Vault	*None*
UAMI1	User-assigned managed identity	Associated with ws1
sp1	Apache Spark pool in Azure Synapse Analytics	Associated with ws1

You need to ensure that you can Spark notebooks in ws1. The solution must ensure secrets from kv1 by using UAMI1. 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:



NEW QUESTION 229
- (Exam Topic 3)
You have an Azure Synapse workspace named MyWorkspace that contains an Apache Spark database named mytestdb.
You run the following command in an Azure Synapse Analytics Spark pool in MyWorkspace. CREATE TABLE mytestdb.myParquetTable(EmployeeID int, EmployeeName string, EmployeeStartDate date) USING Parquet
You then use Spark to insert a row into mytestdb.myParquetTable. The row contains the following data.

EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

One minute later, you execute the following query from a serverless SQL pool in MyWorkspace. SELECT EmployeeID FROM mytestdb.dbo.myParquetTable WHERE name = 'Alice';
What will be returned by the query?

A. 24
B. an error
C. a null value

Answer: B

Explanation:
Once a database has been created by a Spark job, you can create tables in it with Spark that use Parquet as the storage format. Table names will be converted to lower case and need to be queried using the lower case name. These tables will immediately become available for querying by any of the Azure Synapse workspace Spark pools. They can also be used from any of the Spark jobs subject to permissions.
Note: For external tables, since they are synchronized to serverless SQL pool asynchronously, there will be a delay until they appear.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/metadata/table

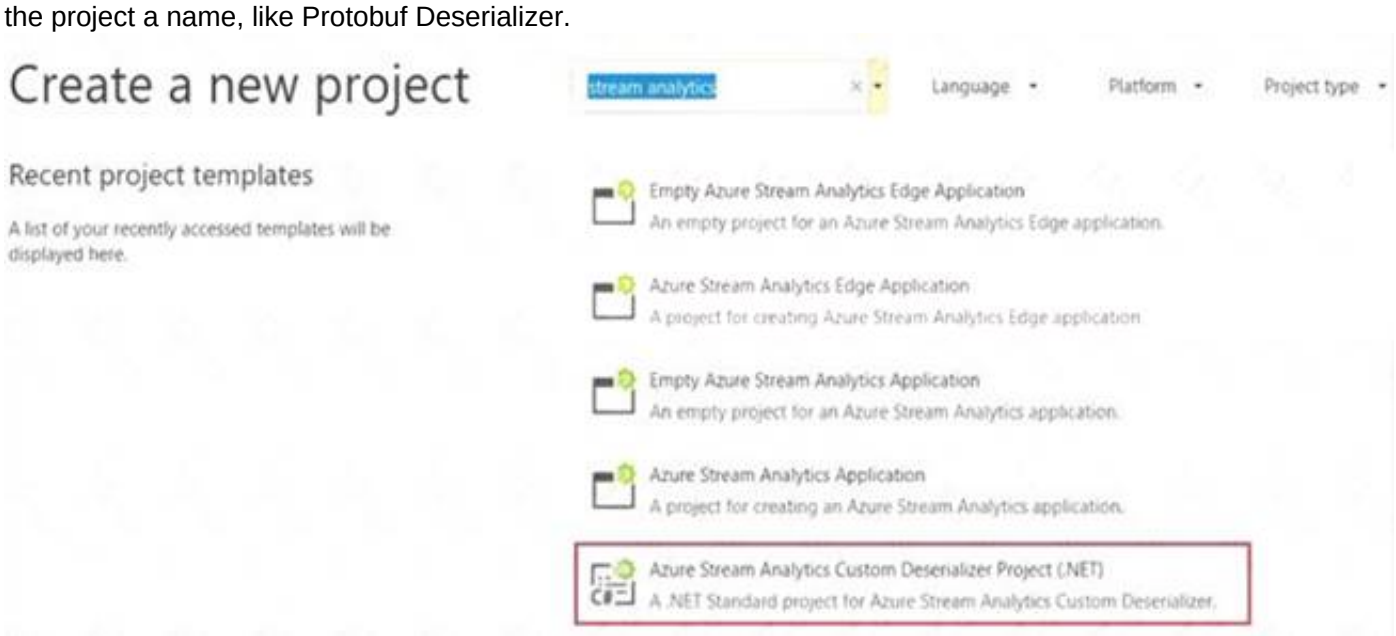
NEW QUESTION 230
- (Exam Topic 3)
You have an Azure Stream Analytics job that is a Stream Analytics project solution in Microsoft Visual Studio. The job accepts data generated by IoT devices in the JSON format.
You need to modify the job to accept data generated by the IoT devices in the Protobuf format.
Which three actions should you perform from Visual Studio on sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

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

A. Mastered
B. Not Mastered

Answer: A

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



* 2. In Solution Explorer, right-click your Protobuf Deserializer project and select Manage NuGet Packages from the menu. Then install the Microsoft.Azure.StreamAnalytics and Google.Protobuf NuGet packages.
* 3. Add the MessageBodyProto class and the MessageBodyDeserializer class to your project.
* 4. Build the Protobuf Deserializer project.
Step 2: Add .NET deserializer code for Protobuf to the custom deserializer project
Azure Stream Analytics has built-in support for three data formats: JSON, CSV, and Avro. With custom .NET deserializers, you can read data from other formats such as Protocol Buffer, Bond and other user defined formats for both cloud and edge jobs.
Step 3: Add an Azure Stream Analytics Application project to the solution Add an Azure Stream Analytics project
In Solution Explorer, right-click the Protobuf Deserializer solution and select Add > New Project. Under Azure Stream Analytics > Stream Analytics, choose Azure Stream Analytics Application. Name it ProtobufCloudDeserializer and select OK.
Right-click References under the ProtobufCloudDeserializer Azure Stream Analytics project. Under Projects, add Protobuf Deserializer. It should be automatically populated for you.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/custom-deserializer

NEW QUESTION 231
- (Exam Topic 3)
You have an Azure Data Factory that contains 10 pipelines.
You need to label each pipeline with its main purpose of either ingest, transform, or load. The labels must be available for grouping and filtering when using the monitoring experience in Data Factory.
What should you add to each pipeline?

A. a resource tag
B. a correlation ID
C. a run group ID
D. an annotation

Answer: D

Explanation:
Annotations are additional, informative tags that you can add to specific factory resources: pipelines, datasets, linked services, and triggers. By adding annotations, you can easily filter and search for specific factory resources.
Reference:
https://www.cathrinewilhelmsen.net/annotations-user-properties-azure-data-factory/

NEW QUESTION 233
- (Exam Topic 3)
You plan to implement an Azure Data Lake Storage Gen2 container that will contain CSV files. The size of the files will vary based on the number of events that occur per hour.
File sizes range from 4.KB to 5 GB.
You need to ensure that the files stored in the container are optimized for batch processing. What should you do?

A. Compress the files.
B. Merge the files.
C. Convert the files to JSON
D. Convert the files to Avro.

Answer: D

Explanation:
Avro supports batch and is very relevant for streaming.
Note: Avro is framework developed within Apache's Hadoop project. It is a row-based storage format which is widely used as a serialization process. AVRO stores its schema in JSON format making it easy to read and interpret by any program. The data itself is stored in binary format by doing it compact and efficient.
Reference:
https://www.adaltas.com/en/2020/07/23/benchmark-study-of-different-file-format/

NEW QUESTION 234
- (Exam Topic 3)
You are building an Azure Synapse Analytics dedicated SQL pool that will contain a fact table for transactions from the first half of the year 2020.
You need to ensure that the table meets the following requirements:
- Minimizes the processing time to delete data that is older than 10 years
- Minimizes the I/O for queries that use year-to-date values

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

Answer Area

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Table Description automatically generated
Box 1: PARTITION
RANGE RIGHT FOR VALUES is used with PARTITION.
Part 2: [TransactionDateID] Partition on the date column.
Example: Creating a RANGE RIGHT partition function on a datetime column
The following partition function partitions a table or index into 12 partitions, one for each month of a year's worth of values in a datetime column.
CREATE PARTITION FUNCTION [myDateRangePF1] (datetime)
AS RANGE RIGHT FOR VALUES ('20030201', '20030301', '20030401',
'20030501', '20030601', '20030701', '20030801',
'20030901', '20031001', '20031101', '20031201');
Reference:

https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql

NEW QUESTION 235
- (Exam Topic 3)
You have an Azure Synapse Analytics workspace named WS1.
You have an Azure Data Lake Storage Gen2 container that contains JSON-formatted files in the following format.

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

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

Values

opendatasource
openjson
openquery
openrowset

Answer Area

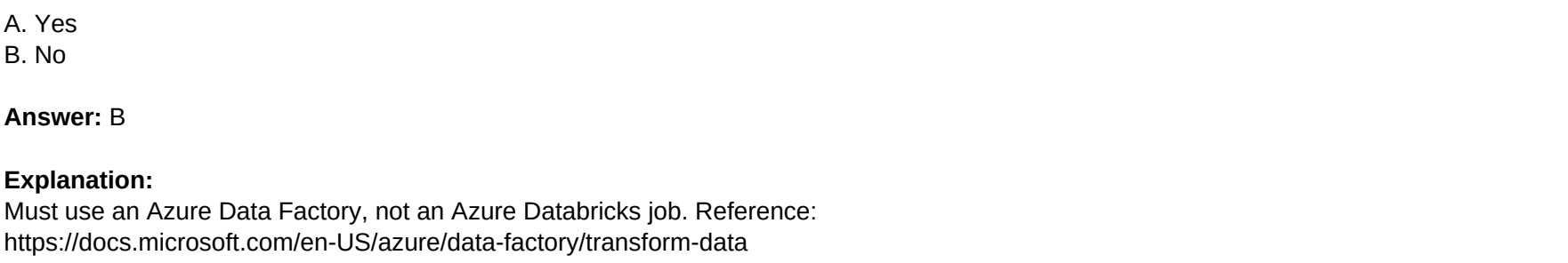
```
select*

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

) as q
cross apply [                ] (contextcustomdimensions)

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

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 240
- (Exam Topic 3)
You need to implement an Azure Databricks cluster that automatically connects to Azure Data lake Storage Gen2 by using Azure Active Directory (Azure AD) integration. How should you configure the new clutter? To answer, select the appropriate options in the answers area. NOTE: Each correct selection is worth one point.



A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 243
- (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 244
- (Exam Topic 3)
You have an Azure Data Lake Storage account that has a virtual network service endpoint configured.
You plan to use Azure Data Factory to extract data from the Data Lake Storage account. The data will then be loaded to a data warehouse in Azure Synapse Analytics by using PolyBase.
Which authentication method should you use to access Data Lake Storage?

A. shared access key authentication
B. managed identity authentication
C. account key authentication
D. service principal authentication

Answer: B

Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse#use-polybase-to-load-d

NEW QUESTION 245
- (Exam Topic 3)
You have an Azure Data Lake Storage Gen2 account named adls2 that is protected by a virtual network. You are designing a SQL pool in Azure Synapse that will use adls2 as a source.
What should you use to authenticate to adls2?

A. a shared access signature (SAS)
B. a managed identity
C. a shared key
D. an Azure Active Directory (Azure AD) user

Answer: B

Explanation:
Managed identity for Azure resources is a feature of Azure Active Directory. The feature provides Azure services with an automatically managed identity in Azure AD. You can use the Managed Identity capability to authenticate to any service that support Azure AD authentication.
Managed Identity authentication is required when your storage account is attached to a VNet. Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-exa

NEW QUESTION 250
- (Exam Topic 3)
You are building a data flow in Azure Data Factory that upserts data into a table in an Azure Synapse Analytics dedicated SQL pool.
You need to add a transformation to the data flow. The transformation must specify logic indicating when a row from the input data must be upserted into the sink.
Which type of transformation should you add to the data flow?

A. join
B. select
C. surrogate key
D. alter row

Answer: D

Explanation:
The alter row transformation allows you to specify insert, update, delete, and upsert policies on rows based on expressions. You can use the alter row transformation to perform upserts on a sink table by matching on a key column and setting the appropriate row policy

NEW QUESTION 252
- (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.

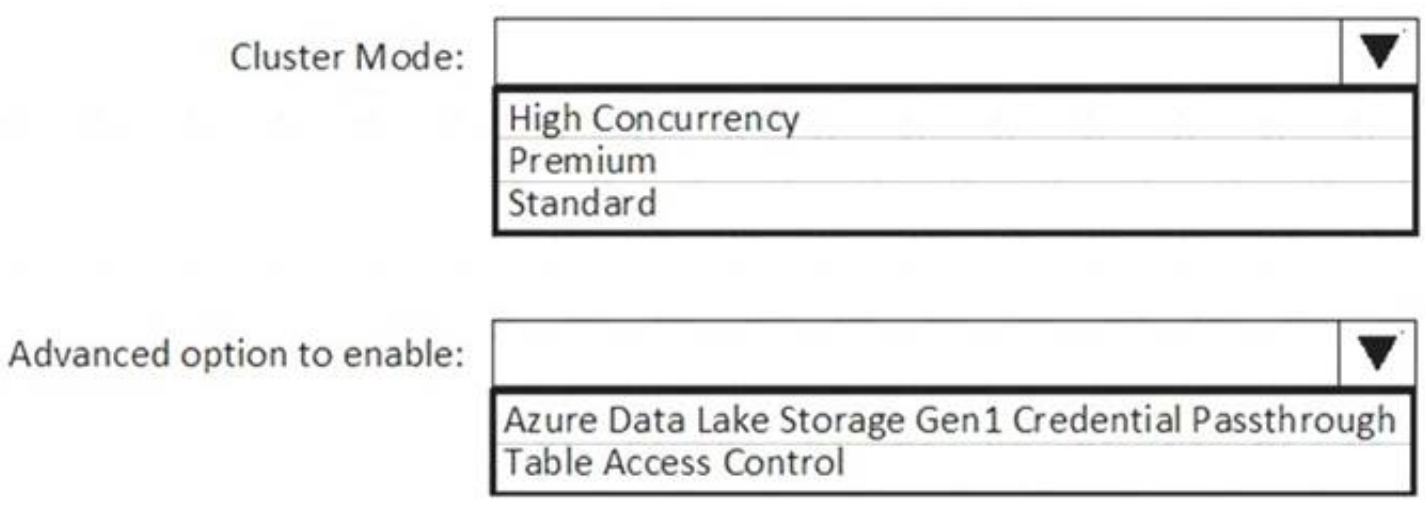
NEW QUESTION 254
- (Exam Topic 3)
You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1.
You need to use an Azure Synapse pipeline 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 255
- (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 256
- (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 261

- (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

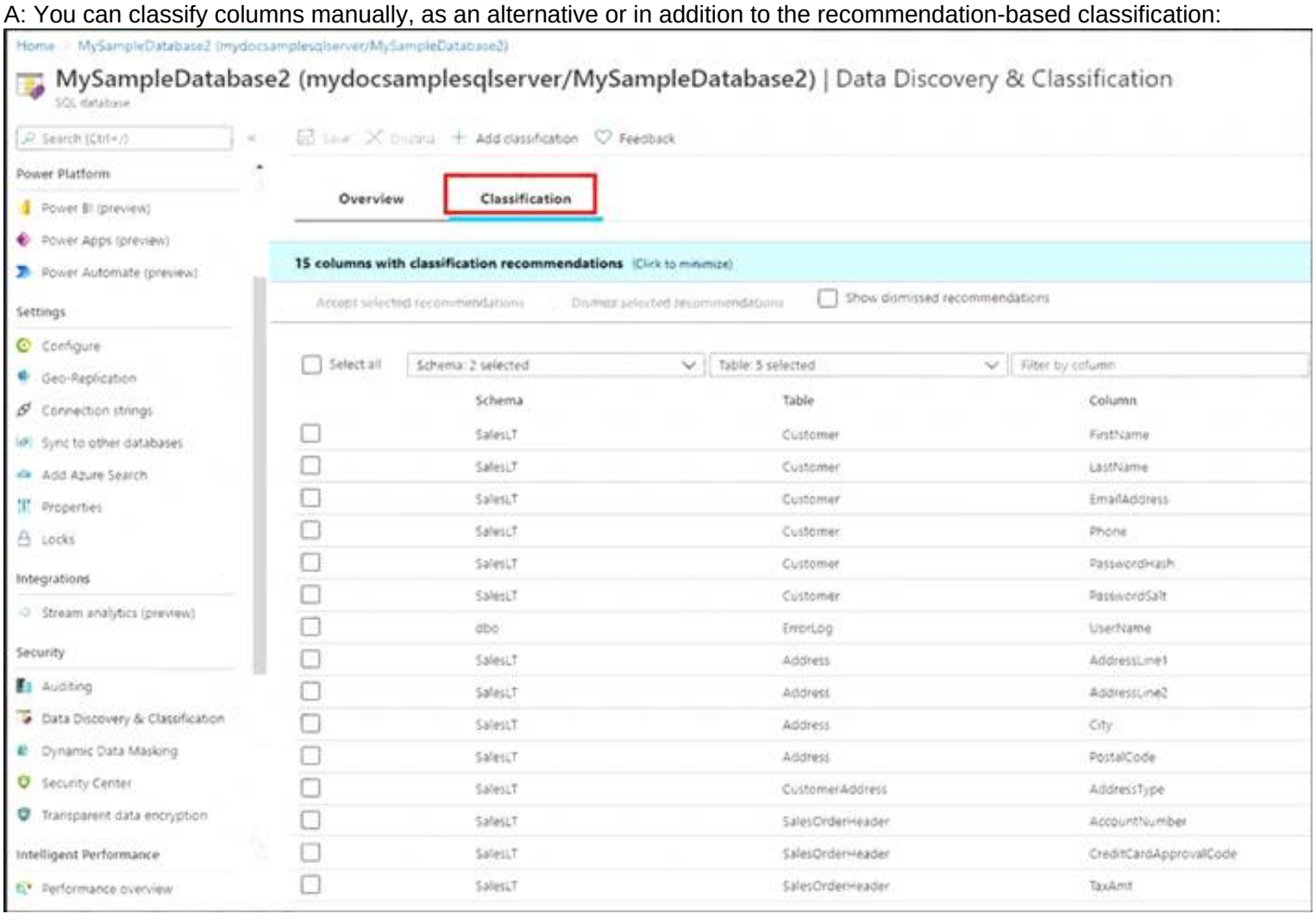
Answer: B

NEW QUESTION 262
- (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

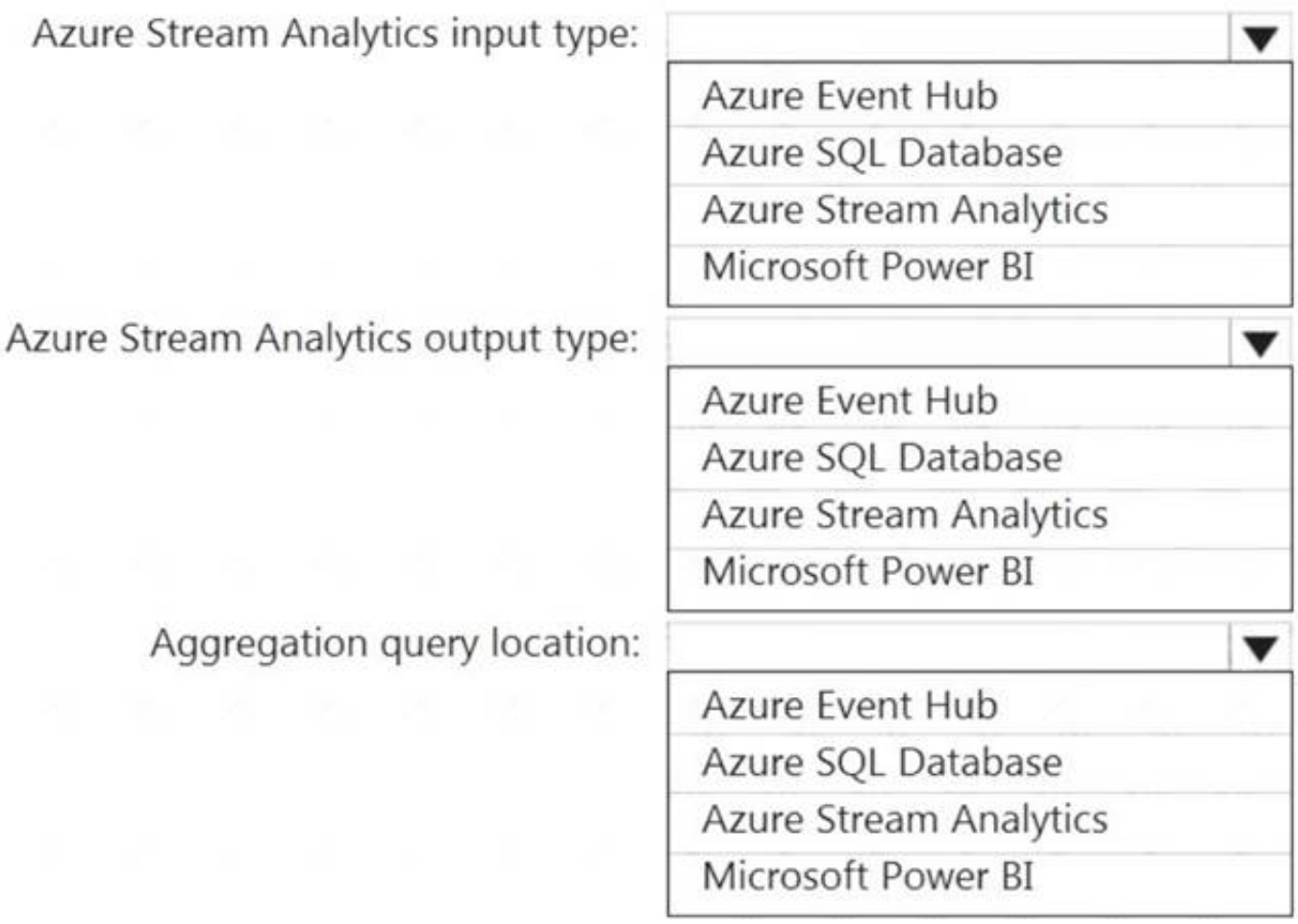
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview

NEW QUESTION 266

- (Exam Topic 3)
You are designing a real-time dashboard solution that will visualize streaming data from remote sensors that connect to the internet. The streaming data must be aggregated to show the average value of each 10-second interval. The data will be discarded after being displayed in the dashboard.
The solution will use Azure Stream Analytics and must meet the following requirements:

- Minimize latency from an Azure Event hub to the dashboard.
- Minimize the required storage.
- Minimize development effort.

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



A. Mastered
B. Not Mastered

Answer: A

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

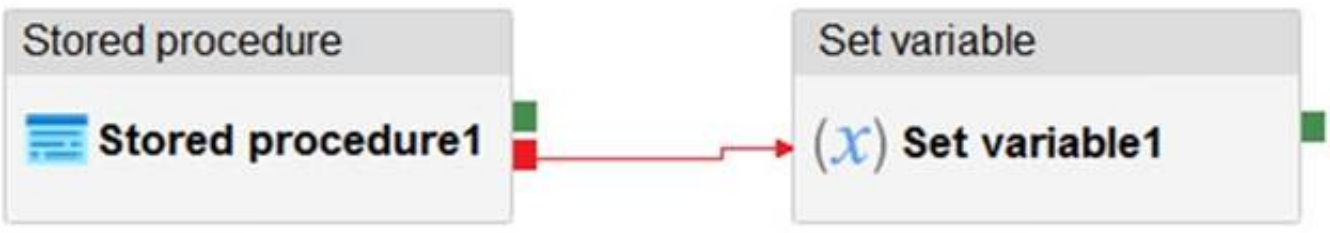
NEW QUESTION 269
- (Exam Topic 3)
You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account.
You need to output the count of tweets during the last five minutes every five minutes. Each tweet must only be counted once.
Which windowing function should you use?

A. a five-minute Session window
B. a five-minute Sliding window
C. a five-minute Tumbling window
D. a five-minute Hopping window that has one-minute hop

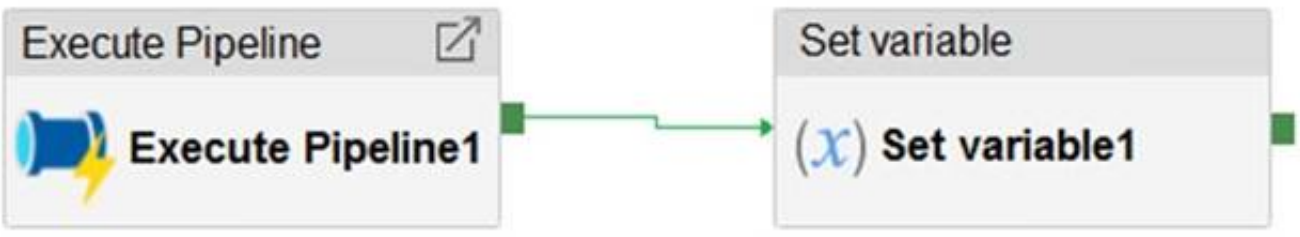
Answer: C

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

NEW QUESTION 271
- (Exam Topic 3)
You have an Azure Data Factory instance that contains two pipelines named Pipeline1 and Pipeline2. Pipeline1 has the activities shown in the following exhibit.



Pipeline2 has the activities shown in the following exhibit.

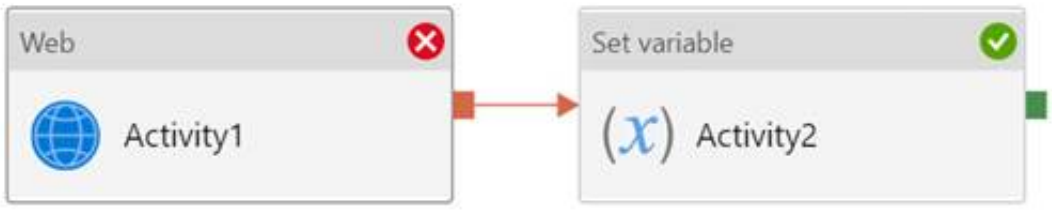


You execute Pipeline2, and Stored procedure1 in Pipeline1 fails. What is the status of the pipeline runs?

A. Pipeline1 and Pipeline2 succeeded.
B. Pipeline1 and Pipeline2 failed.
C. Pipeline1 succeeded and Pipeline2 failed.
D. Pipeline1 failed and Pipeline2 succeeded.

Answer: A

Explanation:
Activities are linked together via dependencies. A dependency has a condition of one of the following: Succeeded, Failed, Skipped, or Completed.
Consider Pipeline1:
If we have a pipeline with two activities where Activity2 has a failure dependency on Activity1, the pipeline
will not fail just because Activity1 failed. If Activity1 fails and Activity2 succeeds, the pipeline will succeed. This scenario is treated as a try-catch block by Data Factory.
Waterfall chart Description automatically generated with medium confidence



The failure dependency means this pipeline reports success. Note:
If we have a pipeline containing Activity1 and Activity2, and Activity2 has a success dependency on Activity1, it will only execute if Activity1 is successful. In this scenario, if Activity1 fails, the pipeline will fail.
Reference:
https://datasavvy.me/category/azure-data-factory/

NEW QUESTION 272
.....

THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual DP-203 Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the DP-203 Product From:

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

Money Back Guarantee

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