Microsoft

Exam Questions DP-203

Data Engineering on Microsoft Azure

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

Commands

CREATE EXTERNAL DATA SOURCE
CREATE EXTERNAL FILE FORMAT
CREATE EXTERNAL TABLE
CREATE EXTERNAL TABLE AS SELECT
CREATE DATABASE SCOPED CREDENTIAL

Answer Area

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 2
- (Exam Topic 1)
You need to implement an Azure Synapse Analytics database object for storing the sales transactions data. The solution must meet the sales transaction dataset requirements.
What solution must meet the sales transaction dataset requirements.
What should you do? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Answer Area

Transact-SQL DDL command to use:	CREATE EXTERNAL TABLE
	CREATE TABLE
	CREATE VIEW

Partitioning option to use in the WITH clause of the DDL statement:	FORMAT_OPTIONS
	FORMAT_TYPE
	RANGE LEFT FOR VALUES
	RANGE RIGHT FOR VALUES

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

Transact-SQL DDL command to use:	CREATE EXTERNAL TABLE
	CREATE TABLE
	CREATE VIEW

Partitioning option to use in the WITH clause of the DDL statement:	FORMAT_OPTIONS
	FORMAT_TYPE
	RANGE LEFT FOR VALUES
	RANGE RIGHT FOR VALUES

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

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

Answer: CE

NEW QUESTION 4
- (Exam Topic 3)
You have files and folders in Azure Data Lake Storage Gen2 for an Azure Synapse workspace as shown in the following exhibit.
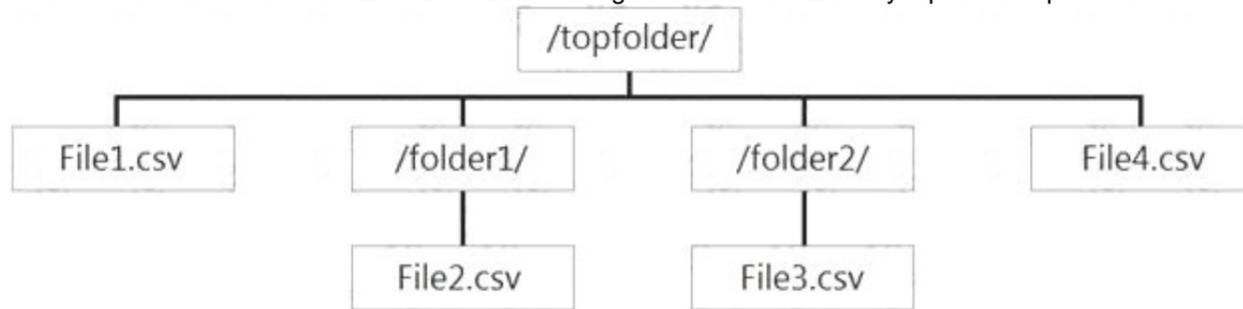
You create an external table named ExtTable that has LOCATION='/topfolder/'.
When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, which files are returned?

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

Answer: C

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

NEW QUESTION 5
- (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 6
- (Exam Topic 3)
You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools. Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company.
You need to move the files to a different folder and transform the data to meet the following requirements: ➢ Provide the fastest possible query times.

➢ Automatically infer the schema from the underlying files.
How should you configure the Data Factory copy activity? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Copy behavior:

Flatten hierarchy
Merge files
Preserve hierarchy

Sink file type:

CSV
JSON
Parquet
TXT

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 7
- (Exam Topic 3)
You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container.
Which type of trigger should you use?

A. on-demand
B. tumbling window
C. schedule
D. event

Answer: D

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.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger

NEW QUESTION 8
- (Exam Topic 3)
What should you recommend using to secure sensitive customer contact information?

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

Answer: B

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

NEW QUESTION 9
- (Exam Topic 3)
You have a Microsoft SQL Server database that uses a third normal form schema.
You plan to migrate the data in the database to a star schema in an A?\ire Synapse Analytics dedicated SQl pool.
You need to design the dimension tables. The solution must optimize read operations.
What should you include in the solution? to answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Answer Area

Transform data for the dimension tables by: ▼

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

Transform data for the dimension tables by: ▼

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

NEW QUESTION 10
- (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

NEW QUESTION 10
- (Exam Topic 3)
You are designing a real-time dashboard solution that will visualize streaming data from remote sensors that connect to the internet. The streaming data must be aggregated to show the average value of each 10-second interval. The data will be discarded after being displayed in the dashboard.
The solution will use Azure Stream Analytics and must meet the following requirements:
≫ Minimize latency from an Azure Event hub to the dashboard.
≫ Minimize the required storage.
≫ Minimize development effort.
What should you include in the solution? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point

Azure Stream Analytics input type:	▼
	Azure Event Hub
	Azure SQL Database
	Azure Stream Analytics
	Microsoft Power BI

Azure Stream Analytics output type:	▼
	Azure Event Hub
	Azure SQL Database
	Azure Stream Analytics
	Microsoft Power BI

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

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 14
- (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.

HubA: ▼

| Stream |
| Reference |

HubB: ▼

| Stream |
| Reference |

Database1: ▼

| Stream |
| Reference |

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

NEW QUESTION 15
- (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.
Does this meet the goal?

A. Yes
B. No

Answer: A

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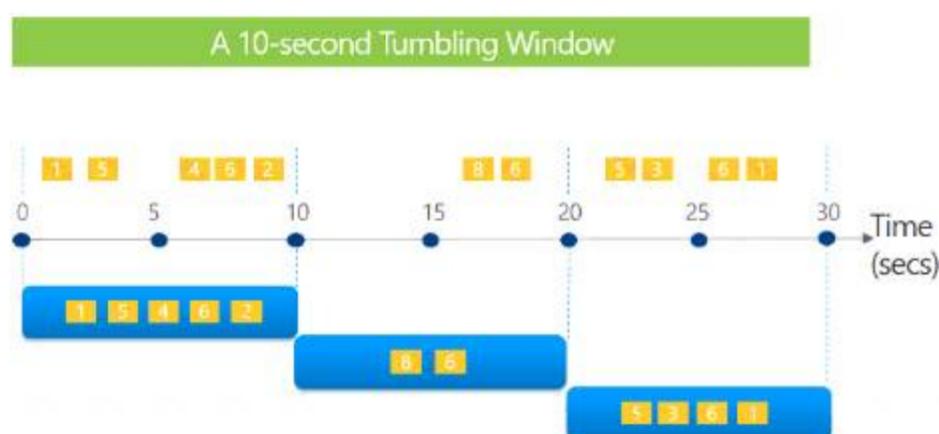

```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

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

NEW QUESTION 16
- (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. NOTE: Each appropriate options in the answer area.

Answer Area

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

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

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

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

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

NEW QUESTION 17
- (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: A

Explanation:
Geo-redundant storage (GRS) copies your data synchronously three times within a single physical location in the primary region using LRS. It then copies your data asynchronously to a single physical location in the secondary region.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy

NEW QUESTION 22
- (Exam Topic 3)
You have an Azure subscription that contains a logical Microsoft SQL server named Server1. Server1 hosts an Azure Synapse Analytics SQL dedicated pool named Pool1.
You need to recommend a Transparent Data Encryption (TDE) solution for Server1. The solution must meet the following requirements:
 Track the usage of encryption keys.
 Maintain the access of client apps to Pool1 in the event of an Azure datacenter outage that affects the availability of the encryption keys.

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

To track encryption key usage: ▼
Always Encrypted
TDE with customer-managed keys
TDE with platform-managed keys

To maintain client app access in
the event of a datacenter outage: ▼
Create and configure Azure key vaults in two Azure regions.
Enable Advanced Data Security on Server1.
Implement the client apps by using a Microsoft .NET Framework data provider.

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 26
- (Exam Topic 3)
You are designing an application that will store petabytes of medical imaging data
When the data is first created, the data will be accessed frequently during the first week. After one month, the data must be accessible within 30 seconds, but files will be accessed infrequently. After one year, the data will be accessed infrequently but must be accessible within five minutes.
You need to select a storage strategy for the data. The solution must minimize costs.
Which storage tier should you use for each time frame? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

First week: ▼
Archive
Cool
Hot

After one month: ▼
Archive
Cool
Hot

After one year: ▼
Archive
Cool
Hot

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 27
- (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 queries.
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:
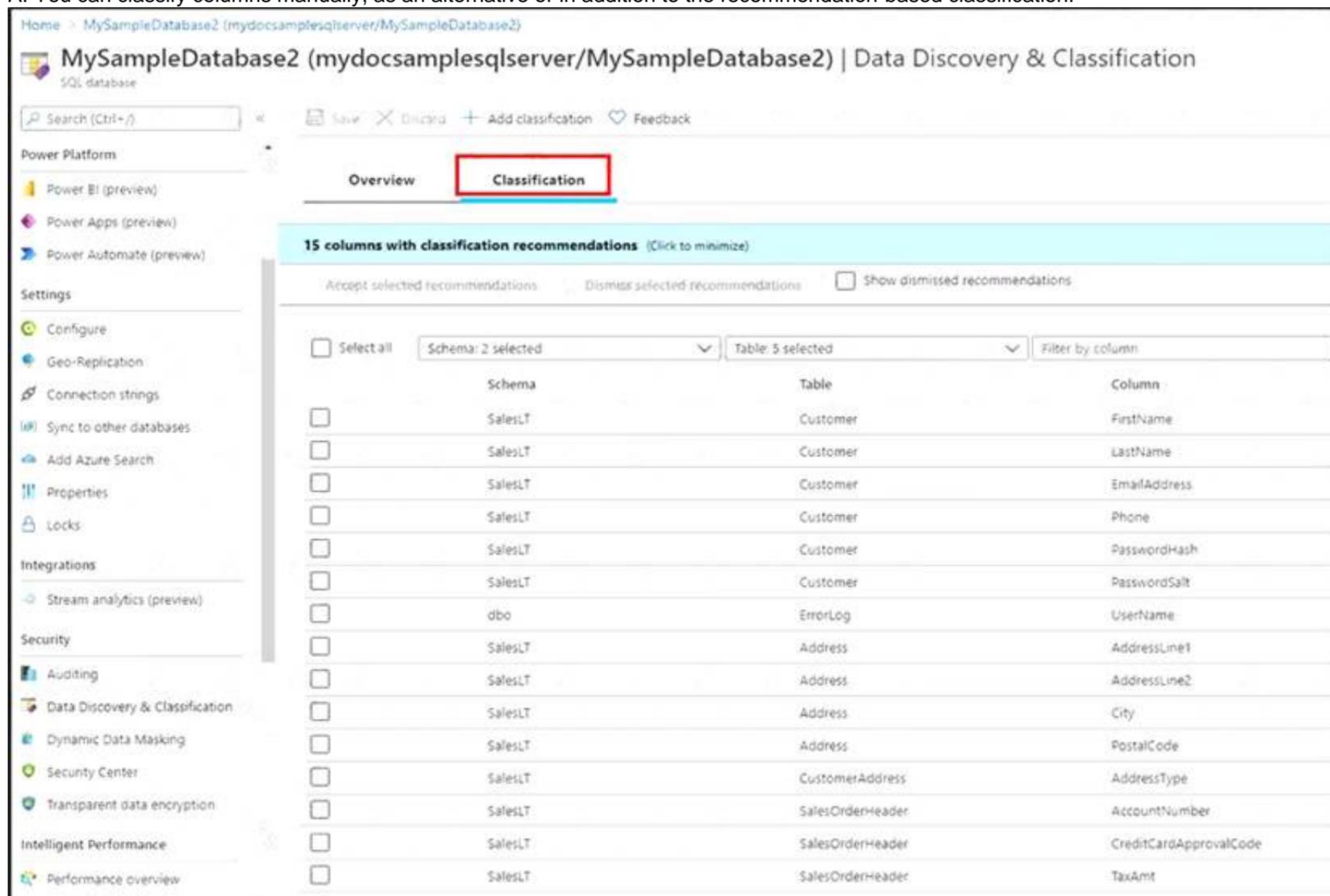
> Select Add classification in the top menu of the pane.

> 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:

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

NEW QUESTION 30
- (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: B

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 31
- (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.

Cluster Mode:

| High Concurrency |
| Premium |
| Standard |

Advanced option to enable:

| Azure Data Lake Storage Gen1 Credential Passthrough |
| Table Access Control |

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 32
- (Exam Topic 3)
You implement an enterprise data warehouse in Azure Synapse Analytics. You have a large fact table that is 10 terabytes (TB) in size.
Incoming queries use the primary key SaleKey column to retrieve data as displayed in the following table:

SaleKey	CityKey	CustomerKey	StockItemKey	InvoiceDateKey	Quantity	UnitPrice	TotalExcludingTax
49309	90858	70	69	10/22/13	8	16	128
49313	55710	126	69	10/22/13	2	16	32
49343	44710	234	68	10/22/13	10	16	160
49352	66109	163	70	10/22/13	4	16	64
49488	65312	230	70	10/22/13	8	16	128
49646	85877	271	70	10/24/13	1	16	16
49798	41238	288	69	10/24/13	1	16	16

You need to distribute the large fact table across multiple nodes to optimize performance of the table. Which technology should you use?

A. hash distributed table with clustered index
B. hash distributed table with clustered Columnstore index
C. round robin distributed table with clustered index
D. round robin distributed table with clustered Columnstore index
E. heap table with distribution replicate

Answer: B

Explanation:
Hash-distributed tables improve query performance on large fact tables.
Columnstore indexes can achieve up to 100x better performance on analytics and data warehousing workloads and up to 10x better data compression than traditional rowstore indexes.
Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-query-performance

NEW QUESTION 37
- (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

| 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. |

Answer Area

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.
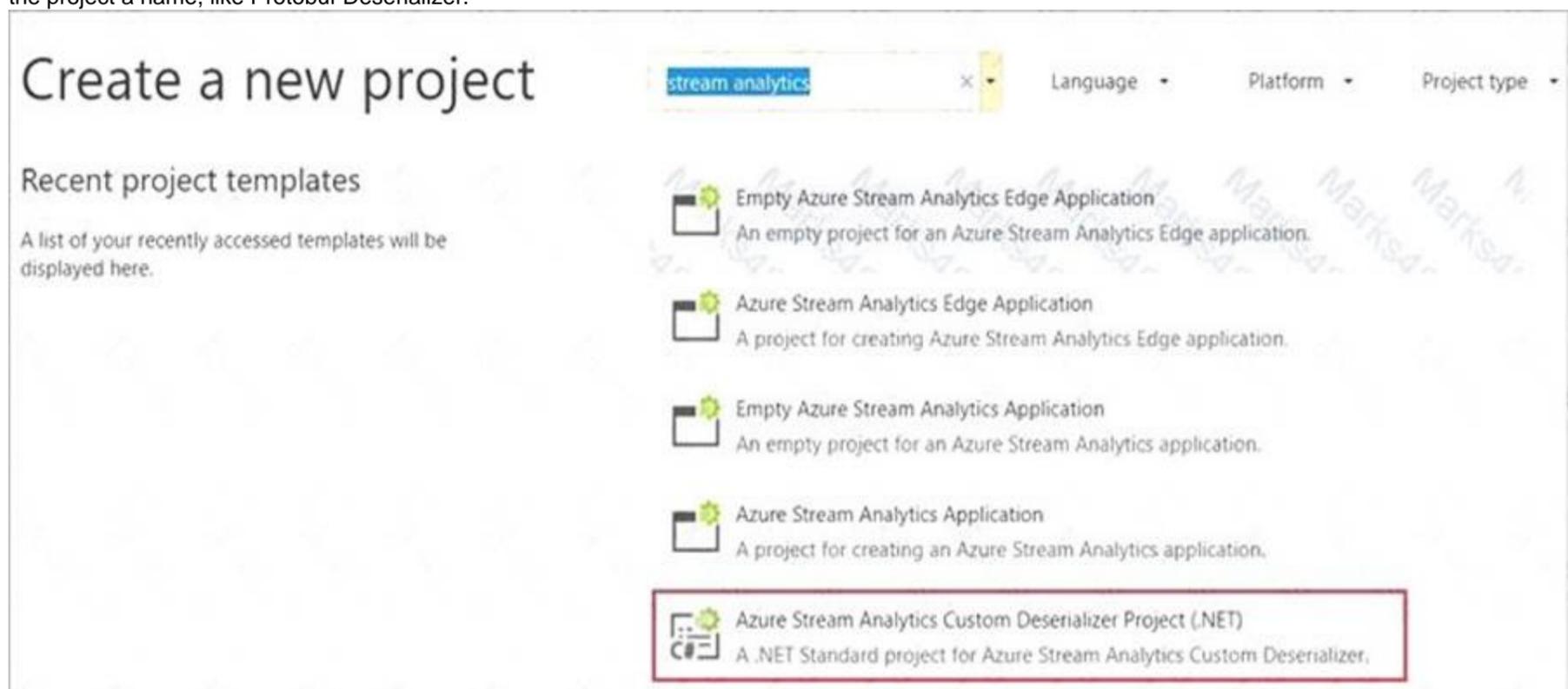
* 2. In Solution Explorer, right-click your Protobuf Deserializer project and select Manage NuGet Packages from the menu. Then install the Microsoft.Azure.StreamAnalytics and Google.Protobuf NuGet packages.
* 3. Add the MessageBodyProto class and the MessageBodyDeserializer class to your project.

* 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
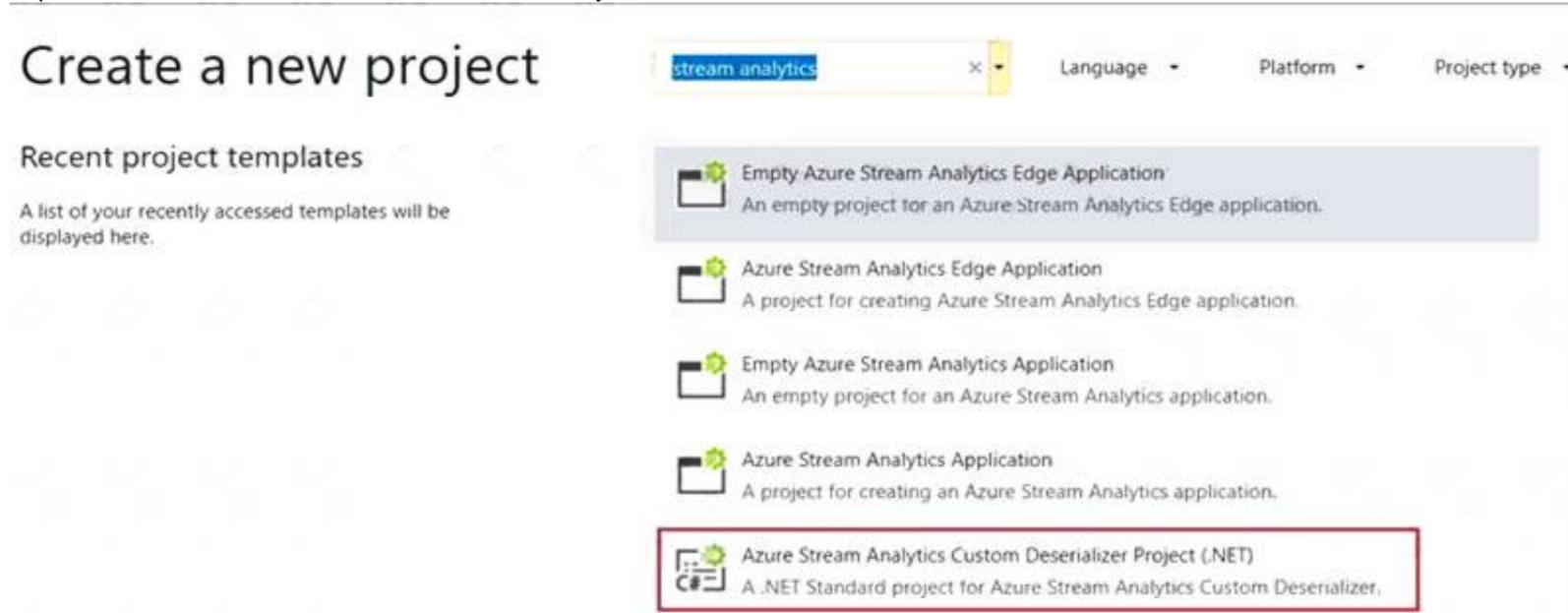
NEW QUESTION 42
- (Exam Topic 3)
You have an Azure data factory.
You need to ensure that pipeline-run data is retained for 120 days. The solution must ensure that you can query the data by using the Kusto query language. Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Step 1: Create an Azure Storage account that has a lifecycle policy
To automate common data management tasks, Microsoft created a solution based on Azure Data Factory. The service, Data Lifecycle Management, makes frequently accessed data available and archives or purges other data according to retention policies. Teams across the company use the service to reduce storage costs, improve app performance, and comply with data retention policies.
Step 2: Create a Log Analytics workspace that has Data Retention set to 120 days.
Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time. With Monitor, you can route diagnostic logs for analysis to multiple different targets, such as a Storage Account: Save your diagnostic logs to a storage account for auditing or manual inspection. You can

use the diagnostic settings to specify the retention time in days.
Step 3: From Azure Portal, add a diagnostic setting. Step 4: Send the data to a log Analytics workspace,
Event Hub: A pipeline that transfers events from services to Azure Data Explorer. Keeping Azure Data Factory metrics and pipeline-run data.
Configure diagnostic settings and workspace.
Create or add diagnostic settings for your data factory.

- ❯ In the portal, go to Monitor. Select Settings > Diagnostic settings.

- ❯ Select the data factory for which you want to set a diagnostic setting.

- ❯ If no settings exist on the selected data factory, you're prompted to create a setting. Select Turn on diagnostics.

- ❯ Give your setting a name, select Send to Log Analytics, and then select a workspace from Log Analytics Workspace.

- ❯ Select Save. Reference:
https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor

NEW QUESTION 43
- (Exam Topic 3)
You are designing an inventory updates table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:
• EventDate: 1 million per day
• EventTypeID: 10 million per event type
• WarehouseID: 100 million per warehouse
• ProductCategoryTypeiD: 25 million per product category type You identify the following usage patterns:
Analyst will most commonly analyze transactions for a warehouse.
Queries will summarize by product category type, date, and/or inventory event type. You need to recommend a partition strategy for the table to minimize query times. On which column should you recommend partitioning the table?

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

Answer: D

NEW QUESTION 44
- (Exam Topic 3)
You are creating dimensions for a data warehouse in an Azure Synapse Analytics dedicated SQL pool. You create a table by using the Transact-SQL statement shown in the following exhibit.

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

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

DimProduct is a **[answer choice]** slowly changing dimension (SCD).

▼
Type 0
Type 1
Type 2

The ProductKey column is **[answer choice]**.

▼
a surrogate key
a business key
an audit column

A. Mastered
B. Not Mastered

Answer: A

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

NEW QUESTION 48
- (Exam Topic 3)
You are designing a fact table named FactPurchase in an Azure Synapse Analytics dedicated SQL pool. The table contains purchases from suppliers for a retail store. FactPurchase will contain the following columns.

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

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

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

Answer: D

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

NEW QUESTION 49
- (Exam Topic 3)
You are planning a streaming data solution that will use Azure Databricks. The solution will stream sales transaction data from an online store. The solution has the following specifications:
* The output data will contain items purchased, quantity, line total sales amount, and line total tax amount.
* Line total sales amount and line total tax amount will be aggregated in Databricks.
* Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.
You need to recommend an output mode for the dataset that will be processed by using Structured Streaming. The solution must minimize duplicate data.
What should you recommend?

A. Append
B. Update
C. Complete

Answer: C

NEW QUESTION 50
- (Exam Topic 3)
You need to output files from Azure Data Factory.
Which file format should you use for each type of output? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Columnar format:
- Avro
- GZip
- Parquet
- TXT

JSON with a timestamp:
- Avro
- GZip
- Parquet
- TXT

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: Parquet
Parquet stores data in columns, while Avro stores data in a row-based format. By their very nature,
column-oriented data stores are optimized for read-heavy analytical workloads, while row-based databases are best for write-heavy transactional workloads.
Box 2: Avro
An Avro schema is created using JSON format.
AVRO supports timestamps.
Note: Azure Data Factory supports the following file formats (not GZip or TXT).
≫ Avro format
≫ Binary format
≫ Delimited text format
≫ Excel format
≫ JSON format
≫ ORC format
≫ Parquet format
≫ XML format
Reference:
https://www.datanami.com/2018/05/16/big-data-file-formats-demystified

NEW QUESTION 51
- (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.

Values

CASE

ELSE

OVER

PARTITION
BY

ROW_NUMBER

Answer Area

SELECT
*,
    Value
        WHEN hire_date >= '2019-01-01' THEN
    'New'   Value   'Standard'
        END AS employee_type
FROM
    employees;

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Values

CASE

ELSE

OVER

PARTITION
BY

ROW_NUMBER

Answer Area

SELECT
*,
    CASE
        WHEN hire_date >= '2019-01-01' THEN
    'New'   PARTITION   'Standard'
        BY
        END AS employee_type
FROM
    employees;

NEW QUESTION 55
- (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 59
- (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 64
- (Exam Topic 3)
You have an Azure Stream Analytics query. The query returns a result set that contains 10,000 distinct values for a column named clusterID.
You monitor the Stream Analytics job and discover high latency. You need to reduce the latency.
Which two actions should you perform? Each correct answer presents a complete solution. NOTE: Each correct selection is worth one point.

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

Answer: CE

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

NEW QUESTION 68
- (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 71
- (Exam Topic 3)
You have an Azure Stream Analytics job that receives clickstream data from an Azure event hub.
You need to define a query in the Stream Analytics job. The query must meet the following requirements: > Count the number of clicks within each 10-second window based on the country of a visitor.

> Ensure that each click is NOT counted more than once. How should you define the Query?

A. SELECT Country, Avg(*) AS AverageFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SlidingWindow(second, 10)
B. SELECT Country, Count(*) AS CountFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, TumblingWindow(second, 10)
C. SELECT Country, Avg(*) AS AverageFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, HoppingWindow(second, 10, 2)
D. SELECT Country, Count(*) AS CountFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SessionWindow(second, 5, 10)

Answer: B

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

NEW QUESTION 75
- (Exam Topic 3)
You develop data engineering solutions for a company.
A project requires the deployment of data to Azure Data Lake Storage.
You need to implement role-based access control (RBAC) so that project members can manage the Azure Data Lake Storage resources.
Which three actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

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

Answer: ADE

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

NEW QUESTION 79
- (Exam Topic 3)
You have a data model that you plan to implement in a data warehouse in Azure Synapse Analytics as shown in the following exhibit.
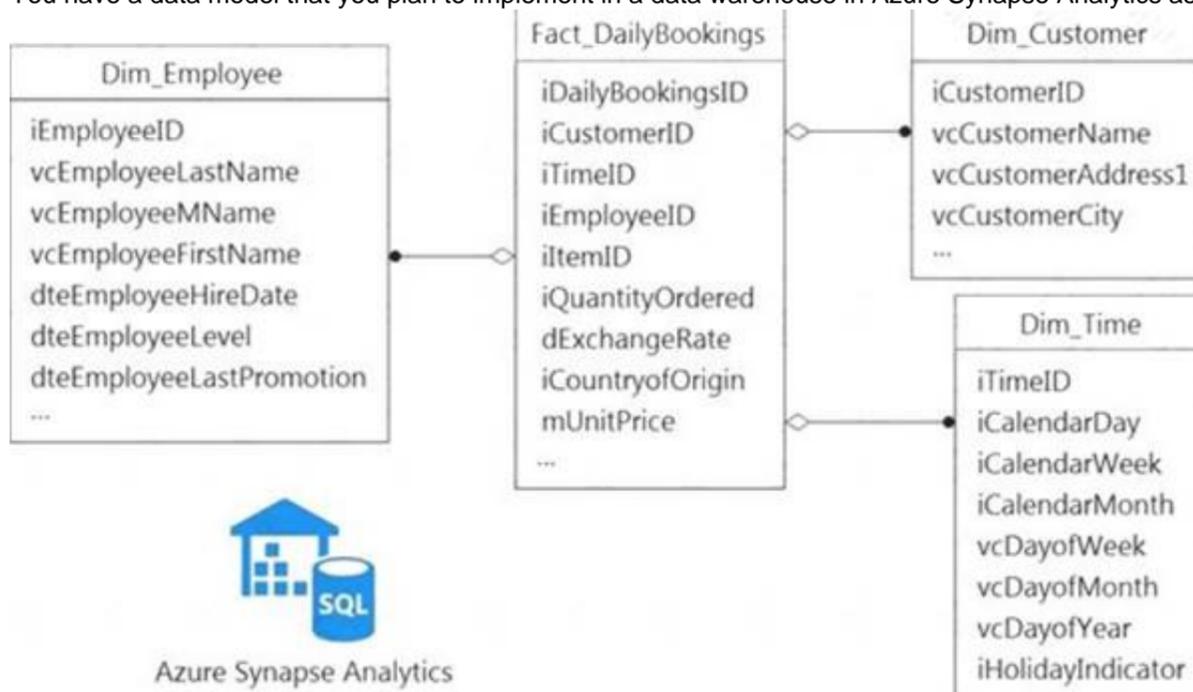
All the dimension tables will be less than 2 GB after compression, and the fact table will be approximately 6 TB.
Which type of table should you use for each table? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area

Dim_Customer: ▼

| Hash distributed |
| Round-robin |
| Replicated |

Dim_Employee: ▼

| Hash distributed |
| Round-robin |
| Replicated |

Dim_Time: ▼

| Hash distributed |
| Round-robin |
| Replicated |

Fact_DailyBookings: ▼

| Hash distributed |
| Round-robin |
| Replicated |

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

Dim_Customer: ▼

| Hash distributed |
| Round-robin |
| _Replicated_ |

Dim_Employee: ▼

| Hash distributed |
| Round-robin |
| _Replicated_ |

Dim_Time: ▼

| Hash distributed |
| Round-robin |
| _Replicated_ |

Fact_DailyBookings: ▼

| _Hash distributed_ |
| Round-robin |
| Replicated |

NEW QUESTION 84
- (Exam Topic 3)

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

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

Answer: C

NEW QUESTION 89
- (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

NEW QUESTION 90
- (Exam Topic 3)
You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements:

» Can return an employee record from a given point in time.

» Maintains the latest employee information.

» Minimizes query complexity.
How should you model the employee data?

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

Answer: D

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

NEW QUESTION 94
- (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.

Answer Area

```
Select TimeZone, count(*) AS MessageCount
FROM
        ┌─────────────────────┐
        │ LAST                │   CreatedAt
        │ OVER                │
        │ SYSTEM.TIMESTAMP()  │
        │ TIMESTAMP BY        │
        └─────────────────────┘
GROUP BY                                MessageStream
        ┌─────────────────────┐
TimeZone, │ HOPPINGWINDOW     │   (second,15)
        │ SESSIONWINDOW       │
        │ SLIDINGWINDOW       │
        │ TUMBLINGWINDOW      │
        └─────────────────────┘
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

```
Select TimeZone, count(*) AS MessageCount
FROM           ┌──────────────────────┐
               │ LAST                 │  CreatedAt
               │ OVER  _ _ _ _ _ _    │
               │ SYSTEM.TIMESTAMP()   │
               │ TIMESTAMP BY         │
               └──────────────────────┘
MessageStream  ┌──────────────────────┐
GROUP BY       │ HOPPINGWINDOW _ _ _  │  (second,15)
TimeZone,      │ SESSIONWINDOW        │
               │ SLIDINGWINDOW        │
               │ TUMBLINGWINDOW       │
               └──────────────────────┘
```

NEW QUESTION 98
- (Exam Topic 3)
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an Azure Stream Analytics solution that will analyze Twitter data.
You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.
Solution: You use a session window that uses a timeout size of 10 seconds. Does this meet the goal?

A. Yes
B. No

Answer: 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 99
......

Thank You for Trying Our Product

We offer two products:

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

2nd - Questons and Answers in PDF Format

DP-203 Practice Exam Features:

* DP-203 Questions and Answers Updated Frequently

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

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

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

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