DUMPSARENA

Data Engineering on Microsoft Azure

Microsoft DP-203

Version Demo

Total Demo Questions: 20

Total Premium Questions: 439

Buy Premium PDF

https://dumpsarena.co

sales@dumpsarena.co

sales@dumpsarena.co
dumpsarena.co

Topic Break Down

Topic	No. of Questions
Topic 2, New Update	227
Topic 3, Case Study 1	7
Topic 4, Case Study 2	2
Topic 5, Mixed Questions	203
Total	439

QUESTION NO: 1 - (HOTSPOT)

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named Account1.

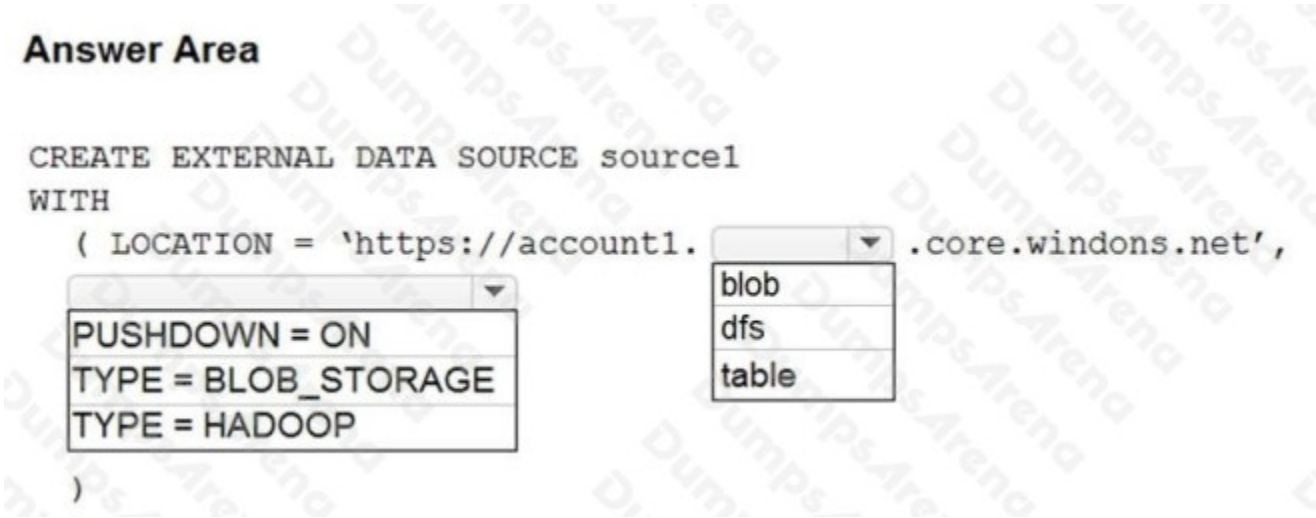
You plan to access the files in Account1 by using an external table.

You need to create a data source in Pool1 that you can reference when you create the external table.

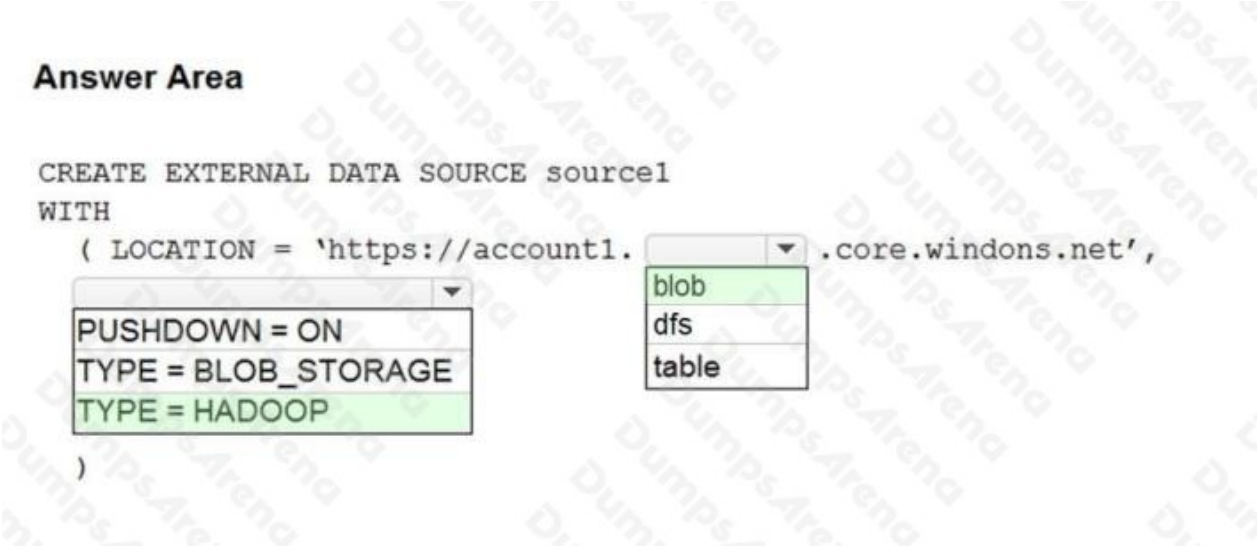
How should you complete the Transact-SQL statement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



ANSWER:



Explanation:

Box 1: blob

The following example creates an external data source for Azure Data Lake Gen2

CREATE EXTERNAL DATA SOURCE YellowTaxi

WITH (LOCATION = 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/', TYPE = HADOOP)

Box 2: HADOOP

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

QUESTION NO: 2

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

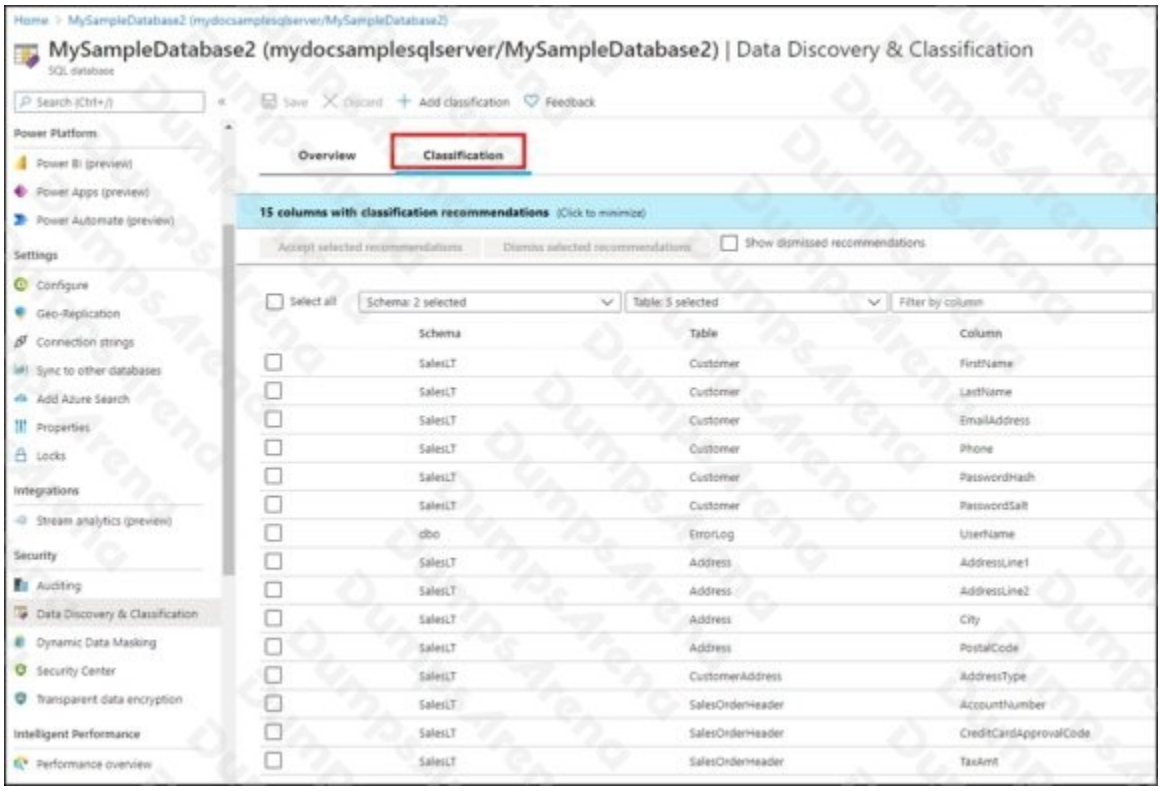
C. audit logs sent to a Log Analytics workspace

D. dynamic data masking for columns that contain confidential information

ANSWER: A C

Explanation:

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



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

QUESTION NO: 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

QUESTION NO: 4

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: B C

QUESTION NO: 5

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: B D

QUESTION NO: 6

You are designing an Azure Synapse solution that will provide a query interface for the data stored in an Azure Storage account. The storage account is only accessible from a virtual network.

You need to recommend an authentication mechanism to ensure that the solution can access the source data.

What should you recommend?

A. a managed identity

B. anonymous public read access

C. a shared key

ANSWER: A

Explanation:

Managed Identity authentication is required when your storage account is attached to a VNet.

Reference:

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-examples

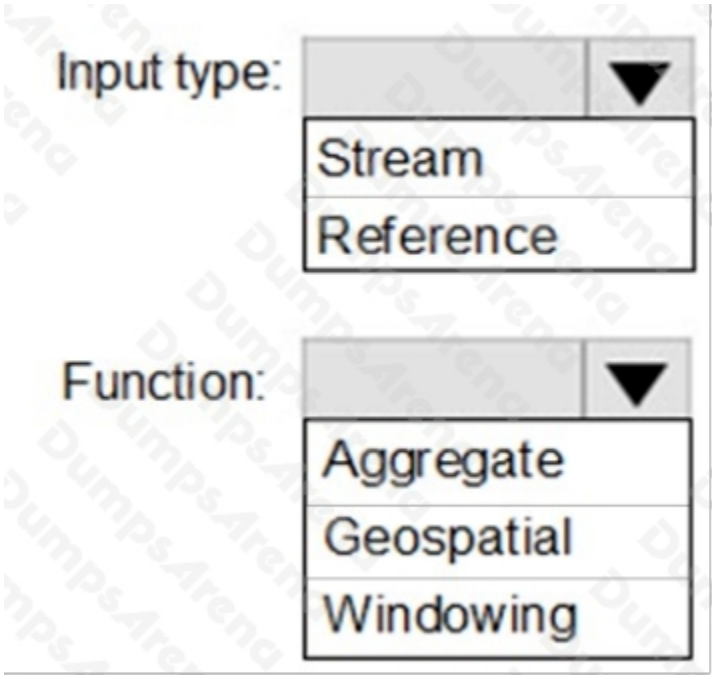
QUESTION NO: 7 - (HOTSPOT)

You plan to create a real-time monitoring app that alerts users when a device travels more than 200 meters away from a designated location.

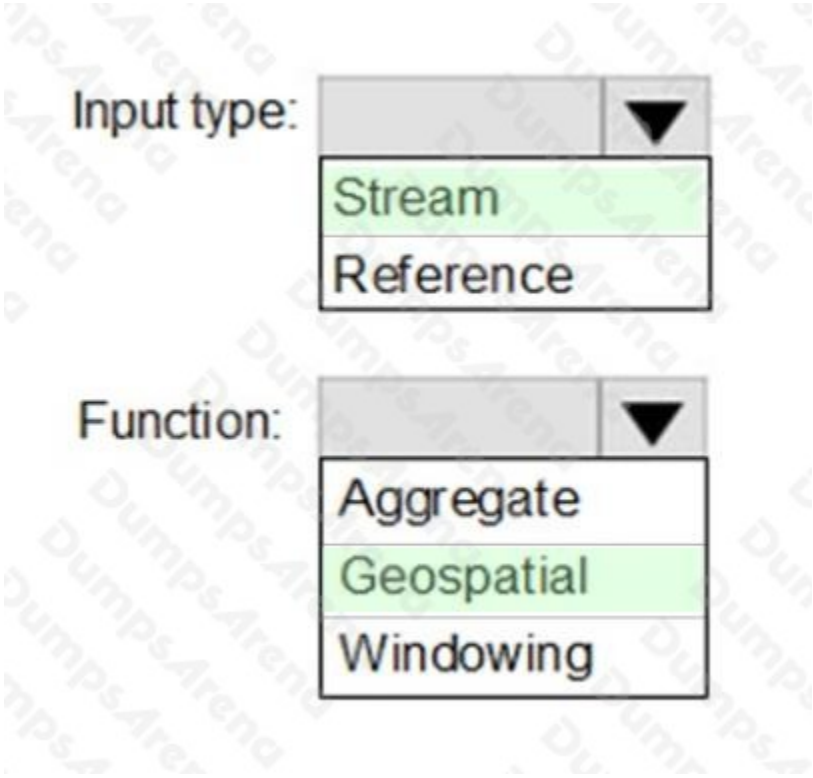
You need to design an Azure Stream Analytics job to process the data for the planned app. The solution must minimize the amount of code developed and the number of technologies used.

What should you include in the Stream Analytics job? To answer, select the appropriate options in the answer area.

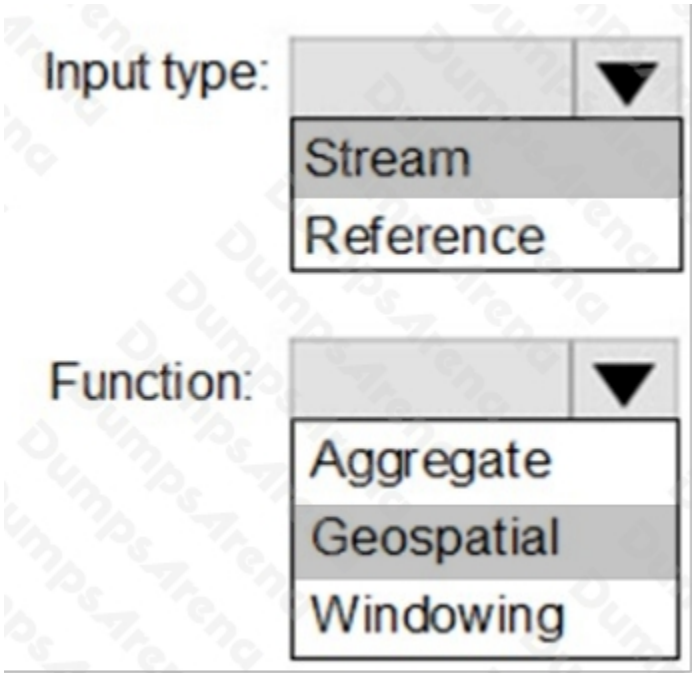
NOTE: Each correct selection is worth one point.



ANSWER:



Explanation:



Input type: Stream

You can process real-time IoT data streams with Azure Stream Analytics.

Function: Geospatial

With built-in geospatial functions, you can use Azure Stream Analytics to build applications for scenarios such as fleet management, ride sharing, connected cars, and asset tracking.

Note: In a real-world scenario, you could have hundreds of these sensors generating events as a stream. Ideally, a gateway device would run code to push these events to Azure Event Hubs or Azure IoT Hubs.

Reference:

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-get-started-with-azure-stream-analytics-to-process-data-from-iot-devices

https://docs.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios

QUESTION NO: 8 - (SIMULATION)

You have an Azure Storage account that generates 200.000 new files daily. The file names have a format of (YYY)/(MM)/(DD)/|HH])/(CustornerID).csv.

You need to design an Azure Data Factory solution that will toad new data from the storage account to an Azure Data lake once hourly. The solution must minimize load times and costs.

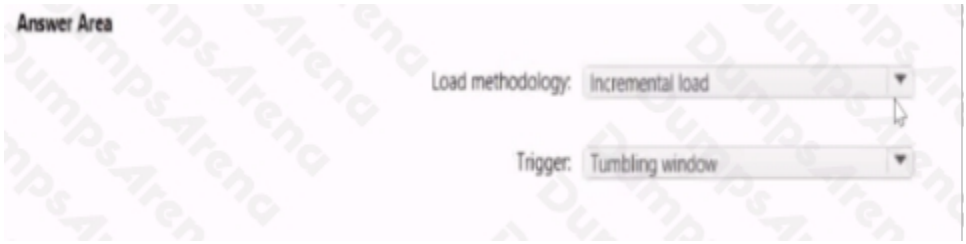
How should you configure the solution? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

ANSWER:

Explanation:

Answer as below



Answer:

QUESTION NO: 9

You haw an Azure data factory named ADF1.

You currently publish all pipeline authoring changes directly to ADF1.

You need to implement version control for the changes made to pipeline artifacts. The solution must ensure that you can apply version control to the resources currently defined m the UX Authoring canvas for ADF1.

Which two actions should you perform? Each correct answer presents part of the solution

NOTE: Each correct selection is worth one point.

A. Create an Azure Data Factory trigger

B. From the UX Authoring canvas, select Set up code repository

C. Create a GitHub action

D. From the UX Authoring canvas, run Publish All.

E. Create a Git repository

F. From the UX Authoring canvas, select Publish

ANSWER: D E

Explanation:

Reference:

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

QUESTION NO: 10 - (DRAG DROP)

DRAG DROP

You are designing an Azure Data Lake Storage Gen2 structure for telemetry data from 25 million devices distributed across seven key geographical regions. Each minute, the devices will send a JSON payload of metrics to Azure Event Hubs.

You need to recommend a folder structure for the data. The solution must meet the following requirements:

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

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

NOTE: Each correct selection is worth one point.

Select and Place:

ANSWER:

Explanation:

Box 1: {raw/regionID}

Box 2: {YYYY}/{MM}/{DD}/{HH}/{mm}

Box 3: {deviceID}

Reference:

https://github.com/paolosalvatori/StreamAnalyticsAzureDataLakeStore/blob/master/README.md

QUESTION NO: 11

You are designing a security model for an Azure Synapse Analytics dedicated SQL pool that will support multiple companies. You need to ensure that users from each company can view only the data of their respective company. Which two objects should you include in the solution? Each correct answer presents part of the solution

NOTE: Each correct selection it worth one point.

A. a custom role-based access control (RBAC) role.

B. asymmetric keys

C. a predicate function

D. a column encryption key

E. a security policy

ANSWER: A E

Explanation:

Reference:

https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security

https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-access-control-overview

QUESTION NO: 12

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

QUESTION NO: 13

You are designing an anomaly detection solution for streaming data from an Azure IoT hub. The solution must meet the following requirements:

Which should you include in the solution?

A. Azure Databricks

B. Azure Stream Analytics

C. Azure SQL Database

ANSWER: B

Explanation:

You can identify anomalies by routing data via IoT Hub to a built-in ML model in Azure Stream Analytics.

Reference:

https://docs.microsoft.com/en-us/learn/modules/data-anomaly-detection-using-azure-iot-hub/

QUESTION NO: 14

You have the following Azure Data Factory pipelines:

- Ingest Data from System1
- Ingest Data from System2
- Populate Dimensions
- Populate Facts

Ingest Data from System1 and Ingest Data from System2 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest Data from System2. Populate Facts must execute after 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. Add a schedule trigger to all four pipelines.

C. Create a patient pipeline that contains the four pipelines and use a schedule trigger.

D. Create a patient pipeline that contains the four pipelines and use an event trigger.

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

QUESTION NO: 15

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-pipelines/3-choose-between-dimension-types

QUESTION NO: 16 - (DRAG DROP)

You have an Azure Data Lake Storage Gen 2 account named storage1.

You need to recommend a solution for accessing the content in storage1. The solution must meet the following requirements:

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

NOTE: Each correct selection is worth one point.

ANSWER:

Explanation:

Box 1: Role-based access control (RBAC) roles

List and read permissions must be granted at the storage account level.

Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.

Role-based access control (Azure RBAC)

Azure RBAC uses role assignments to apply sets of permissions to security principals. A security principal is an object that represents a user, group, service principal, or managed identity that is defined in Azure Active Directory (AD). A permission set can give a security principal a "coarse-grain" level of access such as read or write access to all of the data in a storage account or all of the data in a container.

Box 2: Access control lists (ACLs)

Additional permissions can be applied to individual objects in storage1.

Access control lists (ACLs)

ACLs give you the ability to apply "finer grain" level of access to directories and files. An ACL is a permission construct that contains a series of ACL entries. Each ACL entry associates security principal with an access level.

Reference: https://learn.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control-model

QUESTION NO: 17

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: A D

QUESTION NO: 18

A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU).

You need to optimize performance for the Azure Stream Analytics job.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Implement event ordering.

B. Implement Azure Stream Analytics user-defined functions (UDF).

C. Implement query parallelization by partitioning the data output.

D. Scale the SU count for the job up.

E. Scale the SU count for the job down.

F. Implement query parallelization by partitioning the data input.

ANSWER: D F

Explanation:

Reference:

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

QUESTION NO: 19

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

QUESTION NO: 20

You are designing a financial transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:

You have the following query requirements:

You need to recommend a partition strategy for the table to minimize query times.

On which column should you recommend partitioning the table?

A. CustomerSegment

B. AccountType

C. TransactionType

D. TransactionMonth

ANSWER: C

Explanation:

For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

Example: Any partitioning added to a table is in addition to the distributions created behind the scenes. Using this example, if the sales fact table contained 36 monthly partitions, and given that a dedicated SQL pool has 60 distributions, then the sales fact table should contain 60 million rows per month, or 2.1 billion rows when all months are populated. If a table contains fewer than the recommended minimum number of rows per partition, consider using fewer partitions in order to increase the number of rows per partition.