Fast2Test
FIRST TEST, FIRST PASS!

Fast2Test

Instant Download

After Payment, our system will send you the products you purchase in mailbox in a minute after payment. If not received within 2 hours, please contact us.

365 Days Free Updates

Free update is available within 365 days after your purchase. After 365 days, you will get 50% discounts for updating.

Money Back Guarantee

Full refund if you fail the corresponding exam in 60 days after purchasing. And Free get any another product.

Security & Privacy

We respect customer privacy. We use McAfee's security service to provide you with utmost security for your personal information & peace of mind.

We're not the only ones **happy** about Fast2test Practice Materials ...

62316+ customers in 100+ countries use Fast2test Self Test Engine. Meet our customers.

https://tw.fast2test.com

高效的考試材料是最高通過率的考試題庫

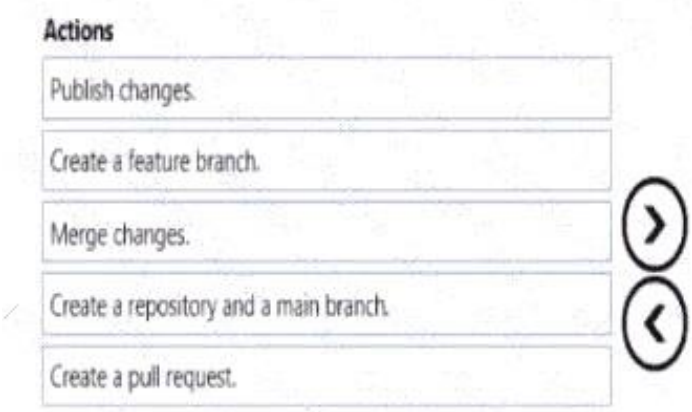
Exam : **DP-203**

Title : Data Engineering on Microsoft Azure

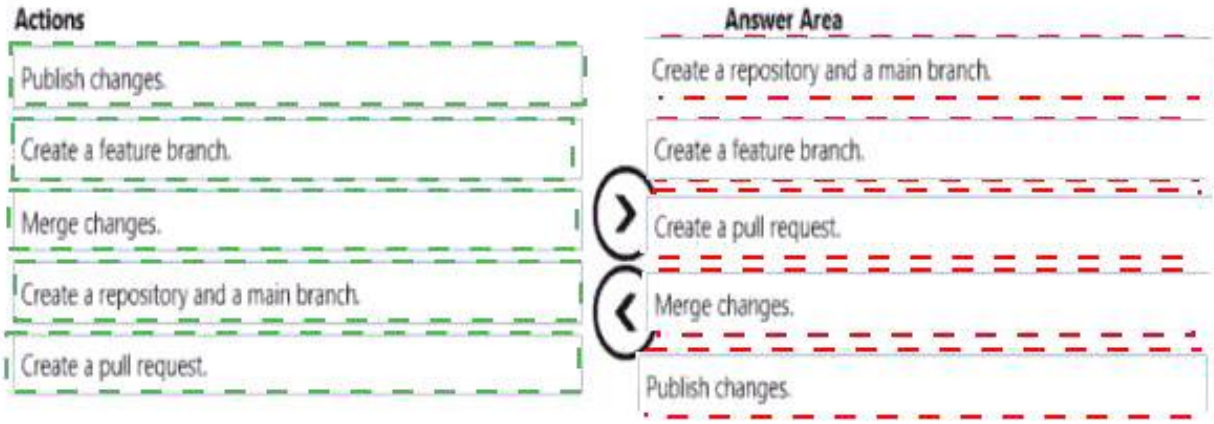
Vendor : Microsoft

Version : DEMO

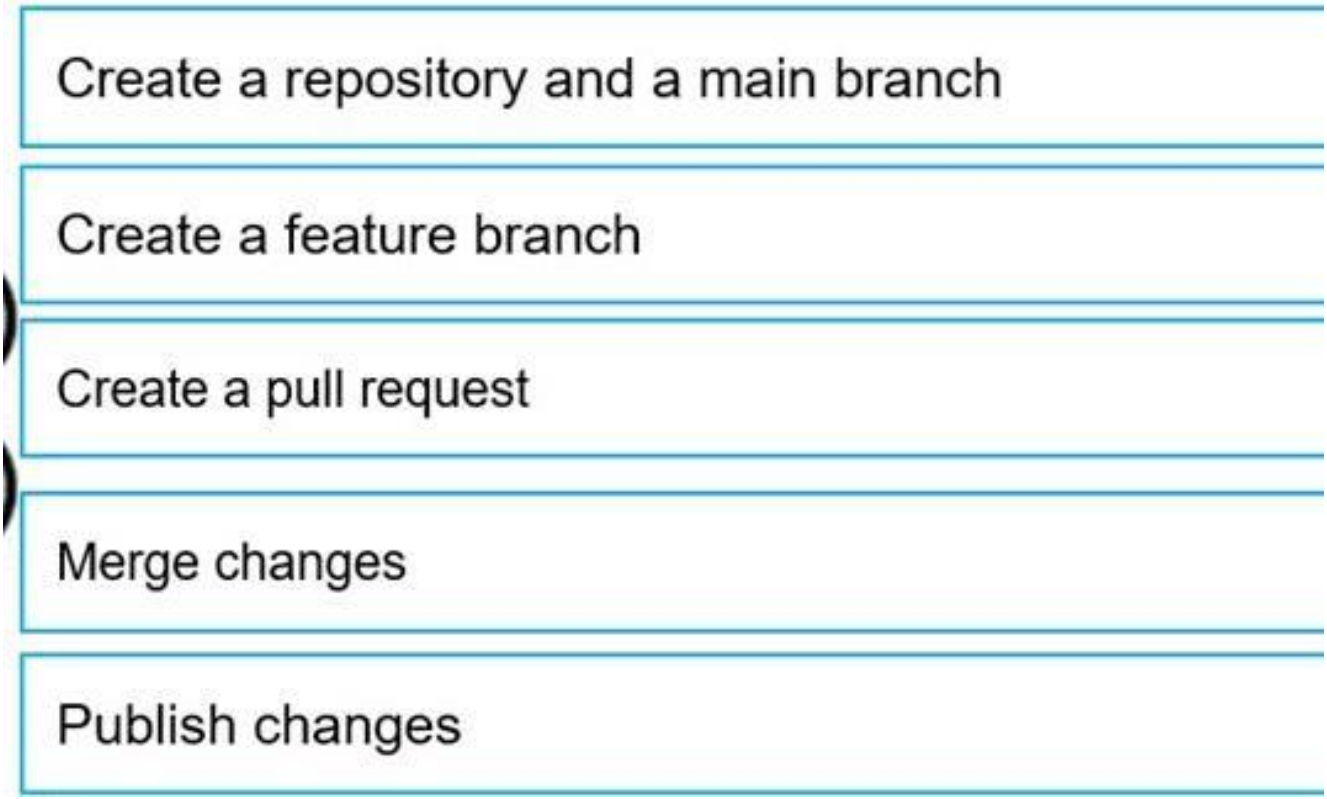
NO.1 You need to implement versioned changes to the integration pipelines. The solution must meet the data integration requirements.
In which order should you perform the actions? To answer, move all actions from the list of actions to the answer area and arrange them in the correct order.



Answer:



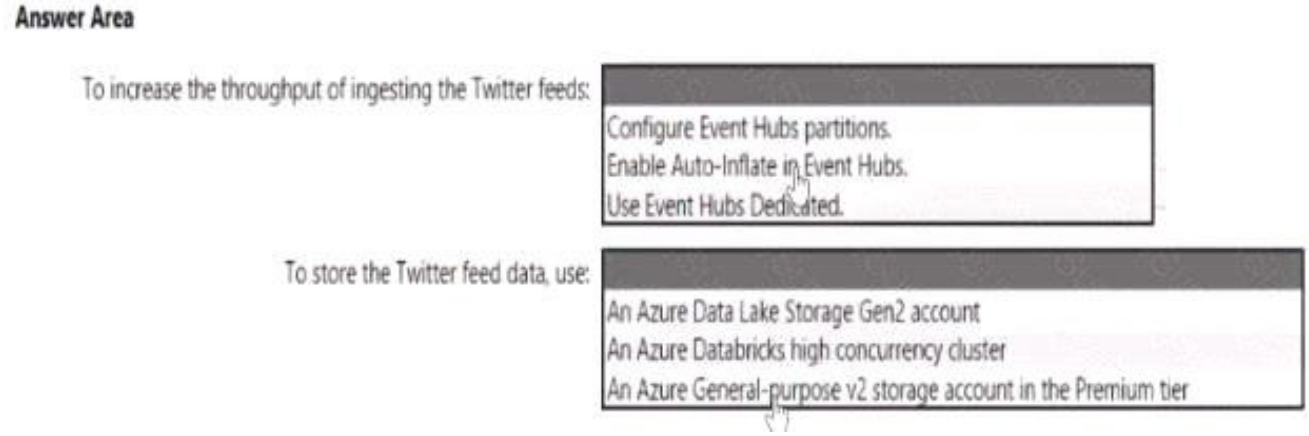
Explanation:



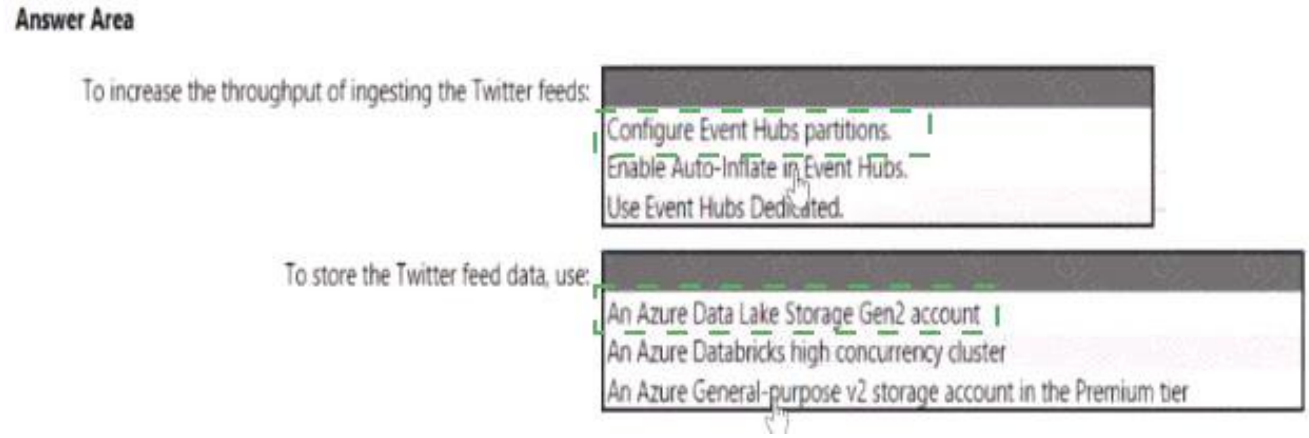
Scenario: Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.
Step 1: Create a repository and a main branch
You need a Git repository in Azure Pipelines, TFS, or GitHub with your app.
Step 2: Create a feature branch
Step 3: Create a pull request
Step 4: Merge changes
Merge feature branches into the main branch using pull requests.
Step 5: Publish changes
Reference:
https://docs.microsoft.com/en-us/azure/devops/pipelines/repos/pipeline-options-for-git
Topic 1, Contoso Case StudyTransactional Date
Contoso has three years of customer, transactional, operation, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL server instances contain data from various operational systems. The data is loaded into the instances by using SQL server integration Services (SSIS) packages.
You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.
Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time period. Sales transaction data that is older than three years will be removed monthly.
You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.
You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.
Streaming Twitter Data

The ecommerce department at Contoso develops and Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.
Planned Changes
Contoso plans to implement the following changes:
* Load the sales transaction dataset to Azure Synapse Analytics.
* Integrate on-premises data stores with Azure Synapse Analytics by using SSIS packages.
* Use Azure Synapse Analytics to analyze Twitter feeds to assess customer sentiments about products.
Sales Transaction Dataset Requirements
Contoso identifies the following requirements for the sales transaction dataset:
* Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong: to the partition on the right.
* Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.
* Implement a surrogate key to account for changes to the retail store addresses.
* Ensure that data storage costs and performance are predictable.
* Minimize how long it takes to remove old records.
Customer Sentiment Analytics Requirement
Contoso identifies the following requirements for customer sentiment analytics:
* Allow Contoso users to use PolyBase in an A/ure 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 A/ureAD credentials.
* Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.
* Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.
* Ensure that the data store supports Azure AD-based access control down to the object level.
* Minimize administrative effort to maintain the Twitter feed data records.
* Purge Twitter feed data records;itftaitJ are older than two years.
Data Integration Requirements
Contoso identifies the following requirements for data integration:
Use an Azure service that leverages the existing SSIS packages to ingest on-premises data into datasets stored in a dedicated SQL pool of Azure Synaps Analytics and transform the data.
Identify a process to ensure that changes to the ingestion and transformation activities can be version controlled and developed independently by multiple data engineers.

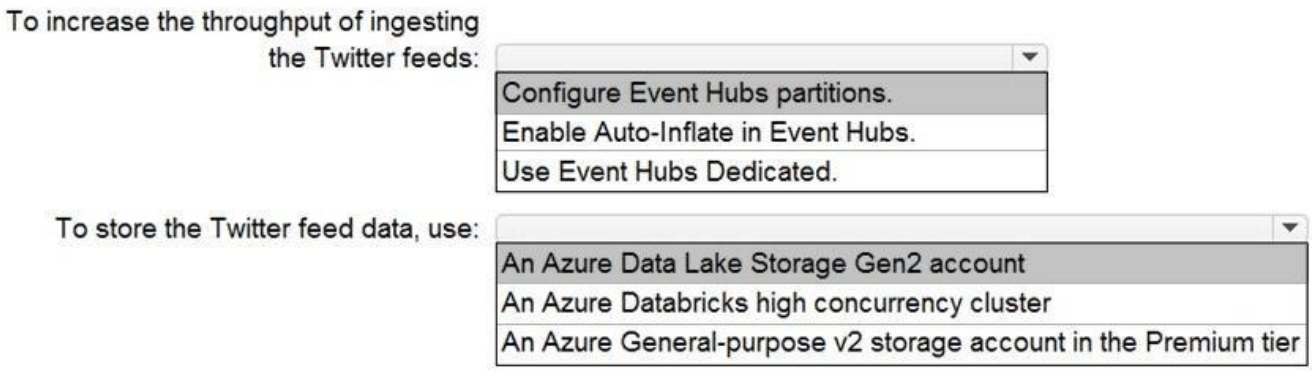
NO.2 You need to design a data ingestion and storage solution for the Twitter feeds. The solution must meet the customer sentiment analytics requirements.
What should you include in the solution? To answer, select the appropriate options in the answer area NOTE: Each correct selection b worth one point.



Answer:



Explanation:



Box 1: Configure Evegent Hubs partitions

Scenario: Maximize the throughput of ingesting Twitter feeds from Event Hubs to Azure Storage without purchasing additional throughput or capacity units.

Event Hubs is designed to help with processing of large volumes of events. Event Hubs throughput is scaled by using partitions and throughput-unit allocations.

Event Hubs traffic is controlled by TUs (standard tier). Auto-inflate enables you to start small with the minimum required TUs you choose. The feature then scales automatically to the maximum limit of TUs you need, depending on the increase in your traffic.

Box 2: An Azure Data Lake Storage Gen2 account

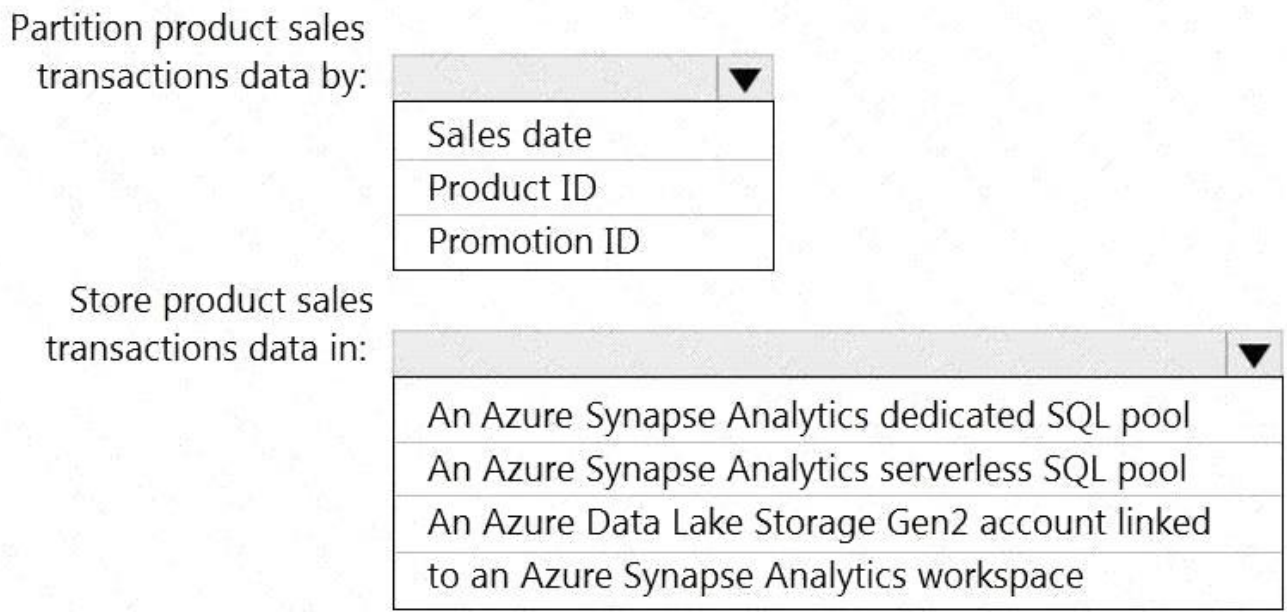
Scenario: Ensure that the data store supports Azure AD-based access control down to the object level.

Azure Data Lake Storage Gen2 implements an access control model that supports both Azure role-based access control (Azure RBAC) and POSIX-like access control lists (ACLs).

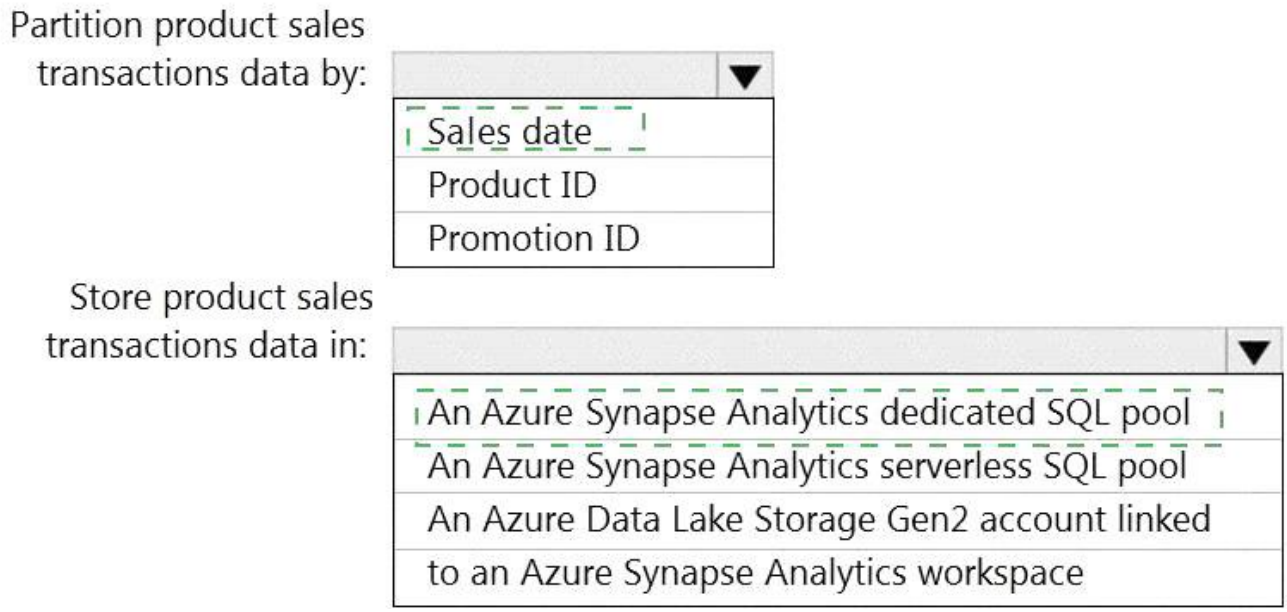
Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control

NO.3 You need to integrate the on-premises data sources and Azure Synapse Analytics. The solution must meet the data integration requirements.
Which type of integration runtime should you use?
A. Azure-SSIS integration runtime
B. self-hosted integration runtime
C. Azure integration runtime
Answer: C

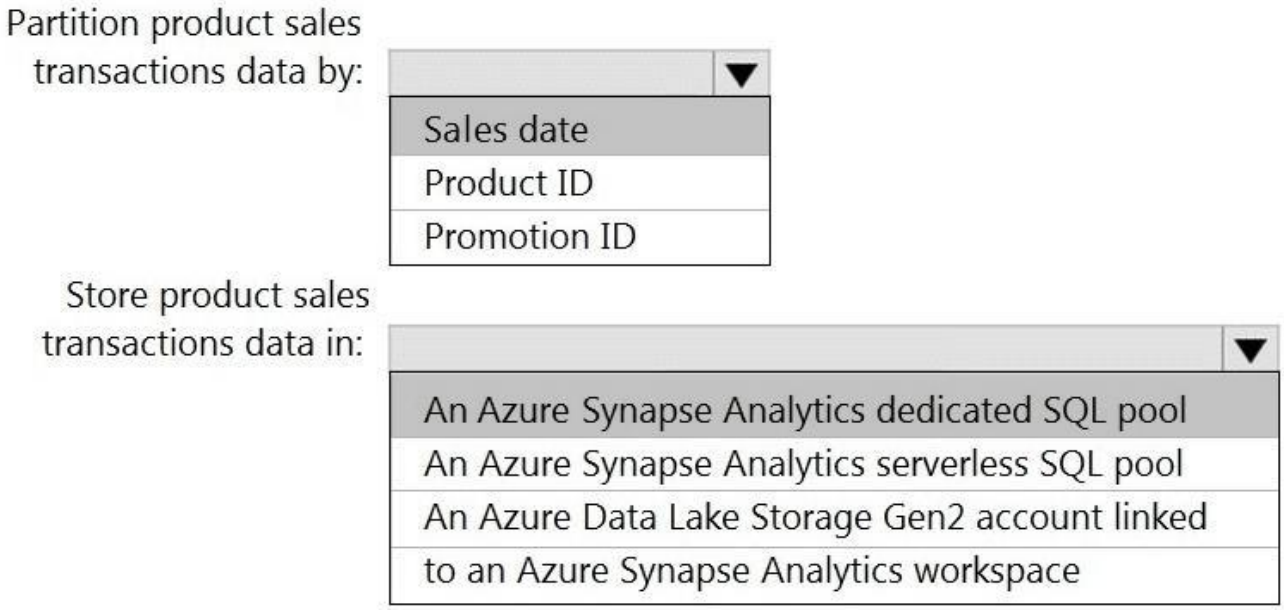
NO.4 You need to design the partitions for the product sales transactions. The solution must meet the sales transaction dataset requirements.
What should you include in the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



Explanation:



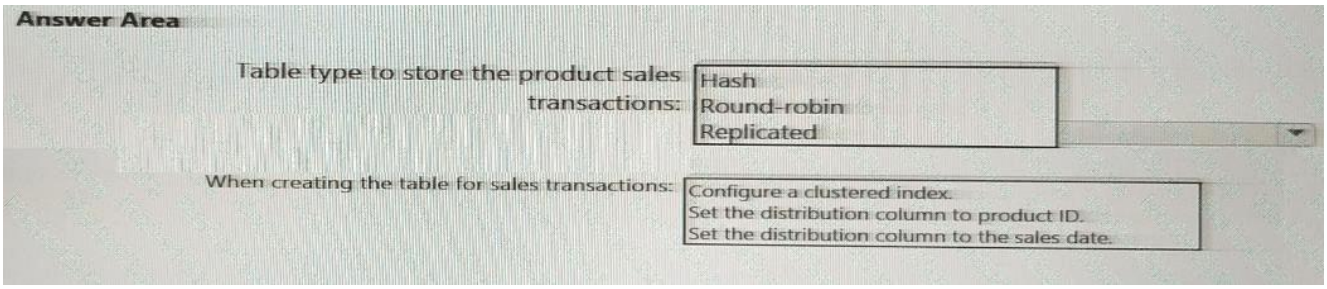
Box 1: Sales date
Scenario: Contoso requirements for data integration include:
Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.
Box 2: An Azure Synapse Analytics Dedicated SQL pool
Scenario: Contoso requirements for data integration include:
Ensure that data storage costs and performance are predictable.
The size of a dedicated SQL pool (formerly SQL DW) is determined by Data Warehousing Units (DWU).
Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage. This format significantly reduces the data storage costs, and improves query performance.
Synapse analytics dedicated sql pool
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-overview- what-is

NO.5 You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1.
You use Azure Monitor.
You need to monitor the performance of queries executed in Pool1.
Which log should you query?
A. SynapseSqlPoolWaits
B. SynapseSqlPoolSqlRequests
C. SynapseSqlPoolExecRequests
D. SynapseSqlPoolRequestSteps
Answer: C

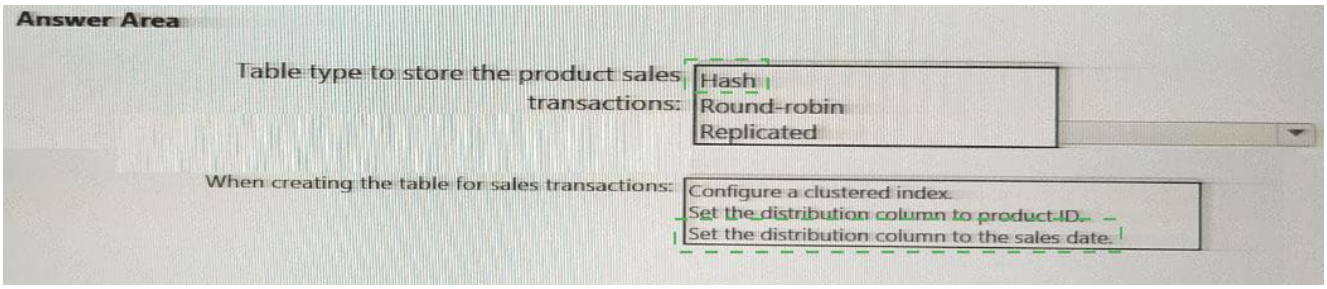
NO.6 You need to design a data retention solution for the Twitter teed data records. The solution must meet the customer sentiment analytics requirements.
Which Azure Storage functionality should you include in the solution?
A. time-based retention
B. change feed
C. soft delete
D. lifecycle management
Answer: C

NO.7 You need to implement the surrogate key for the retail store table. The solution must meet the sales transaction dataset requirements.
What should you create?
A. a table that has an IDENTITY property
B. a system-versioned temporal table
C. a user-defined SEQUENCE object
D. a table that has a FOREIGN KEY constraint
Answer: A
Explanation:
Scenario: Implement a surrogate key to account for changes to the retail store addresses.
A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables- identity

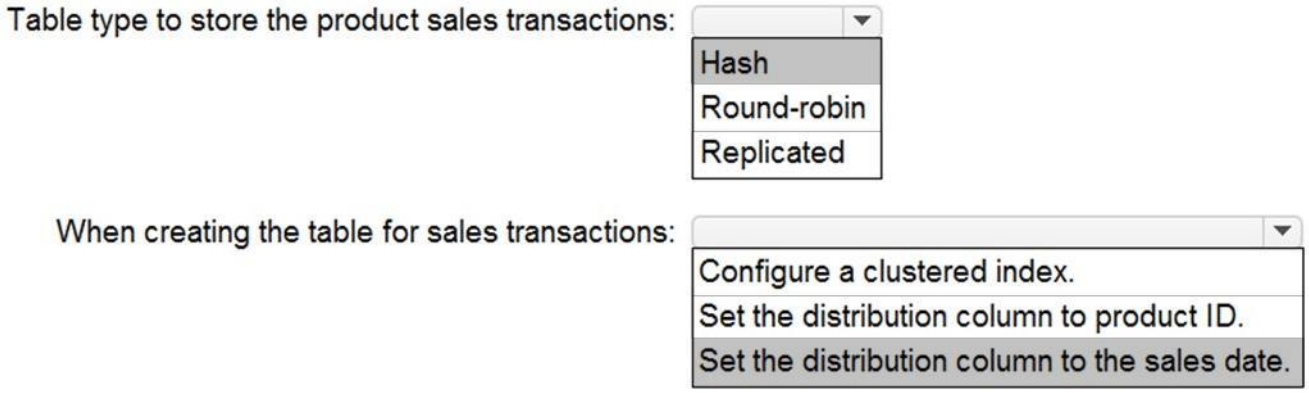
NO.8 You need to design a data storage structure for the product sales transactions. The solution must meet the sales transaction dataset requirements.
What should you include in the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:

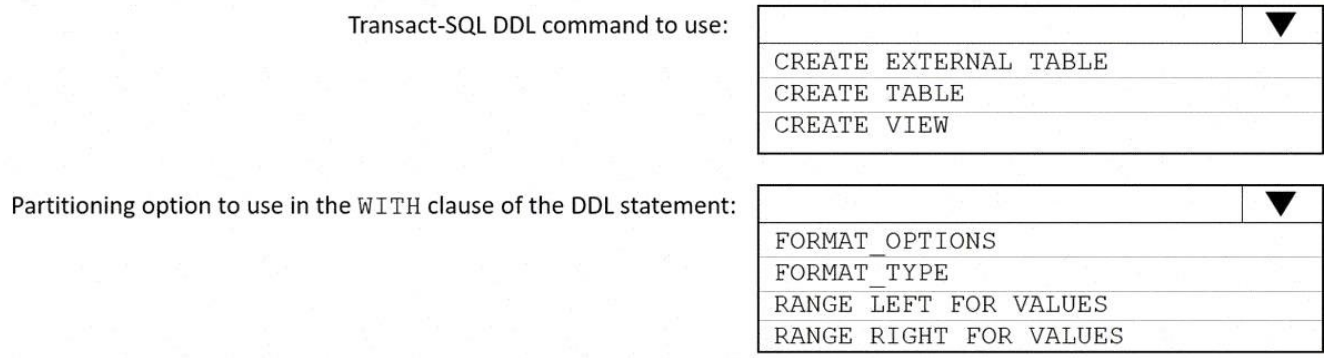


Explanation:

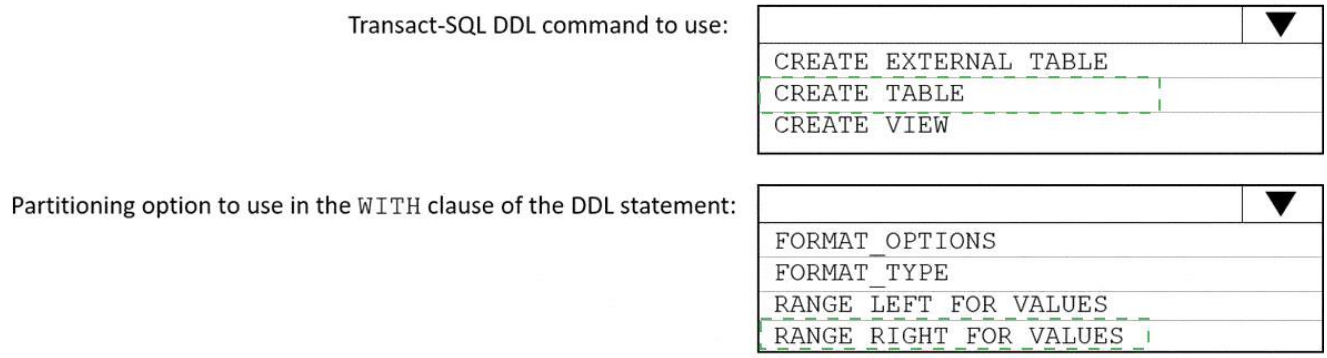


Box 1: Hash
Scenario:
Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.
A hash distributed table can deliver the highest query performance for joins and aggregations on large tables.
Box 2: Set the distribution column to the sales date.
Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.
Reference:
https://rajanieshkaushikk.com/2020/09/09/how-to-choose-right-data-distribution-strategy-for-azure-synapse/

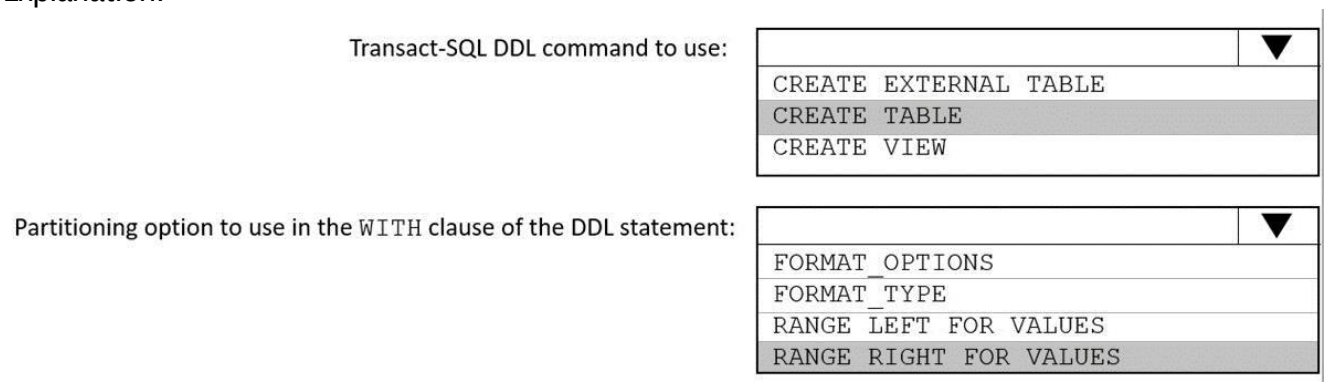
NO.9 You need to implement an Azure Synapse Analytics database object for storing the sales transactions data.
The solution must meet the sales transaction dataset requirements.
What solution must meet the sales transaction dataset requirements.
What should you do? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



Explanation:

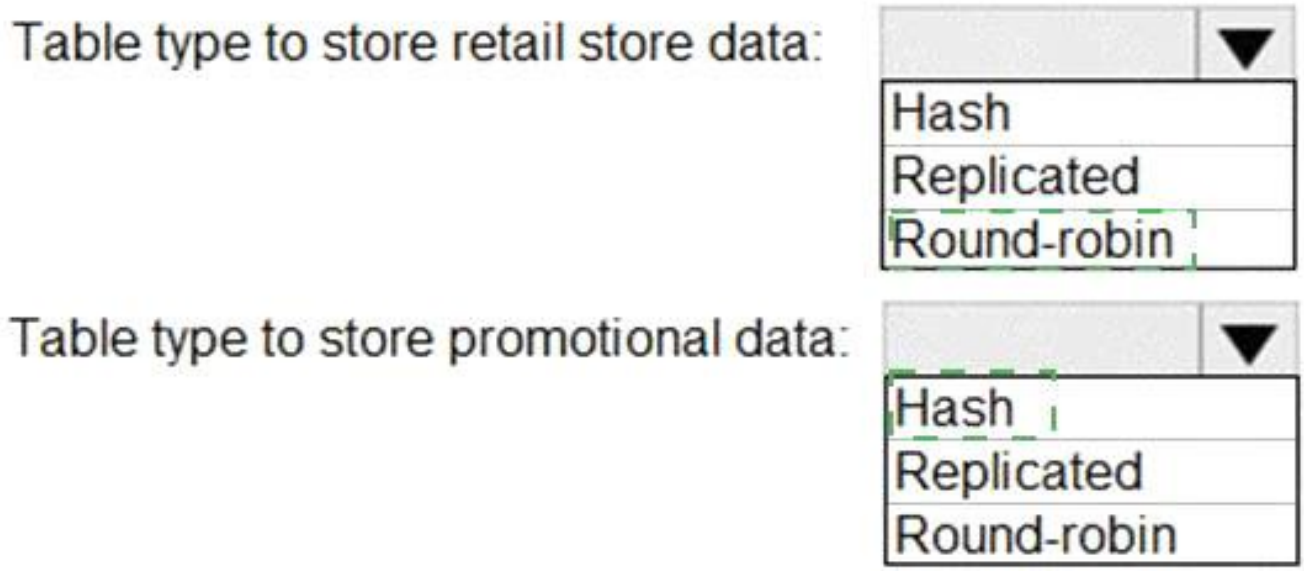


Box 1: Create table
Scenario: Load the sales transaction dataset to Azure Synapse Analytics Box 2: RANGE RIGHT FOR VALUES Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.
RANGE RIGHT: Specifies the boundary value belongs to the partition on the right (higher values).
FOR VALUES (boundary_value [,...n]): Specifies the boundary values for the partition.
Scenario: Load the sales transaction dataset to Azure Synapse Analytics.
Contoso identifies the following requirements for the sales transaction dataset:
Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.
Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.
Implement a surrogate key to account for changes to the retail store addresses.
Ensure that data storage costs and performance are predictable.
Minimize how long it takes to remove old records.
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse

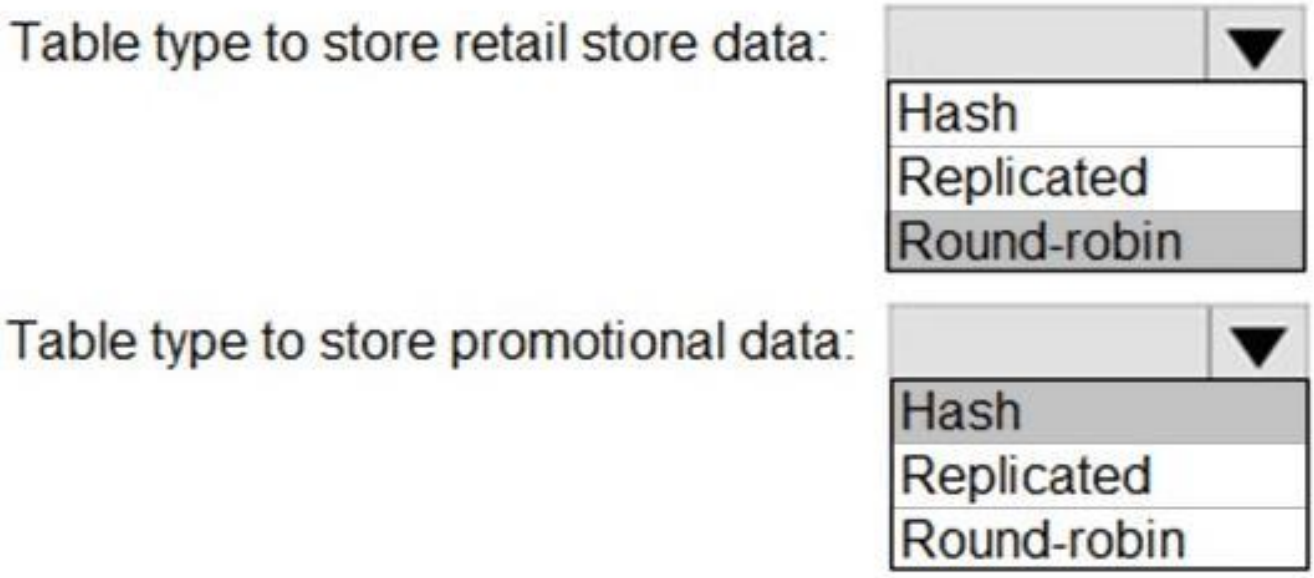
NO.10 You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements.
Which Azure Storage functionality should you include in the solution?
A. change feed
B. soft delete
C. time-based retention
D. lifecycle management
Answer: B
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

NO.11 You need to design an analytical storage solution for the transactional data. The solution must meet the sales transaction dataset requirements.
What should you include in the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer:



Explanation:



Box 1: Round-robin
Round-robin tables are useful for improving loading speed.
Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month.
Box 2: Hash
Hash-distributed tables improve query performance on large fact tables.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

NO.12 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

Answer:

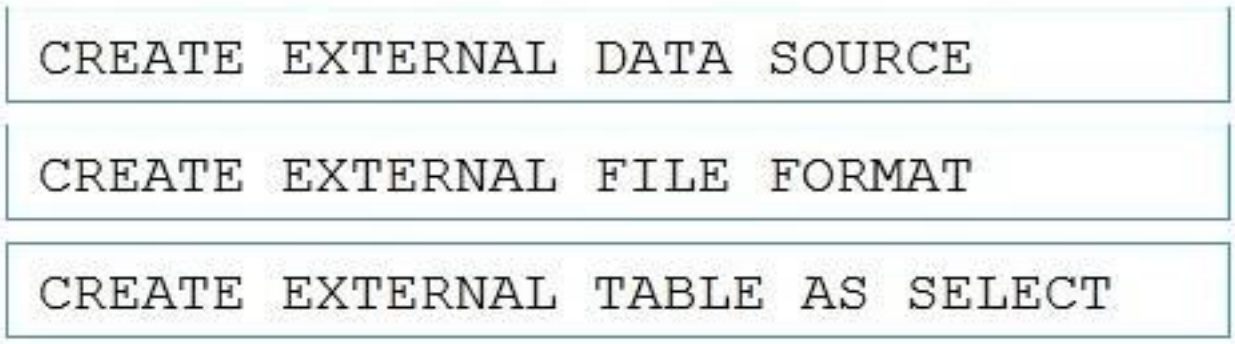
Commands

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

Answer Area

- CREATE EXTERNAL DATA SOURCE
- CREATE EXTERNAL FILE FORMAT
- CREATE EXTERNAL TABLE AS SELECT

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.

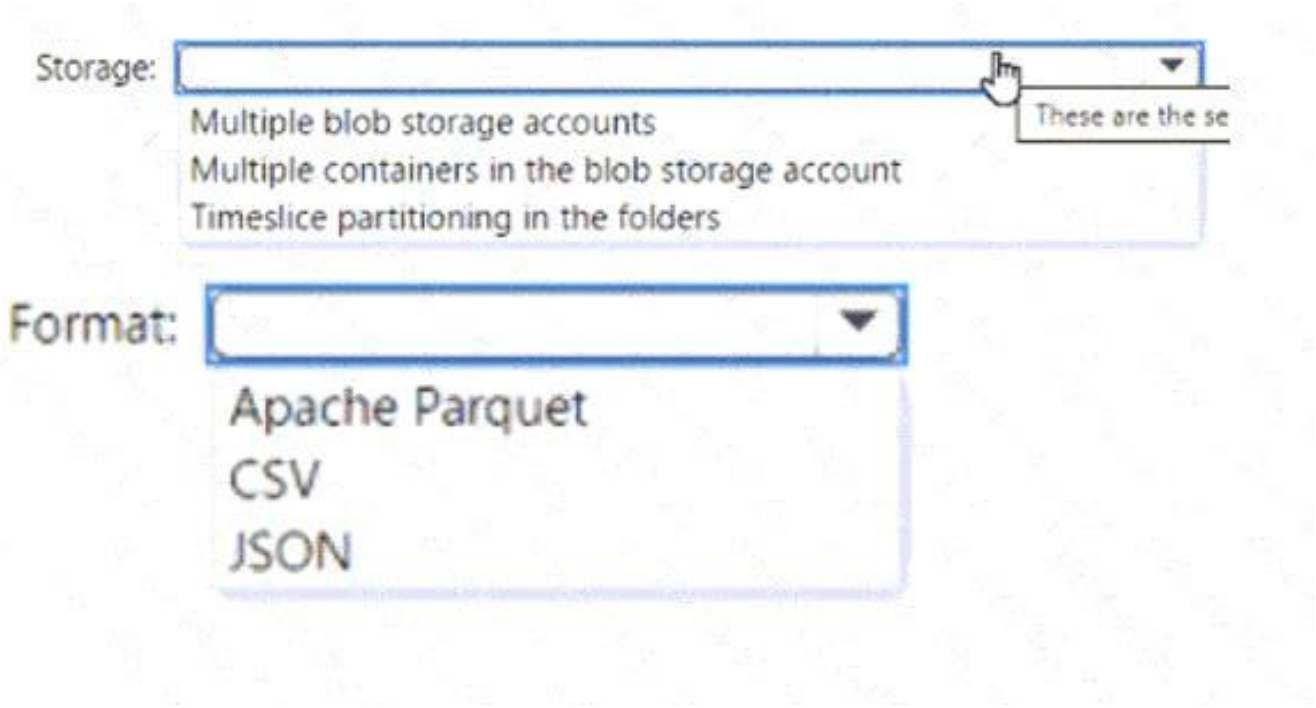
Reference:

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

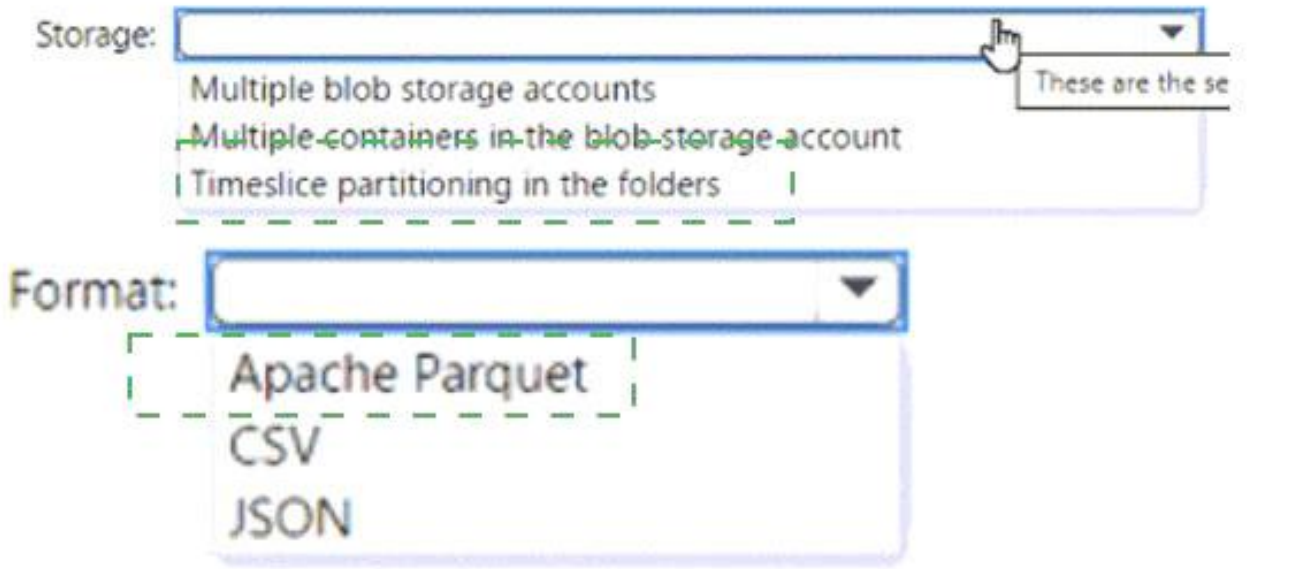
NO.13 You have an Azure Synapse Analytics dedicated SQL pool named Pcol1. Pool1 contains a table named tablet You load 5 TB of data into table1.
You need to ensure that column store compression is maximized for table1.
Which statement should you execute?
A. DBCC IIDEXDEFRAG (pool1, table1)
B. ALTER INDEX ALL on table REORGANIZE
C. DBCC DBREINOEX (table)
D. ALTER INDEX ALL on table REBUILD
Answer: C
Topic 2, Litware, inc. Case Study
Case study
This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.
To start the case study
To display the first question in this case study, click the Next button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the Question button to return to the question.
Overview
Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.
Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.
Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.
Requirements
Business Goals
Litware wants to create a new analytics environment in Azure to meet the following requirements:
See inventory levels across the stores. Data must be updated as close to real time as possible.
Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.
Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.
Technical Requirements

Litware identifies the following technical requirements:
Minimize the number of different Azure services needed to achieve the business goals.
Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.
Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.
Use Azure Active Directory (Azure AD) authentication whenever possible.
Use the principle of least privilege when designing security.
Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store.
Litware wants to remove transient data from Data Lake Storage once the data is no longer in use.
Files that have a modified date that is older than 14 days must be removed.
Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.
Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.
Planned Environment
Litware plans to implement the following environment:
The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.
Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, 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.
Daily inventory data comes from a Microsoft SQL server located on a private network.
Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.
Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.
Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

NO.14 You have an Azure Blob storage account that contains a folder. The folder contains 120,000 files. Each file contains 62 columns.
Each day, 1,500 new files are added to the folder.
You plan to incrementally load five data columns from each new file into an Azure Synapse Analytics workspace.
You need to minimize how long it takes to perform the incremental loads.
What should you use to store the files and format?



Answer:



Explanation:
Box 1 = timeslice partitioning in the foldersThis means that you should organize your files into folders based on a time attribute, such as year, month, day, or hour. For example, you can have a folder structure like /yyyy
/mm/dd/file.csv. This way, you can easily identify and load only the new files that are added each day by using a time filter in your Azure Synapse pipeline12. Timeslice partitioning can also improve the performance of data loading and querying by reducing the number of files that need to be scanned
Box = 2 Apache Parquet This is because Parquet is a columnar file format that can efficiently store and compress data with many columns. Parquet files can also be partitioned by a time attribute, which can improve the performance of incremental loading and querying by reducing the number of files that need to be scanned123. Parquet files are supported by both dedicated SQL pool and

serverless SQL pool in Azure Synapse Analytics2.

NO.15 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

NO.16 You have an Azure subscription that contains an Azure data factory.

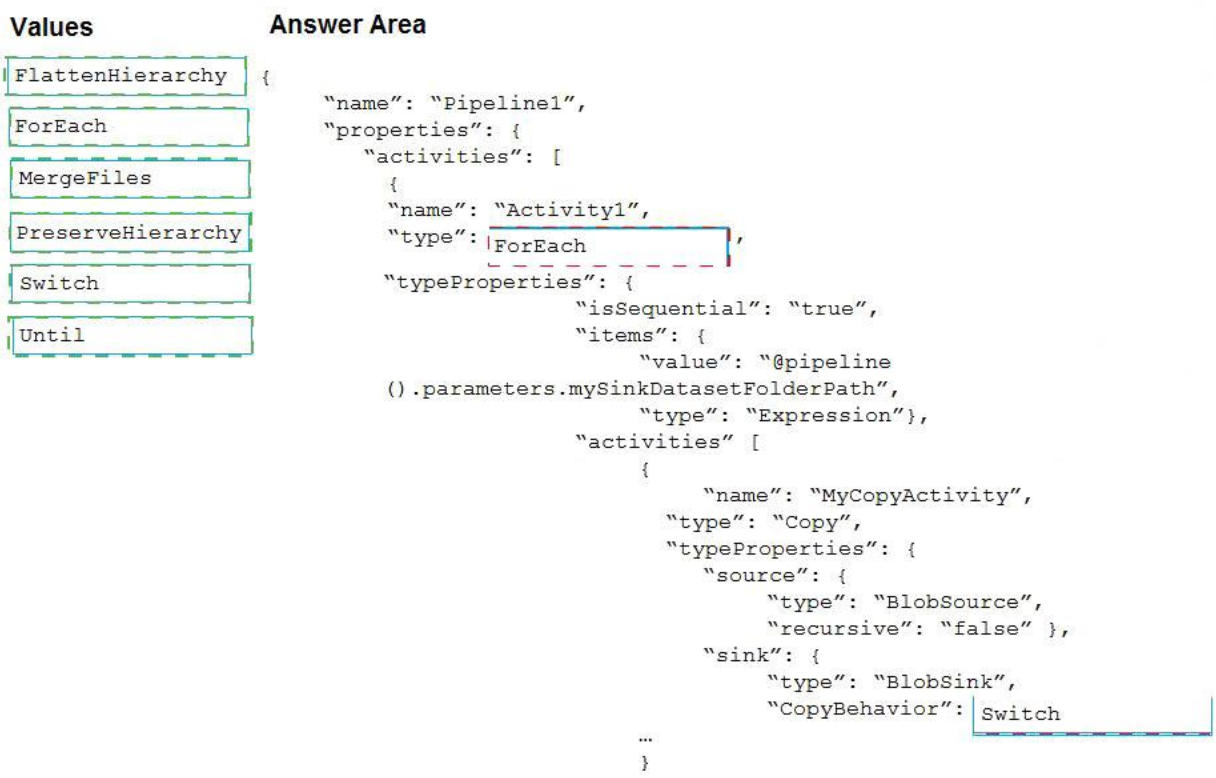
You are editing an Azure Data Factory activity JSON.

The script needs to copy a file from Azure Blob Storage to multiple destinations. The solution must ensure that the source and destination files have consistent folder paths.

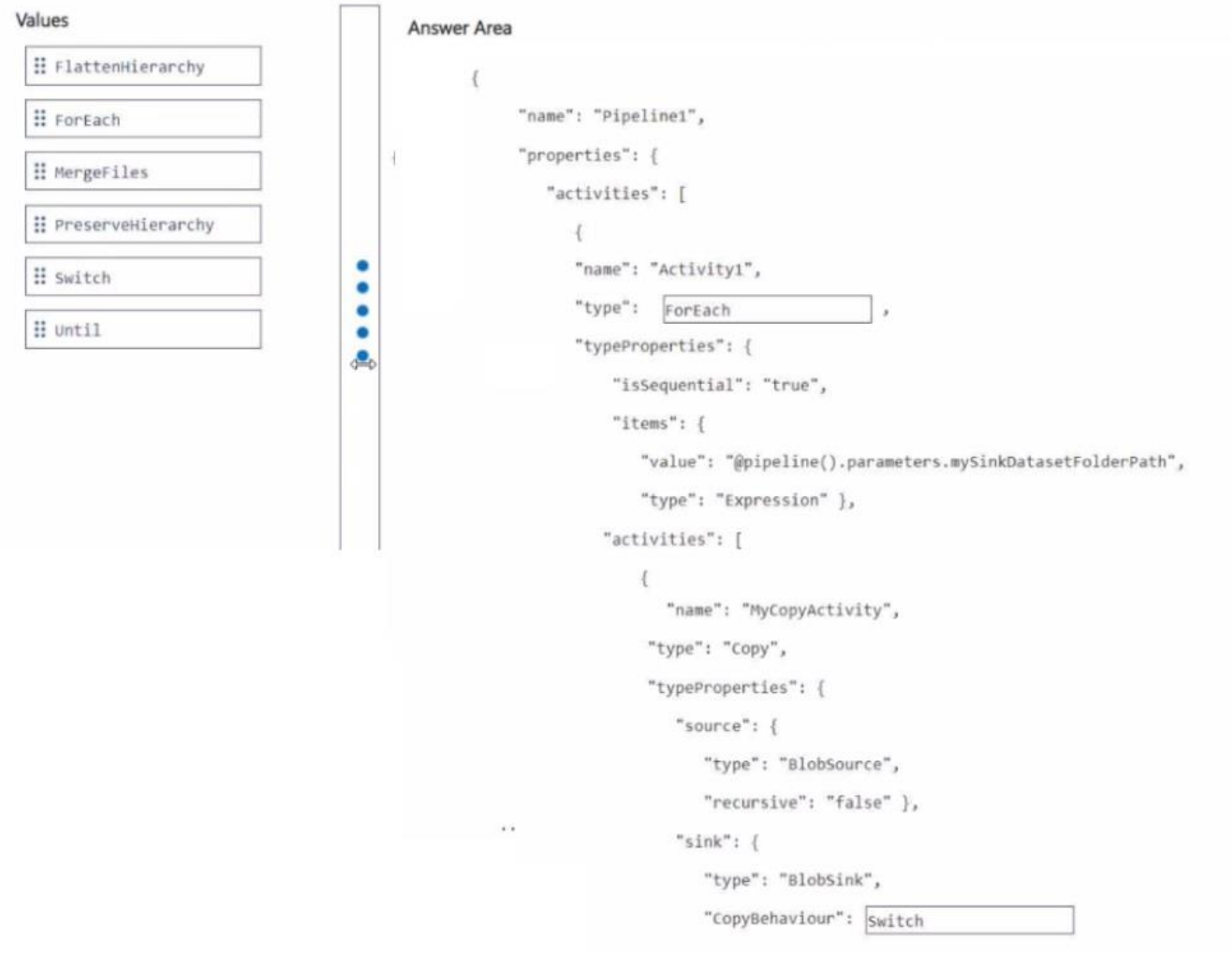
How should you complete the script? To answer, drag the appropriate values to the correct targets Each value may be used once, more than once, or not at all You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point

Answer:



Explanation:



NO.17 What should you recommend to prevent users outside the Litware on-premises network from accessing the analytical data store?

A. a server-level virtual network rule

B. a database-level virtual network rule

C. a database-level firewall IP rule

D. a server-level firewall IP rule

Answer: A

Explanation:

Virtual network rules are one firewall security feature that controls whether the database server for your single databases and elastic pool in Azure SQL Database or for your databases in SQL Data Warehouse accepts communications that are sent from particular subnets in virtual networks.

Server-level, not database-level: Each virtual network rule applies to your whole Azure SQL Database server, not just to one particular database on the server. In other words, virtual network rule applies at the serverlevel, not at the database-level.

References:

https://docs.microsoft.com/en-us/azure/sql-database/sql-database-vnet-service-endpoint-rule-overview

NO.18 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

NO.19 You have an Azure subscription that contains an Azure Cosmos DB analytical store and an Azure Synapse Analytics workspace named WS 1. WS1 has a serverless SQL pool name Pool1.
You execute the following query by using Pool1.

```
WITH IDENTITY = 'SHARED A
SECRET = 'fed434747987242343356365345 6345ddfa==';

SELECT clientID AS ClientID,
       client AS ClientName
FROM OPENROWSET
(
    PROVIDER = 'CosmosDB',
    CONNECTION = 'Account=account1;Database=database1',
                 OBJECT = 'clients',
                 SERVER_CREDENTIAL = 'AccountCred'
)
WITH

(
    clientID int,
    client varchar(50),
    streetAddress varchar(100)
) AS c;
```

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

Answer Area

Statements	Yes	No
The query returns three columns.	○	○
The container being queried is named `clients`.	○	○
Authentication is performed by using an account key.	○	○

Answer:

Answer Area

Statements	Yes	No
The query returns three columns.	○	○
The container being queried is named `clients`.	○	○
Authentication is performed by using an account key.	○	○

Explanation:

Answer Area

Statements	Yes	No
The query returns three columns.	○	●
The container being queried is named `clients`.	●	○
Authentication is performed by using an account key.	○	●

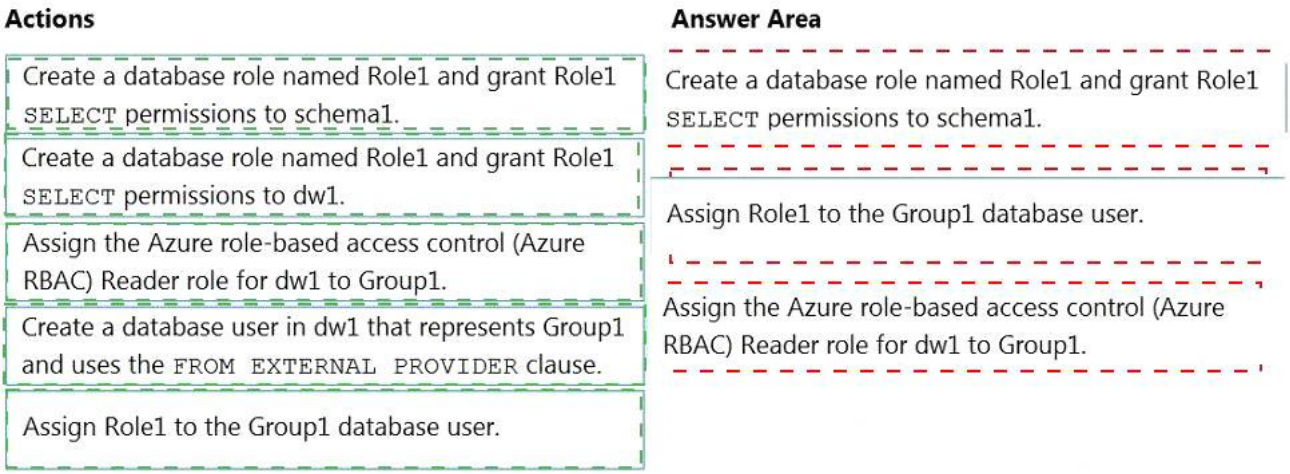
NO.20 You have an Azure Active Directory (Azure AD) tenant that contains a security group named Group1. You have an Azure Synapse Analytics dedicated SQL pool named dw1 that contains a schema named schema1.
You need to grant Group1 read-only permissions to all the tables and views in schema1. The solution must use the principle of least privilege.
Which three 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.

Actions

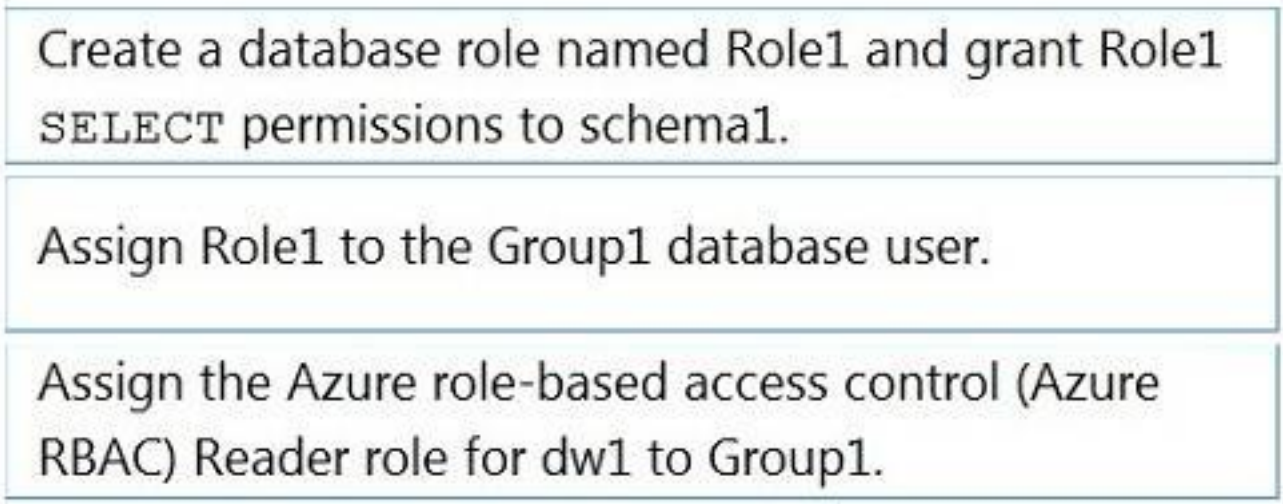
- Create a database role named Role1 and grant Role1 `SELECT` permissions to schema1.
- Create a database role named Role1 and grant Role1 `SELECT` permissions to dw1.
- Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1.
- Create a database user in dw1 that represents Group1 and uses the `FROM EXTERNAL PROVIDER` clause.
- Assign Role1 to the Group1 database user.

Answer Area

Answer:

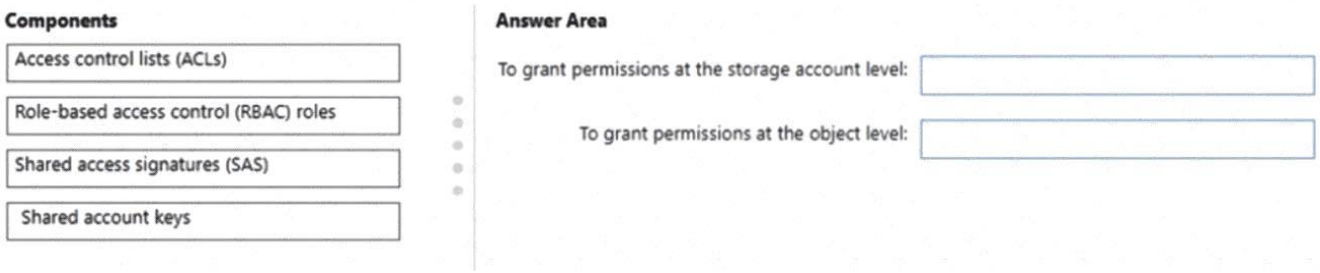


Explanation:

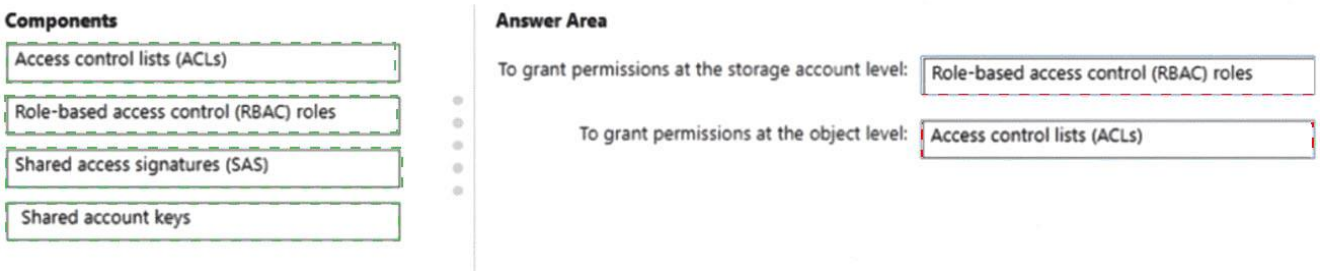


Step 1: Create a database role named Role1 and grant Role1 SELECT permissions to schema You need to grant Group1 read-only permissions to all the tables and views in schema1.
Place one or more database users into a database role and then assign permissions to the database role.
Step 2: Assign Rol1 to the Group database user
Step 3: Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1
Reference:
https://docs.microsoft.com/en-us/azure/data-share/how-to-share-from-sql

NO.21 You have an Azure Data Lake Storage Gen 2 account named storage1.
You need to recommend a solution for accessing the content in storage1. The solution must meet the following requirements:
List and read permissions must be granted at the storage account level.
Additional permissions can be applied to individual objects in storage1.
Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.
What should you use? To answer, drag the appropriate components to the correct requirements.
Each component may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.



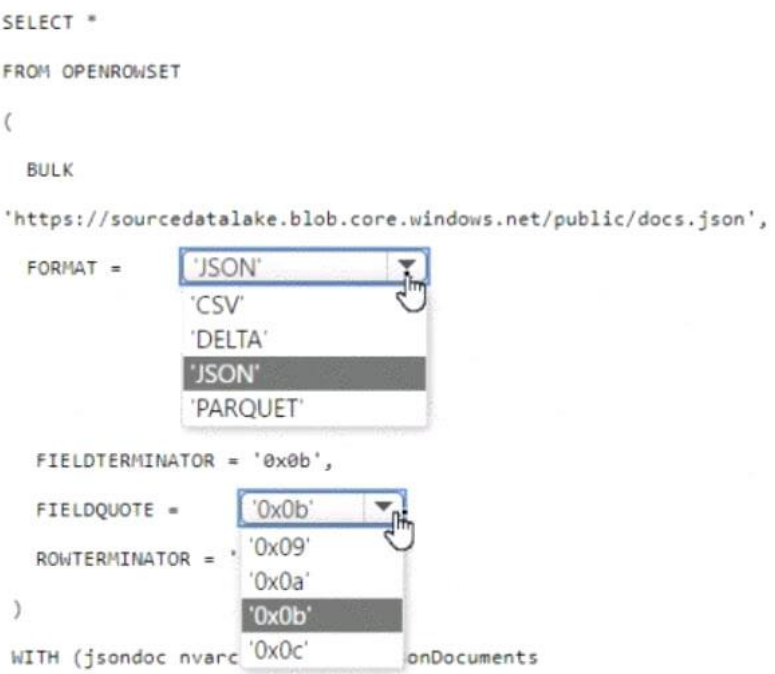
Answer:



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

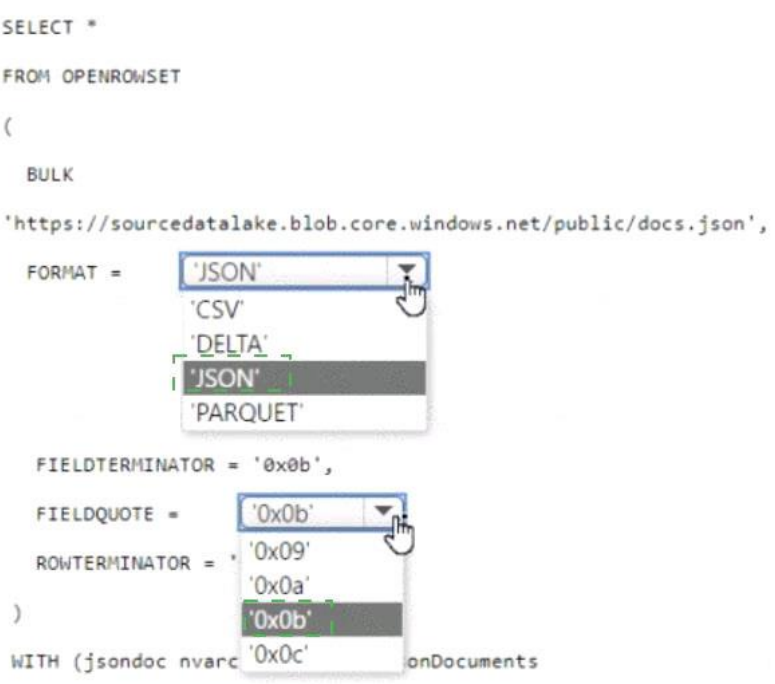
NO.22 You have an Azure Synapse serverless SQL pool.
You need to read JSON documents from a file by using the OPENROWSET function.
How should you complete the query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area



Answer:

Answer Area



Explanation:

Answer Area

```
SELECT *
FROM OPENROWSET
(
  BULK
'https://sourcedatalake.blob.core.windows.net/public/docs.json',
  FORMAT =      'JSON'
  FIELDTERMINATOR = '0x0b',
  FIELDQUOTE =      '0x0b'
  ROWTERMINATOR = '0x0b'
)
WITH (jsondoc nvarchar(max) AS JsonDocuments
```

NO.23 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

NO.24 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

NO.25 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

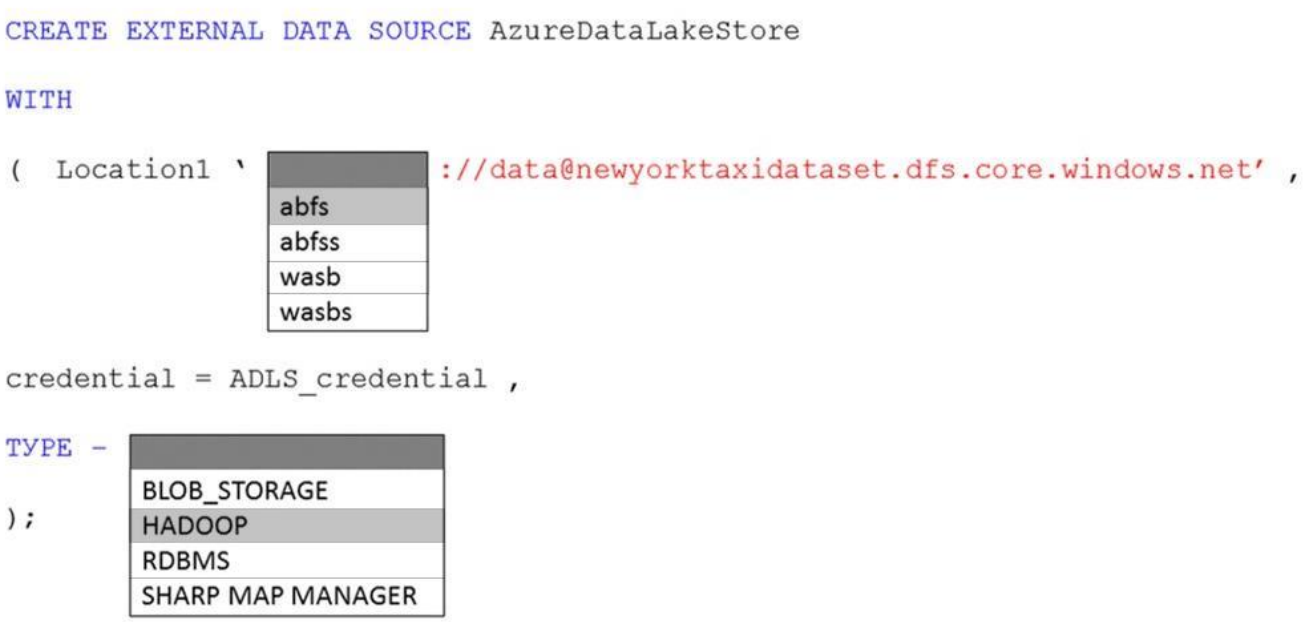
```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore

WITH

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

credential = ADLS_credential ,

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

Answer:

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
```

Explanation:



Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure- sqldw-latest&preserve-view=true&tabs=dedicated

NO.26 You are designing 2 solution that will use tables in Delta Lake on Azure Databricks.
You need to minimize how long it takes to perform the following:
*Queries against non-partitioned tables
* Joins on non-partitioned columns
Which two options should you include in the solution? Each correct answer presents part of the solution.
(Choose Correct Answer and Give Explanation and References to Support the answers based from Data Engineering on Microsoft Azure)

A. Z-Ordering

B. Apache Spark caching

C. dynamic file pruning (DFP)

D. the clone command

Answer: A C

Explanation:
According to the information I found on the web, two options that you should include in the solution to minimize how long it takes to perform queries and joins on non-partitioned tables are:
Z-Ordering: This is a technique to colocate related information in the same set of files. This co-locality is automatically used by Delta Lake in data-skipping algorithms. This behavior dramatically reduces the amount of data that Delta Lake on Azure Databricks needs to read123.
Apache Spark caching: This is a feature that allows you to cache data in memory or on disk for faster access.
Caching can improve the performance of repeated queries and joins on the same data. You can cache Delta tables using the CACHE TABLE or CACHE LAZY commands.
To minimize the time it takes to perform queries against non-partitioned tables and joins on non-partitioned columns in Delta Lake on Azure Databricks, the following options should be included in

the solution:
A:Z-Ordering: Z-Ordering improves query performance by co-locating data that share the same column values in the same physical partitions. This reduces the need for shuffling data across nodes during query execution. By using Z-Ordering, you can avoid full table scans and reduce the amount of data processed.
B:Apache Spark caching: Caching data in memory can improve query performance by reducing the amount of data read from disk. This helps to speed up subsequent queries that need to access the same data. When you cache a table, the data is read from the data source and stored in memory. Subsequent queries can then read the data from memory, which is much faster than reading it from disk.
References:
Delta Lake on Databricks: https://docs.databricks.com/delta/index.html
Best Practices for Delta Lake on Databricks: https://databricks.com/blog/2020/05/14/best-practices-for-delta- lake-on-databricks.html

NO.27 You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named storage1 and an Azure Synapse Analytics workspace named Workspace1. Workspace1 has a serverless SQL pool.
You use the serverless SQL pool to query customer orders from the files in storage1.
You run the following query.
SELECT *
FROM OPENROWSET(BULK 'https://storage1.blob.core.windows.net/data/orders/year =* /month =* / *.* ', FORMAT = 'parquet') AS customerorders WHERE customerorders. filepath(1) = '2024' AND customerorders.filepath(2) IN ('3','4'); For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.

Answer Area

Statements	Yes	No
storage1 provides a hierarchical namespace.	○	○
Files from March 2025 will be included.	○	○
Only files that have a Parquet file extension will be included.	○	○

Answer:

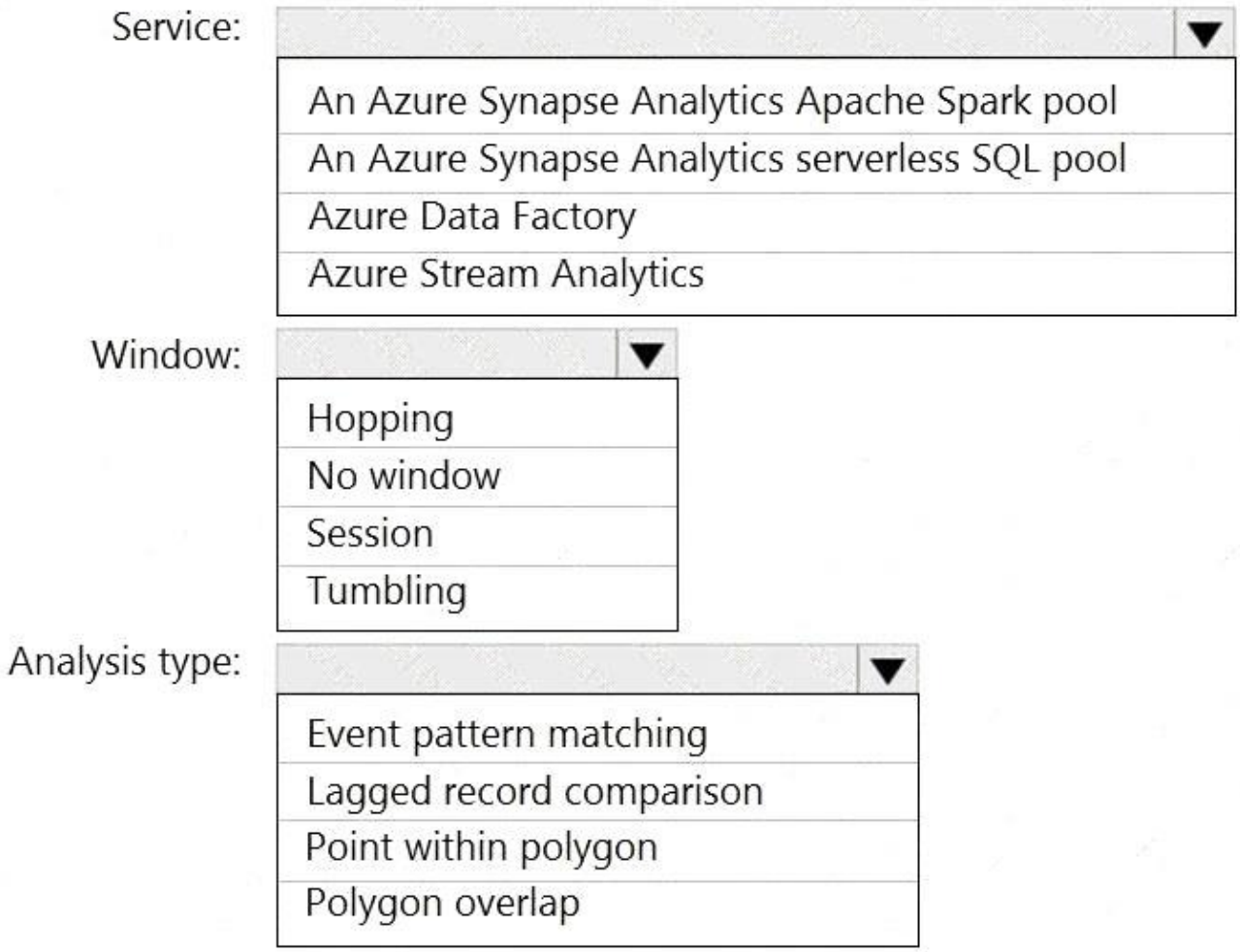
Answer Area

Statements	Yes	No
storage1 provides a hierarchical namespace.	○	○
Files from March 2025 will be included.	○	○
Only files that have a Parquet file extension will be included.	○	○

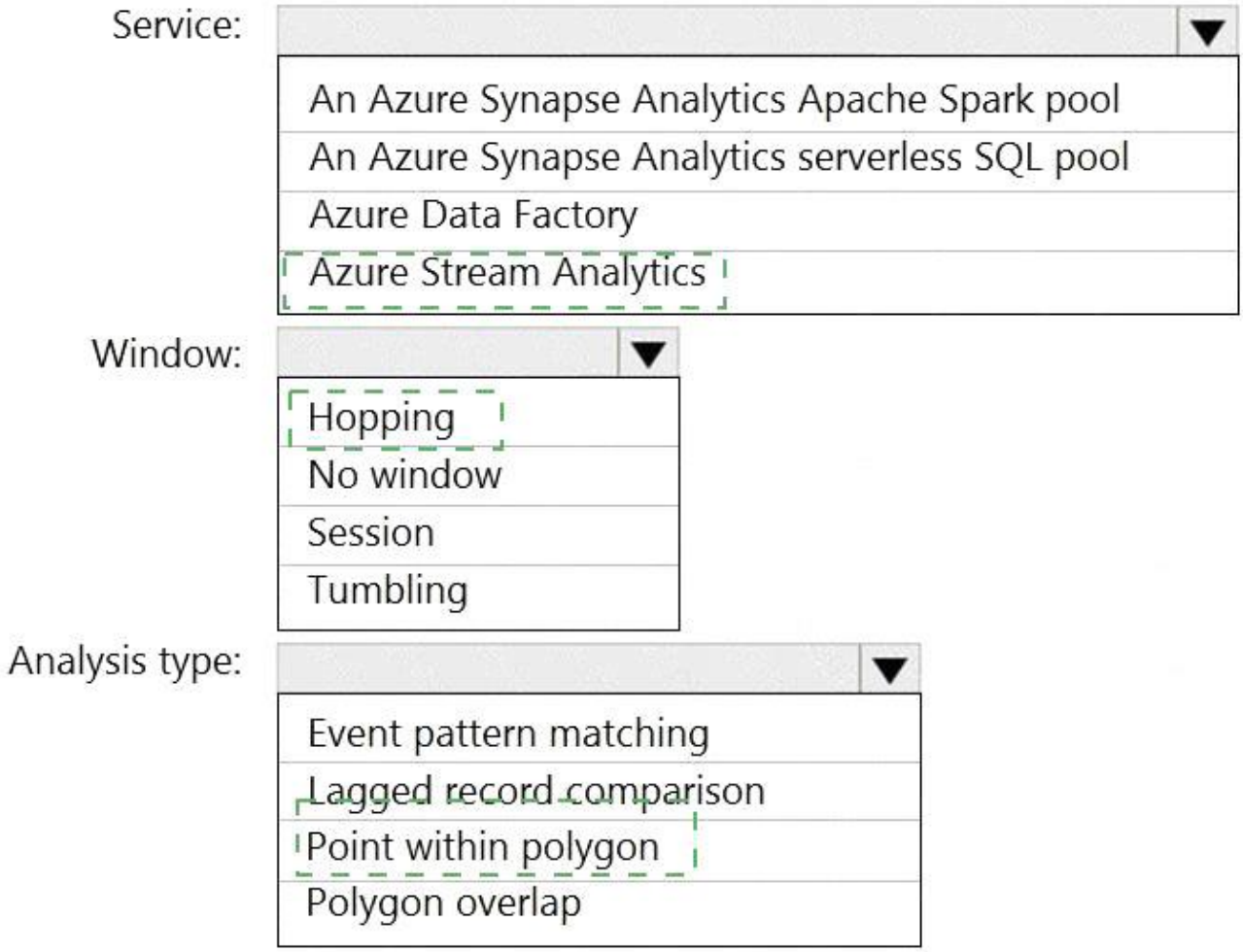
Explanation:
Storage1 provides a hierarchical namespace: Yes
Files from March 2025 will be included: No
Only files that have a Parquet file extension will be included: Yes

Query Breakdown
* Data Source:
* The OPENROWSET function queries data stored in Azure Data Lake Storage Gen2 (storage1) using the serverless SQL pool in Synapse Analytics.
* The data is stored in Parquet files in the folder structure data/orders/year=YYYY/month=MM/.
* Query Filter:
* The filter conditions in the query are:
* customerorders.filepath(1) = '2024': Limits the query to files in the folder year=2024.
* customerorders.filepath(2) IN ('3', '4'): Limits the query to files in the subfolders month=3 or month=4.
* File Format:
* The FORMAT = 'parquet' clause specifies that only Parquet files will be queried.
Statements Analysis
* Storage1 provides a hierarchical namespace.answer: Yes
* Azure Data Lake Storage Gen2 supports a hierarchical namespace, which enables folder-based organization.
* The folder structure (e.g., data/orders/year=2024/month=3/) demonstrates the use of a hierarchical namespace.
* Files from March 2025 will be included.answer: No
* The query explicitly filters for year=2024, so files from 2025 will not be included in the results.
* Only files that have a Parquet file extension will be included.answer: Yes
* The FORMAT = 'parquet' clause in the query ensures that only Parquet files are queried. Files with other extensions (e.g., .csv or .json) will not be included.

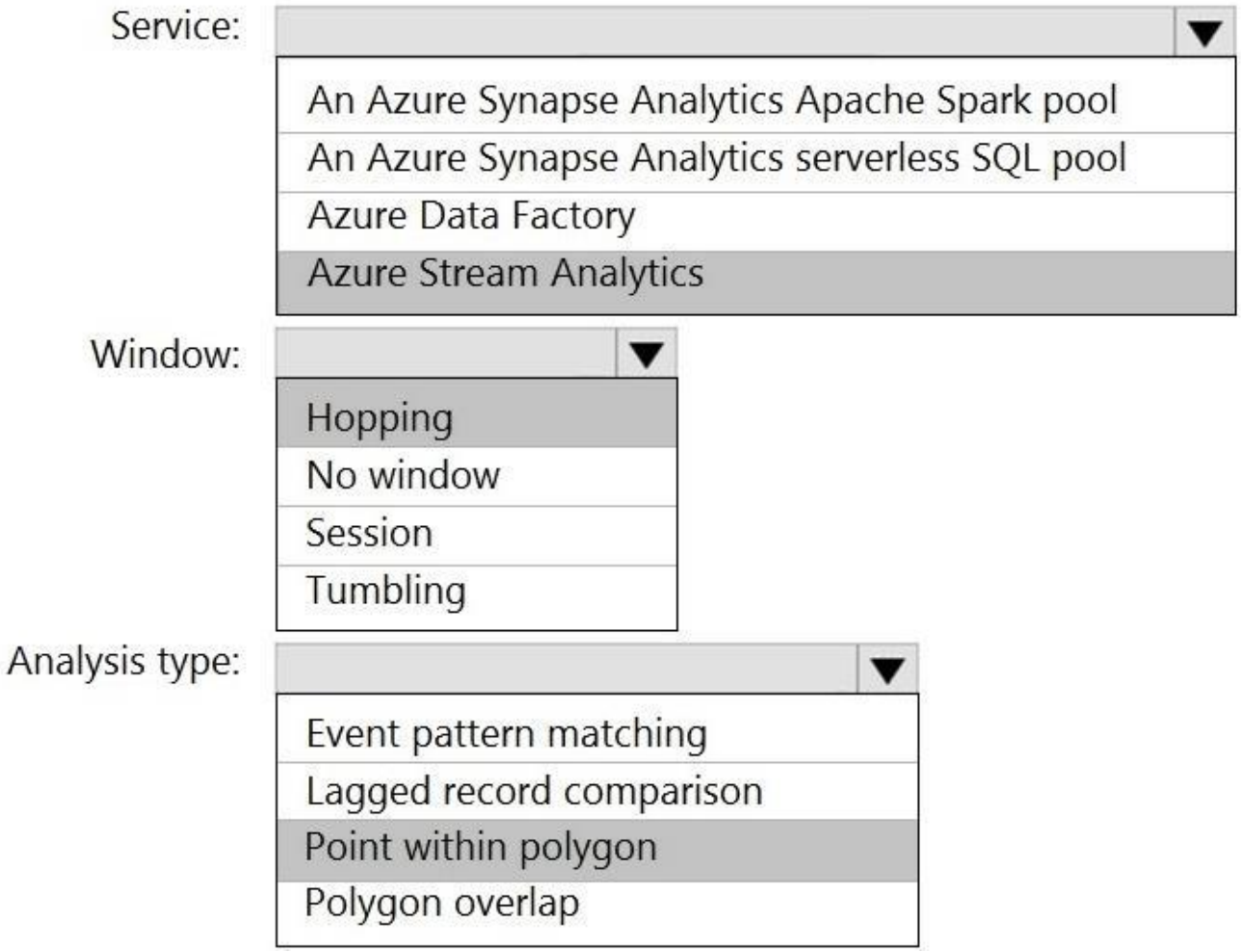
NO.28 You are designing a monitoring solution for a fleet of 500 vehicles. Each vehicle has a GPS tracking device that sends data to an Azure event hub once per minute.
You have a CSV file in an Azure Data Lake Storage Gen2 container. The file maintains the expected geographical area in which each vehicle 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.


Service:
An Azure Synapse Analytics Apache Spark pool
An Azure Synapse Analytics serverless SQL pool
Azure Data Factory
Azure Stream Analytics
Window:
Hopping
No window
Session
Tumbling
Analysis type:
Event pattern matching
Lagged record comparison
Point within polygon
Polygon overlap

Answer:



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

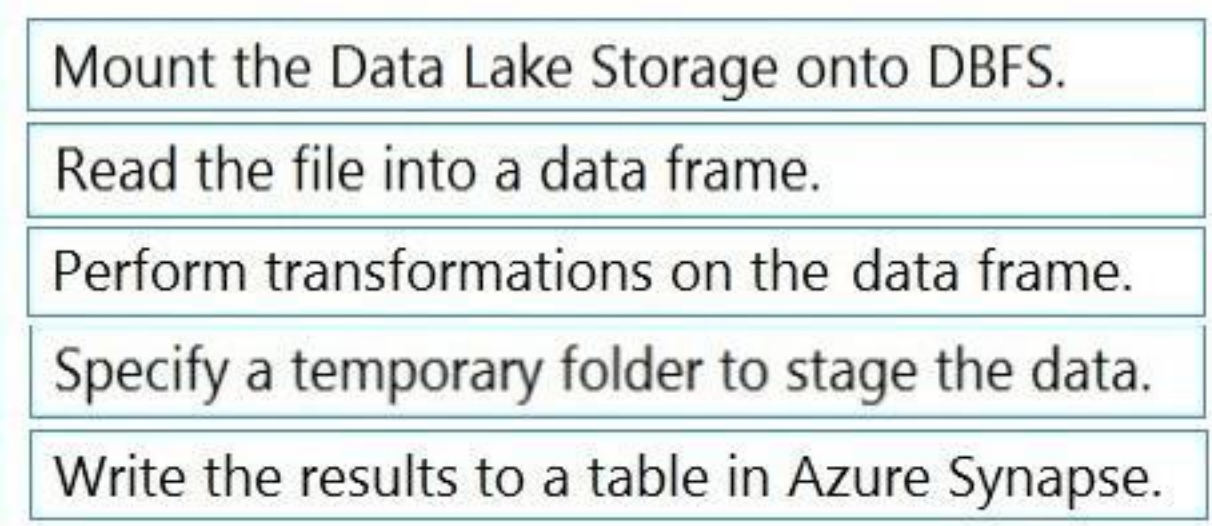
NO.29 You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers. The file contains two attributes named FirstName and LastName.
You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the FirstName and LastName values.
You create the following components:
A destination table in Azure Synapse
An Azure Blob storage container
A service principal
In which order should you perform the actions? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

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

Answer:

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

Explanation:



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

NO.30 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

NO.31 You have an Azure Data lake Storage account that contains a staging zone.
You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes an Azure Databricks notebook, and then inserts the data into the data warehouse.
Dow this meet the goal?

A. Yes

B. No

Answer: B

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

NO.32 You are designing an Azure Data Lake Storage solution that will transform raw JSON files for use in an analytical workload.
You need to recommend a format for the transformed files. The solution must meet the following

requirements:
Contain information about the data types of each column in the files.
Support querying a subset of columns in the files.
Support read-heavy analytical workloads.
Minimize the file size.
What should you recommend?
A. JSON
B. CSV
C. Apache Avro
D. Apache Parquet
Answer: D
Explanation:
Parquet, an open-source file format for Hadoop, stores nested data structures in a flat columnar format.
Compared to a traditional approach where data is stored in a row-oriented approach, Parquet file format is more efficient in terms of storage and performance.
It is especially good for queries that read particular columns from a "wide" (with many columns) table since only needed columns are read, and IO is minimized.
Reference: https://www.clairvoyant.ai/blog/big-data-file-formats

NO.33 Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values.
75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly.
Solution: You convert the files to compressed delimited text files.
Does this meet the goal?
A. Yes
B. No
Answer: A
Explanation:
All file formats have different performance characteristics. For the fastest load, use compressed delimited text files.
Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

NO.34 You have an Apache Spark DataFrame named temperatures. A sample of the data is shown in the following table.

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

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

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

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

Values

CAST
COLLATE
CONVERT
FLATTEN
PIVOT
UNPIVOT

Answer Area

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

Answer:

Values

CAST
COLLATE
CONVERT
FLATTEN
PIVOT
UNPIVOT

Answer Area

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

Explanation:

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

Box 1: PIVOT

PIVOT rotates a table-valued expression by turning the unique values from one column in the expression into multiple columns in the output. And PIVOT runs aggregations where they're required on any remaining column values that are wanted in the final output.

Reference:

https://learnsql.com/cookbook/how-to-convert-an-integer-to-a-decimal-in-sql-server/

https://docs.microsoft.com/en-us/sql/t-sql/queries/from-using-pivot-and-unpivot

NO.35 A company uses the Azure Data Lake Storage Gen2 service.

You need to design a data archiving solution that meets the following requirements:

Data that is older than five years is accessed infrequency but must be available within one second when requested.

Data that is older than seven years in NOT accessed.

Costs must be minimized while maintaining the required availability.
How should you manage the data? To answer, select the appropriate option in he answers area.
NOTE: Each correct selection is worth one point.

Answer:

Explanation:

Answer Area


Data over five years old:
Delete the blob.
Move to archive storage.
Move to cool storage.
Move to hot storage.
Data over seven years old:
Delete the blob.
Move to archive storage.
Move to cool storage.
Move to hot storage.