Microsof.DP-203.vMay-2024.by.ShinamotoHkato.139q

Number: DP-203
Passing Score: 800
Time Limit: 120
File Version: 22.0

Exam Code: DP-203
Exam Name: Data Engineering on Microsoft Azure

Case 01-Design and implement data storage

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

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest it integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment

Transactional Data

Contoso has three years of customer, transactional, operational, 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 periods. 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 an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

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 Requirements

Contoso identifies the following requirements for customer sentiment analytics:

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.

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

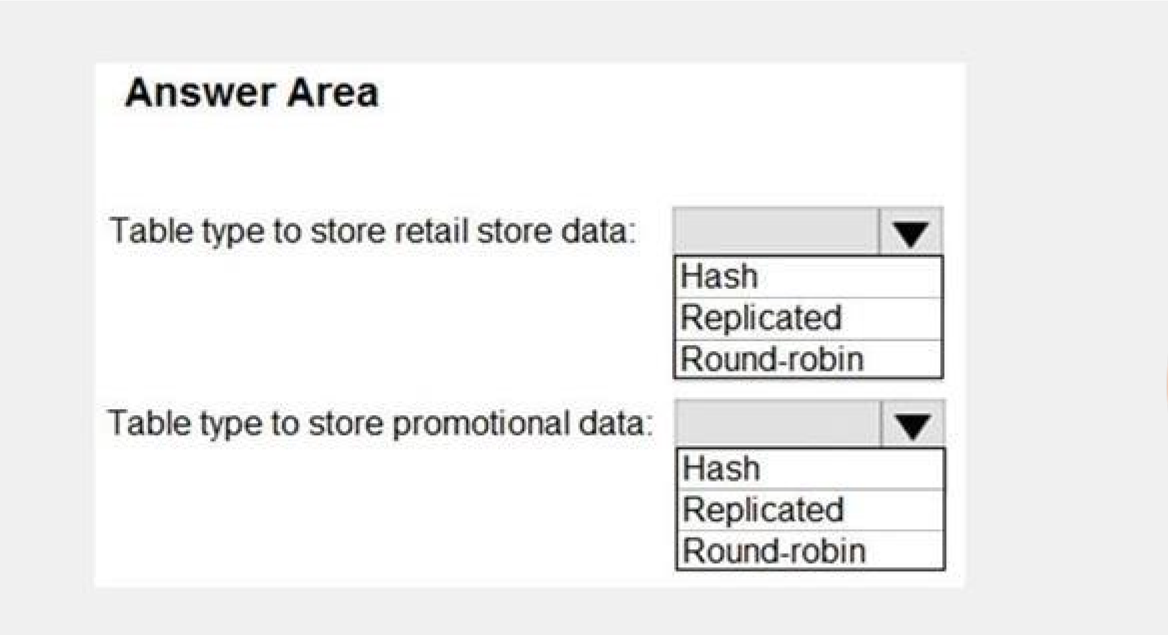
Store Twitter feeds in Azure Storage by using Event Hubs Capture. The feeds will be converted into Parquet files.

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 that 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 Synapse 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.

QUESTION 1
HOTSPOT
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.

Hot Area:

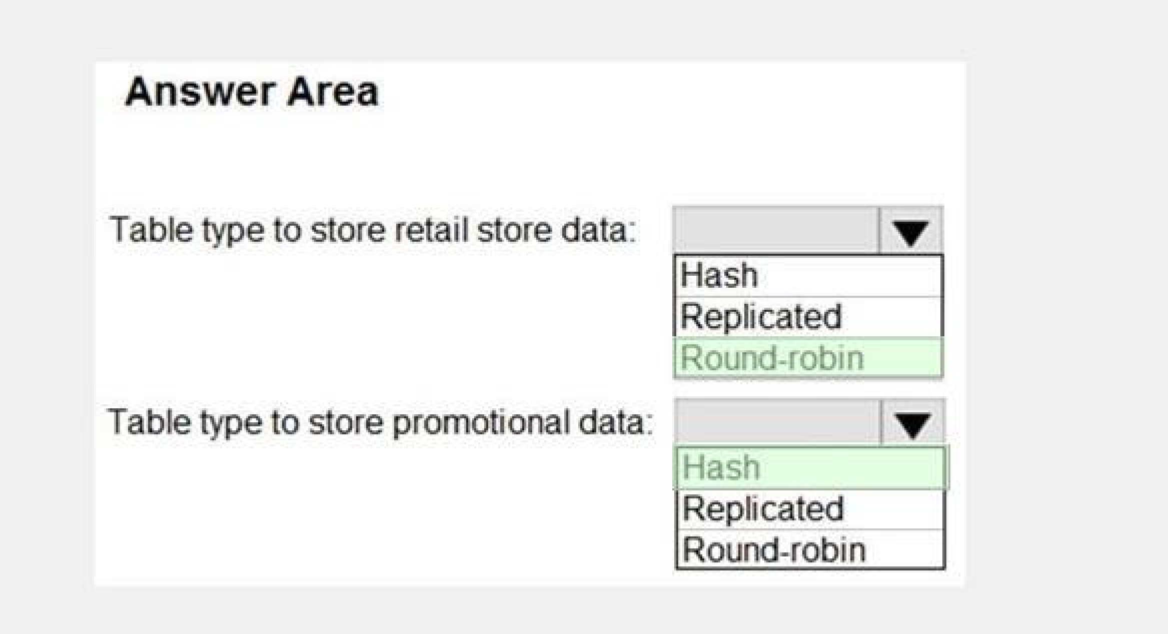


Answer Area

Table type to store retail store data:
- Hash
- Replicated
- Round-robin

Table type to store promotional data:
- Hash
- Replicated
- Round-robin

Answer Area:



Answer Area

Table type to store retail store data:
- Hash
- Replicated
- [✓ Round-robin]

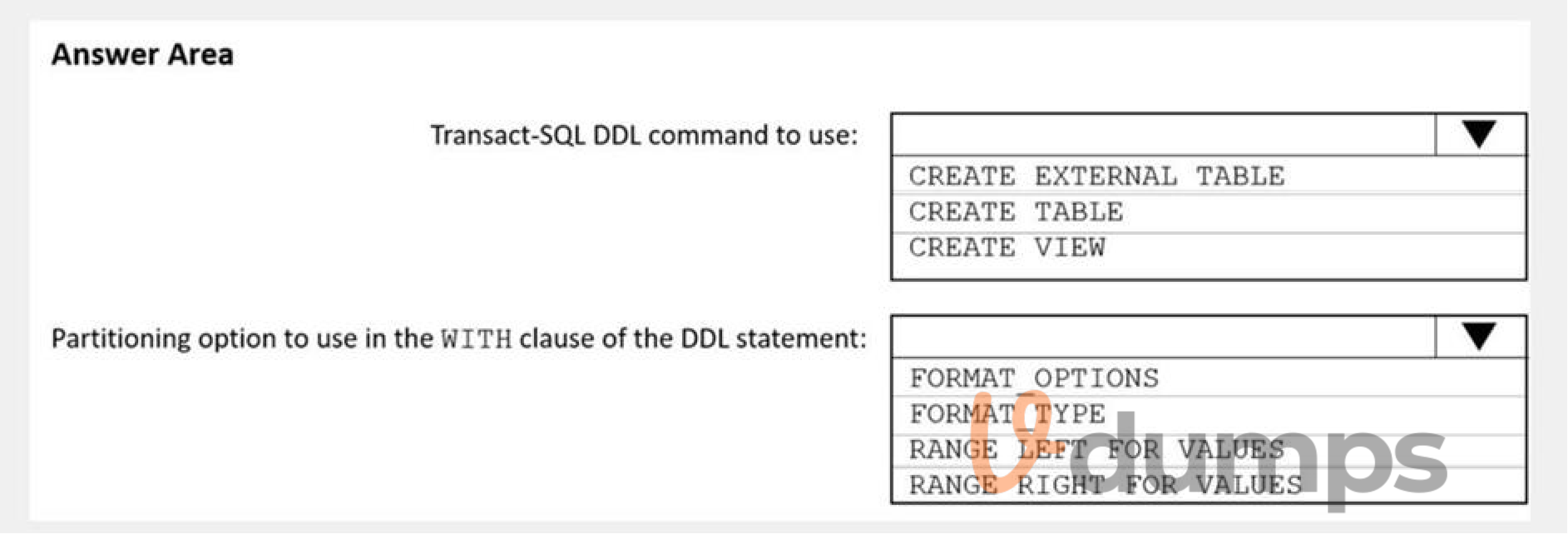
Table type to store promotional data:
- [✓ Hash]
- Replicated
- Round-robin

Section:

Explanation:
https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/design-guidance-for-replicated-tables#what-is-a-replicated-table

QUESTION 2
HOTSPOT
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 should you do? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:

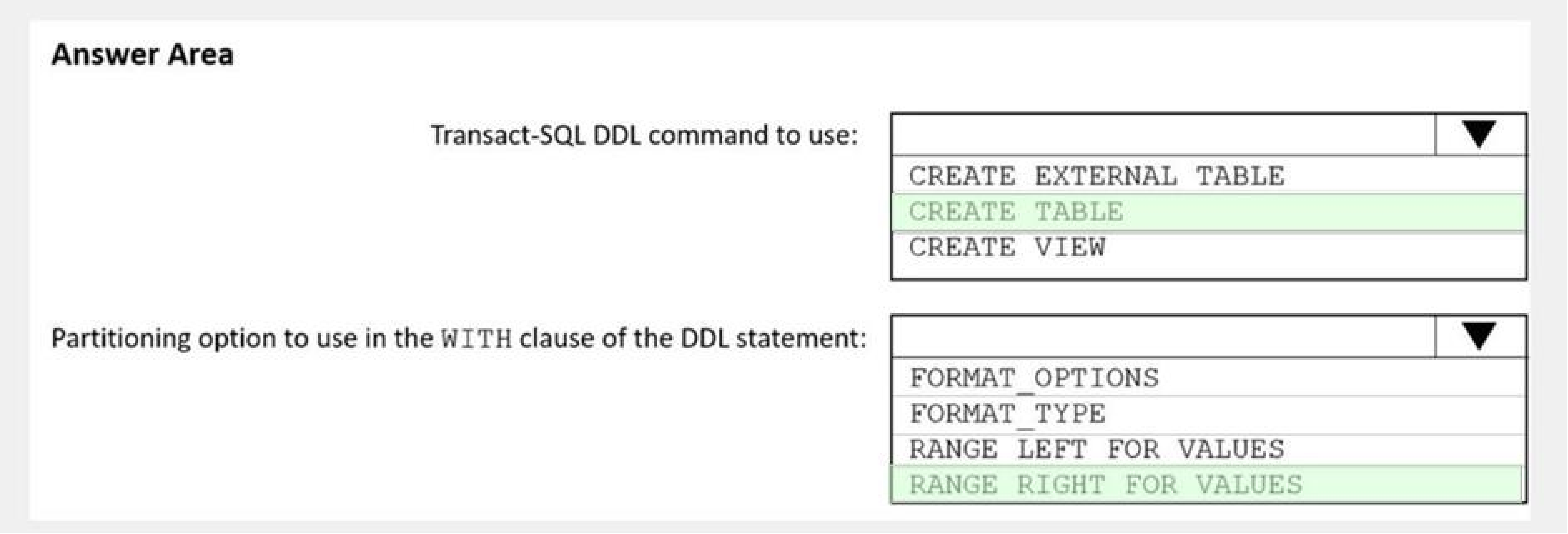


Answer Area

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

Answer Area:



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]

Section:
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

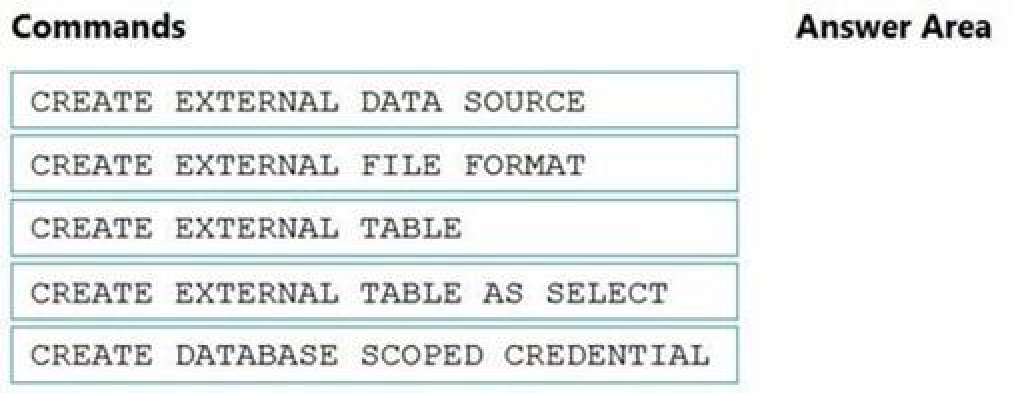
QUESTION 3
You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements. Which Azure Storage functionality should you include in the solution?

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

Correct Answer: B
Section:
Explanation:

QUESTION 4
DRAG DROP
You need to ensure that the Twitter feed data can be analyzed in the dedicated SQL pool. The solution must meet the customer sentiment analytic requirements.
Which three Transact-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.

Select and Place:

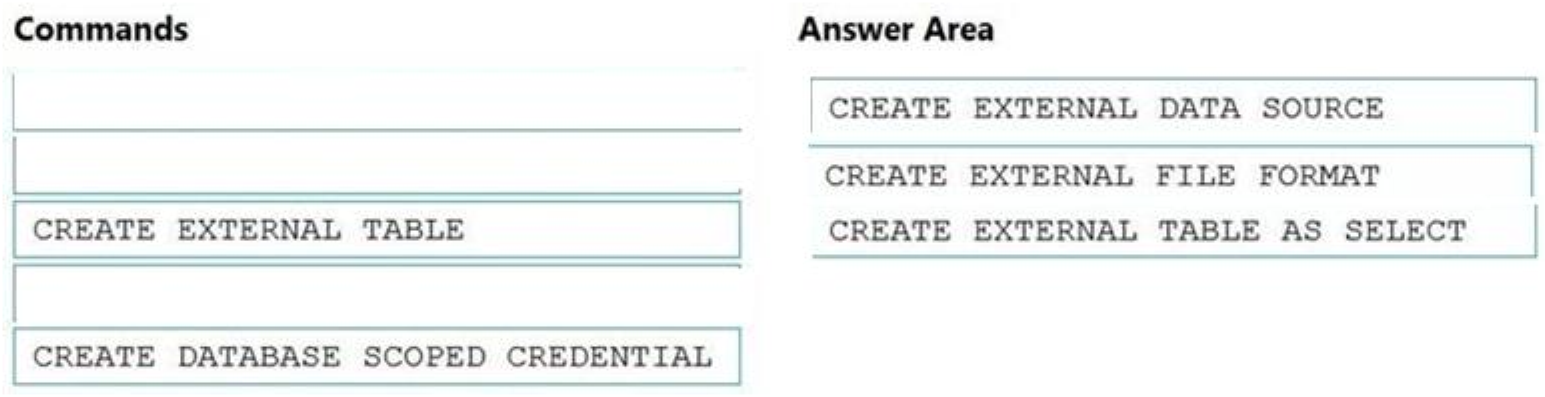


Commands

Answer Area

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

Correct Answer:



Commands

CREATE EXTERNAL TABLE

CREATE DATABASE SCOPED CREDENTIAL

Answer Area

CREATE EXTERNAL DATA SOURCE

CREATE EXTERNAL FILE FORMAT

CREATE EXTERNAL TABLE AS SELECT

Section:

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.

Incorrect Answers:

CREATE EXTERNAL TABLE

The CREATE EXTERNAL TABLE command creates an external table for Synapse SQL to access data stored in Azure Blob Storage or Azure Data Lake Storage.

Reference:

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

QUESTION 5

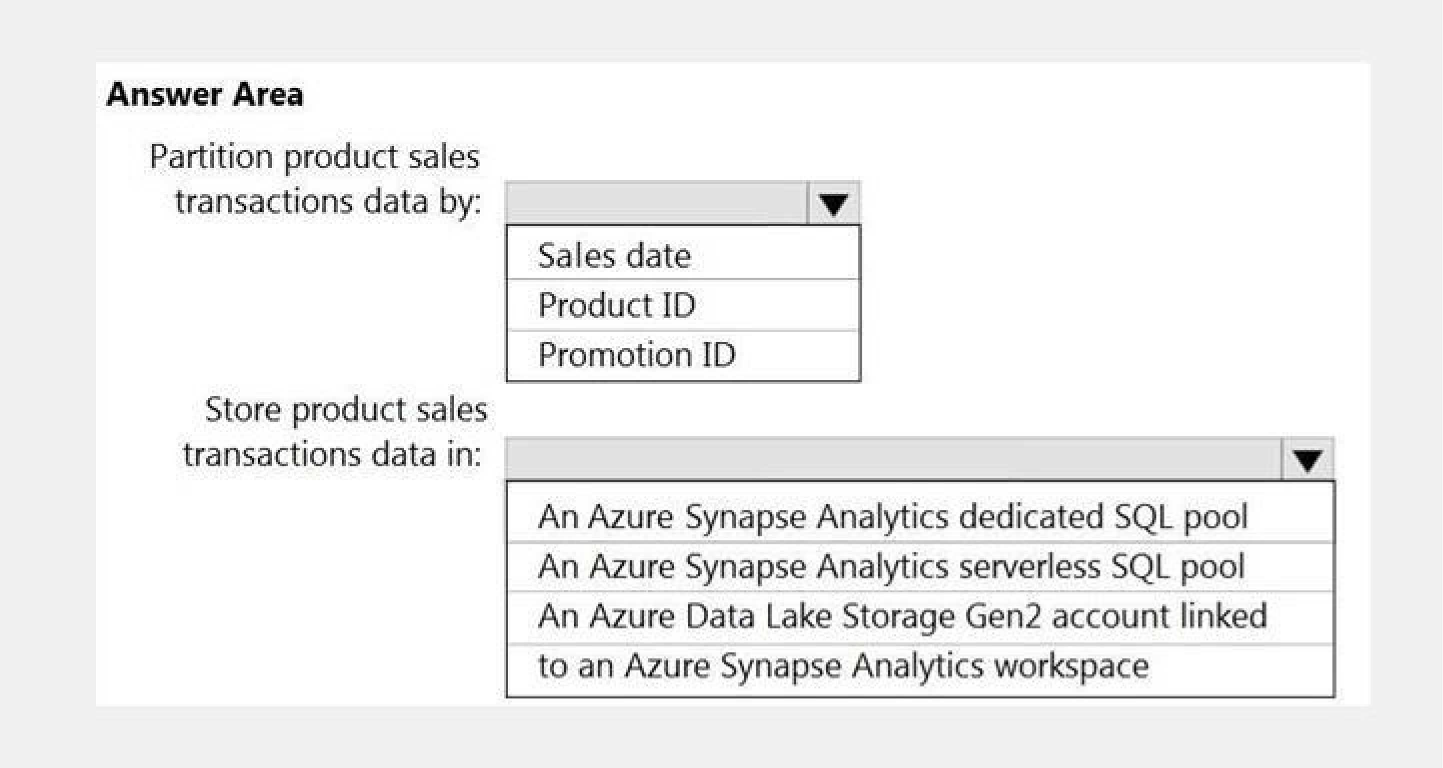
HOTSPOT

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.

Hot Area:

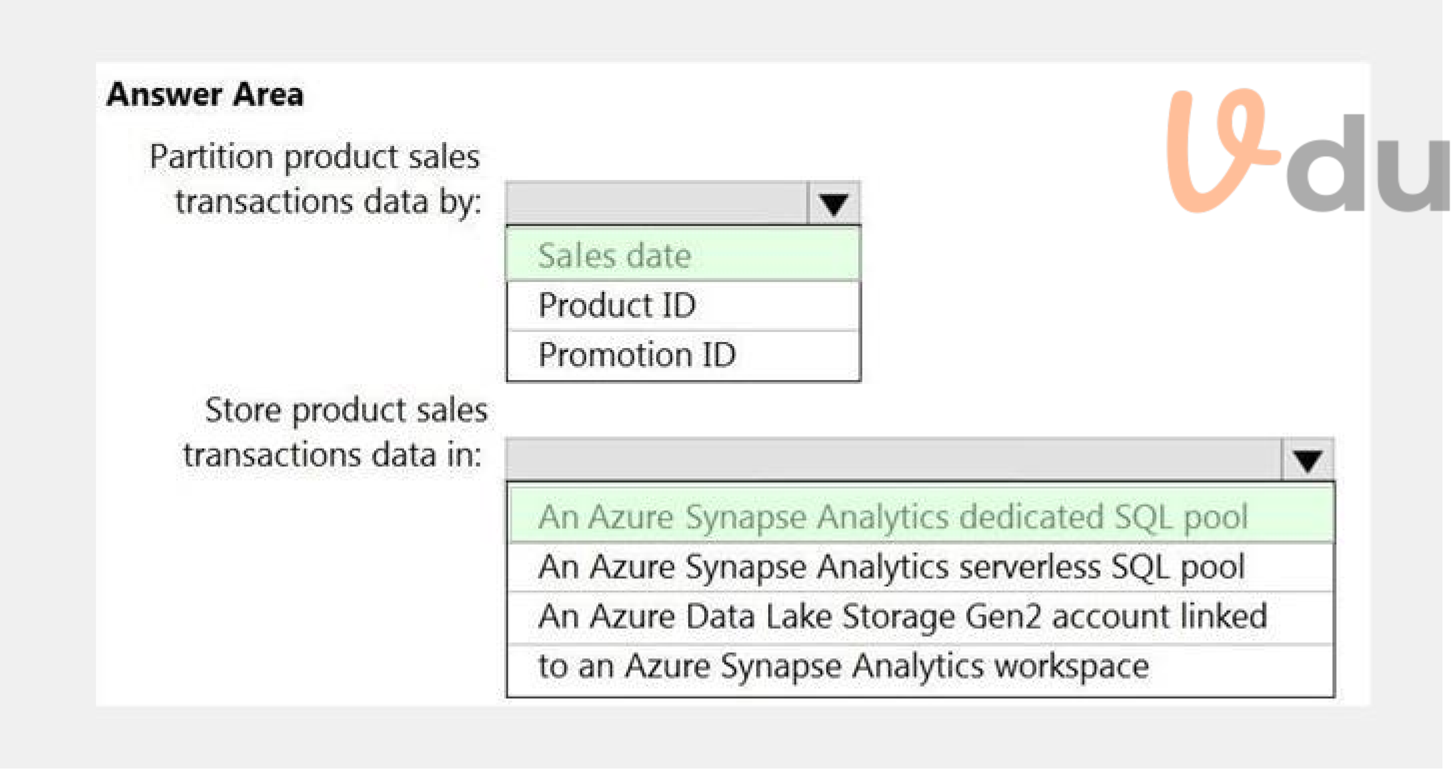


Answer Area

Partition product sales transactions data by: ▼
- Sales date
- Product ID
- Promotion ID

Store product sales transactions data in: ▼
- An Azure Synapse Analytics dedicated SQL pool
- An Azure Synapse Analytics serverless SQL pool
- An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace

Answer Area:



Answer Area

Partition product sales transactions data by: ▼
- [✓ Sales date]
- Product ID
- Promotion ID

Store product sales transactions data in: ▼
- [✓ An Azure Synapse Analytics dedicated SQL pool]
- An Azure Synapse Analytics serverless SQL pool
- An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace

Section:

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.

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.

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

QUESTION 6
You have an Azure subscription that contains a Microsoft Purview account named MP1, an Azure data factory named DF1, and a storage account named storage. MP1 is configured 10 scan storage1. DF1 is connected to MP1 and contains 3 dataset named DS1. DS1 references 2 file in storage.In DF1, you plan to create a pipeline that will process data from DS1.You need to review the schema and lineage information in MP1 for the data referenced by DS1.Which two features can you use to locate the information? Each correct answer presents a complete solution. NOTE: Each correct answer is worth one point.

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

Correct Answer: C, D
Section:
Explanation:
The search bar in the Microsoft Purview governance portal: This is a feature that allows you to search for assets in your data estate using keywords, filters, and facets. You can use the search bar to findthe files in storage1 that are referenced by DS1, and then view their schema and lineage information in the asset details page12.The search bar in Azure Data Factory Studio: This is a feature that allows you to search for datasets, linked services, pipelines, and other resources in your data factory. You can use the search bar to find DS1 in DF1, and then view its schema and lineage information in the dataset details page. You canalso click on the Open in Purview button to open the corresponding asset in MP13.The two features that can be used to locate the schema and lineage information for the data referenced by DS1 are the search bar in Azure Data Factory Studio and the search bar in the Microsoft Purview governance portal.The search bar in Azure Data Factory Studio allows you to search for the dataset DS1 and view its properties and lineage. This can help you locate information about the source and destination data stores, as well as the transformations that were applied to the data.The search bar in the Microsoft Purview governance portal allows you to search for the storage account and view its metadata, including schema and lineage information. This can help you understand the different data assets that are stored in the storage account and how they are related to each other.The Storage browser of storage1 in the Azure portal may allow you to view the files that are stored in the storage account, but it does not provide lineage or schema information for those files. Similarly, the search bar in the Azure portal may allow you to search for resources in the Azure subscription, but it does not provide detailed information about the data assets themselves.Reference:What is Azure Purview?Use Azure Data Factory Studio

QUESTION 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

Correct Answer: A
Section:
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

02-Design and implement data storage

QUESTION 1
You are designing a partition strategy for a fact table in an Azure Synapse Analytics dedicated SQL pool. The table has the following specifications:
Contain sales data for 20,000 products.
Use hash distribution on a column named ProductID.

Contain 2.4 billion records for the years 2019 and 2020.
Which number of partition ranges provides optimal compression and performance for the clustered columnstore index?

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

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

QUESTION 2
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 Storage account that contains 100 GB of files. The files contain rows of 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 modify the files to ensure that each row is more than 1 MB.
Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section:
Explanation:

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

Correct Answer: B
Section:
Explanation:
Avro supports batch and is very relevant for streaming.

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

QUESTION 4
You are designing a financial transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:
TransactionType: 40 million rows per transaction type
CustomerSegment: 4 million per customer segment
TransactionMonth: 65 million rows per month AccountType: 500 million per account type You have the following query requirements:
Analysts will most commonly analyze transactions for a given month. Transactions analysis will typically summarize transactions by transaction type, customer segment, and/or account type You need to recommend a partition strategy for the table to minimize query times. On which column should you recommend partitioning the table?

A. CustomerSegment
B. AccountType
C. TransactionType
D. TransactionMonth

Correct Answer: D
Section:

QUESTION 5
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 Databricks 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 files can be queried quickly and that the data type information is retained.
What should you recommend?

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

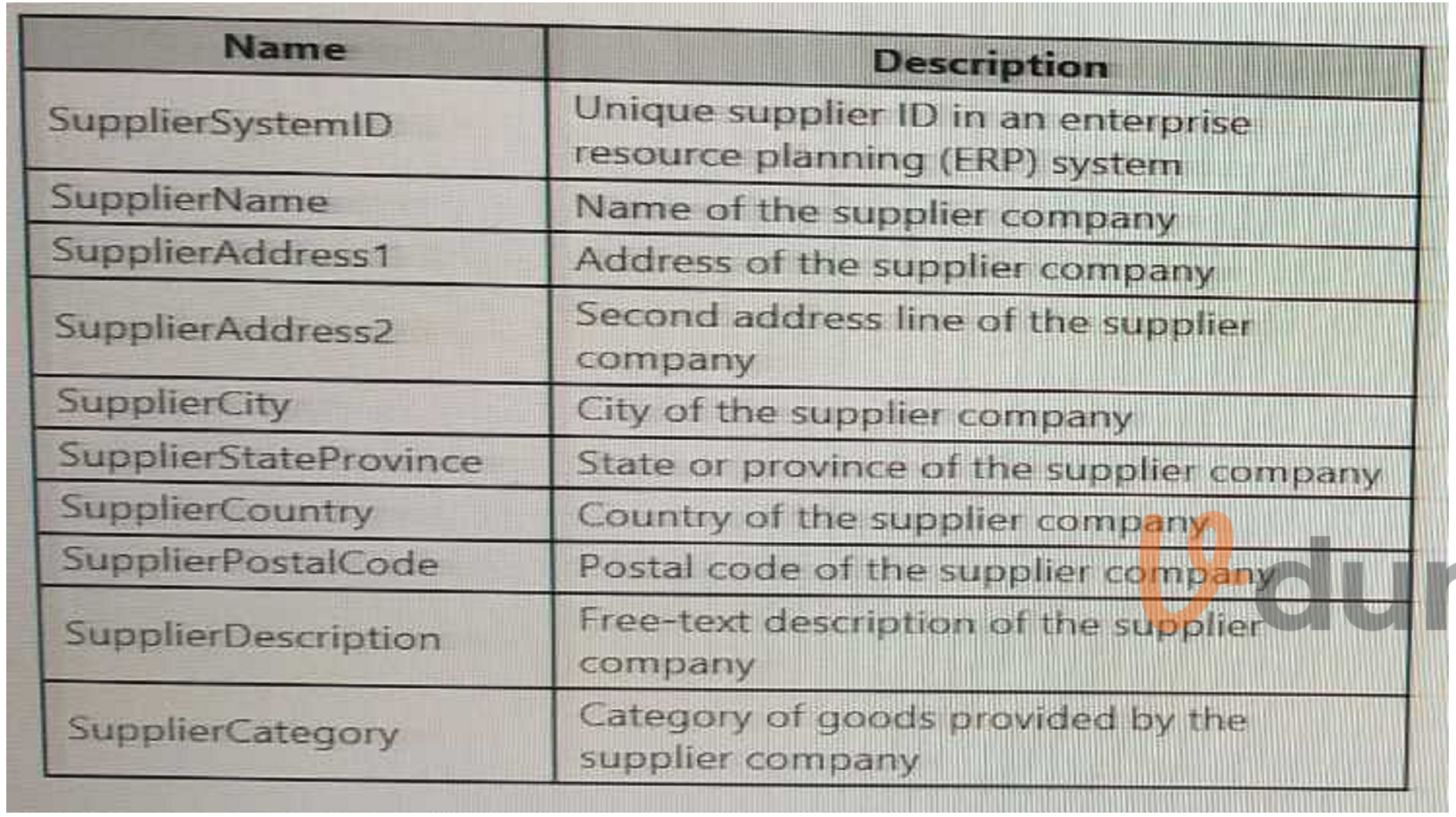
Correct Answer: B
Section:
Explanation:
Need Parquet to support both Databricks and PolyBase.
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-file-format-transact-sql

QUESTION 6
You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a partitioned fact table named dbo.Sales and a staging table named stg.Sales that has the matching table and partition definitions. You need to overwrite the content of the first partition in dbo.Sales with the content of the same partition in stg.Sales. The solution must minimize load times. What should you do?

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

Correct Answer: C
Section:
Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

QUESTION 7
You are designing a slowly changing dimension (SCD) for supplier data in an Azure Synapse Analytics dedicated SQL pool.
You plan to keep a record of changes to the available fields.
The supplier data contains the following columns.



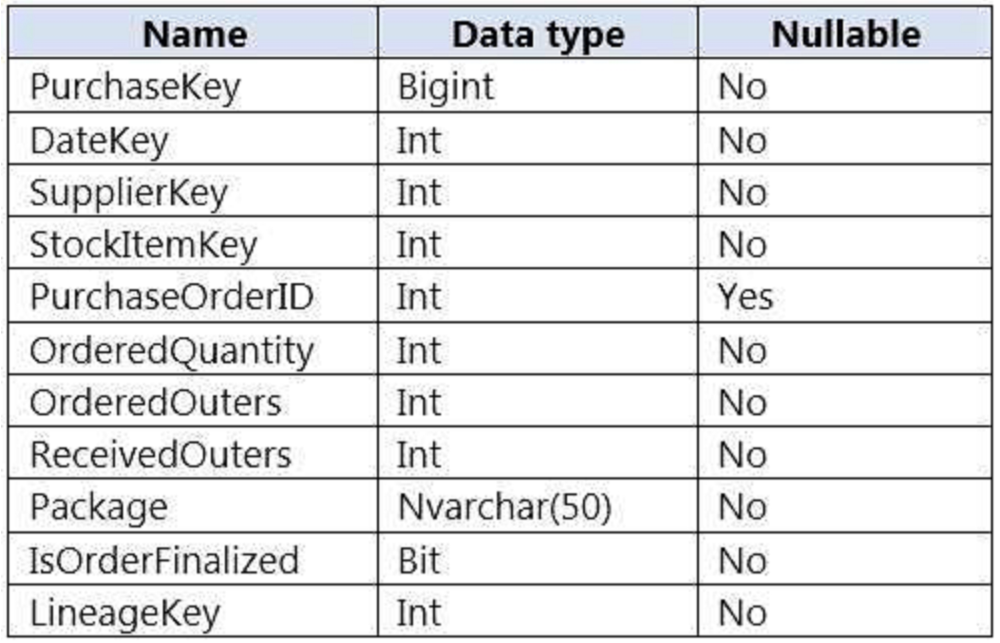
Name	Description
SupplierSystemID	Unique supplier ID in an enterprise resource planning (ERP) system
SupplierName	Name of the supplier company
SupplierAddress1	Address of the supplier company
SupplierAddress2	Second address line of the supplier company
SupplierCity	City of the supplier company
SupplierStateProvince	State or province of the supplier company
SupplierCountry	Country of the supplier company
SupplierPostalCode	Postal code of the supplier company
SupplierDescription	Free-text description of the supplier company
SupplierCategory	Category of goods provided by the supplier company

Which three additional columns should you add to the data to create a Type 2 SCD? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. surrogate primary key
B. effective start date
C. business key
D. last modified date
E. effective end date
F. foreign key

Correct Answer: A, B, E
Section:
Explanation:
https://learn.microsoft.com/en-us/training/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types

QUESTION 8
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.

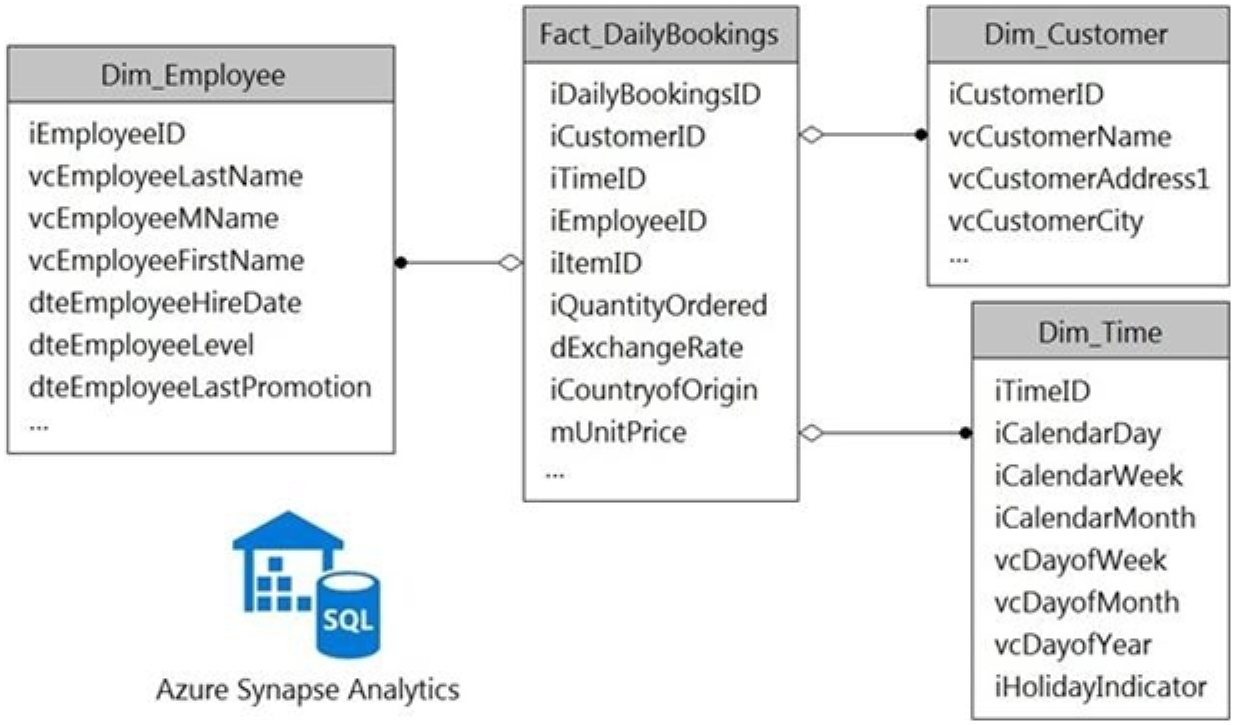
```
SELECT
SupplierKey, StockItemKey, COUNT(*)
FROM FactPurchase
WHERE DateKey >= 20210101
AND DateKey <= 20210131
GROUP By SupplierKey, StockItemKey
```

Which table distribution will minimize query times?

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

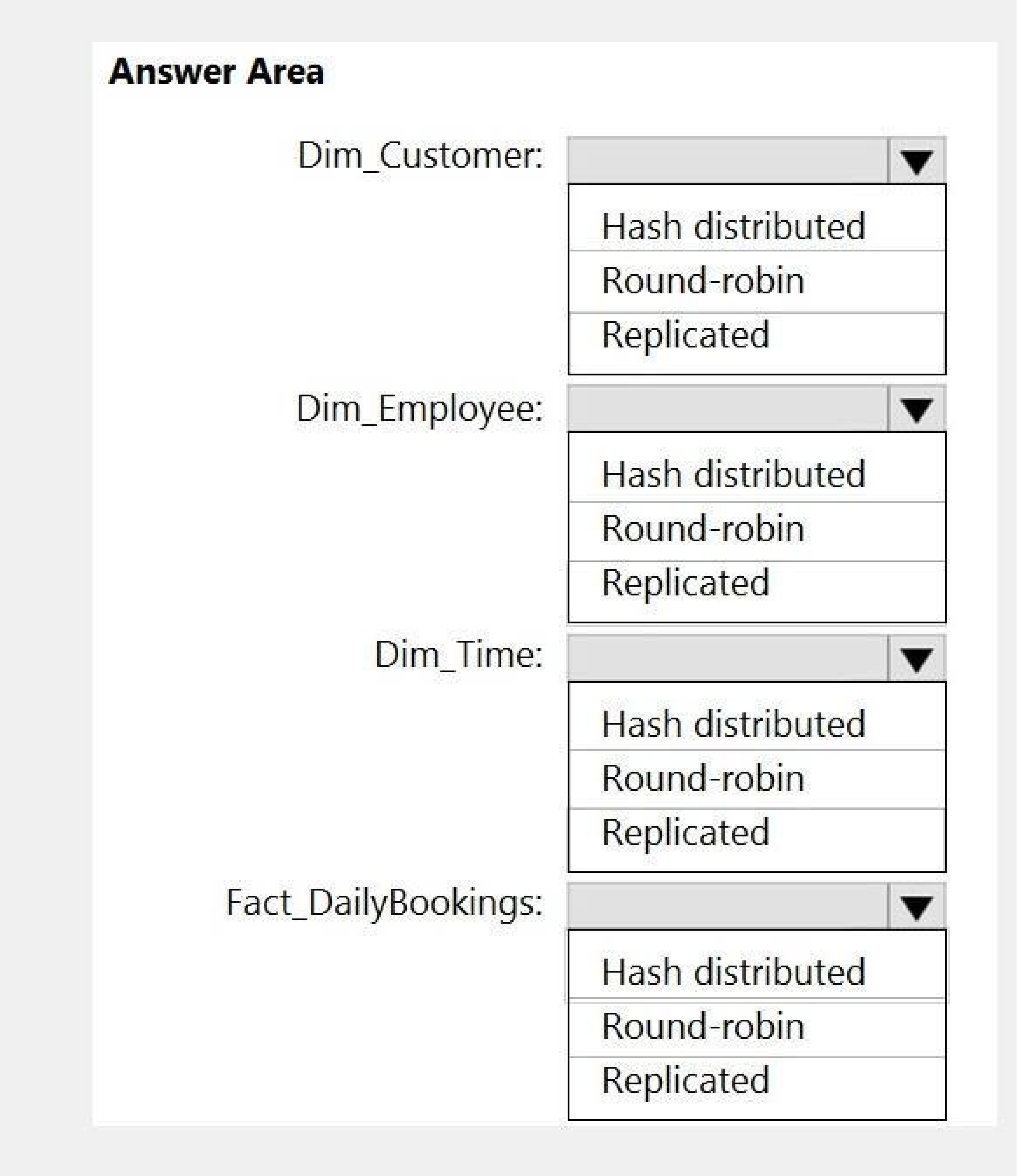
Correct Answer: B
Section:
Explanation:
Hash-distributed tables improve query performance on large fact tables, and are the focus of this article. Round-robin tables are useful for improving loading speed. Incorrect:
Not D: Do not use a date column. . All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

QUESTION 9
HOTSPOT
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. The dimension tables will be relatively static with very few data inserts and updates.
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.

Hot Area:



Answer Area

Dim_Customer: ▼

- Hash distributed
- Round-robin
- Replicated

Dim_Employee: ▼

- Hash distributed
- Round-robin
- Replicated

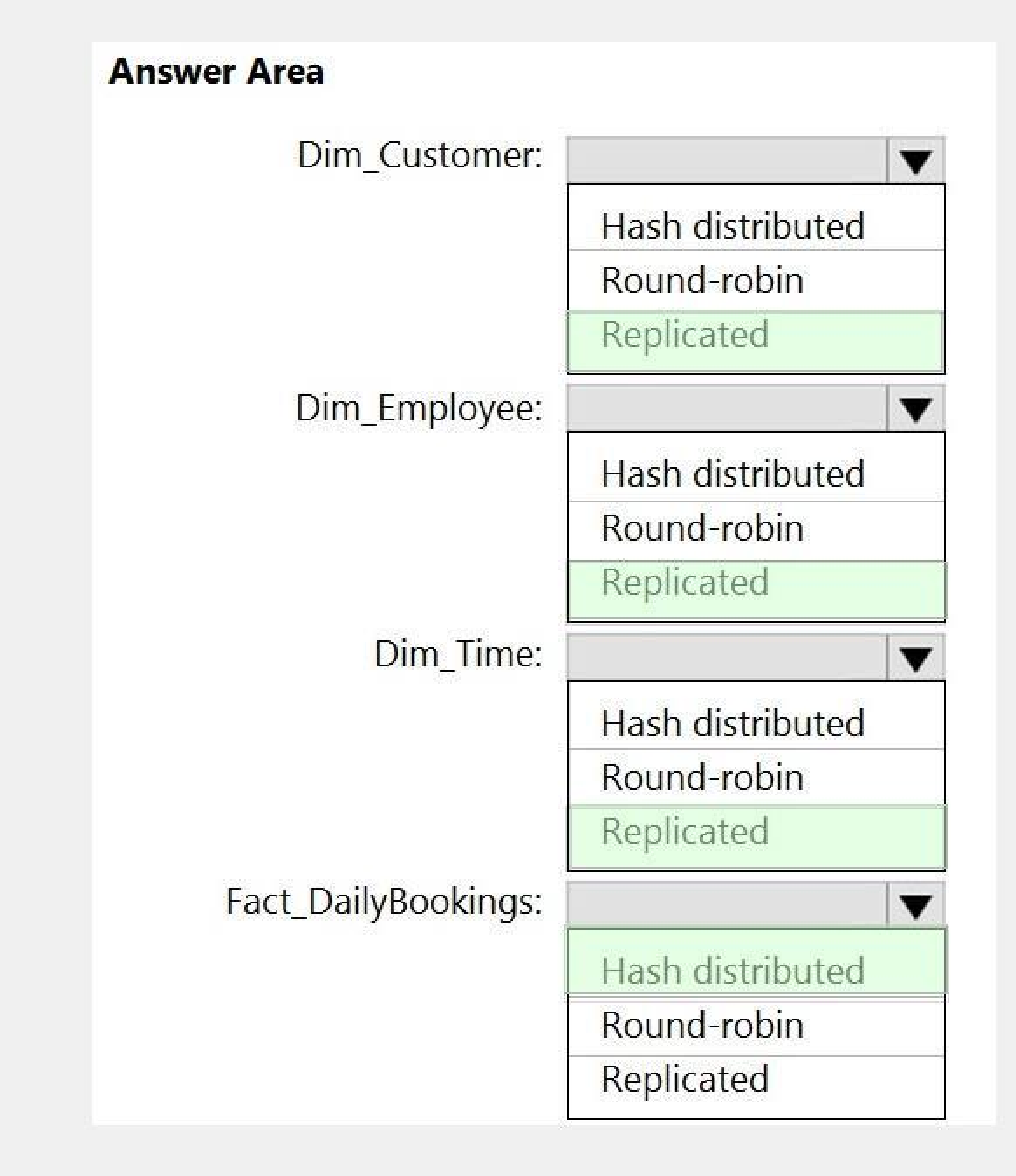
Dim_Time: ▼

- Hash distributed
- Round-robin
- Replicated

Fact_DailyBookings: ▼

- Hash distributed
- Round-robin
- Replicated

Answer Area:



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

Section:
Explanation:
Box 1: Replicated
Replicated tables are ideal for small star-schema dimension tables, because the fact table is often distributed on a column that is not compatible with the connected dimension tables. If this case applies to your schema, consider changing small dimension tables currently implemented as round-robin to replicated.
Box 2: Replicated
Box 3: Replicated
Box 4: Hash-distributed
For Fact tables use hash-distribution with clustered columnstore index. Performance improves when two hash tables are joined on the same distribution column.
Reference:
https://azure.microsoft.com/en-us/updates/reduce-data-movement-and-make-your-queries-more-efficient-with-the-general-availability-of-replicated-tables/
https://azure.microsoft.com/en-us/blog/replicated-tables-now-generally-available-in-azure-sql-data-warehouse/

QUESTION 10
HOTSPOT

You have an Azure Data Lake Storage Gen2 container.

Data is ingested into the container, and then transformed by a data integration application. The data is NOT modified after that. Users can read files in the container but cannot modify the files.

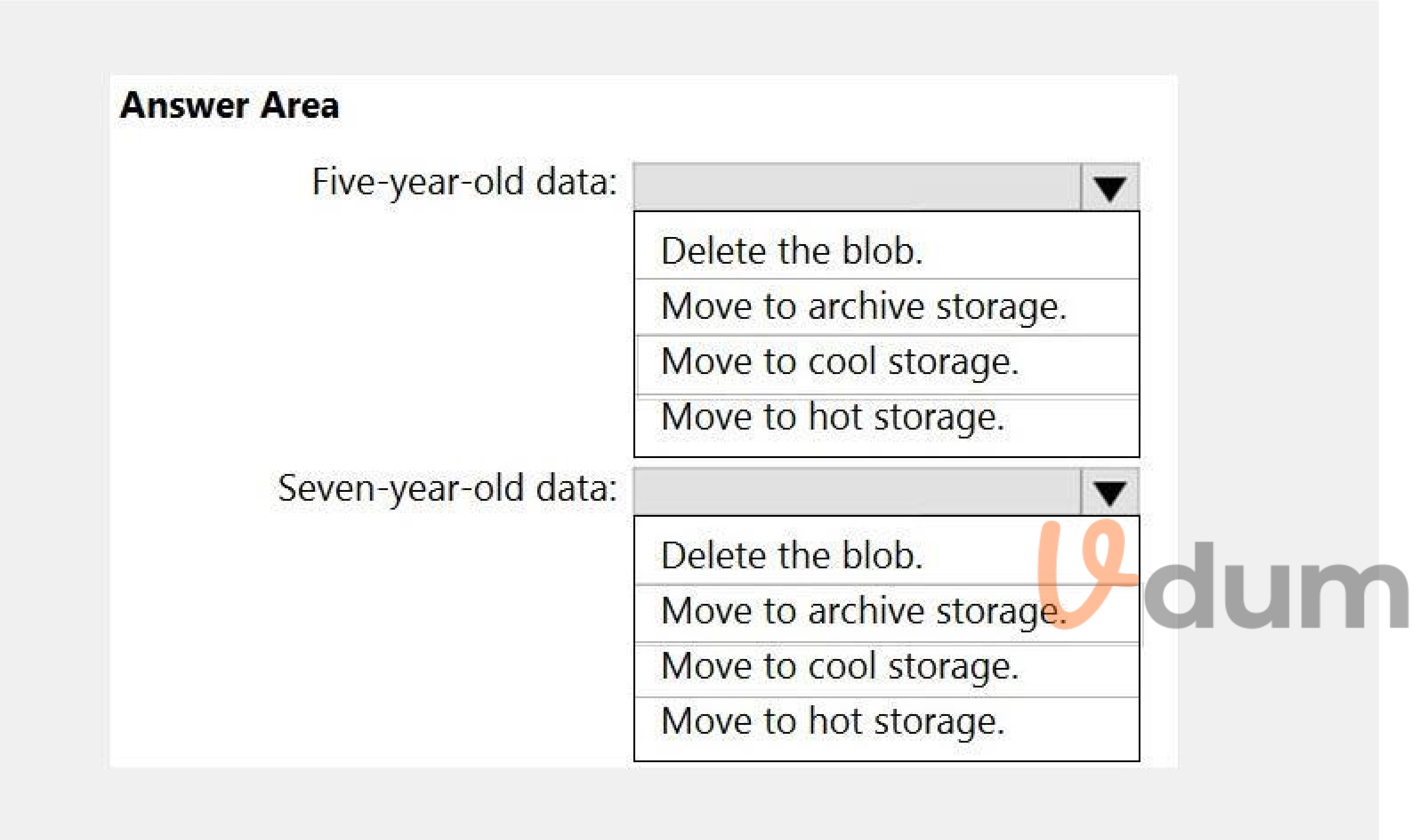
You need to design a data archiving solution that meets the following requirements:

New data is accessed frequently and must be available as quickly as possible. Data that is older than five years is accessed infrequently but must be available within one second when requested. Data that is older than seven years is NOT accessed. After seven years, the data must be persisted at the lowest cost possible. Costs must be minimized while maintaining the required availability.

How should you manage the data? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point

Hot Area:



Answer Area

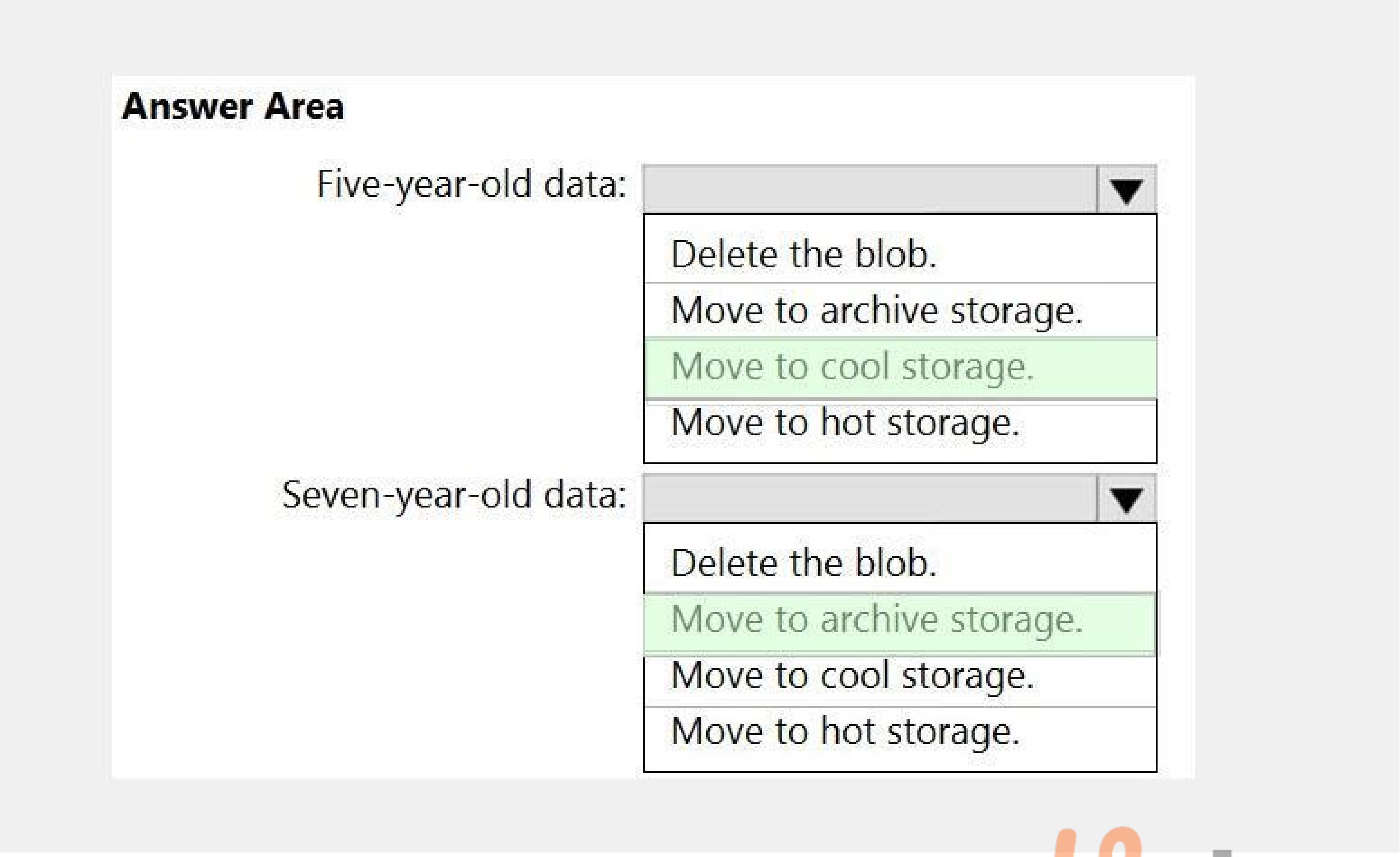
Five-year-old data:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Seven-year-old data:

- Delete the blob.
- Move to archive storage.
- Move to cool storage.
- Move to hot storage.

Answer Area:



Answer Area

Five-year-old data:

- Delete the blob.
- Move to archive storage.
- [✓ Move to cool storage.]
- Move to hot storage.

Seven-year-old data:

- Delete the blob.
- [✓ Move to archive storage.]
- Move to cool storage.
- Move to hot storage.

Section:
Explanation:
HOTSPOT
You have an Azure Data Lake Storage Gen2 container.
Data is ingested into the container, and then transformed by a data integration application. The data is NOT modified after that. Users can read files in the container but cannot modify the files.
You need to design a data archiving solution that meets the following requirements:
New data is accessed frequently and must be available as quickly as possible. Data that is older than five years is accessed infrequently but must be available within one second when requested. Data that is older than seven years is NOT accessed. After seven years, the data must be persisted at the lowest cost possible. Costs must be minimized while maintaining the required availability.
How should you manage the data? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point

QUESTION 11

You are designing the folder structure for an Azure Data Lake Storage Gen2 container. Users will query data by using a variety of services including Azure Databricks and Azure Synapse Analytics serverless SQL pools. The data will be secured by subject area. Most queries will include data from the current year or current month.
Which folder structure should you recommend to support fast queries and simplified folder security?

A. /{SubjectArea}/{DataSource}/{DD}/{MM}/{YYYY}/{FileData}_{YYYY}_{MM}_{DD}.csv
B. /{DD}/{MM}/{YYYY}/{SubjectArea}/{DataSource}/{FileData}_{YYYY}_{MM}_{DD}.csv
C. /{YYYY}/{MM}/{DD}/{SubjectArea}/{DataSource}/{FileData}_{YYYY}_{MM}_{DD}.csv
D. /{SubjectArea}/{DataSource}/{YYYY}/{MM}/{DD}/{FileData}_{YYYY}_{MM}_{DD}.csv

Correct Answer: D
Section:
Explanation:
There's an important reason to put the date at the end of the directory structure. If you want to lock down certain regions or subject matters to users/groups, then you can easily do so with the POSIX permissions. Otherwise,

if there was a need to restrict a certain security group to viewing just the UK data or certain planes, with the date structure in front a separate permission would be required for numerous directories under every hour directory. Additionally, having the date structure in front would exponentially increase the number of directories as time went on. Note: In IoT workloads, there can be a great deal of data being landed in the data store that spans across numerous products, devices, organizations, and customers. It's important to pre-plan the directory layout for organization, security, and efficient processing of the data for down-stream consumers. A general template to consider might be the following layout: {Region}/{SubjectMatter(s)}/{yyyy}/{mm}/{dd}/{hh}/

QUESTION 12
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

Correct Answer: D
Section:
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 13
You have an enterprise-wide Azure Data Lake Storage Gen2 account. The data lake is accessible only through an Azure virtual network named VNET1. You are building a SQL pool in Azure Synapse that will use data from the data lake. Your company has a sales team. All the members of the sales team are in an Azure Active Directory group named Sales. POSIX controls are used to assign the Sales group access to the files in the data lake. You plan to load data to the SQL pool every hour.
You need to ensure that the SQL pool can load the sales data from the data lake. Which three actions should you perform? Each correct answer presents part of the solution. NOTE: Each area selection is worth one point.

A. Add the managed identity to the Sales group.
B. Use the managed identity as the credentials for the data load process.
C. Create a shared access signature (SAS).
D. Add your Azure Active Directory (Azure AD) account to the Sales group.
E. Use the shared access signature (SAS) as the credentials for the data load process.
F. Create a managed identity.

Correct Answer: B, D, F
Section:
Explanation:
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-identity

QUESTION 14
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 Storage account that contains 100 GB of files. The files contain rows of 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

Correct Answer: A
Section:
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

QUESTION 15

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 Storage account that contains 100 GB of files. The files contain rows of 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 copy the files to a table that has a columnstore index. Does this meet the goal?

A. Yes

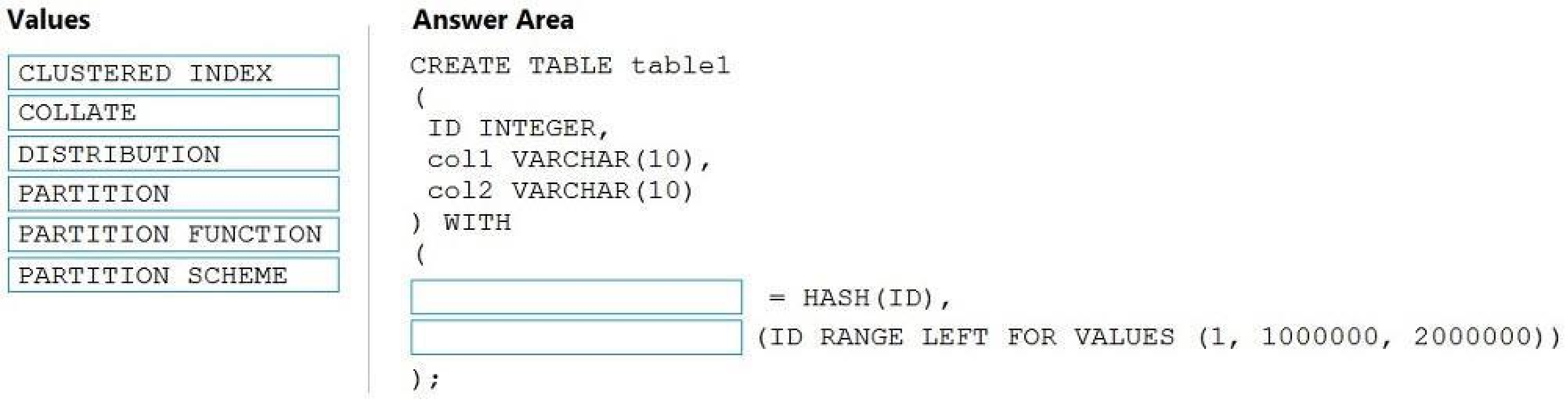
B. No

Correct Answer: B
Section:
Explanation:
Instead convert the files to compressed delimited text files.
Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

QUESTION 16

DRAG DROP
You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.
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.

Select and Place:



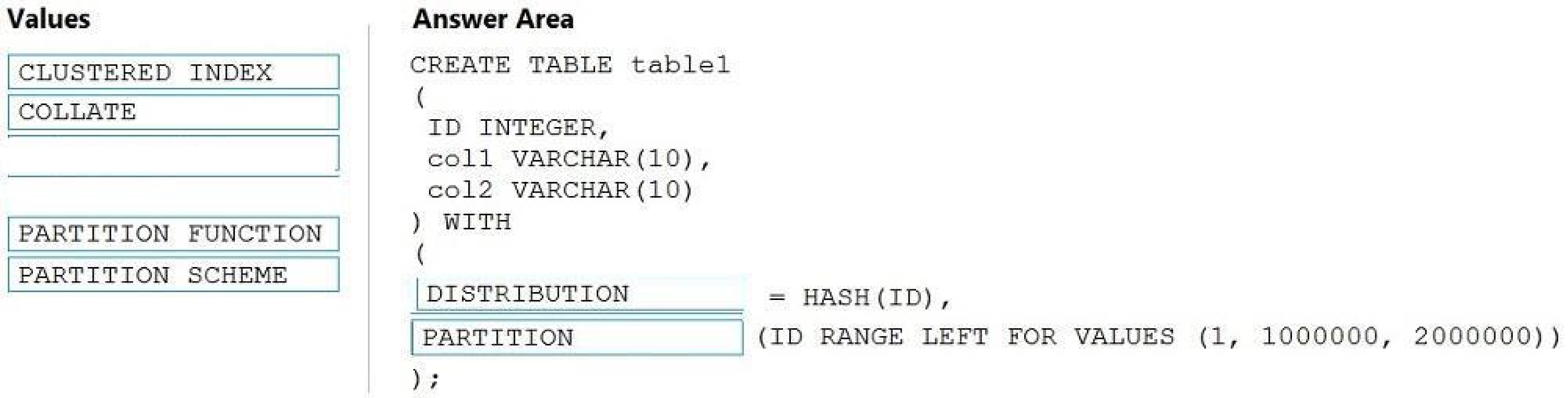
Values

- CLUSTERED INDEX
- COLLATE
- DISTRIBUTION
- PARTITION
- PARTITION FUNCTION
- PARTITION SCHEME

Answer Area

```
CREATE TABLE table1
(
 ID INTEGER,
 col1 VARCHAR(10),
 col2 VARCHAR(10)
) WITH
(
[                    ] = HASH(ID),
[                    ] (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
);
```

Correct Answer:



Values

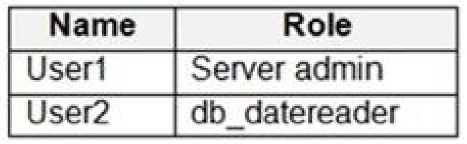
- CLUSTERED INDEX
- COLLATE
- PARTITION FUNCTION
- PARTITION SCHEME

Answer Area

```
CREATE TABLE table1
(
 ID INTEGER,
 col1 VARCHAR(10),
 col2 VARCHAR(10)
) WITH
(
 DISTRIBUTION  = HASH(ID),
PARTITION  (ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
);
```

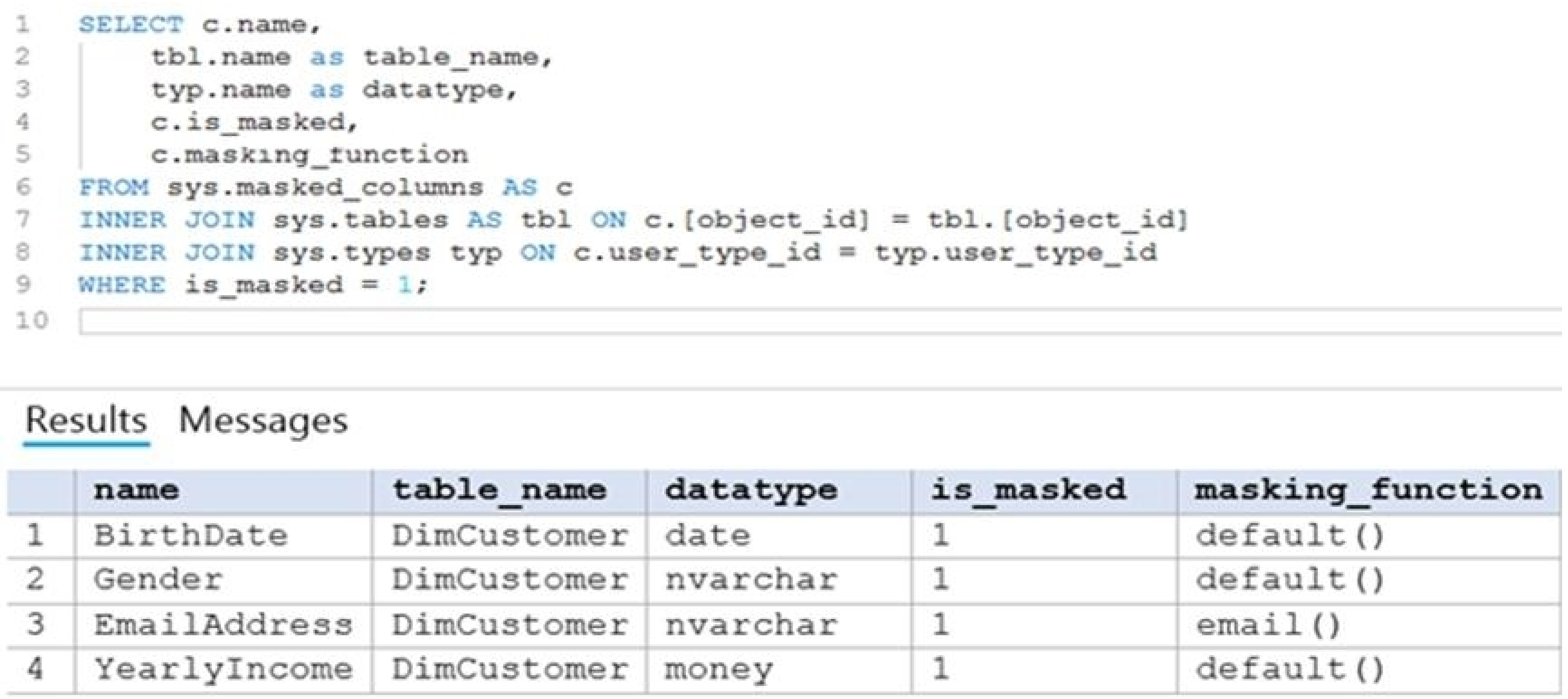
Section:
Explanation:
Box 1: DISTRIBUTION
Table distribution options include DISTRIBUTION = HASH (distribution_column_name), assigns each row to one distribution by hashing the value stored in distribution_column_name.
Box 2: PARTITION
Table partition options. Syntax:
PARTITION (partition_column_name RANGE [LEFT | RIGHT] FOR VALUES ([boundary_value [,...n]]))
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse?

QUESTION 17
HOTSPOT
You have an Azure Synapse Analytics dedicated SQL pool that contains the users shown in the following table.



Name	Role
User1	Server admin
User2	db_datereader

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



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

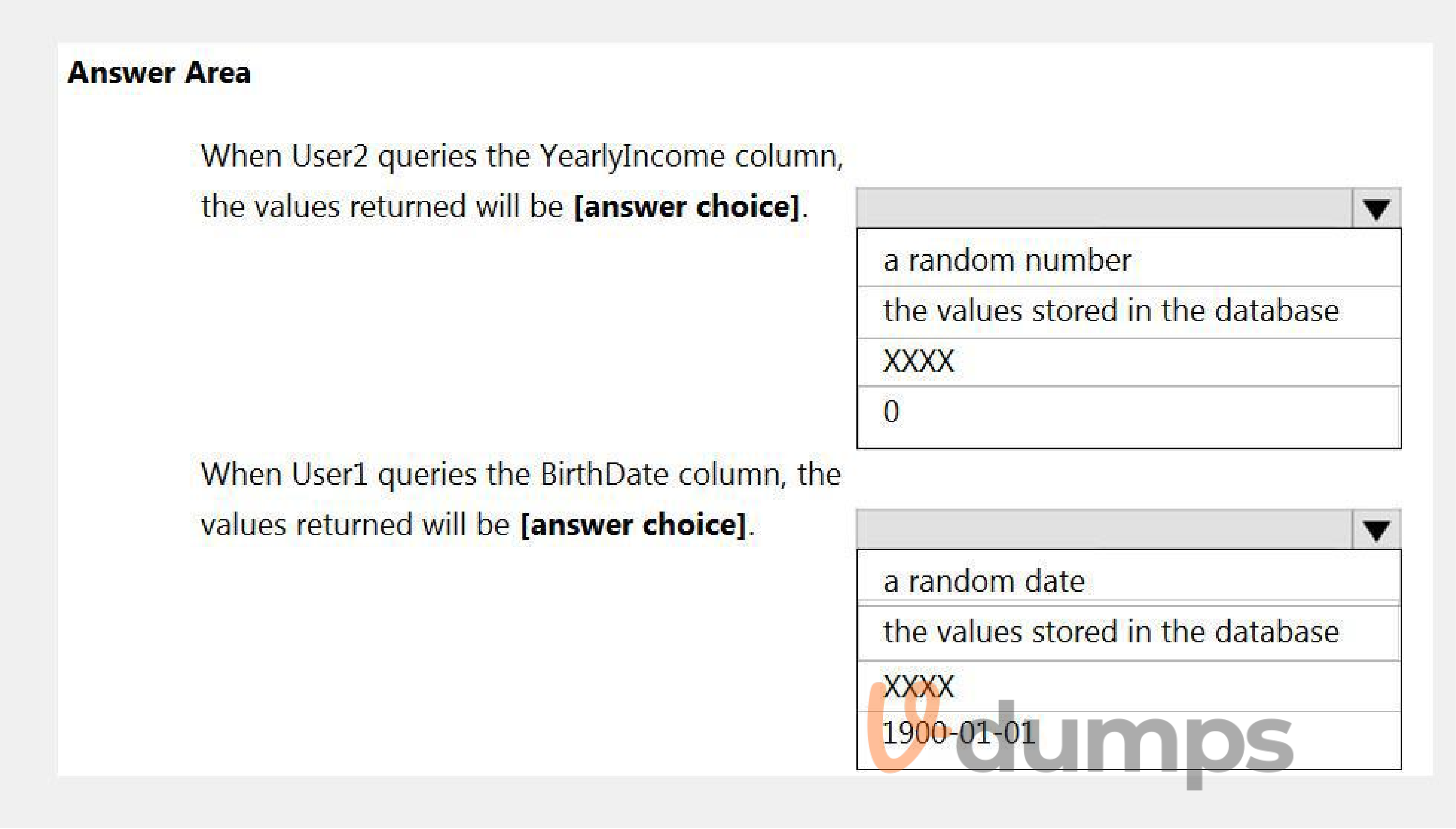
	name	table_name	datatype	is_masked	masking_function
1	BirthDate	DimCustomer	date	1	default()
2	Gender	DimCustomer	nvarchar	1	default()
3	EmailAddress	DimCustomer	nvarchar	1	email()
4	YearlyIncome	DimCustomer	money	1	default()

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

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

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

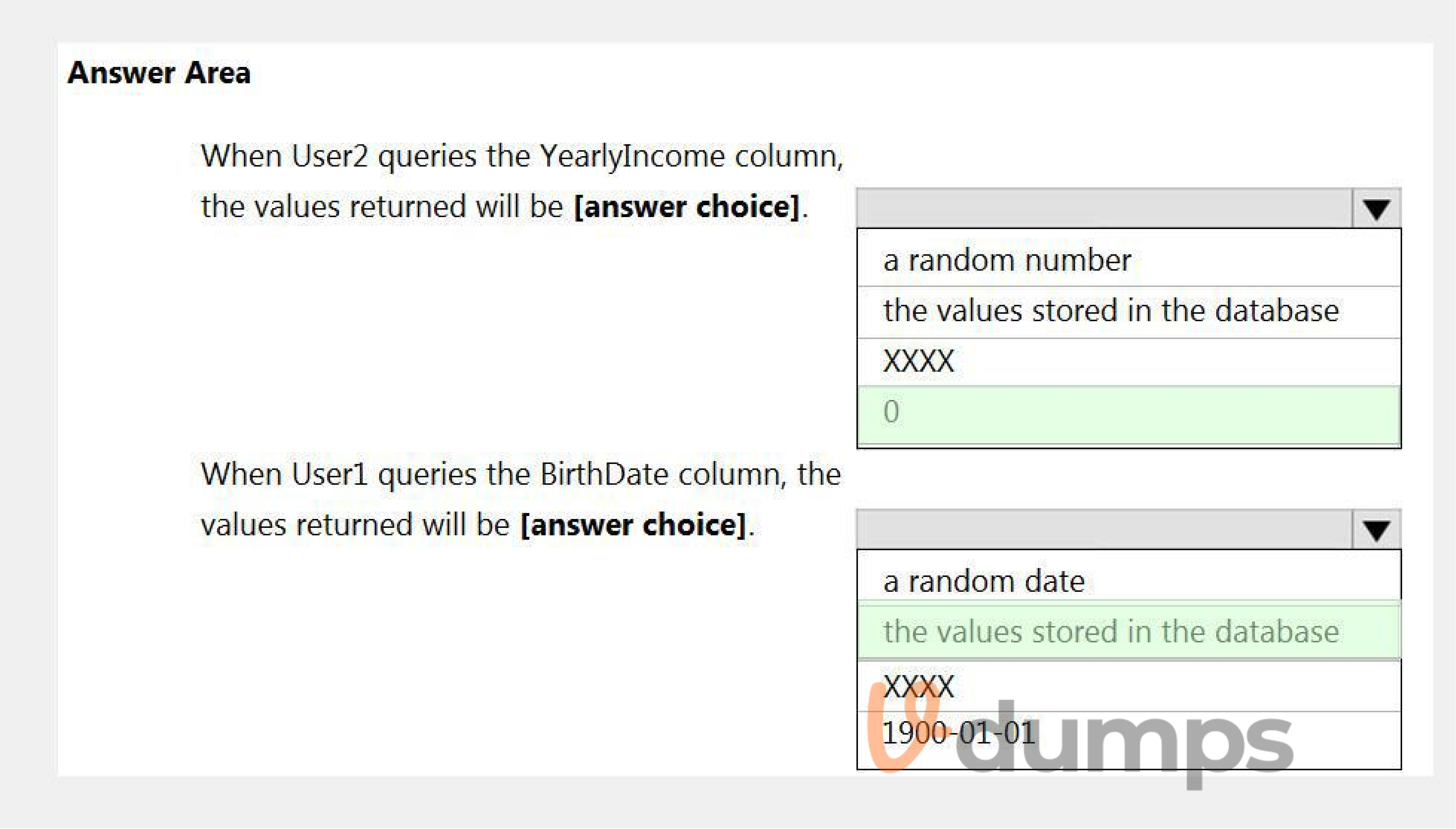
When User2 queries the YearlyIncome column, the values returned will be **[answer choice]**.

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

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

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

Answer Area:



Answer Area

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

- a random number
- the values stored in the database
- XXXX
- [✓ 0]

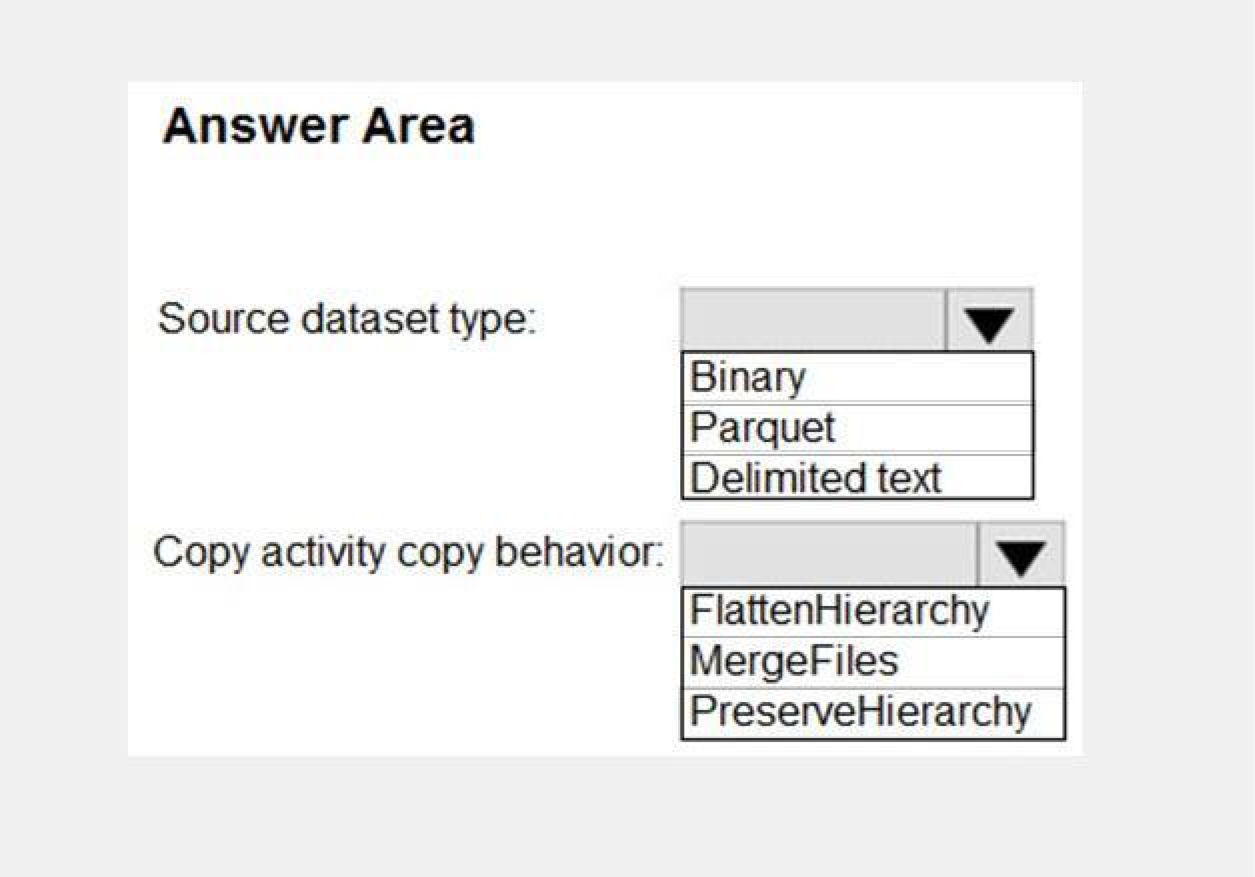
When User1 queries the BirthDate column, the values returned will be **[answer choice]**.

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

Section:
Explanation:
Box 1: 0
The YearlyIncome column is of the money data type.
The Default masking function: Full masking according to the data types of the designated fields Use a zero value for numeric data types (bigint, bit, decimal, int, money, numeric, smallint, smallmoney, tinyint, float, real).
Box 2: the values stored in the database
Users with administrator privileges are always excluded from masking, and see the original data without any mask.
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

QUESTION 18
HOTSPOT
You have two Azure Storage accounts named Storage1 and Storage2. Each account holds one container and has the hierarchical namespace enabled. The system has files that contain data stored in the Apache Parquet format.
You need to copy folders and files from Storage1 to Storage2 by using a Data Factory copy activity. The solution must meet the following requirements:
No transformations must be performed.
The original folder structure must be retained.
Minimize time required to perform the copy activity.
How should you configure the copy activity? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:

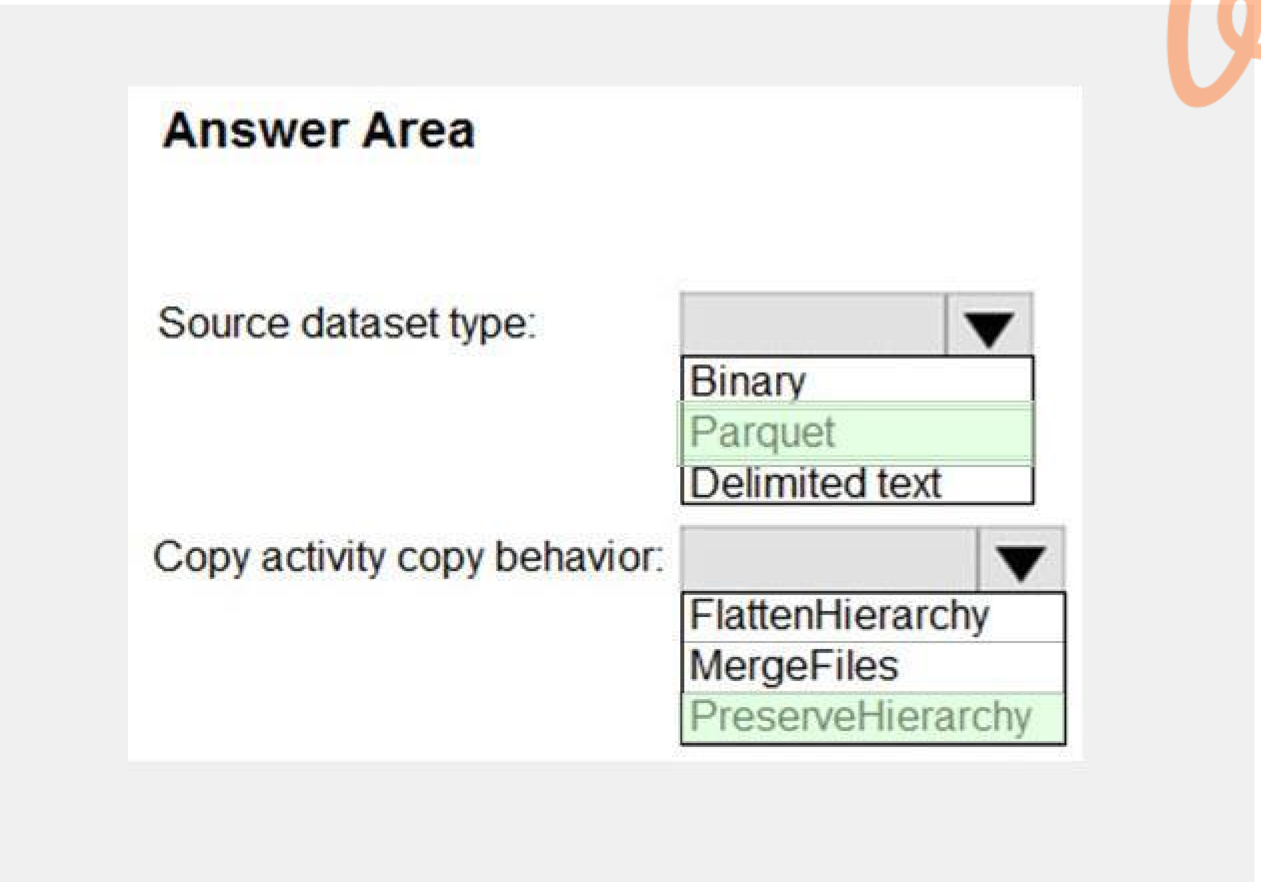


Answer Area

Source dataset type:
- Binary
- Parquet
- Delimited text

Copy activity copy behavior:
- FlattenHierarchy
- MergeFiles
- PreserveHierarchy

Answer Area:



Answer Area

Source dataset type:
- Binary
- [✓ Parquet]
- Delimited text

Copy activity copy behavior:
- FlattenHierarchy
- MergeFiles
- [✓ PreserveHierarchy]

Section:

Explanation:

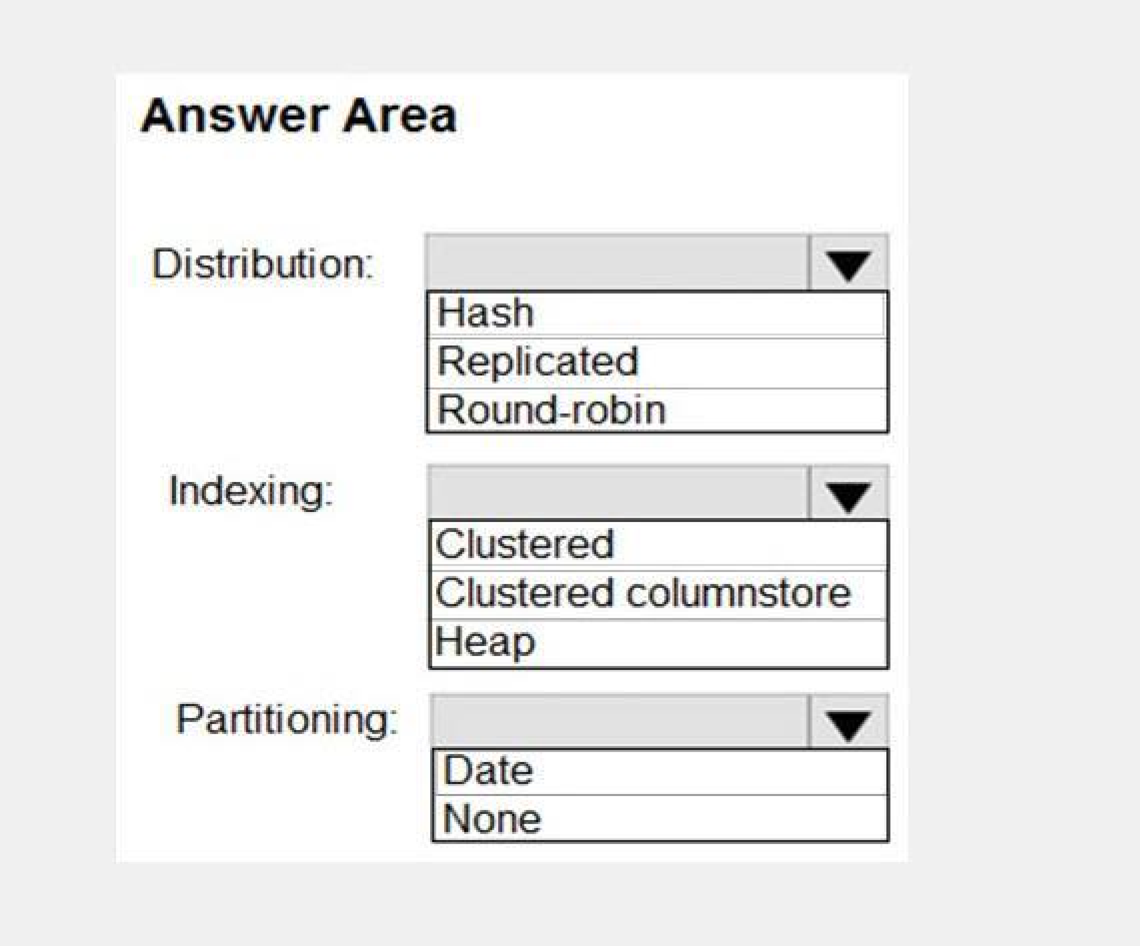
Box 1: Parquet

For Parquet datasets, the type property of the copy activity source must be set to ParquetSource.
Box 2: PreserveHierarchy
PreserveHierarchy (default): Preserves the file hierarchy in the target folder. The relative path of the source file to the source folder is identical to the relative path of the target file to the target folder. Incorrect Answers: FlattenHierarchy: All files from the source folder are in the first level of the target folder. The target files have autogenerated names. MergeFiles: Merges all files from the source folder to one file. If the file name is specified, the merged file name is the specified name. Otherwise, it's an autogenerated file name.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/format-parquet
https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage

QUESTION 19

HOTSPOT
You have a SQL pool in Azure Synapse.
You plan to load data from Azure Blob storage to a staging table. Approximately 1 million rows of data will be loaded daily. The table will be truncated before each daily load.
You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.
How should you configure the table? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:



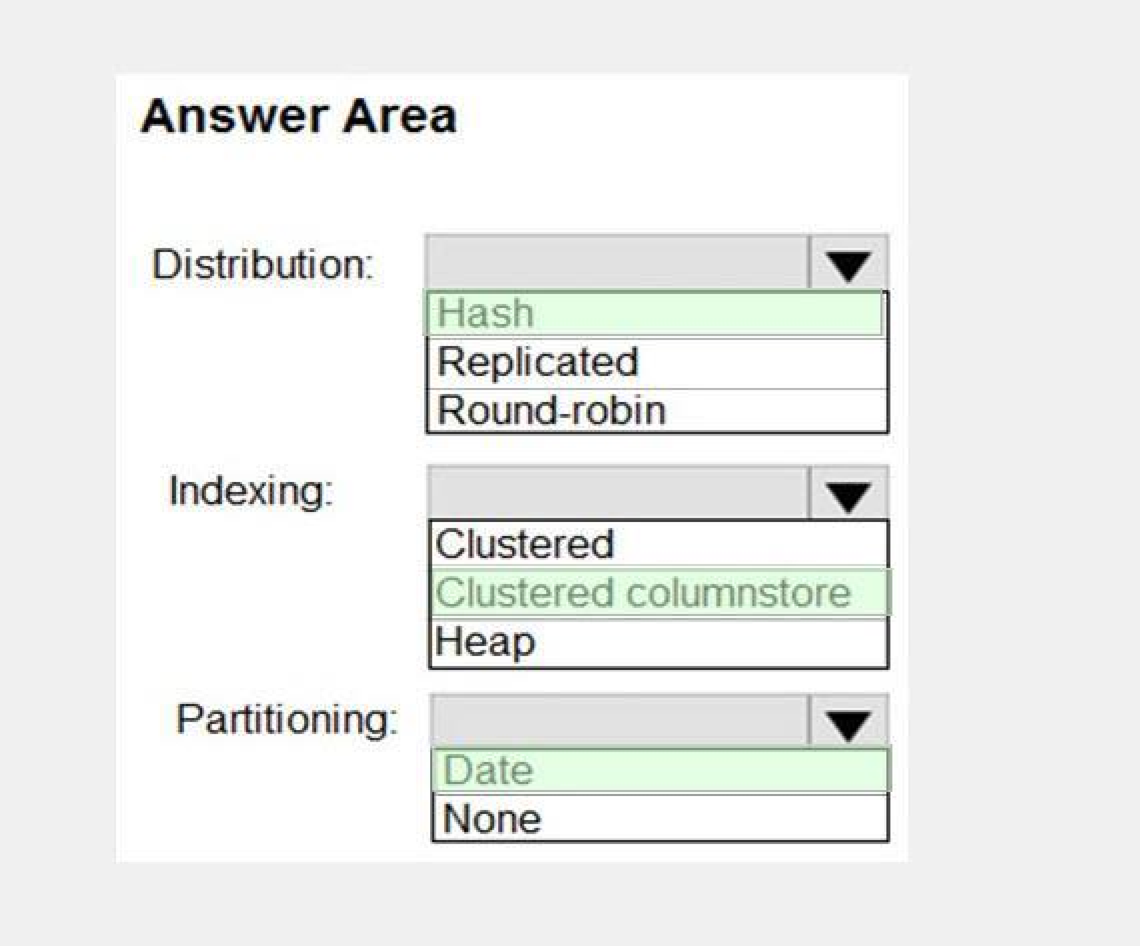
Answer Area

Distribution: ▼
- Hash
- Replicated
- Round-robin

Indexing: ▼
- Clustered
- Clustered columnstore
- Heap

Partitioning: ▼
- Date
- None

Answer Area:



Answer Area

Distribution: ▼
- Hash
- Replicated
- Round-robin

Indexing: ▼
- Clustered
- Clustered columnstore
- Heap

Partitioning: ▼
- Date
- None

Section:

Explanation:

Box 1: Hash

Hash-distributed tables improve query performance on large fact tables. They can have very large numbers of rows and still achieve high performance.

Incorrect Answers:

Round-robin tables are useful for improving loading speed.

Box 2: Clustered columnstore

When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed.

Box 3: Date

Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column. Partition switching can be used to quickly remove or replace a section of a table.

Reference:

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

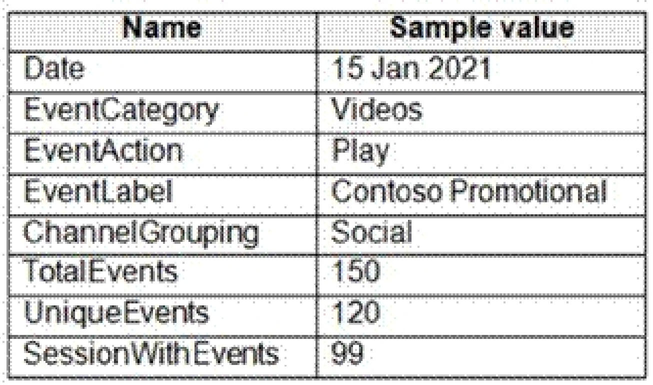
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

QUESTION 20

HOTSPOT

From a website analytics system, you receive data extracts about user interactions such as downloads, link clicks, form submissions, and video plays.

The data contains the following columns.



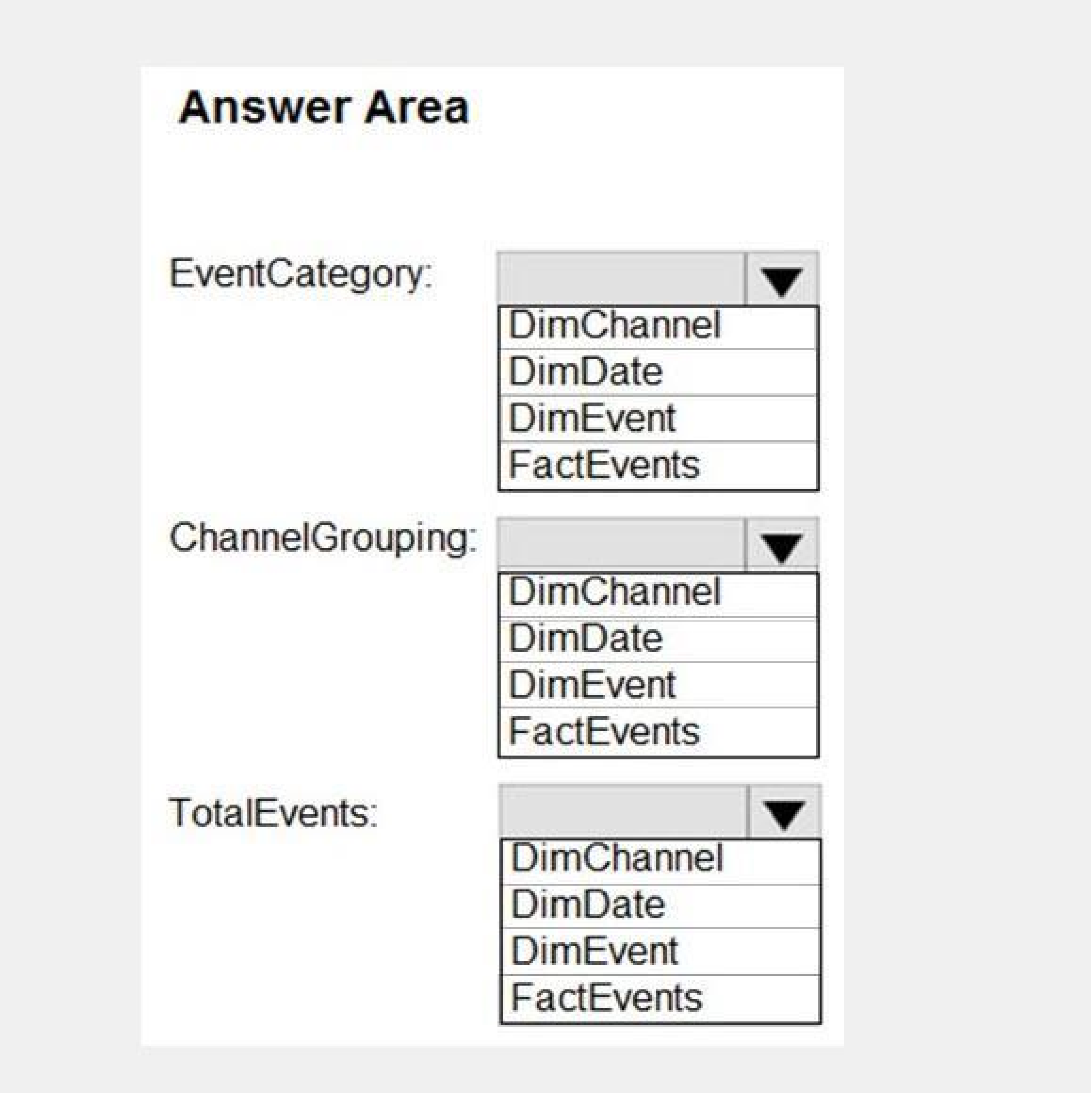
Name	Sample value
Date	15 Jan 2021
EventCategory	Videos
EventAction	Play
EventLabel	Contoso Promotional
ChannelGrouping	Social
TotalEvents	150
UniqueEvents	120
SessionWithEvents	99

You need to design a star schema to support analytical queries of the data. The star schema will contain four tables including a date dimension.

To which table should you add each column? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



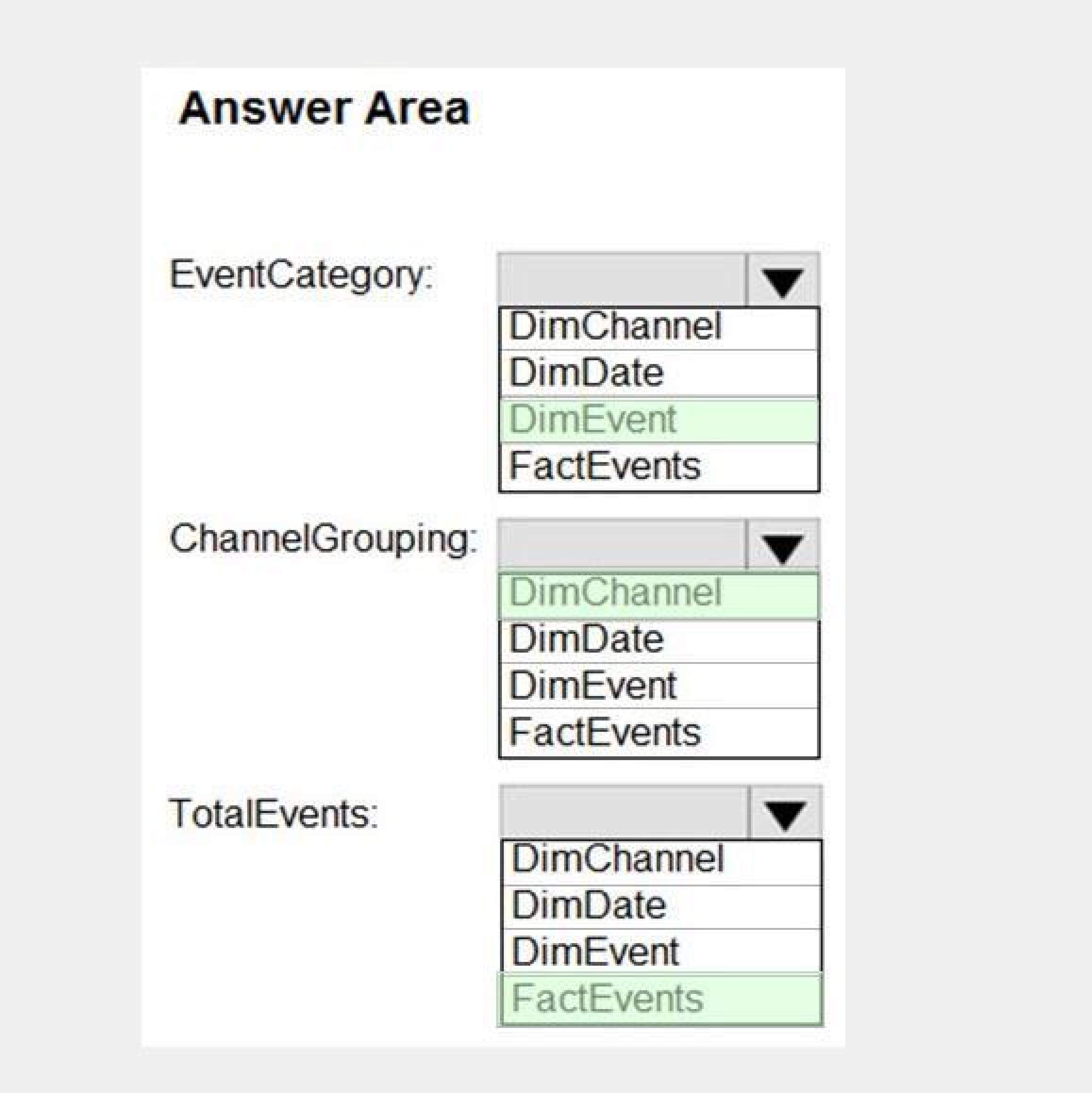
Answer Area

EventCategory:
- DimChannel
- DimDate
- DimEvent
- FactEvents

ChannelGrouping:
- DimChannel
- DimDate
- DimEvent
- FactEvents

TotalEvents:
- DimChannel
- DimDate
- DimEvent
- FactEvents

Answer Area:



Answer Area

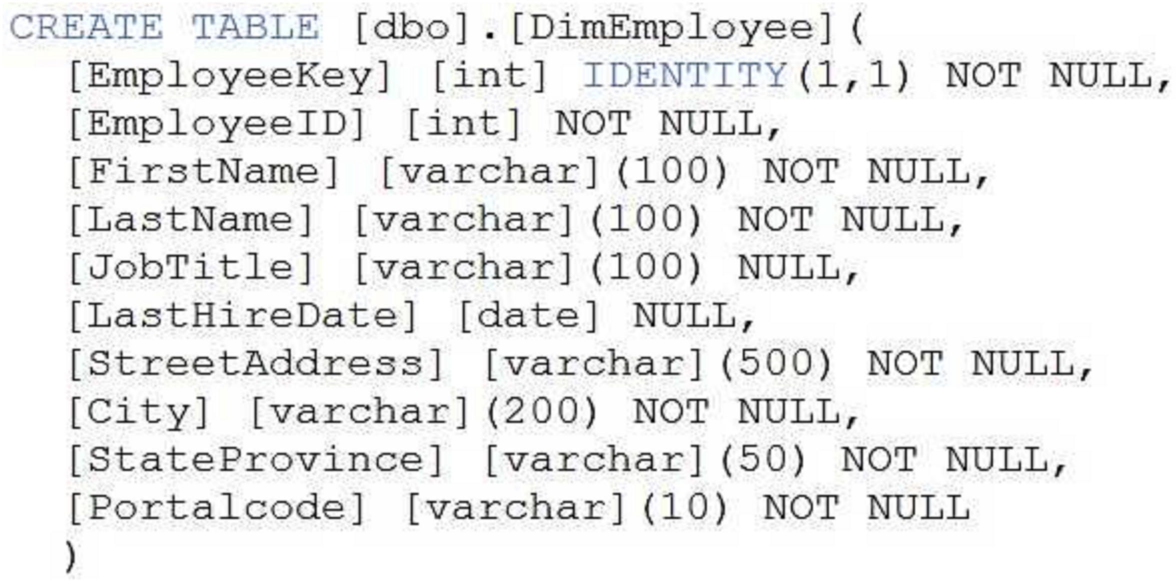
EventCategory:
- DimChannel
- DimDate
- [✓ DimEvent]
- FactEvents

ChannelGrouping:
- [✓ DimChannel]
- DimDate
- DimEvent
- FactEvents

TotalEvents:
- DimChannel
- DimDate
- DimEvent
- [✓ FactEvents]

Section:
Explanation:
Box 1: DimEvent
Box 2: DimChannel
Box 3: FactEvents
Fact tables store observations or events, and can be sales orders, stock balances, exchange rates, temperatures, etc
Reference:
https://docs.microsoft.com/en-us/power-bi/guidance/star-schema

QUESTION 21
You have a table in an Azure Synapse Analytics dedicated SQL pool. The table was created by using the following Transact-SQL statement.



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

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

A. [ManagerEmployeeID] [smallint] NULL

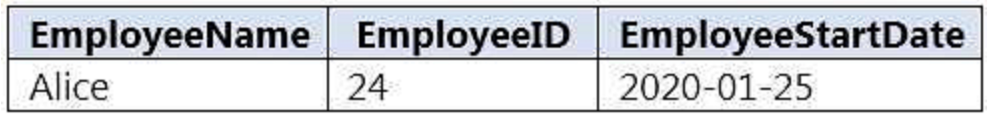
B. [ManagerEmployeeKey] [smallint] NULL

C. [ManagerEmployeeKey] [int] NULL

D. [ManagerName] [varchar](200) NULL

Correct Answer: C
Section:
Explanation:
We need an extra column to identify the Manager. Use the data type as the EmployeeKey column, an int column.
Reference: https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular

QUESTION 22
You have an Azure Synapse workspace named MyWorkspace that contains an Apache Spark database named mytestdb.
You run the following command in an Azure Synapse Analytics Spark pool in MyWorkspace.
CREATE TABLE mytestdb.myParquetTable(
EmployeeID int,
EmployeeName string,
EmployeeStartDate date)
USING Parquet
You then use Spark to insert a row into mytestdb.myParquetTable. The row contains the following data.



EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

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

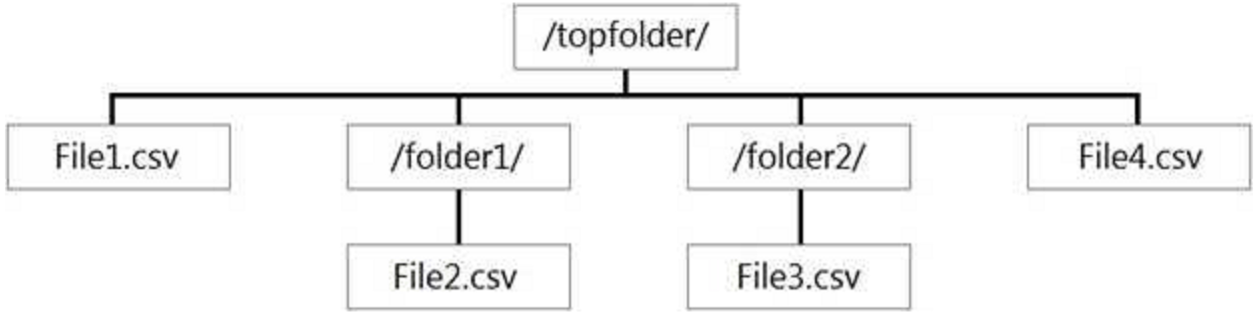
A. 24

B. an error

C. a null value

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

QUESTION 23
You have files and folders in Azure Data Lake Storage Gen2 for an Azure Synapse workspace as shown in the following exhibit.



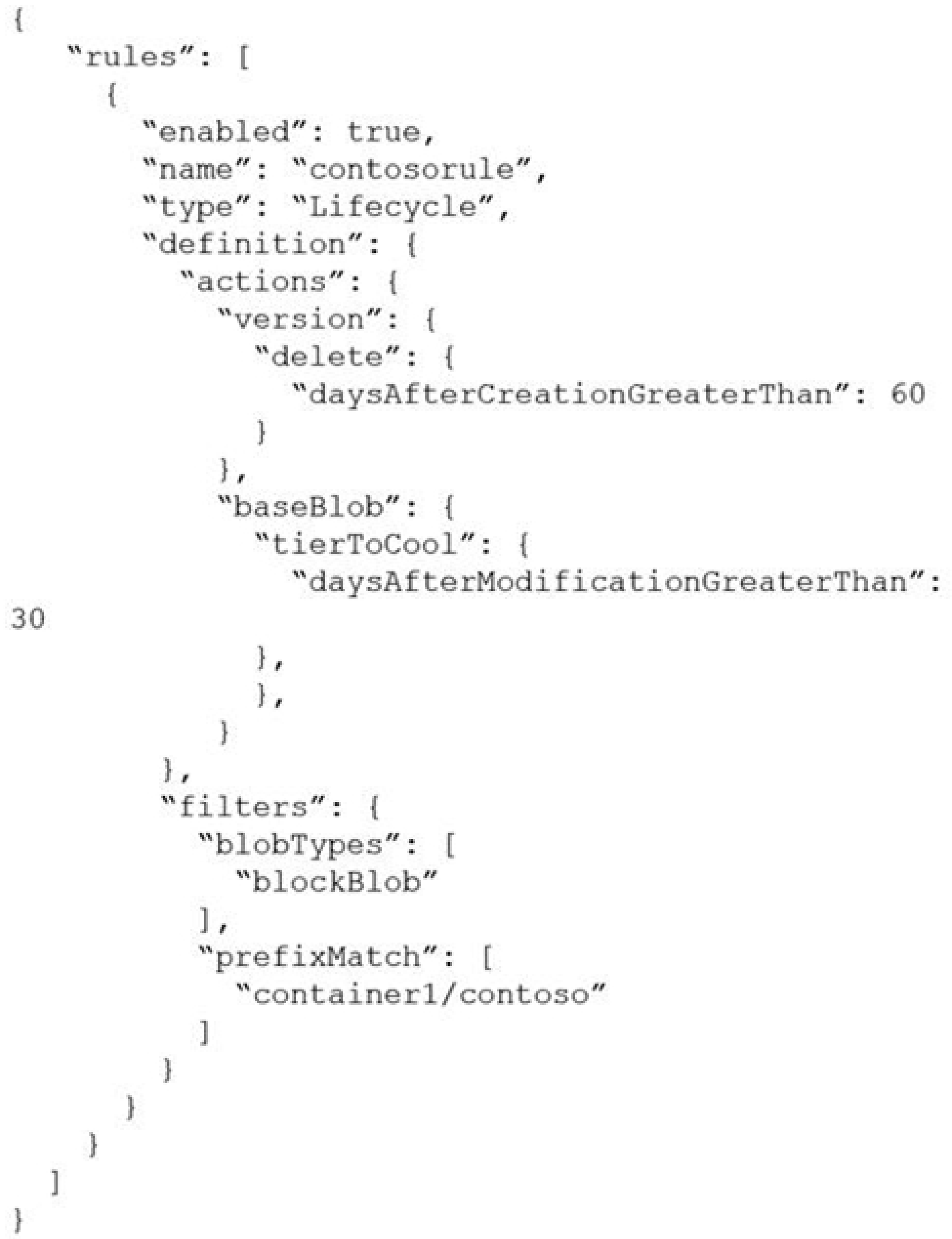
You create an external table named ExtTable that has LOCATION='/topfolder/'.
When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, which files are returned?

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

Correct Answer: B
Section:
Explanation:

Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-data-storage#query-multiple-files-or-folders

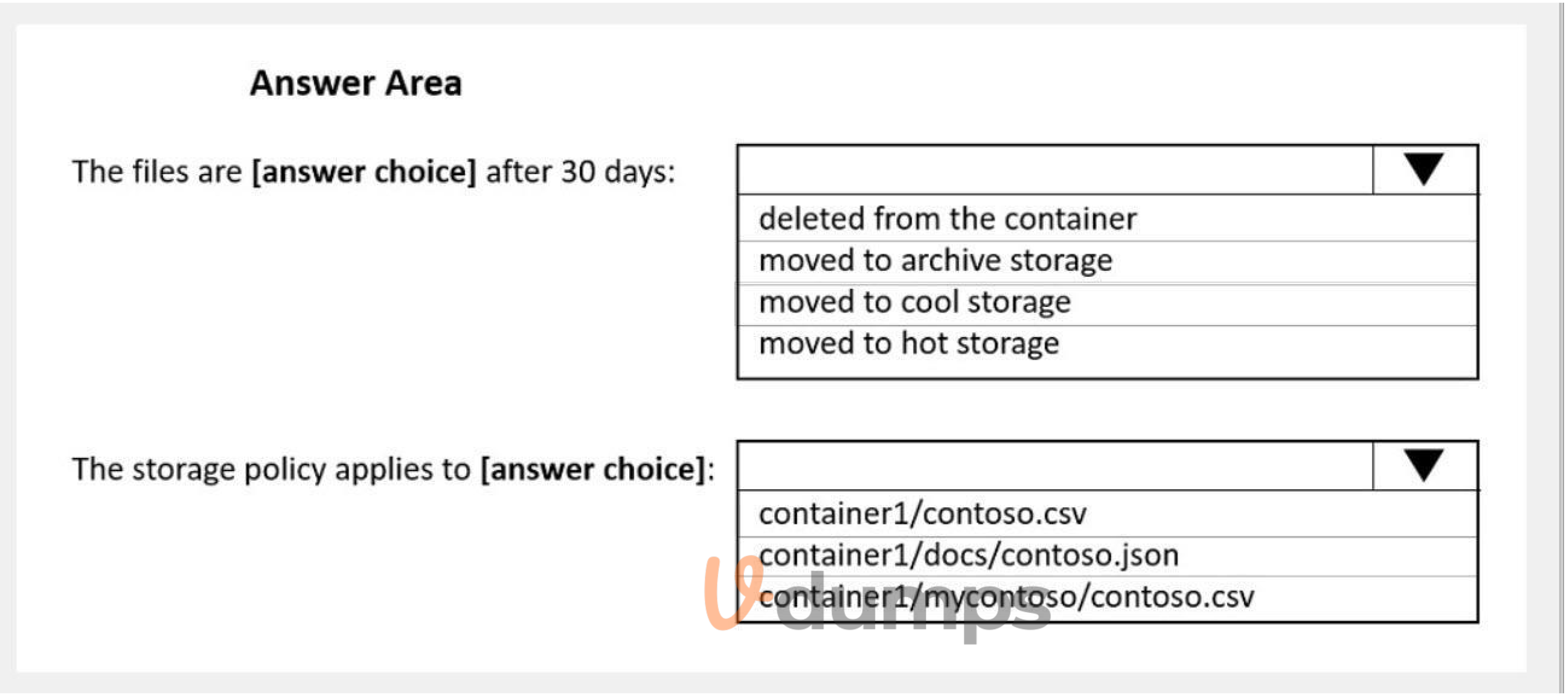
QUESTION 24
HOTSPOT
You store files in an Azure Data Lake Storage Gen2 container. The container has the storage policy shown in the following exhibit.



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

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

Hot Area:



Answer Area

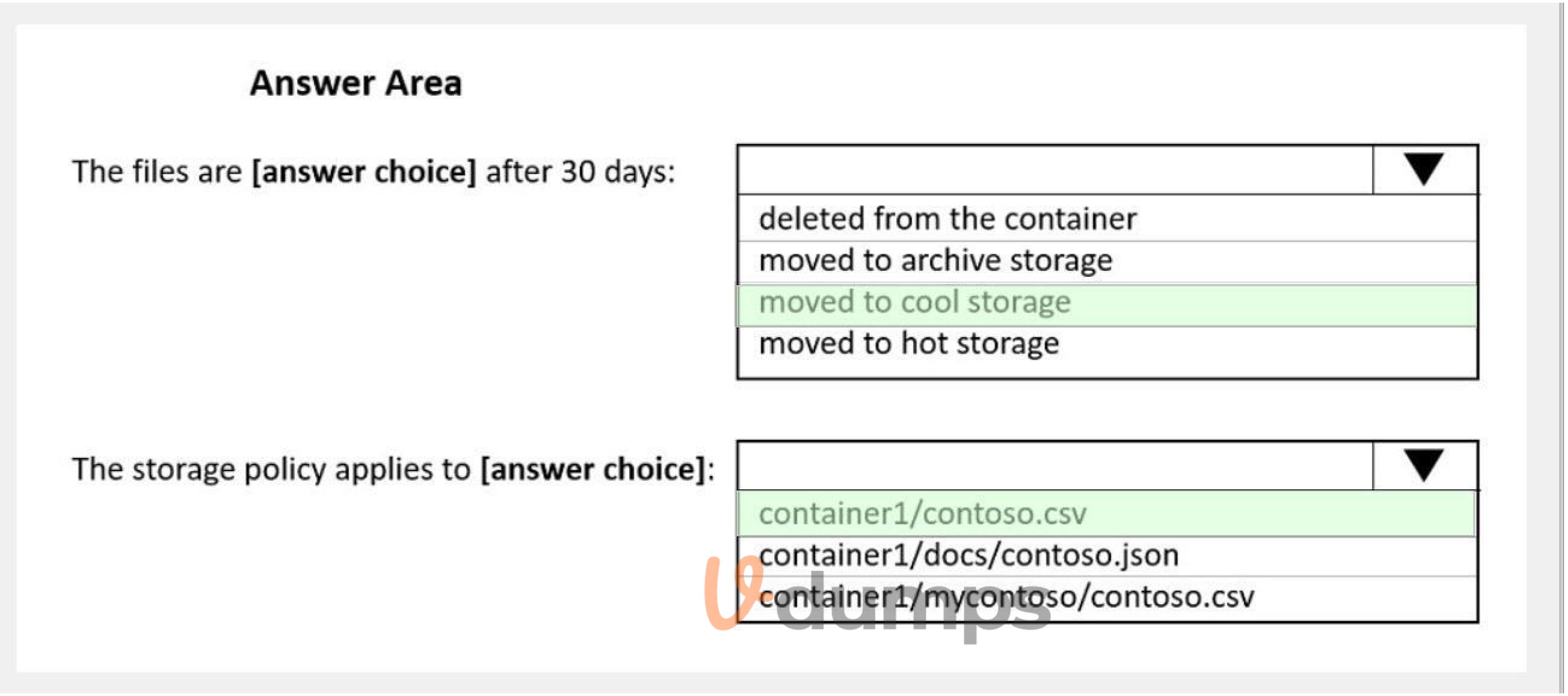
The files are **[answer choice]** after 30 days:

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

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

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

Answer Area:



Answer Area

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

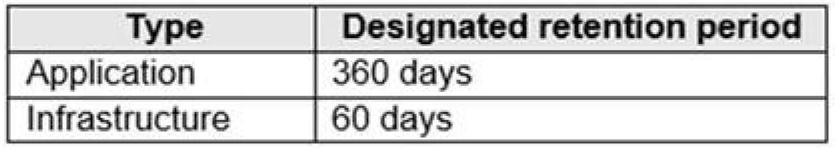
- deleted from the container
- moved to archive storage
- [✓ moved to cool storage]
- moved to hot storage

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

- [✓ container1/contoso.csv]
- container1/docs/contoso.json
- container1/mycontoso/contoso.csv

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

QUESTION 25
HOTSPOT
You have an Azure Data Lake Storage Gen2 account named account1 that stores logs as shown in the following table.

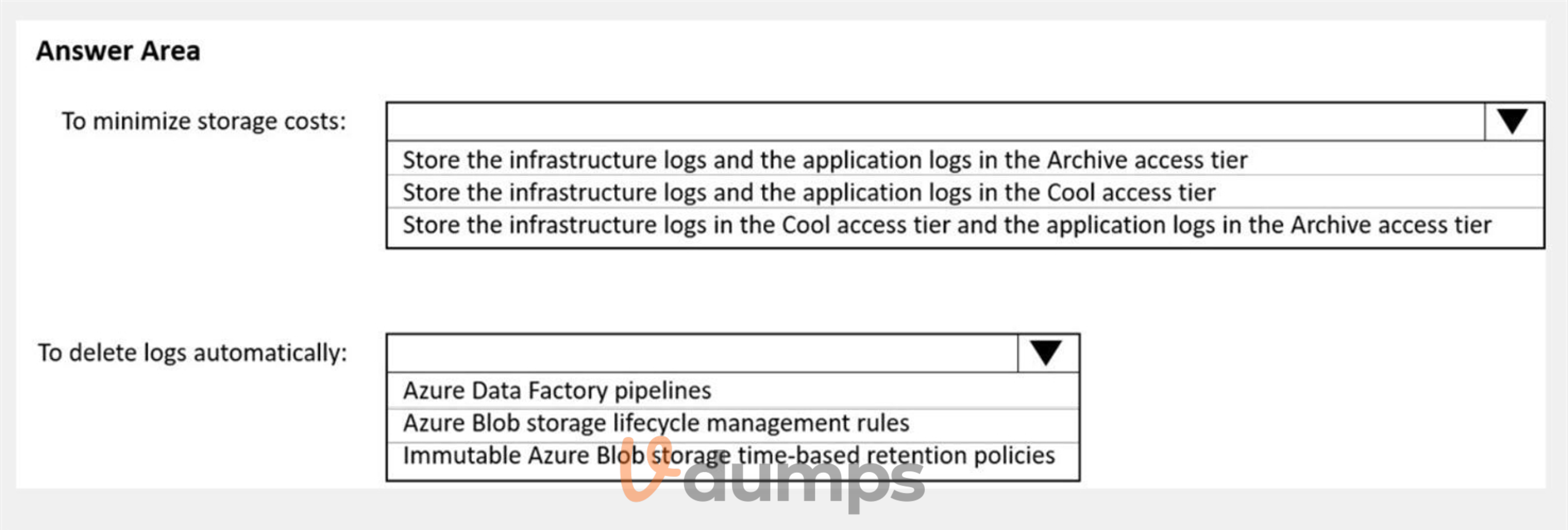


Type	Designated retention period
Application	360 days
Infrastructure	60 days

You do not expect that the logs will be accessed during the retention periods.
You need to recommend a solution for account1 that meets the following requirements:
Automatically deletes the logs at the end of each retention period Minimizes storage costs
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.

Hot Area:



Answer Area

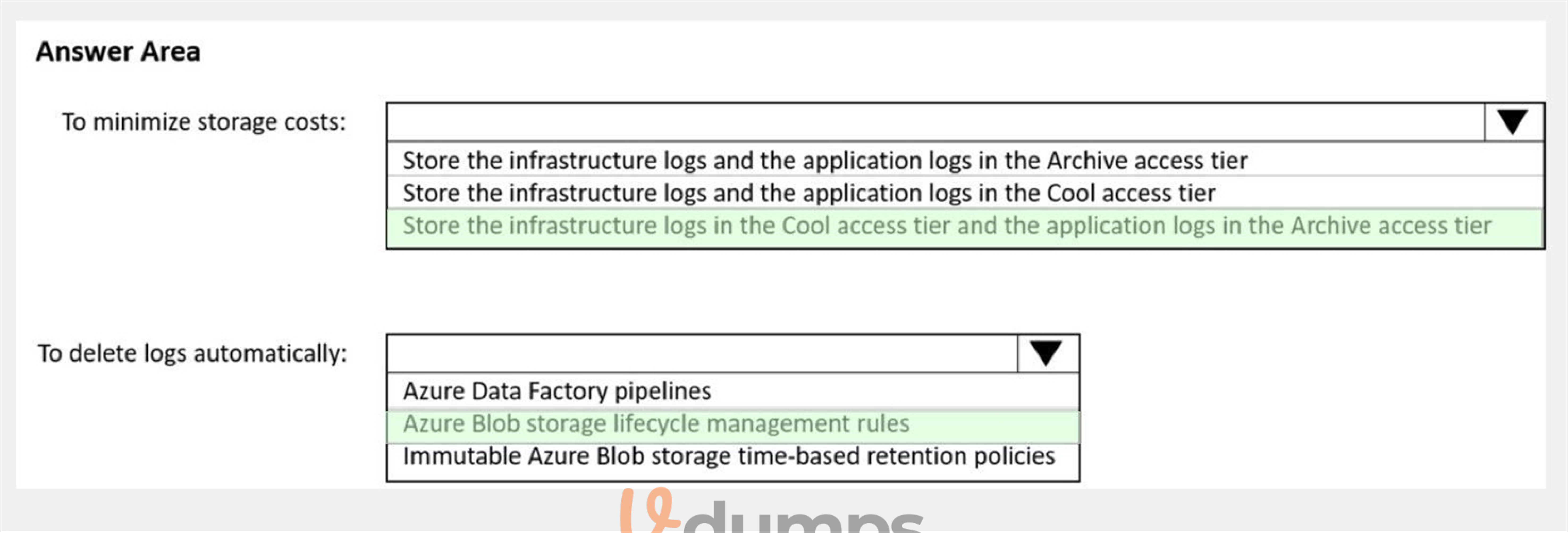
To minimize storage costs:

- Store the infrastructure logs and the application logs in the Archive access tier
- Store the infrastructure logs and the application logs in the Cool access tier
- Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier

To delete logs automatically:

- Azure Data Factory pipelines
- Azure Blob storage lifecycle management rules
- Immutable Azure Blob storage time-based retention policies

Answer Area:



Answer Area

To minimize storage costs:

- Store the infrastructure logs and the application logs in the Archive access tier
- Store the infrastructure logs and the application logs in the Cool access tier
- [✓ Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier]

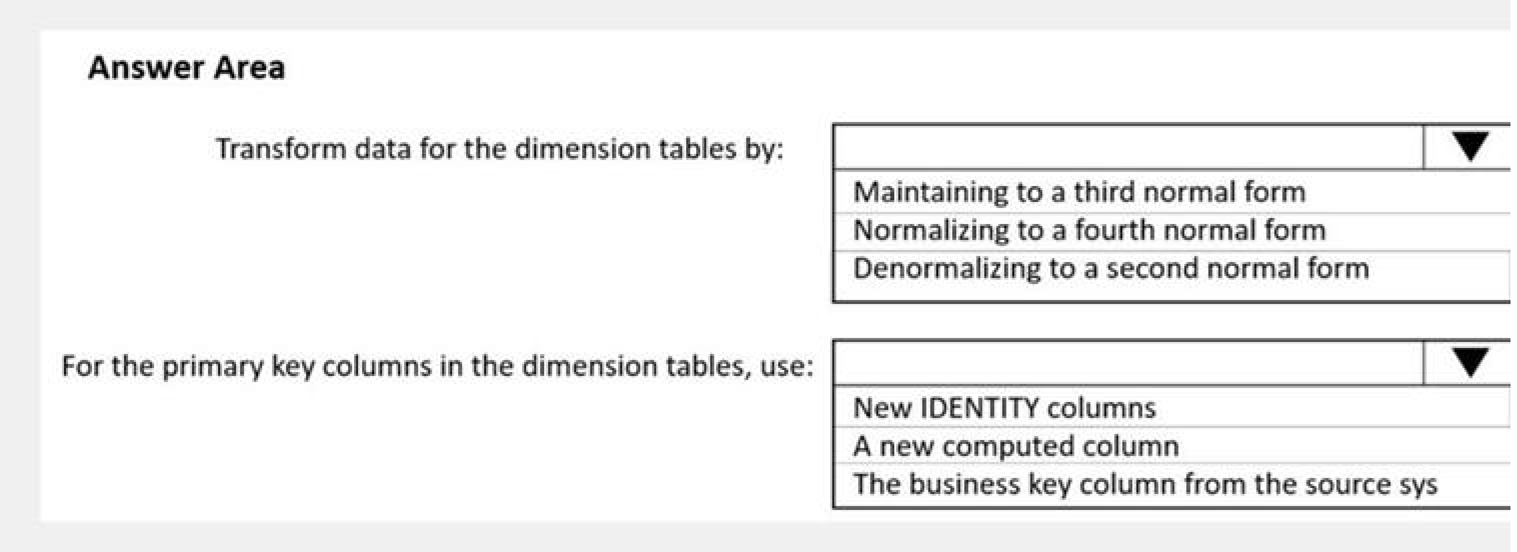
To delete logs automatically:

- Azure Data Factory pipelines
- [✓ Azure Blob storage lifecycle management rules]
- Immutable Azure Blob storage time-based retention policies

Section:
Explanation:
Box 1: Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier For infrastructure logs: Cool tier - An online tier optimized for storing data that is infrequently accessed or modified. Data in the cool tier should be stored for a minimum of 30 days. The cool tier has lower storage costs and higher access costs compared to the hot tier.
For application logs: Archive tier - An offline tier optimized for storing data that is rarely accessed, and that has flexible latency requirements, on the order of hours. Data in the archive tier should be stored for a minimum of 180 days.
Box 2: Azure Blob storage lifecycle management rules
Blob storage lifecycle management offers a rule-based policy that you can use to transition your data to the desired access tier when your specified conditions are met. You can also use lifecycle management to expire data at the end of its life.
Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview

QUESTION 26
HOTSPOT
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 Azure 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.

Hot Area:



Answer Area

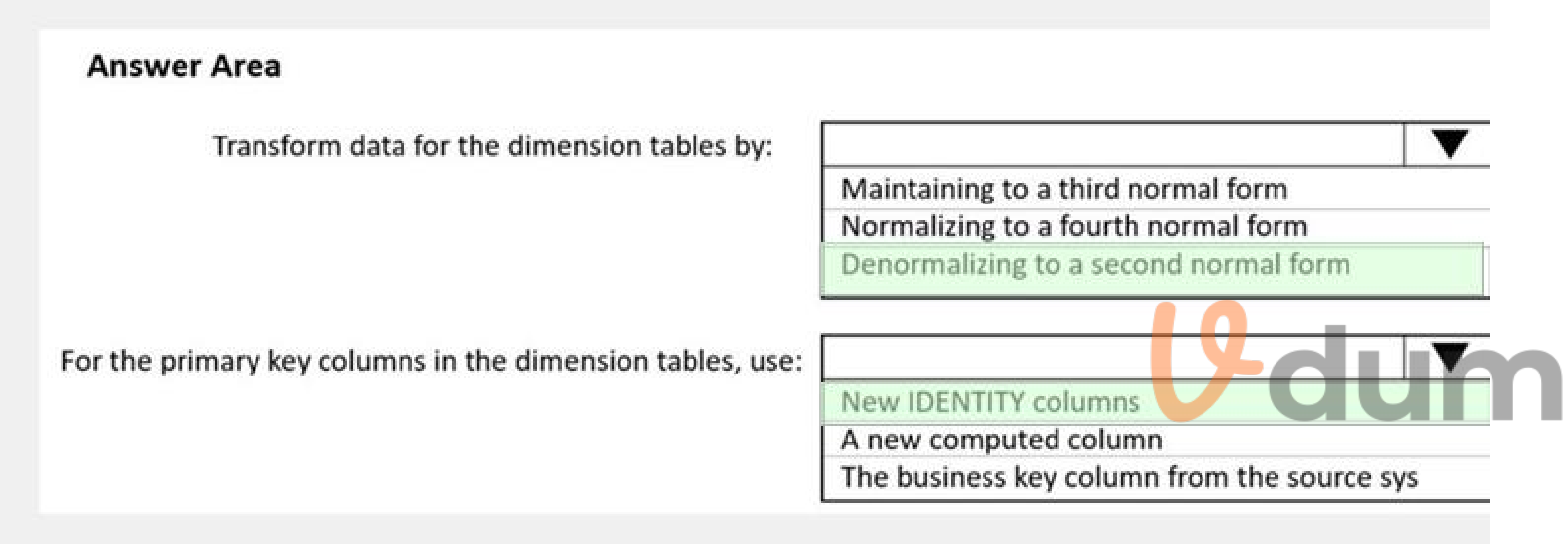
Transform data for the dimension tables by:

- Maintaining to a third normal form
- Normalizing to a fourth normal form
- Denormalizing to a second normal form

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

Answer Area:



Answer Area

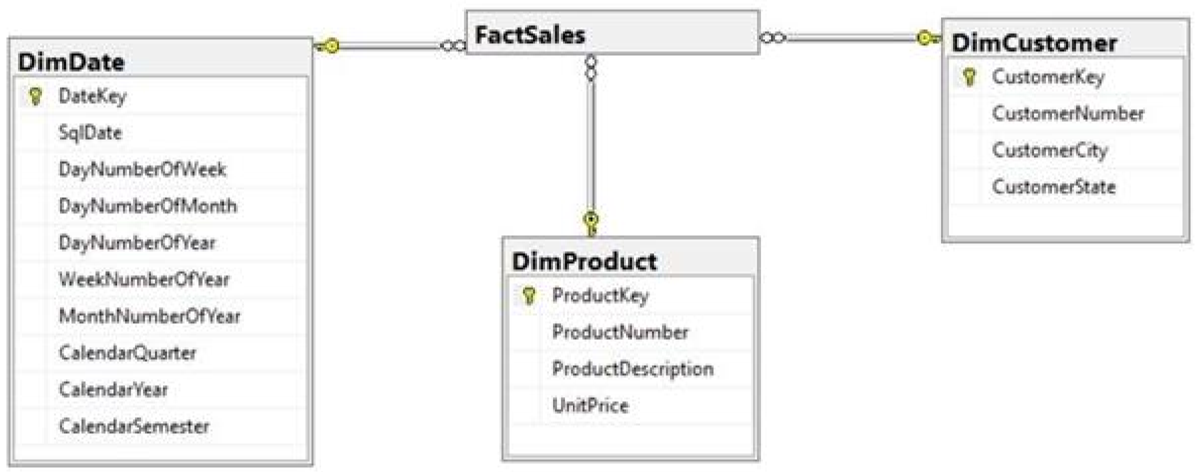
Transform data for the dimension tables by:

- Maintaining to a third normal form
- Normalizing to a fourth normal form
- [✓ Denormalizing to a second normal form]

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

Section:
Explanation:
Box 1: Denormalize to a second normal form
Denormalization is the process of transforming higher normal forms to lower normal forms via storing the join of higher normal form relations as a base relation. Denormalization increases the performance in data retrieval at cost of bringing update anomalies to a database.
Box 2: New identity columns
The collapsing relations strategy can be used in this step to collapse classification entities into component entities to obtain ?at dimension tables with single-part keys that connect directly to the fact table. The single-part key is a surrogate key generated to ensure it remains unique over time.
Example:



Note: 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://www.mssqltips.com/sqlservertip/5614/explore-the-role-of-normal-forms-in-dimensional-modeling/
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identity

QUESTION 27

HOTSPOT

You plan to develop a dataset named Purchases by using Azure Databricks. Purchases will contain the following columns:

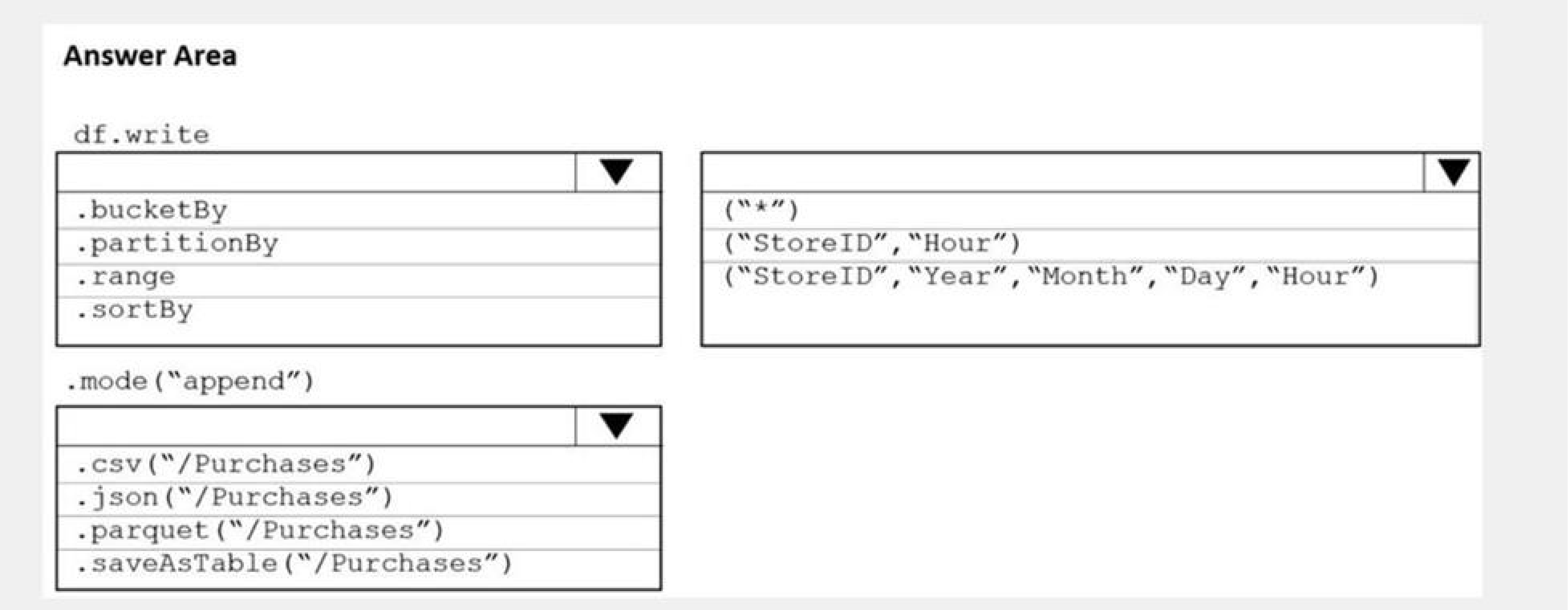
ProductID
ItemPrice
LineTotal
Quantity
StoreID
Minute
Month
Hour
Year
Day

You need to store the data to support hourly incremental load pipelines that will vary for each Store ID. The solution must minimize storage costs.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

df.write

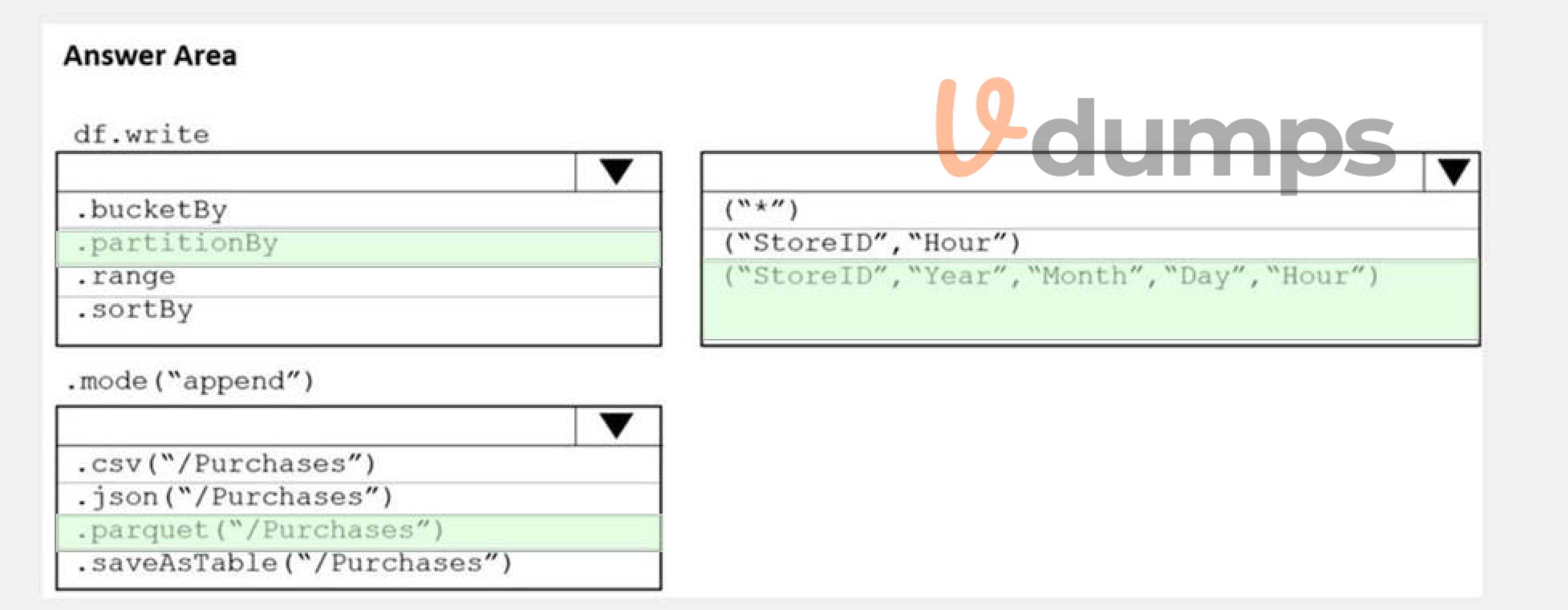
- .bucketBy
- .partitionBy
- .range
- .sortBy

- ("*")
- ("StoreID","Hour")
- ("StoreID","Year","Month","Day","Hour")

.mode("append")

- .csv("/Purchases")
- .json("/Purchases")
- .parquet("/Purchases")
- .saveAsTable("/Purchases")

Answer Area:



Answer Area

df.write

- .bucketBy
- .partitionBy
- .range
- .sortBy

- ("*")
- ("StoreID","Hour")
- ("StoreID","Year","Month","Day","Hour")

.mode("append")

- .csv("/Purchases")
- .json("/Purchases")
- .parquet("/Purchases")
- .saveAsTable("/Purchases")

Section:

Explanation:

Box 1: partitionBy

We should overwrite at the partition level.

Example:

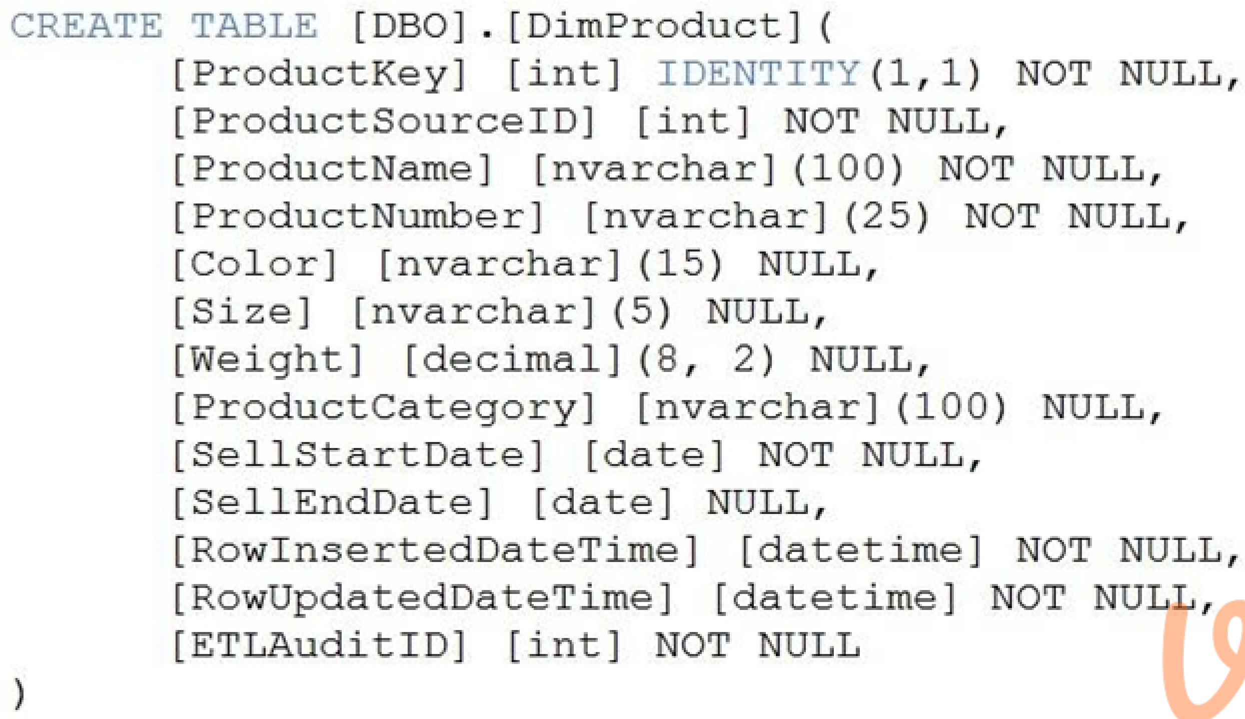
df.write.partitionBy("y","m","d")
.mode(SaveMode.Append)
.parquet("/data/hive/warehouse/db_name.db/" + tableName)

Box 2: ("StoreID", "Year", "Month", "Day", "Hour", "StoreID")

Box 3: parquet("/Purchases")
Reference:
https://intellipaat.com/community/11744/how-to-partition-and-write-dataframe-in-spark-without-deleting-partitions-with-no-new-data

QUESTION 28

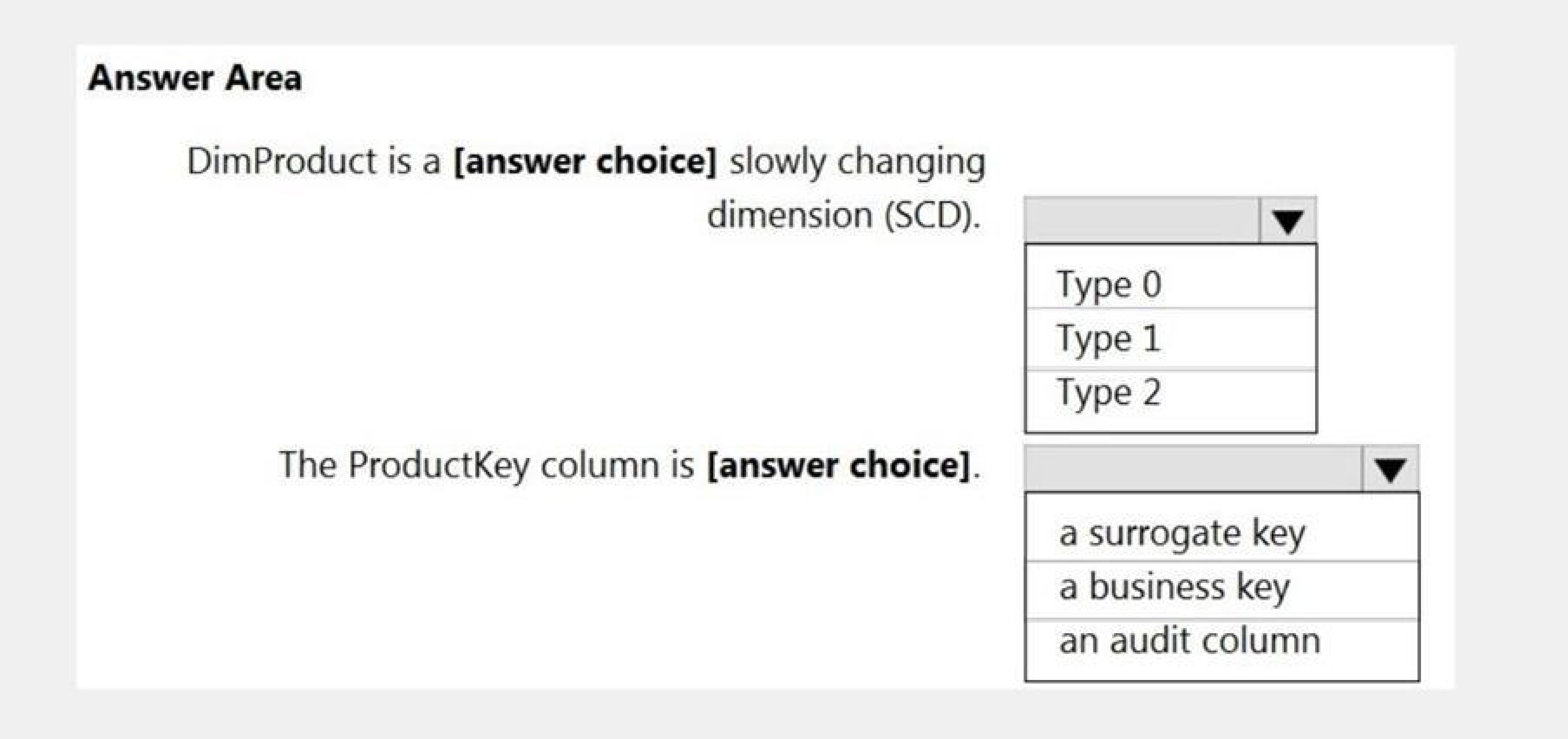
HOTSPOT
You are creating dimensions for a data warehouse in an Azure Synapse Analytics dedicated SQL pool.
You create a table by using the Transact-SQL statement shown in the following exhibit.



```
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.

Hot Area:



Answer Area

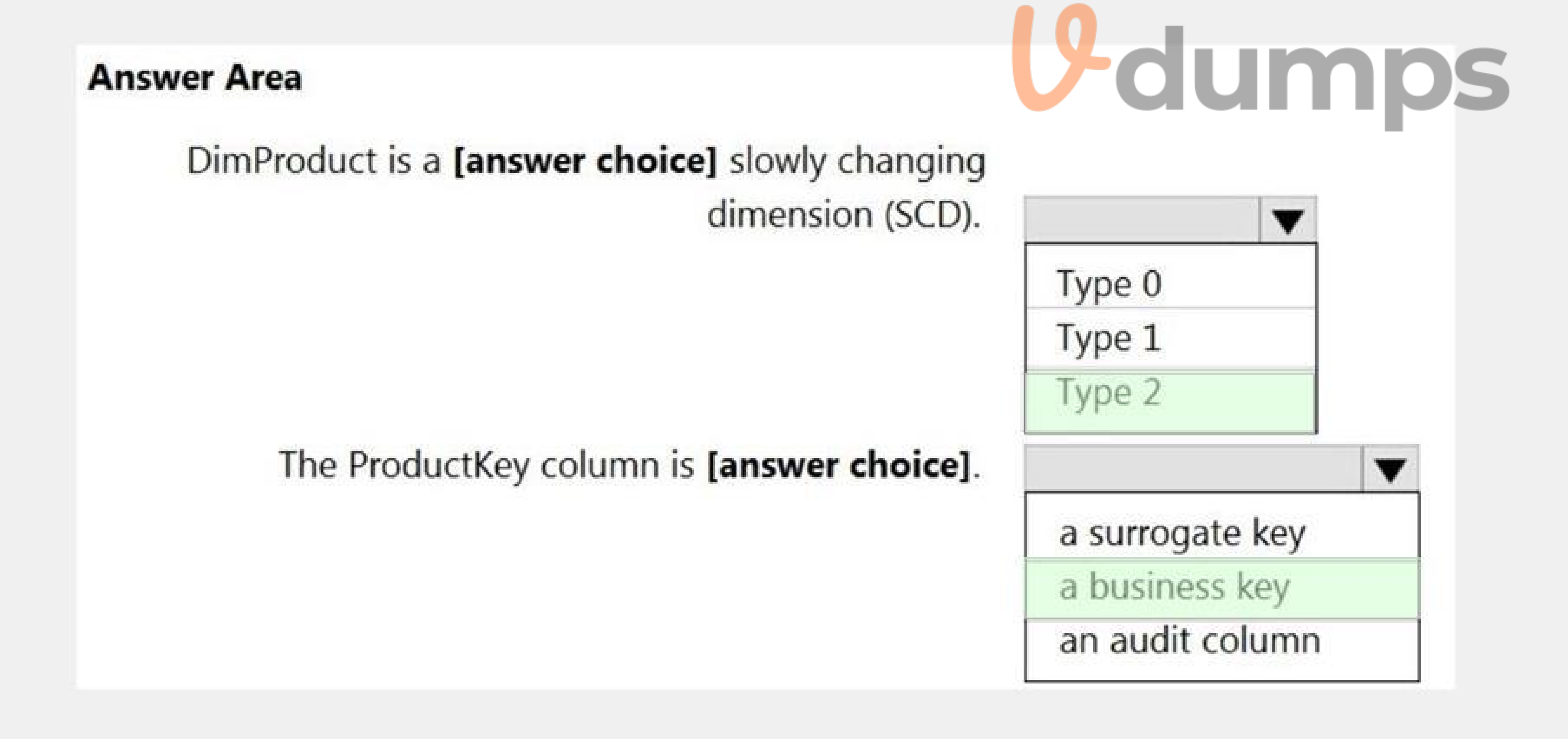
DimProduct is a **[answer choice]** slowly changing dimension (SCD).

- Type 0
- Type 1
- Type 2

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

- a surrogate key
- a business key
- an audit column

Answer Area:



Answer Area

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

Section:
Explanation:

QUESTION 29

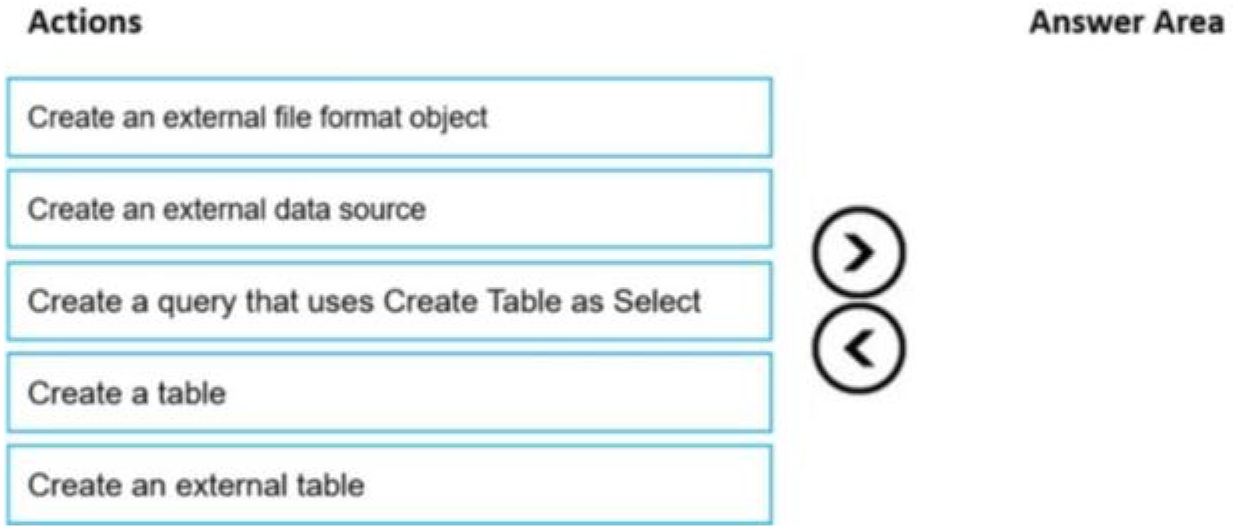
DRAG DROP

You need to build a solution to ensure that users can query specific files in an Azure Data Lake Storage Gen2 account from an Azure Synapse Analytics serverless SQL pool.

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.

Select and Place:

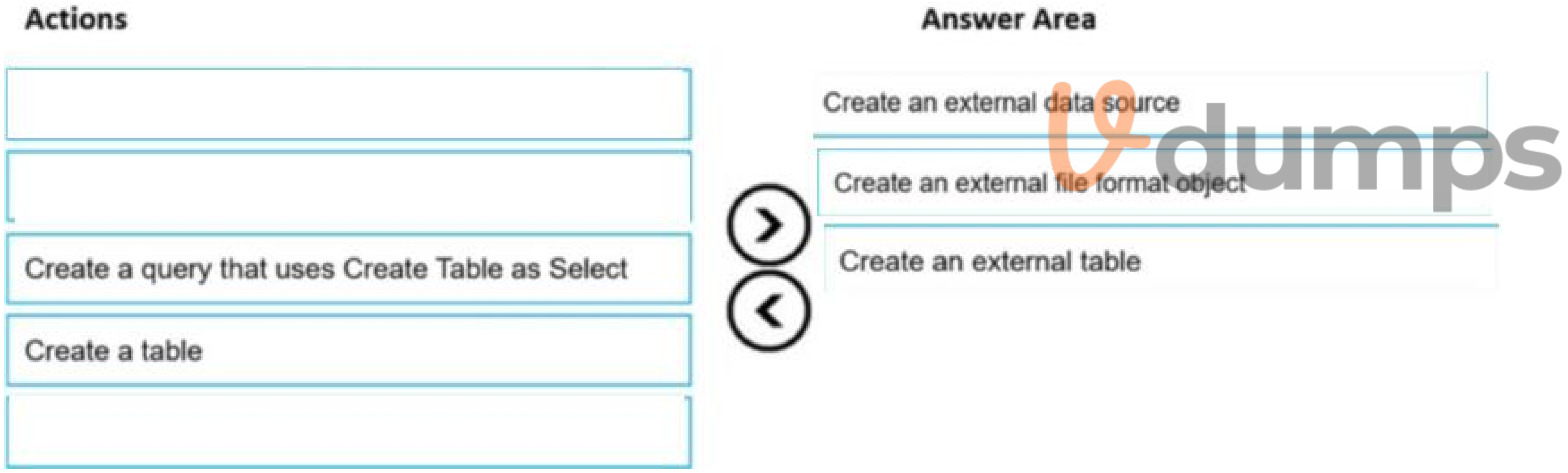


Actions

- Create an external file format object
- Create an external data source
- Create a query that uses Create Table as Select
- Create a table
- Create an external table

Answer Area

Correct Answer:



Actions

- Create a query that uses Create Table as Select
- Create a table

Answer Area

- Create an external data source
- Create an external file format object
- Create an external table

Section:

Explanation:

Step 1: Create an external data source

You can create external tables in Synapse SQL pools via the following steps:

CREATE EXTERNAL DATA SOURCE to reference an external Azure storage and specify the credential that should be used to access the storage. CREATE EXTERNAL FILE FORMAT to describe format of CSV or Parquet files. CREATE EXTERNAL TABLE on top of the files placed on the data source with the same file format.

Step 2: Create an external file format object

Creating an external file format is a prerequisite for creating an external table.

Step 3: Create an external table

Reference:

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

QUESTION 30

You build a data warehouse in an Azure Synapse Analytics dedicated SQL pool. Analysts write a complex SELECT query that contains multiple JOIN and CASE statements to transform data for use in inventory reports. The inventory reports will use the data and additional WHERE parameters depending on the report. The reports will be produced once daily.

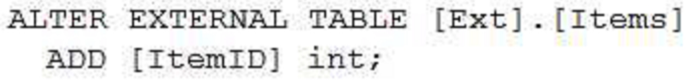
You need to implement a solution to make the dataset available for the reports. The solution must minimize query times. What should you implement?

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

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

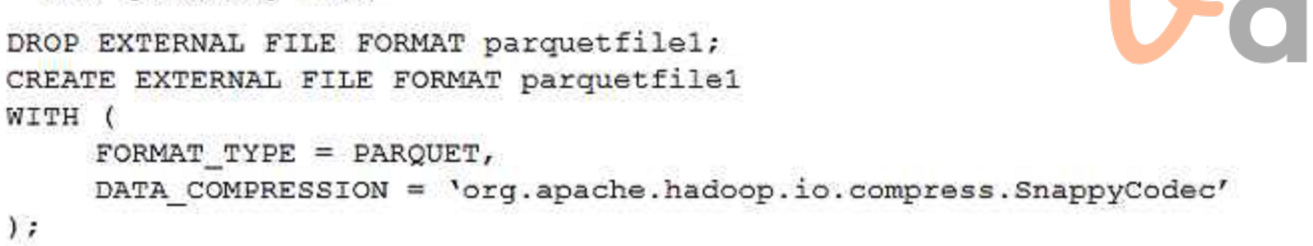
QUESTION 31
You have an enterprise data warehouse in Azure Synapse Analytics. Using PolyBase, you create an external table named [Ext].[Items] to query Parquet files stored in Azure Data Lake Storage Gen2 without importing the data to the data warehouse. The external table has three columns.
You discover that the Parquet files have a fourth column named ItemID. Which command should you run to add the ItemID column to the external table?

A.



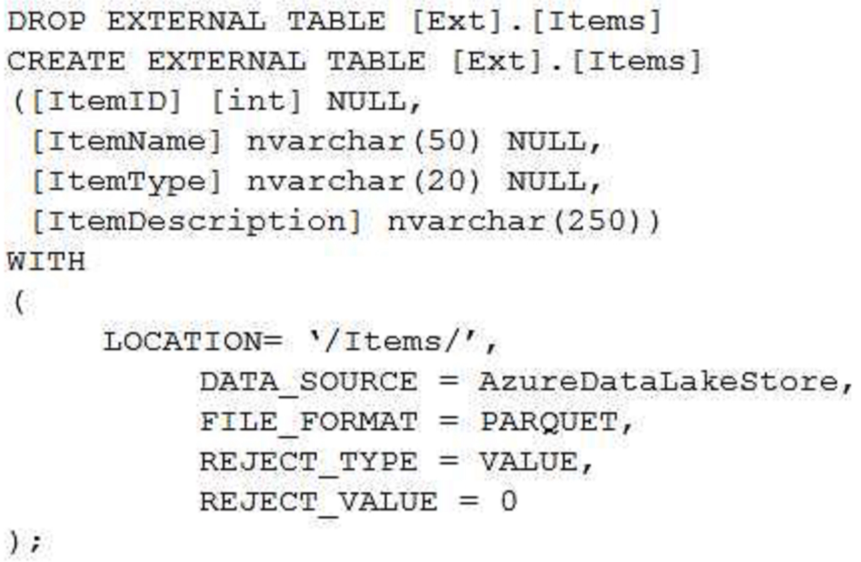
```
ALTER EXTERNAL TABLE [Ext].[Items]
  ADD [ItemID] int;
```

B.



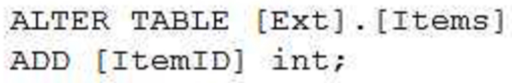
```
DROP EXTERNAL FILE FORMAT parquetfile1;
CREATE EXTERNAL FILE FORMAT parquetfile1
WITH (
     FORMAT_TYPE = PARQUET,
     DATA_COMPRESSION = 'org.apache.hadoop.io.compress.SnappyCodec'
);
```

C.



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

D.



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

A. Option A

B. Option B

C. Option C

D. Option D

Correct Answer: C
Section:
Explanation:
Incorrect Answers:
A, D: Only these Data Definition Language (DDL) statements are allowed on external tables:
CREATE TABLE and DROP TABLE
CREATE STATISTICS and DROP STATISTICS CREATE VIEW and DROP VIEW
Reference: https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql

QUESTION 32
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. geo-redundant storage (GRS)
B. read-access geo-redundant storage (RA-GRS)
C. zone-redundant storage (ZRS)
D. locally-redundant storage (LRS)

Correct Answer: B
Section:
Explanation:
Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages. However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region. When you enable read access to the secondary region, your data is available to be read at all times, including in a situation where the primary region becomes unavailable. Incorrect Answers:
A: While Geo-redundant storage (GRS) is cheaper than Read-Access Geo-Redundant Storage (RA-GRS), GRS does NOT initiate automatic failover. C, D: Locally redundant storage (LRS) and Zone-redundant storage (ZRS) provides redundancy within a single region.
Reference: https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy

QUESTION 33
You plan to implement an Azure Data Lake Gen 2 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?

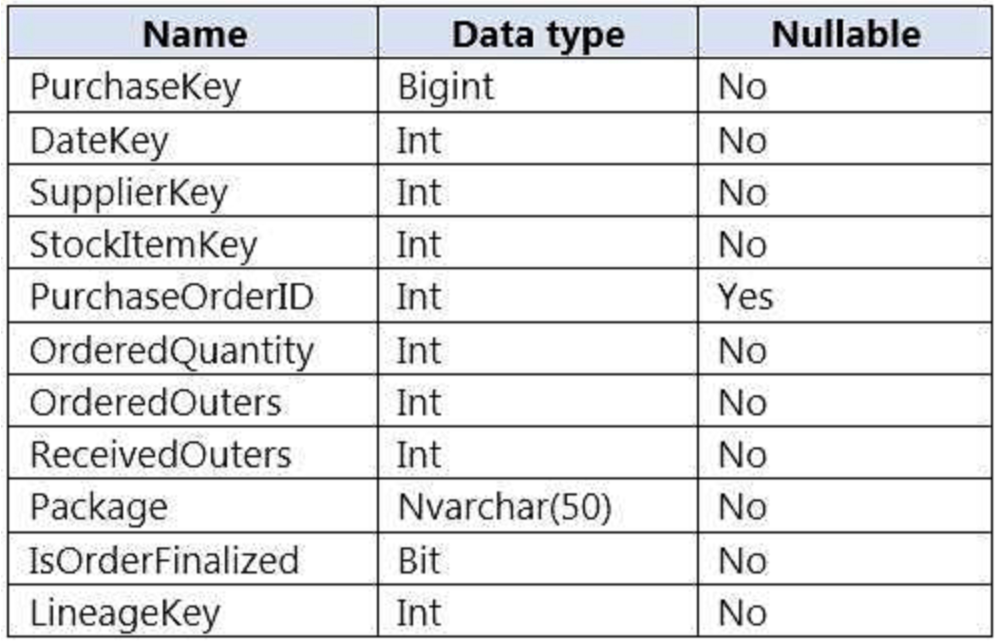
A. geo-redundant storage (GRS)
B. geo-zone-redundant storage (GZRS)
C. locally-redundant storage (LRS)
D. zone-redundant storage (ZRS)

Correct Answer: D
Section:
Explanation:

Reference: https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy

QUESTION 34

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, IsOrderFinalized, COUNT(*)
FROM FactPurchase
WHERE DateKey >= 20210101
AND DateKey <= 20210131
GROUP By SupplierKey, StockItemKey, IsOrderFinalized
```

Which table distribution will minimize query times?

A. replicated

B. hash-distributed on PurchaseKey

C. round-robin

D. hash-distributed on IsOrderFinalized

Correct Answer: B

Section:

Explanation:

Hash-distributed tables improve query performance on large fact tables. To balance the parallel processing, select a distribution column that:

Has many unique values. The column can have duplicate values. All rows with the same value are assigned to the same distribution. Since there are 60 distributions, some distributions can have > 1 unique values while others may end with zero values.

Does not have NULLs, or has only a few NULLs. Is not a date column. Incorrect Answers:

C: 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-distribute

QUESTION 35

You have an Azure Synapse Analytics workspace named WS1 that contains an Apache Spark pool named Pool1. You plan to create a database named DB1 in Pool1.

You need to ensure that when tables are created in DB1, the tables are available automatically as external tables to the built-in serverless SQL pool. Which format should you use for the tables in DB1?

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

Correct Answer: D
Section:
Explanation:
Serverless SQL pool can automatically synchronize metadata from Apache Spark. A serverless SQL pool database will be created for each database existing in serverless Apache Spark pools. For each Spark external table based on Parquet or CSV and located in Azure Storage, an external table is created in a serverless SQL pool database.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-storage-files-spark-tables

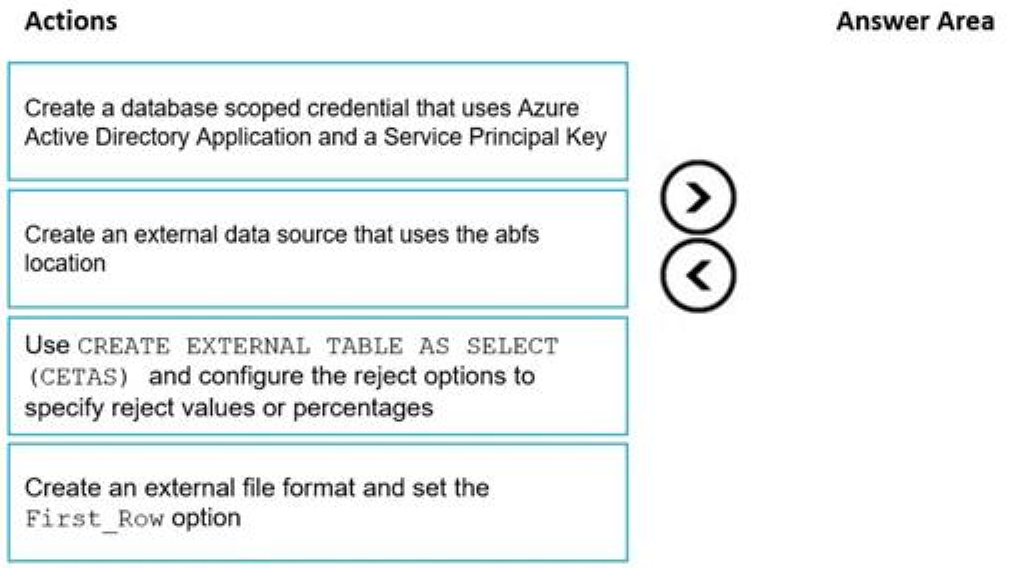
QUESTION 36
You are planning a solution to aggregate streaming data that originates in Apache Kafka and is output to Azure Data Lake Storage Gen2. The developers who will implement the stream processing solution use Java. Which service should you recommend using to process the streaming data?

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

Correct Answer: D
Section:
Explanation:

QUESTION 37
DRAG DROP
You have data stored in thousands of CSV files in Azure Data Lake Storage Gen2. Each file has a header row followed by a properly formatted carriage return (/r) and line feed (/n).
You are implementing a pattern that batch loads the files daily into an enterprise data warehouse in Azure Synapse Analytics by using PolyBase.
You need to skip the header row when you import the files into the data warehouse. Before building the loading pattern, you need to prepare the required database objects in Azure Synapse Analytics.
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: Each correct selection is worth one point

Select and Place:



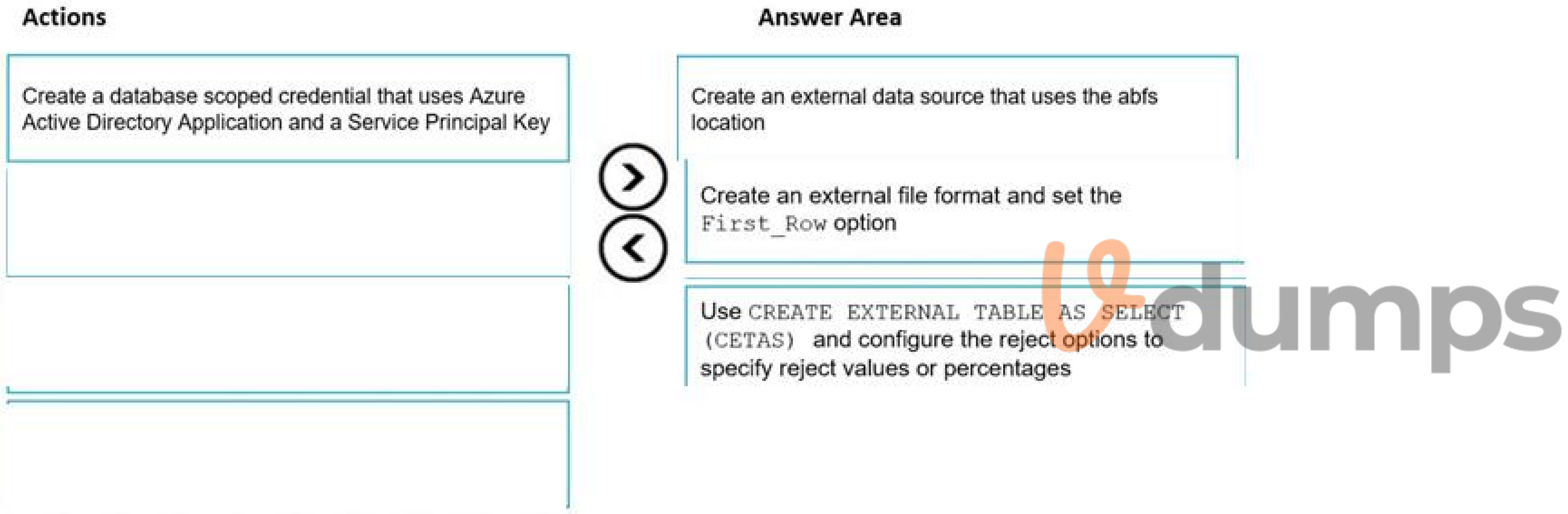
Create a database scoped credential that uses Azure Active Directory Application and a Service Principal Key

Create an external data source that uses the abfs location

Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages

Create an external file format and set the First_Row option

Correct Answer:



Create a database scoped credential that uses Azure Active Directory Application and a Service Principal Key

Create an external data source that uses the abfs location

Create an external file format and set the First_Row option

Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages

Section:

Explanation:

Step 1: Create an external data source that uses the abfs location Create External Data Source to reference Azure Data Lake Store Gen 1 or 2

Step 2: Create an external file format and set the First_Row option. Create External File Format.

Step 3: Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages To use PolyBase, you must create external tables to reference your external data. Use reject options.

Note: REJECT options don't apply at the time this CREATE EXTERNAL TABLE AS SELECT statement is run. Instead, they're specified here so that the database can use them at a later time when it imports data from the external table. Later, when the CREATE TABLE AS SELECT statement selects data from the external table, the database will use the reject options to determine the number or percentage of rows that can fail to import before it stops the import.

Reference:

https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects

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

QUESTION 38

HOTSPOT

You are building an Azure Synapse Analytics dedicated SQL pool that will contain a fact table for transactions from the first half of the year 2020.

You need to ensure that the table meets the following requirements:

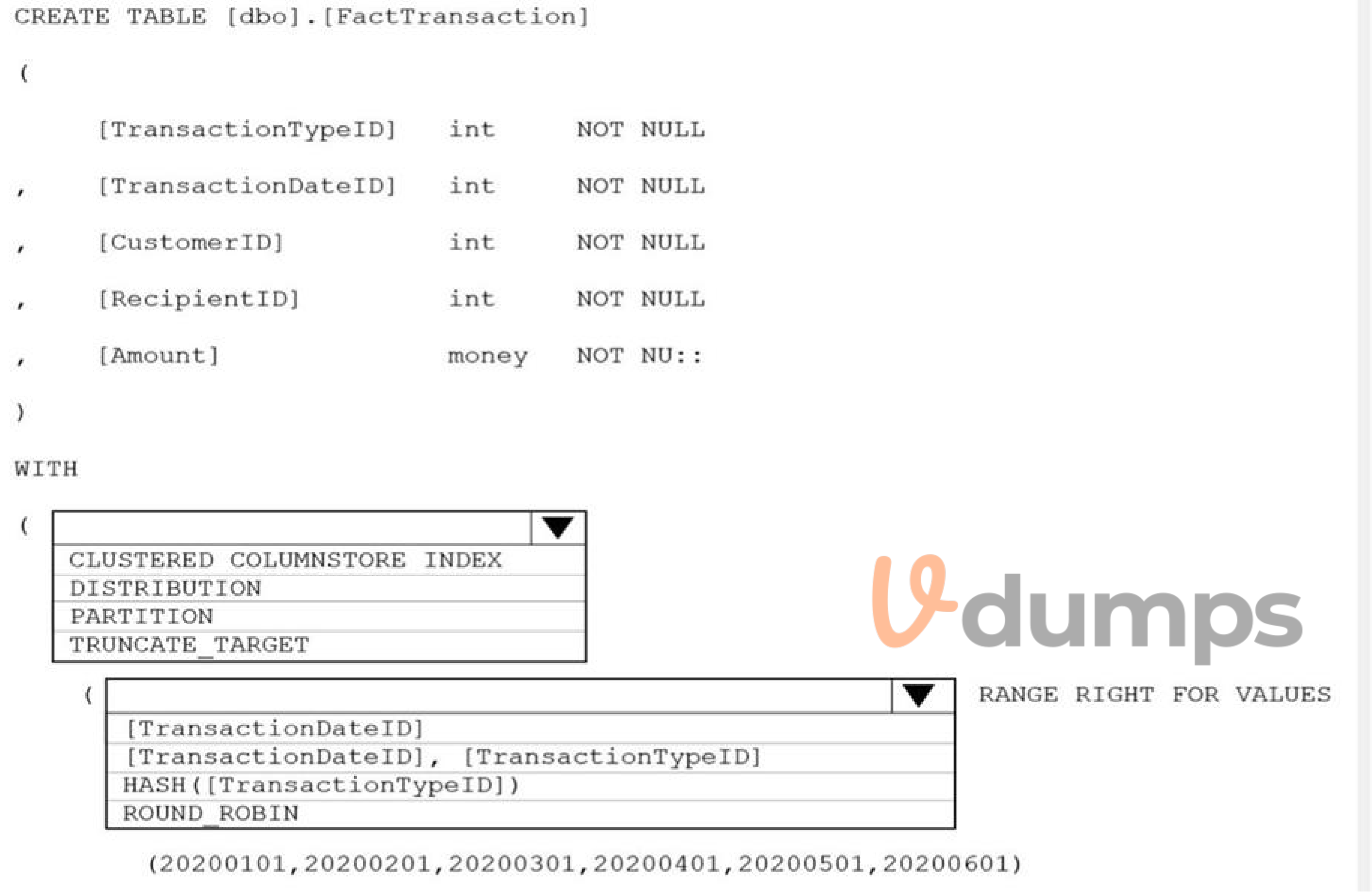
Minimizes the processing time to delete data that is older than 10 years

Minimizes the I/O for queries that use year-to-date values

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

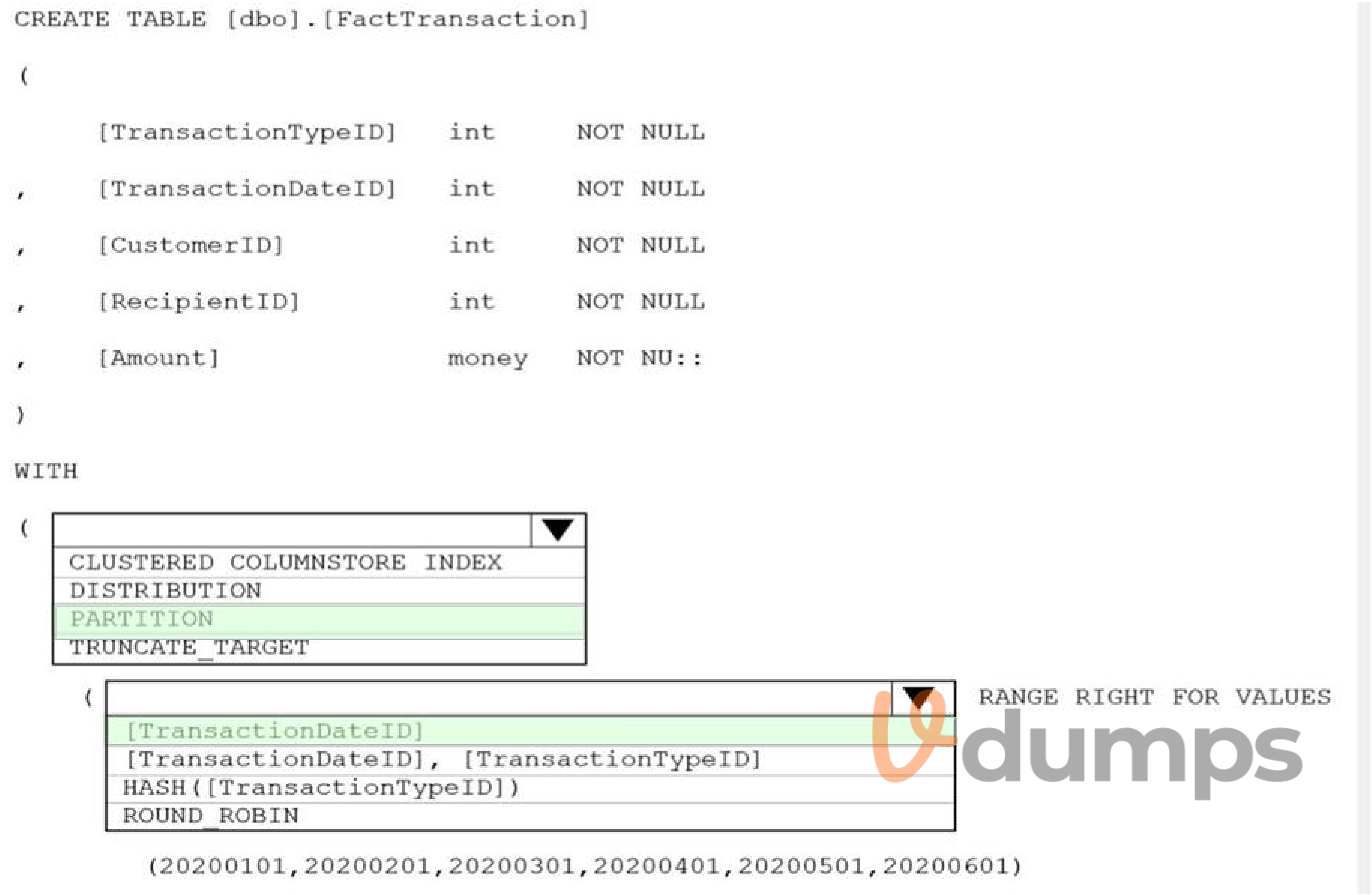
NOTE: Each correct selection is worth one point.

Hot Area:



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

Answer Area:



```
CREATE TABLE [dbo].[FactTransaction]
(
      [TransactionTypeID]   int     NOT NULL
,     [TransactionDateID]   int     NOT NULL
,     [CustomerID]          int     NOT NULL
,     [RecipientID]         int     NOT NULL
,     [Amount]              money   NOT NU::
)
WITH
(
    (                                        RANGE RIGHT FOR VALUES
        (20200101,20200201,20200301,20200401,20200501,20200601)
```

CLUSTERED COLUMNSTORE INDEX
DISTRIBUTION
PARTITION
TRUNCATE_TARGET
[TransactionDateID]
[TransactionDateID], [TransactionTypeID]
HASH([TransactionTypeID])
ROUND_ROBIN

Section:

Explanation:

Box 1: PARTITION

RANGE RIGHT FOR VALUES is used with PARTITION.

Part 2: [TransactionDateID]

Partition on the date column.

Example: Creating a RANGE RIGHT partition function on a datetime column The following partition function partitions a table or index into 12 partitions, one for each month of a year's worth of values in a datetime column.

CREATE PARTITION FUNCTION [myDateRangePF1] (datetime)
AS RANGE RIGHT FOR VALUES ('20030201', '20030301', '20030401', '20030501', '20030601', '20030701', '20030801',
'20030901', '20031001', '20031101', '20031201');

Reference:

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

Case 01 - Design and develop data processing

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.

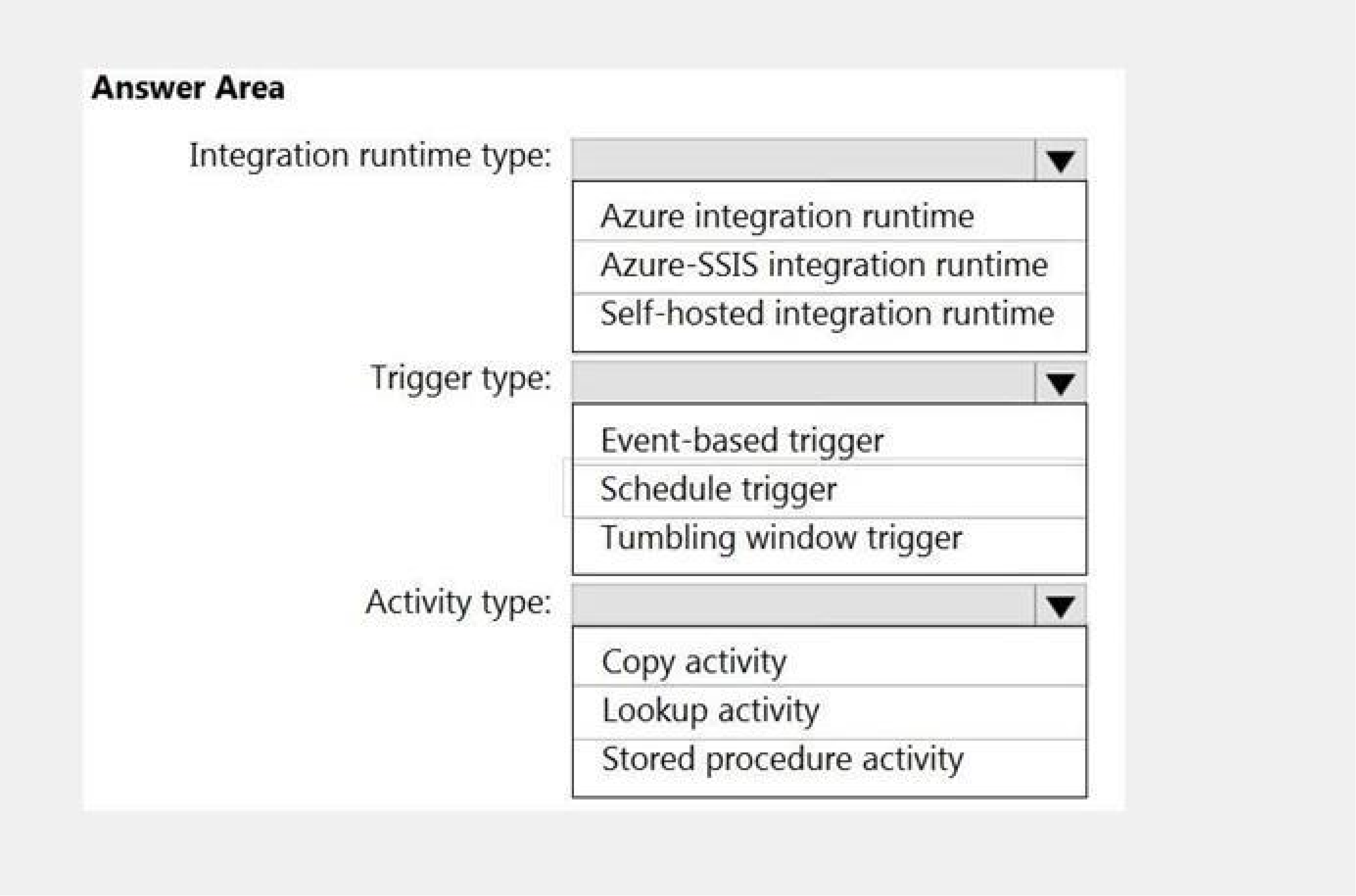
Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table. Product data, including product ID, name, and category, comes from Salesforce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

QUESTION 1
HOTSPOT

Which Azure Data Factory components should you recommend using together to import the daily inventory data from the SQL server to Azure Data Lake Storage? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

Integration runtime type:

- Azure integration runtime
- Azure-SSIS integration runtime
- Self-hosted integration runtime

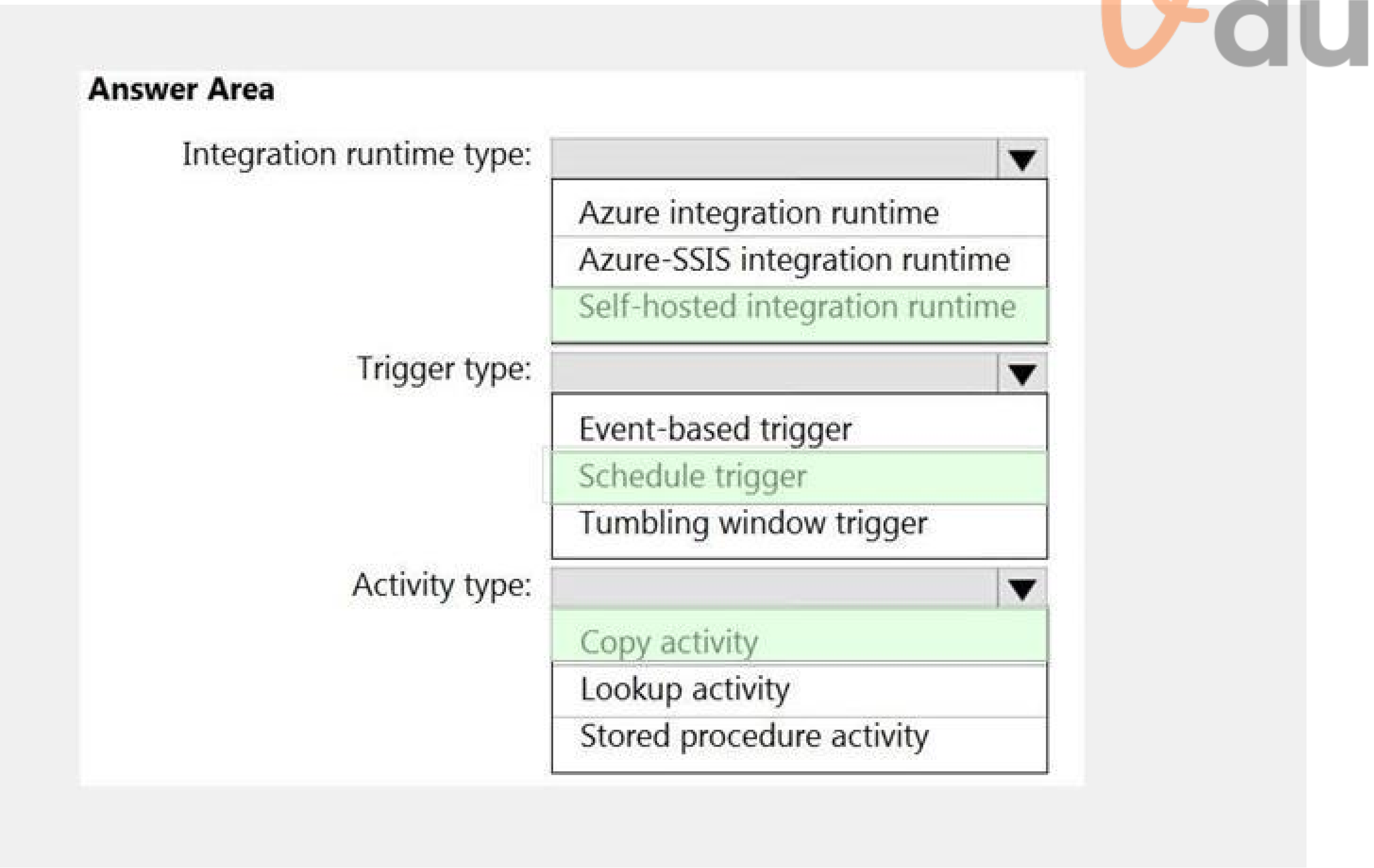
Trigger type:

- Event-based trigger
- Schedule trigger
- Tumbling window trigger

Activity type:

- Copy activity
- Lookup activity
- Stored procedure activity

Answer Area:



Answer Area

Integration runtime type:

- Azure integration runtime
- Azure-SSIS integration runtime
- [✓ Self-hosted integration runtime]

Trigger type:

- Event-based trigger
- [✓ Schedule trigger]
- Tumbling window trigger

Activity type:

- [✓ Copy activity]
- Lookup activity
- Stored procedure activity

Section:

Explanation:

Explanation:

Box 1: Self-hosted integration runtime
A self-hosted IR is capable of running copy activity between a cloud data stores and a data store in private network.

Box 2: Schedule trigger
Schedule every 8 hours

Box 3: Copy activity

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

Case 02 - Design and develop data processing

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

Contoso, Ltd. is a clothing retailer based in Seattle. The company has 2,000 retail stores across the United States and an emerging online presence.

The network contains an Active Directory forest named contoso.com. The forest it integrated with an Azure Active Directory (Azure AD) tenant named contoso.com. Contoso has an Azure subscription associated to the contoso.com Azure AD tenant.

Existing Environment

Transactional Data

Contoso has three years of customer, transactional, operational, 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 periods. 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 an Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

Planned Changes and Requirements

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 Requirements

Contoso identifies the following requirements for customer sentiment analytics:

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.

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 that 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 Synapse 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.

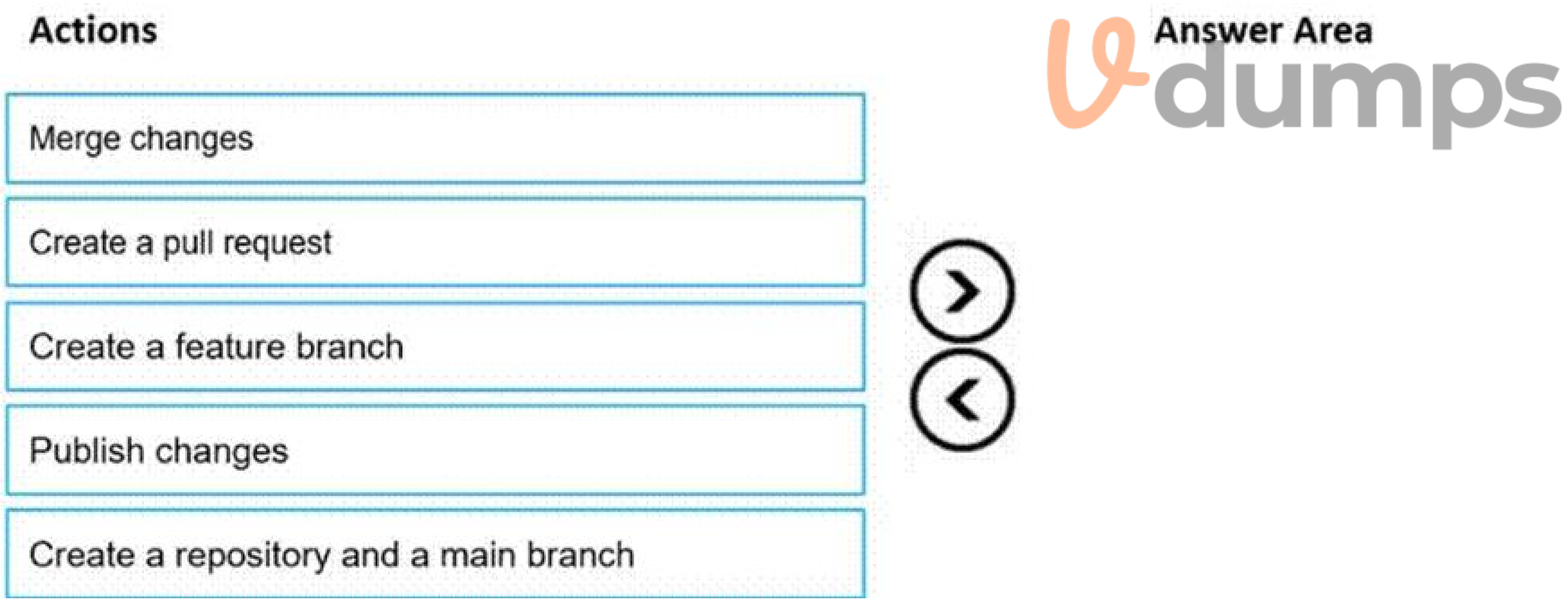
QUESTION 1

DRAG DROP

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 the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

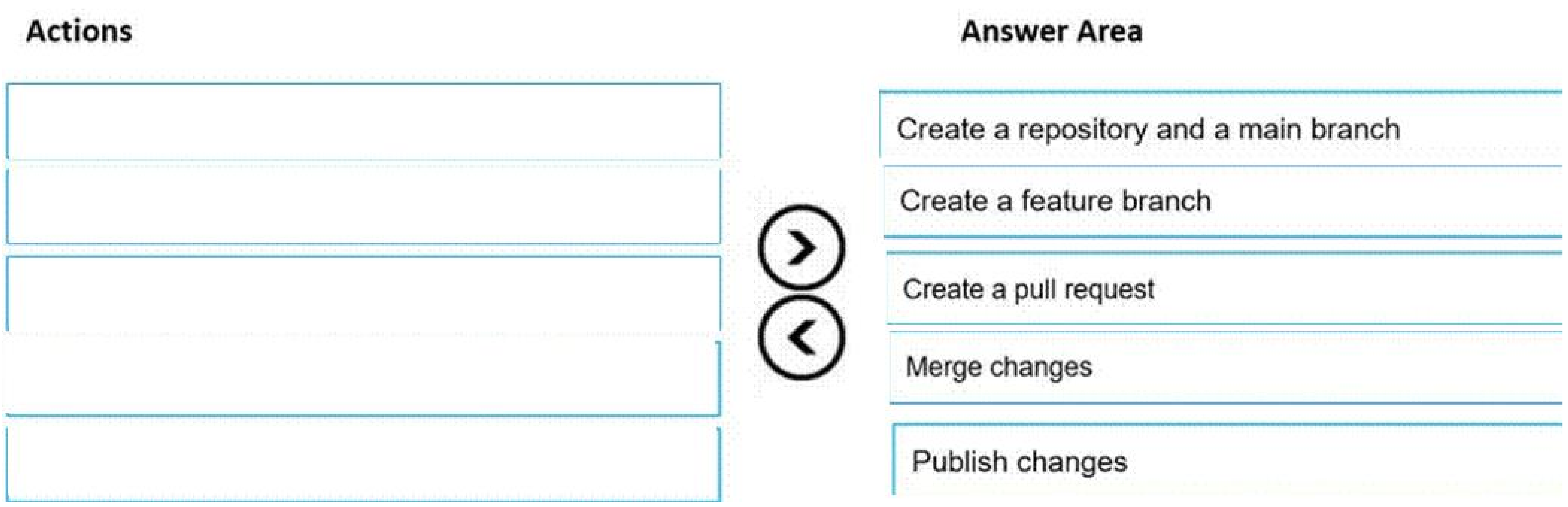


Actions

- Merge changes
- Create a pull request
- Create a feature branch
- Publish changes
- Create a repository and a main branch

Answer Area

Correct Answer:



Actions

Answer Area

Create a repository and a main branch

Create a feature branch

Create a pull request

Merge changes

Publish changes

Section:
Explanation:

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

02 - Design and implement data security

QUESTION 1
You have an Azure Synapse Analytics dedicated SQL pool.
You need to ensure that data in the pool is encrypted at rest. The solution must NOT require modifying applications that query the data. What should you do?

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

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

QUESTION 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

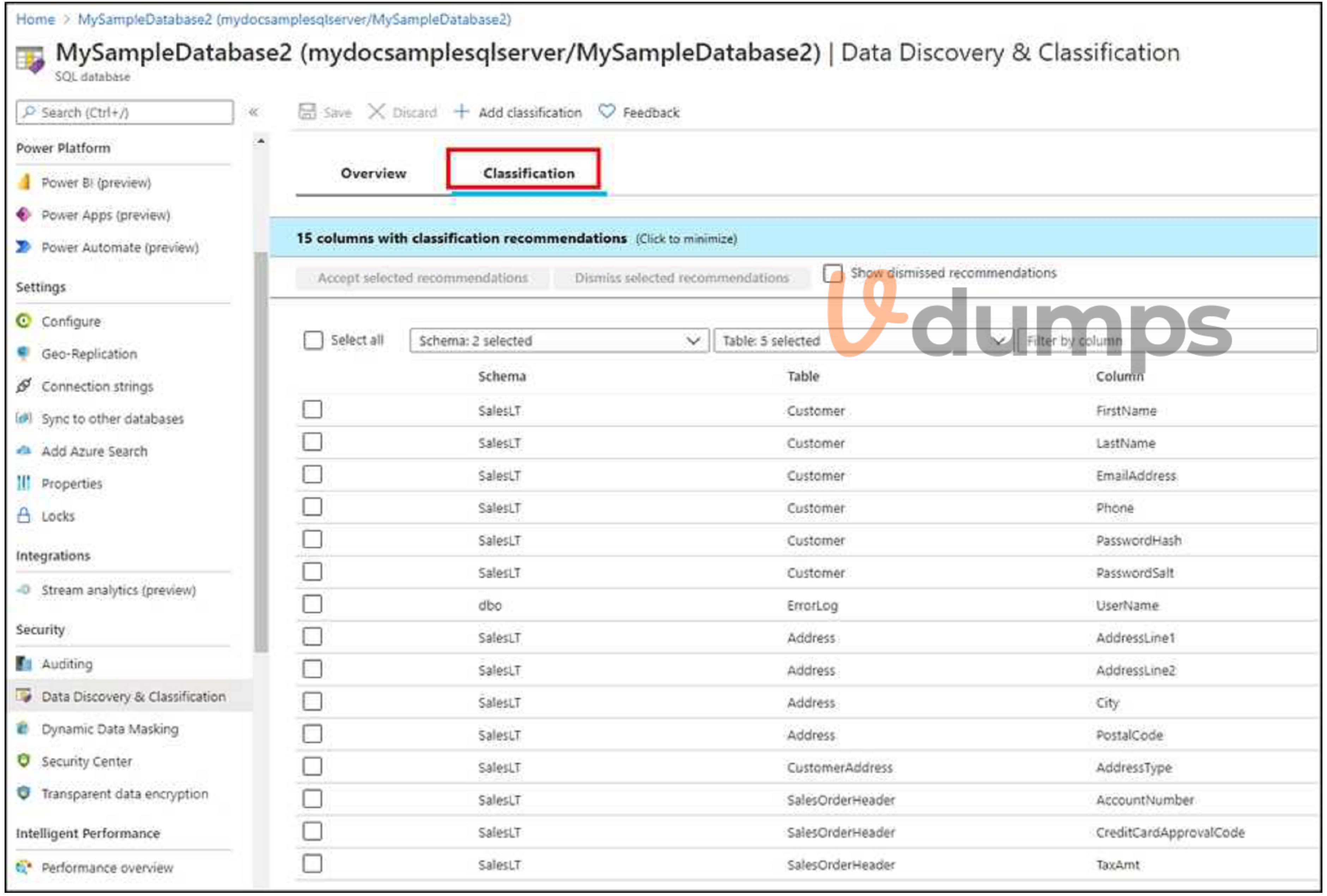
D. dynamic data masking for columns that contain confidential information

Correct Answer: A, C

Section:

Explanation:

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

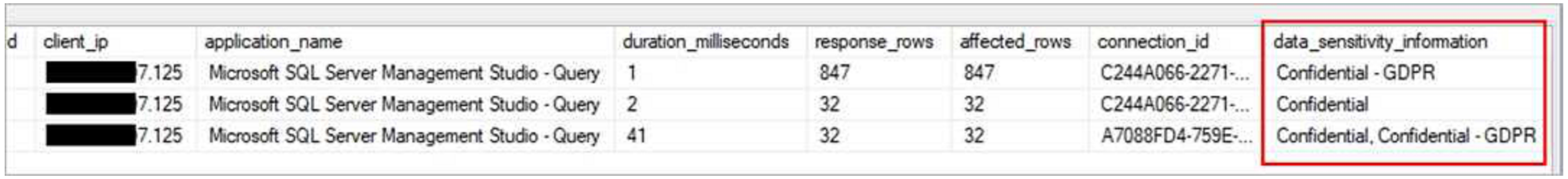


	Schema	Table	Column
□	SalesLT	Customer	FirstName
□	SalesLT	Customer	LastName
□	SalesLT	Customer	EmailAddress
□	SalesLT	Customer	Phone
□	SalesLT	Customer	PasswordHash
□	SalesLT	Customer	PasswordSalt
□	dbo	ErrorLog	UserName
□	SalesLT	Address	AddressLine1
□	SalesLT	Address	AddressLine2
□	SalesLT	Address	City
□	SalesLT	Address	PostalCode
□	SalesLT	CustomerAddress	AddressType
□	SalesLT	SalesOrderHeader	AccountNumber
□	SalesLT	SalesOrderHeader	CreditCardApprovalCode
□	SalesLT	SalesOrderHeader	TaxAmt

Select Add classification in the top menu of the pane.

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

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



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

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

QUESTION 3
You are designing an enterprise data warehouse in Azure Synapse Analytics that will contain a table named Customers. Customers will contain credit card information. You need to recommend a solution to provide salespeople with the ability to view all the entries in Customers. The solution must prevent all the salespeople from viewing or inferring the credit card information. What should you include in the recommendation?

A. data masking
B. Always Encrypted
C. column-level security
D. row-level security

Correct Answer: A
Section:
Explanation:
SQL Database dynamic data masking limits sensitive data exposure by masking it to non-privileged users. The Credit card masking method exposes the last four digits of the designated fields and adds a constant string as a prefix in the form of a credit card. Example: XXXX-XXXX-XXXX-1234
Reference: https://docs.microsoft.com/en-us/azure/sql-database/sql-database-dynamic-data-masking-get-started

QUESTION 4
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. Create security groups in Azure Active Directory (Azure AD) and add project members.
B. Configure end-user authentication for the Azure Data Lake Storage account.
C. Assign Azure AD security groups to Azure Data Lake Storage.
D. Configure Service-to-service authentication for the Azure Data Lake Storage account.
E. Configure access control lists (ACL) for the Azure Data Lake Storage account.

Correct Answer: A, C, E
Section:
Explanation:
AC: Create security groups in Azure Active Directory. Assign users or security groups to Data Lake Storage Gen1 accounts. E: Assign users or security groups as ACLs to the Data Lake Storage Gen1 file system
Reference: https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-secure-data

QUESTION 5
You have an Azure Data Factory version 2 (V2) resource named Df1. Df1 contains a linked service. You have an Azure Key vault named vault1 that contains an encryption key named key1. You need to encrypt Df1 by using key1.
What should you do first?

A. Add a private endpoint connection to vaul1.

B. Enable Azure role-based access control on vault1.

C. Remove the linked service from Df1.

D. Create a self-hosted integration runtime.

Correct Answer: C
Section:
Explanation:
Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources.
Incorrect Answers:
D: A self-hosted integration runtime copies data between an on-premises store and cloud storage.
Reference: https://docs.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key https://docs.microsoft.com/en-us/azure/data-factory/concepts-linked-services https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime

QUESTION 6
You are designing an Azure Synapse Analytics dedicated SQL pool. You need to ensure that you can audit access to Personally Identifiable Information (PII). What should you include in the solution?

A. column-level security

B. dynamic data masking

C. row-level security (RLS)

D. sensitivity classifications

Correct Answer: D
Section:
Explanation:
Data Discovery & Classification is built into Azure SQL Database, Azure SQL Managed Instance, and Azure Synapse Analytics. It provides basic capabilities for discovering, classifying, labeling, and reporting the sensitive data in your databases.
Your most sensitive data might include business, financial, healthcare, or personal information. Discovering and classifying this data can play a pivotal role in your organization's information-protection approach. It can serve as infrastructure for:
Helping to meet standards for data privacy and requirements for regulatory compliance. Various security scenarios, such as monitoring (auditing) access to sensitive data. Controlling access to and hardening the security of databases that contain highly sensitive data.
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview

QUESTION 7
You have a data warehouse in Azure Synapse Analytics.
You need to ensure that the data in the data warehouse is encrypted at rest. What should you enable?

A. Advanced Data Security for this database

B. Transparent Data Encryption (TDE)

C. Secure transfer required

D. Dynamic Data Masking

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

QUESTION 8
You are designing a streaming data solution that will ingest variable volumes of data. You need to ensure that you can change the partition count after creation. Which service should you use to ingest the data?

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

Correct Answer: B
Section:

QUESTION 9
You are designing a date dimension table in an Azure Synapse Analytics dedicated SQL pool. The date dimension table will be used by all the fact tables. Which distribution type should you recommend to minimize data movement?

A. HASH
B. REPLICATE
C. ROUND_ROBIN

Correct Answer: B
Section:
Explanation:
A replicated table has a full copy of the table available on every Compute node. Queries run fast on replicated tables since joins on replicated tables don't require data movement. Replication requires extra storage, though, and isn't practical for large tables.
Incorrect Answers:
A: A hash distributed table is designed to achieve high performance for queries on large tables. C: A round-robin table distributes table rows evenly across all distributions. The rows are distributed randomly. Loading data into a round-robin table is fast. Keep in mind that queries can require more data movement than the other distribution methods.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overview

QUESTION 10
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 is worth one point.

A. a security policy
B. a custom role-based access control (RBAC) role
C. a function
D. a column encryption key
E. asymmetric keys

Correct Answer: A, B
Section:
Explanation:
A: Row-Level Security (RLS) enables you to use group membership or execution context to control access to rows in a database table. Implement RLS by using the CREATE SECURITY POLICYTransact-SQL statement. B: Azure Synapse provides a comprehensive and fine-grained access control system, that integrates:
Azure roles for resource management and access to data in storage, Synapse roles for managing live access to code and execution, SQL roles for data plane access to data in SQL pools.
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 11

HOTSPOT

You develop a dataset named DBTBL1 by using Azure Databricks.

DBTBL1 contains the following columns:

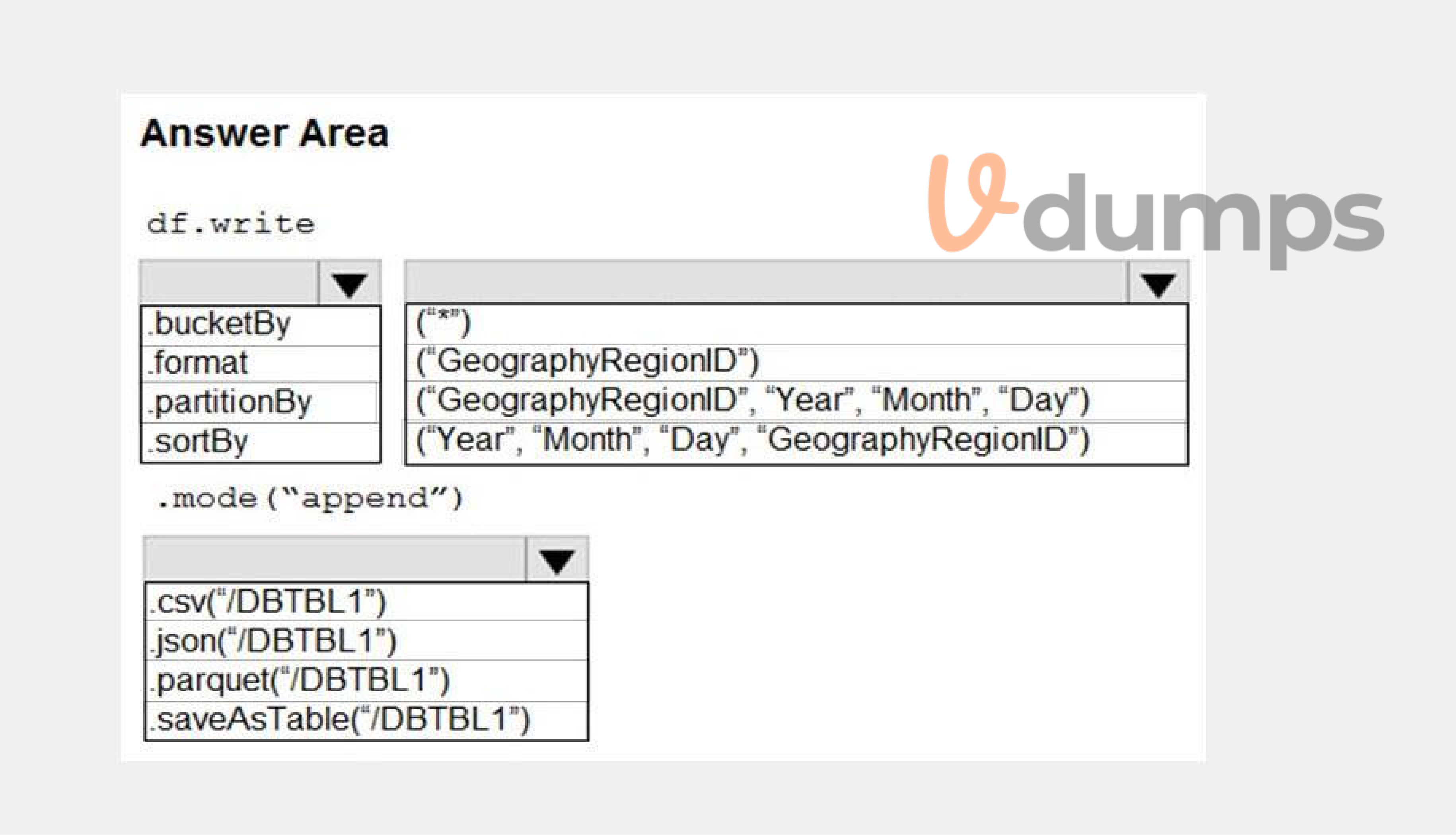
SensorTypeID
GeographyRegionID
Year
Month
Day
Hour
Minute
Temperature
WindSpeed
Other

You need to store the data to support daily incremental load pipelines that vary for each GeographyRegionID. The solution must minimize storage costs.

How should you complete the code? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

```
df.write
```

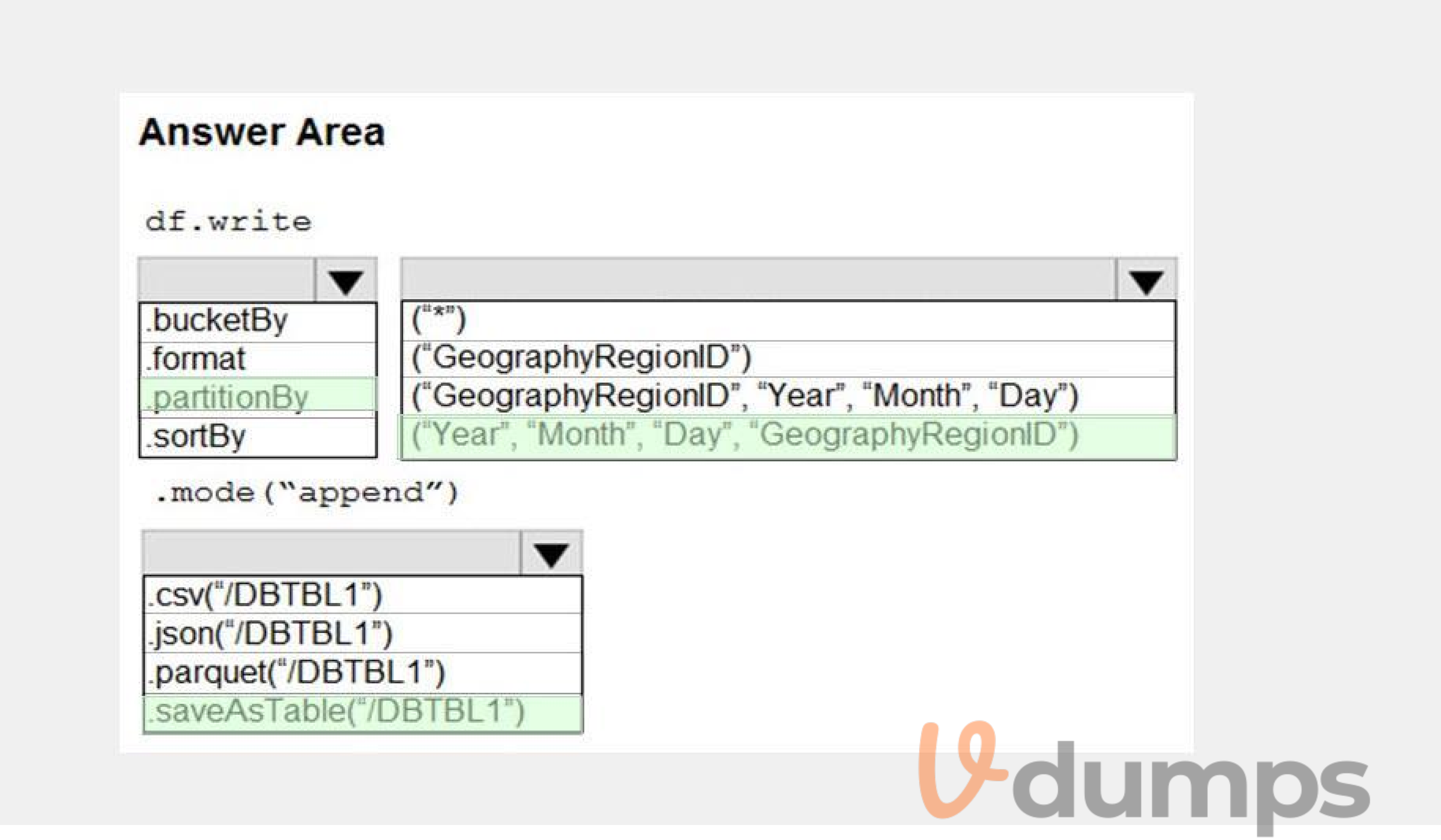
- .bucketBy
- .format
- .partitionBy
- .sortBy

- ("*")
- ("GeographyRegionID")
- ("GeographyRegionID", "Year", "Month", "Day")
- ("Year", "Month", "Day", "GeographyRegionID")

```
.mode("append")
```

- .csv("/DBTBL1")
- .json("/DBTBL1")
- .parquet("/DBTBL1")
- .saveAsTable("/DBTBL1")

Answer Area:



Answer Area

```
df.write
```

- .bucketBy
- .format
- .partitionBy
- .sortBy

- ("*")
- ("GeographyRegionID")
- ("GeographyRegionID", "Year", "Month", "Day")
- ("Year", "Month", "Day", "GeographyRegionID")

```
.mode("append")
```

- .csv("/DBTBL1")
- .json("/DBTBL1")
- .parquet("/DBTBL1")
- .saveAsTable("/DBTBL1")

Section:
Explanation:
Box 1: .partitionBy
Incorrect Answers:
.format:
Method: format():
Arguments: "parquet", "csv", "txt", "json", "jdbc", "orc", "avro", etc.
.bucketBy:
Method: bucketBy()
Arguments: (numBuckets, col, col..., coln)
The number of buckets and names of columns to bucket by. Uses Hive's bucketing scheme on a filesystem.
Box 2: ("Year", "Month", "Day","GeographyRegionID")
Specify the columns on which to do the partition. Use the date columns followed by the GeographyRegionID column.
Box 3: .saveAsTable("/DBTBL1")
Method: saveAsTable()
Argument: "table_name"
The table to save to.
Reference:
https://www.oreilly.com/library/view/learning-spark-2nd/9781492050032/ch04.html
https://docs.microsoft.com/en-us/azure/databricks/delta/delta-batch

QUESTION 12
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

Correct Answer: A
Section:
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

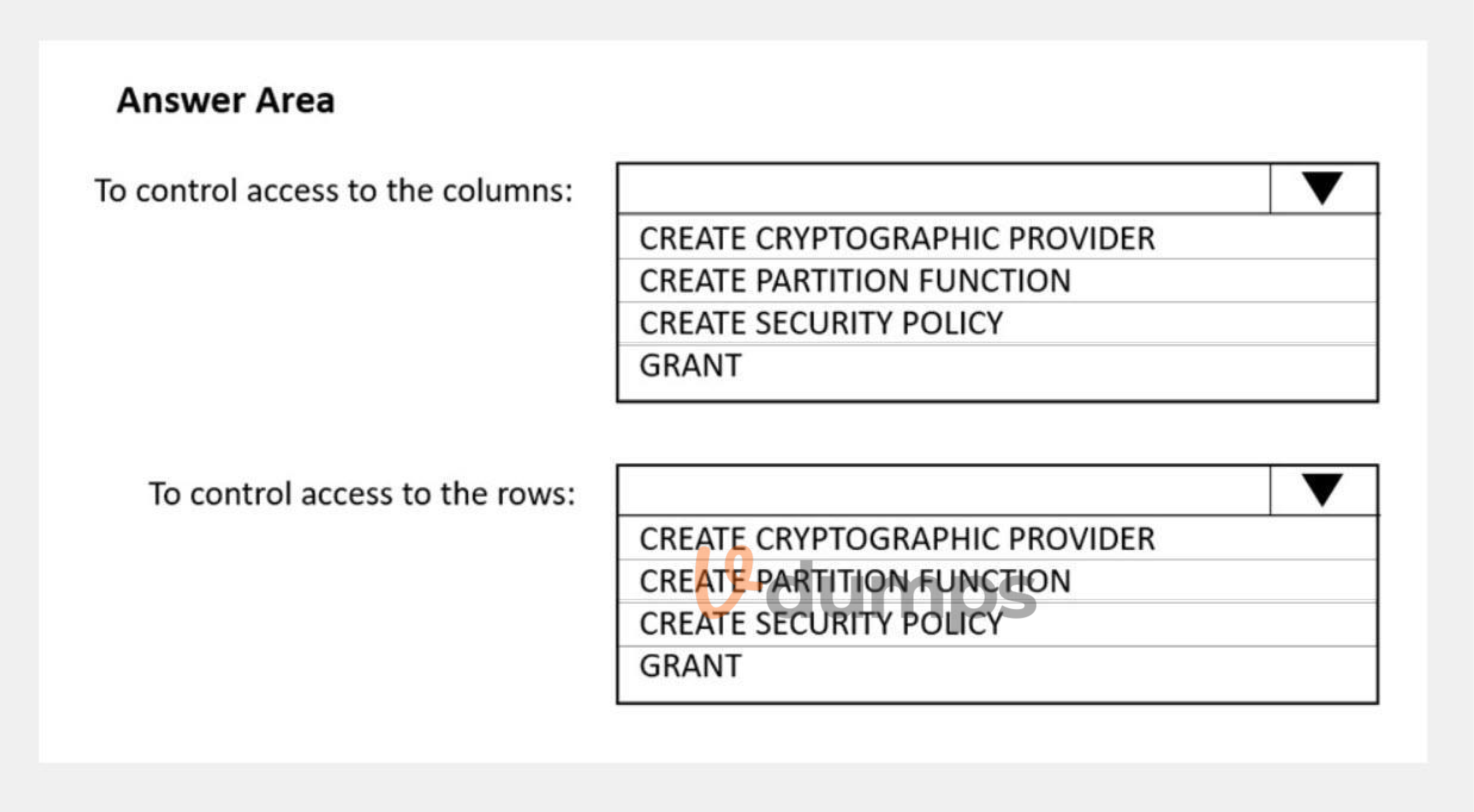
QUESTION 13
You are developing an application that uses Azure Data Lake Storage Gen2. You need to recommend a solution to grant permissions to a specific application for a limited time period. What should you include in the recommendation?

A. role assignments
B. shared access signatures (SAS)
C. Azure Active Directory (Azure AD) identities
D. account keys

Correct Answer: B
Section:
Explanation:
A shared access signature (SAS) provides secure delegated access to resources in your storage account. With a SAS, you have granular control over how a client can access your data. For example:
What resources the client may access.
What permissions they have to those resources. How long the SAS is valid.
Reference: https://docs.microsoft.com/en-us/azure/storage/common/storage-sas-overview

QUESTION 14
HOTSPOT
You have an Azure Synapse Analytics SQL pool named Pool1. In Azure Active Directory (Azure AD), you have a security group named Group1.
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 correct selection is worth one point.

Hot Area:



Answer Area

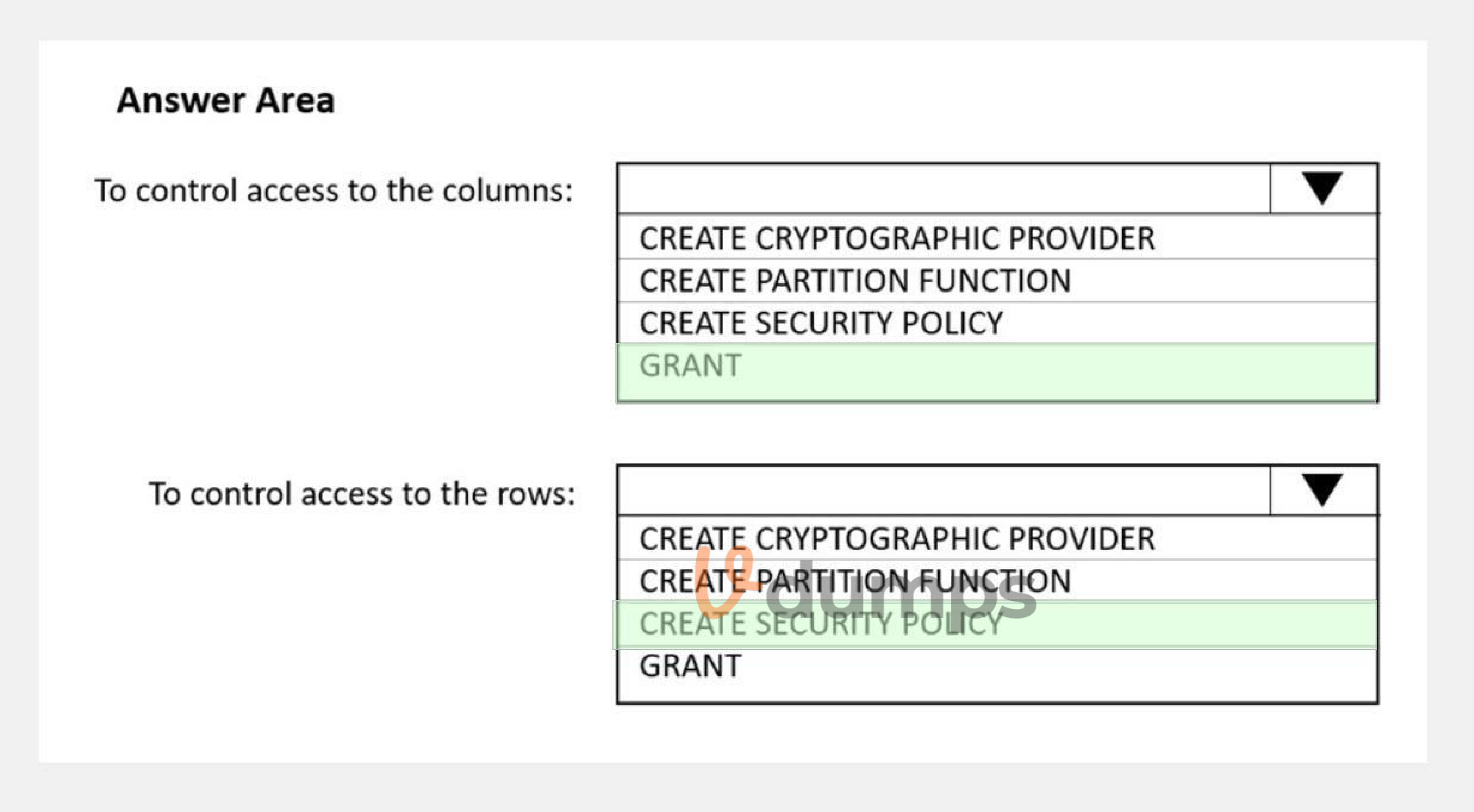
To control access to the columns:

- 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

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:

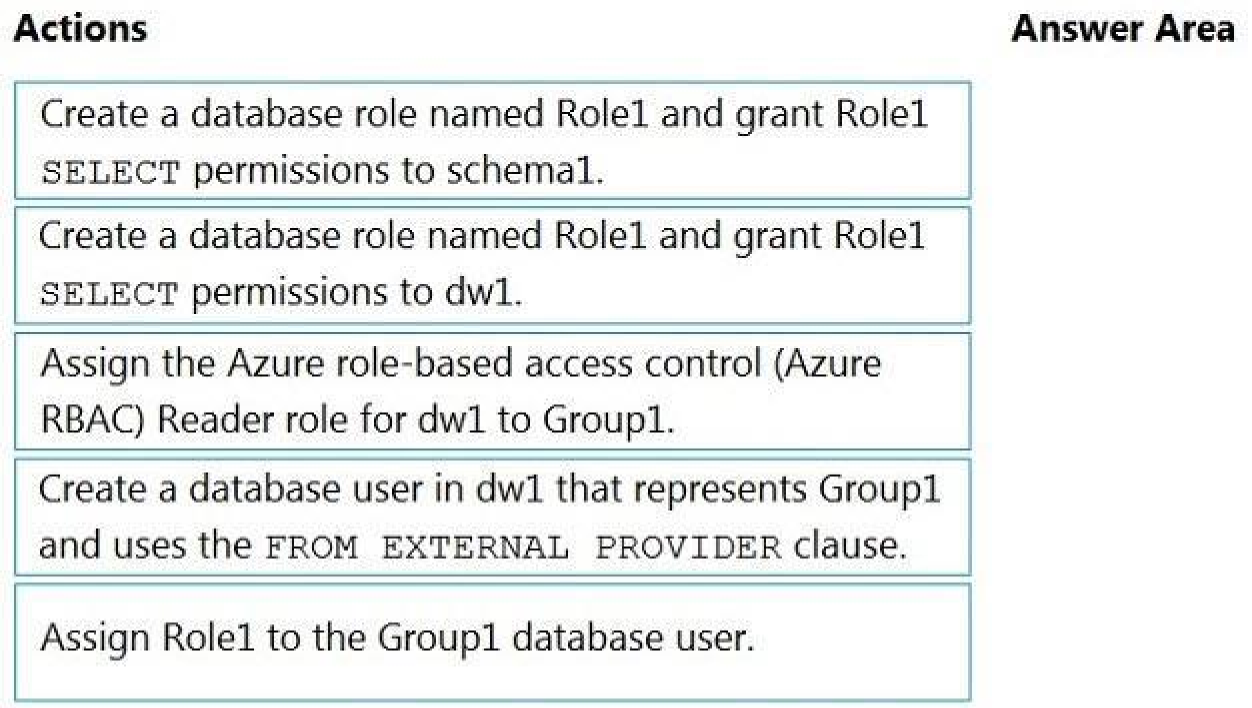
- CREATE CRYPTOGRAPHIC PROVIDER
- CREATE PARTITION FUNCTION
- [✓ CREATE SECURITY POLICY]
- GRANT

Section:
Explanation:
Box 1: GRANT
You can implement column-level security with the GRANT T-SQL statement. With this mechanism, both SQL and Azure Active Directory (Azure AD) authentication are supported.
Box 2: CREATE SECURITY POLICY
Implement RLS by using the CREATE SECURITY POLICY Transact-SQL statement, and predicates created as inline table-valued functions.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security
https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security

QUESTION 15
DRAG DROP
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.

Select and Place:

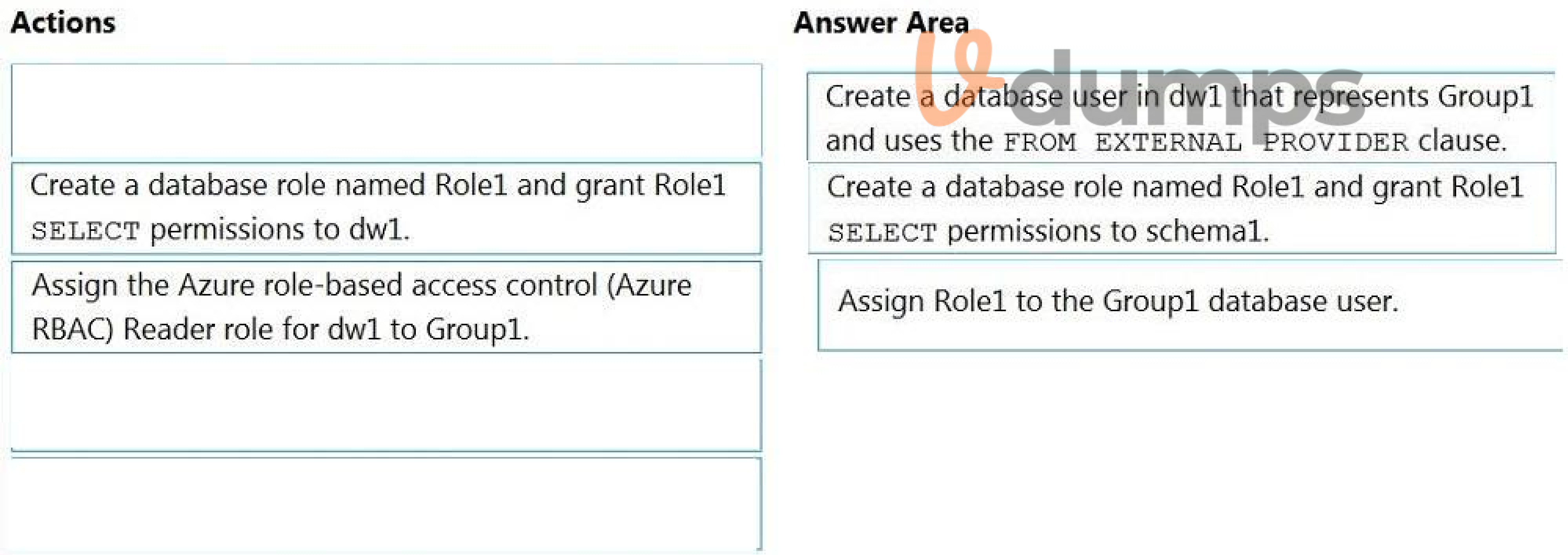


Actions

- 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

Correct Answer:



Actions

- 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.

Answer Area

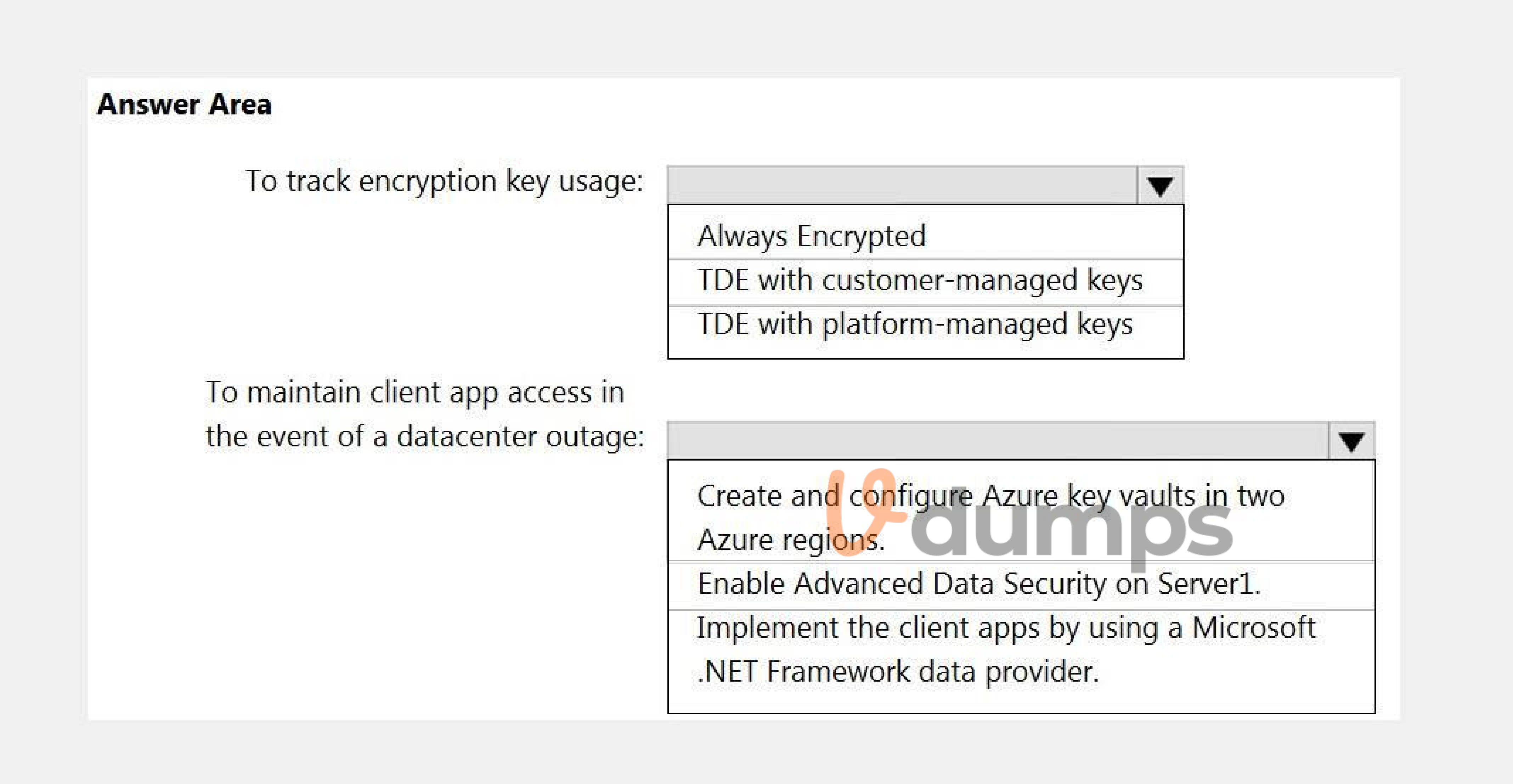
- Create a database user in dw1 that represents Group1 and uses the `FROM EXTERNAL PROVIDER` clause.
- Create a database role named Role1 and grant Role1 `SELECT` permissions to schema1.
- Assign Role1 to the Group1 database user.

Section:
Explanation:
Step 1: Create a database user named dw1 that represents Group1 and use the FROM EXTERNAL PROVIDER clause. Step 2: Create a database role named Role1 and grant Role1 SELECT permissions to schema1. Step 3: Assign Role1 to the Group1 database user.
Reference: https://docs.microsoft.com/en-us/azure/data-share/how-to-share-from-sql

QUESTION 16
HOTSPOT
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.

Hot Area:

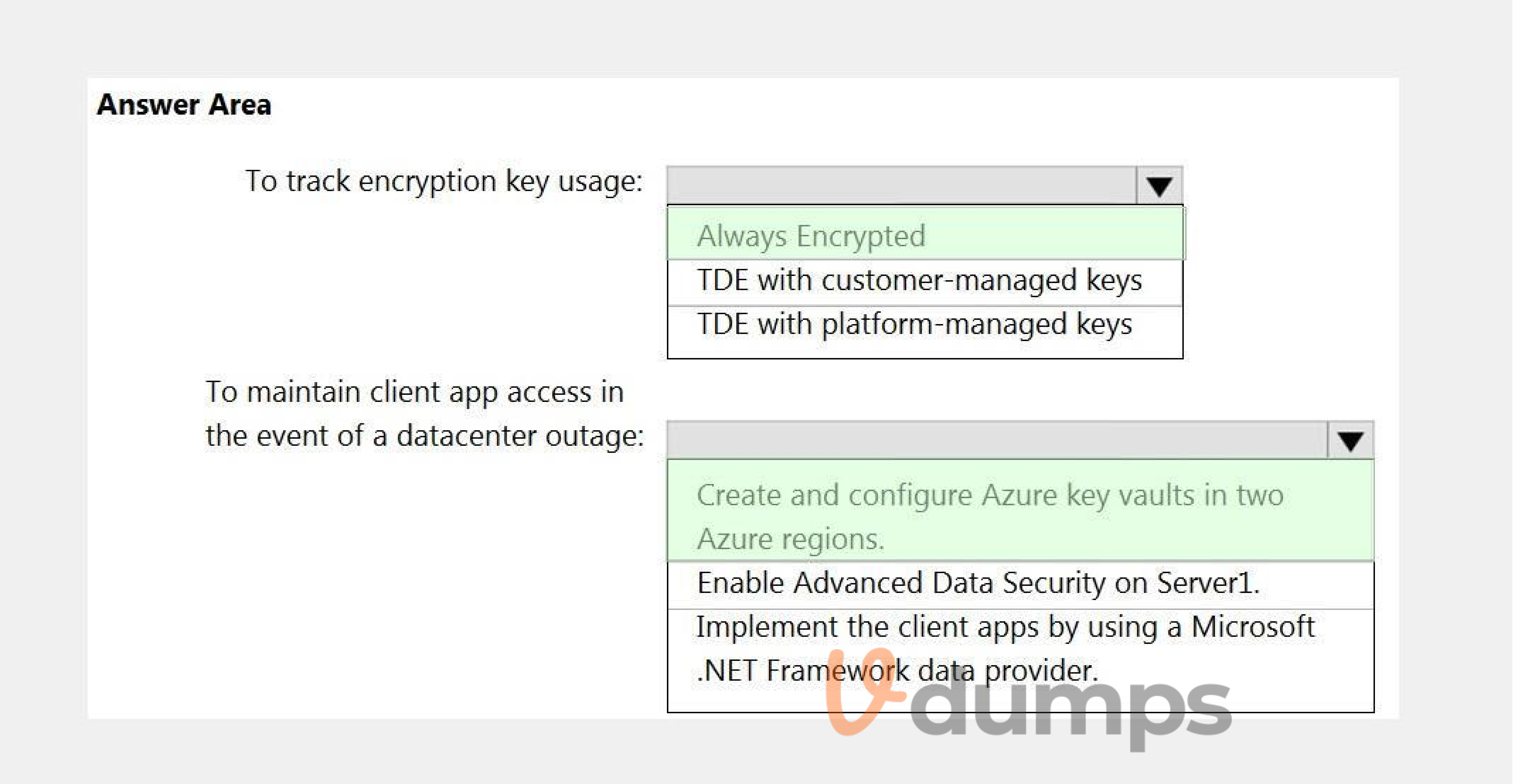


Answer Area

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.

Answer Area:



Answer Area

To track encryption key usage: ▼

- [✓ Always Encrypted]
- TDE with customer-managed keys
- TDE with platform-managed keys

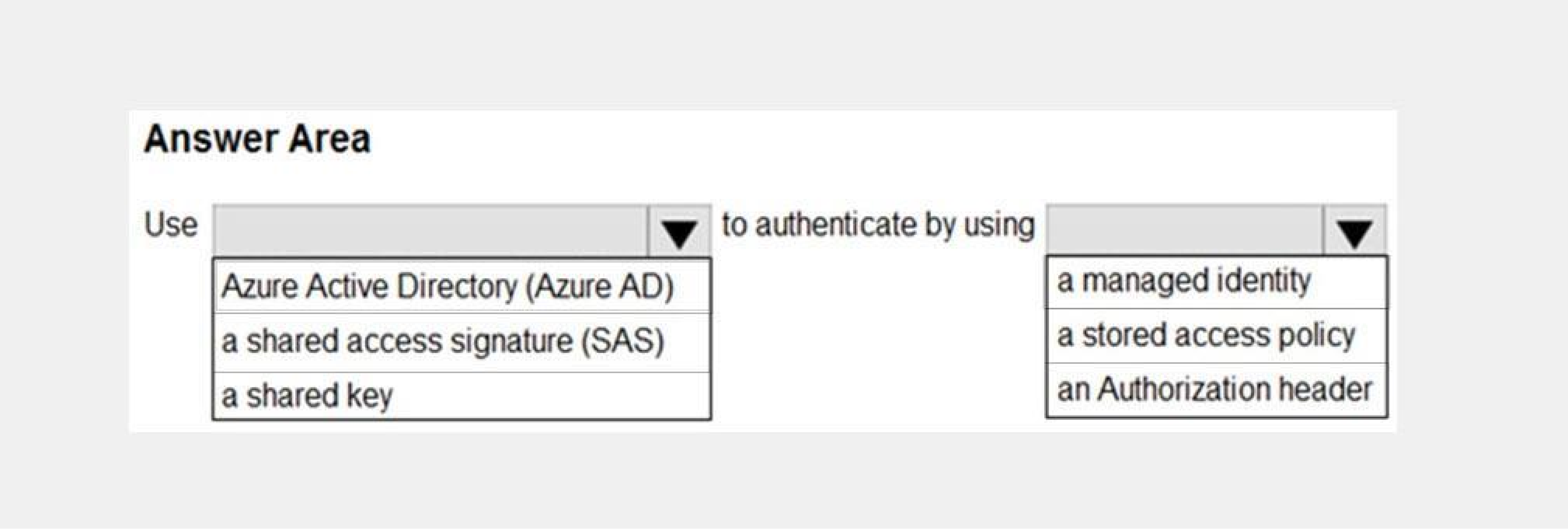
To maintain client app access in the event of a datacenter outage: ▼

- [✓ 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.

Section:
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

QUESTION 17
HOTSPOT
You have an Azure subscription that contains an Azure Data Lake Storage account. The storage account contains a data lake named DataLake1.
You plan to use an Azure data factory to ingest data from a folder in DataLake1, transform the data, and land the data in another folder.
You need to ensure that the data factory can read and write data from any folder in the DataLake1 file system. The solution must meet the following requirements:
Minimize the risk of unauthorized user access.
Use the principle of least privilege.
Minimize maintenance effort.
How should you configure access to the storage account for the data factory? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

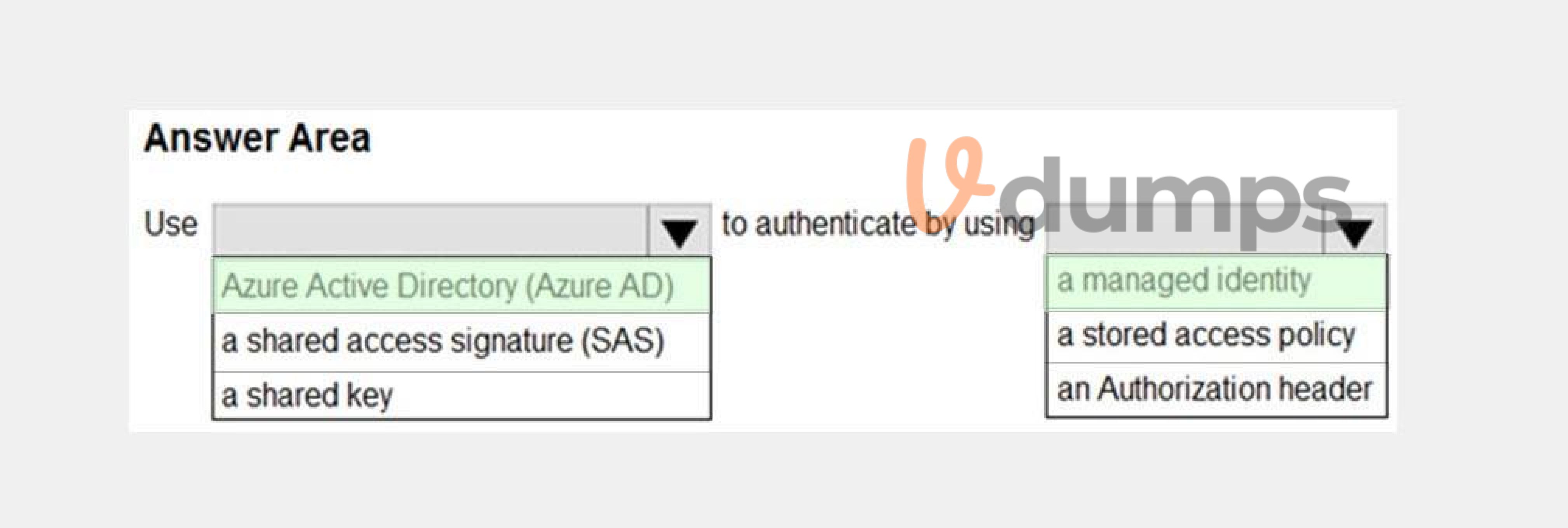
Use

- Azure Active Directory (Azure AD)
- a shared access signature (SAS)
- a shared key

to authenticate by using

- a managed identity
- a stored access policy
- an Authorization header

Answer Area:



Answer Area

Use

- [✓ Azure Active Directory (Azure AD)]
- a shared access signature (SAS)
- a shared key

to authenticate by using

- [✓ a managed identity]
- a stored access policy
- an Authorization header

Section:
Explanation:
Box 1: Azure Active Directory (Azure AD)

On Azure, managed identities eliminate the need for developers having to manage credentials by providing an identity for the Azure resource in Azure AD and using it to obtain Azure Active Directory (Azure AD) tokens.

Box 2: a managed identity

A data factory can be associated with a managed identity for Azure resources, which represents this specific data factory. You can directly use this managed identity for Data Lake Storage Gen2 authentication, similar to using your own service principal. It allows this designated factory to access and copy data to or from your Data Lake Storage Gen2.

Note: The Azure Data Lake Storage Gen2 connector supports the following authentication types. Account key authentication

Service principal authentication

Managed identities for Azure resources authentication

Reference:

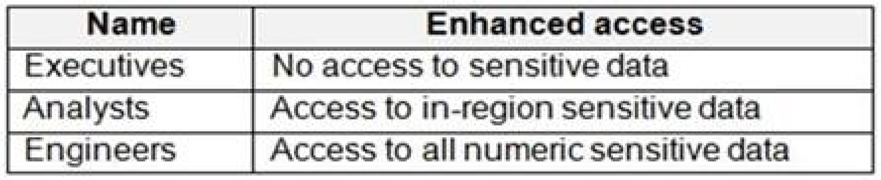
https://docs.microsoft.com/en-us/azure/active-directory/managed-identities-azure-resources/overview https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage

QUESTION 18

HOTSPOT

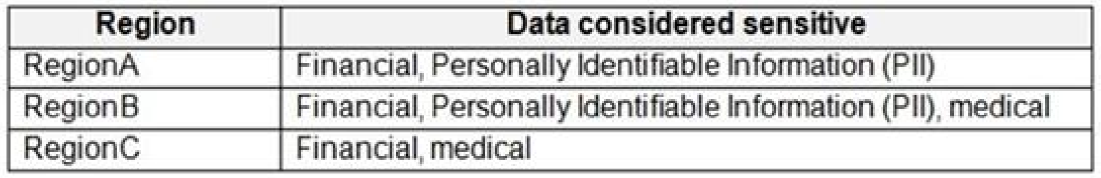
You are designing an Azure Synapse Analytics dedicated SQL pool.

Groups will have access to sensitive data in the pool as shown in the following table.



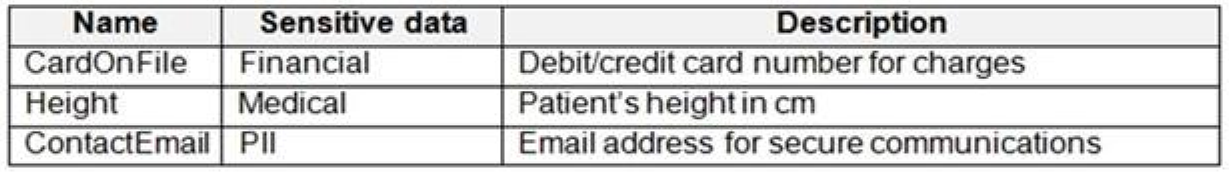
Name	Enhanced access
Executives	No access to sensitive data
Analysts	Access to in-region sensitive data
Engineers	Access to all numeric sensitive data

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



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

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



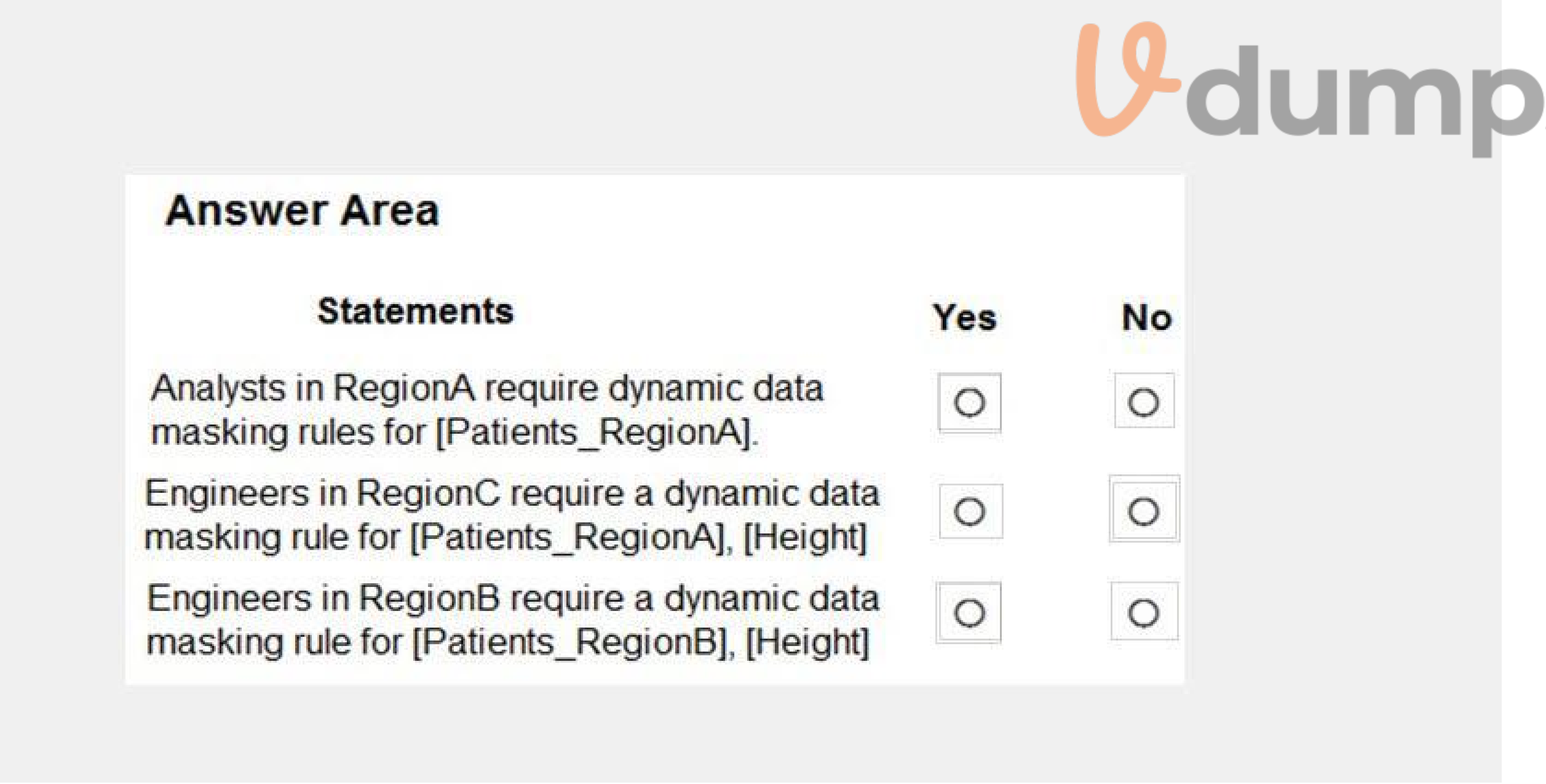
Name	Sensitive data	Description
CardOnFile	Financial	Debit/credit card number for charges
Height	Medical	Patient's height in cm
ContactEmail	PII	Email address for secure communications

You are designing dynamic data masking to maintain compliance.

For each of the following statements, select Yes if the statement is true. Otherwise, select No.

NOTE: Each correct selection is worth one point.

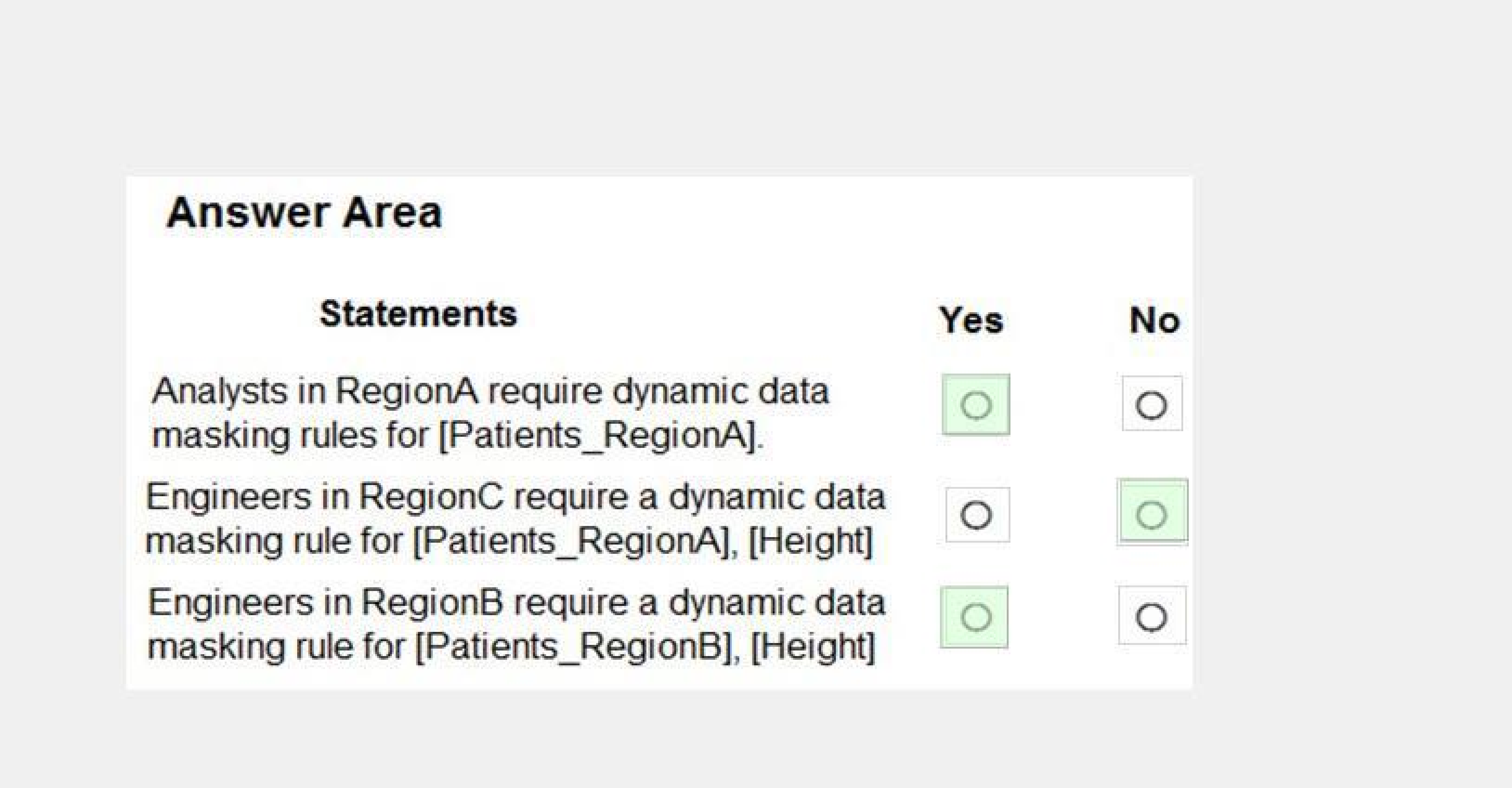
Hot Area:



Answer Area

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

Answer Area:



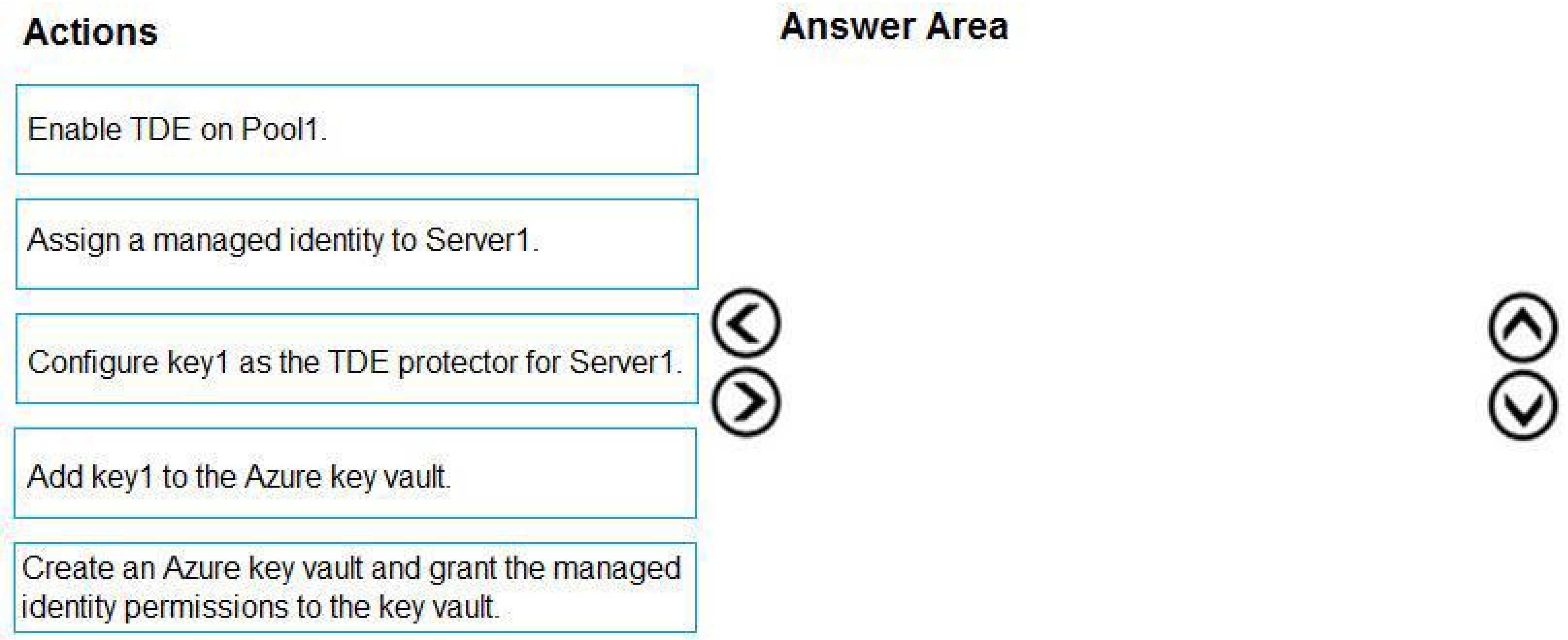
Answer Area

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

Section:
Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

QUESTION 19
DRAG DROP
You have an Azure Synapse Analytics SQL pool named Pool1 on a logical Microsoft SQL server named Server1.
You need to implement Transparent Data Encryption (TDE) on Pool1 by using a custom key named key1.
Which five actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

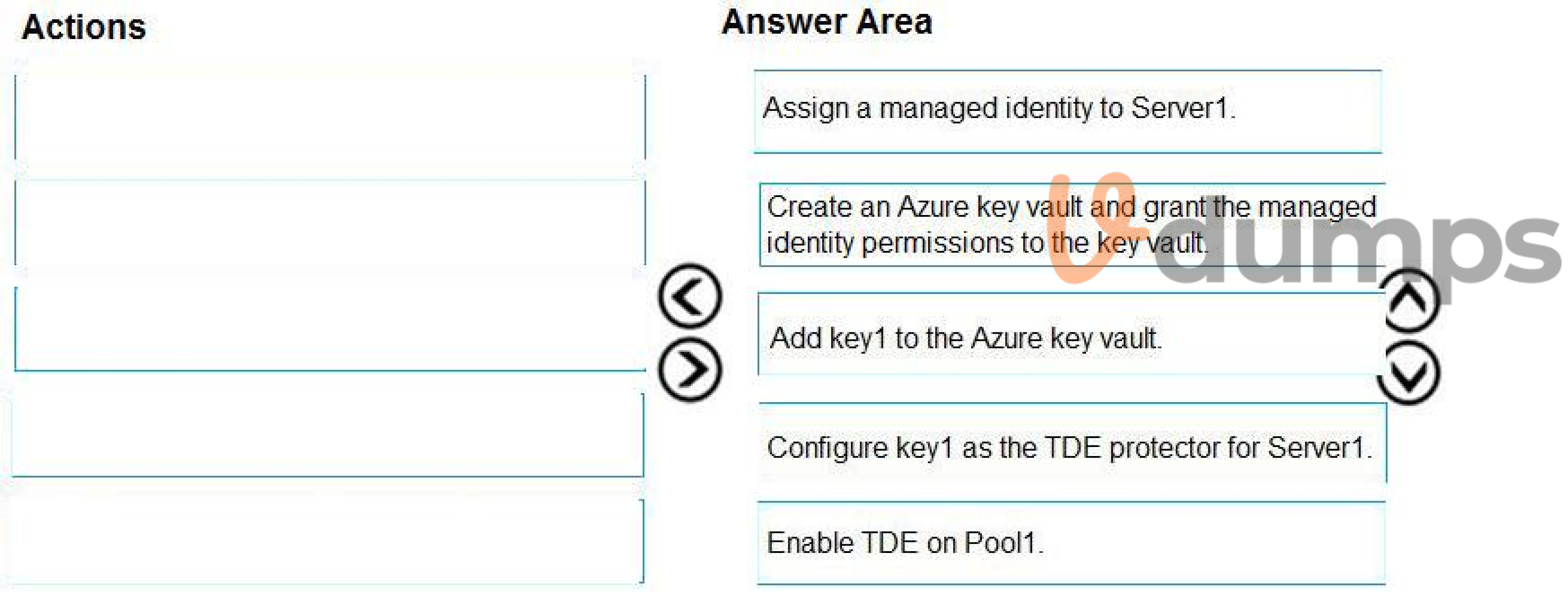


Actions

- Enable TDE on Pool1.
- Assign a managed identity to Server1.
- Configure key1 as the TDE protector for Server1.
- Add key1 to the Azure key vault.
- Create an Azure key vault and grant the managed identity permissions to the key vault.

Answer Area

Correct Answer:



Actions

Answer Area

- Assign a managed identity to Server1.
- Create an Azure key vault and grant the managed identity permissions to the key vault.
- Add key1 to the Azure key vault.
- Configure key1 as the TDE protector for Server1.
- Enable TDE on Pool1.

Section:

Explanation:

Step 1: Assign a managed identity to Server1

You will need an existing Managed Instance as a prerequisite.

Step 2: Create an Azure key vault and grant the managed identity permissions to the vault Create Resource and setup Azure Key Vault.

Step 3: Add key1 to the Azure key vault

The recommended way is to import an existing key from a .pfx file or get an existing key from the vault. Alternatively, generate a new key directly in Azure Key Vault.

Step 4: Configure key1 as the TDE protector for Server1

Provide TDE Protector key

Step 5: Enable TDE on Pool1

Reference:

https://docs.microsoft.com/en-us/azure/azure-sql/managed-instance/scripts/transparent-data-encryption-byok-powershell

QUESTION 20

HOTSPOT

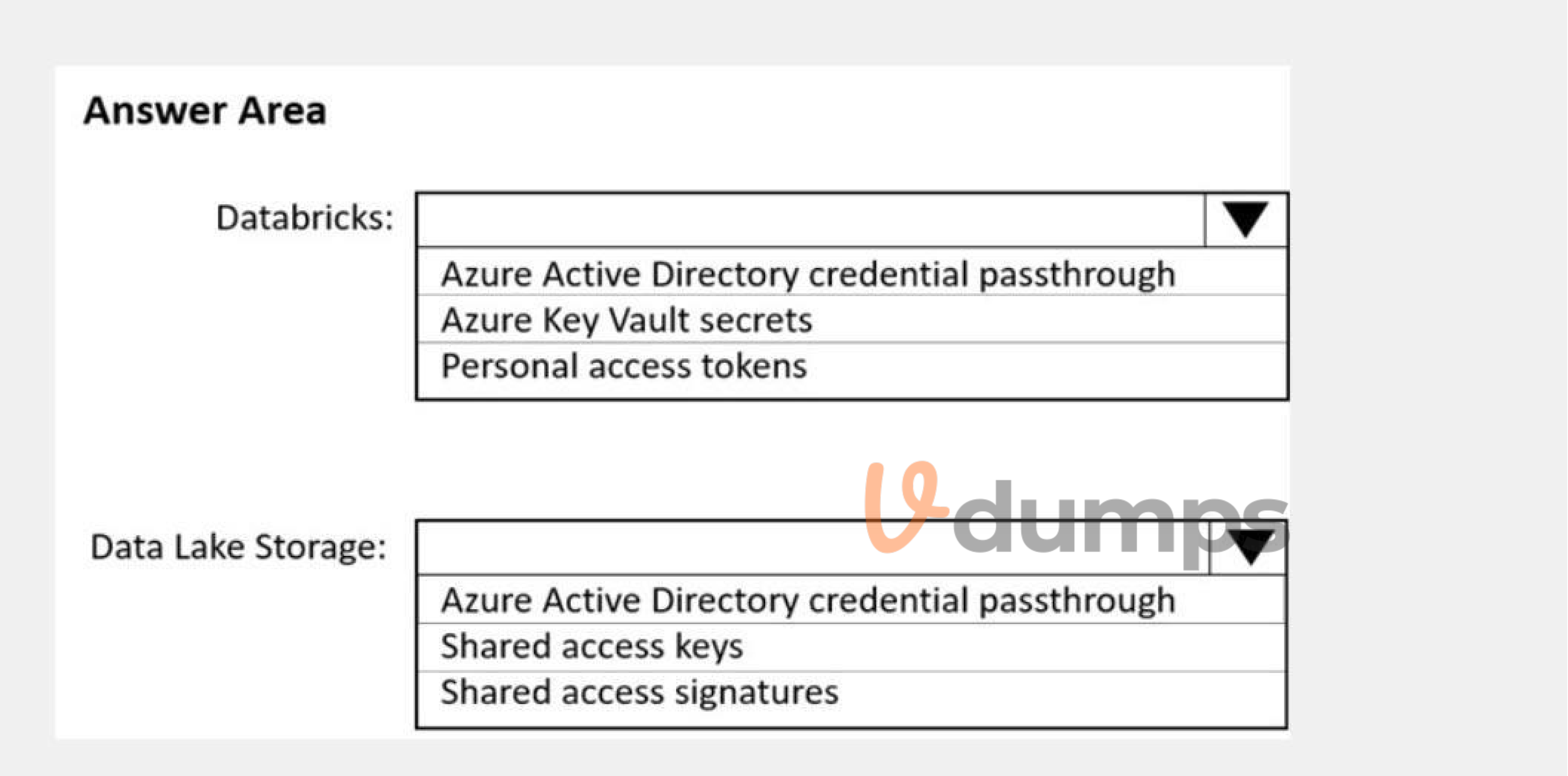
You use Azure Data Lake Storage Gen2 to store data that data scientists and data engineers will query by using Azure Databricks interactive notebooks. Users will have access only to the Data Lake Storage folders that relate to the projects on which they work.

You need to recommend which authentication methods to use for Databricks and Data Lake Storage to provide the users with the appropriate access. The solution must minimize administrative effort and development effort.

Which authentication method should you recommend for each Azure service? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

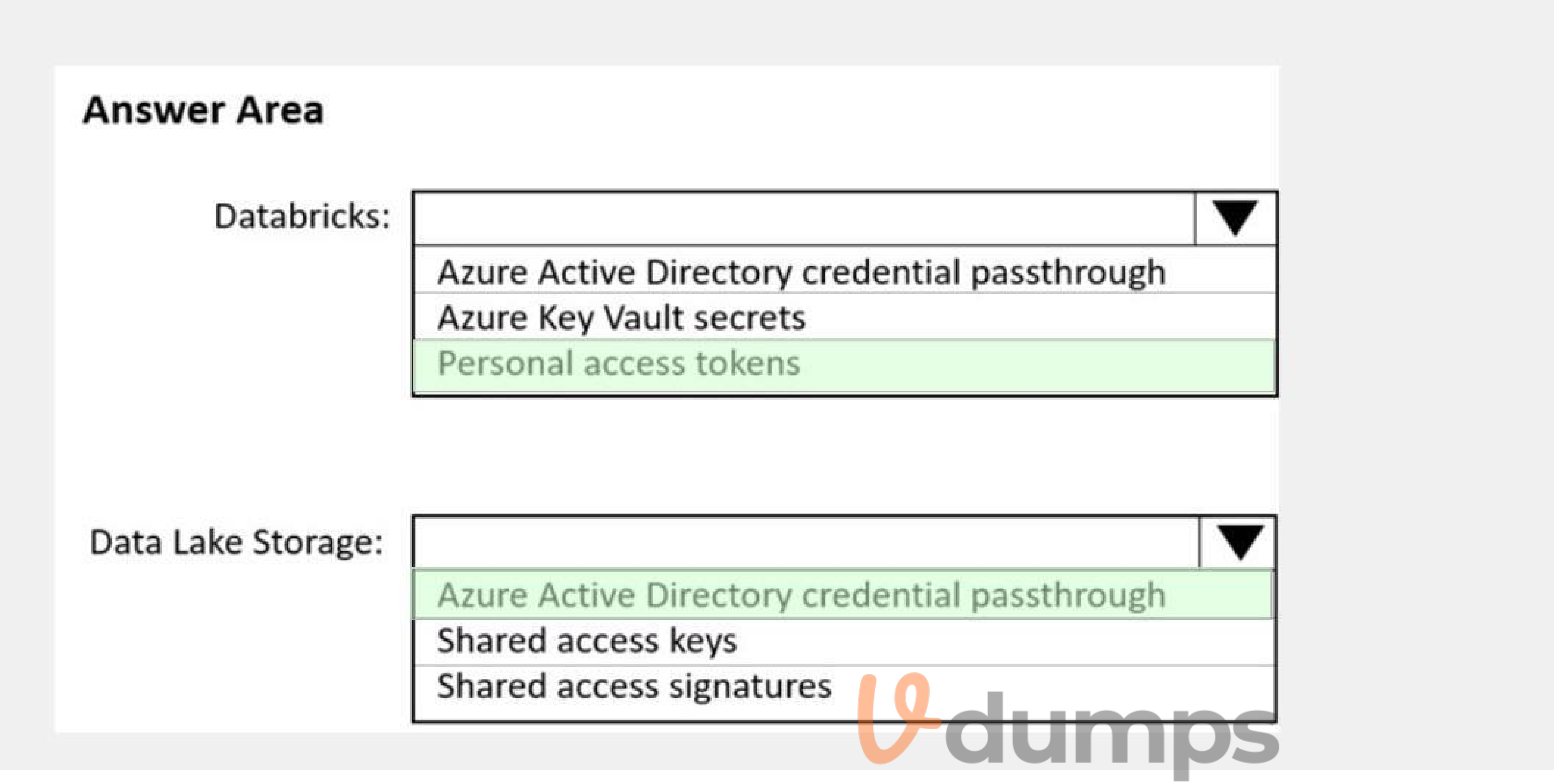


Answer Area

Databricks:
- Azure Active Directory credential passthrough
- Azure Key Vault secrets
- Personal access tokens

Data Lake Storage:
- Azure Active Directory credential passthrough
- Shared access keys
- Shared access signatures

Answer Area:



Answer Area

Databricks:

- Azure Active Directory credential passthrough
- Azure Key Vault secrets
- Personal access tokens

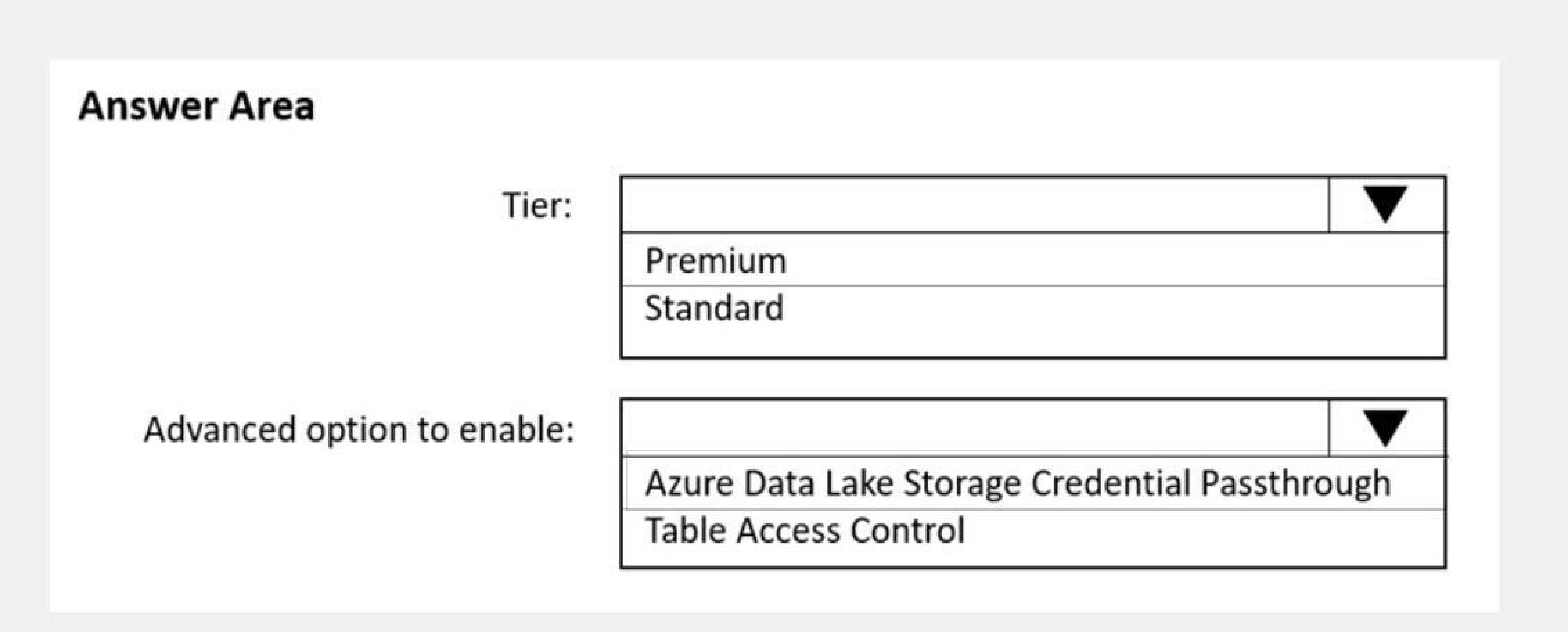
Data Lake Storage:

- Azure Active Directory credential passthrough
- Shared access keys
- Shared access signatures

Section:
Explanation:
Box 1: Personal access tokens
You can use storage shared access signatures (SAS) to access an Azure Data Lake Storage Gen2 storage account directly. With SAS, you can restrict access to a storage account using temporary tokens with fine-grained access control.
You can add multiple storage accounts and configure respective SAS token providers in the same Spark session.
Box 2: Azure Active Directory credential passthrough
You can authenticate automatically to Azure Data Lake Storage Gen1 (ADLS Gen1) and Azure Data Lake Storage Gen2 (ADLS 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.
After configuring Azure Data Lake Storage credential passthrough and creating storage containers, you can access data directly in Azure Data Lake Storage Gen1 using an adl:// path and Azure Data Lake Storage Gen2 using an abfss:// path:
Reference:
https://docs.microsoft.com/en-us/azure/databricks/data/data-sources/azure/adls-gen2/azure-datalake-gen2-sas-access
https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough

QUESTION 21
HOTSPOT
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.

Hot Area:



Answer Area

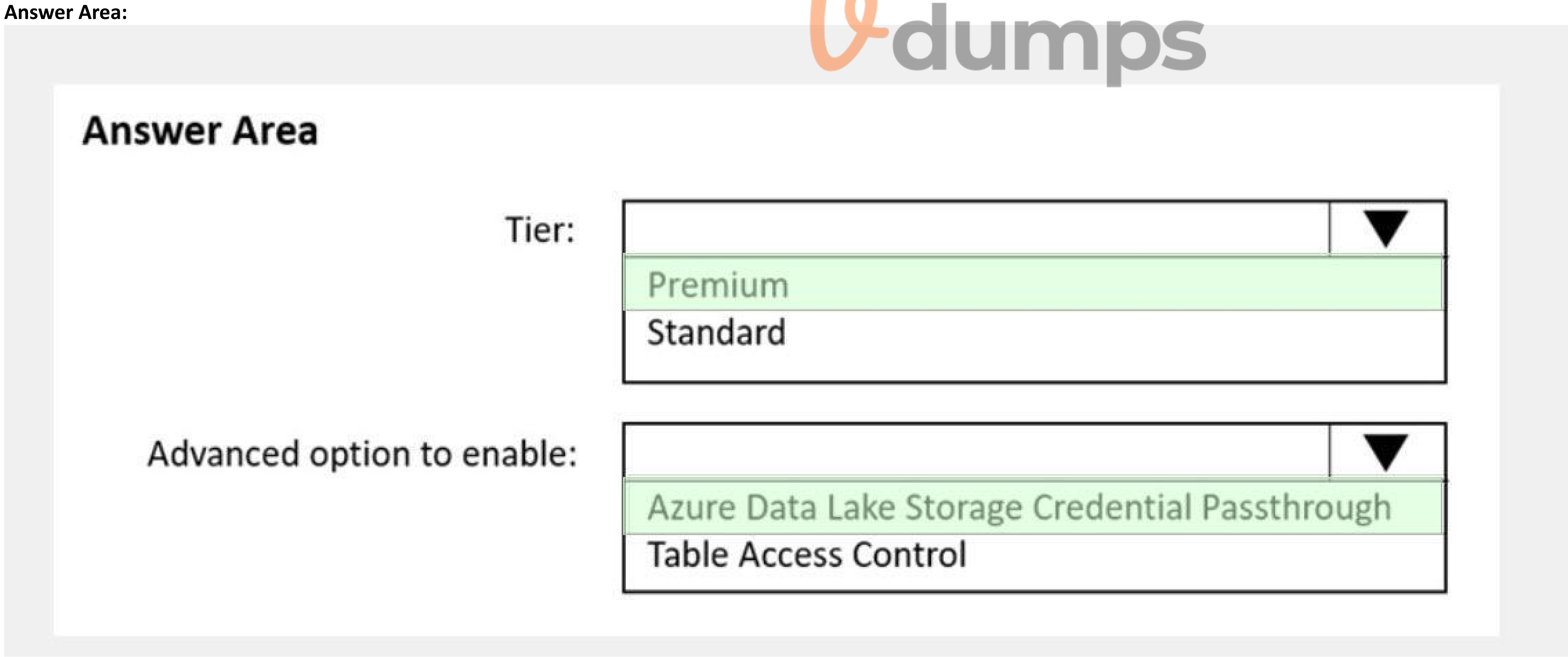
Tier:

- Premium
- Standard

Advanced option to enable:

- Azure Data Lake Storage Credential Passthrough
- Table Access Control

Answer Area:



Answer Area

Tier:

- [✓ Premium]
- Standard

Advanced option to enable:

- [✓ Azure Data Lake Storage Credential Passthrough]
- Table Access Control

Section:

Explanation:

Box 1: Premium
Credential passthrough requires an Azure Databricks Premium Plan
Box 2: Azure Data Lake Storage credential passthrough
You can access Azure Data Lake Storage using Azure Active Directory credential passthrough. 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.
Reference:
https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough

QUESTION 22
You have a SQL pool in Azure Synapse that contains a table named dbo.Customers. The table contains a column name Email. You need to prevent nonadministrative users from seeing the full email addresses in the Email column. The users must see values in a format of a XXX@XXXX.com instead. What should you do?

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

Correct Answer: D
Section:
Explanation:
The Email masking method, which exposes the first letter and replaces the domain with XXX.com using a constant string prefix in the form of an email address. aXX@XXXX.com
Reference: https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

QUESTION 23
You have an Azure Data Lake Storage Gen2 account named adls2 that is protected by a virtual network. You are designing a SQL pool in Azure Synapse that will use adls2 as a source. What should you use to authenticate to adls2?

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

Correct Answer: D
Section:
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

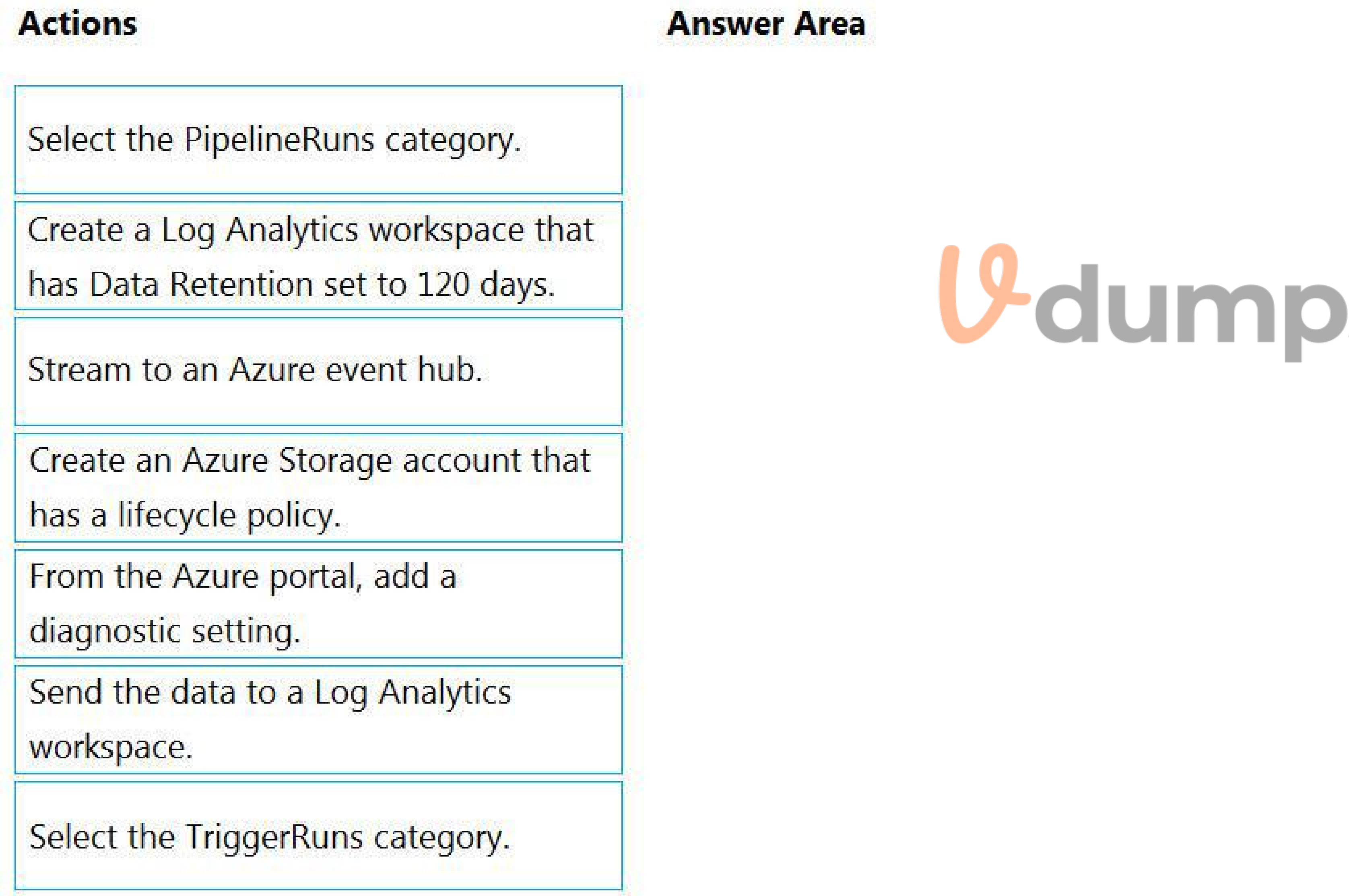
QUESTION 24
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Contacts. Contacts contains a column named Phone. You need to ensure that users in a specific role only see the last four digits of a phone number when querying the Phone column. What should you include in the solution?

A. table partitions
B. a default value
C. row-level security (RLS)
D. column encryption
E. dynamic data masking

Correct Answer: E
Section:
Explanation:
Dynamic data masking helps prevent unauthorized access to sensitive data by enabling customers to designate how much of the sensitive data to reveal with minimal impact on the application layer. It's a policy-based security feature that hides the sensitive data in the result set of a query over designated database fields, while the data in the database is not changed.
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

QUESTION 25
DRAG DROP
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.

Select and Place:

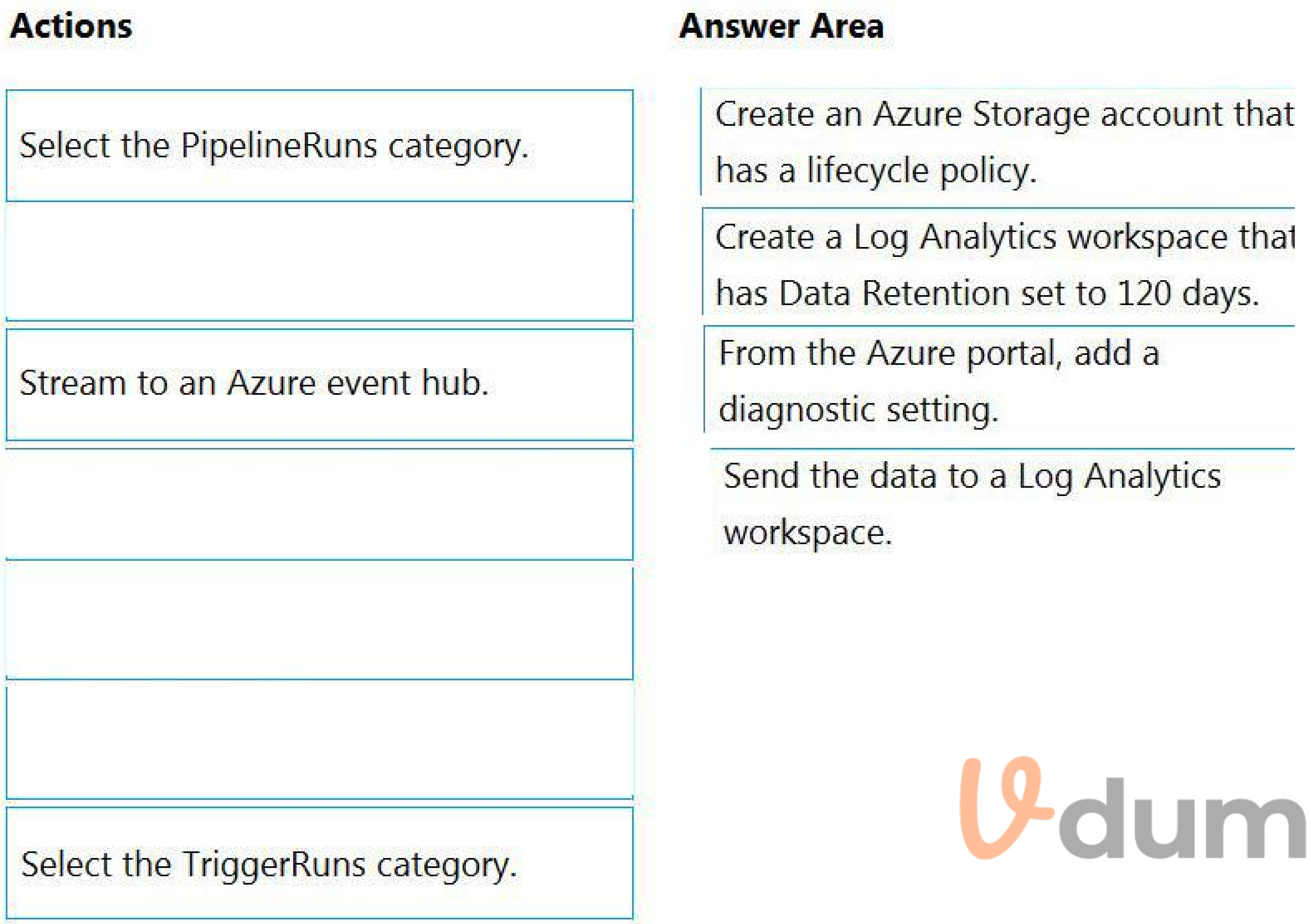


Actions

- Select the PipelineRuns category.
- Create a Log Analytics workspace that has Data Retention set to 120 days.
- Stream to an Azure event hub.
- Create an Azure Storage account that has a lifecycle policy.
- From the Azure portal, add a diagnostic setting.
- Send the data to a Log Analytics workspace.
- Select the TriggerRuns category.

Answer Area

Correct Answer:



Actions

Select the PipelineRuns category.

Stream to an Azure event hub.

Select the TriggerRuns category.

Answer Area

Create an Azure Storage account that has a lifecycle policy.

Create a Log Analytics workspace that has Data Retention set to 120 days.

From the Azure portal, add a diagnostic setting.

Send the data to a Log Analytics workspace.

Section:

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.

1. In the portal, go to Monitor. Select Settings > Diagnostic settings.
2. Select the data factory for which you want to set a diagnostic setting.
3. If no settings exist on the selected data factory, you're prompted to create a setting. Select Turn on diagnostics.
4. Give your setting a name, select Send to Log Analytics, and then select a workspace from Log Analytics Workspace.
5. Select Save.

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor

QUESTION 26

HOTSPOT

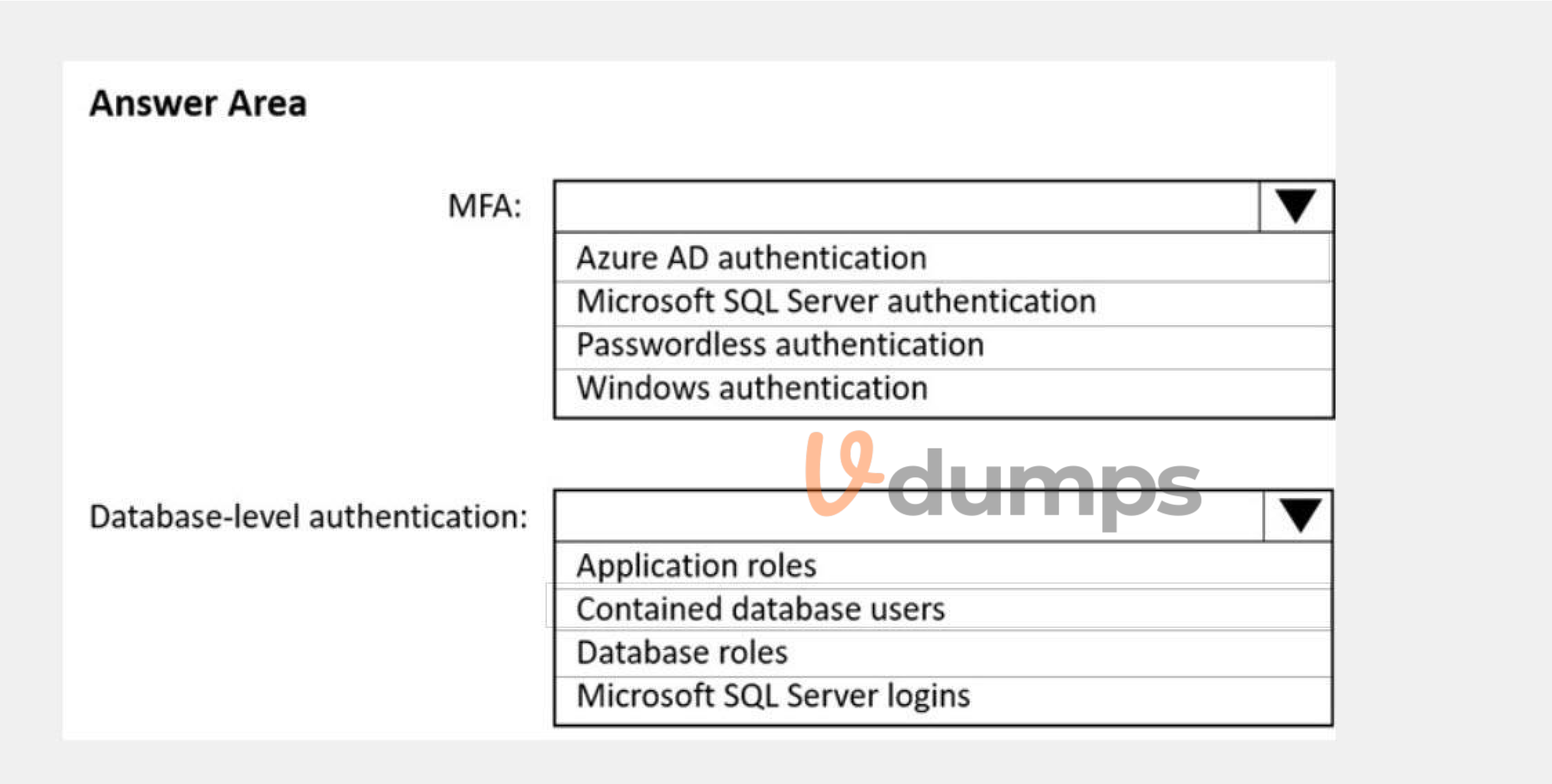
You have an Azure subscription that is linked to a hybrid Azure Active Directory (Azure AD) tenant. The subscription contains an Azure Synapse Analytics SQL pool named Pool1.

You need to recommend an authentication solution for Pool1. The solution must support multi-factor authentication (MFA) and database-level authentication.

Which authentication solution or solutions should you include m the recommendation? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

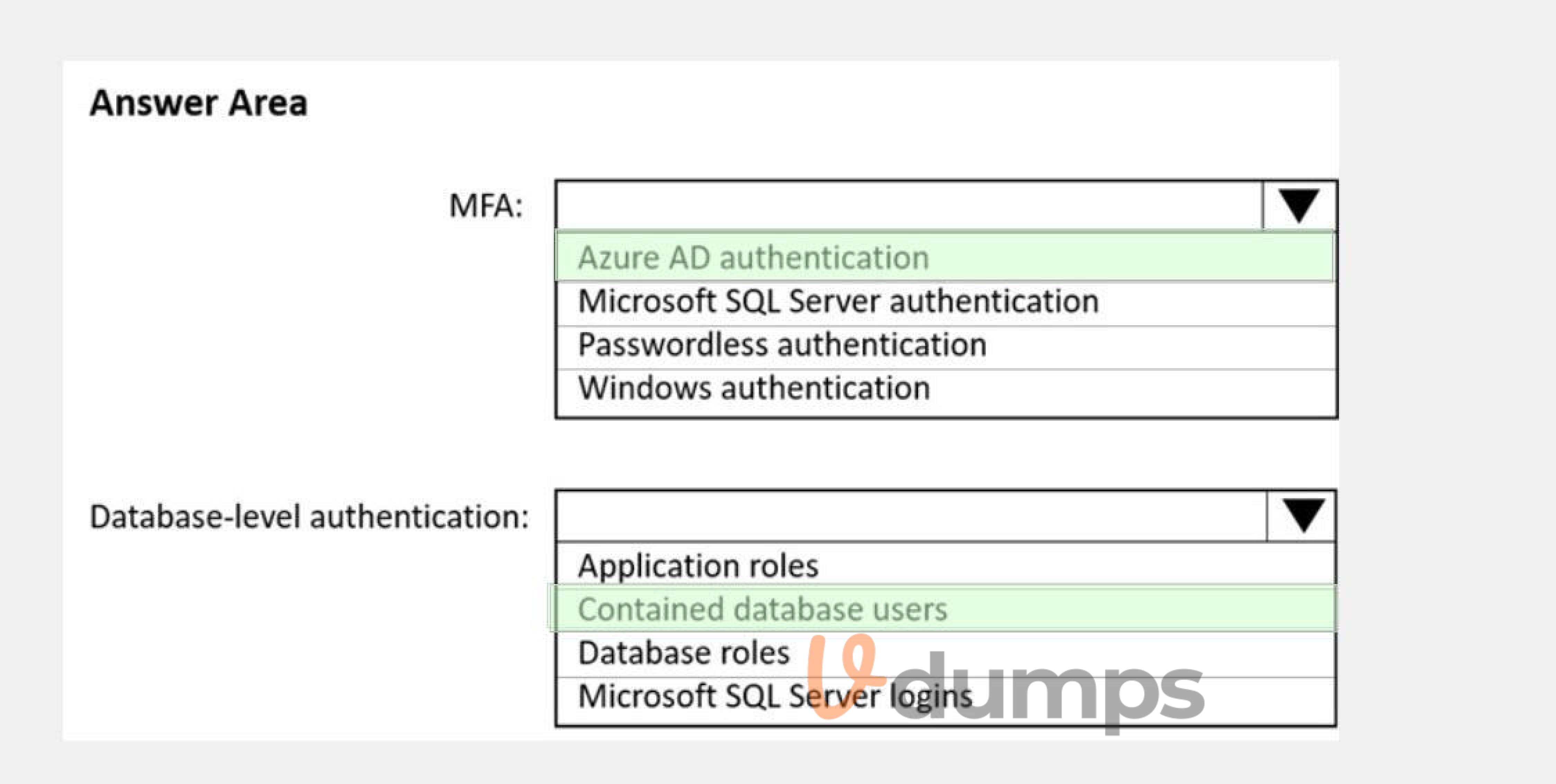
MFA: ▼

- Azure AD authentication
- Microsoft SQL Server authentication
- Passwordless authentication
- Windows authentication

Database-level authentication: ▼

- Application roles
- Contained database users
- Database roles
- Microsoft SQL Server logins

Answer Area:



Answer Area

MFA:

- Azure AD authentication
- Microsoft SQL Server authentication
- Passwordless authentication
- Windows authentication

Database-level authentication:

- Application roles
- Contained database users
- Database roles
- Microsoft SQL Server logins

Section:
Explanation:
Box 1: Azure AD authentication
Azure AD authentication has the option to include MFA.
Box 2: Contained database users
Azure AD authentication uses contained database users to authenticate identities at the database level.
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/authentication-mfa-ssms-overview
https://docs.microsoft.com/en-us/azure/azure-sql/database/authentication-aad-overview

Exam F

QUESTION 1
Note: This question-is part of a series of questions that present the same scenario. Each question-in the series contains a unique solution that might meet the stated goals. Some question-sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question-in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You have an Azure Data Lake Storage account that contains a staging zone. You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics. Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data flow, and then inserts the data into the data warehouse. Does this meet the goal?

A. Yes
B. No

Correct Answer: A
Section:
Explanation:

QUESTION 2

Note: This question-is part of a series of questions that present the same scenario. Each question-in the series contains a unique solution that might meet the stated goals. Some question-sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question-in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You have an Azure Data Lake Storage account that contains a staging zone. You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics. Solution: You schedule an Azure Databricks job that executes an R notebook, and then inserts the data into the data warehouse. Does this meet the goal?

A. Yes
B. No

Correct Answer: A
Section:
Explanation:

QUESTION 3

You plan to create an Azure Data Factory pipeline that will include a mapping data flow. You have JSON data containing objects that have nested arrays. You need to transform the JSON-formatted data into a tabular dataset. The dataset must have one row for each item in the arrays. Which transformation method should you use in the mapping data flow?

A. new branch
B. unpivot
C. alter row
D. flatten

Correct Answer: D
Section:
Explanation:
Use the flatten transformation to take array values inside hierarchical structures such as JSON and unroll them into individual rows. This process is known as denormalization.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-flatten

QUESTION 4

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

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

Correct Answer: D
Section:
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,

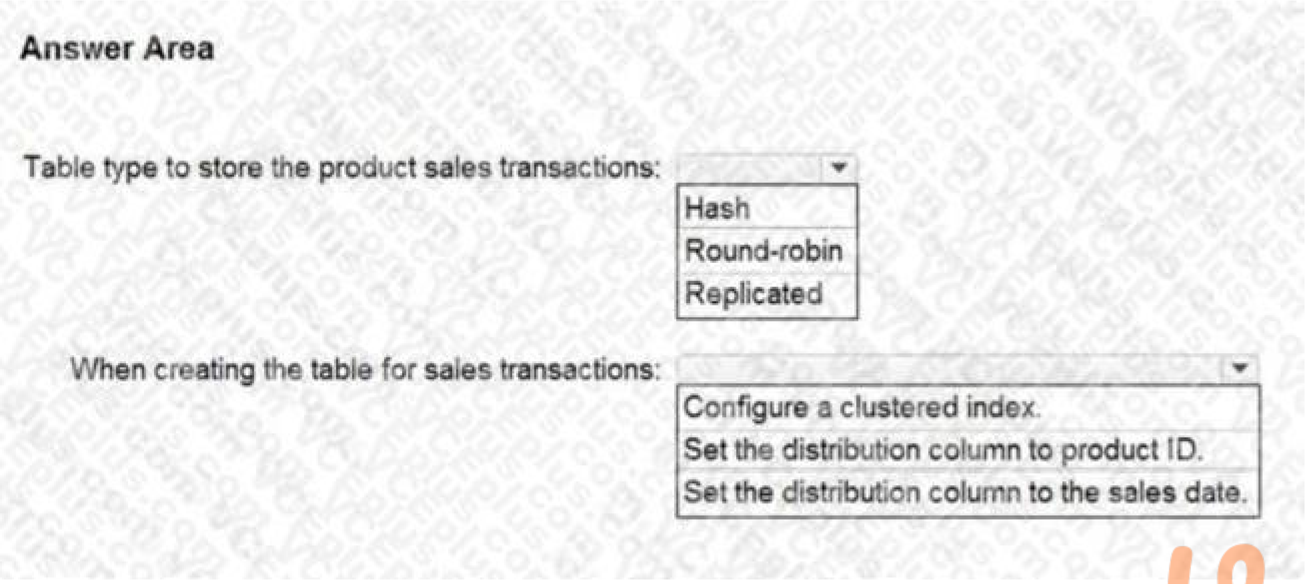
do not overlap, and an event cannot belong to more than one tumbling window.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

QUESTION 5

HOTSPOT

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.

Hot Area:

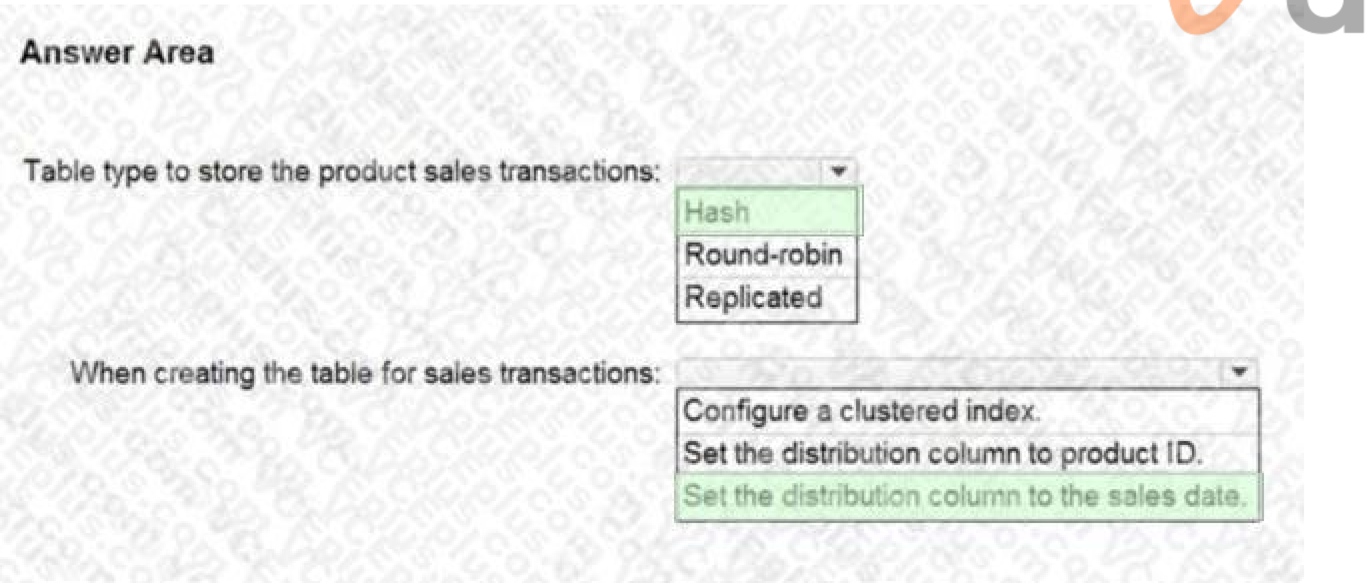


Answer Area

Table type to store the product sales transactions:
- Hash
- Round-robin
- Replicated

When creating the table for sales transactions:
- Configure a clustered index.
- Set the distribution column to product ID.
- Set the distribution column to the sales date.

Answer Area:



Answer Area

Table type to store the product sales transactions:
- [✓ Hash]
- Round-robin
- Replicated

When creating the table for sales transactions:
- Configure a clustered index.
- Set the distribution column to product ID.
- [✓ Set the distribution column to the sales date.]

Section:

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.

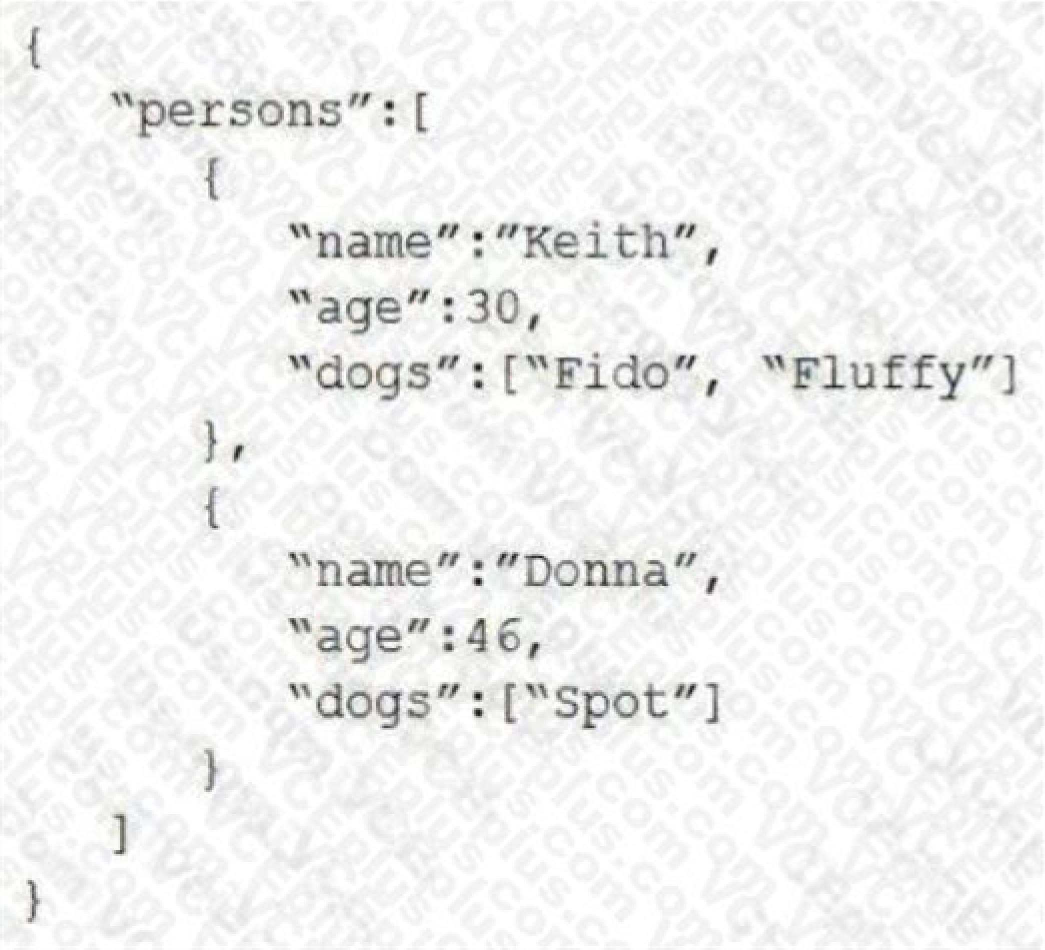
Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

Reference: https://rajanieshkaushikk.com/2020/09/09/how-to-choose-right-data-distribution-strategy-for-azure-synapse/

QUESTION 6

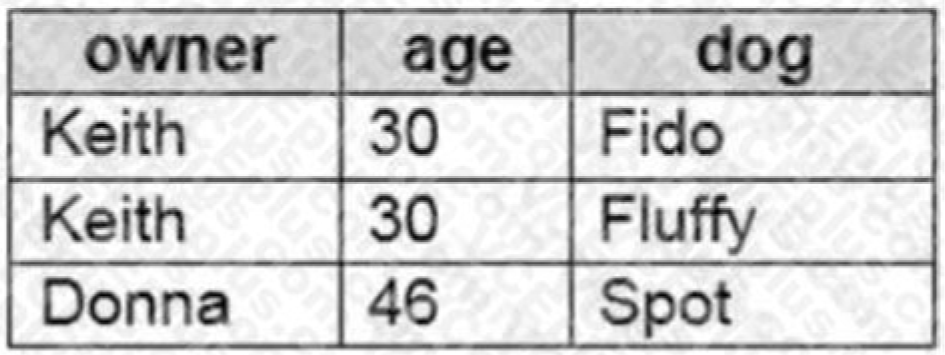
DRAG DROP

You use PySpark in Azure Databricks to parse the following JSON input.



```
{
    "persons":[
        {
            "name":"Keith",
            "age":30,
            "dogs":["Fido", "Fluffy"]
        },
        {
            "name":"Donna",
            "age":46,
            "dogs":["Spot"]
        }
    ]
}
```

You need to output the data in the following tabular format.

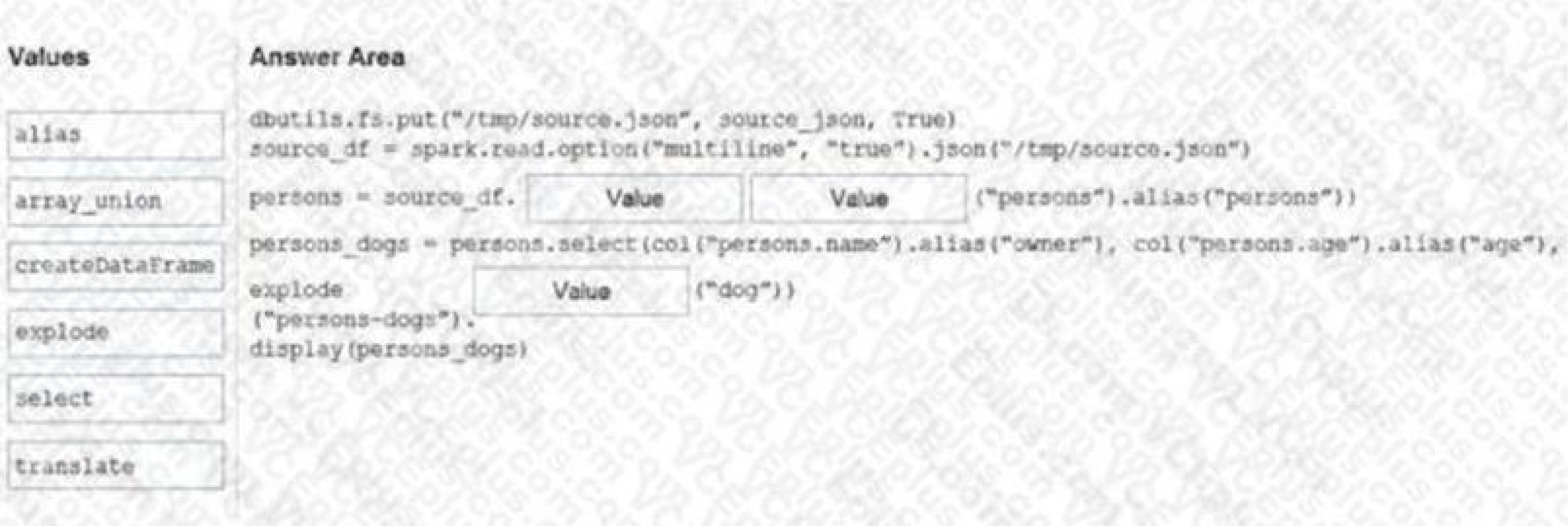


owner	age	dog
Keith	30	Fido
Keith	30	Fluffy
Donna	46	Spot

How should you complete the PySpark code? 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 spit bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:



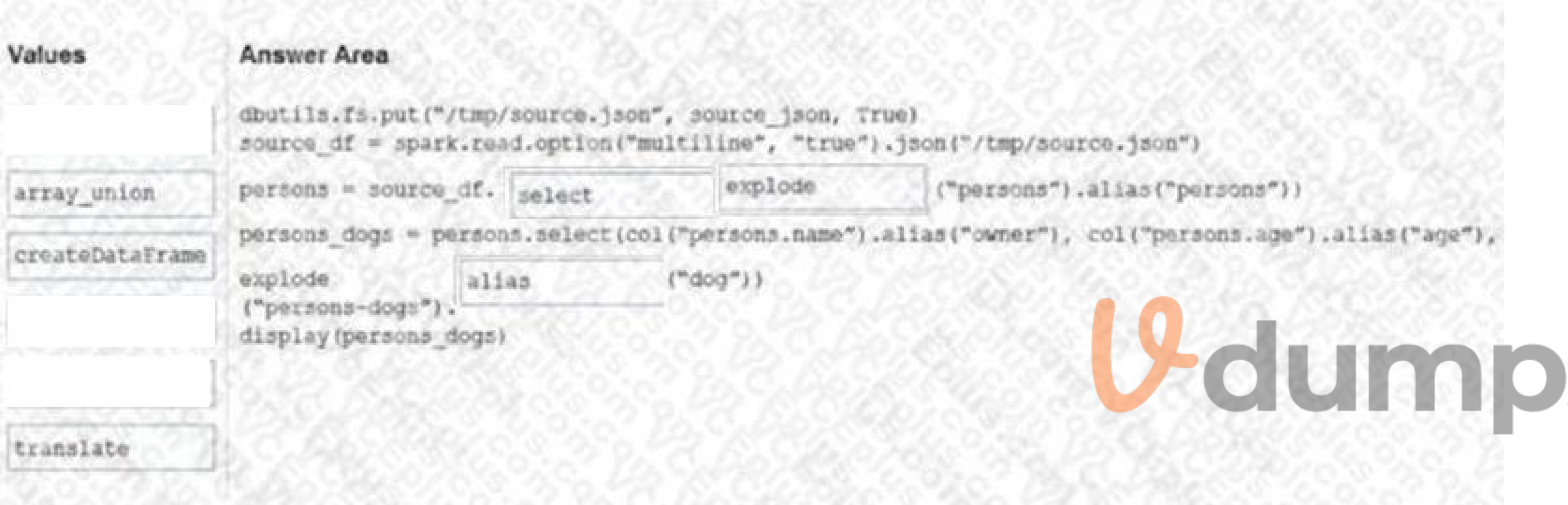
Values

alias
array_union
createDataFrame
explode
select
translate

Answer Area

```
dbutils.fs.put("/tmp/source.json", source_json, True)
source_df = spark.read.option("multiline", "true").json("/tmp/source.json")
persons = source_df. [Value] [Value] ("persons").alias("persons"))
persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"),
explode [Value] ("dog"))
("persons-dogs").
display(persons_dogs)
```

Correct Answer:



Values

array_union
createDataFrame
translate

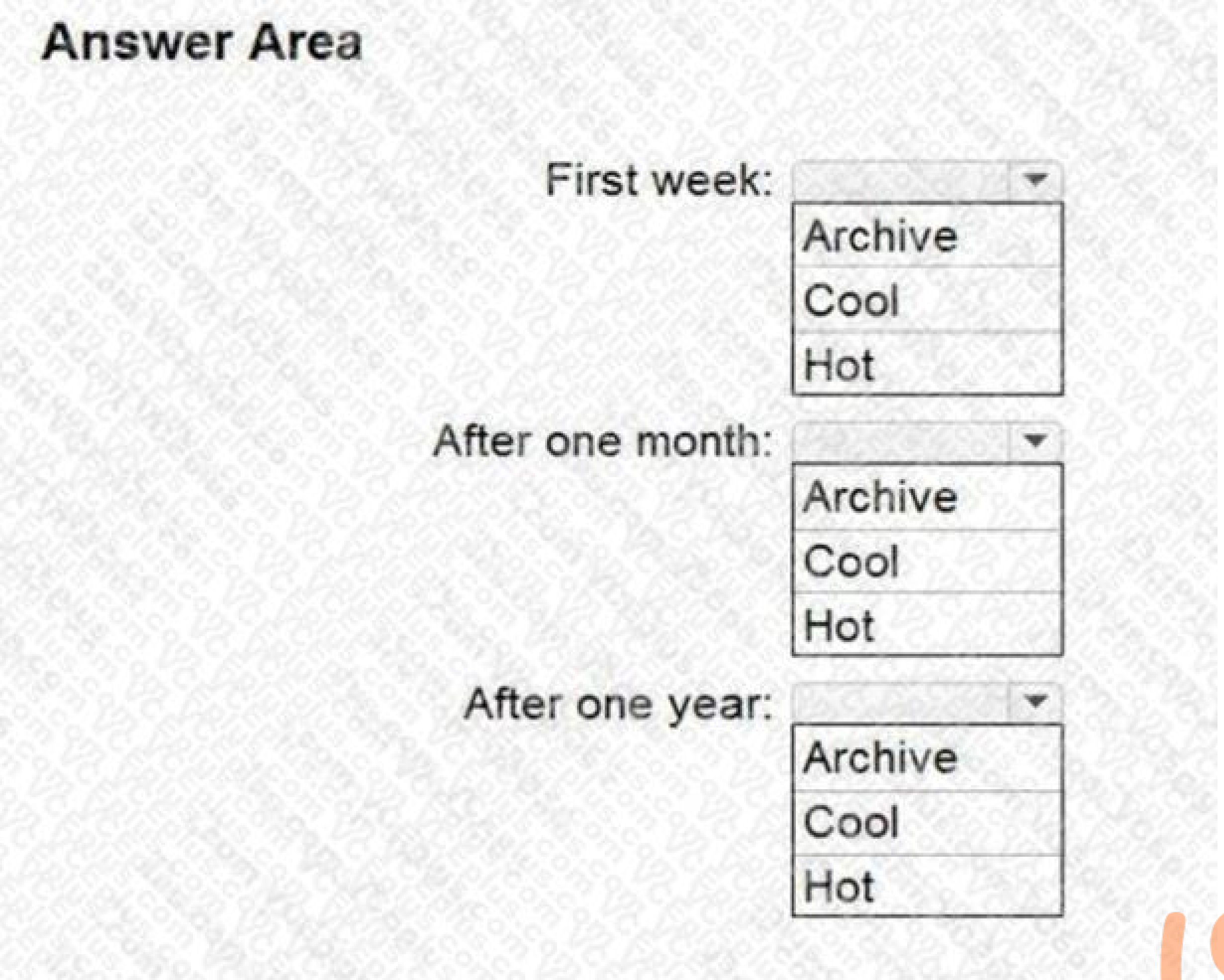
Answer Area

```
dbutils.fs.put("/tmp/source.json", source_json, True)
source_df = spark.read.option("multiline", "true").json("/tmp/source.json")
persons = source_df. select explode ("persons").alias("persons"))
persons_dogs = persons.select(col("persons.name").alias("owner"), col("persons.age").alias("age"),
explode alias ("dog"))
("persons-dogs").
display(persons_dogs)
```

Section:
Explanation:
Box 1: select
Box 2: explode
Bop 3: alias pyspark.sql.Column.alias returns this column aliased with a new name or names (in the case of expressions that return more than one column, such as explode).
Reference:
https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.sql.Column.alias.html
https://docs.microsoft.com/en-us/azure/databricks/sql/language-manual/functions/explode

QUESTION 7
HOTSPOT
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.

Hot Area:



Answer Area

First week:

- Archive
- Cool
- Hot

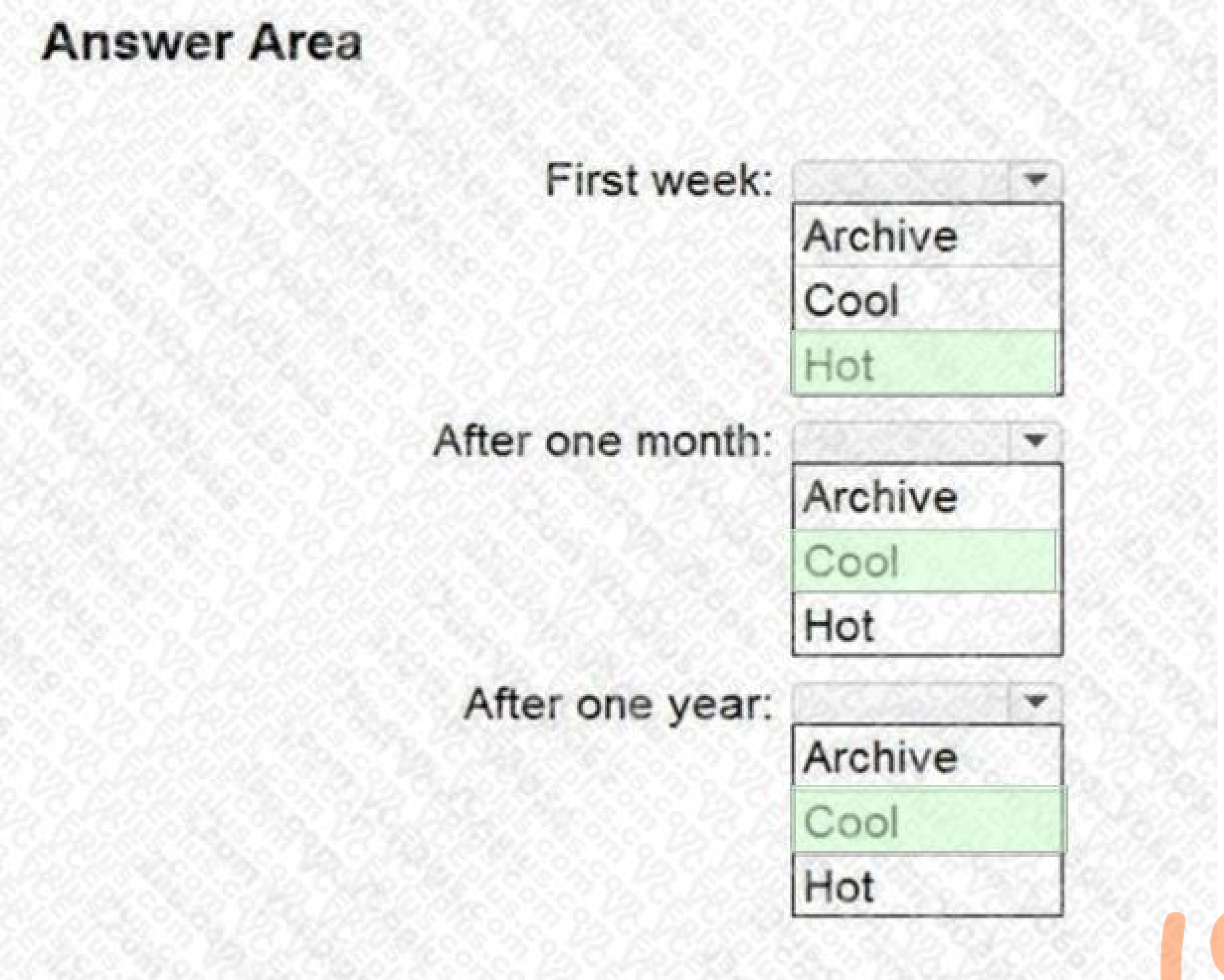
After one month:

- Archive
- Cool
- Hot

After one year:

- Archive
- Cool
- Hot

Answer Area:



Answer Area

First week:
Archive
Cool
Hot

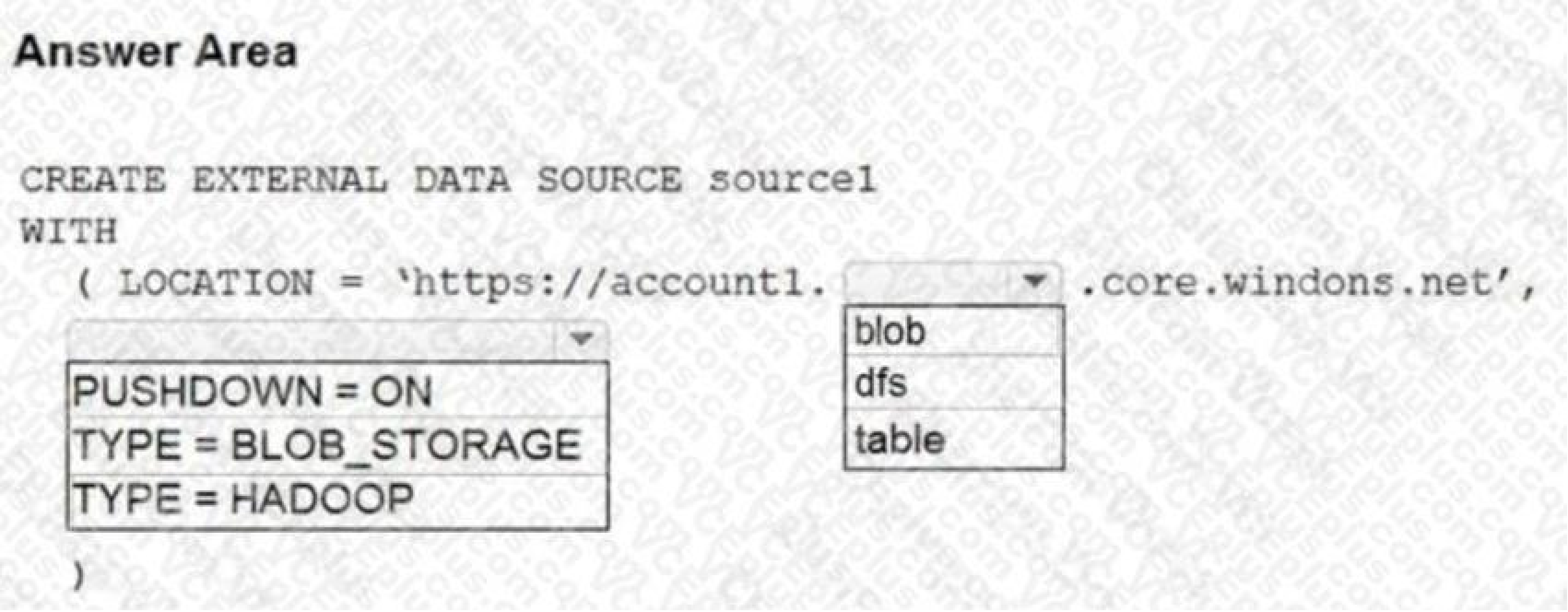
After one month:
Archive
Cool
Hot

After one year:
Archive
Cool
Hot

Section:
Explanation:
Box 1: Hot
Hot tier - An online tier optimized for storing data that is accessed or modified frequently. The Hot tier has the highest storage costs, but the lowest access costs. Box 2: Cool
Cool tier - An online tier optimized for storing data that is infrequently accessed or modified. Data in the Cool tier should be stored for a minimum of 30 days. The Cool tier has lower storage costs and higher access costs compared to the Hot tier.
Box 3: Cool
Not Archive tier - An offline tier optimized for storing data that is rarely accessed, and that has flexible latency requirements, on the order of hours. Data in the Archive tier should be stored for a minimum of 180 days.
Reference: https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview https://www.altaro.com/hyperv/azure-archive-storage/

QUESTION 8
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.

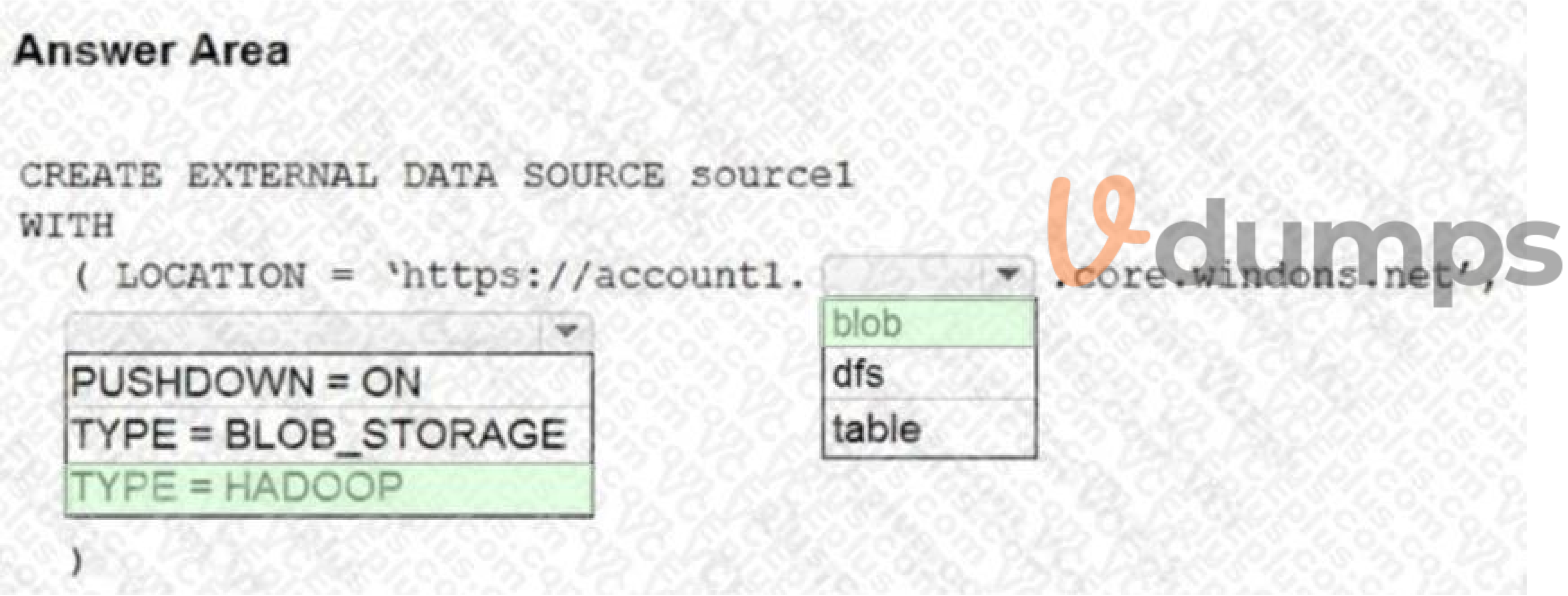
Hot Area:



Answer Area

```
CREATE EXTERNAL DATA SOURCE source1
WITH
   ( LOCATION = 'https://account1.          .core.windons.net',

   )
```

Answer Area:



Answer Area

```
CREATE EXTERNAL DATA SOURCE source1
WITH
   ( LOCATION = 'https://account1.          .core.windons.net',

   )
```

Section:

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

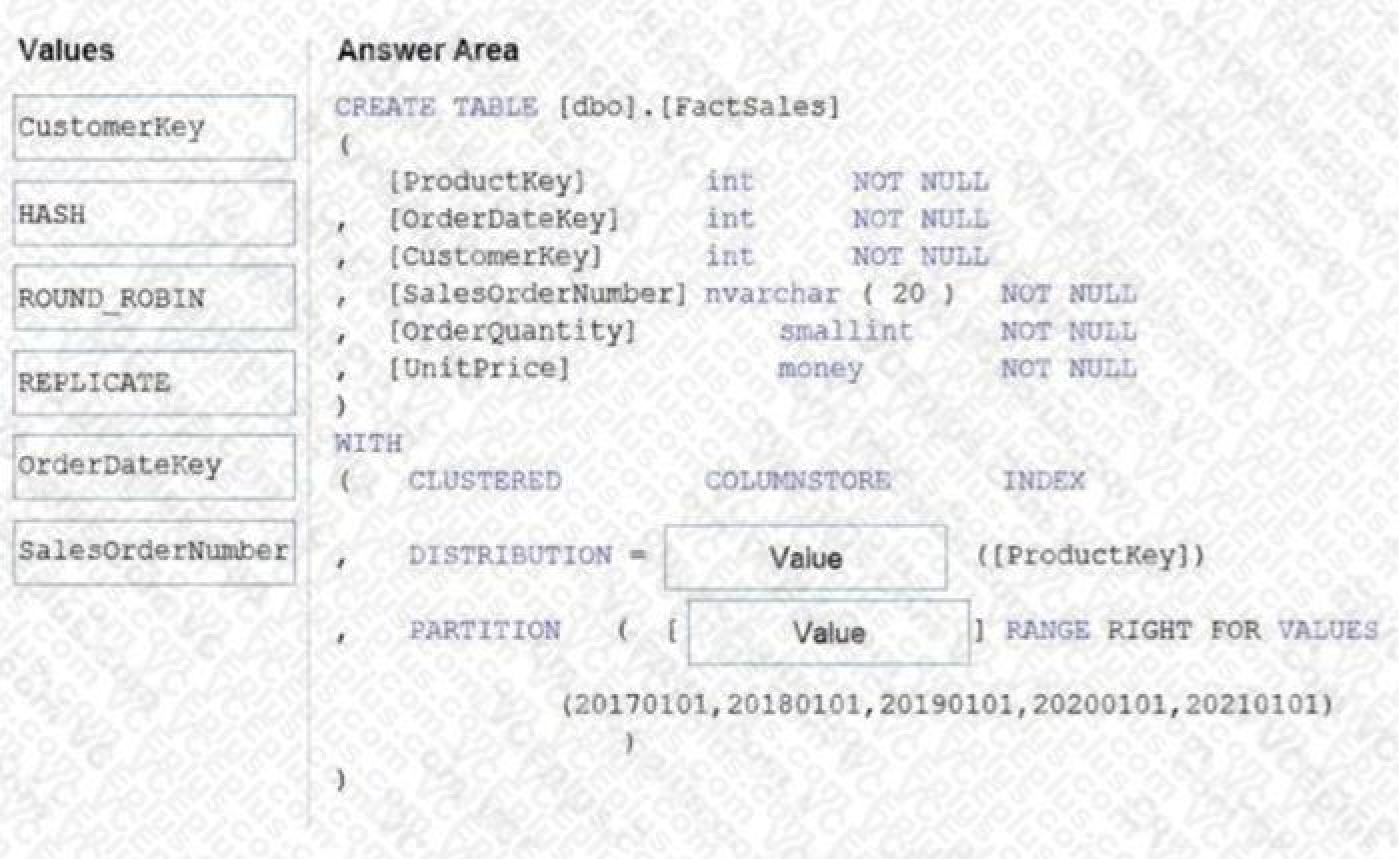
QUESTION 9

DRAG DROP

You plan to create a table in an Azure Synapse Analytics dedicated SQL pool. Data in the table will be retained for five years. Once a year, data that is older than five years will be deleted. You need to ensure that the data is distributed evenly across partitions. The solution must minimize the amount of time required to delete old data. 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.

Select and Place:



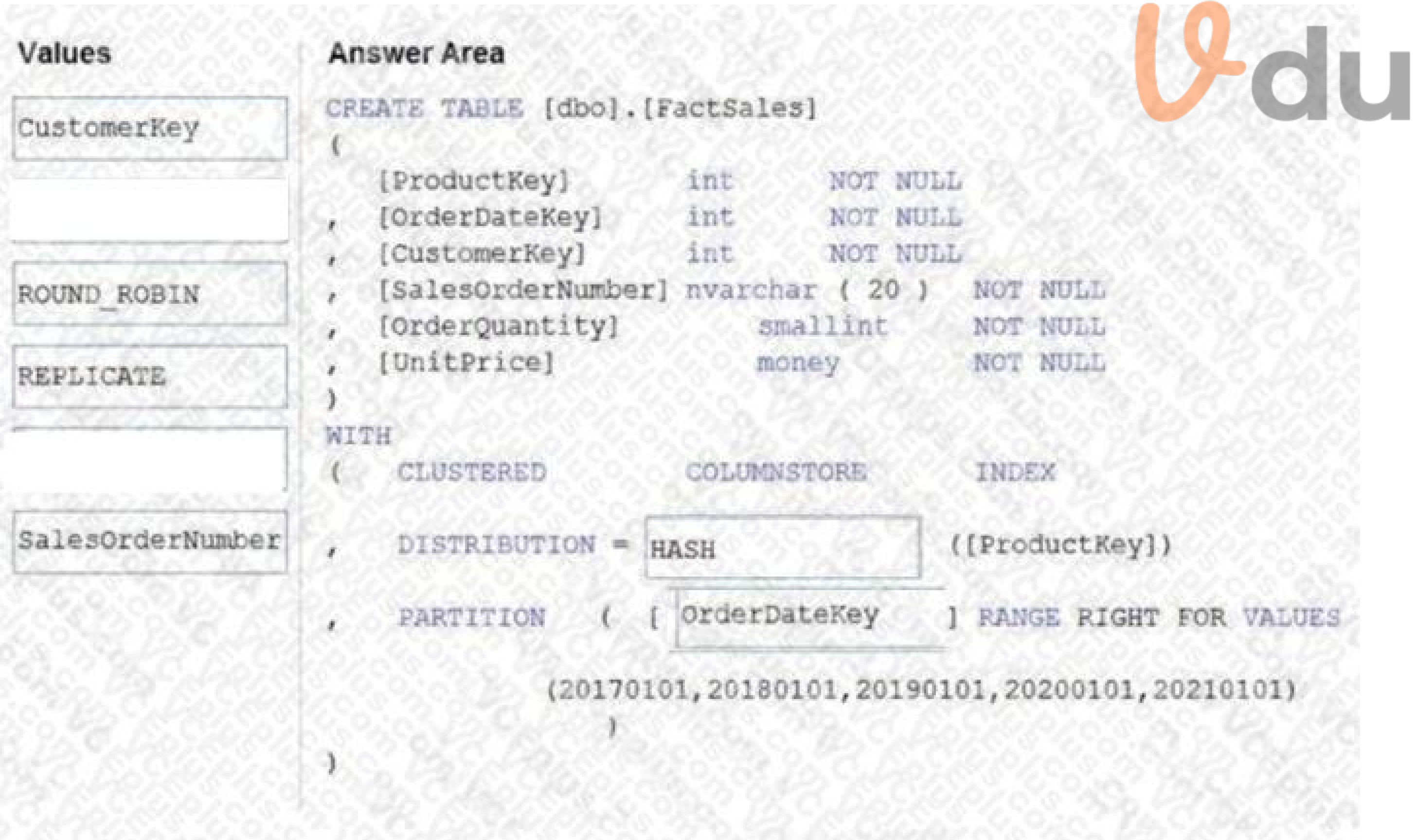
Values

- CustomerKey
- HASH
- ROUND_ROBIN
- REPLICATE
- OrderDateKey
- SalesOrderNumber

Answer Area

```
CREATE TABLE [dbo].[FactSales]
(
    [ProductKey]        int       NOT NULL
,   [OrderDateKey]      int       NOT NULL
,   [CustomerKey]       int       NOT NULL
,   [SalesOrderNumber] nvarchar ( 20 )  NOT NULL
,   [OrderQuantity]         smallint    NOT NULL
,   [UnitPrice]             money       NOT NULL
)
WITH
(   CLUSTERED         COLUMNSTORE       INDEX

,   DISTRIBUTION =  Value    ([ProductKey])

,   PARTITION    (  [  Value  ] RANGE RIGHT FOR VALUES

                (20170101,20180101,20190101,20200101,20210101)
                     )
)
```

Correct Answer:



Values

- CustomerKey
- ROUND_ROBIN
- REPLICATE
- SalesOrderNumber

Answer Area

```
CREATE TABLE [dbo].[FactSales]
(
    [ProductKey]        int       NOT NULL
,   [OrderDateKey]      int       NOT NULL
,   [CustomerKey]       int       NOT NULL
,   [SalesOrderNumber] nvarchar ( 20 )  NOT NULL
,   [OrderQuantity]         smallint    NOT NULL
,   [UnitPrice]             money       NOT NULL
)
WITH
(   CLUSTERED         COLUMNSTORE       INDEX

,   DISTRIBUTION =  HASH    ([ProductKey])

,   PARTITION    (  [  OrderDateKey  ] RANGE RIGHT FOR VALUES

                (20170101,20180101,20190101,20200101,20210101)
                     )
)
```

Section:

Explanation:

Box 1: HASH

Box 2: OrderDateKey

In most cases, table partitions are created on a date column. A way to eliminate rollbacks is to use Metadata Only operations like partition switching for data management. For example, rather than execute a DELETE statement to delete all rows in a table where the order_date was in October of 2001, you could partition your data early. Then you can switch out the partition with data for an empty partition from another table.
Reference: https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

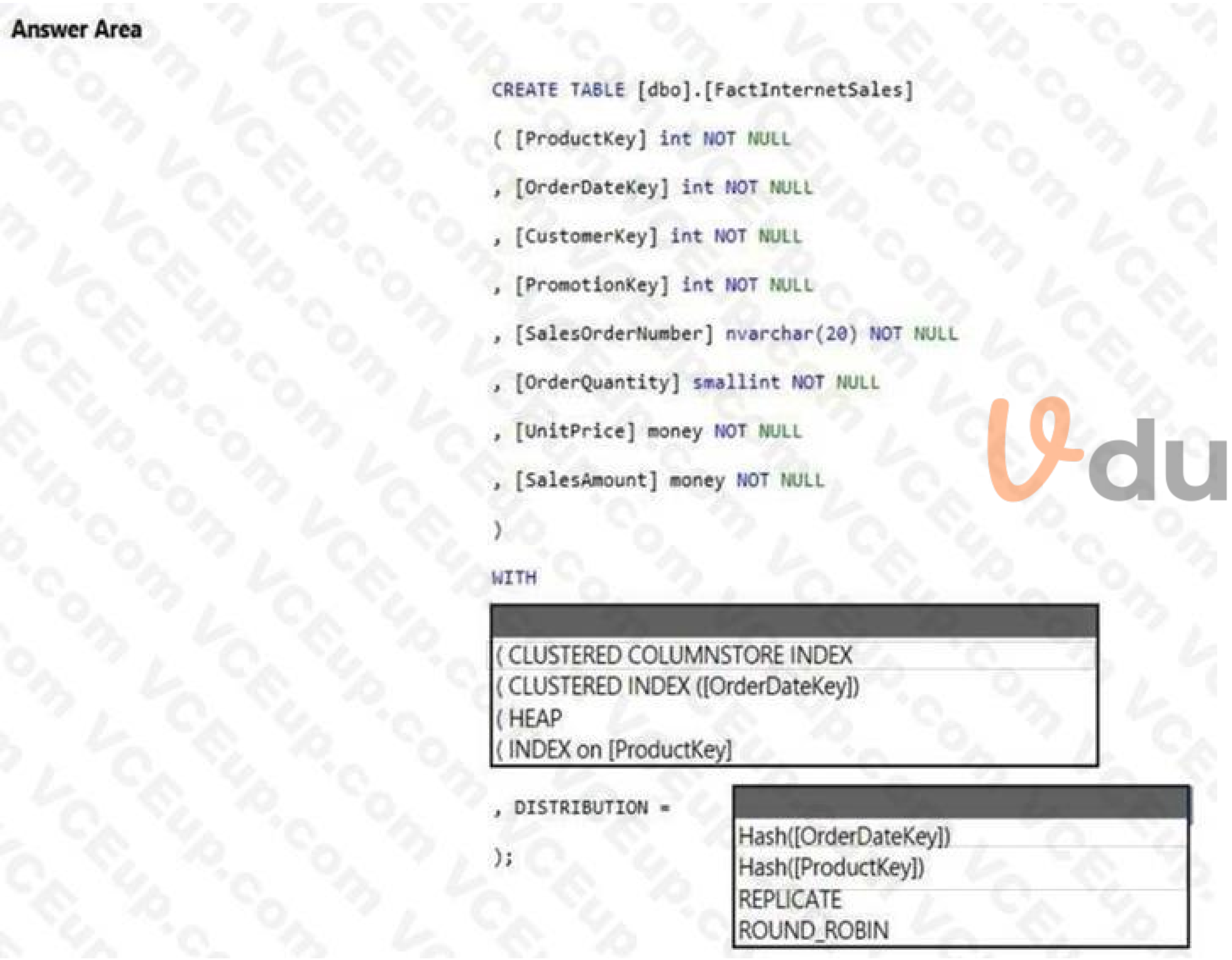
QUESTION 10

HOTSPOT

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a table named FactInternetSales that will be a large fact table in a dimensional model. FactInternetSales will contain 100 million rows and two columns named SalesAmount and OrderQuantity. Queries executed on FactInternetSales will aggregate the values in SalesAmount and OrderQuantity from the last year for a specific product. The solution must minimize the data size and query execution time. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

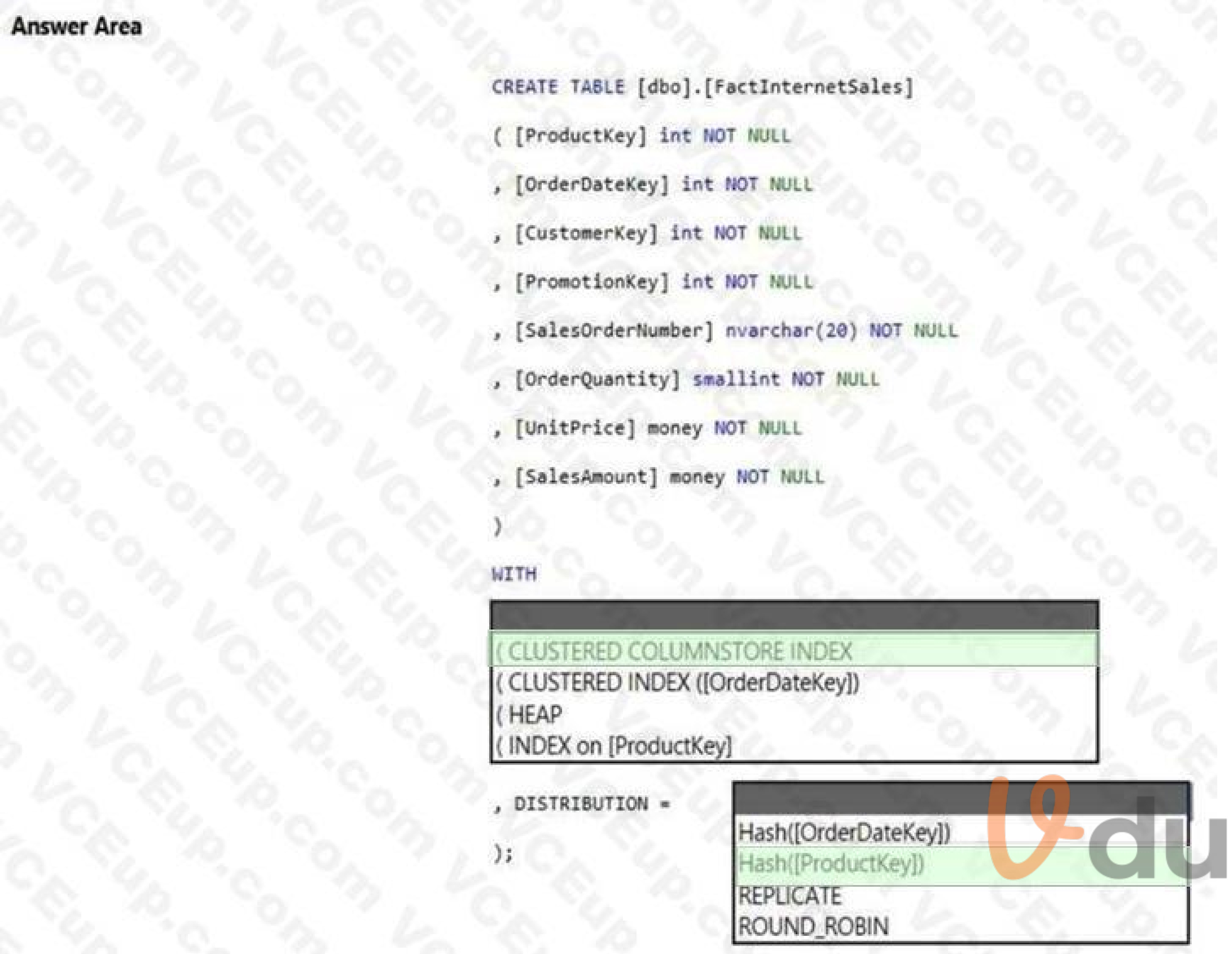
```
CREATE TABLE [dbo].[FactInternetSales]
( [ProductKey] int NOT NULL
, [OrderDateKey] int NOT NULL
, [CustomerKey] int NOT NULL
, [PromotionKey] int NOT NULL
, [SalesOrderNumber] nvarchar(20) NOT NULL
, [OrderQuantity] smallint NOT NULL
, [UnitPrice] money NOT NULL
, [SalesAmount] money NOT NULL
)
WITH
```

(CLUSTERED COLUMNSTORE INDEX
(CLUSTERED INDEX ([OrderDateKey])
(HEAP
(INDEX on [ProductKey]

```
, DISTRIBUTION =
);
```

Hash([OrderDateKey])
Hash([ProductKey])
REPLICATE
ROUND_ROBIN

Answer Area:



Answer Area

```
CREATE TABLE [dbo].[FactInternetSales]
( [ProductKey] int NOT NULL
, [OrderDateKey] int NOT NULL
, [CustomerKey] int NOT NULL
, [PromotionKey] int NOT NULL
, [SalesOrderNumber] nvarchar(20) NOT NULL
, [OrderQuantity] smallint NOT NULL
, [UnitPrice] money NOT NULL
, [SalesAmount] money NOT NULL
)
WITH
```

[✓ (CLUSTERED COLUMNSTORE INDEX]
(CLUSTERED INDEX ([OrderDateKey])
(HEAP
(INDEX on [ProductKey]

```
, DISTRIBUTION =
);
```

Hash([OrderDateKey])
[✓ Hash([ProductKey])]
REPLICATE
ROUND_ROBIN

Section:

Explanation:

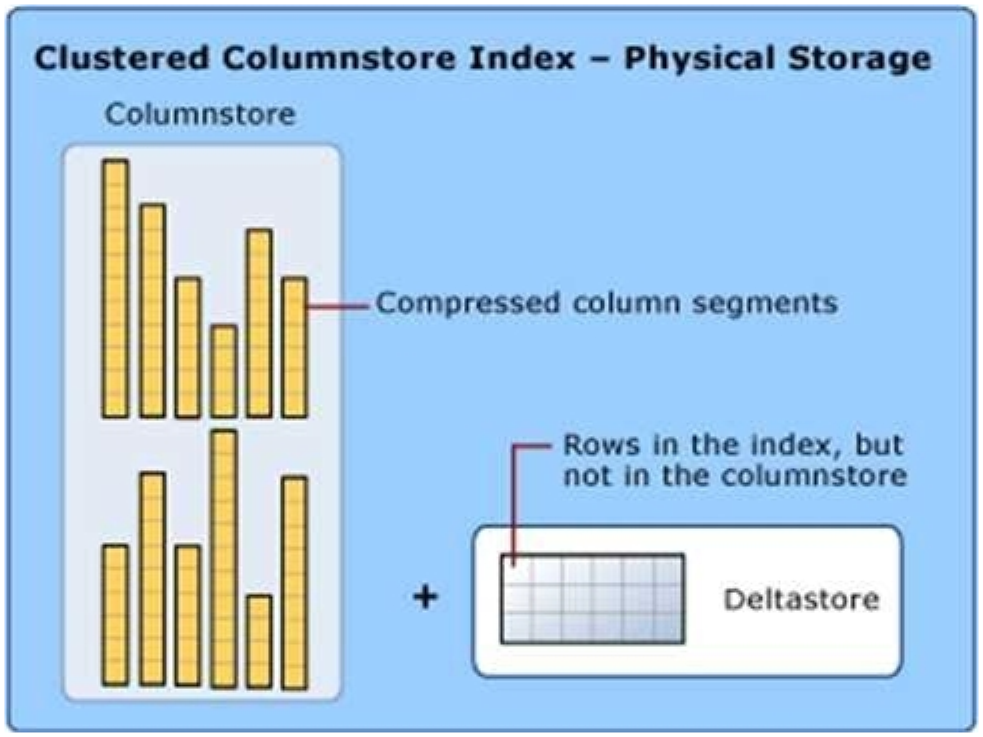
Box 1: (CLUSTERED COLUMNSTORE INDEX

CLUSTERED COLUMNSTORE INDEX

Columnstore indexes are the standard for storing and querying large data warehousing fact tables. This index uses column-based data storage and query processing to achieve gains up to 10 times the query performance in your data warehouse over traditional row-oriented storage. You can also achieve gains up to 10 times the data compression over the uncompressed data size. Beginning with SQL Server 2016 (13.x) SP1, columnstore indexes enable operational analytics: the ability to run performant real-time analytics on a transactional workload.

Note: Clustered columnstore index

A clustered columnstore index is the physical storage for the entire table.


Clustered Columnstore Index – Physical Storage
Columnstore
Compressed column segments
Rows in the index, but not in the columnstore
+
Deltastore

To reduce fragmentation of the column segments and improve performance, the columnstore index might store some data temporarily into a clustered index called a deltastore and a B-tree list of IDs for deleted rows. The deltastore operations are handled behind the scenes. To return the correct query results, the clustered columnstore index combines query results from both the columnstore and the deltastore.

Box 2: HASH([ProductKey])

A hash distributed table distributes rows based on the value in the distribution column. A hash distributed table is designed to achieve high performance for queries on large tables. Choose a distribution column with data that distributes evenly

Reference:

https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-overview
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehousetables-overview https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehousetables-distribute

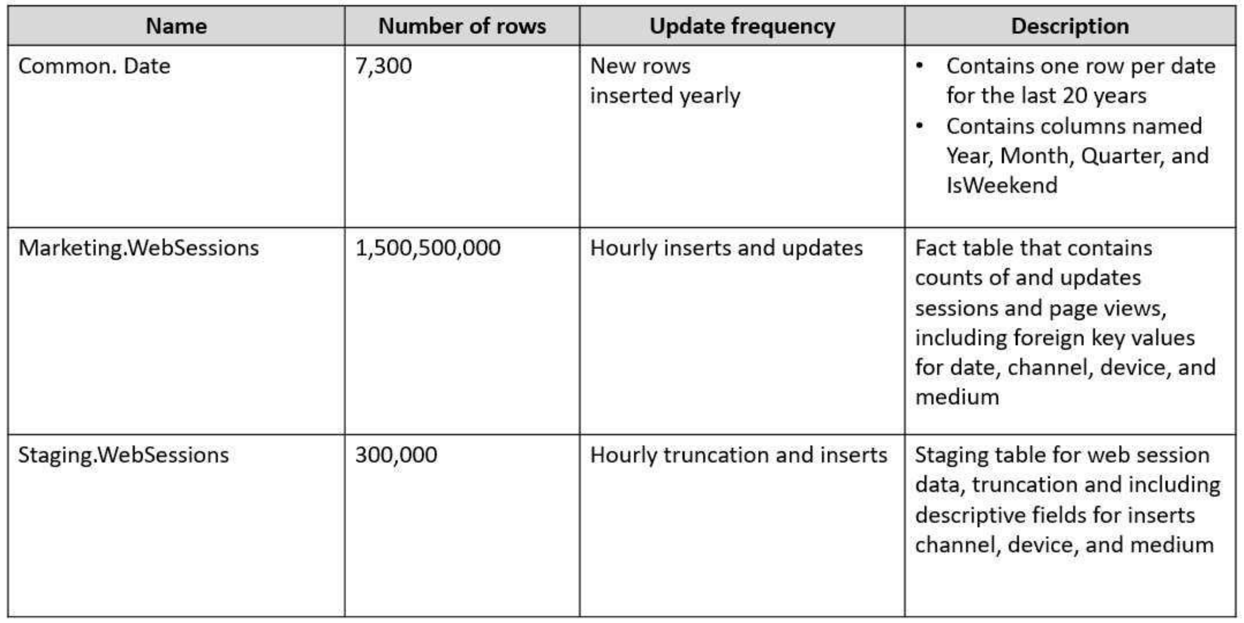
QUESTION 11

DRAG DROP

You have an Azure subscription.

You plan to build a data warehouse in an Azure Synapse Analytics dedicated SQL pool named pool1 that will contain staging tables and a dimensional model. Pool1 will contain the following tables. You need to design the table storage for pool1. The solution must meet the following requirements:

Maximize the performance of data loading operations to Staging.WebSessions. Minimize query times for reporting queries against the dimensional model. Which type of table distribution should you use for each table? To answer, drag the appropriate table distribution types to the correct tables.



Name	Number of rows	Update frequency	Description
Common. Date	7,300	New rows inserted yearly	• Contains one row per date for the last 20 years • Contains columns named Year, Month, Quarter, and IsWeekend
Marketing.WebSessions	1,500,500,000	Hourly inserts and updates	Fact table that contains counts of and updates sessions and page views, including foreign key values for date, channel, device, and medium
Staging.WebSessions	300,000	Hourly truncation and inserts	Staging table for web session data, truncation and including descriptive fields for inserts channel, device, and medium

Each table distribution type 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:

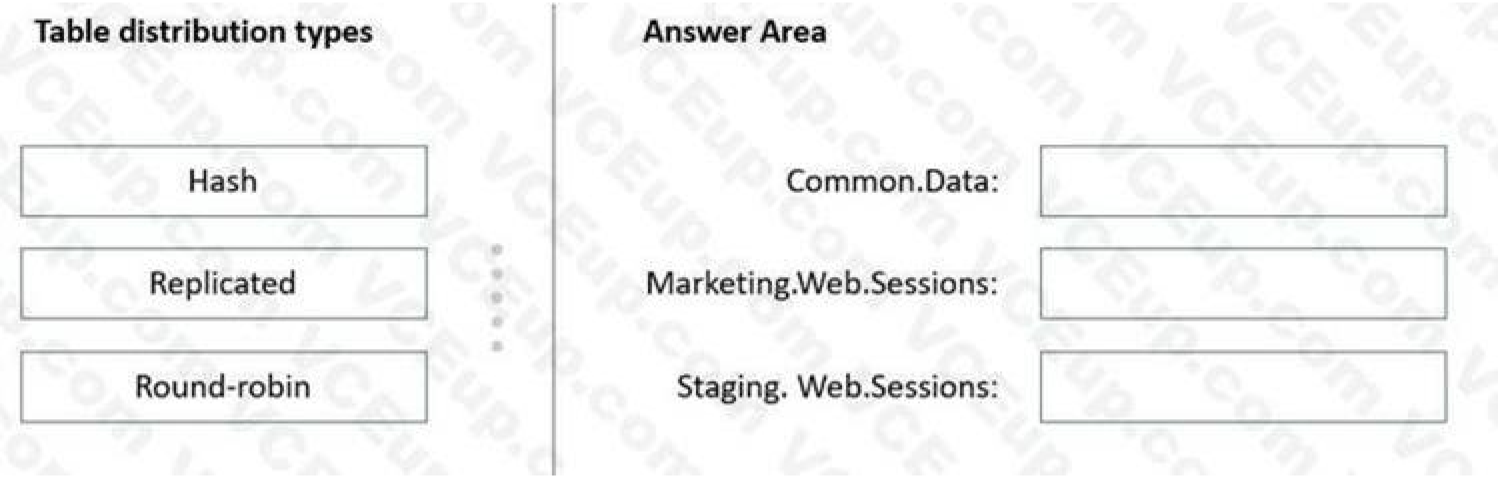


Table distribution types

Hash

Replicated

Round-robin

Answer Area

Common.Data:

Marketing.Web.Sessions:

Staging. Web.Sessions:

Correct Answer:

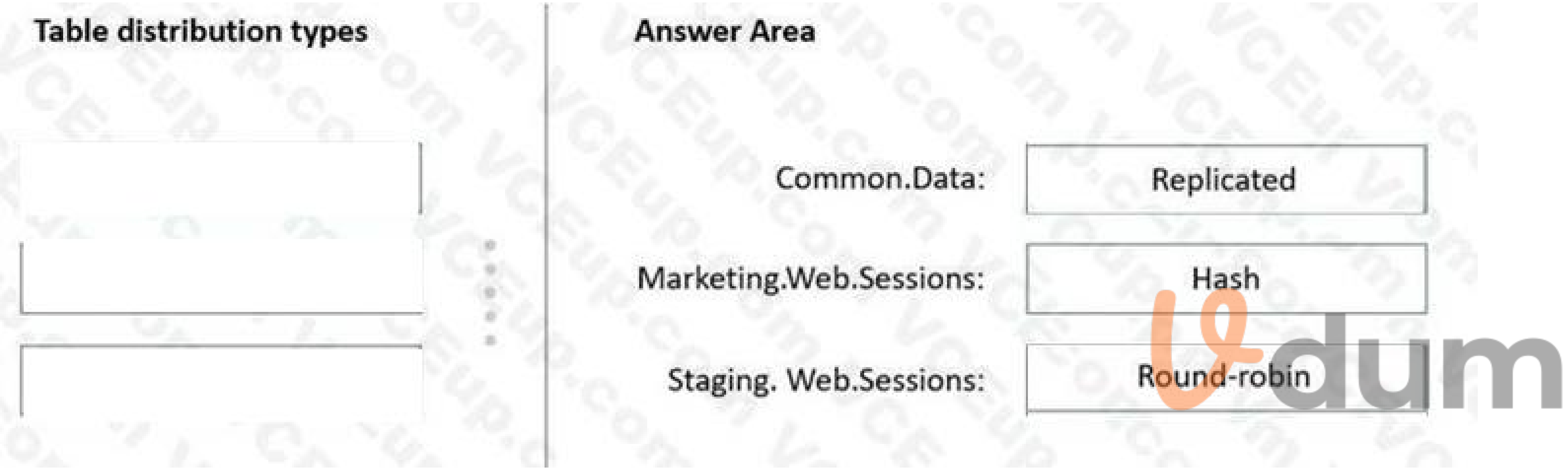


Table distribution types

Answer Area

Common.Data: Replicated

Marketing.Web.Sessions: Hash

Staging. Web.Sessions: Round-robin

Section:

Explanation:

Box 1: Replicated

The best table storage option for a small table is to replicate it across all the Compute nodes. Box 2: Hash

Hash-distribution improves query performance on large fact tables. Box 3: Round-robin

Round-robin distribution is useful for improving loading speed.

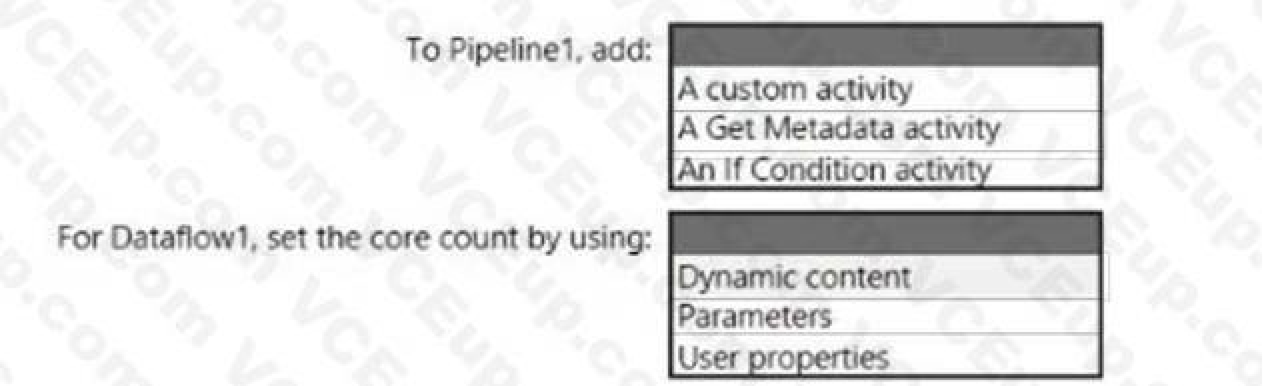
Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-datawarehouse-tables-distribute

QUESTION 12

HOTSPOT

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

Hot Area:



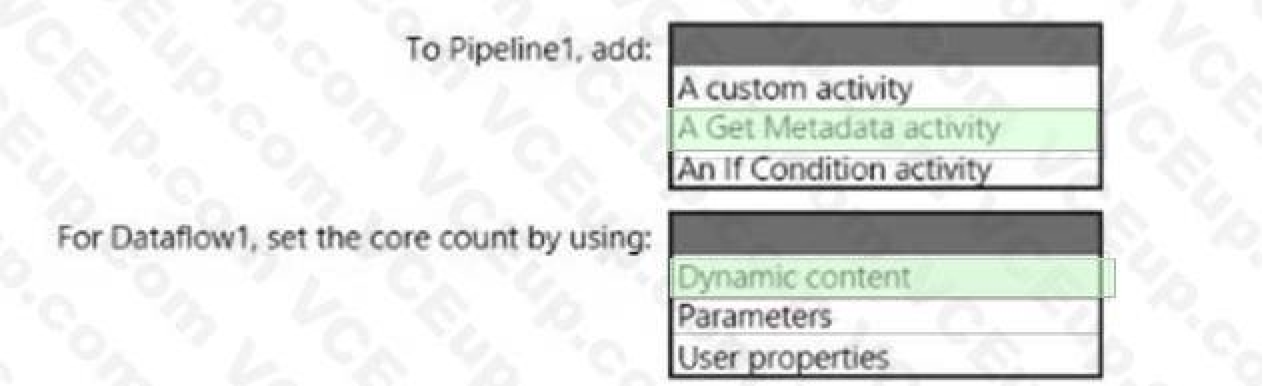
To Pipeline1, add:

- A custom activity
- A Get Metadata activity
- An If Condition activity

For Dataflow1, set the core count by using:

- Dynamic content
- Parameters
- User properties

Answer Area:



To Pipeline1, add:

- A custom activity
- [✓ A Get Metadata activity]
- An If Condition activity

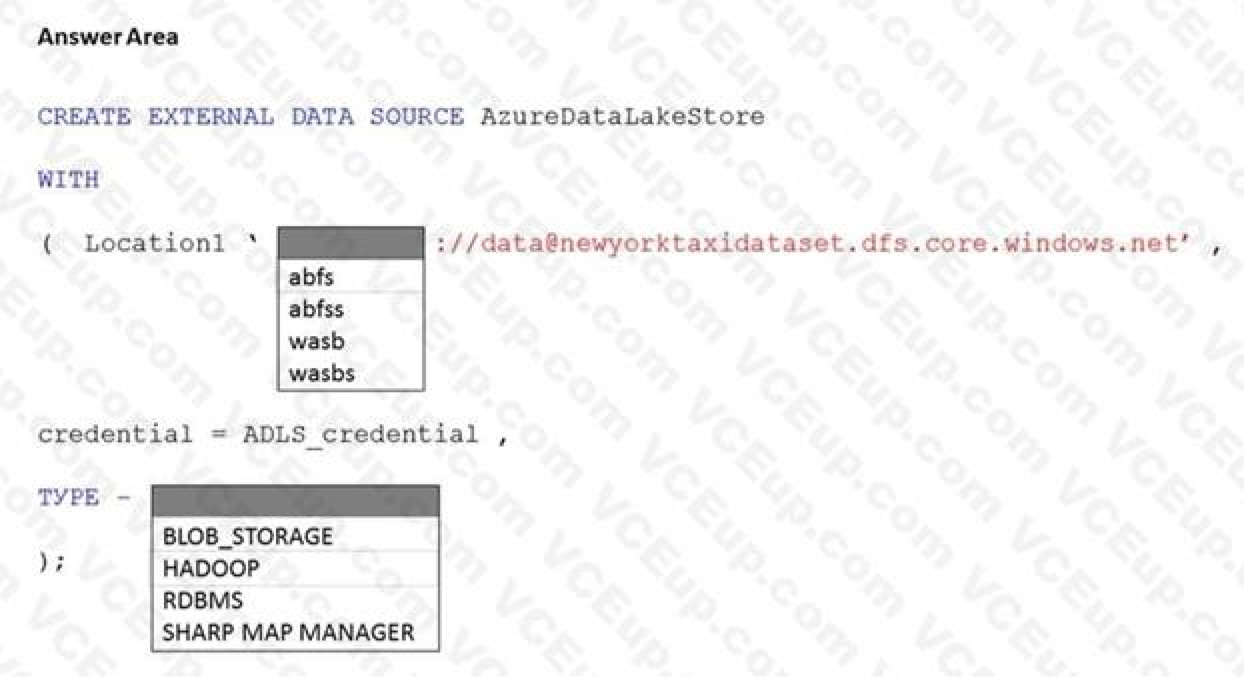
For Dataflow1, set the core count by using:

- [✓ Dynamic content]
- Parameters
- User properties

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

QUESTION 13
HOTSPOT
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.

Hot Area:



Answer Area

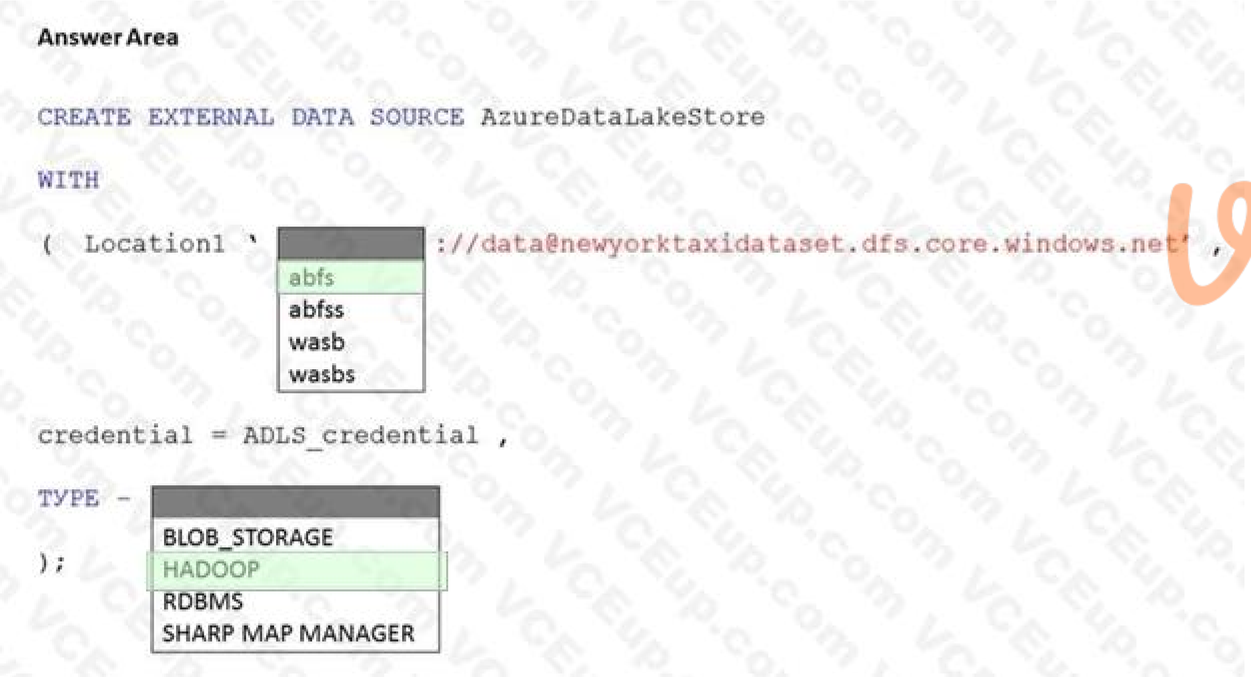
```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore

WITH

(  Location1 '              ://data@newyorktaxidataset.dfs.core.windows.net' ,
```

- abfs
- abfss
- wasb
- wasbs

```
credential = ADLS_credential ,

TYPE -

);
```

- BLOB_STORAGE
- HADOOP
- RDBMS
- SHARP MAP MANAGER

Answer Area:



Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore

WITH

(  Location1 '              ://data@newyorktaxidataset.dfs.core.windows.net' ,
```

- abfs
- abfss
- wasb
- wasbs

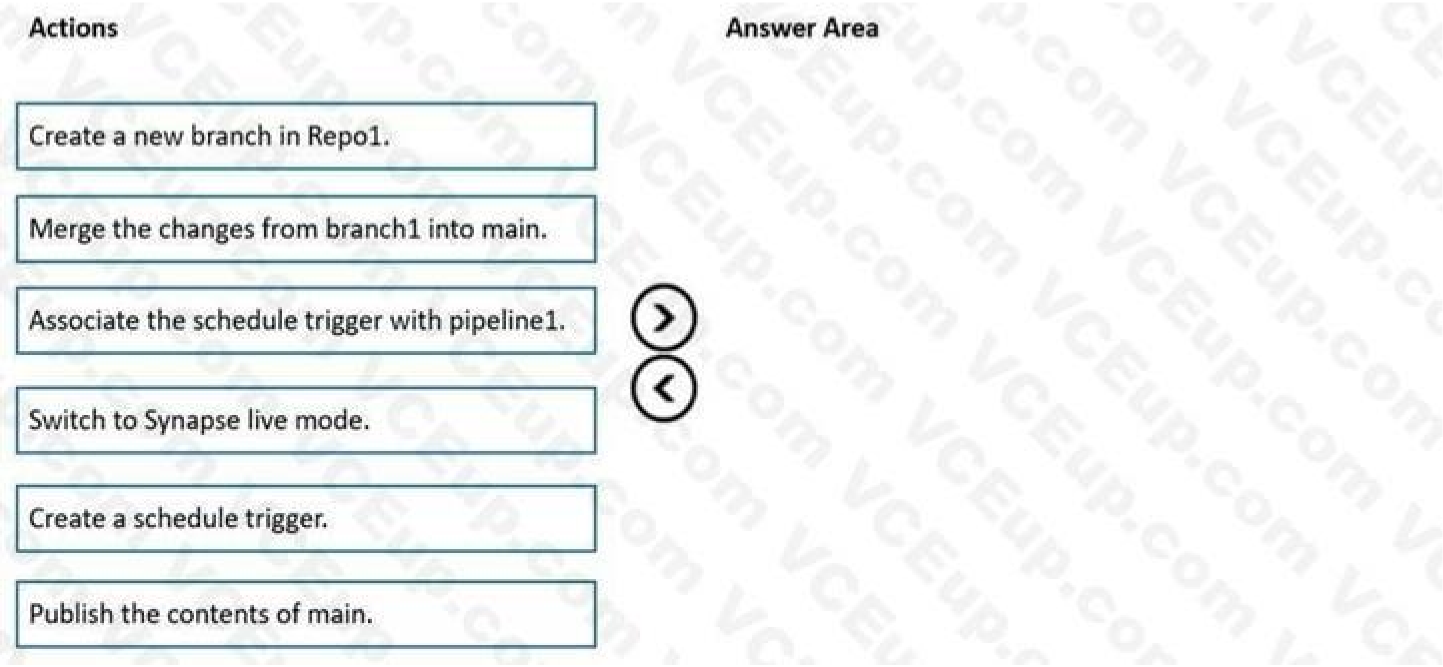
```
credential = ADLS_credential ,

TYPE -

);
```

- BLOB_STORAGE
- HADOOP
- RDBMS
- SHARP MAP MANAGER

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

QUESTION 14
DRAG DROP
You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1. Branch1 contains an
Azure Synapse pipeline named pipeline1. In workspace1, you complete testing of pipeline1. You need to schedule pipeline1 to run daily at 6 AM.
Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order. NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Select and Place:



Actions

Answer Area

Create a new branch in Repo1.

Merge the changes from branch1 into main.

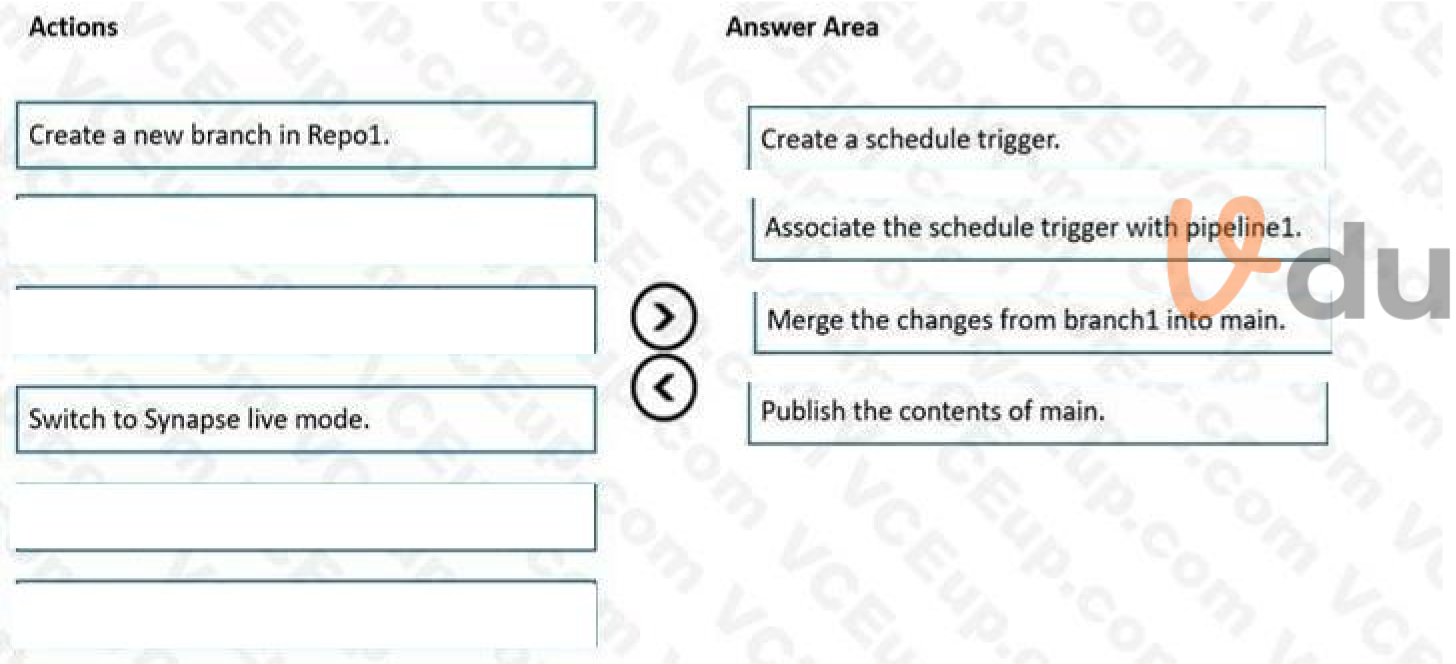
Associate the schedule trigger with pipeline1.

Switch to Synapse live mode.

Create a schedule trigger.

Publish the contents of main.

Correct Answer:



Actions

Create a new branch in Repo1.

Switch to Synapse live mode.

Answer Area

Create a schedule trigger.

Associate the schedule trigger with pipeline1.

Merge the changes from branch1 into main.

Publish the contents of main.

Section:

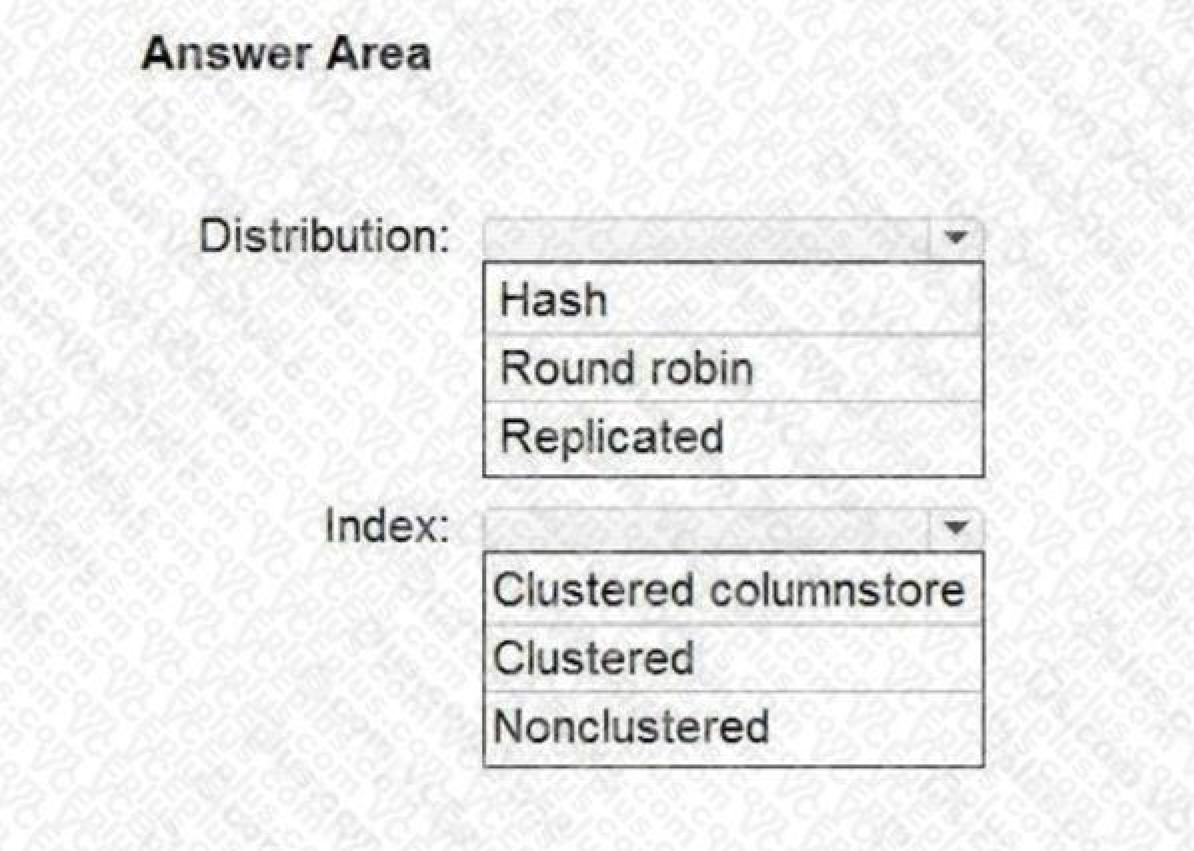
Explanation:

QUESTION 15

HOTSPOT

You are designing an enterprise data warehouse in Azure Synapse Analytics that will store website traffic analytics in a star schema. You plan to have a fact table for website visits. The table will be approximately 5 GB. You need to recommend which distribution type and index type to use for the table. The solution must provide the fastest query performance. What should you recommend? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

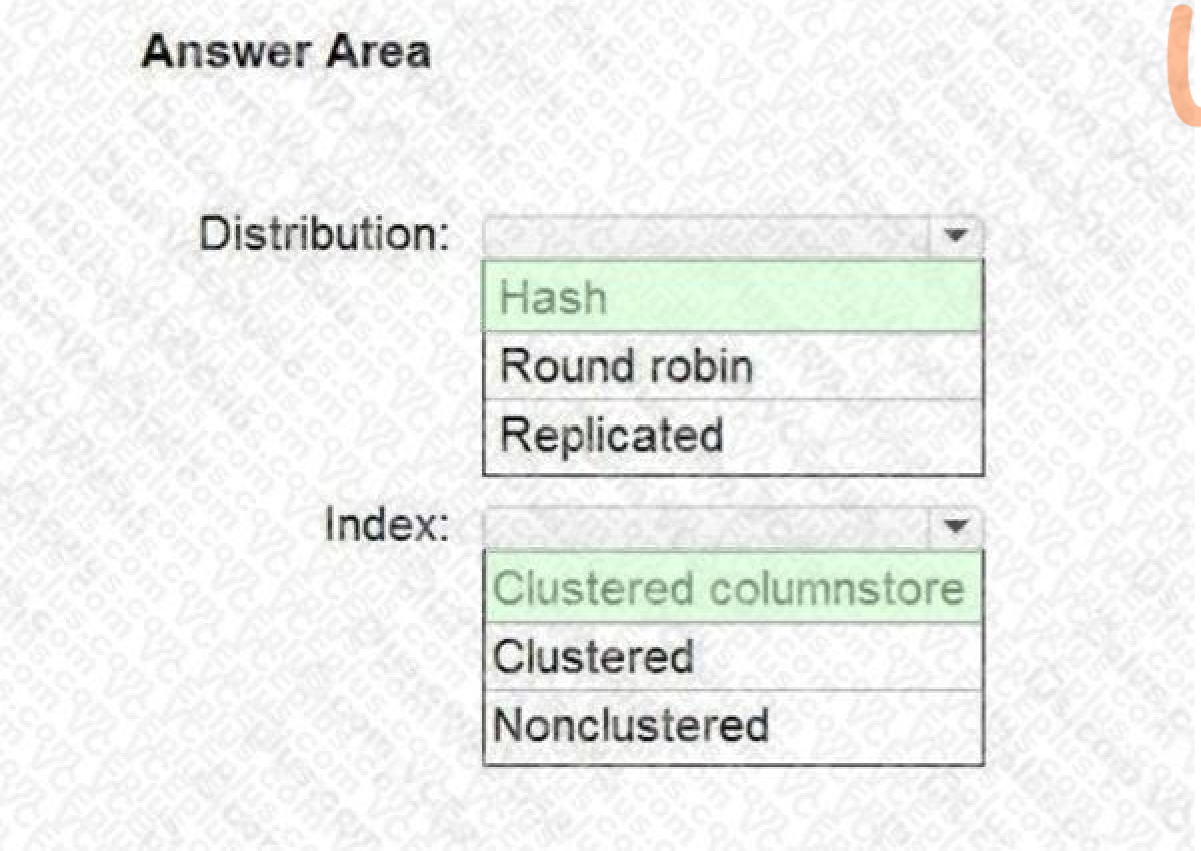
Distribution:

- Hash
- Round robin
- Replicated

Index:

- Clustered columnstore
- Clustered
- Nonclustered

Answer Area:



Answer Area

Distribution:

- Hash
- Round robin
- Replicated

Index:

- Clustered columnstore
- Clustered
- Nonclustered

Section:
Explanation:
Box 1: Hash
Consider using a hash-distributed table when:
The table size on disk is more than 2 GB.

The table has frequent insert, update, and delete operations. Box 2: Clustered columnstore
Clustered columnstore tables offer both the highest level of data compression and the best overall query performance.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-index

QUESTION 16

DRAG DROP

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:

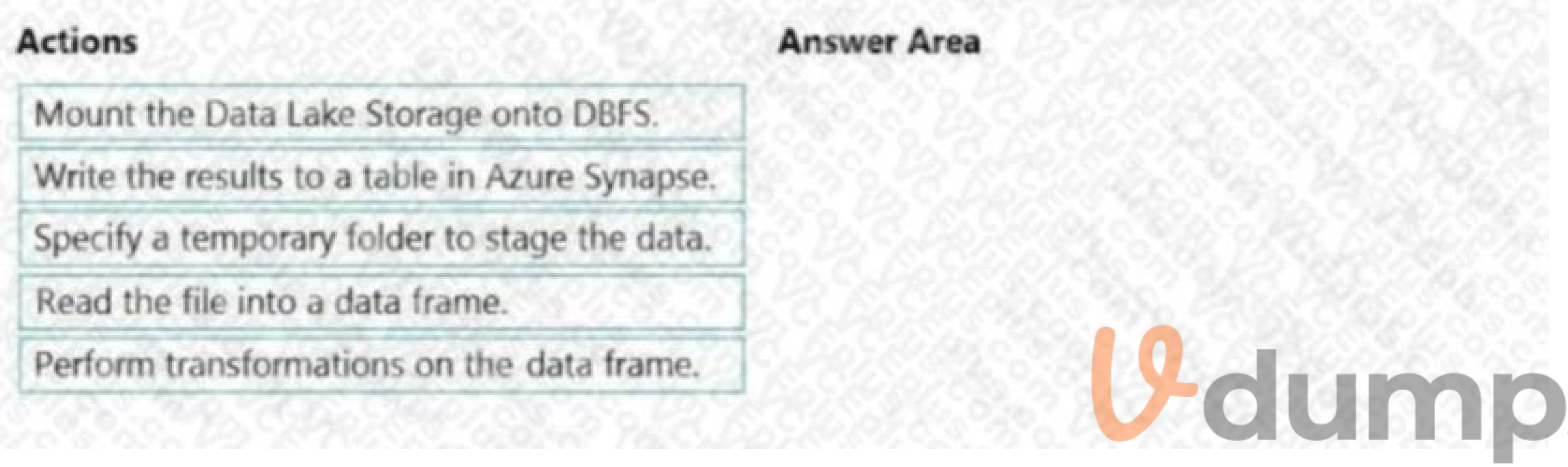
A destination table in Azure Synapse

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.

Select and Place:



Actions

Mount the Data Lake Storage onto DBFS.

Write the results to a table in Azure Synapse.

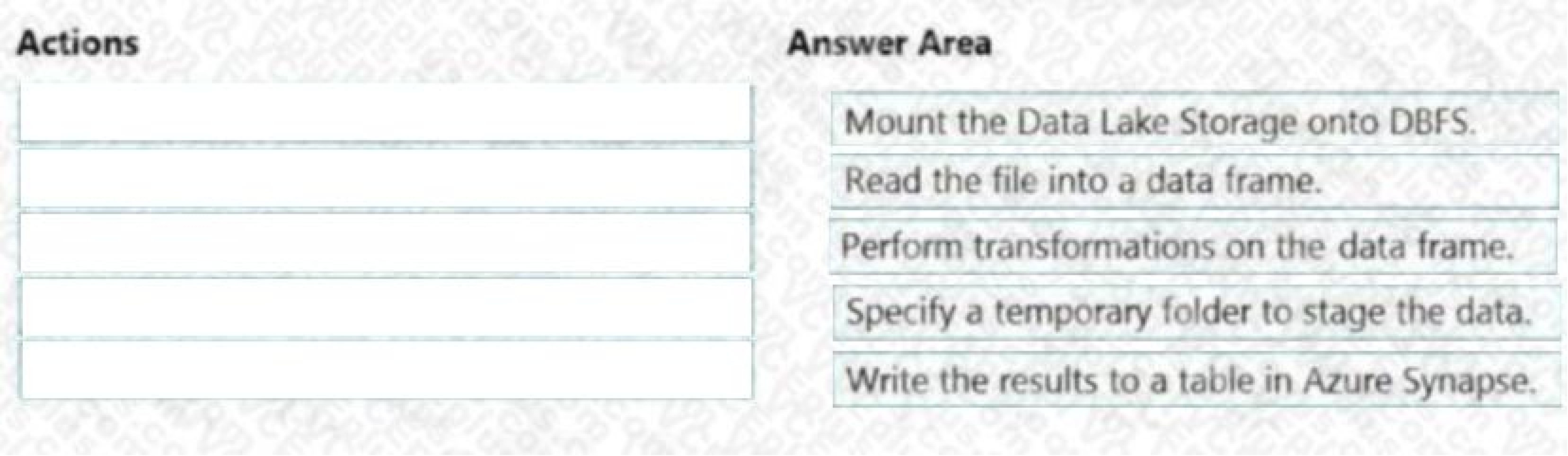
Specify a temporary folder to stage the data.

Read the file into a data frame.

Perform transformations on the data frame.

Answer Area

Correct Answer:



Actions

Answer Area

Mount the Data Lake Storage onto DBFS.

Read the file into a data frame.

Perform transformations on the data frame.

Specify a temporary folder to stage the data.

Write the results to a table in Azure Synapse.

Section:

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

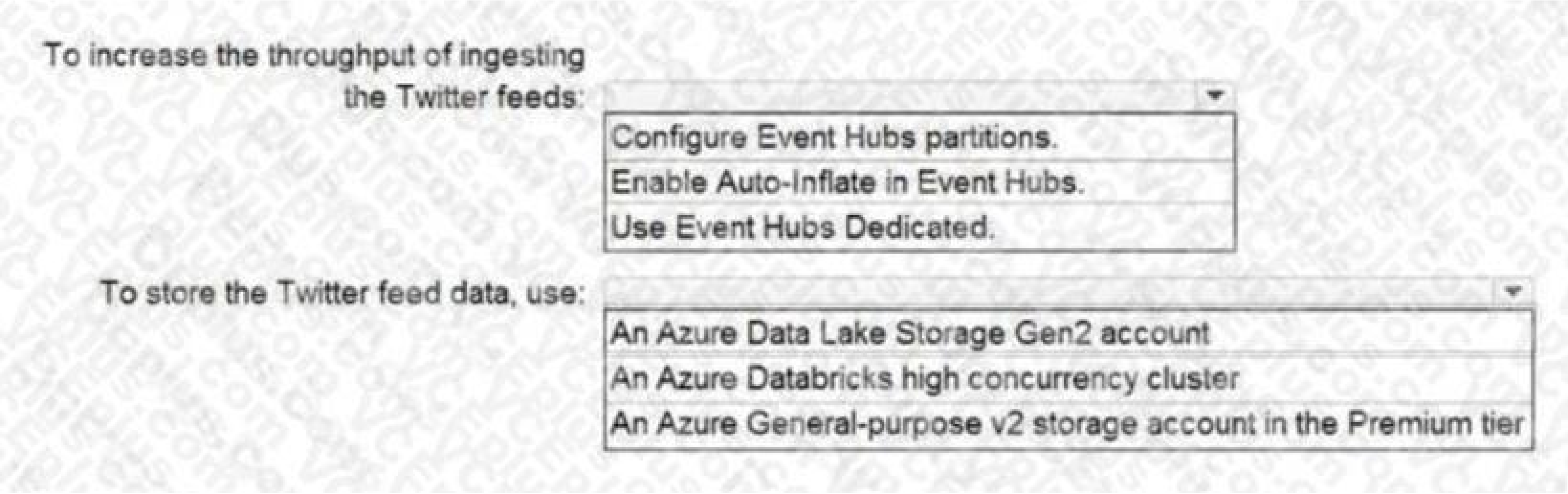
QUESTION 17

HOTSPOT

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 is worth one point.

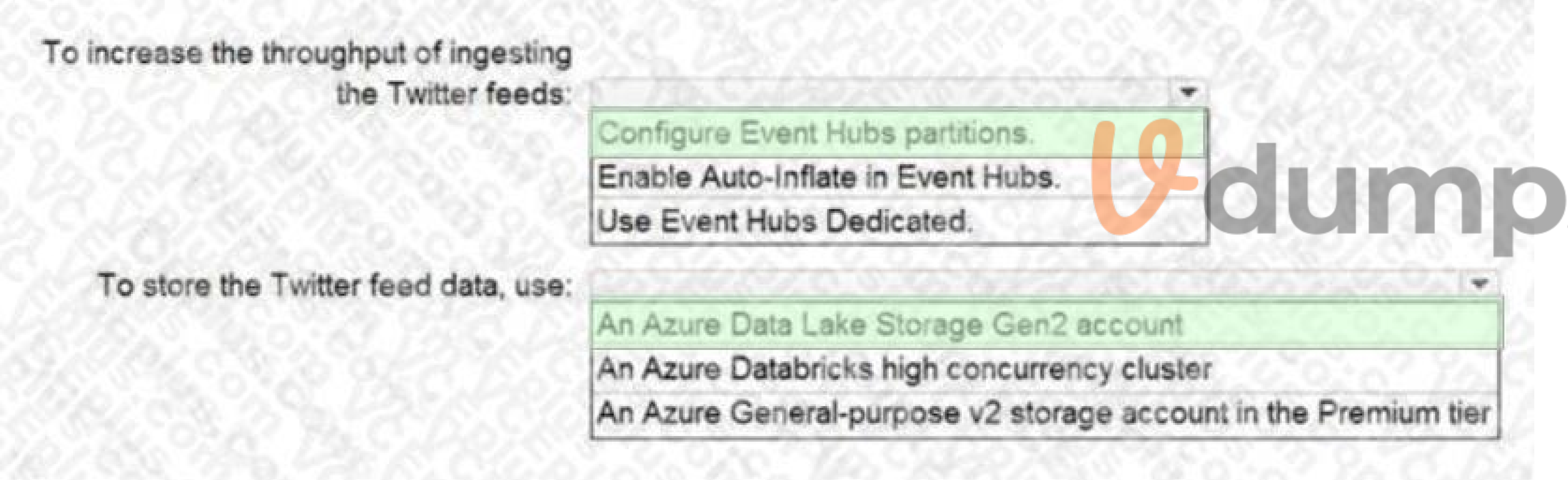
Hot Area:



To increase the throughput of ingesting the Twitter feeds:
- Configure Event Hubs partitions.
- Enable Auto-Inflate in Event Hubs.
- Use Event Hubs Dedicated.

To store the Twitter feed data, use:
- An Azure Data Lake Storage Gen2 account
- An Azure Databricks high concurrency cluster
- An Azure General-purpose v2 storage account in the Premium tier

Answer Area:



To increase the throughput of ingesting the Twitter feeds:
- [✓ Configure Event Hubs partitions.]
- Enable Auto-Inflate in Event Hubs.
- Use Event Hubs Dedicated.

To store the Twitter feed data, use:
- [✓ An Azure Data Lake Storage Gen2 account]
- An Azure Databricks high concurrency cluster
- An Azure General-purpose v2 storage account in the Premium tier

Section:

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.

Incorrect Answers:

Event Hubs Dedicated: Event Hubs clusters offer single-tenant deployments for customers with the most demanding streaming needs. This single-tenant offering has a guaranteed 99.99% SLA and is available only on our Dedicated pricing tier.

Auto-Inflate: The Auto-inflate feature of Event Hubs automatically scales up by increasing the number of TUs, to meet usage needs.

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).

Incorrect Answers:

Azure Databricks: An Azure administrator with the proper permissions can configure Azure Active Directory conditional access to control where and when users are permitted to sign in to Azure Databricks. Azure Storage supports using Azure Active Directory (Azure AD) to authorize requests to blob data. You can scope access to Azure blob resources at the following levels, beginning with the narrowest scope:

- An individual container. At this scope, a role assignment applies to all of the blobs in the container, as well as container properties and metadata.
- The storage account. At this scope, a role assignment applies to all containers and their blobs. - The resource group. At this scope, a role assignment applies to all of the containers in all of the storage accounts in the resource group.
- The subscription. At this scope, a role assignment applies to all of the containers in all of the storage accounts in all of the resource groups in the subscription. - A management group.
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

QUESTION 18
You have an Azure Data Factory pipeline that is triggered hourly. The pipeline has had 100% success for the past seven days.
The pipeline execution fails, and two retries that occur 15 minutes apart also fail. The third failure returns the following error.
ErrorCode=UserErrorFileNotFound,'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,Message=ADL S Gen2 operation failed for: Operation returned an invalid status code 'NotFound'. Account: 'contosoproduksouth'. Filesystem: wwi. Path: 'BIKES/CARBON/year=2021/month=01/day=10/hour=06'. ErrorCode: 'PathNotFound'. Message: 'The specified path does not exist.'. RequestId: '6d269b78-901f-001b-4924-e7a7bc000000'. TimeStamp: 'Sun, 10 Jan 2021 07:45:05
What is a possible cause of the error?

A. The parameter used to generate year=2021/month=01/day=10/hour=06 was incorrect.
B. From 06:00 to 07:00 on January 10, 2021, there was no data in wwi/BIKES/CARBON.
C. From 06:00 to 07:00 on January 10, 2021, the file format of data in wwi/BIKES/CARBON was incorrect.
D. The pipeline was triggered too early.

Correct Answer: C
Section:
Explanation:
A file is missing.

QUESTION 19
You have an Azure Synapse Analytics job that uses Scala.
You need to view the status of the job.
What should you do?

A. From Synapse Studio, select the workspace. From Monitor, select SQL requests.
B. From Azure Monitor, run a Kusto query against the AzureDiagnostics table.
C. From Synapse Studio, select the workspace. From Monitor, select Apache Sparks applications.
D. From Azure Monitor, run a Kusto query against the SparkLoggingEvent_CL table.

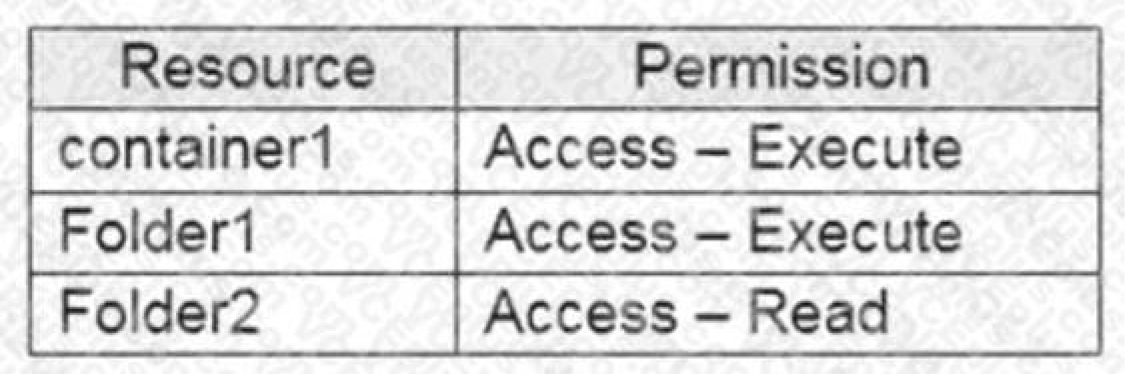
Correct Answer: C
Section:
Explanation:
Use Synapse Studio to monitor your Apache Spark applications. To monitor running Apache Spark application Open Monitor, then select Apache Spark applications. To view the details about the Apache Spark applications that are running, select the submitting Apache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

QUESTION 20
You are designing database for an Azure Synapse Analytics dedicated SQL pool to support workloads for detecting ecommerce transaction fraud. Data will be combined from multiple ecommerce sites and can include sensitive financial information such as credit card numbers. You need to recommend a solution that meets the following requirements:
Users must be able to identify potentially fraudulent transactions. Users must be able to use credit cards as a potential feature in models. Users must NOT be able to access the actual credit card numbers. What should you include in the recommendation?

A. Transparent Data Encryption (TDE)
B. row-level security (RLS)
C. column-level encryption
D. Azure Active Directory (Azure AD) pass-through authentication

Correct Answer: B
Section:
Explanation:

QUESTION 21
You have an Azure subscription linked to an Azure Active Directory (Azure AD) tenant that contains a service principal named ServicePrincipal1. The subscription contains an Azure Data Lake Storage account named adls1. Adls1 contains a folder named Folder2 that has a URI of https://adls1.dfs.core.windows.net/container1/Folder1/Folder2/. ServicePrincipal1 has the access control list (ACL) permissions shown in the following table.



Resource	Permission
container1	Access – Execute
Folder1	Access – Execute
Folder2	Access – Read

You need to ensure that ServicePrincipal1 can perform the following actions:
Traverse child items that are created in Folder2. Read files that are created in Folder2. The solution must use the principle of least privilege.
Which two permissions should you grant to ServicePrincipal1 for Folder2? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. Access - Read
B. Access - Write
C. Access - Execute
D. Default - Read
E. Default - Write
F. Default - Execute

Correct Answer: D, F
Section:
Explanation:
Execute (X) permission is required to traverse the child items of a folder. There are two kinds of access control lists (ACLs), Access ACLs and Default ACLs. Access ACLs: These control access to an object. Files and folders both have Access ACLs. Default ACLs: A "template" of ACLs associated with a folder that determine the Access ACLs for any child items that are created under that folder. Files do not have Default ACLs. Reference:
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control

QUESTION 22
You manage an enterprise data warehouse in Azure Synapse Analytics. Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries. You need to monitor resource utilization to determine the source of the performance issues. Which metric should you monitor?

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

D. CPU percentage

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

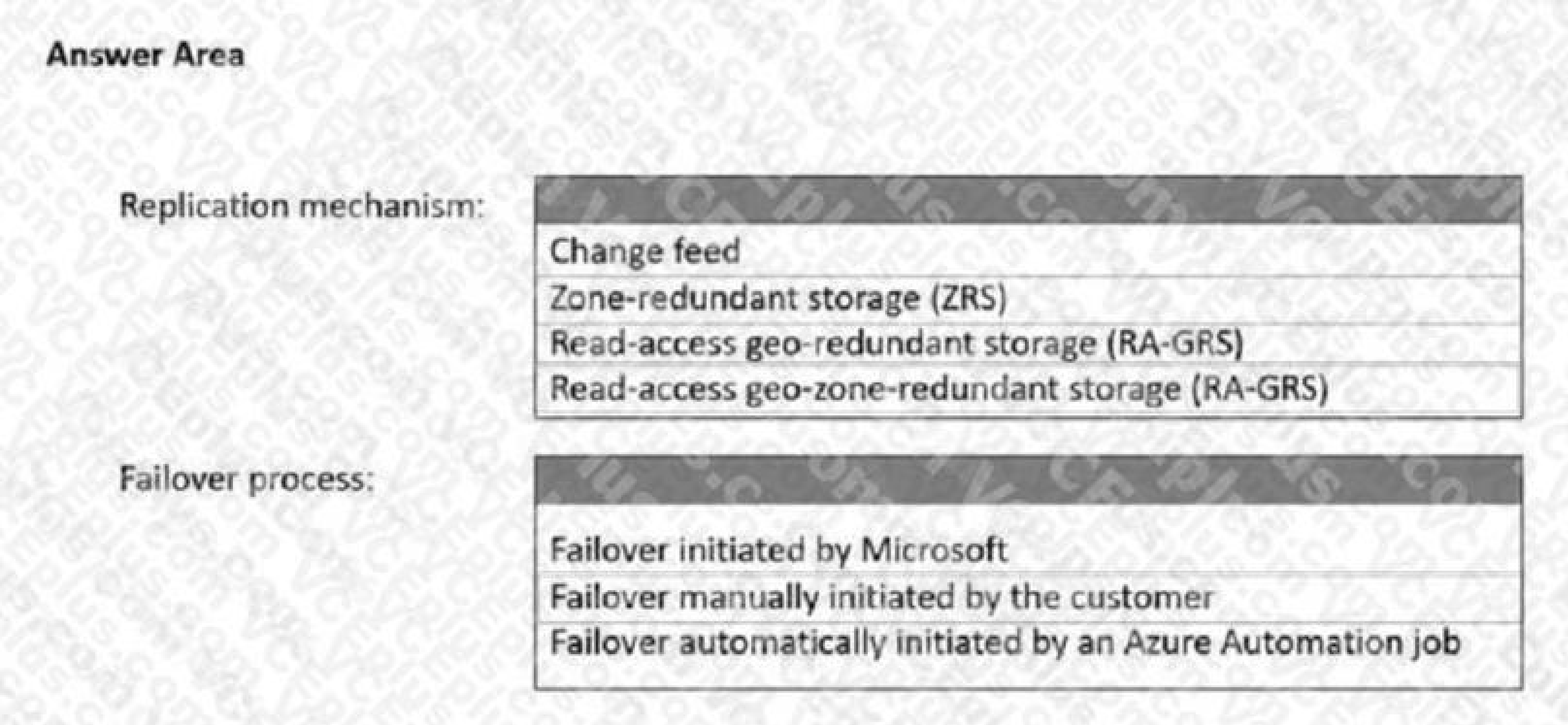
QUESTION 23
You have an Azure data factory.
You need to examine the pipeline failures from the last 180 days. What should you use?

A. the Activity log blade for the Data Factory resource
B. Pipeline runs in the Azure Data Factory user experience
C. the Resource health blade for the Data Factory resource
D. Azure Data Factory activity runs in Azure Monitor

Correct Answer: D
Section:
Explanation:
Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor

QUESTION 24
HOTSPOT
You plan to create an Azure Data Lake Storage Gen2 account. You need to recommend a storage solution that meets the following requirements:
Provides the highest degree of data resiliency
Ensures that content remains available for writes if a primary data center fails 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.

Hot Area:



Answer Area

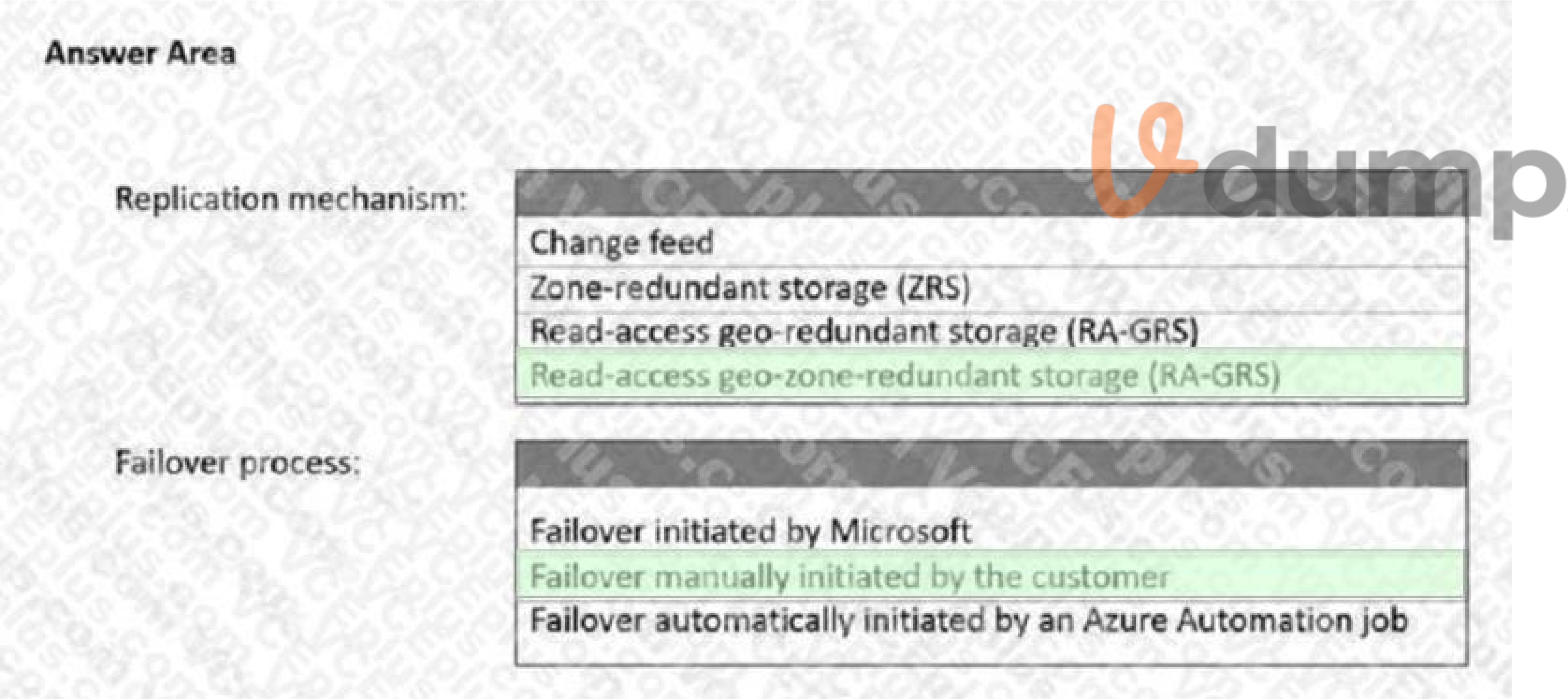
Replication mechanism:

- Change feed
- Zone-redundant storage (ZRS)
- Read-access geo-redundant storage (RA-GRS)
- Read-access geo-zone-redundant storage (RA-GRS)

Failover process:

- Failover initiated by Microsoft
- Failover manually initiated by the customer
- Failover automatically initiated by an Azure Automation job

Answer Area:



Answer Area

Replication mechanism:

- Change feed
- Zone-redundant storage (ZRS)
- Read-access geo-redundant storage (RA-GRS)
- [✓ Read-access geo-zone-redundant storage (RA-GRS)]

Failover process:

- Failover initiated by Microsoft
- [✓ Failover manually initiated by the customer]
- Failover automatically initiated by an Azure Automation job

Section:
Explanation:
https://docs.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance?toc=/azure/storage/blobs/toc.json https://docs.microsoft.com/en-us/answers/questions/32583/azure-data-lakegen2-disaster-recoverystorage-acco.html

QUESTION 25
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. Update

B. Complete

C. Append

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

QUESTION 26
You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1. You need to determine the size of the transaction log file for each distribution of DW1. What should you do?

A. On DW1, execute a query against the sys.database_files dynamic management view.

B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.

C. Execute a query against the logs of DW1 by using the Get-AzOperationalInsightsSearchResult PowerShell cmdlet.

D. On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.

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

QUESTION 27
You are designing an anomaly detection solution for streaming data from an Azure IoT hub. The solution must meet the following requirements:
Send the output to Azure Synapse.
Identify spikes and dips in time series data.
Minimize development and configuration effort.
Which should you include in the solution?

A. Azure Databricks

B. Azure Stream Analytics

C. Azure SQL Database

Correct Answer: B
Section:
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 28
A company uses Azure Stream Analytics to monitor devices.
The company plans to double the number of devices that are monitored. You need to monitor a Stream Analytics job to ensure that there are enough processing resources to handle the additional load. Which metric should you monitor?

A. Early Input Events
B. Late Input Events
C. Watermark delay
D. Input Deserialization Errors

Correct Answer: A
Section:

QUESTION 29
You have an Azure Stream Analytics job.
You need to ensure that the job has enough streaming units provisioned. You configure monitoring of the SU % Utilization metric.
Which two additional metrics should you monitor? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

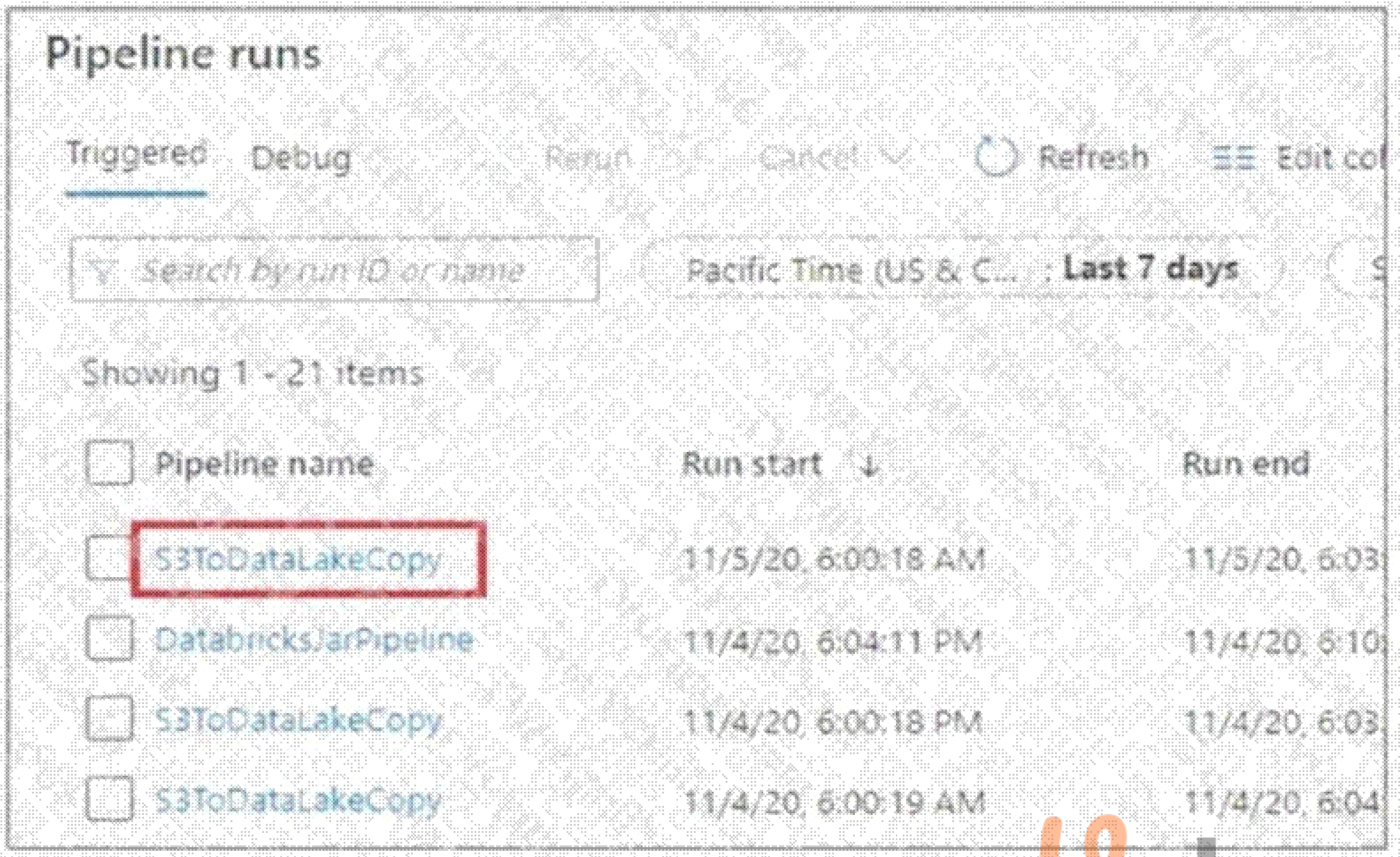
A. Backlogged Input Events
B. Watermark Delay
C. Function Events
D. Out of order Events
E. Late Input Events

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

QUESTION 30
You have an activity in an Azure Data Factory pipeline. The activity calls a stored procedure in a data warehouse in Azure Synapse Analytics and runs daily.
You need to verify the duration of the activity when it ran last. What should you use?

A. activity runs in Azure Monitor
B. Activity log in Azure Synapse Analytics
C. the sys.dm_pdw_wait_stats data management view in Azure Synapse Analytics
D. an Azure Resource Manager template

Correct Answer: A
Section:
Explanation:
Monitor activity runs. To get a detailed view of the individual activity runs of a specific pipeline run, click on the pipeline name. Example:

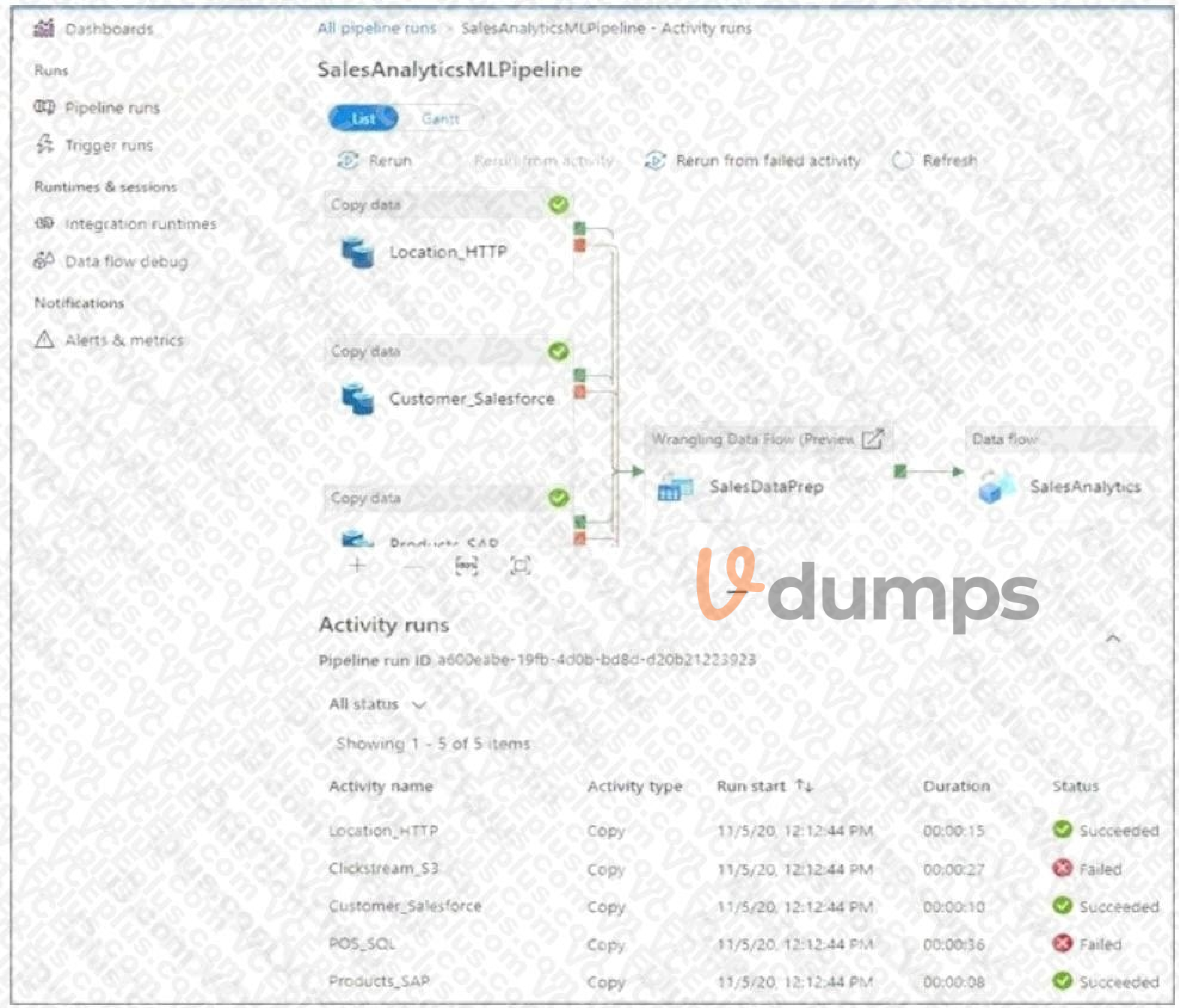


Pipeline runs

Showing 1 - 21 items

□ Pipeline name	Run start ↓	Run end
□ S3ToDataLakeCopy	11/5/20, 6:00:18 AM	11/5/20, 6:03
□ DatabricksJarPipeline	11/4/20, 6:04:11 PM	11/4/20, 6:10
□ S3ToDataLakeCopy	11/4/20, 6:00:18 PM	11/4/20, 6:03
□ S3ToDataLakeCopy	11/4/20, 6:00:19 AM	11/4/20, 6:04

The list view shows activity runs that correspond to each pipeline run. Hover over the specific activity run to get run-specific information such as the JSON input, JSON output, and detailed activity-specific monitoring experiences.



SalesAnalyticsMLPipeline

Activity runs

Pipeline run ID a600eabe-19fb-4d0b-bd8d-d20b21223923

All status

Showing 1 - 5 of 5 items

Activity name	Activity type	Run start ↑↓	Duration	Status
Location_HTTP	Copy	11/5/20, 12:12:44 PM	00:00:15	✓ Succeeded
Clickstream_S3	Copy	11/5/20, 12:12:44 PM	00:00:27	✗ Failed
Customer_Salesforce	Copy	11/5/20, 12:12:44 PM	00:00:10	✓ Succeeded
POS_SQL	Copy	11/5/20, 12:12:44 PM	00:00:36	✗ Failed
Products_SAP	Copy	11/5/20, 12:12:44 PM	00:00:08	✓ Succeeded

You can check the Duration.
Incorrect Answers:
C: sys.dm_pdw_wait_stats holds information related to the SQL Server OS state related to instances running on the different nodes.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/monitor-visually

QUESTION 31

]

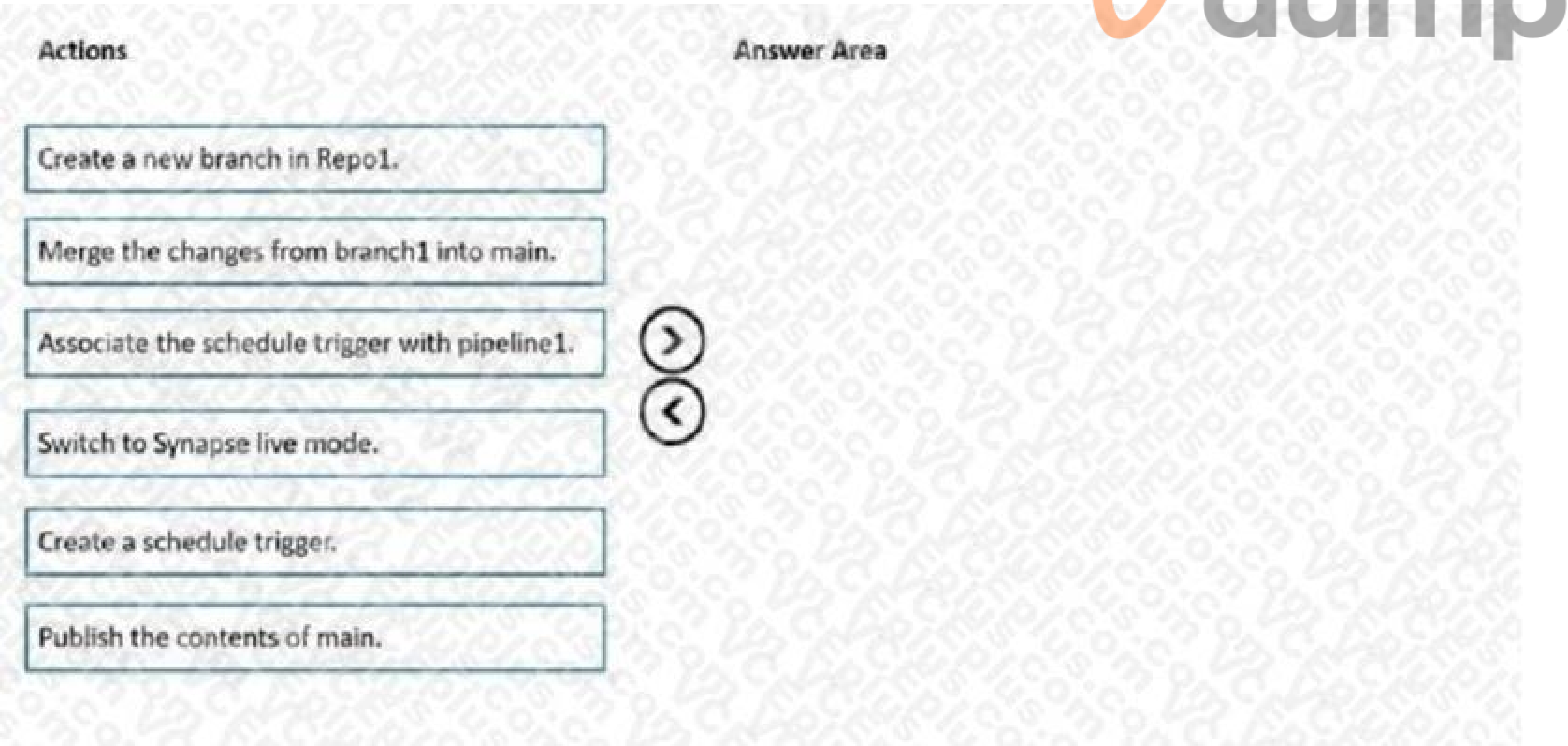
You have an Azure data factory named ADF1.
You currently publish all pipeline authoring changes directly to ADF1. You need to implement version control for the changes made to pipeline artifacts. The solution must ensure that you can apply version control to the resources currently defined in the UX 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. From the UX Authoring canvas, select Set up code repository.
B. Create a Git repository.
C. Create a GitHub action.
D. Create an Azure Data Factory trigger.
E. From the UX Authoring canvas, select Publish.
F. From the UX Authoring canvas, run Publish All.

Correct Answer: B, F
Section:
Explanation:
https://docs.microsoft.com/en-us/azure/data-factory/source-control

QUESTION 32
DRAG DROP
You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1. Branch1 contains an Azure Synapse pipeline named pipeline1. In workspace1, you complete testing of pipeline1.
You need to schedule pipeline1 to run daily at 6 AM.
Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order. NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Select and Place:



Actions

Answer Area

Create a new branch in Repo1.

Merge the changes from branch1 into main.

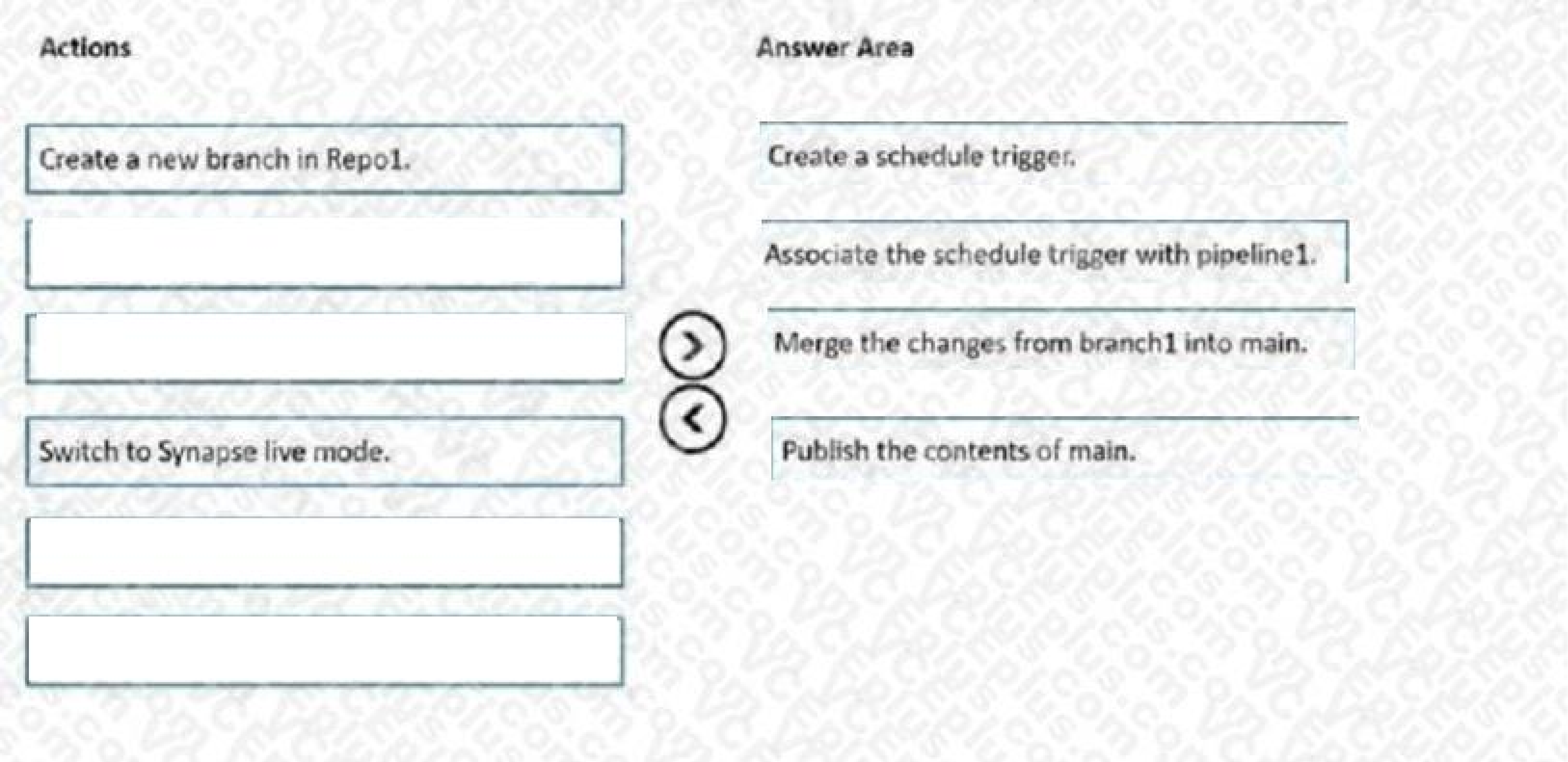
Associate the schedule trigger with pipeline1.

Switch to Synapse live mode.

Create a schedule trigger.

Publish the contents of main.

Correct Answer:



Actions

- Create a new branch in Repo1.
- Switch to Synapse live mode.

Answer Area

1. Create a schedule trigger.
2. Associate the schedule trigger with pipeline1.
3. Merge the changes from branch1 into main.
4. Publish the contents of main.

Section:
Explanation:

QUESTION 33
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

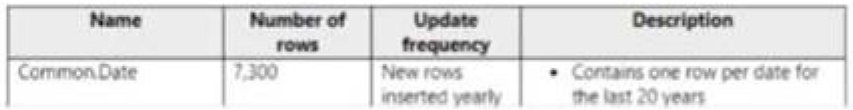
Correct Answer: D
Section:

QUESTION 34
You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a table named table1. You load 5 TB of data intotable1.
You need to ensure that columnstore compression is maximized for table1. Which statement should you execute?

A. ALTER INDEX ALL on table1 REORGANIZE
B. ALTER INDEX ALL on table1 REBUILD
C. DBCC DBREINOEX (table1)
D. DBCC INDEXDEFRAG (pool1,tablel)

Correct Answer: C
Section:
Explanation:

QUESTION 35
DRAG DROP
You have an Azure subscription.
You plan to build a data warehouse in an Azure Synapse Analytics dedicated SQL pool named pool1 that will contain staging tables and a dimensional model Pool1 will contain the following tables.



Name	Number of rows	Update frequency	Description
Common.Date	7,300	New rows inserted yearly	• Contains one row per date for the last 20 years

Select and Place:

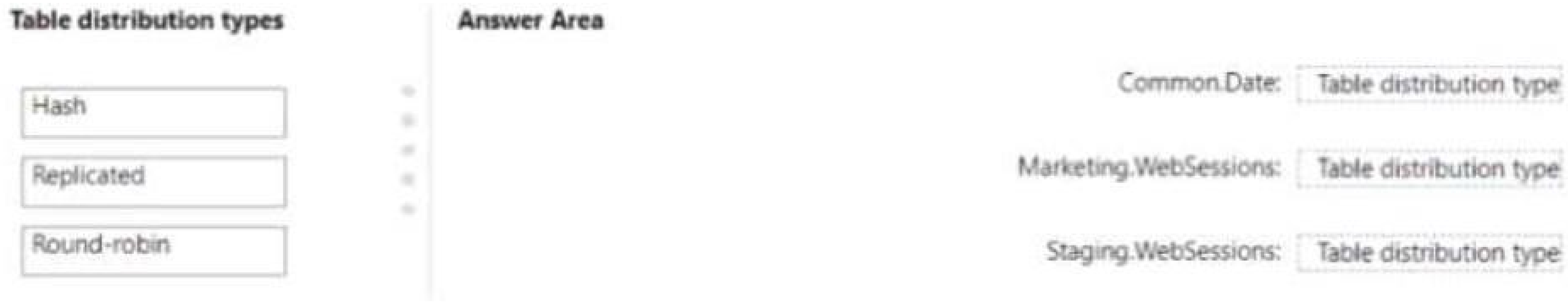


Table distribution types

Hash
Replicated
Round-robin

Answer Area

Common.Date: Table distribution type
Marketing.WebSessions: Table distribution type
Staging.WebSessions: Table distribution type

Correct Answer:

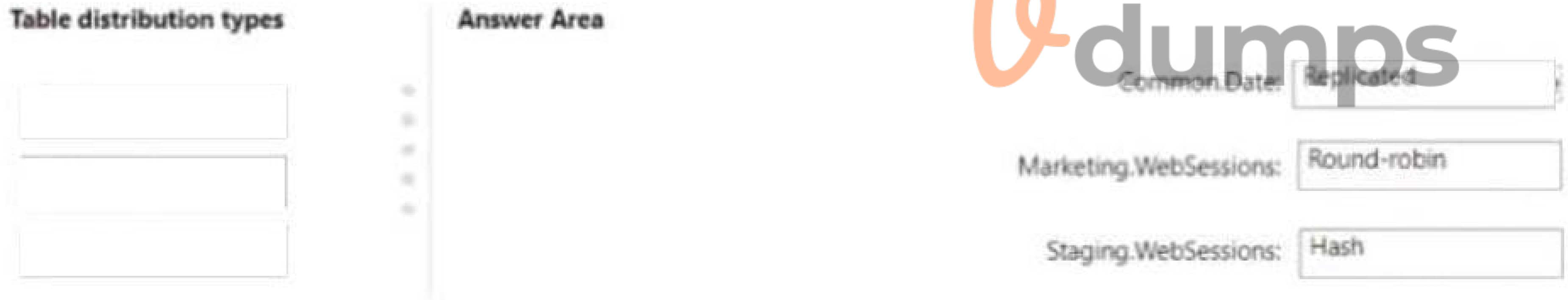


Table distribution types

Answer Area

Common.Date: Replicated
Marketing.WebSessions: Round-robin
Staging.WebSessions: Hash

Section:
Explanation:

QUESTION 36
You plan to create a dimension table in Azure Synapse Analytics that will be less than 1 GB. You need to create the table to meet the following requirements:
• Provide the fastest Query time.
• Minimize data movement during queries.
Which type of table should you use?

A. hash distributed
B. heap
C. replicated
D. round-robin

Correct Answer: C
Section:

Explanation:
A replicated table has a full copy of the table accessible on each Compute node. Replicating a table removes the need to transfer data among Compute nodes before a join or aggregation. Since the table has multiple copies, replicated tables work best when the table size is less than 2 GB compressed. 2 GB is not a hard limit.
Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/designguidance-for-replicated-tables

QUESTION 37
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

Correct Answer: B, E
Section:
Explanation:
https://docs.microsoft.com/en-us/azure/data-factory/source-control

QUESTION 38
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

Correct Answer: B, D
Section:

QUESTION 39
You have an Azure Synapse Analytics dedicated SQL pool.
You need to Create a fact table named Table1 that will store sales data from the last three years. The solution must be optimized for the following query operations:
Show order counts by week.
• Calculate sales totals by region.
• Calculate sales totals by product.
• Find all the orders from a given month.
Which data should you use to partition Table1?

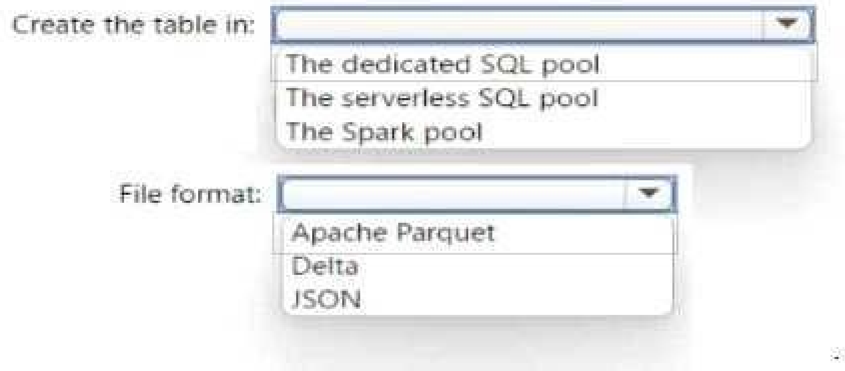
A. region

B. product
C. week
D. month

Correct Answer: C
Section:

QUESTION 40
HOTSPOT
You have an Azure Synapse Analytics serverless SQL pool, an Azure Synapse Analytics dedicated SQL pool, an Apache Spark pool, and an Azure Data Lake Storage Gen2 account. You need to create a table in a lake database. The table must be available to both the serverless SQL pool and the Spark pool. Where should you create the table, and Which file format should you use for data in the table? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

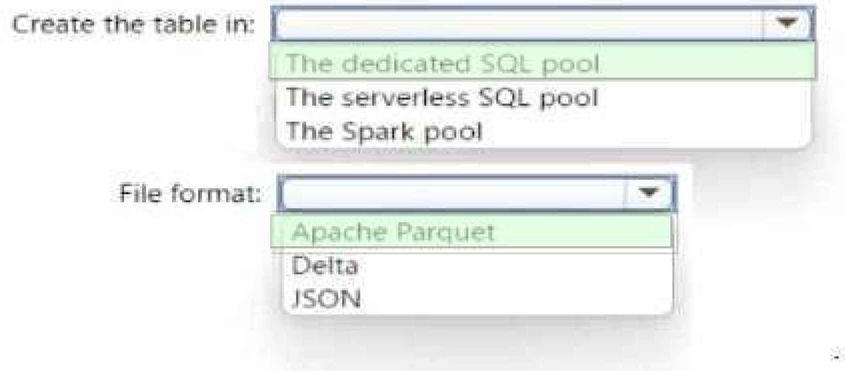
Hot Area:



Create the table in:
- The dedicated SQL pool
- The serverless SQL pool
- The Spark pool

File format:
- Apache Parquet
- Delta
- JSON

Answer Area:



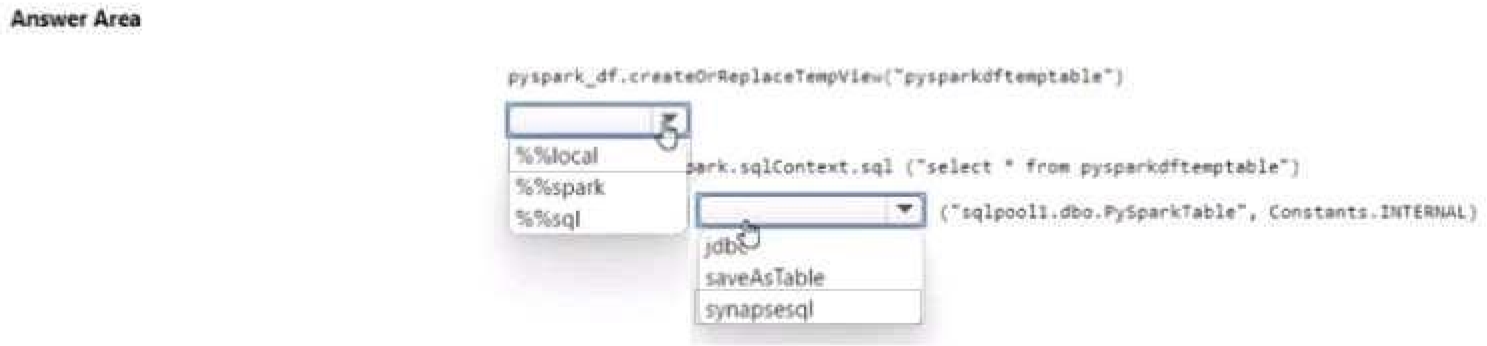
Create the table in:
- [✓ The dedicated SQL pool]
- The serverless SQL pool
- The Spark pool

File format:
- [✓ Apache Parquet]
- Delta
- JSON

Section:
Explanation:

QUESTION 41
HOTSPOT
You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 contains a dedicated SQL pool named SQL Pool and an Apache Spark pool named sparkpool. Sparkpool1 contains a DataFrame named pyspark.df.
You need to write the contents of pyspark_df to a tabte in SQLPooM by using a PySpark notebook. How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Hot Area:



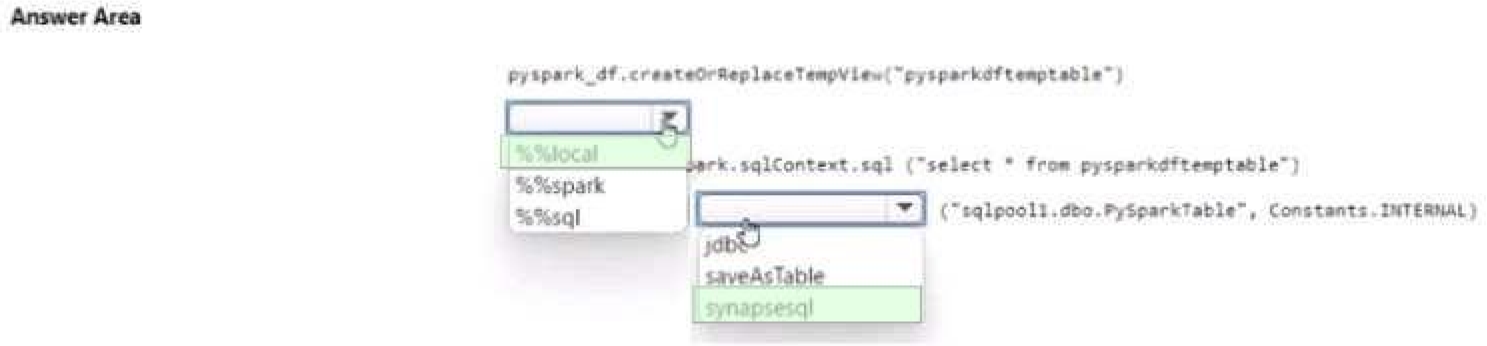
Answer Area

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")
```

```
park.sqlContext.sql ("select * from pysparkdftemptable")
```

```
("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)
```

Answer Area:



Answer Area

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")
```

```
park.sqlContext.sql ("select * from pysparkdftemptable")
```

```
("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)
```

Section:

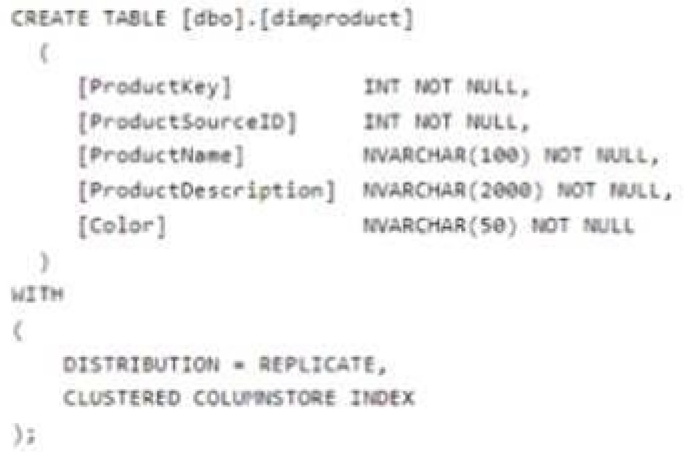
Explanation:

QUESTION 42

You are implementing a star schema in an Azure Synapse Analytics dedicated SQL pool. You plan to create a table named DimProduct.

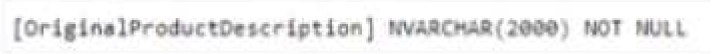
DimProduct must be a Type 3 slowly changing dimension (SCO) table that meets the following requirements:

- The values in two columns named ProductKey and ProductSourceID will remain the same.
- The values in three columns named ProductName, ProductDescription, and Color can change. You need to add additional columns to complete the following table definition.



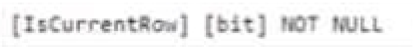
```
CREATE TABLE [dbo].[dimproduct]
  (
     [ProductKey]             INT NOT NULL,
     [ProductSourceID]        INT NOT NULL,
     [ProductName]            NVARCHAR(100) NOT NULL,
     [ProductDescription]     NVARCHAR(2000) NOT NULL,
     [Color]                  NVARCHAR(50) NOT NULL
  )
WITH
(
    DISTRIBUTION = REPLICATE,
    CLUSTERED COLUMNSTORE INDEX
);
```

A.



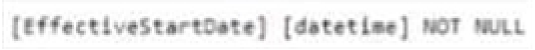
```
[OriginalProductDescription] NVARCHAR(2000) NOT NULL
```

B.



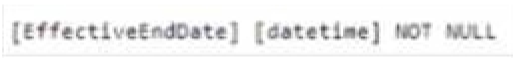
```
[IsCurrentRow] [bit] NOT NULL
```

C.



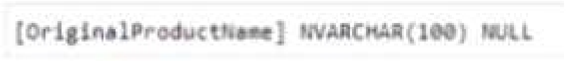
```
[EffectiveStartDate] [datetime] NOT NULL
```

D.



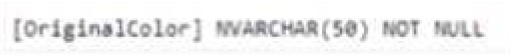
```
[EffectiveEndDate] [datetime] NOT NULL
```

E.



```
[OriginalProductName] NVARCHAR(100) NULL
```

F.



```
[OriginalColor] NVARCHAR(50) NOT NULL
```

Correct Answer: A, B, C
Section:

QUESTION 43
You have an Azure Databricks workspace that contains a Delta Lake dimension table named Tablet. Table1 is a Type 2 slowly changing dimension (SCD) table. You need to apply updates from a source table to Table1. Which Apache Spark SQL operation should you use?

A. CREATE
B. UPDATE
C. MERGE
D. ALTER

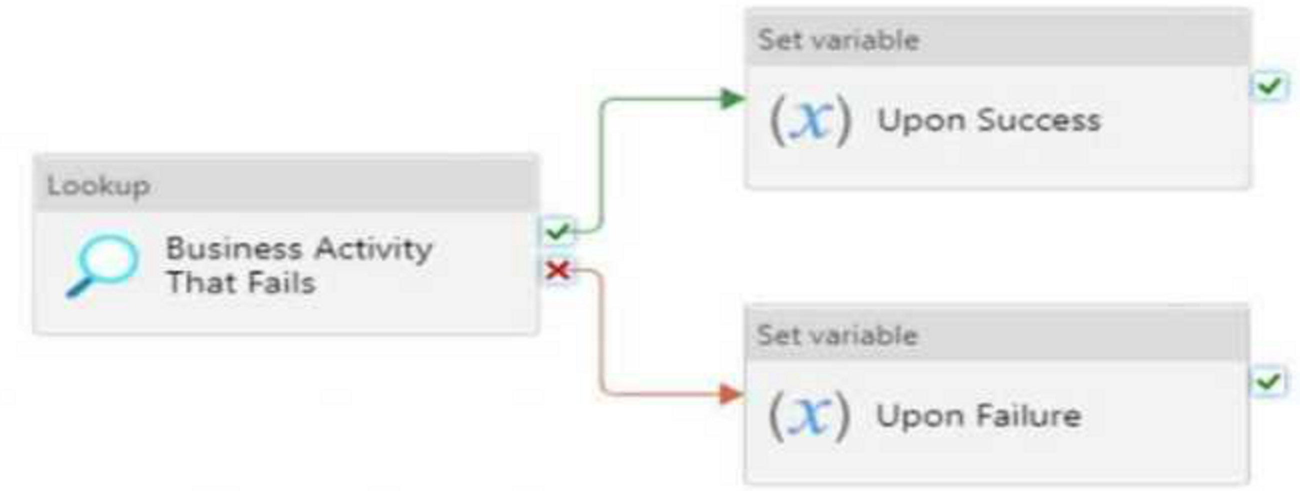
Correct Answer: C
Section:
Explanation:
The Delta provides the ability to infer the schema for data input which further reduces the effort required in managing the schema changes. The Slowly Changing Data(SCD) Type 2 records all the changes made to each key in the dimensional table. These operations require updating the existing rows to mark the previous values of the keys as old and then inserting new rows as the latest values. Also, Given a source table with the updates and the target table with dimensional data, SCD Type 2 can be expressed with the merge.
Example:

```
// Implementing SCD Type 2 operation using merge function
customersTable
.as("customers")
.merge(
stagedUpdates.as("staged_updates"),
"customers.customerId = mergeKey")
.whenMatched("customers.current = true AND customers.address <> staged_updates.address") .updateExpr(Map(
"current" -> "false",
"endDate" -> "staged_updates.effectiveDate"))
.whenNotMatched()
.insertExpr(Map(
"customerid" -> "staged_updates.customerId",
"address" -> "staged_updates.address",
"current" -> "true",
"effectiveDate" -> "staged_updates.effectiveDate",
"endDate" -> "null"))
.execute()
}
```

Reference:
https://www.projectpro.io/recipes/what-is-slowly-changing-data-scd-type-2-operation-delta-tabledatabricks

QUESTION 44

You have the Azure Synapse Analytics pipeline shown in the following exhibit.



You need to add a set variable activity to the pipeline to ensure that after the pipeline's completion, the status of the pipeline is always successful.
What should you configure for the set variable activity?

A. a success dependency on the Business Activity That Fails activity

B. a failure dependency on the Upon Failure activity

C. a skipped dependency on the Upon Success activity

D. a skipped dependency on the Upon Failure activity

Correct Answer: B
Section:
Explanation:
A failure dependency means that the activity will run only if the previous activity fails. In this case, setting a failure dependency on the Upon Failure activity will ensure that the set variable activity will run after the pipeline fails and set the status of the pipeline to successful.
https://www.validexamdumps.com

QUESTION 45
You are building a data flow in Azure Data Factory that upserts data into a table in an Azure Synapse Analytics dedicated SQL pool.You need to add a transformation to the data flow. The transformation must specify logic indicating when a row from the input data must be upserted into the sink.
Which type of transformation should you add to the data flow?

A. join

B. select

C. surrogate key

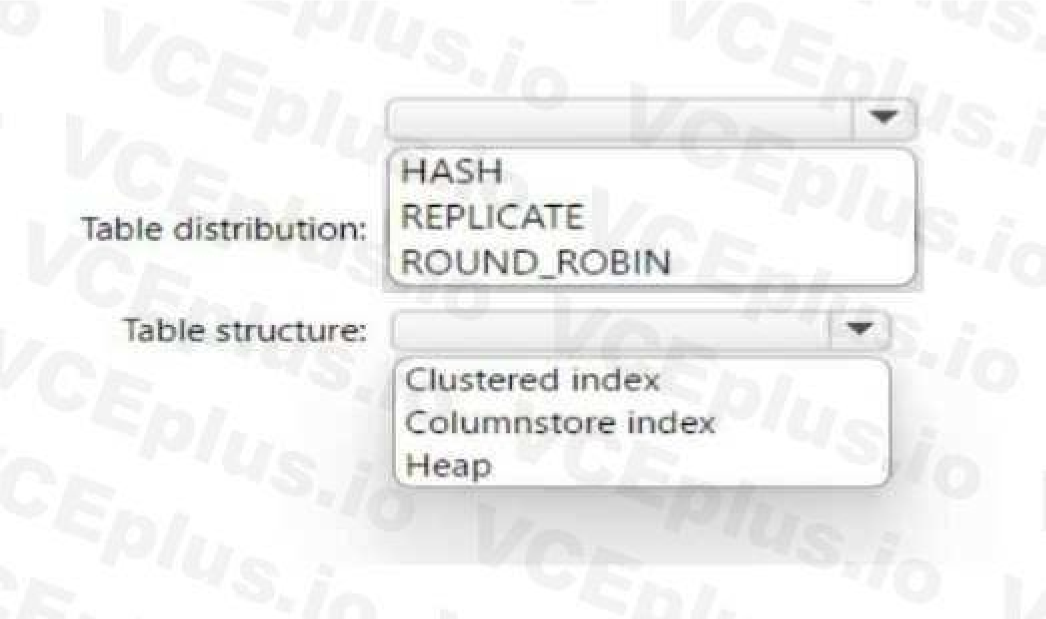
D. alter row

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

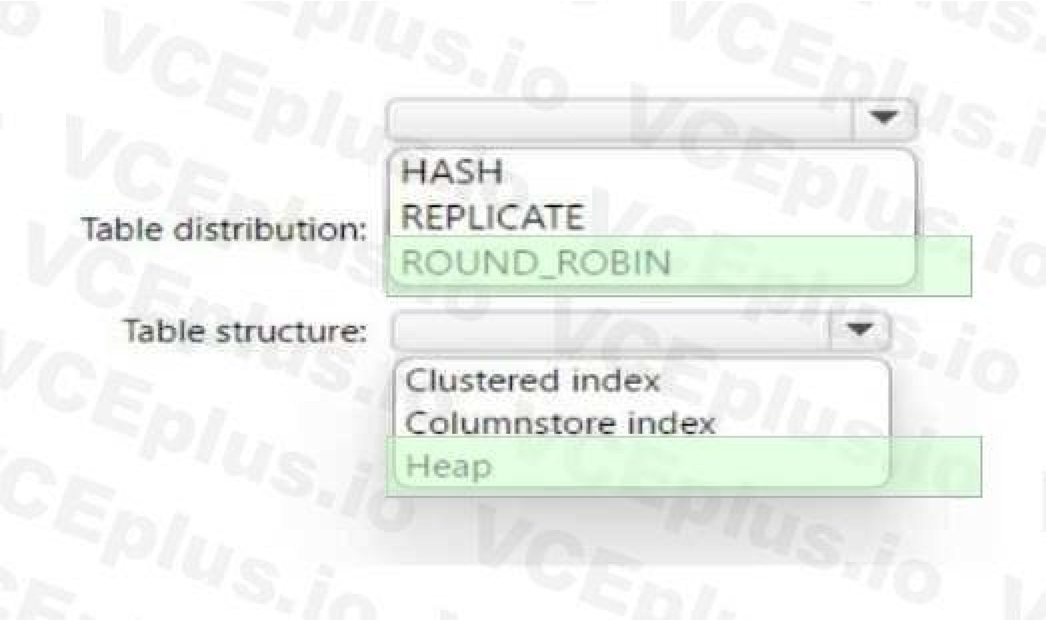
QUESTION 46
HOTSPOT
You are incrementally loading data into fact tables in an Azure Synapse Analytics dedicated SQL pool.
Each batch of incoming data is staged before being loaded into the fact tables.

You need to ensure that the incoming data is staged as quickly as possible.
How should you configure the staging tables? To answer, select the appropriate options in theanswer area.

Hot Area:



Answer Area:



Section:

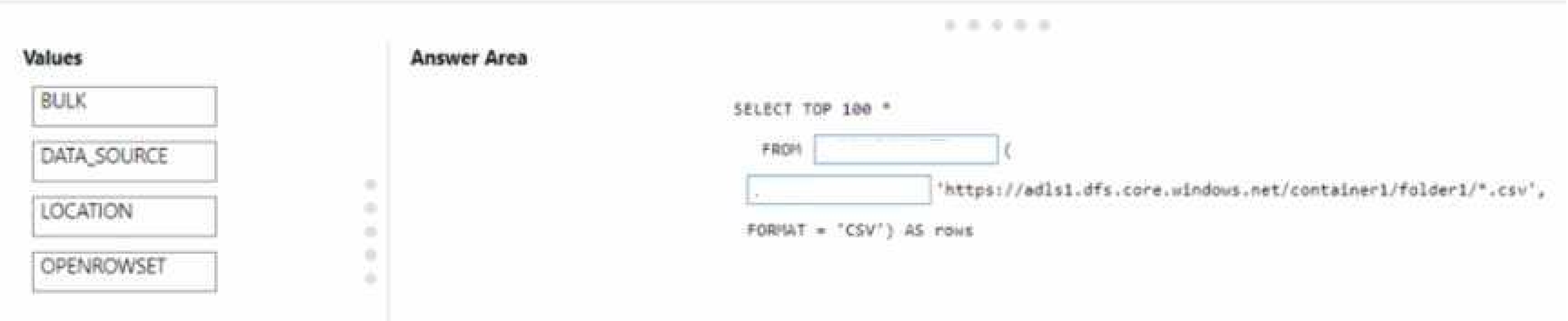
Explanation:

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

QUESTION 47

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

Select and Place:



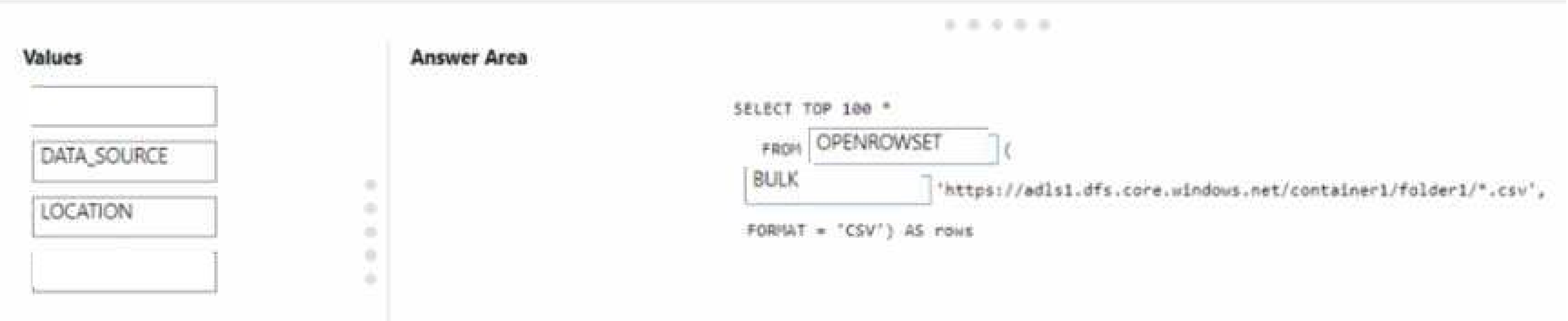
Values

- BULK
- DATA_SOURCE
- LOCATION
- OPENROWSET

Answer Area

```
SELECT TOP 100 *
  FROM [          ] (
[          ] 'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',
FORMAT = 'CSV') AS rows
```

Correct Answer:



Values

- DATA_SOURCE
- LOCATION

Answer Area

```
SELECT TOP 100 *
  FROM OPENROWSET (
BULK 'https://adls1.dfs.core.windows.net/container1/folder1/*.csv',
FORMAT = 'CSV') AS rows
```

Section:

Explanation:

QUESTION 48

HOTSPOT

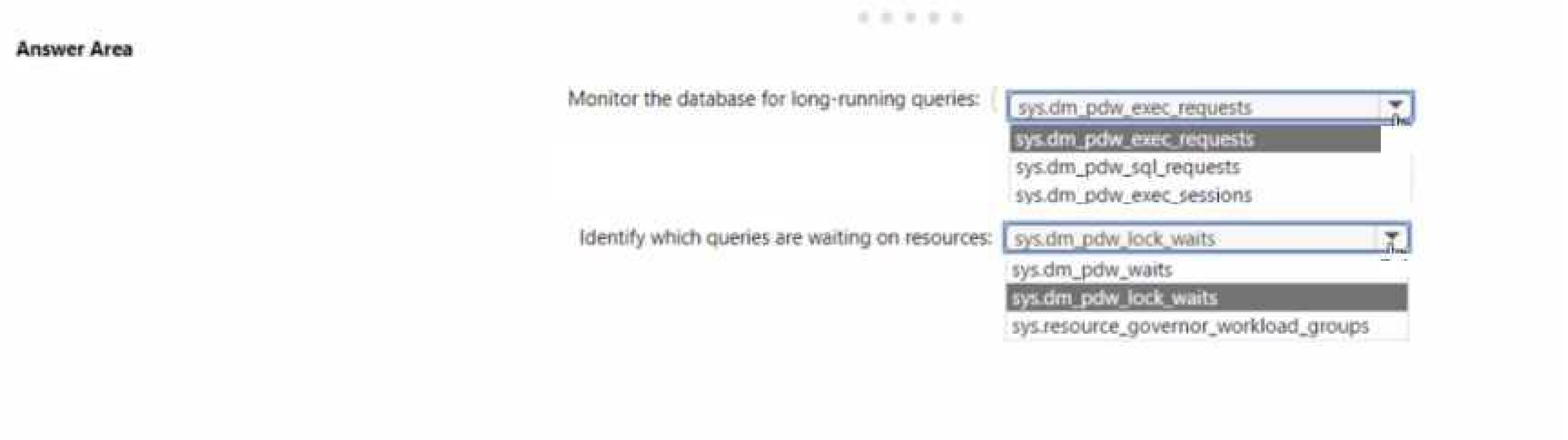
You have an Azure Synapse Analytics dedicated SQL pool.

You need to monitor the database for long-running queries and identify which queries are waiting on resources

Which dynamic management view should you use for each requirement? To answer, select the appropriate options in the answer area.

NOTE; Each correct answer is worth one point.

Hot Area:



Answer Area

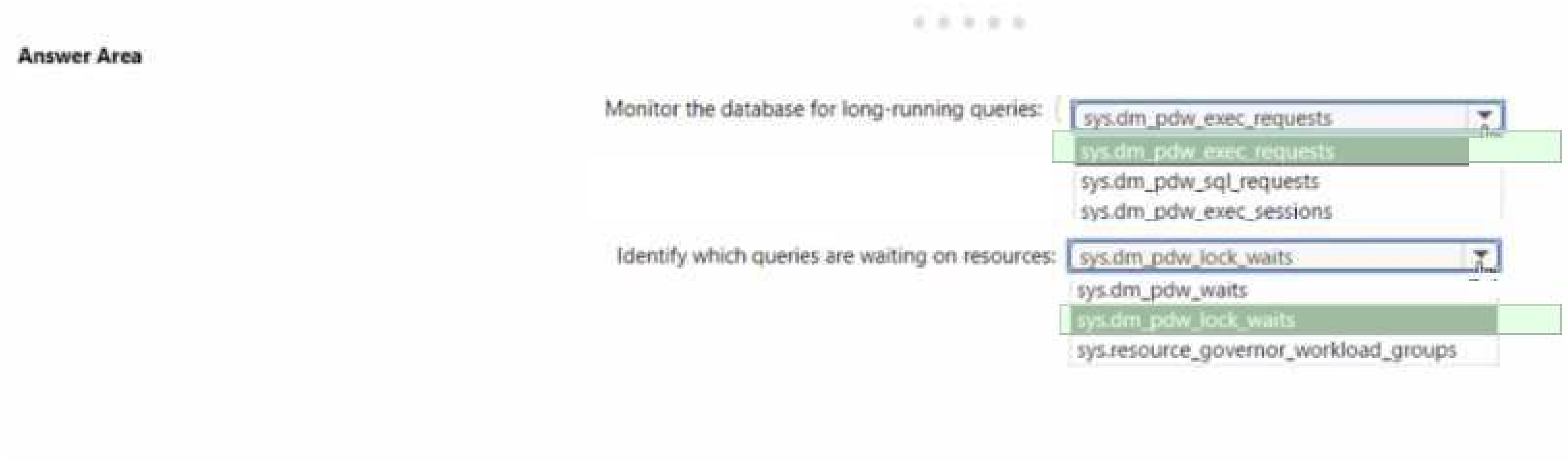
Monitor the database for long-running queries: sys.dm_pdw_exec_requests

- sys.dm_pdw_exec_requests
- sys.dm_pdw_sql_requests
- sys.dm_pdw_exec_sessions

Identify which queries are waiting on resources: sys.dm_pdw_lock_waits

- sys.dm_pdw_waits
- sys.dm_pdw_lock_waits
- sys.resource_governor_workload_groups

Answer Area:



Answer Area

Monitor the database for long-running queries: sys.dm_pdw_exec_requests

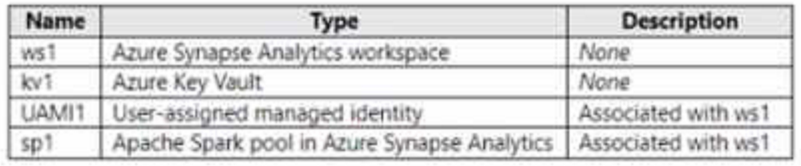
- sys.dm_pdw_exec_requests
- sys.dm_pdw_sql_requests
- sys.dm_pdw_exec_sessions

Identify which queries are waiting on resources: sys.dm_pdw_lock_waits

- sys.dm_pdw_waits
- sys.dm_pdw_lock_waits
- sys.resource_governor_workload_groups

Section:
Explanation:

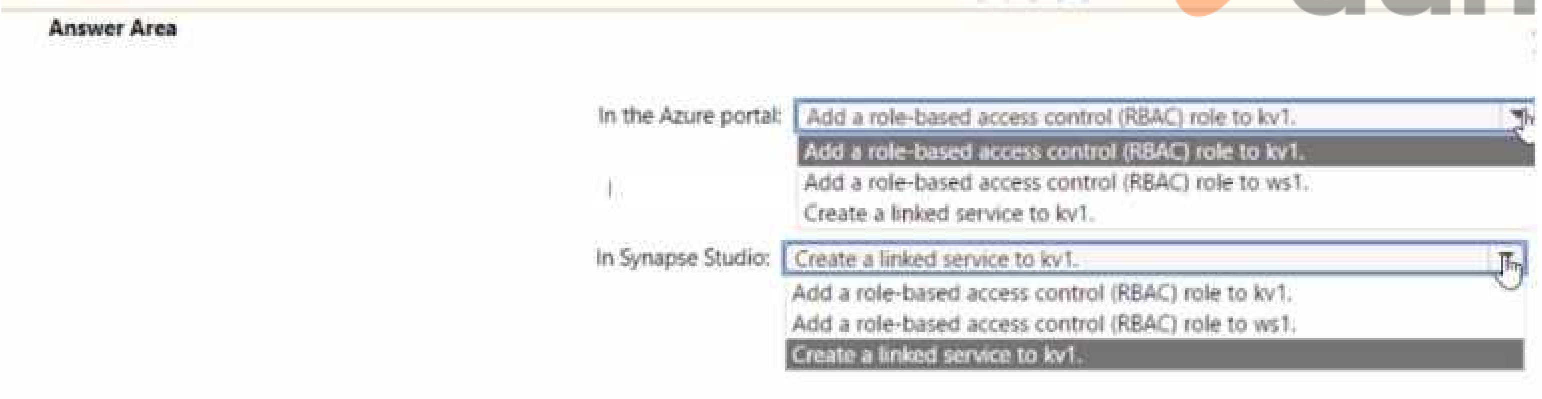
QUESTION 49
HOTSPOT
You have an Azure subscription that contains the resources shown in the following table.



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

You need to ensure that you can Spark notebooks in ws1. The solution must ensure secrets from kv1 by using UAMI1. What should you do? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

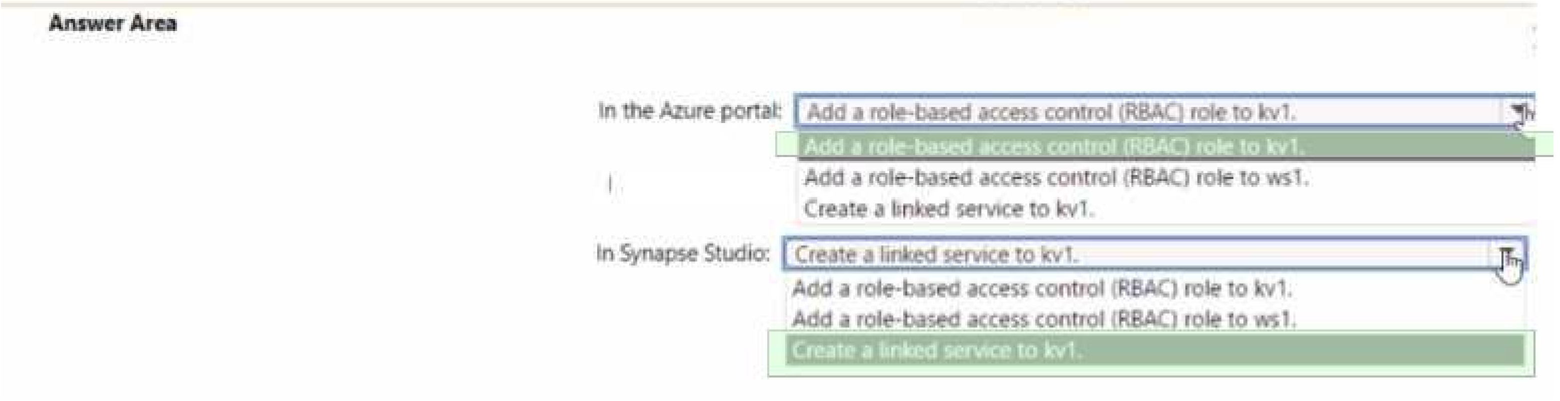
In the Azure portal: Add a role-based access control (RBAC) role to kv1.

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.
- Create a linked service to kv1.

In Synapse Studio: Create a linked service to kv1.

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.
- Create a linked service to kv1.

Answer Area:



Answer Area

In the Azure portal: Add a role-based access control (RBAC) role to kv1.

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.
- Create a linked service to kv1.

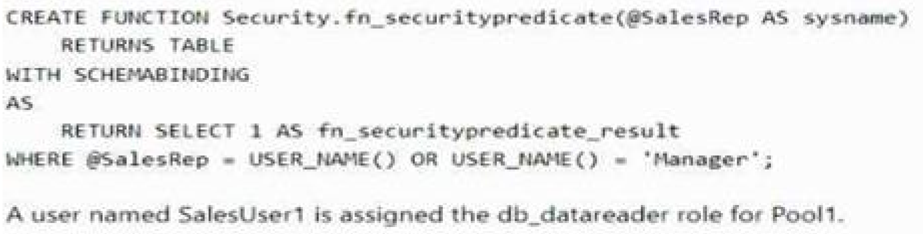
In Synapse Studio: Create a linked service to kv1.

- Add a role-based access control (RBAC) role to kv1.
- Add a role-based access control (RBAC) role to ws1.
- Create a linked service to kv1.

Section:

Explanation:

QUESTION 50
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains a table named Sales. Sales has row-level security (RLS) applied. RLS uses the following predicate filter.



```
CREATE FUNCTION Security.fn_securitypredicate(@SalesRep AS sysname)
    RETURNS TABLE
WITH SCHEMABINDING
AS
    RETURN SELECT 1 AS fn_securitypredicate_result
WHERE @SalesRep = USER_NAME() OR USER_NAME() = 'Manager';
```

A user named SalesUser1 is assigned the db_datareader role for Pool1.

A user named SalesUser1 is assigned the db_datareader role for Pool1. Which rows in the Sales table are returned when SalesUser1 queries the table?

A. only the rows for which the value in the User_Name column is SalesUser1
B. all the rows
C. only the rows for which the value in the SalesRep column is Manager
D. only the rows for which the value in the SalesRep column is SalesUser1

Correct Answer: A
Section:

QUESTION 51
You are designing a dimension table in an Azure Synapse Analytics dedicated SQL pool. You need to create a surrogate key for the table. The solution must provide the fastest query performance. What should you use for the surrogate key?

A. a GUID column
B. a sequence object
C. an IDENTITY column

Correct Answer: C
Section:
Explanation:
Use IDENTITY to create surrogate keys using dedicated SQL pool in AzureSynapse Analytics. Note: 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-datawarehouse-tables-identity

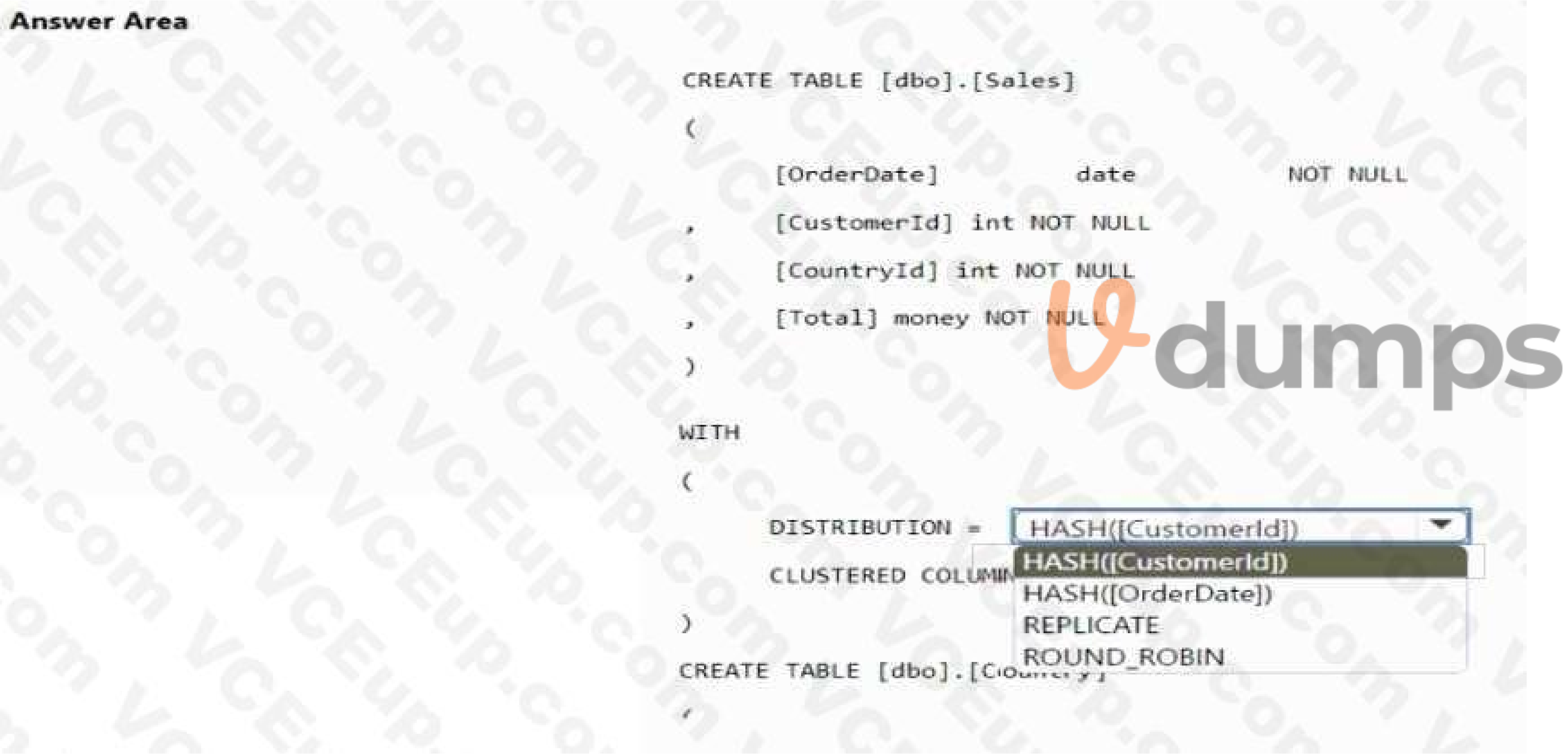
QUESTION 52
You are designing an Azure Data Lake Storage solution that will transform raw JSON files for use in an analytical workload. You need to recommend a format for the transformed files. The solution must meet the following requirements:
Contain information about the data types of each column 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

Correct Answer: D
Section:
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

QUESTION 53
HOTSPOT
You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool. You plan to deploy a solution that will analyze sales data and include the following:
- A table named Country that will contain 195 rows
- A table named Sales that will contain 100 million rows
- A query to identify total sales by country and customer from the past 30 days You need to create the tables. The solution must maximize query performance. How should you complete the script? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

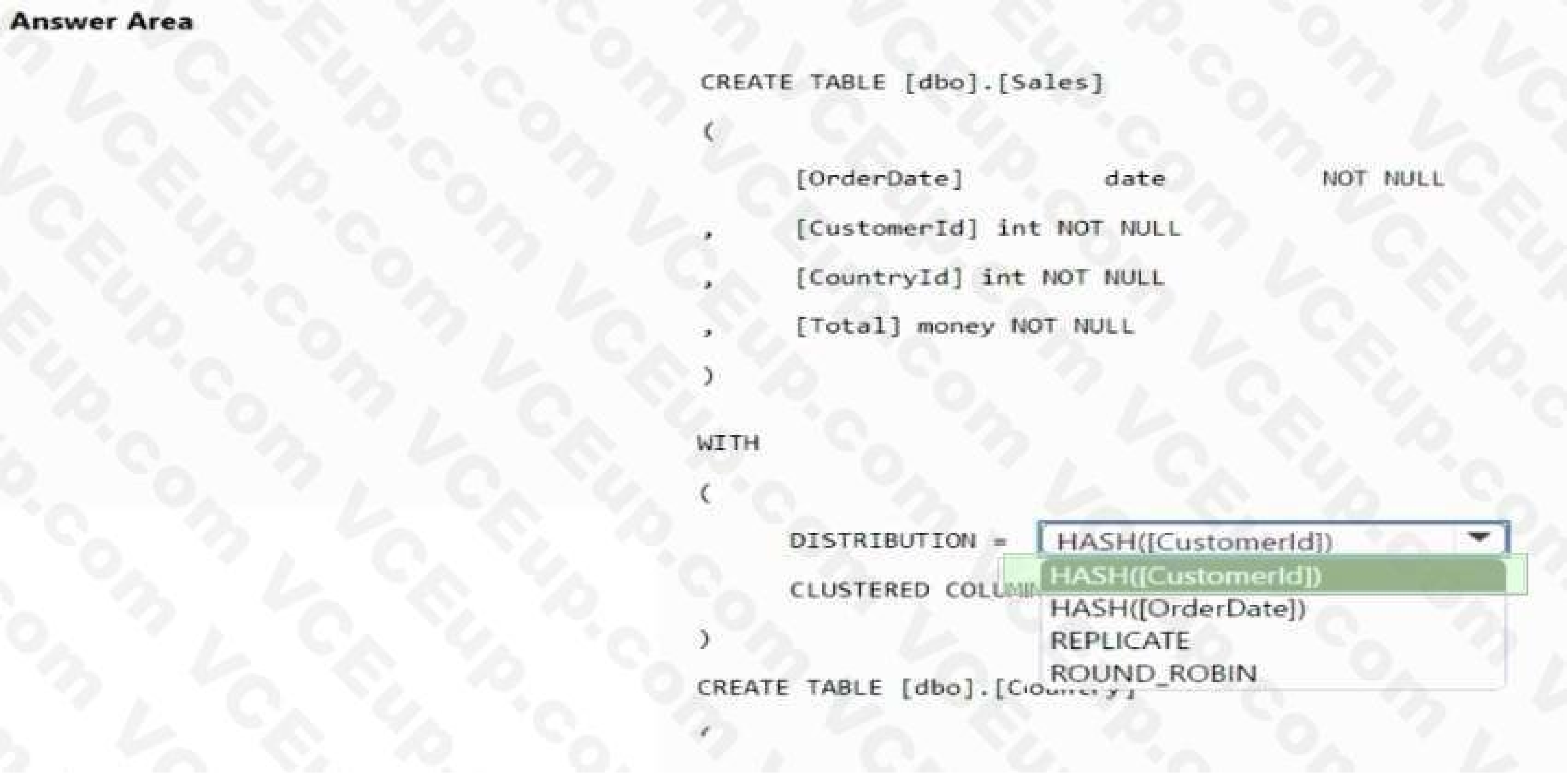
Hot Area:



Answer Area

```
CREATE TABLE [dbo].[Sales]
(
    [OrderDate]         date          NOT NULL
,   [CustomerId] int NOT NULL
,   [CountryId] int NOT NULL
,   [Total] money NOT NULL
)

WITH
(
    DISTRIBUTION = HASH([CustomerId])
    CLUSTERED COLUMN
)
CREATE TABLE [dbo].[Country]
(
```

Answer Area:

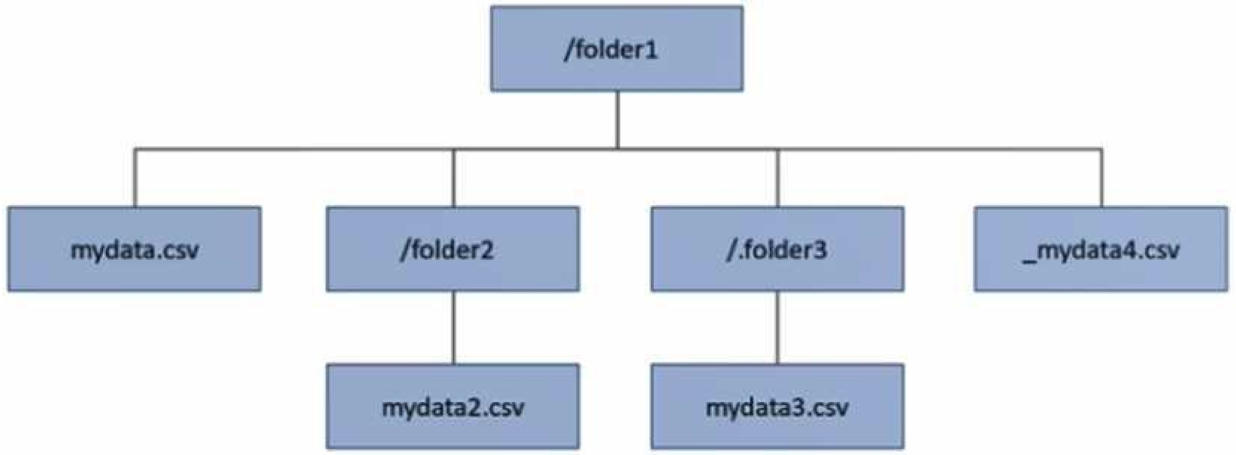


Answer Area

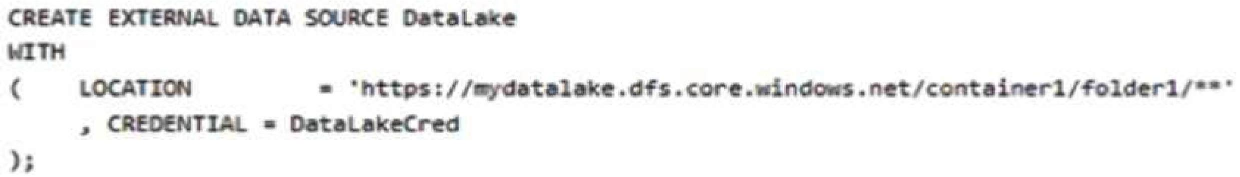
```
CREATE TABLE [dbo].[Sales]
(
        [OrderDate]          date           NOT NULL
,       [CustomerId] int NOT NULL
,       [CountryId] int NOT NULL
,       [Total] money NOT NULL
)

WITH
(
       DISTRIBUTION =  HASH([CustomerId])
       CLUSTERED COLUM
)
CREATE TABLE [dbo].[Co
(
```

Section:
Explanation:

QUESTION 54
HOTSPOT
You have an Azure Data Lake Storage Gen2 account that contains a container named container1. You have an Azure Synapse Analytics serverless SQL pool that contains a native external table named dbo.Table1. The source data for dbo.Table1 is stored in container1. The folder structure of container1 is shown in the following exhibit.



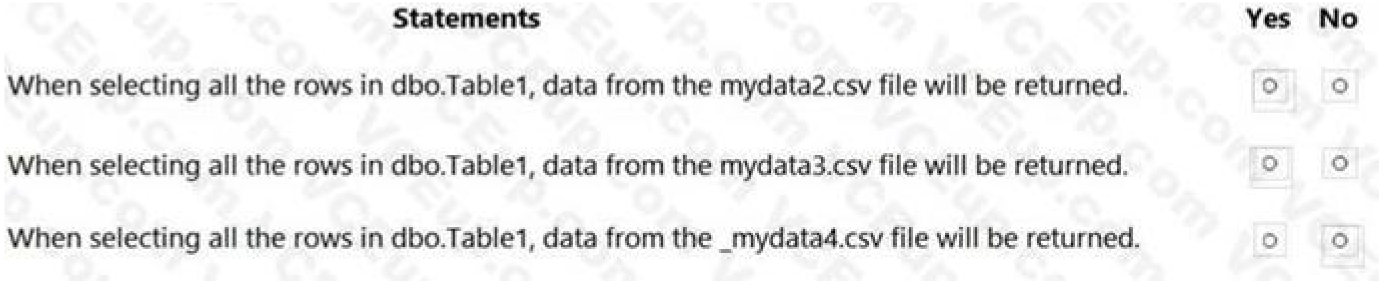
The external data source is defined by using the following statement.



```
CREATE EXTERNAL DATA SOURCE DataLake
WITH
(    LOCATION         = 'https://mydatalake.dfs.core.windows.net/container1/folder1/**'
   , CREDENTIAL = DataLakeCred
);
```

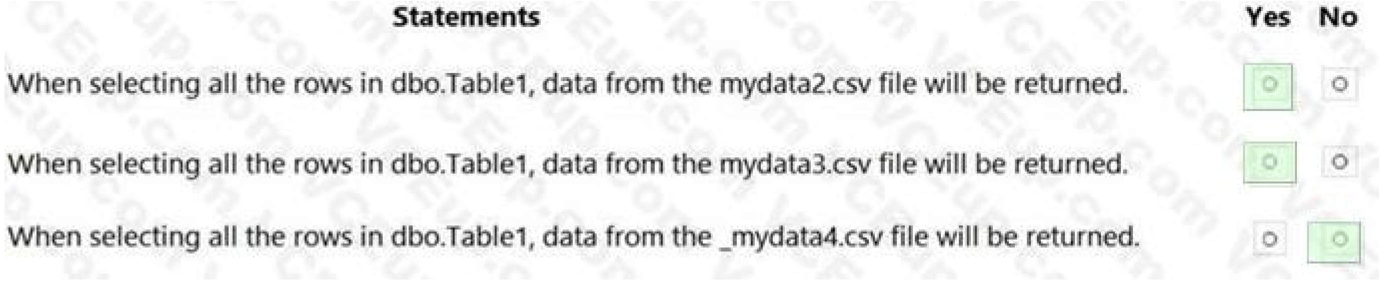
For each of the following statements, select Yes if the statement is true. Otherwise, select No. NOTE: Each correct selection is worth one point.

Hot Area:



Statements	Yes	No
When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.	○	○
When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	○	○
When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	○	○

Answer Area:



Statements	Yes	No
When selecting all the rows in dbo.Table1, data from the mydata2.csv file will be returned.	●	○
When selecting all the rows in dbo.Table1, data from the mydata3.csv file will be returned.	●	○
When selecting all the rows in dbo.Table1, data from the _mydata4.csv file will be returned.	○	●

Section:

Explanation:

Box 1: Yes

In the serverless SQL pool you can also use recursive wildcards /logs/** to reference Parquet or CSV files in any sub-folder beneath the referenced folder.

Box 2: Yes

Box 3: No

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

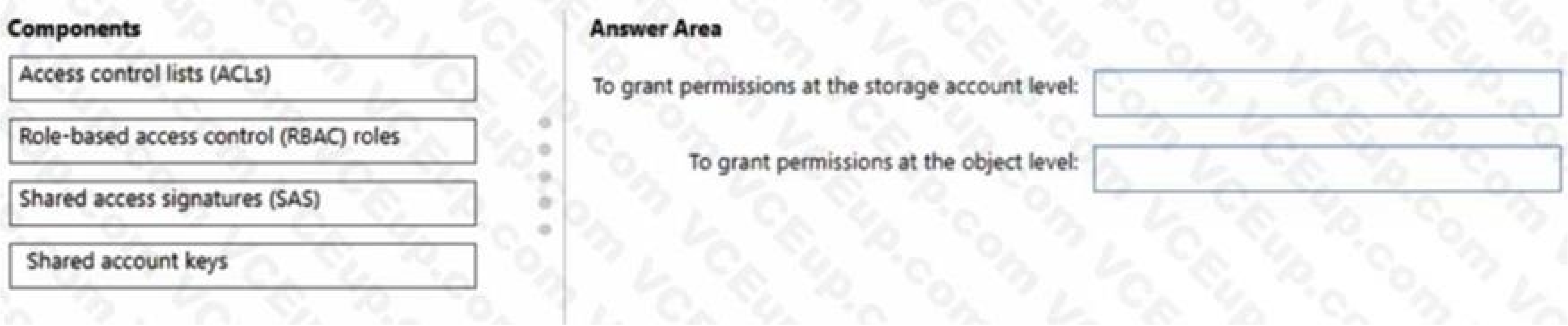
QUESTION 55

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:

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.

Select and Place:



Components

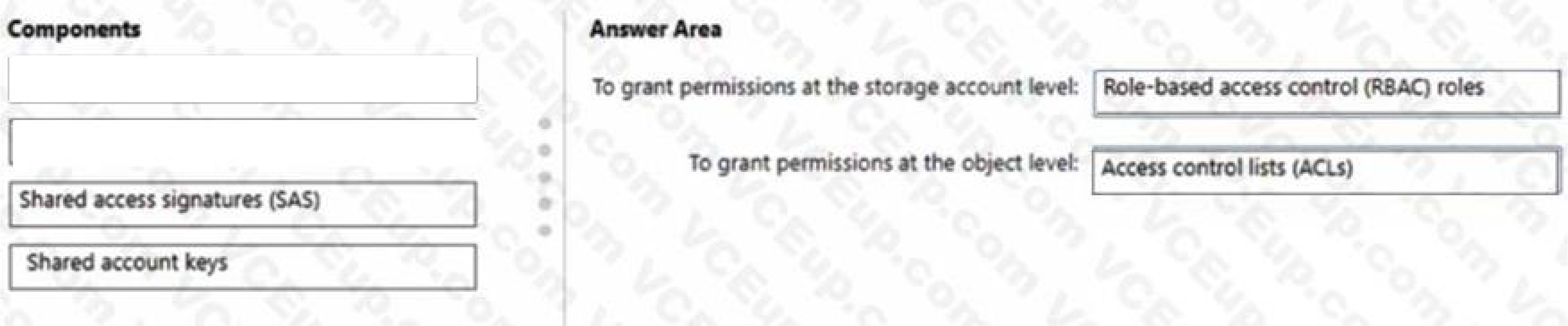
- Access control lists (ACLs)
- Role-based access control (RBAC) roles
- Shared access signatures (SAS)
- Shared account keys

Answer Area

To grant permissions at the storage account level:

To grant permissions at the object level:

Correct Answer:



Components

Shared access signatures (SAS)

Shared account keys

Answer Area

To grant permissions at the storage account level: Role-based access control (RBAC) roles

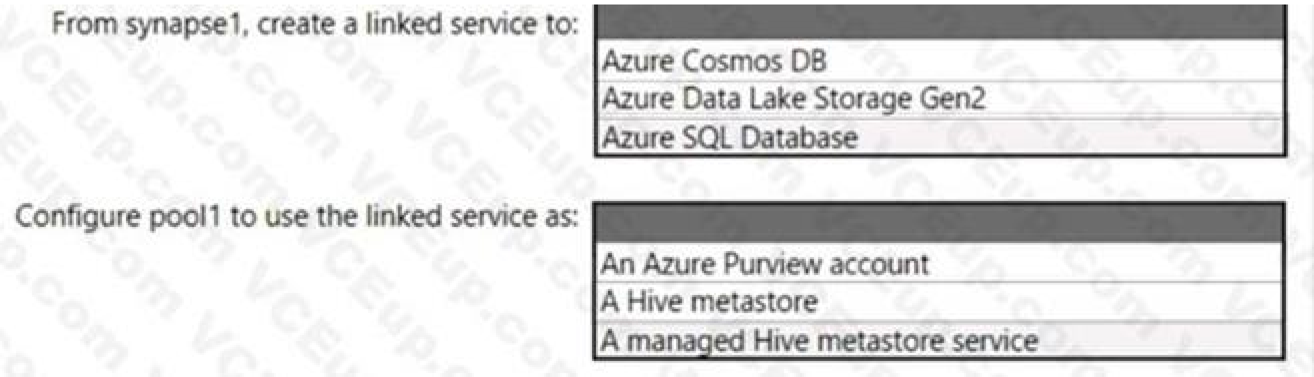
To grant permissions at the object level: Access control lists (ACLs)

Section:
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-controlmodel

QUESTION 56

HOTSPOT
You have an Azure subscription that contains an Azure Databricks workspace named databricks1 and an Azure Synapse Analytics workspace named synapse1. The synapse1 workspace contains an Apache Spark pool named pool1.
You need to share an Apache Hive catalog of pool1 with databricks1. What should you do? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

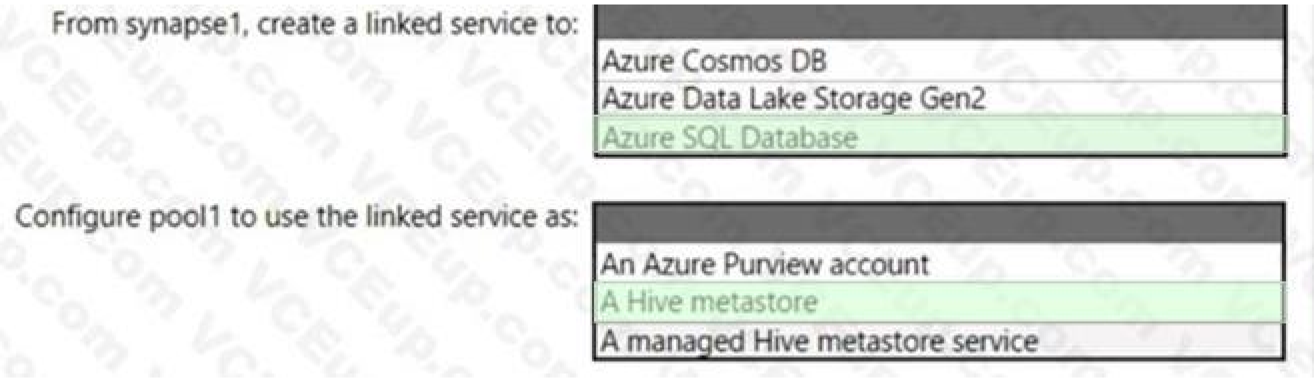
Hot Area:



From synapse1, create a linked service to:
- Azure Cosmos DB
- Azure Data Lake Storage Gen2
- Azure SQL Database

Configure pool1 to use the linked service as:
- An Azure Purview account
- A Hive metastore
- A managed Hive metastore service

Answer Area:



From synapse1, create a linked service to:
- Azure Cosmos DB
- Azure Data Lake Storage Gen2
- [✓ Azure SQL Database]

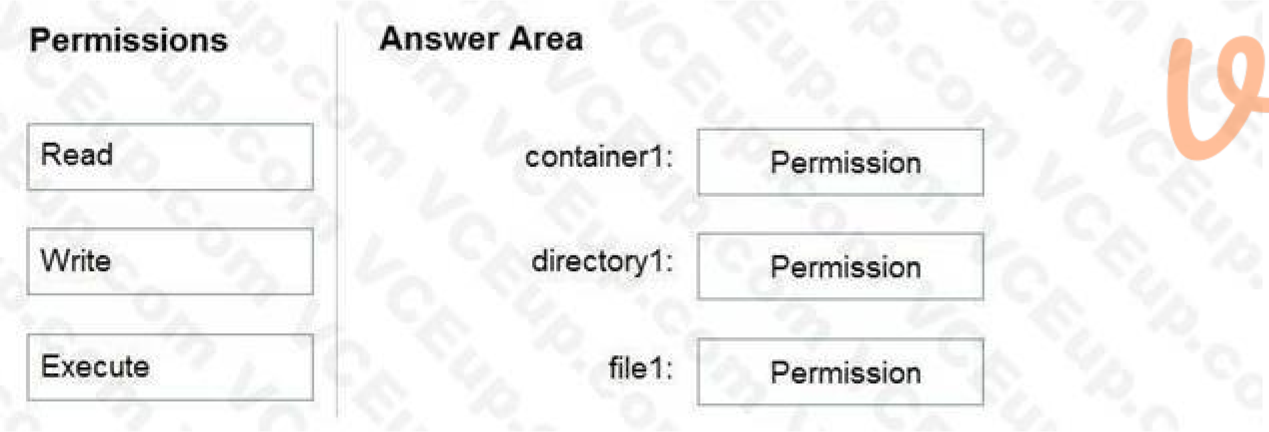
Configure pool1 to use the linked service as:
- An Azure Purview account
- [✓ A Hive metastore]
- A managed Hive metastore service

Section:

Explanation:
Box 1: Azure SQL Database
Use external Hive Metastore for Synapse Spark Pool
Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog.
Set up linked service to Hive Metastore
Follow below steps to set up a linked service to the external Hive Metastore in Synapse workspace. Open Synapse Studio, go to Manage > Linked services at left, click New to create a new linked service. Set up Hive Metastore linked service
Choose Azure SQL Database or Azure Database for MySQL based on your database type, click Continue. Provide Name of the linked service. Record the name of the linked service, this info will be used to configure Spark shortly. You can either select Azure SQL Database/Azure Database for MySQL for the external Hive Metastore from Azure subscription list, or enter the info manually. Provide User name and Password to set up the connection. Test connection to verify the username and password.
Click Create to create the linked service.
Box 2: A Hive Metastore
Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/spark/apache-spark-externalmetastore

QUESTION 57
DRAG DROP
You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named storage1. Storage1 contains a container named container1. Container1 contains a directory named directory1. Directory1 contains a file named file1.
You have an Azure Active Directory (Azure AD) user named User1 that is assigned the Storage Blob Data Reader role for storage1. You need to ensure that User1 can append data to file1. The solution must use the principle of least privilege. Which permissions should you grant? To answer, drag the appropriate permissions to the correct resources. Each permission 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.

Select and Place:



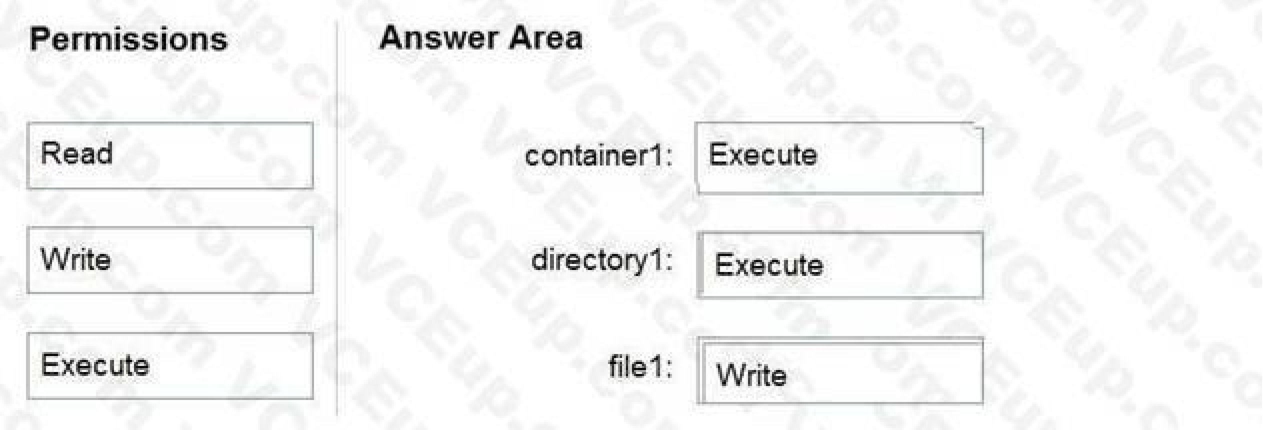
Permissions

Read
Write
Execute

Answer Area

container1: Permission
directory1: Permission
file1: Permission

Correct Answer:



Permissions

Read
Write
Execute

Answer Area

container1: Execute
directory1: Execute
file1: Write

Section:
Explanation:
Box 1: Execute
If you are granting permissions by using only ACLs (no Azure RBAC), then to grant a security principal read or write access to a file, you'll need to give the security principal Execute permissions to the root folder of the

container, and to each folder in the hierarchy of folders that lead to the file. Box 2: Execute
On Directory: Execute (X): Required to traverse the child items of a directory Box 3: Write
On file: Write (W): Can write or append to a file.
Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control

QUESTION 58
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. Table1 contains the following:
One billion rows
A clustered columnstore index
A hash-distributed column named Product Key
A column named Sales Date that is of the date data type and cannot be null Thirty million rows will be added to Table1 each month. You need to partition Table1 based on the Sales Date column. The solution must optimize query performance and data loading. How often should you create a partition?

A. once per month
B. once per year
C. once per day
D. once per week

Correct Answer: B
Section:
Explanation:
Need a minimum 1 million rows per distribution. Each table is 60 distributions. 30 millions rows is added each month. Need 2 months to get a minimum of 1 million rows per distribution in a new partition. Note: When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributions. 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.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehousetables-partition

QUESTION 59
You are creating an Apache Spark job in Azure Databricks that will ingest JSON-formatted data. You need to convert a nested JSON string into a DataFrame that will contain multiple rows. Which Spark SQL function should you use?

A. explode
B. filter
C. coalesce
D. extract

Correct Answer: A
Section:
Explanation:
Convert nested JSON to a flattened DataFrame
You can to flatten nested JSON, using only $"column.*" and explode methods. Note: Extract and flatten
Use $"column.*" and explode methods to flatten the struct and array types before displaying the flattened DataFrame. Scala

```
display(DF.select($"id" as "main_id",$"name",$"batters",$"ppu",explode($"topping")) // Exploding the topping column using explode as it is an array type
.withColumn("topping_id",$"col.id") // Extracting topping_id from col using DOT form .withColumn("topping_type",$"col.type") // Extracting topping_tytpe from col using DOT form .drop($"col")
.select($"*",$"batters.*") // Flattened the struct type batters tto array type which is batter .drop($"batters")
.select($"*",explode($"batter"))
.drop($"batter")
.withColumn("batter_id",$"col.id") // Extracting batter_id from col using DOT form .withColumn("battter_type",$"col.type") // Extracting battter_type from col using DOT form .drop($"col")
)
```

Reference: https://learn.microsoft.com/en-us/azure/databricks/kb/scala/flatten-nested-columnsdynamically

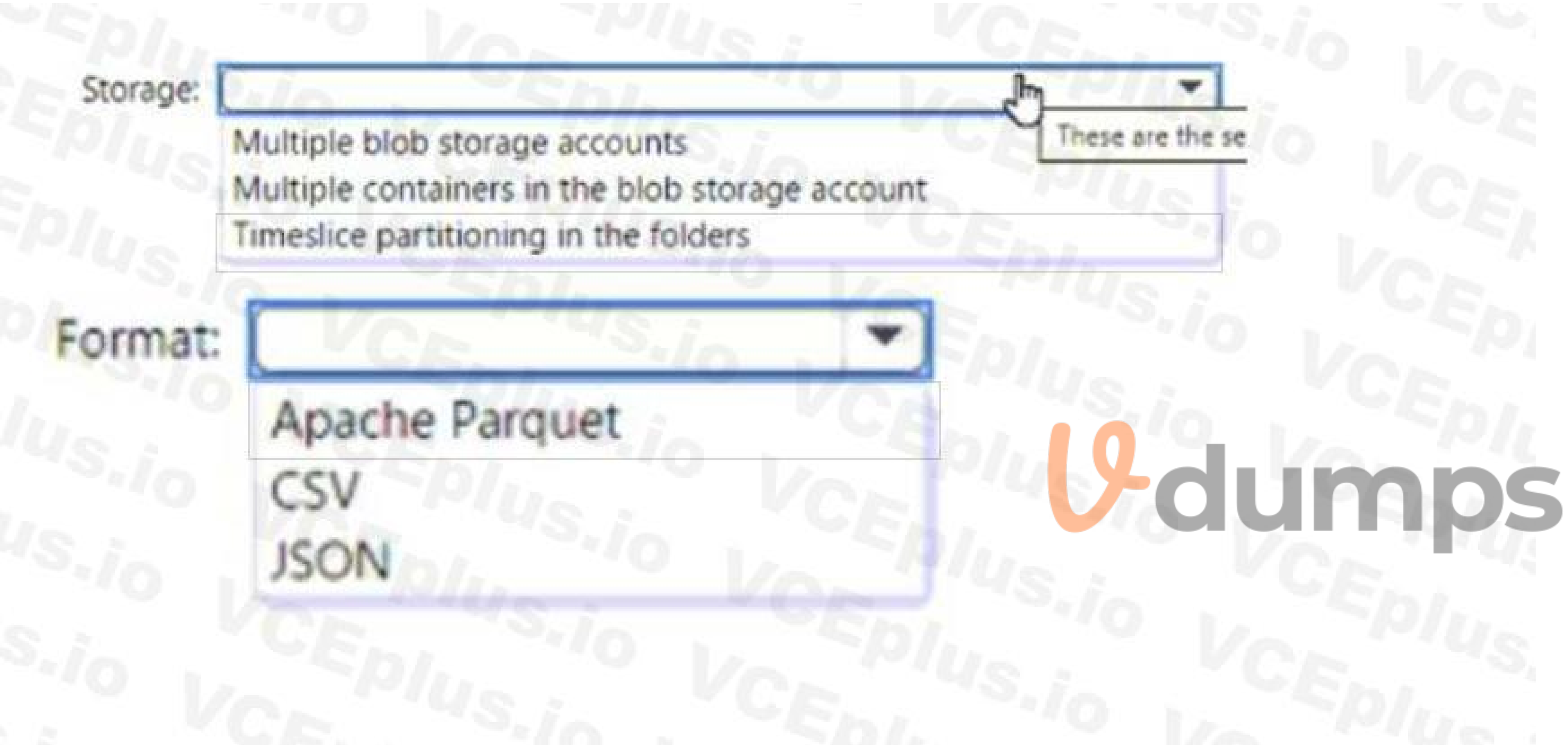
QUESTION 60

HOTSPOT

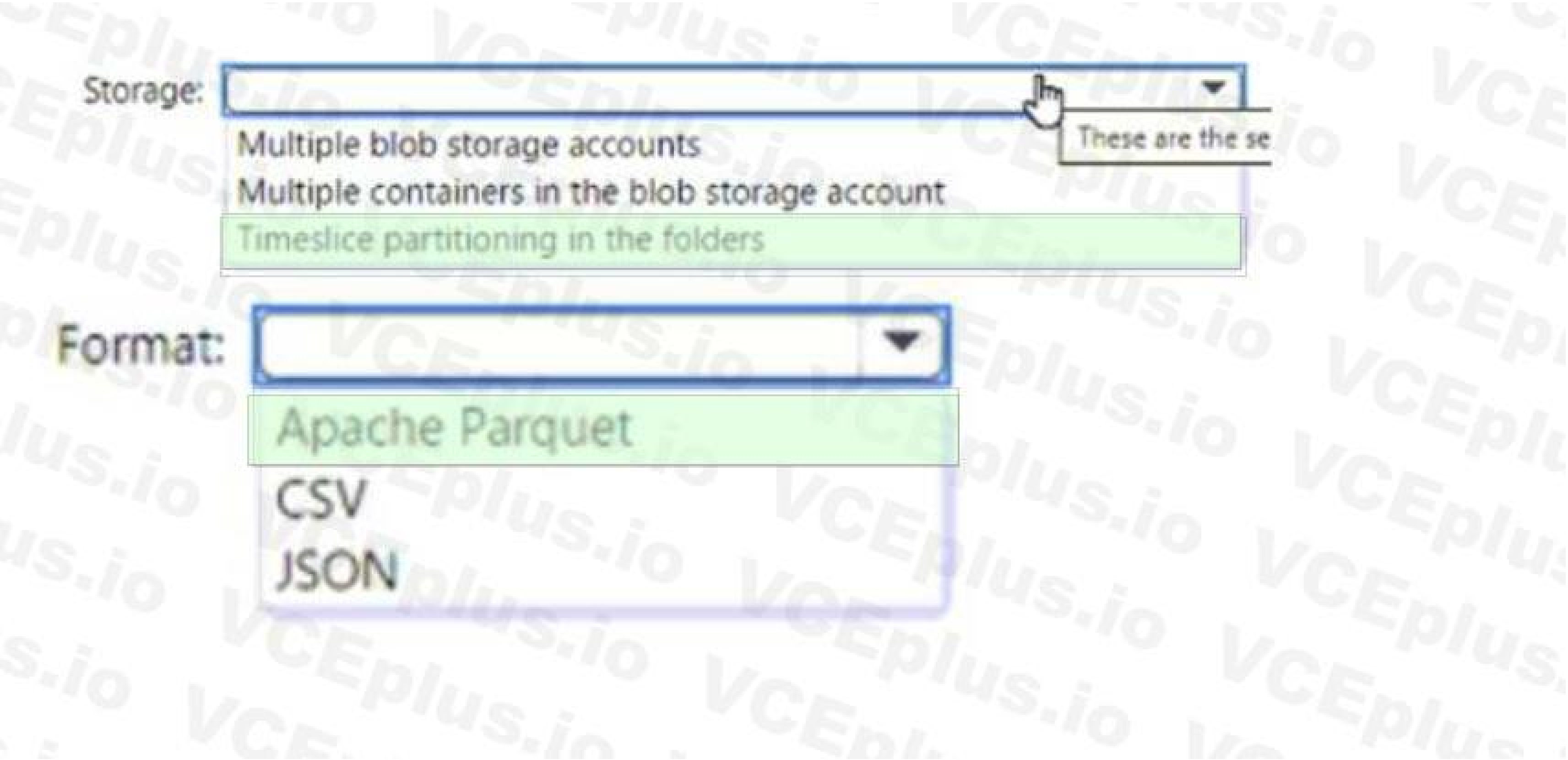
You have an Azure Blob storage account that contains a folder. The folder contains 120,000 files. Each file contains 62 columns. Each day, 1,500 new files are added to the folder.

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?

Hot Area:


Storage:
Multiple blob storage accounts
Multiple containers in the blob storage account
Timeslice partitioning in the folders
These are the se
Format:
Apache Parquet
CSV
JSON

Answer Area:


Storage:
Multiple blob storage accounts
Multiple containers in the blob storage account
Timeslice partitioning in the folders
These are the se
Format:
Apache Parquet
CSV
JSON

Section:

Explanation:

Box 1 = timeslice partitioning in the folders

This 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.

QUESTION 61

HOTSPOT

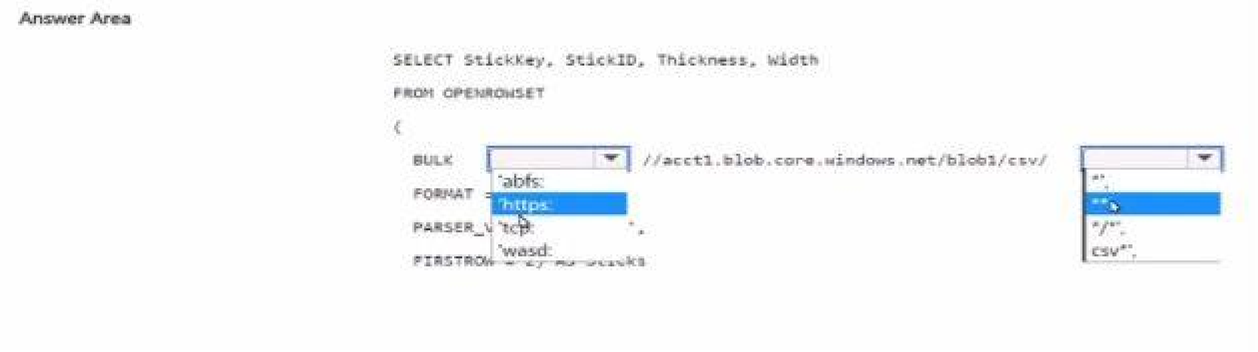
You have an Azure subscription that contains a storage account. The account contains a blob container named blob1 and an Azure Synapse Analytic serve-less SQL pool

You need to Query the CSV files stored in blob1. The solution must ensure that all the files in a (older named csv and all its subfolders are queried

How should you complete the query? to answer, select the appropriate options in the answer area

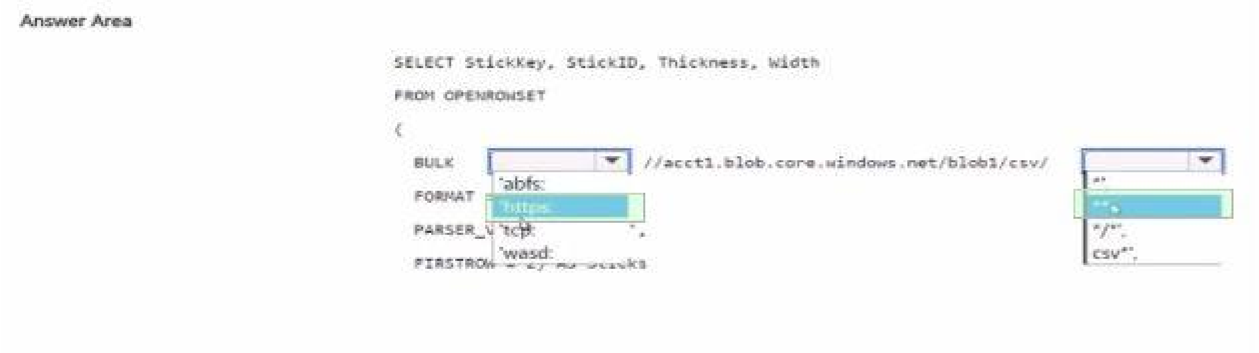
NOTE: Each correct selection is worth one point.

Hot Area:



```
SELECT StickKey, StickID, Thickness, Width
FROM OPENROWSET
(
  BULK  [          ▼] //acct1.blob.core.windows.net/blob1/csv/  [          ▼]
  FORMAT =
  PARSER_V                 ',
  FIRSTROW = 2) AS Sticks
```

Answer Area:



Answer Area

```
SELECT StickKey, StickID, Thickness, Width
FROM OPENROWSET
(
  BULK                //acct1.blob.core.windows.net/blob1/csv/
  FORMAT
  PARSER_V                           ',
  FIRSTROW                AS Sticks
```

'abfs:
'https:
'tcp:
'wasd:

*'
**',
/',
csv*',

Section:
Explanation:

QUESTION 62
You have an Azure subscription that contains a Microsoft Purview account.
You need to search the Microsoft Purview Data Catalog to identify assets that have an assetType property of Table or View
Which query should you run?

A. assetType IN (Table', 'View')
B. assetType:Table OR assetType:View
C. assetType - (Table or view)
D. assetType:(Table OR View)

Correct Answer: B
Section:

QUESTION 63
You have an Azure Synapse Analytics workspace.
You plan to deploy a lake database by using a database template in Azure Synapse.
Which two elements ate included in the template? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point

A. relationships
B. table definitions
C. table permissions
D. linked services
E. data formats

Correct Answer: A, B
Section:

QUESTION 64
You have an Azure subscription that contains an Azure data factory named ADF1.
From Azure Data Factory Studio, you build a complex data pipeline in ADF1.
You discover that the Save button is unavailable and there are validation errors that prevent the pipeline from being published.
You need to ensure that you can save the logic of the pipeline.
Solution: You export ADF1 as an Azure Resource Manager (ARM) template.

A. Yes

B. No

Correct Answer: B
Section:

QUESTION 65
You have a Microsoft Entra tenant.
The tenant contains an Azure Data Lake Storage Gen2 account named storage! that has two containers named fs1 and fs2. You have a Microsoft Entra group named Oepartment A . You need to meet the following requirements:
* OepartmentA must be able to read, write, and list all the files in fs1.
* OepartmentA must be prevented from accessing any files in fs2
* The solution must use the principle of least privilege.
Which role should you assign to DepartmentA?

A. Contributor for fsl

B. Storage Blob Data Owner for fsl

C. Storage Blob Data Contributor for storage1

D. Storage Blob Data Contributor for fsl

Correct Answer: D
Section:

QUESTION 66
You have an Azure Stream Analytics job that read data from an Azure event hub.
You need to evaluate whether the job processes data as quickly as the data arrives or cannot keep up.
Which metric should you review?

A. InputEventLastPunctuationTime

B. Input Sources Receive

C. Late input Events

D. Backlogged input Events

Correct Answer: B
Section: