Number: DP-203
Passing Score: 800
Time Limit: 120
File Version: 31.0

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
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

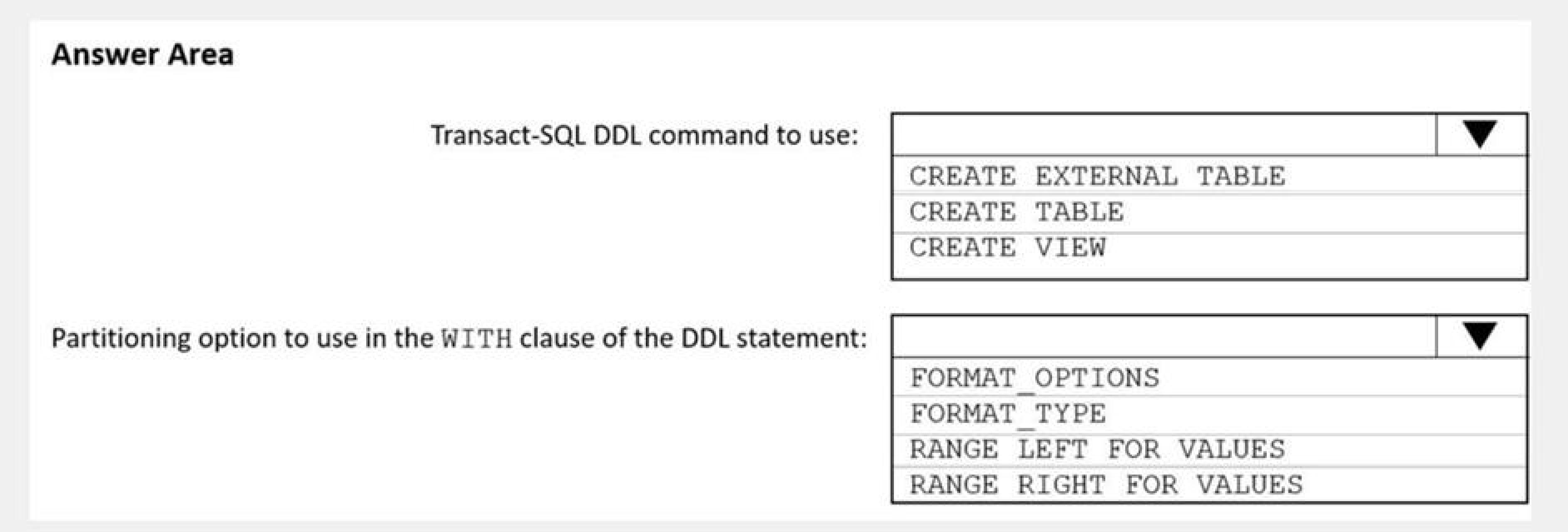
QUESTION 2
You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements. Which Azure Storage functionality should you include in the solution?

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

Correct Answer: B
Section:
Explanation:

QUESTION 3
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

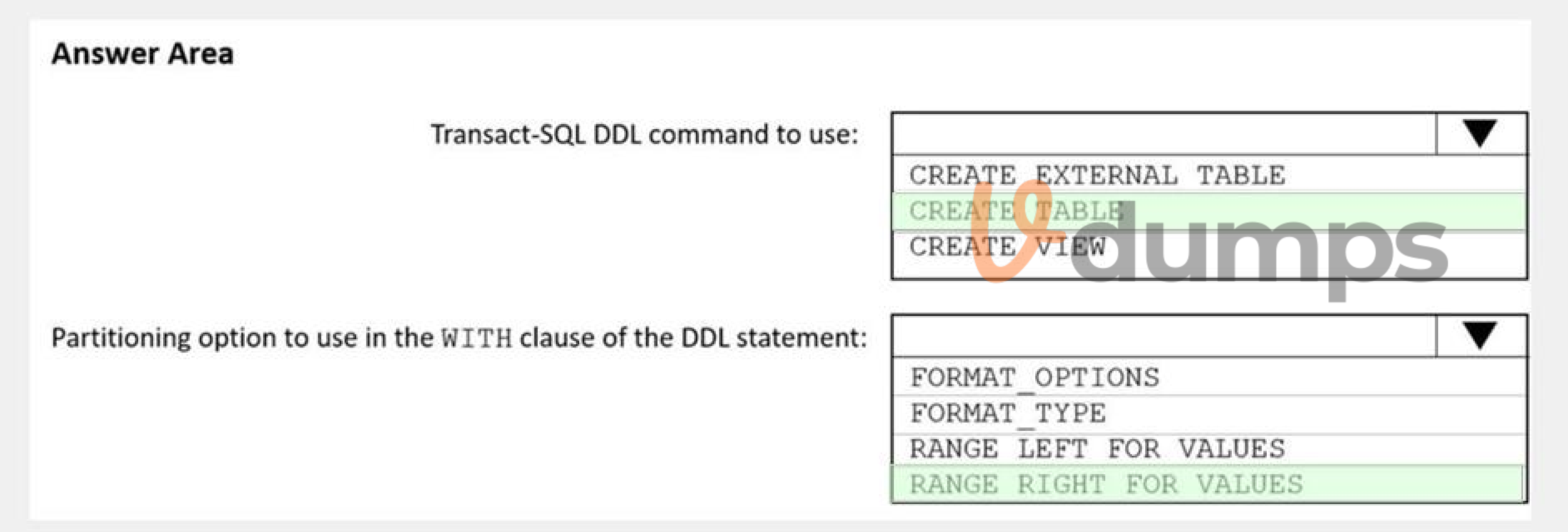
Transact-SQL DDL command to use:

- 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 4
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

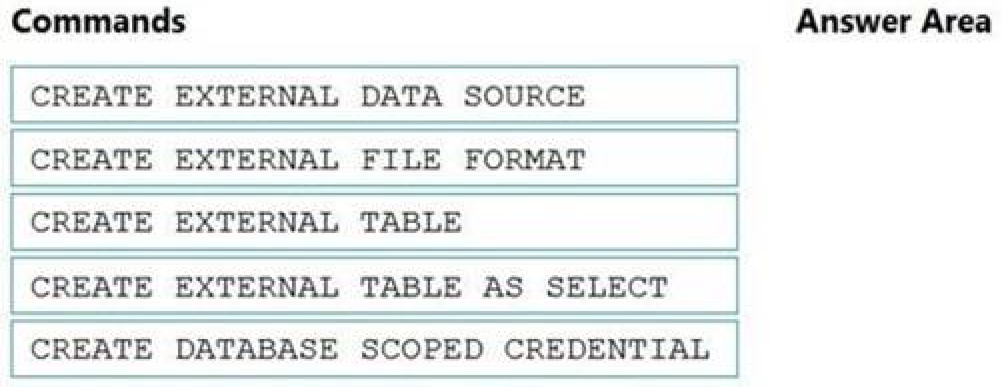
and contains 3 dataset named DS1. DS1 references 2 file in storage.In DF1, you plan to create a pipeline that will process data from DS1.You need to review the schema and lineage information in MP1 for the data referenced by DS1.Which two features can you use to locate the information? Each correct answer 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 5
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:

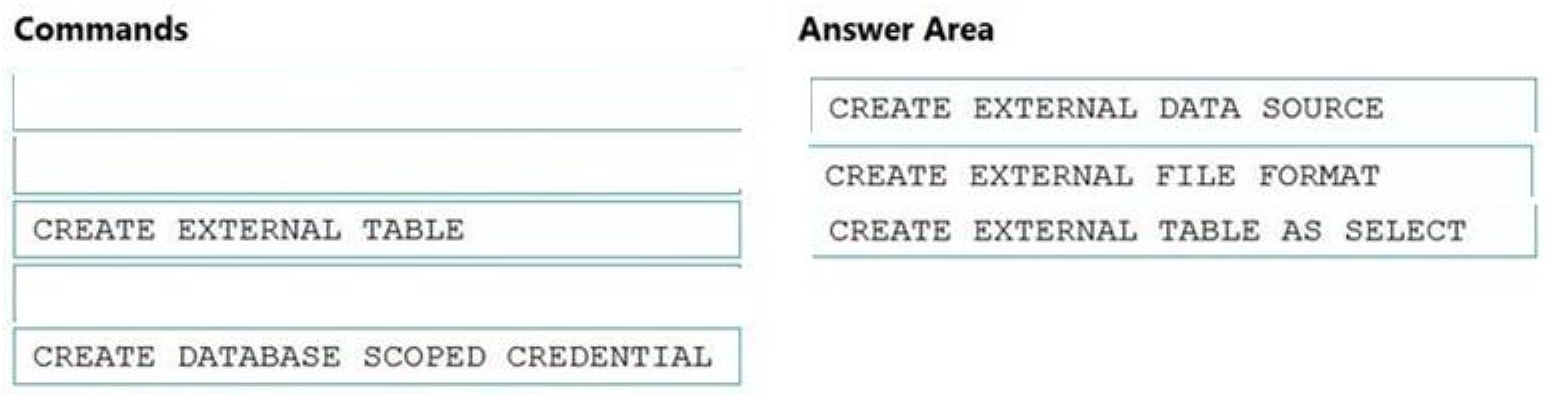


Commands

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 6

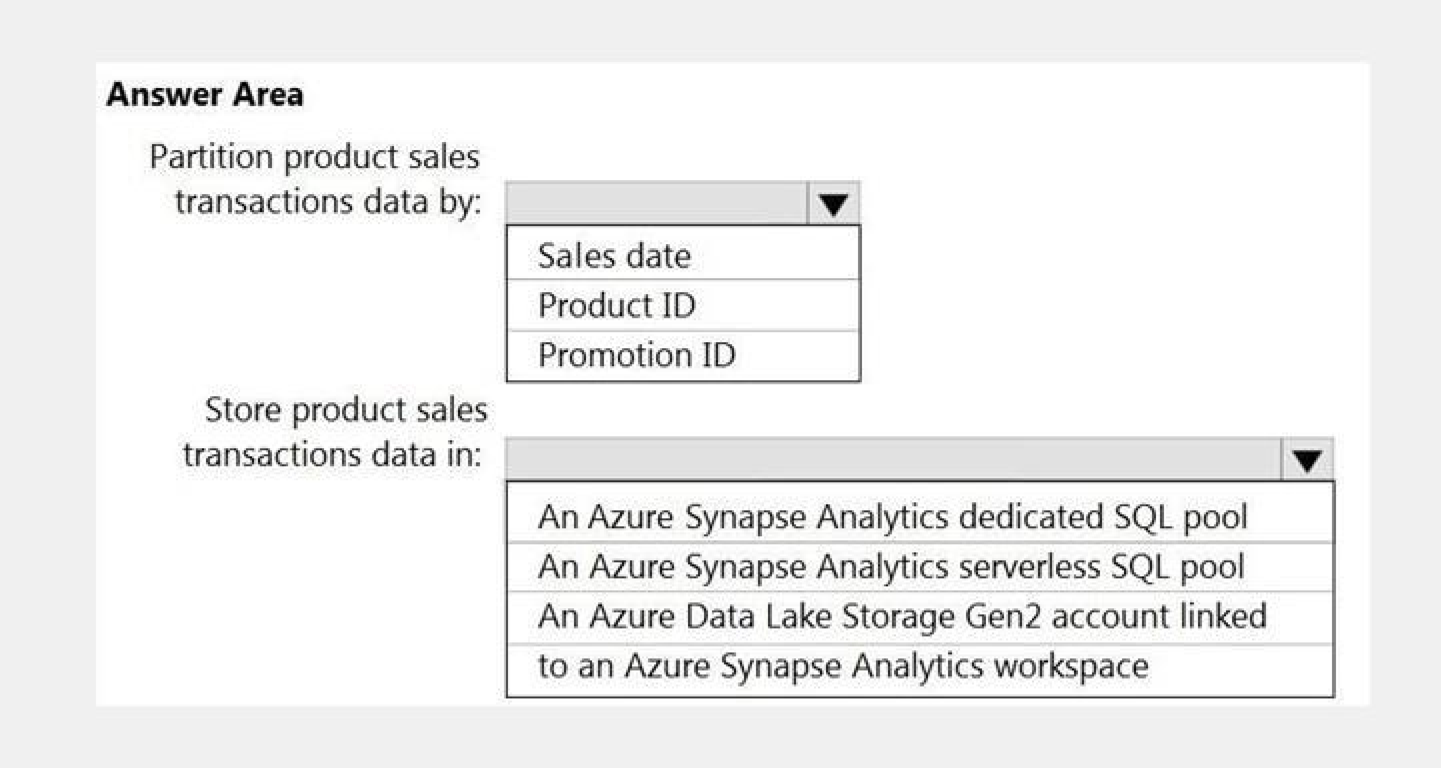
HOTSPOT

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

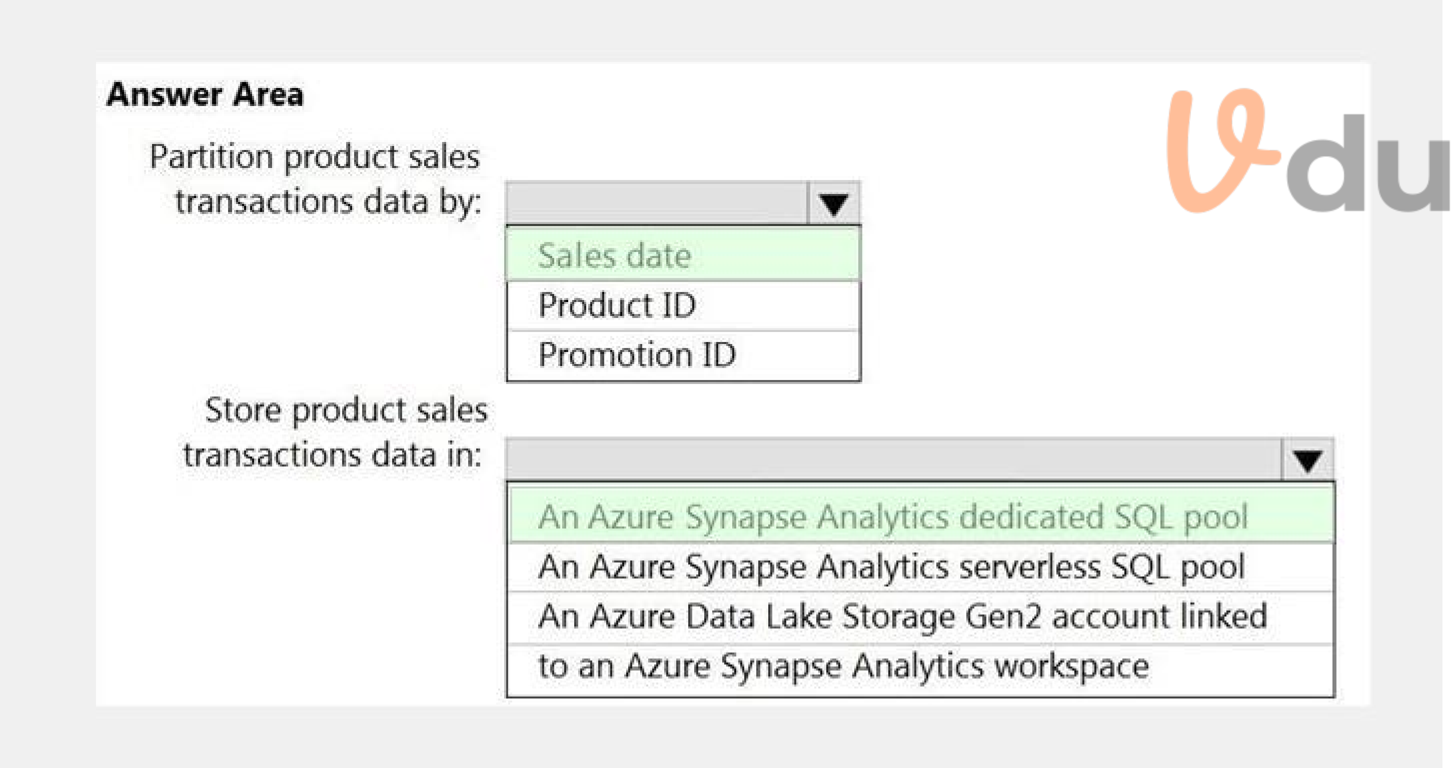
Partition product sales transactions data by: ▼

- 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:

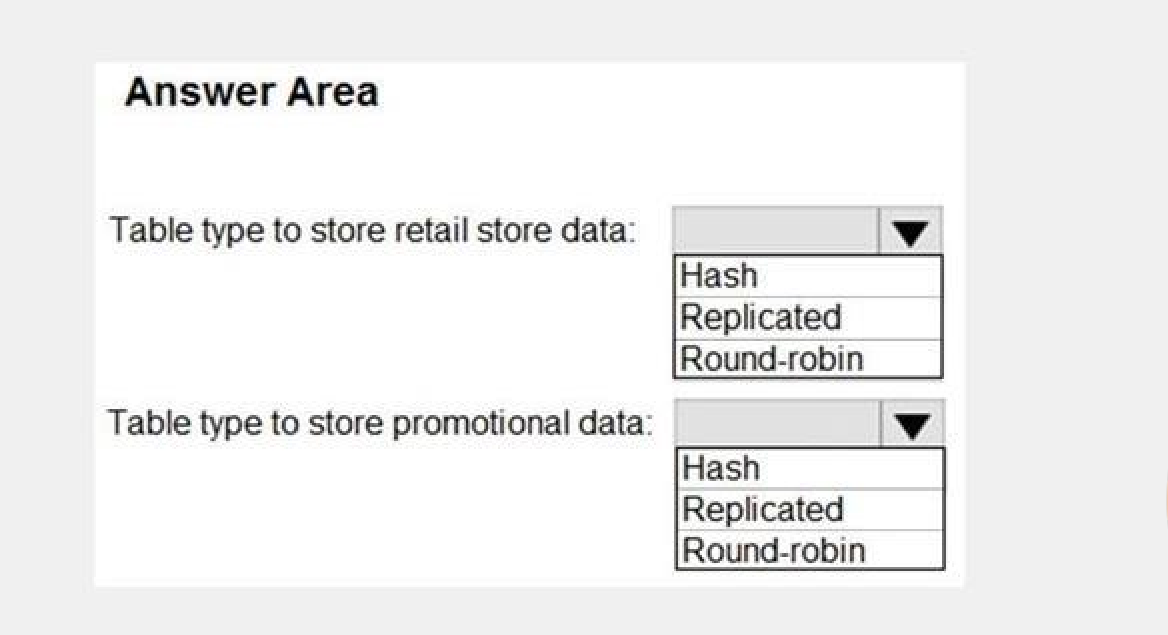
- Ensure that data storage costs and performance are predictable.

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 7
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:

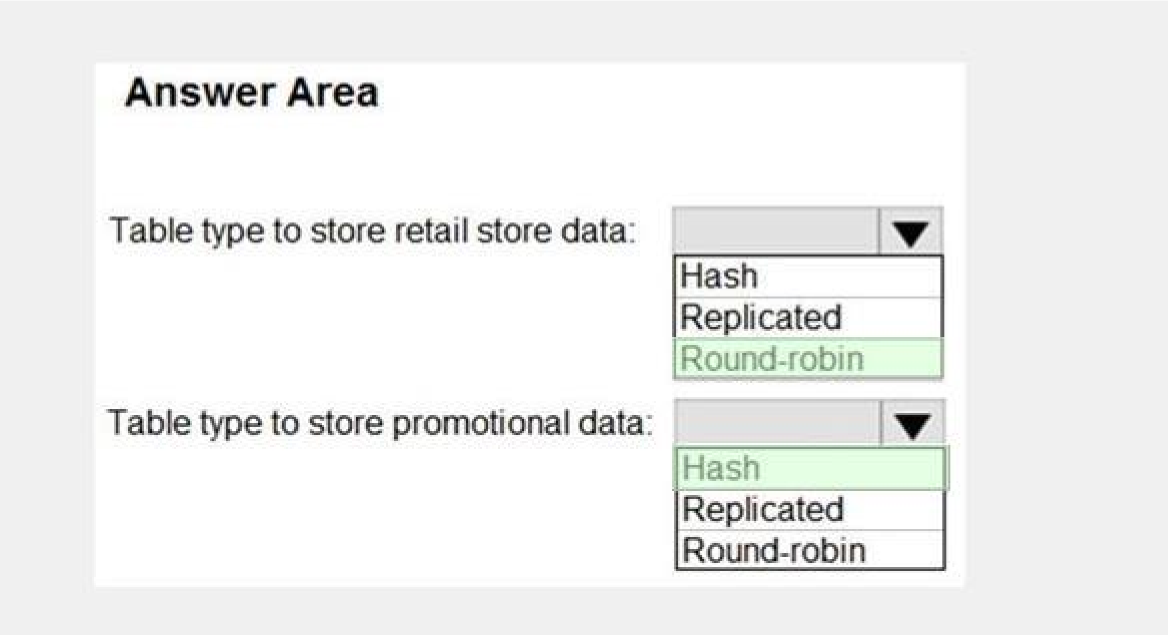


Answer Area

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

02-Design and implement data storage

QUESTION 1

DRAG DROP

You have a table named SalesFact in an enterprise data warehouse in Azure Synapse Analytics. SalesFact contains sales data from the past 36 months and has the following characteristics:

Is partitioned by month

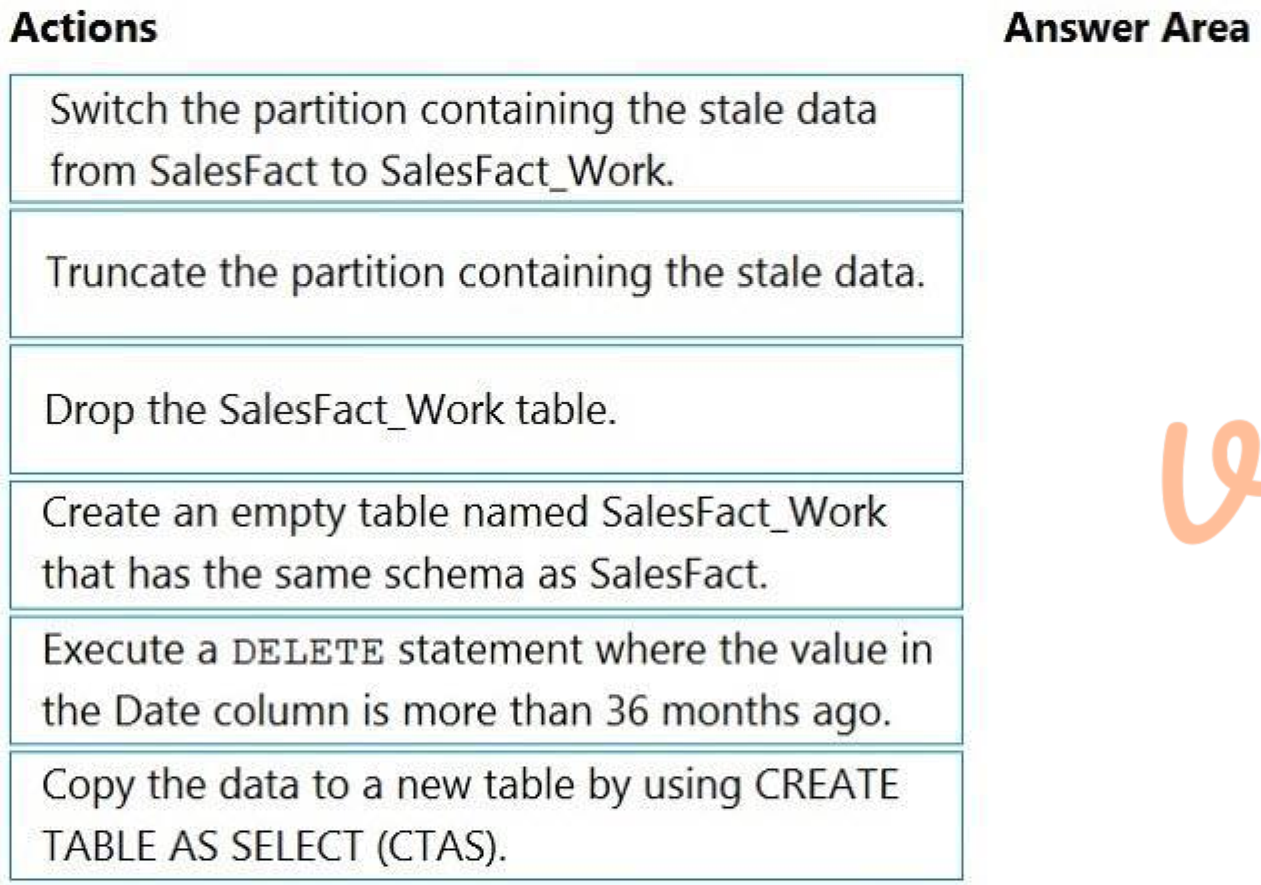
Contains one billion rows

Has clustered columnstore indexes

At the beginning of each month, you need to remove data from SalesFact that is older than 36 months as quickly as possible.

Which three actions should you perform in sequence in a stored procedure? 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:

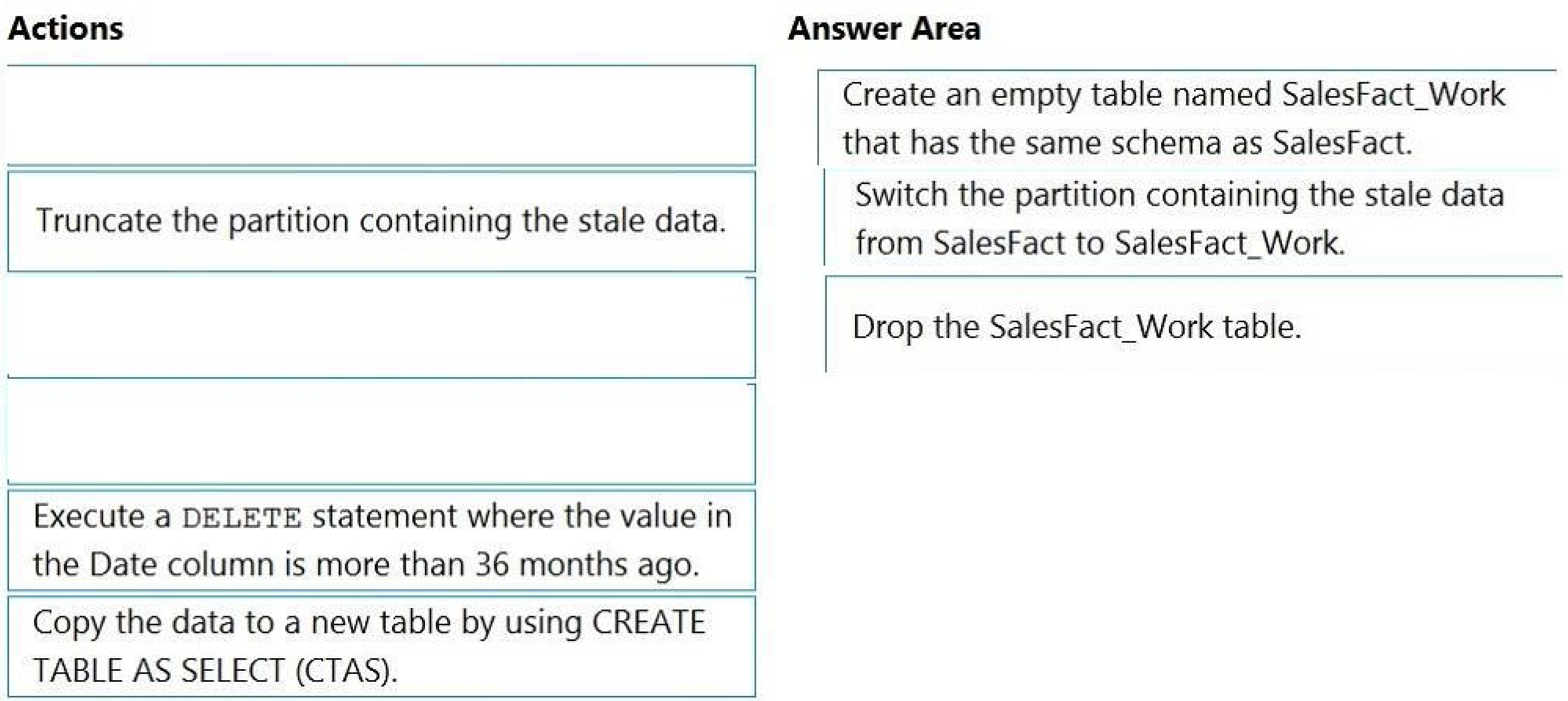


Actions

- Switch the partition containing the stale data from SalesFact to SalesFact_Work.
- Truncate the partition containing the stale data.
- Drop the SalesFact_Work table.
- Create an empty table named SalesFact_Work that has the same schema as SalesFact.
- Execute a DELETE statement where the value in the Date column is more than 36 months ago.
- Copy the data to a new table by using CREATE TABLE AS SELECT (CTAS).

Answer Area

Correct Answer:



Actions

Truncate the partition containing the stale data.

Execute a DELETE statement where the value in the Date column is more than 36 months ago.

Copy the data to a new table by using CREATE TABLE AS SELECT (CTAS).

Answer Area

Create an empty table named SalesFact_Work that has the same schema as SalesFact.

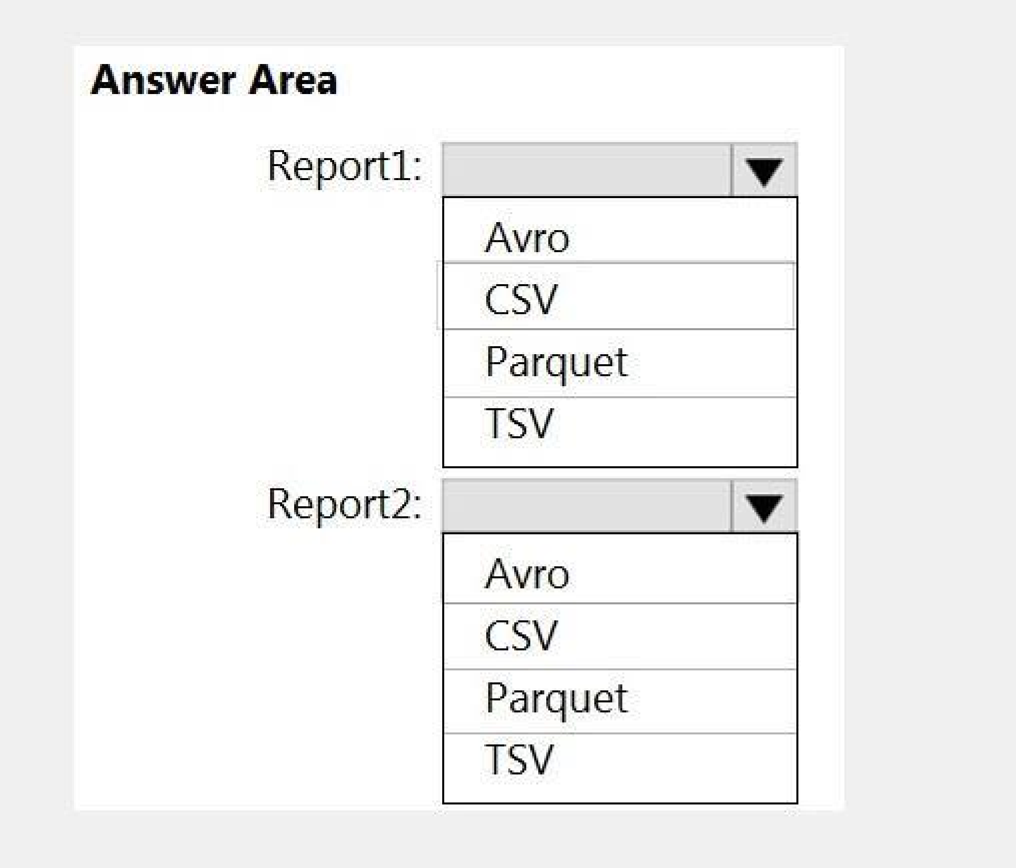
Switch the partition containing the stale data from SalesFact to SalesFact_Work.

Drop the SalesFact_Work table.

Section:
Explanation:
Step 1: Create an empty table named SalesFact_work that has the same schema as SalesFact.
Step 2: Switch the partition containing the stale data from SalesFact to SalesFact_Work. SQL Data Warehouse supports partition splitting, merging, and switching. To switch partitions between two tables, you must ensure that the partitions align on their respective boundaries and that the table definitions match.
Loading data into partitions with partition switching is a convenient way stage new data in a table that is not visible to users the switch in the new data. Step 3: Drop the SalesFact_Work table.
Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-partition

QUESTION 2
HOTSPOT
You are planning the deployment of Azure Data Lake Storage Gen2.
You have the following two reports that will access the data lake:
Report1: Reads three columns from a file that contains 50 columns. Report2: Queries a single record based on a timestamp.
You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.
What should you recommend for each report? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

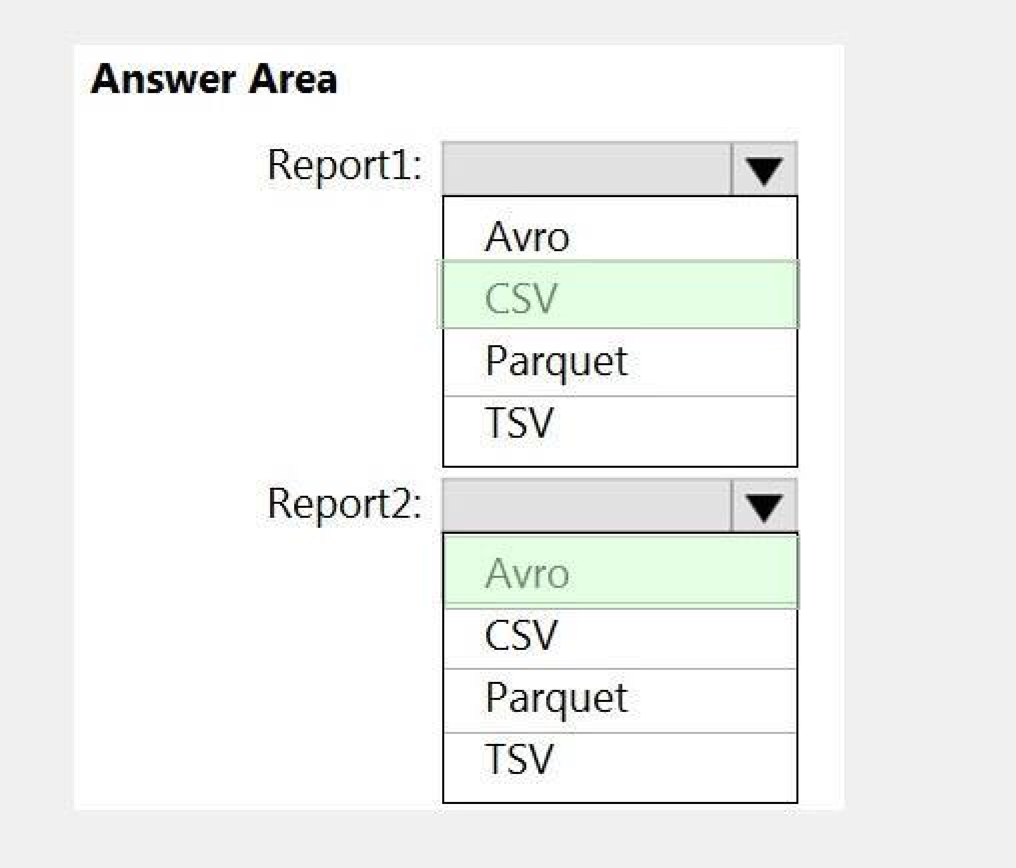
Report1:

- Avro
- CSV
- Parquet
- TSV

Report2:

- Avro
- CSV
- Parquet
- TSV

Answer Area:



Answer Area

Report1:

- Avro
- [✓ CSV]
- Parquet
- TSV

Report2:

- [✓ Avro]
- CSV
- Parquet
- TSV

Section:
Explanation:

QUESTION 3
You are implementing a batch dataset in the Parquet format. Data files will be produced be using Azure Data Factory and stored in Azure Data Lake Storage Gen2. The files will be consumed by an Azure Synapse Analytics serverless SQL pool. You need to minimize storage costs for the solution.
What should you do?

A. Use Snappy compression for files.
B. Use OPENROWSET to query the Parquet files.
C. Create an external table that contains a subset of columns from the Parquet files.
D. Store all data as string in the Parquet files.

Correct Answer: C
Section:
Explanation:
An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

QUESTION 4
You are designing a data mart for the human resources (HR) department at your company. The data mart will contain employee information and employee transactions.
From a source system, you have a flat extract that has the following fields:
EmployeeID
FirstName
LastName
Recipient
GrossAmount
TransactionID
GovernmentID
NetAmountPaid
TransactionDate
You need to design a star schema data model in an Azure Synapse Analytics dedicated SQL pool for the data mart.
Which two tables should you create? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. a dimension table for Transaction
B. a dimension table for EmployeeTransaction
C. a dimension table for Employee
D. a fact table for Employee
E. a fact table for Transaction

Correct Answer: C, E
Section:
Explanation:
C: Dimension tables contain attribute data that might change but usually changes infrequently. For example, a customer's name and address are stored in a dimension table and updated only when the customer's profile changes. To minimize the size of a large fact table, the customer's name and address don't need to be in every row of a fact table. Instead, the fact table and the dimension table can share a customer ID. A query can join the two tables to associate a customer's profile and transactions.
E: Fact tables contain quantitative data that are commonly generated in a transactional system, and then loaded into the dedicated SQL pool. For example, a retail business generates sales transactions every day, and then loads the data into a dedicated SQL pool fact table for analysis.

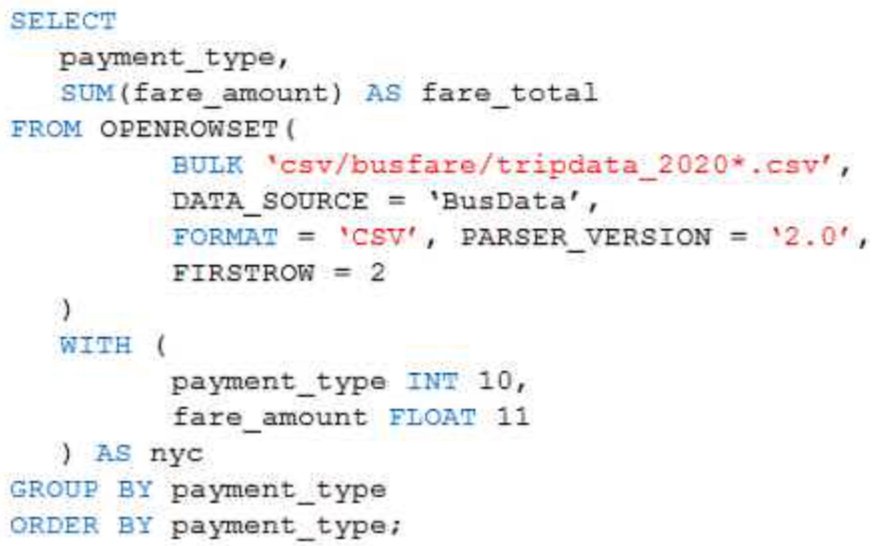
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overview

QUESTION 5
You are designing a dimension table for a data warehouse. The table will track the value of the dimension attributes over time and preserve the history of the data by adding new rows as the data changes. Which type of slowly changing dimension (SCD) should you use?

A. Type 0
B. Type 1
C. Type 2
D. Type 3

Correct Answer: C
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.
Incorrect Answers:
B: A Type 1 SCD always reflects the latest values, and when changes in source data are detected, the dimension table data is overwritten. D: A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD.
Reference:
https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types

QUESTION 6
You are performing exploratory analysis of the bus fare data in an Azure Data Lake Storage Gen2 account by using an Azure Synapse Analytics serverless SQL pool.
You execute the Transact-SQL query shown in the following exhibit.



```
SELECT
   payment_type,
   SUM(fare_amount) AS fare_total
FROM OPENROWSET(
         BULK 'csv/busfare/tripdata_2020*.csv',
         DATA_SOURCE = ‘BusData’,
         FORMAT = 'CSV', PARSER_VERSION = '2.0',
         FIRSTROW = 2
   )
   WITH (
         payment_type INT 10,
         fare_amount FLOAT 11
   ) AS nyc
GROUP BY payment_type
ORDER BY payment_type;
```

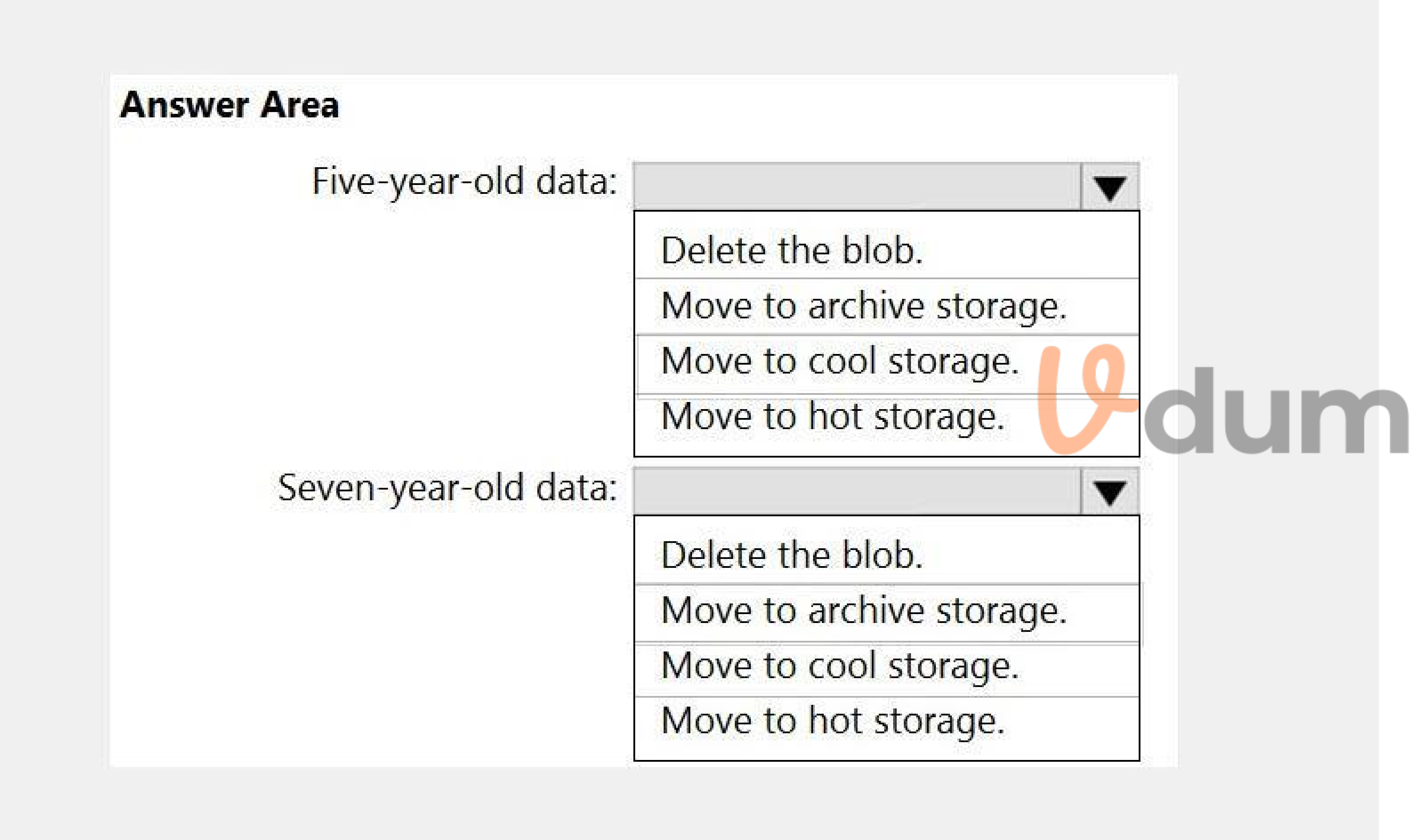
What do the query results include?

A. Only CSV files in the tripdata_2020 subfolder.
B. All files that have file names that beginning with "tripdata_2020".
C. All CSV files that have file names that contain "tripdata_2020".
D. Only CSV that have file names that beginning with "tripdata_2020".

Correct Answer: D
Section:

QUESTION 7
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:

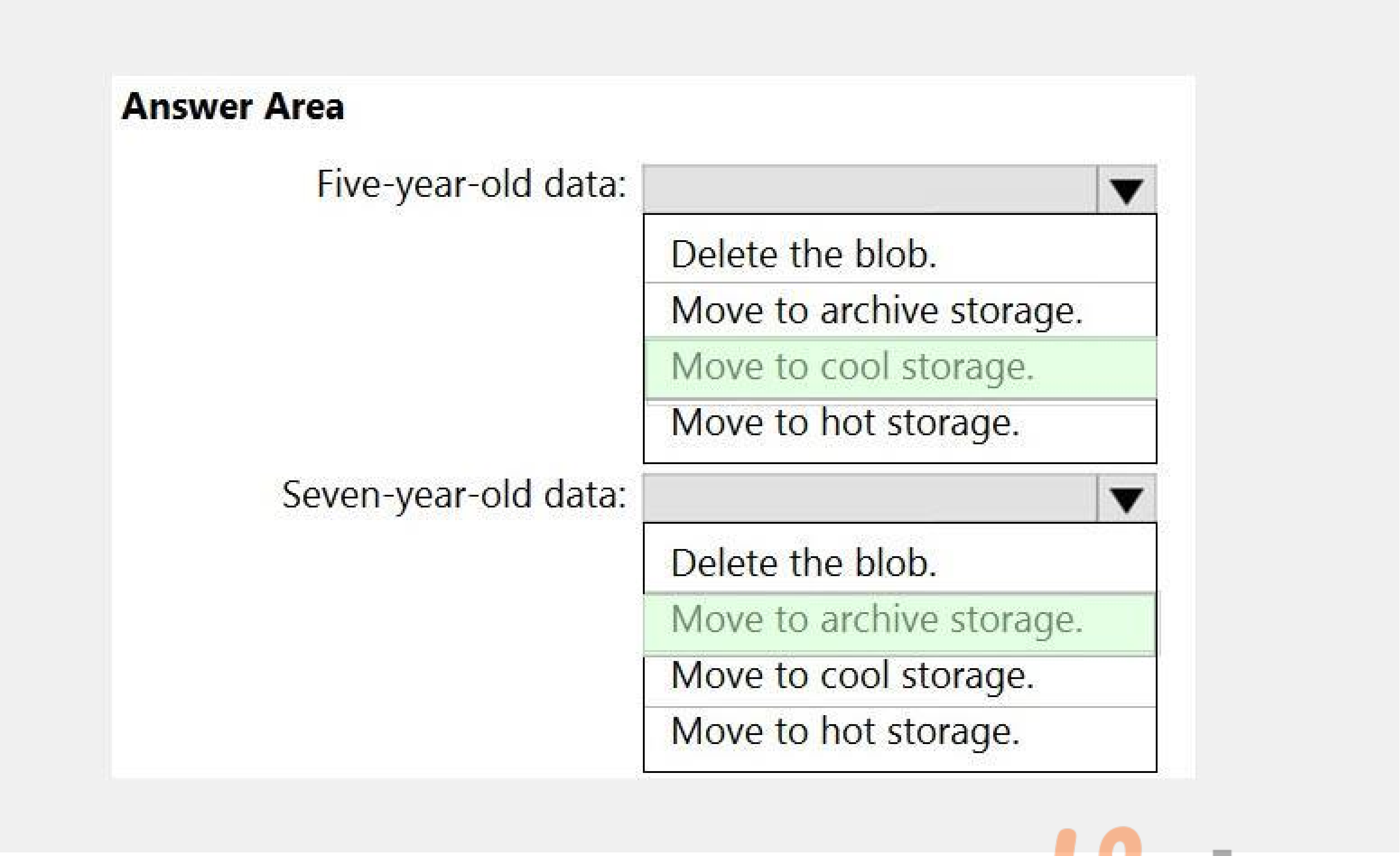


Answer Area

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

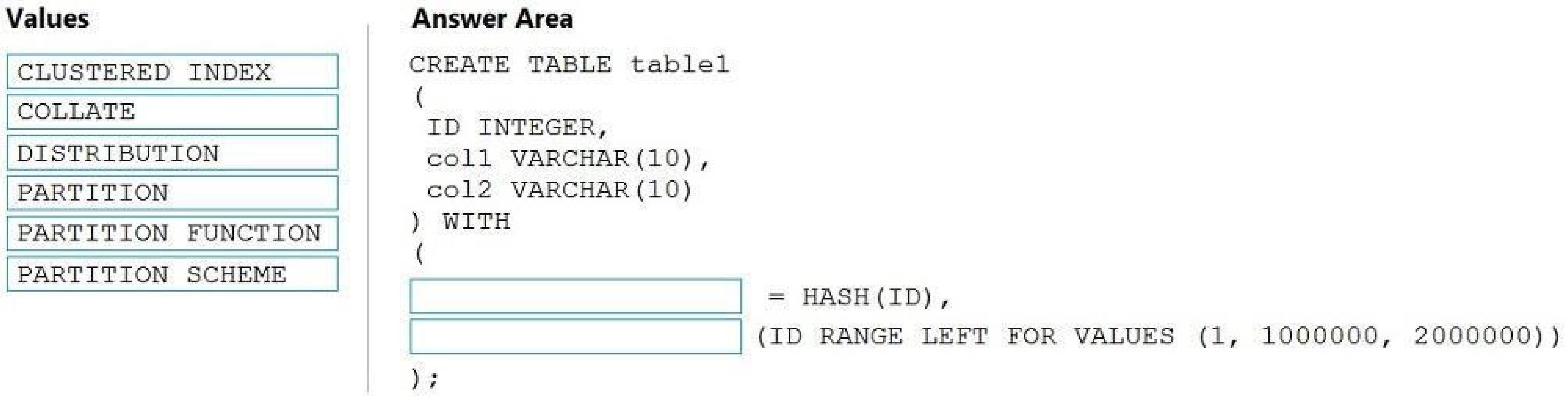
Five-year-old data:
- Delete the blob.
- Move to archive storage.
- [✓ Move to cool storage.]
- Move to hot storage.

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 8
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:



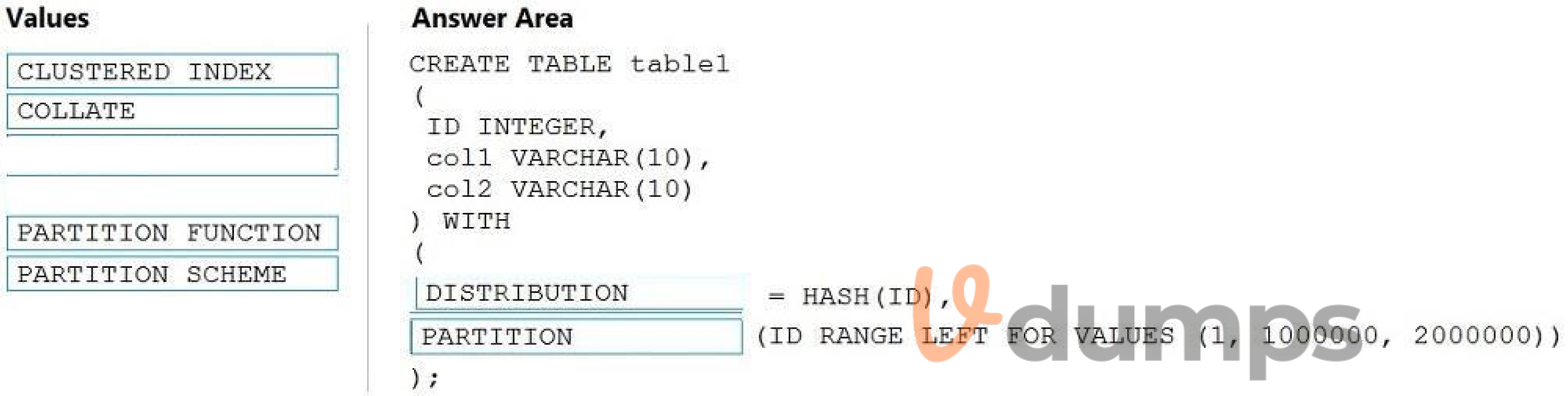
Values

- 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

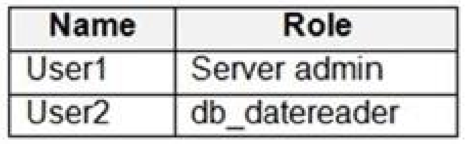
- CLUSTERED INDEX
- COLLATE
-
- PARTITION FUNCTION
- PARTITION SCHEME

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))
);
```

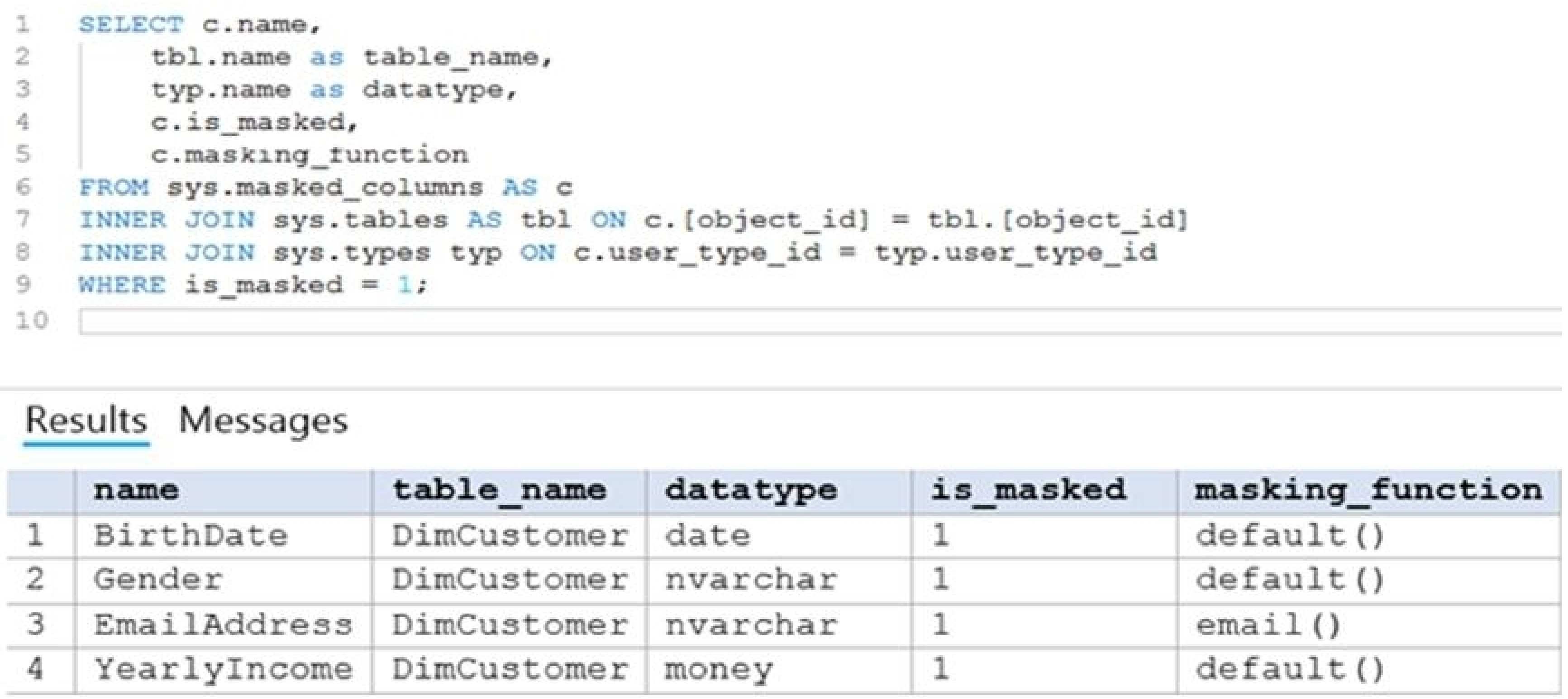
Section:
Explanation:
Box 1: DISTRIBUTION
Table distribution options include DISTRIBUTION = HASH (distribution_column_name), assigns each row to one distribution by hashing the value stored in distribution_column_name.
Box 2: PARTITION
Table partition options. Syntax:
PARTITION (partition_column_name 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 9
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;
```

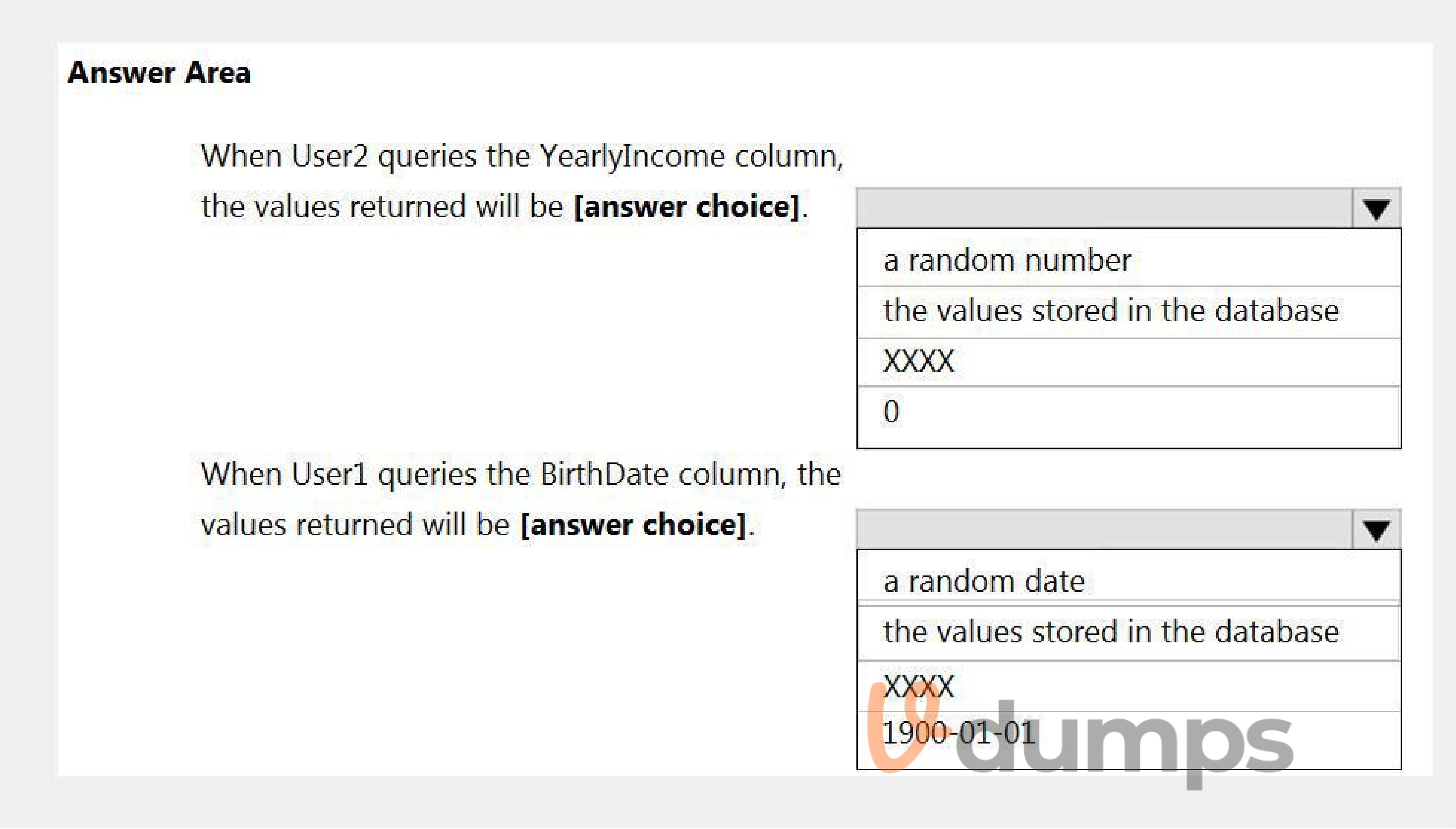
	name	table_name	datatype	is_masked	masking_function
1	BirthDate	DimCustomer	date	1	default()
2	Gender	DimCustomer	nvarchar	1	default()
3	EmailAddress	DimCustomer	nvarchar	1	email()
4	YearlyIncome	DimCustomer	money	1	default()

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

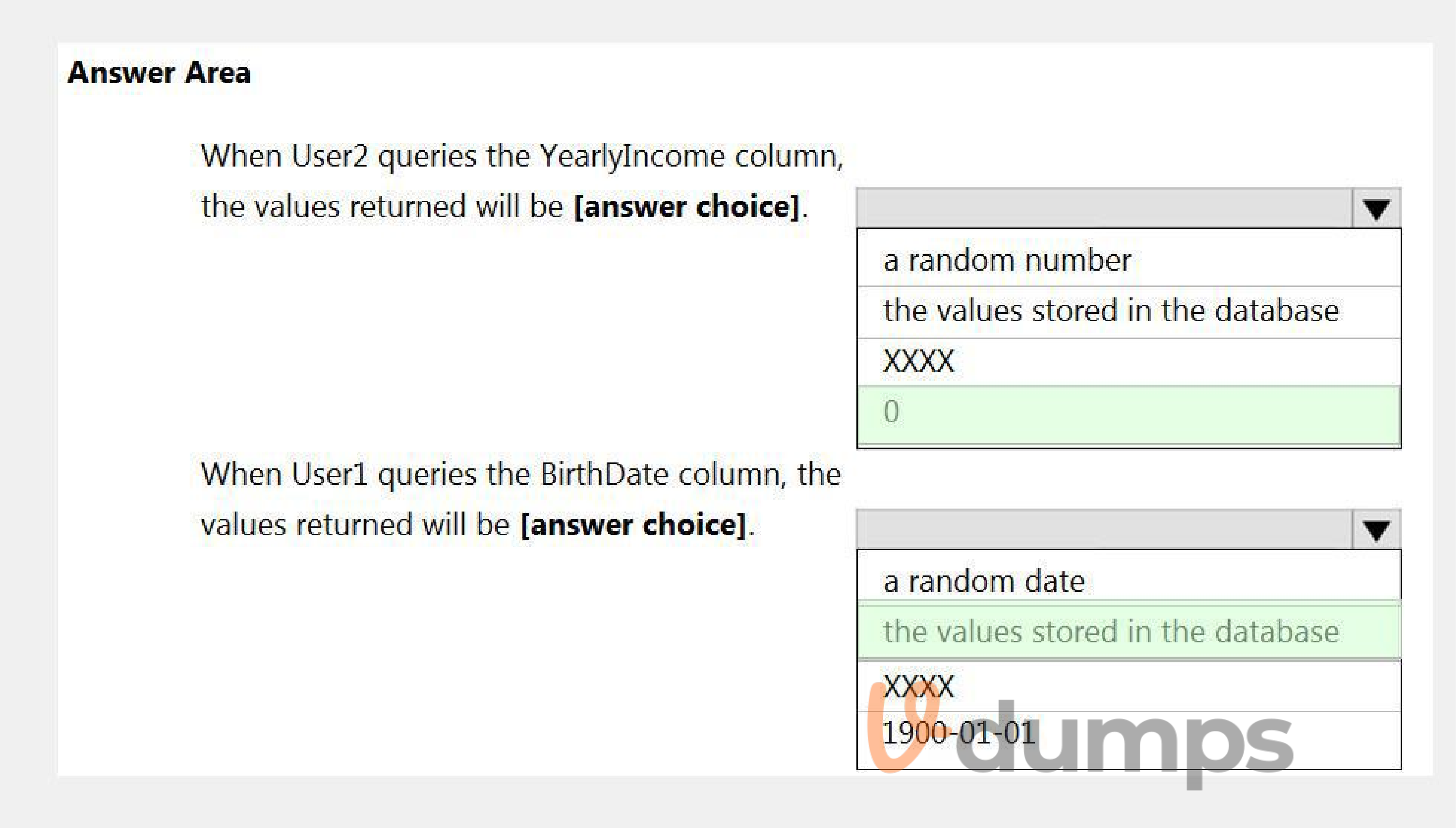
When User2 queries the YearlyIncome column, the values returned will be **[answer choice]**.

- 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]

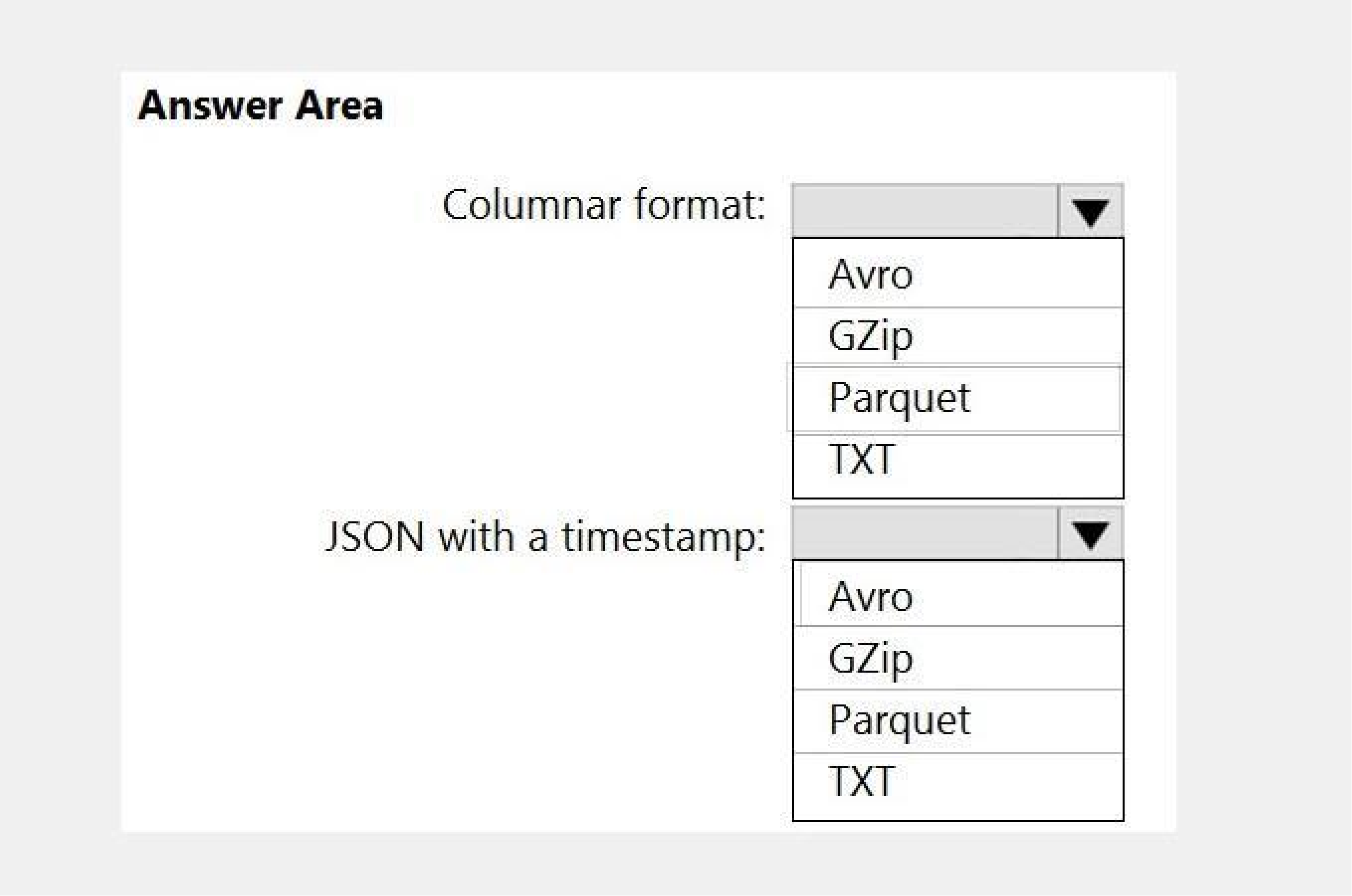
When User1 queries the BirthDate column, the values returned will be **[answer choice]**.

- 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 10
HOTSPOT
You need to output files from Azure Data Factory.
Which file format should you use for each type of output? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:

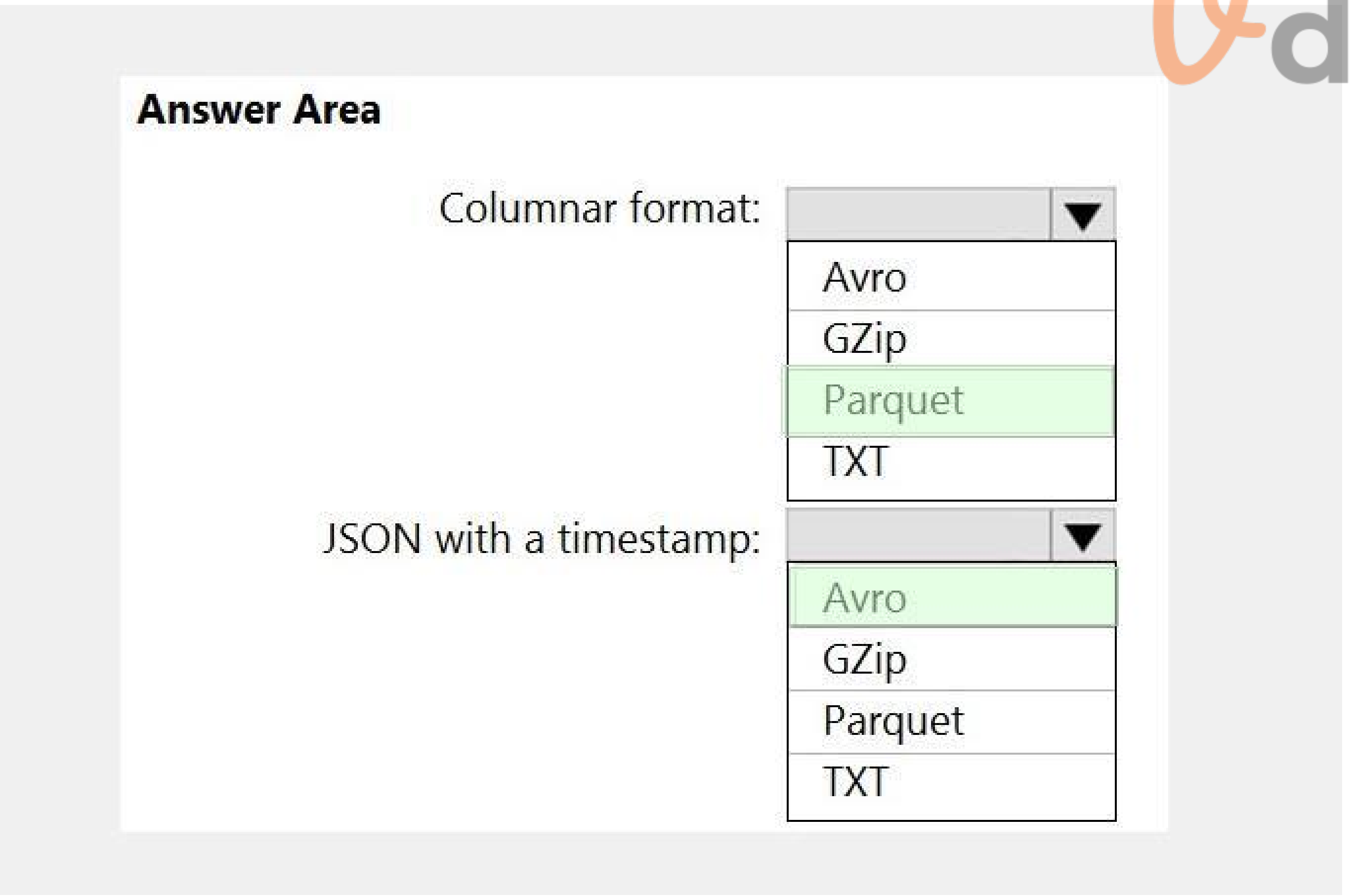


Answer Area

Columnar format: Avro / GZip / Parquet / TXT

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

Answer Area:



Answer Area

Columnar format: Avro / GZip / [✓ Parquet] / TXT

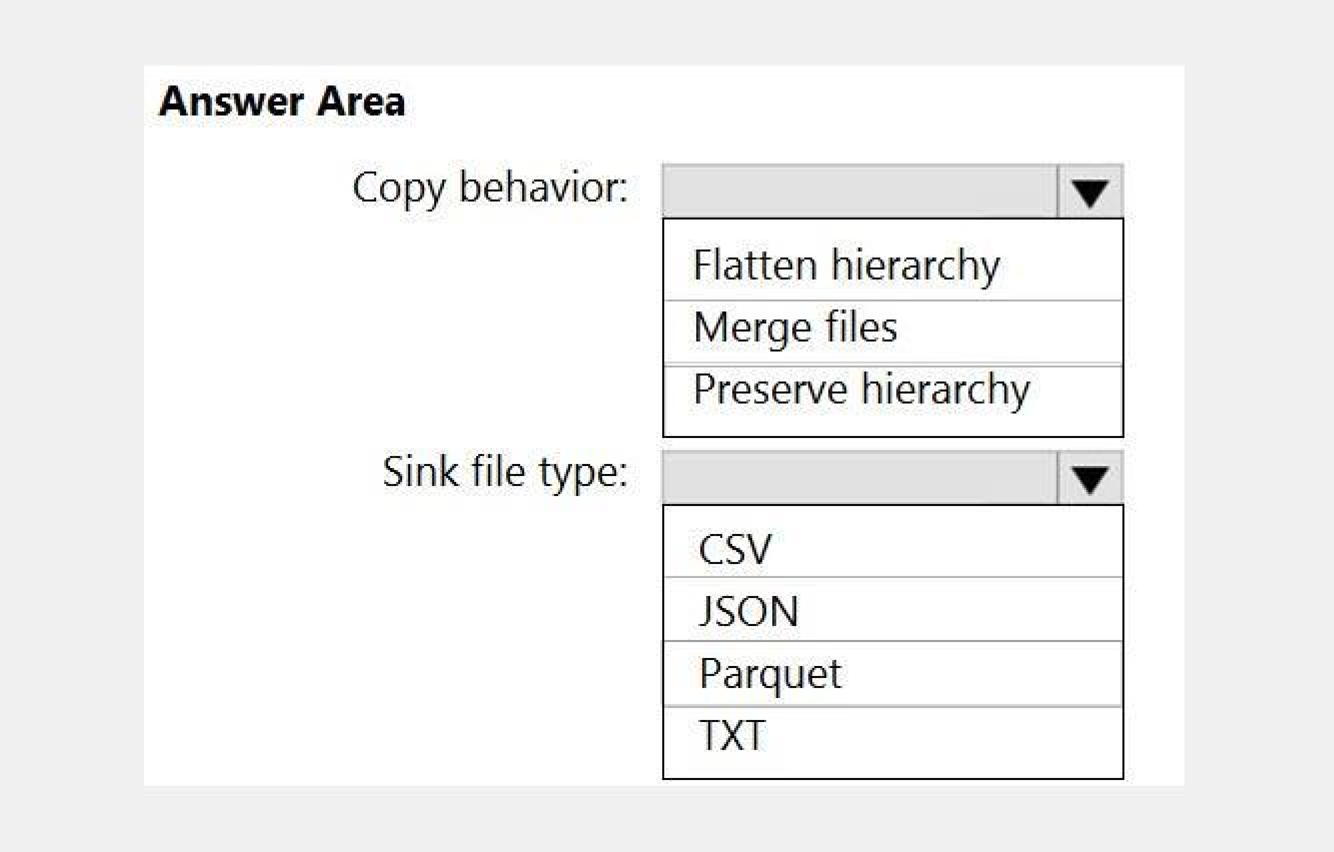
JSON with a timestamp: [✓ Avro] / GZip / Parquet / TXT

Section:
Explanation:

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

QUESTION 11
HOTSPOT
You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools.
Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company.
You need to move the files to a different folder and transform the data to meet the following requirements:
Provide the fastest possible query times.
Automatically infer the schema from the underlying files.
How should you configure the Data Factory copy activity? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:

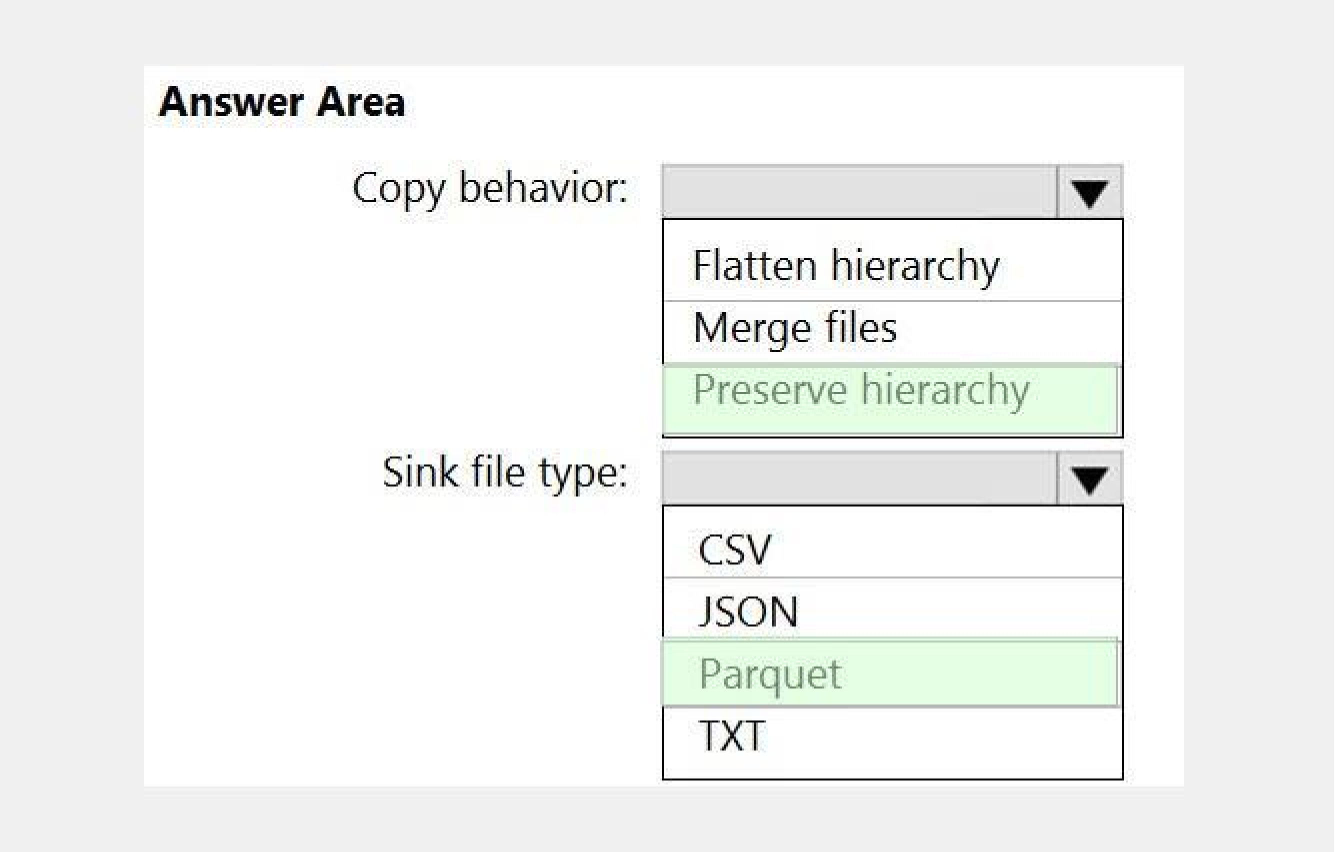


Answer Area

Copy behavior:
- Flatten hierarchy
- Merge files
- Preserve hierarchy

Sink file type:
- CSV
- JSON
- Parquet
- TXT

Answer Area:



Answer Area

Copy behavior:

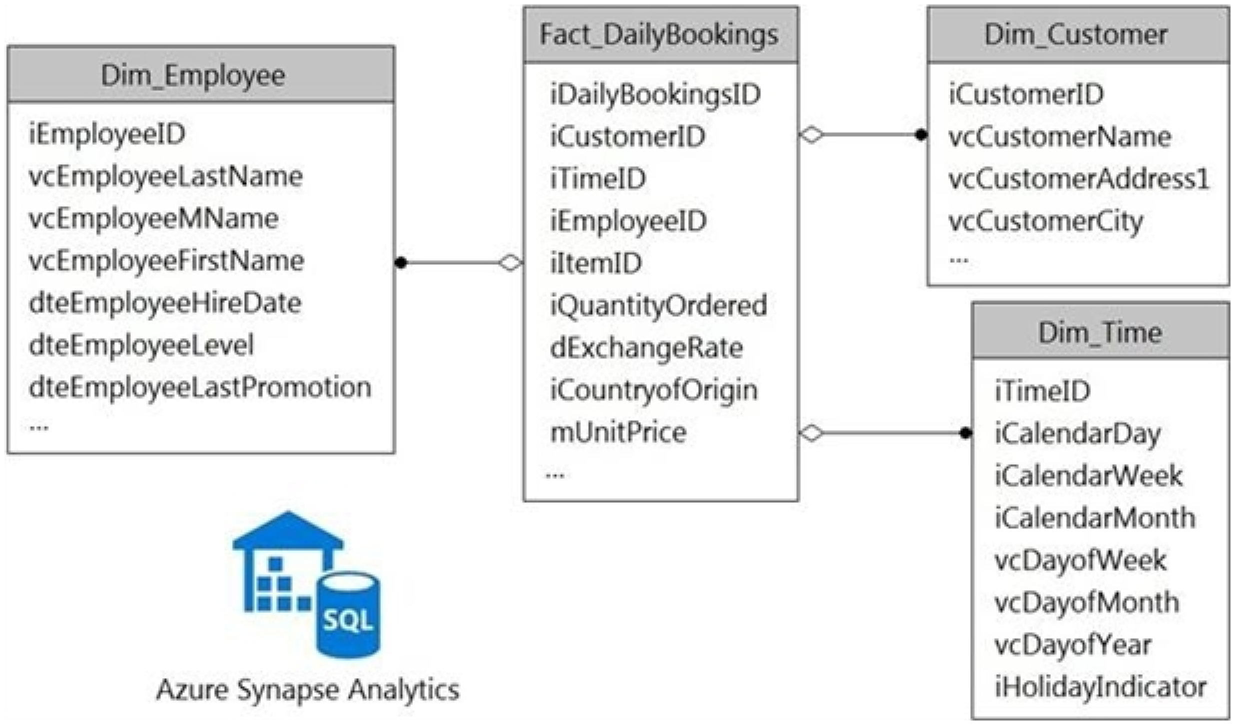
Flatten hierarchy
Merge files
[✓ Preserve hierarchy]

Sink file type:

CSV
JSON
[✓ Parquet]
TXT

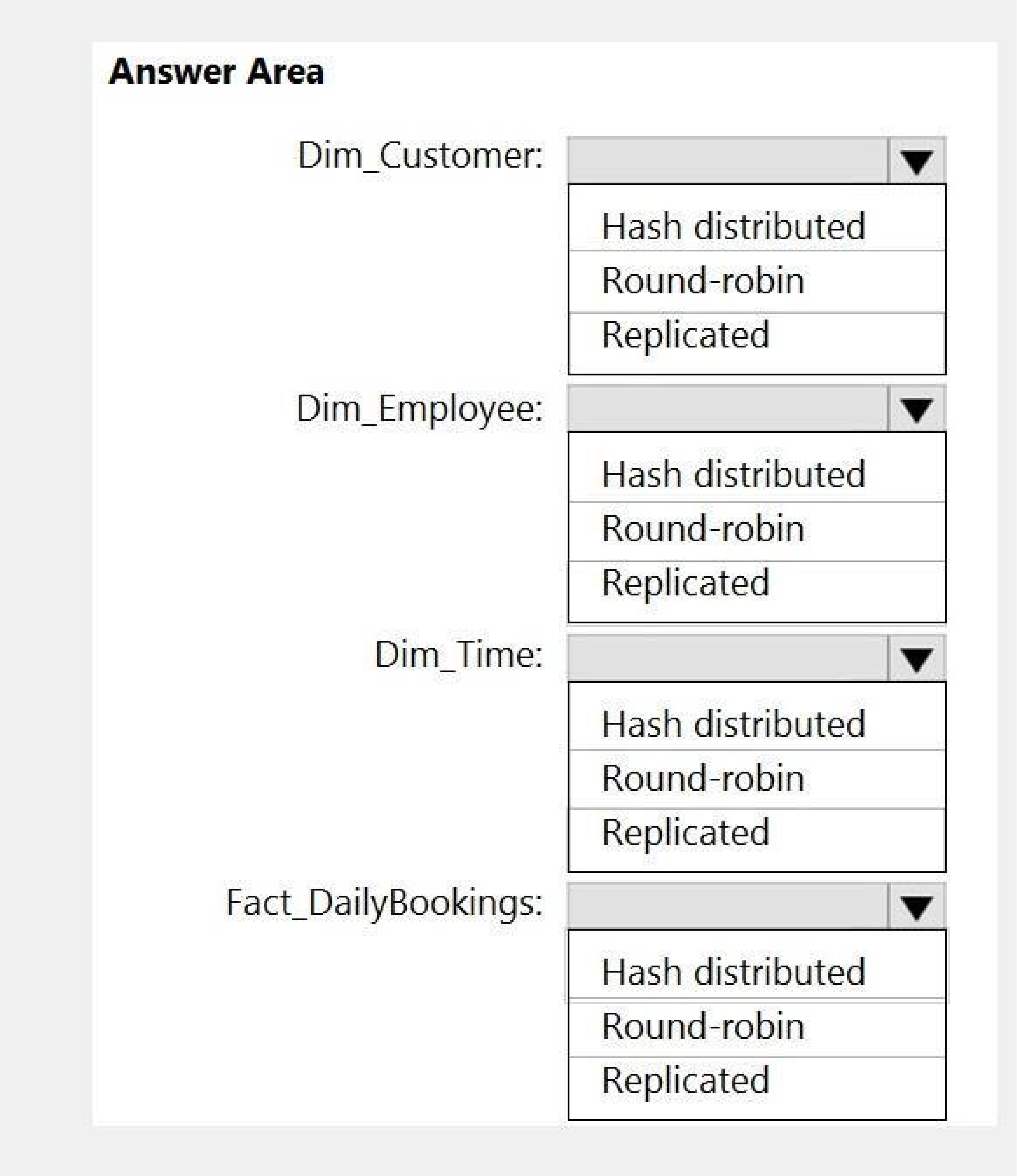
Section:
Explanation:
Box 1: Preserver herarchy
Compared to the flat namespace on Blob storage, the hierarchical namespace greatly improves the performance of directory management operations, which improves overall job performance.
Box 2: Parquet
Azure Data Factory parquet format is supported for Azure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

QUESTION 12
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

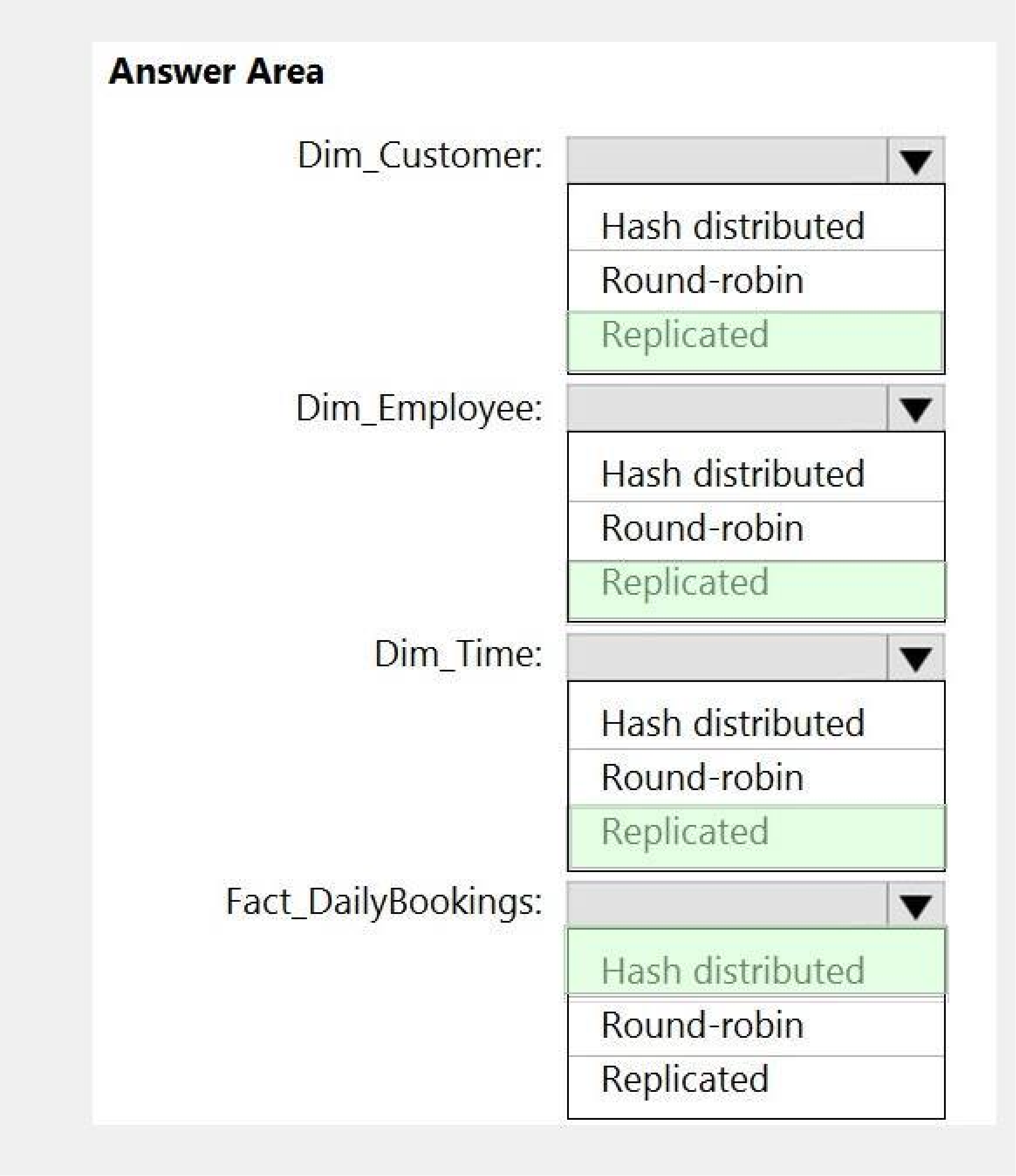
Dim_Time: ▼

- Hash distributed
- Round-robin
- Replicated

Fact_DailyBookings: ▼

- Hash distributed
- Round-robin
- Replicated

Answer Area:



Answer Area

Dim_Customer:
- Hash distributed
- Round-robin
- [✓ Replicated]

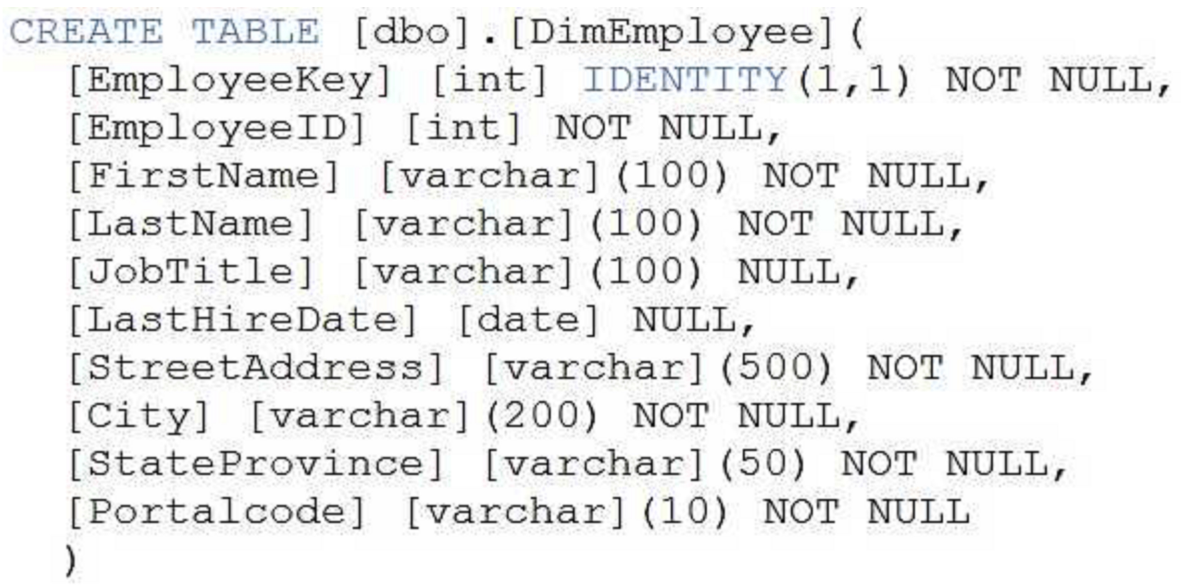
Dim_Employee:
- Hash distributed
- Round-robin
- [✓ Replicated]

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 13
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

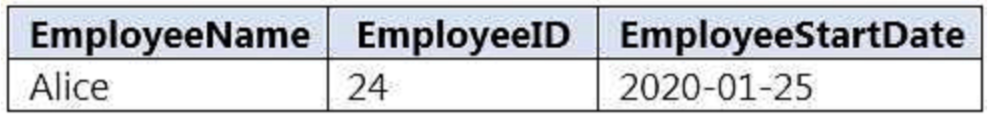
C. [ManagerEmployeeKey] [int] NULL

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 14

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?

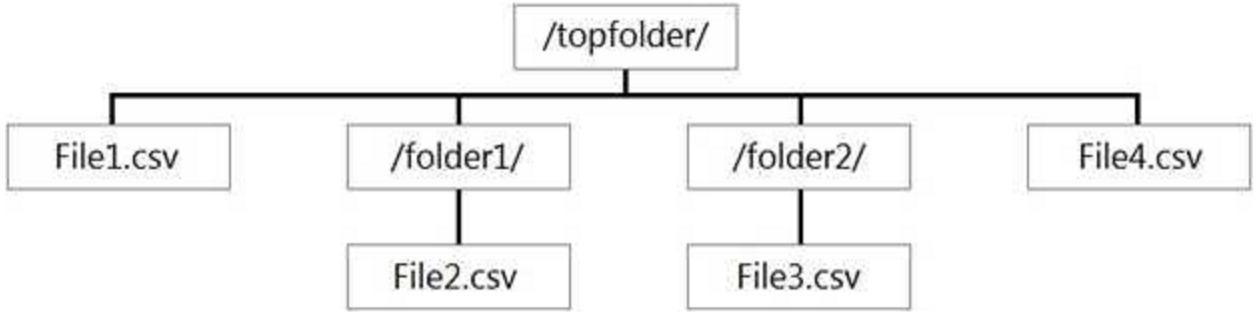
A. 24

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 15
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 16
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 17
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

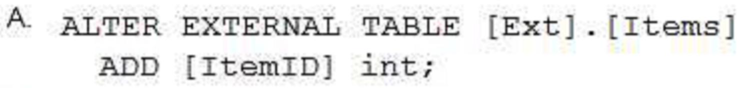
Correct Answer: D
Section:
Explanation:
A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique 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 18
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.

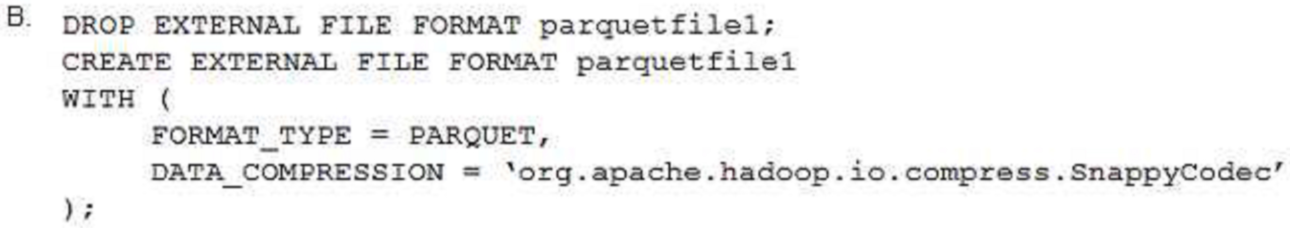
A. Add the managed identity to the Sales group.
B. Use the managed identity as the credentials for the data load process.
C. Create a shared access signature (SAS).
D. Add your Azure Active Directory (Azure AD) account to the Sales group.
E. Use the shared access signature (SAS) as the credentials for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

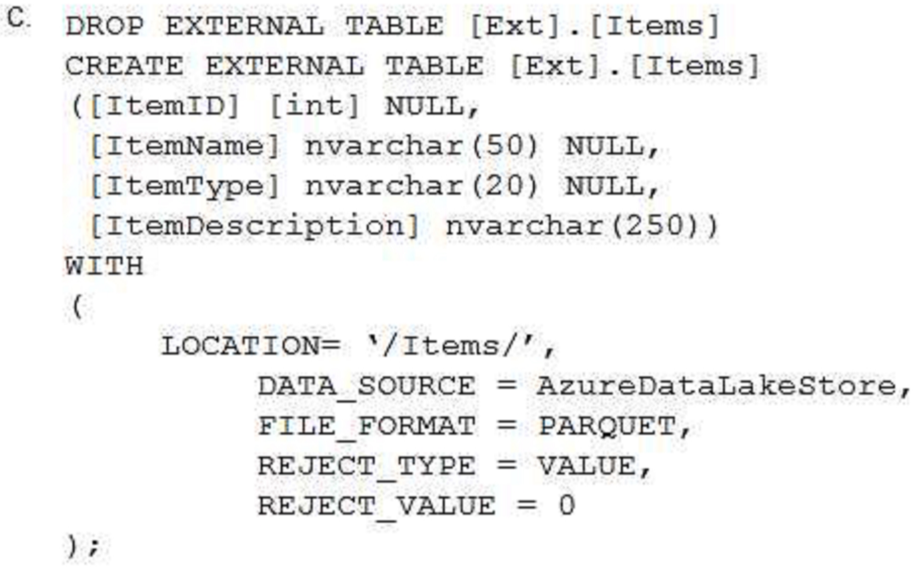
QUESTION 19
You have an enterprise data warehouse in Azure Synapse Analytics. Using PolyBase, you create an external table named [Ext].[Items] to query Parquet files stored in Azure Data Lake Storage Gen2 without importing the data to the data warehouse. The external table has three columns.
You discover that the Parquet files have a fourth column named ItemID. Which command should you run to add the ItemID column to the external table?



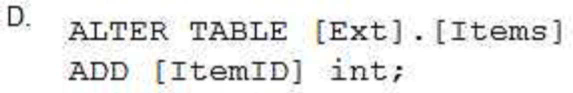
A.
```
ALTER EXTERNAL TABLE [Ext].[Items]
  ADD [ItemID] int;
```



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 20
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

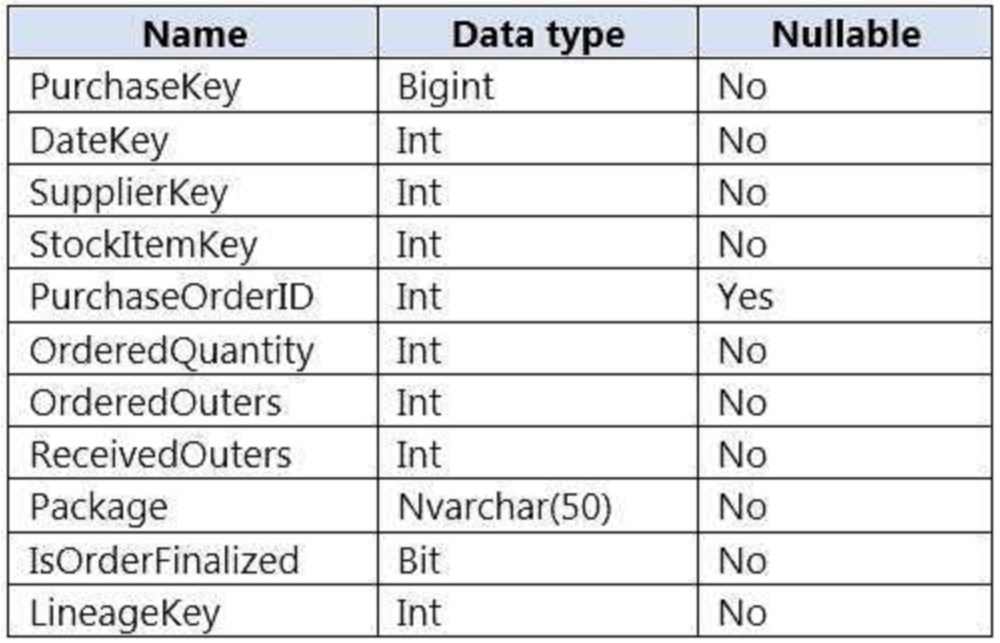
QUESTION 21
You plan to implement an Azure Data Lake Gen 2 storage account. You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs. Which type of replication should you use for the storage account?

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 22
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 23
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 24
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 25

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 26

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 27

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 28
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 29
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 30
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 31
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

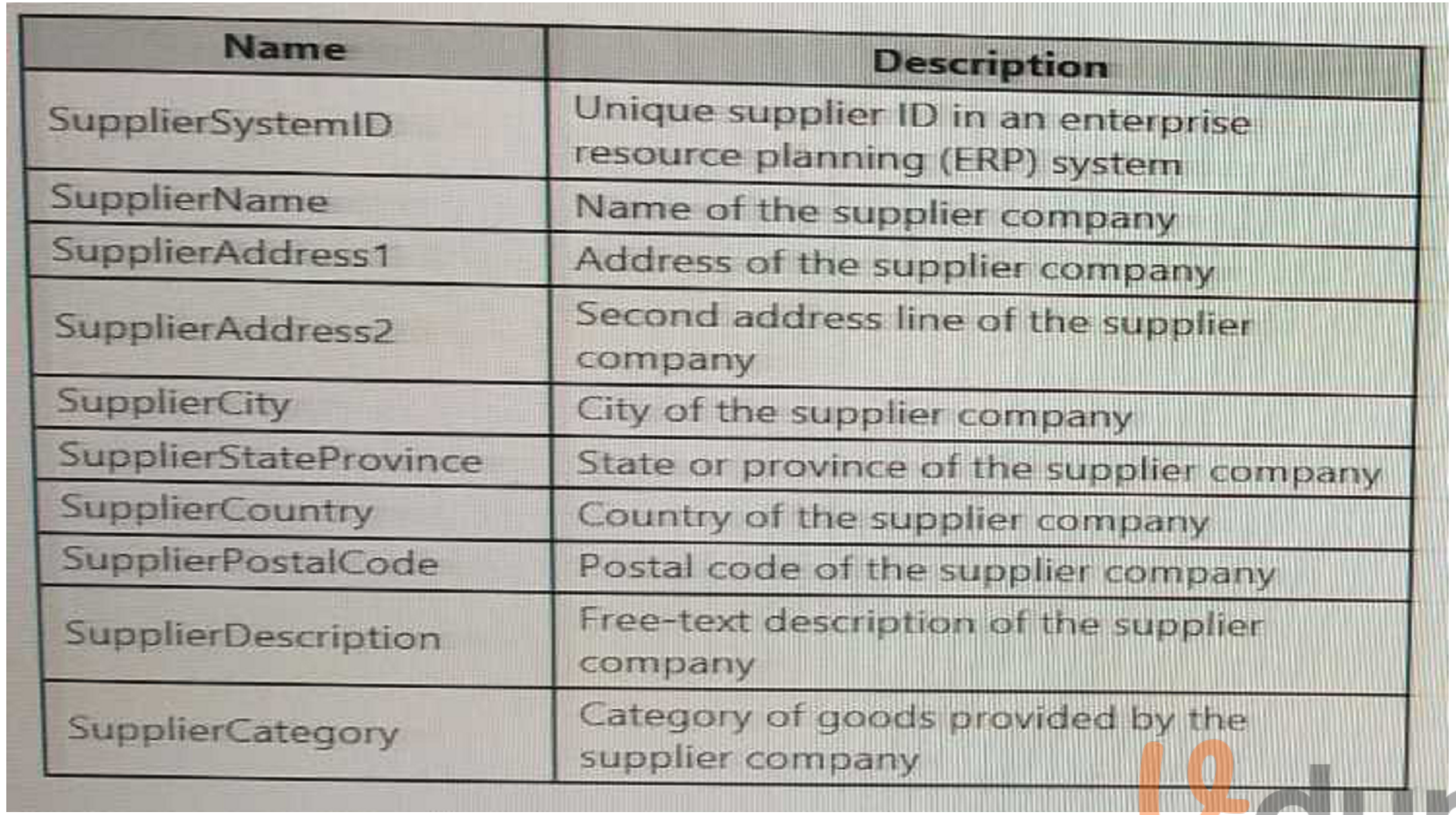
QUESTION 32
You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a partitioned fact table named dbo.Sales and a staging table named stg.Sales that has the matching table and partition definitions. You need to overwrite the content of the first partition in dbo.Sales with 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 33

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 34

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?

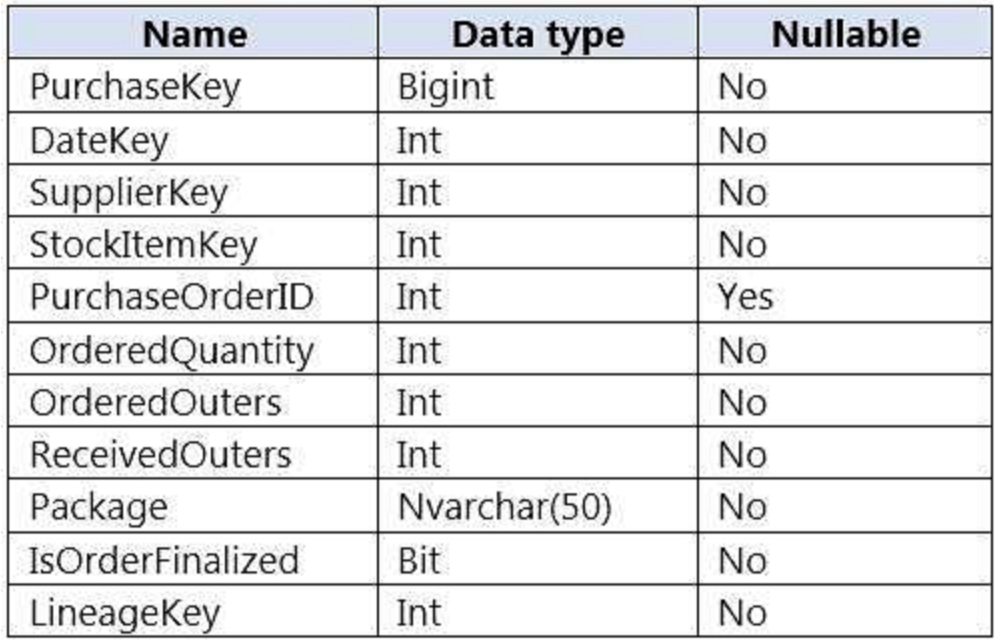
A. 40
B. 240

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 35
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

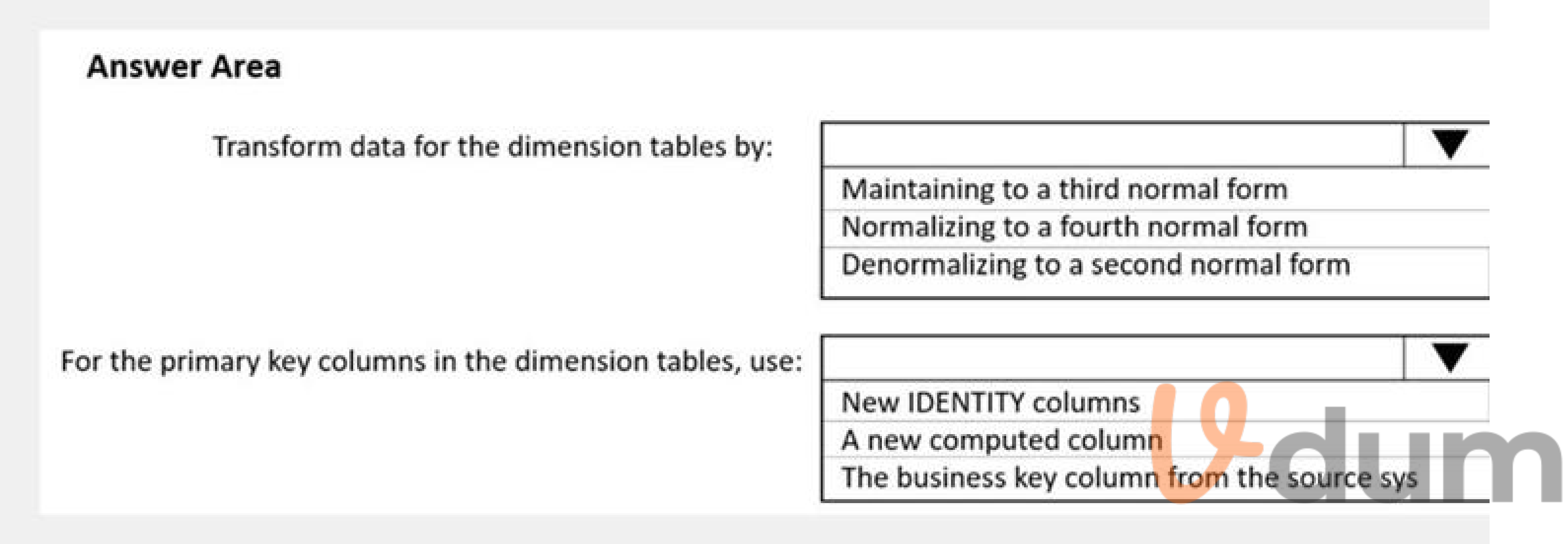
D. hash-distributed on DateKey

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 36
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:

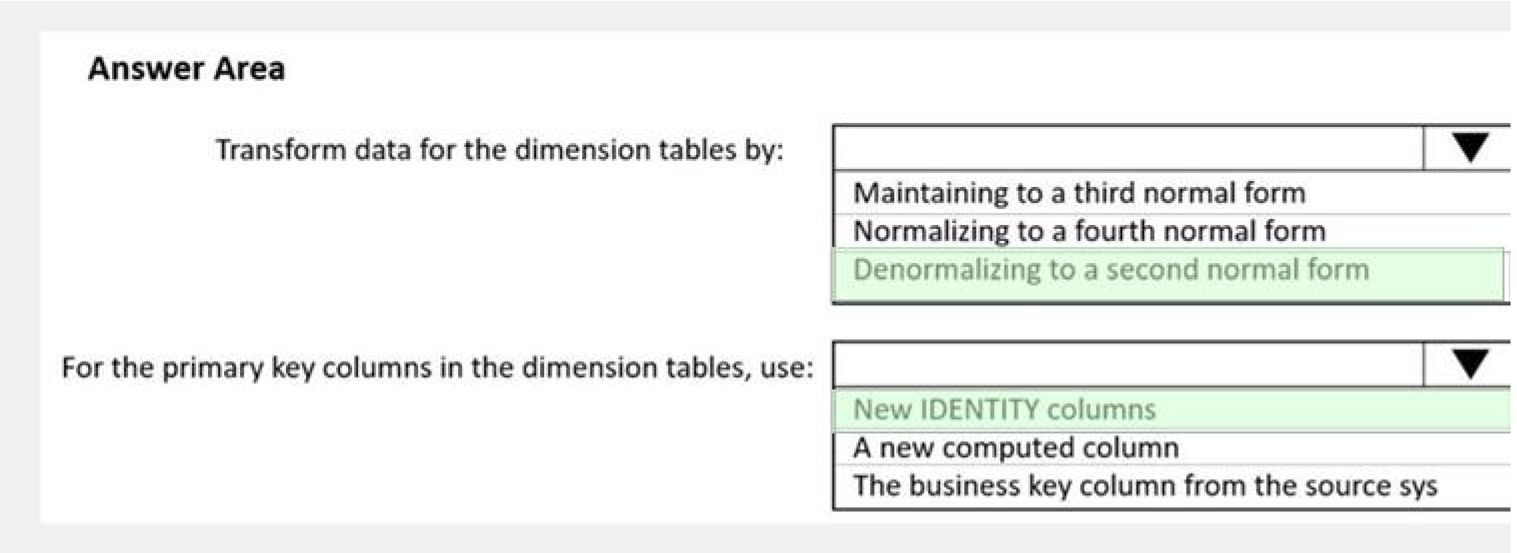


Answer Area

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:



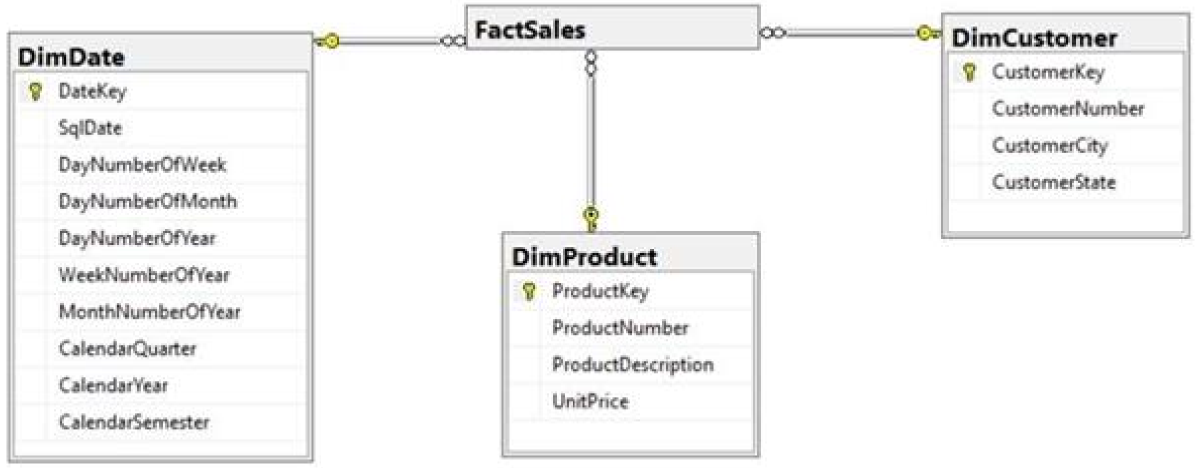
Answer Area

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

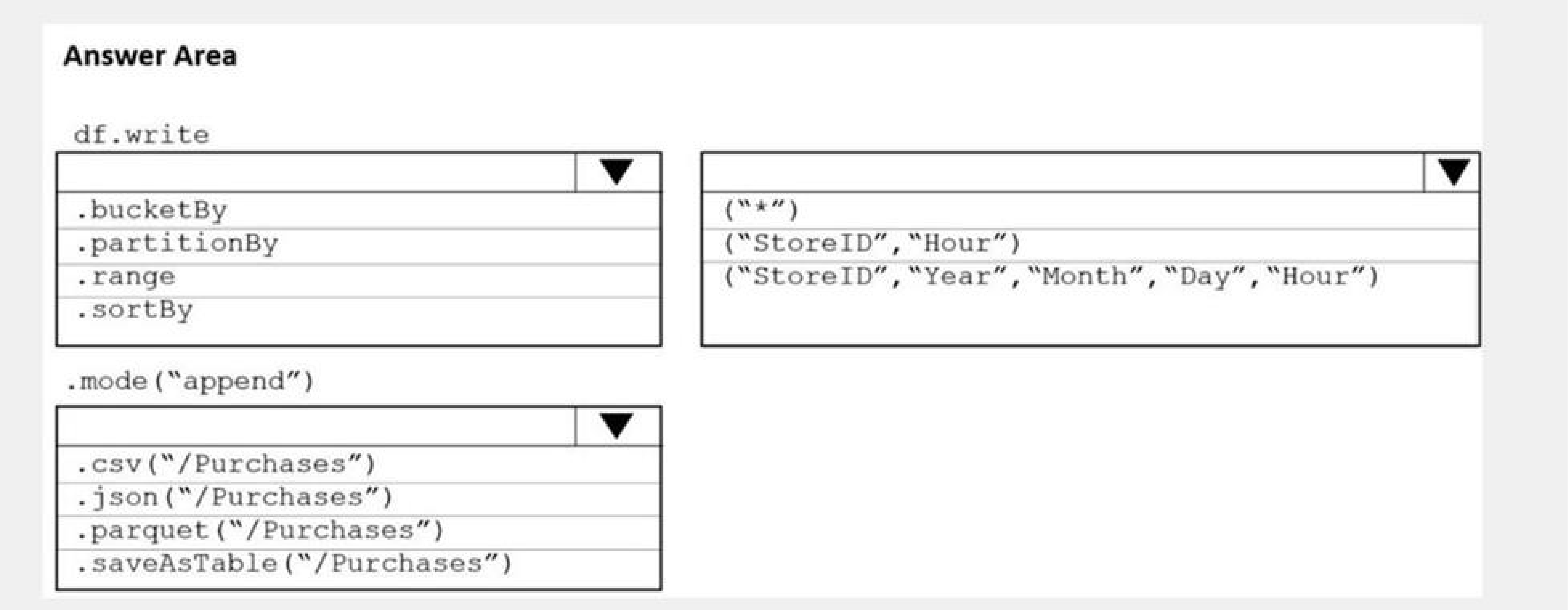
is a surrogate key generated to ensure it remains unique over time.
Example:



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 37
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

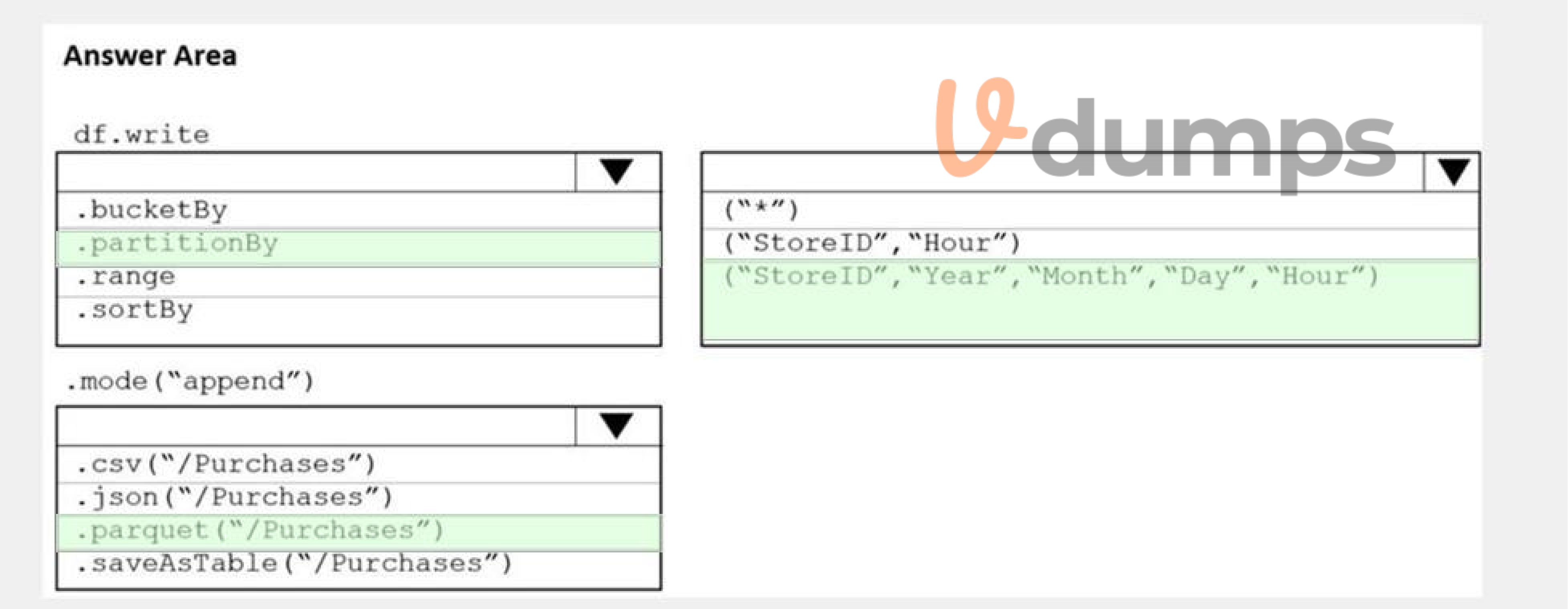
- .bucketBy
- .partitionBy
- .range
- .sortBy

- ("*")
- ("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.

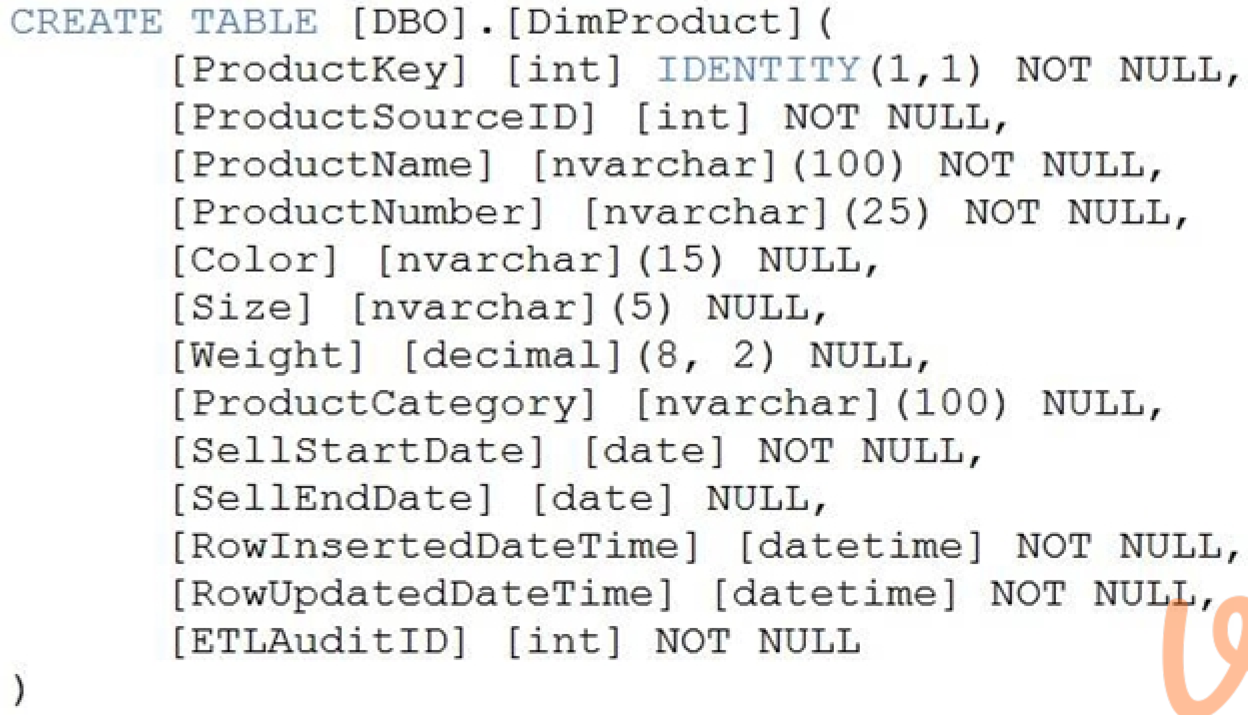
Example:

```
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

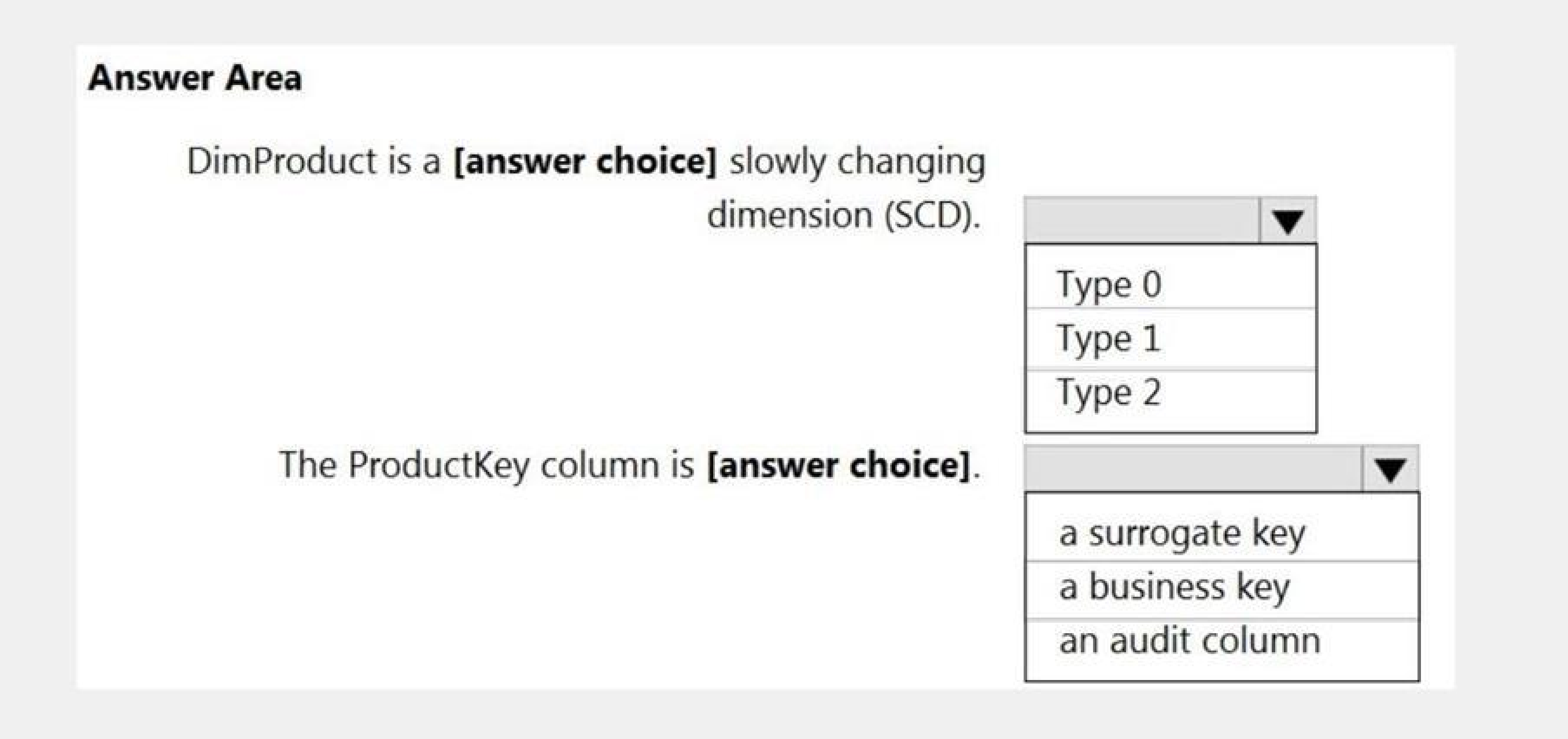
QUESTION 38
HOTSPOT
You are creating dimensions for a data warehouse in an Azure Synapse Analytics dedicated SQL pool.
You create a table by using the Transact-SQL statement shown in the following exhibit.



```
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

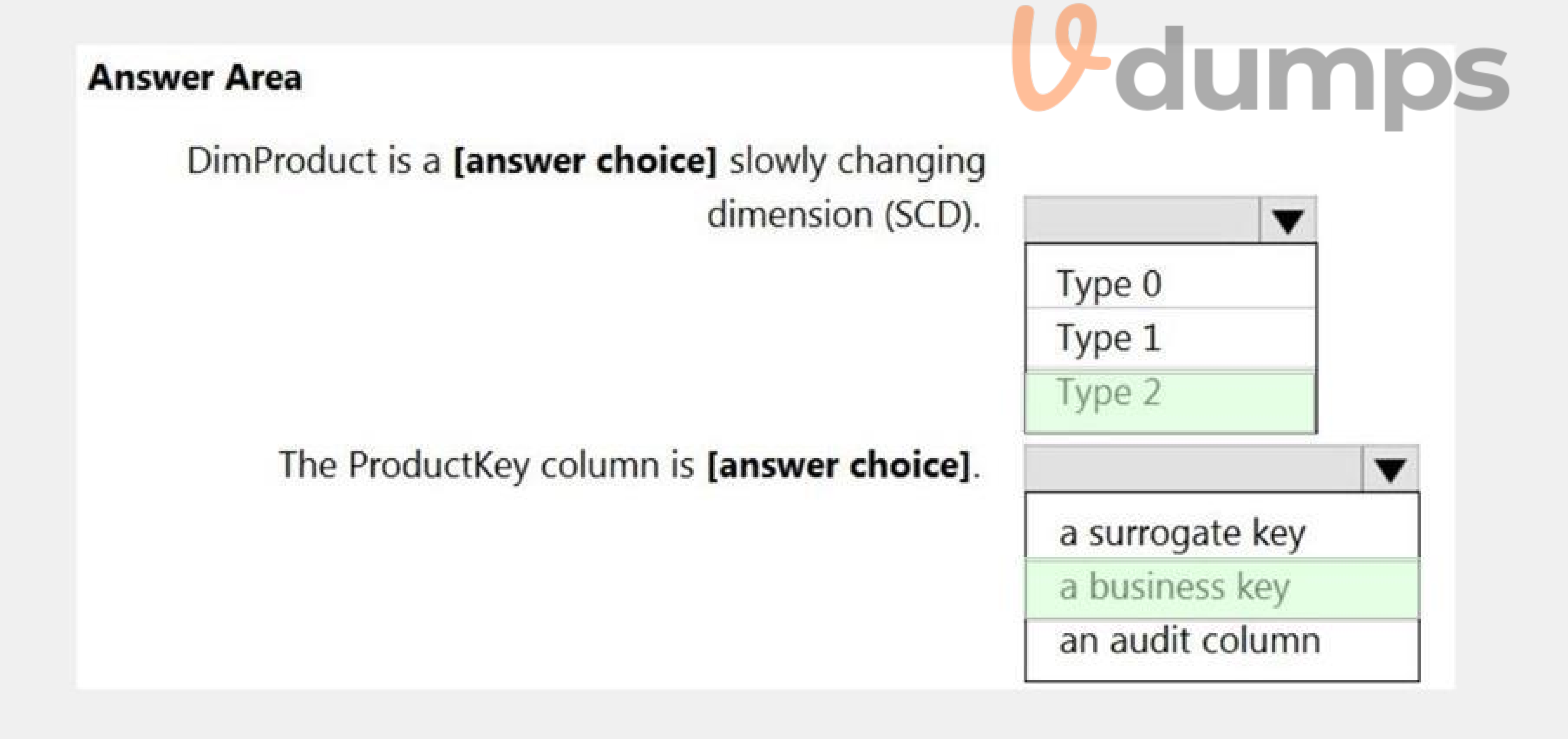
DimProduct is a **[answer choice]** slowly changing dimension (SCD).

- 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]**.

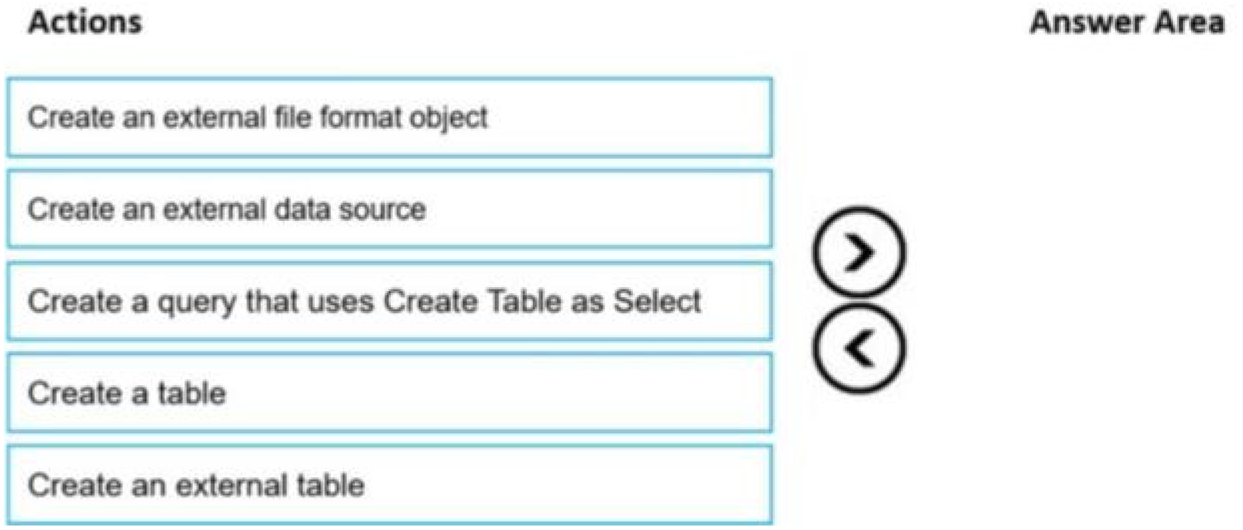
- a surrogate key
- [✓ a business key]
- an audit column

Section:
Explanation:

QUESTION 39

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:

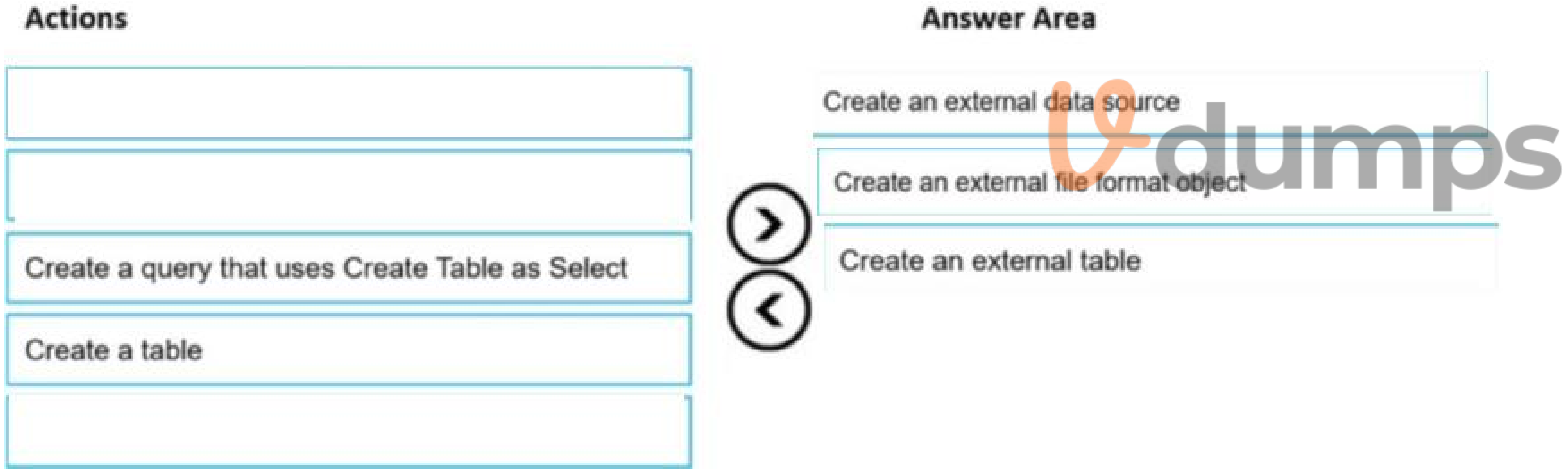


Actions

- 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

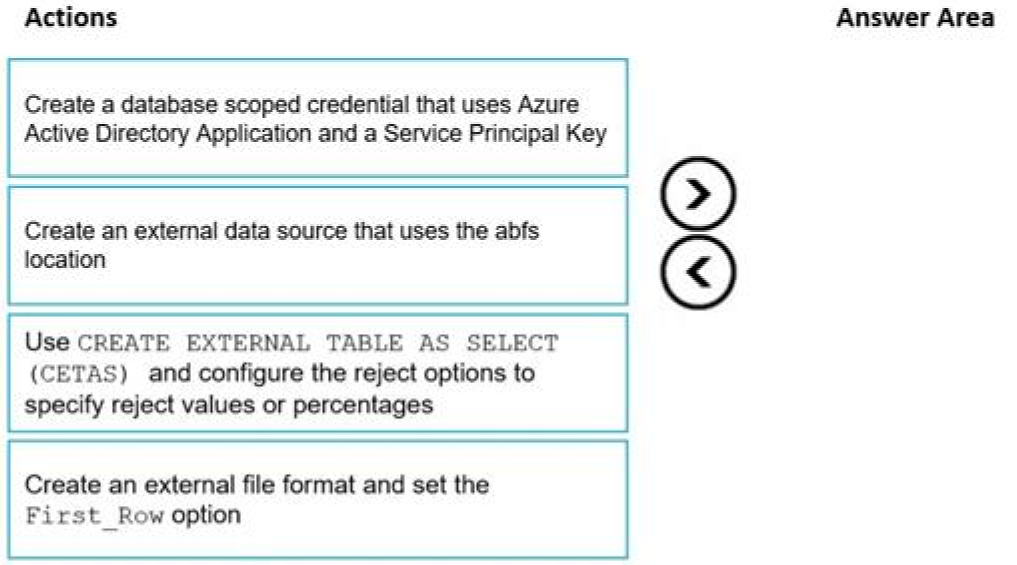
- Create an external data source
- Create an external file format object
- Create an external table

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 40
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:

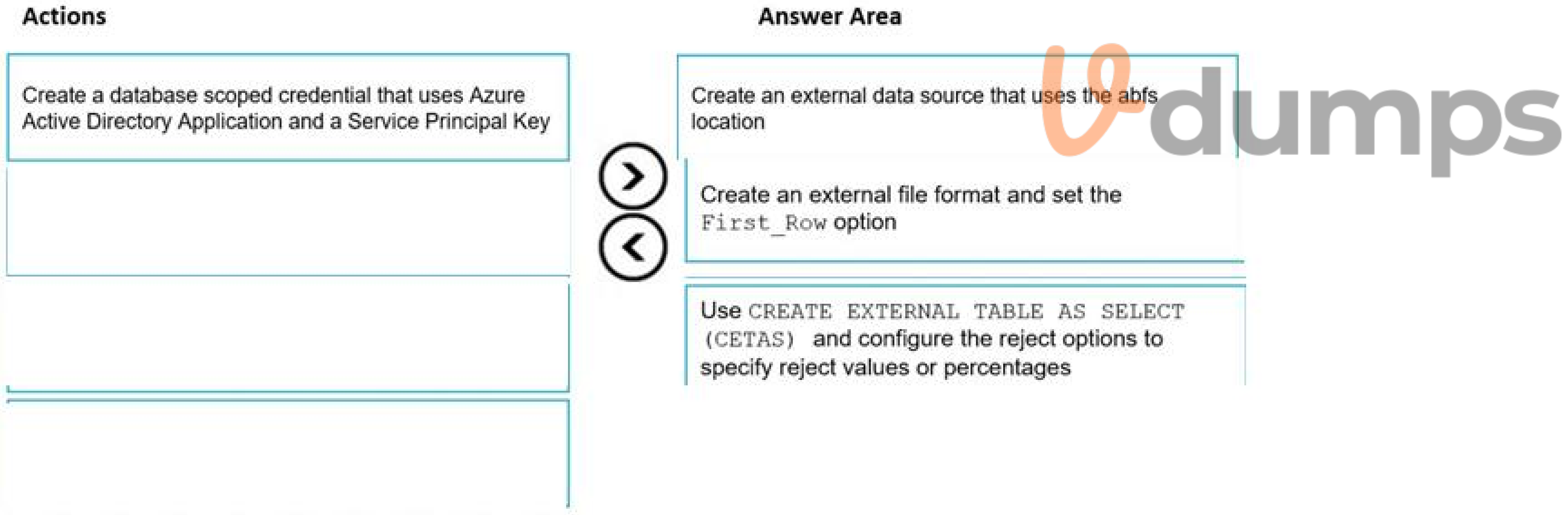


Actions

- 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

Answer Area

Correct Answer:



Actions

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

Answer Area

- 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 41

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:

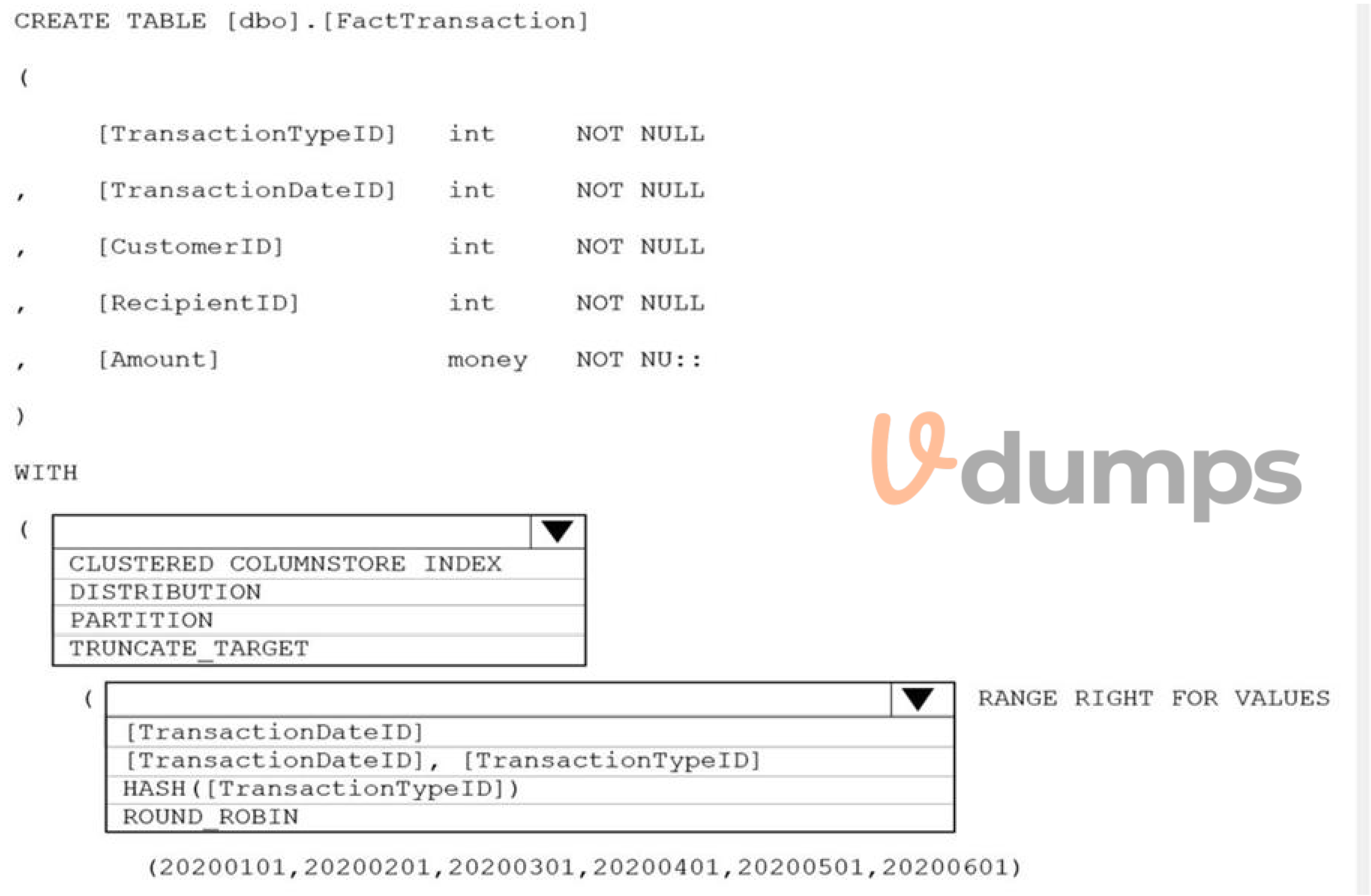
Minimizes the processing time to delete data that is older than 10 years

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.

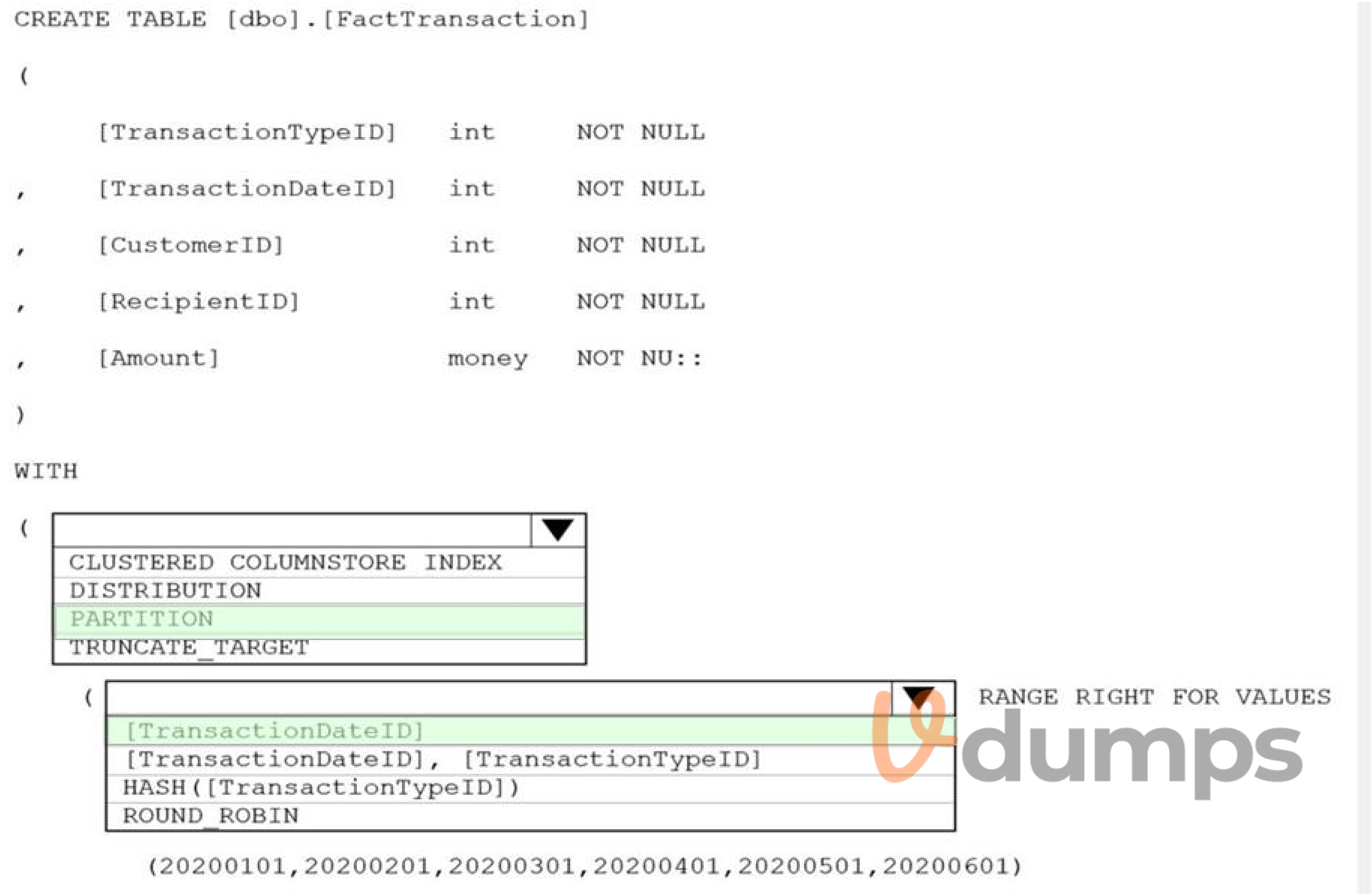
NOTE: Each correct selection is worth one point.

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
(   [                                ▼]
    CLUSTERED COLUMNSTORE INDEX
    DISTRIBUTION
    PARTITION
    TRUNCATE_TARGET
      (  [                                          ▼]  RANGE RIGHT FOR VALUES
         [TransactionDateID]
         [TransactionDateID], [TransactionTypeID]
         HASH([TransactionTypeID])
         ROUND_ROBIN
         (20200101,20200201,20200301,20200401,20200501,20200601)
```

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.

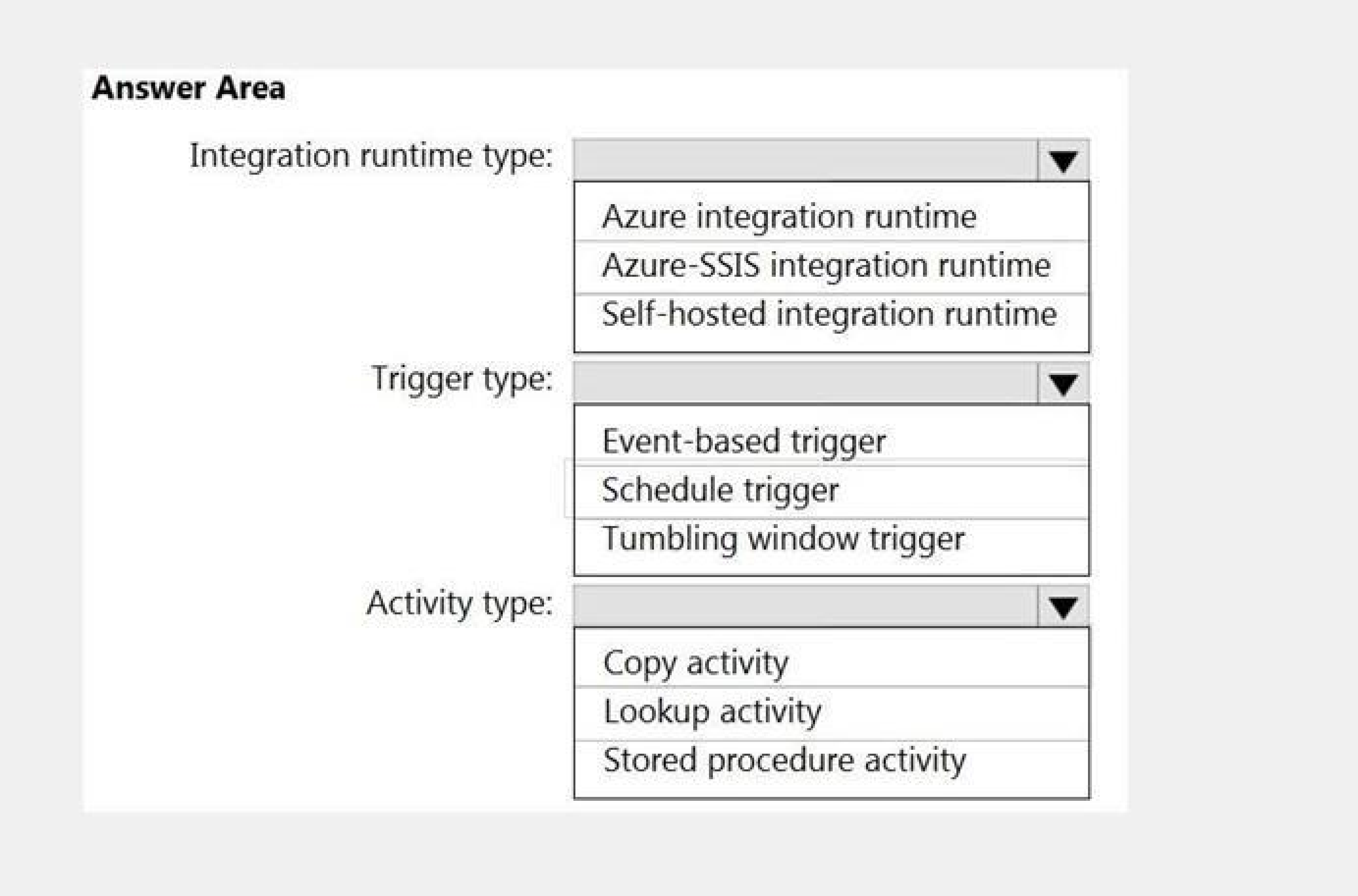
Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table. Product data, including product ID, name, and category, comes from Salesforce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

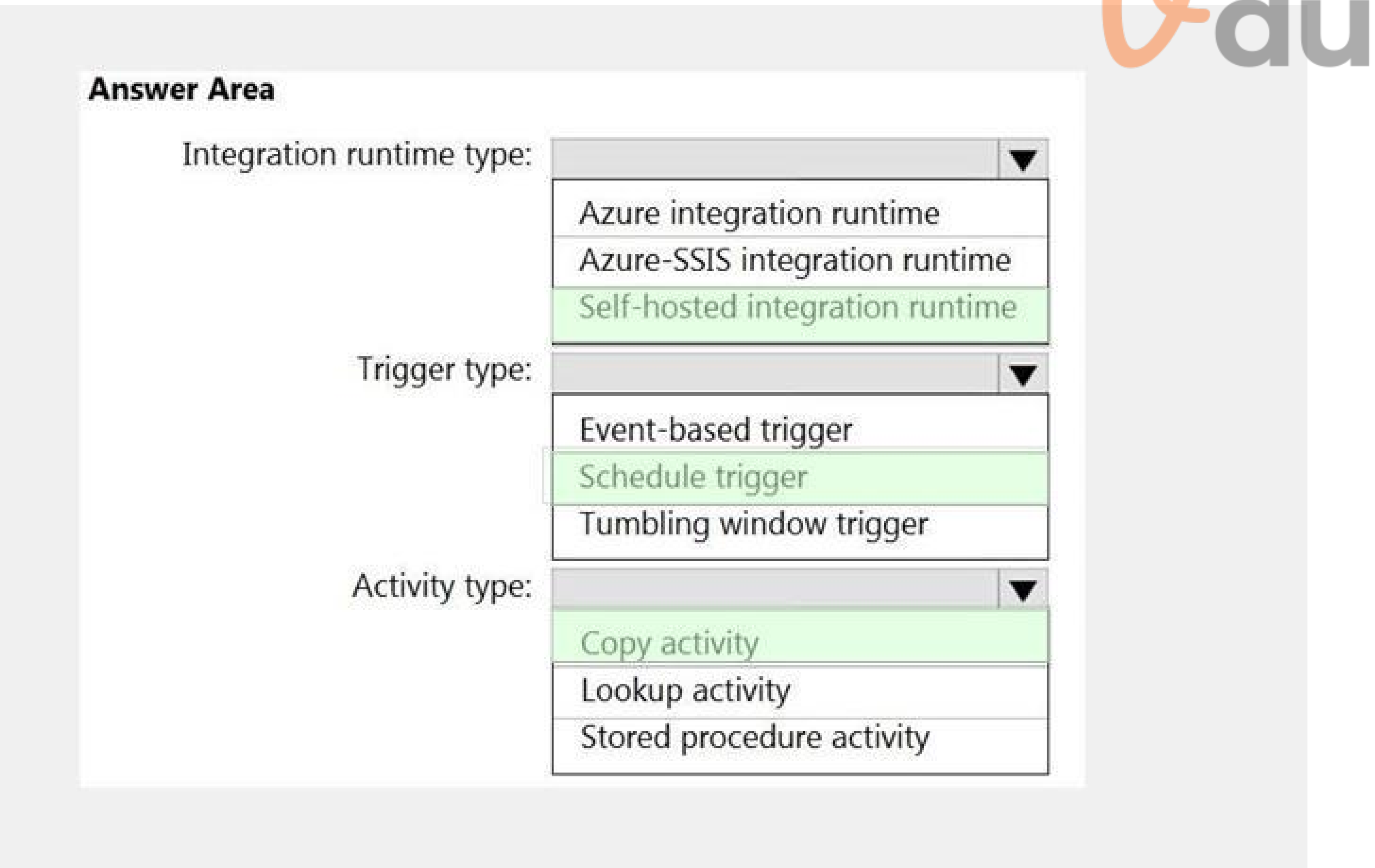
Trigger type: ▼

- 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.

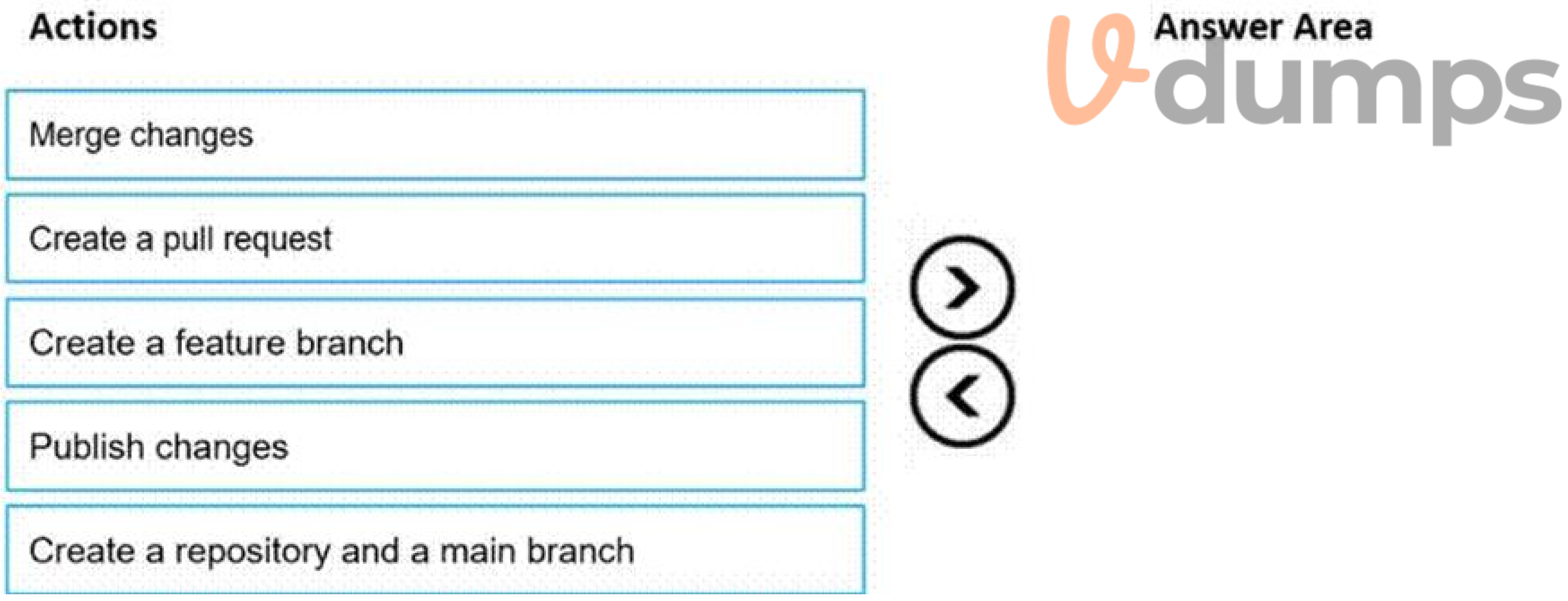
QUESTION 1

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:

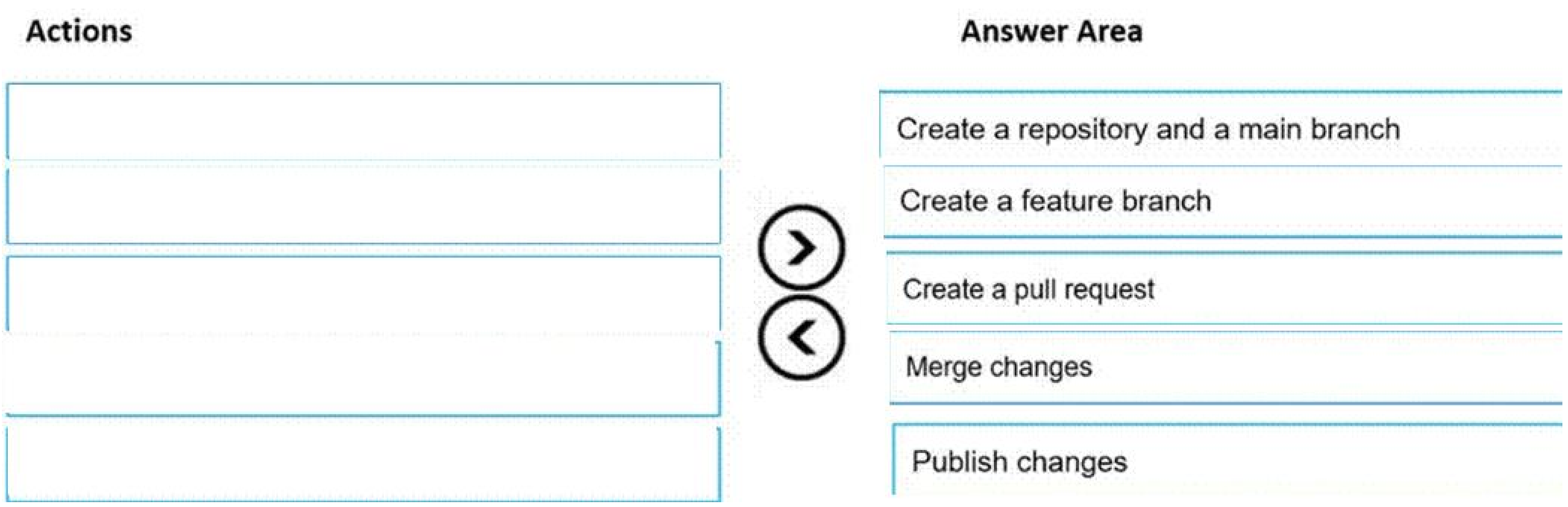


Actions

- 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.

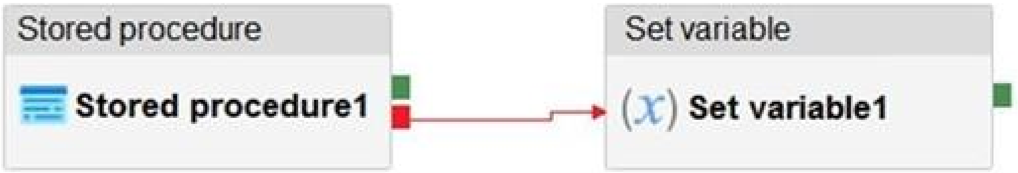
Step 5: Publish changes

Reference:

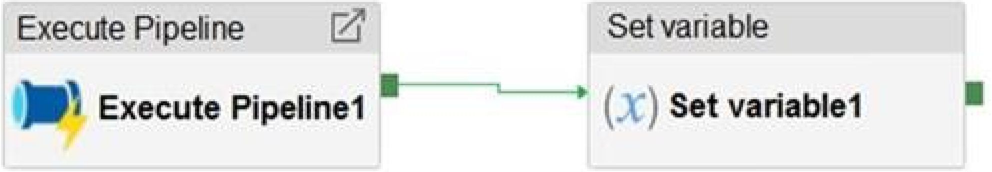
https://docs.microsoft.com/en-us/azure/devops/pipelines/repos/pipeline-options-for-git

03 - Design and develop data processing

QUESTION 1
You have an Azure Data Factory instance that contains two pipelines named Pipeline1 and Pipeline2.
Pipeline1 has the activities shown in the following exhibit.



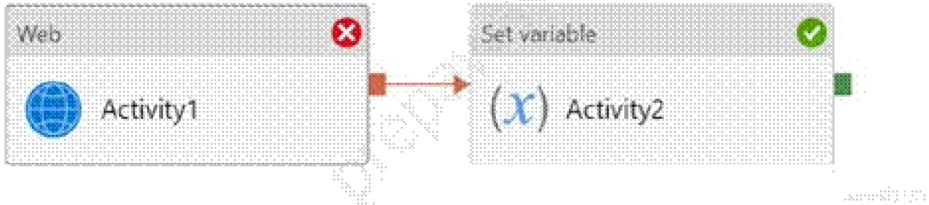
Pipeline2 has the activities shown in the following exhibit.



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

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

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



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

QUESTION 2
You have an Azure Data Factory that contains 10 pipelines.
You need to label each pipeline with its main purpose of either ingest, transform, or load. The labels must be available for grouping and filtering when using the monitoring experience in Data Factory. What should you add to each pipeline?

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

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

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

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

D. event

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

QUESTION 4
You have two Azure Data Factory instances named ADFdev and ADFprod. ADFdev connects to an Azure DevOps Git repository. You publish changes from the main branch of the Git repository to ADFdev. You need to deploy the artifacts from ADFdev to ADFprod.
What should you do first?

A. From ADFdev, modify the Git configuration.
B. From ADFdev, create a linked service.
C. From Azure DevOps, create a release pipeline.
D. From Azure DevOps, update the main branch.

Correct Answer: C
Section:
Explanation:
In Azure Data Factory, continuous integration and delivery (CI/CD) means moving Data Factory pipelines from one environment (development, test, production) to another. Note: The following is a guide for setting up an Azure Pipelines release that automates the deployment of a data factory to multiple environments. In Azure DevOps, open the project that's configured with your data factory. On the left side of the page, select Pipelines, and then select Releases. Select New pipeline, or, if you have existing pipelines, select New and then New release pipeline. In the Stage name box, enter the name of your environment. Select Add artifact, and then select the git repository configured with your development data factory. Select the publish branch of the repository for the Default branch. By default, this publish branch is adf_publish. Select the Empty job template.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment

QUESTION 5
You are developing a solution that will stream to Azure Stream Analytics. The solution will have both streaming data and reference data. Which input type should you use for the reference data?

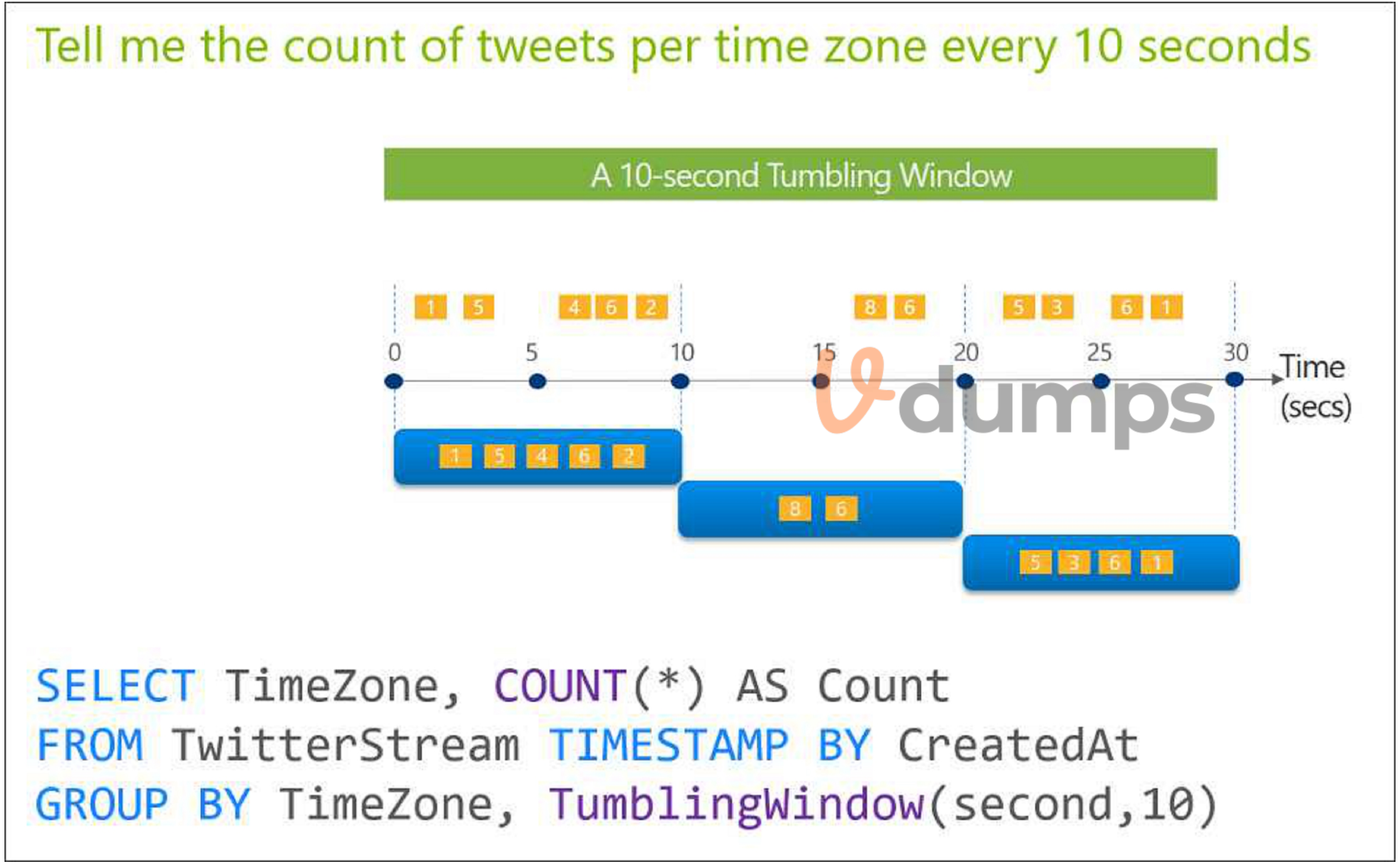
A. Azure Cosmos DB
B. Azure Blob storage
C. Azure IoT Hub
D. Azure Event Hubs

Correct Answer: B
Section:
Explanation:
Stream Analytics supports Azure Blob storage and Azure SQL Database as the storage layer for Reference Data.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data

QUESTION 6
You are designing an Azure Stream Analytics job to process incoming events from sensors in retail environments. You need to process the events to produce a running average of shopper counts during the previous 15 minutes, calculated at five-minute intervals. Which type of window should you use?

A. snapshot
B. tumbling
C. hopping
D. sliding

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



Tell me the count of tweets per time zone every 10 seconds

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

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

QUESTION 7
You are designing an Azure Databricks table. The table will ingest an average of 20 million streaming events per day. You need to persist the events in the table for use in incremental load pipeline jobs in Azure Databricks. The solution must minimize storage costs and incremental load times. What should you include in the solution?

A. Partition by DateTime fields.

B. Sink to Azure Queue storage.

C. Include a watermark column.

D. Use a JSON format for physical data storage.

Correct Answer: B
Section:
Explanation:
The Databricks ABS-AQS connector uses Azure Queue Storage (AQS) to provide an optimized file source that lets you find new files written to an Azure Blob storage (ABS) container without repeatedly listing all of the files.
This provides two major advantages:
Lower latency: no need to list nested directory structures on ABS, which is slow and resource intensive. Lower costs: no more costly LIST API requests made to ABS.
Reference:
https://docs.microsoft.com/en-us/azure/databricks/spark/latest/structured-streaming/aqs

QUESTION 8
You have an Azure Databricks workspace named workspace1 in the Standard pricing tier. You need to configure workspace1 to support autoscaling all-purpose clusters. The solution must meet the following requirements:
Automatically scale down workers when the cluster is underutilized for three minutes. Minimize the time it takes to scale to the maximum number of workers. Minimize costs. What should you do first?

A. Enable container services for workspace1.

B. Upgrade workspace1 to the Premium pricing tier.

C. Set Cluster Mode to High Concurrency.

D. Create a cluster policy in workspace1.

Correct Answer: B
Section:
Explanation:
For clusters running Databricks Runtime 6.4 and above, optimized autoscaling is used by all-purpose clusters in the Premium plan Optimized autoscaling:
Scales up from min to max in 2 steps.
Can scale down even if the cluster is not idle by looking at shuffle file state. Scales down based on a percentage of current nodes.
On job clusters, scales down if the cluster is underutilized over the last 40 seconds. On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds. The spark.databricks.aggressiveWindowDownS Spark configuration property specifies in seconds how often a cluster makes down-scaling decisions. Increasing the value causes a cluster to scale down more slowly. The maximum value is 600.
Note: Standard autoscaling
Starts with adding 8 nodes. Thereafter, scales up exponentially, but can take many steps to reach the max. You can customize the first step by setting the spark.databricks.autoscaling.standardFirstStepUp Spark configuration property. Scales down only when the cluster is completely idle and it has been underutilized for the last 10 minutes. Scales down exponentially, starting with 1 node.
Reference:
https://docs.databricks.com/clusters/configure.html

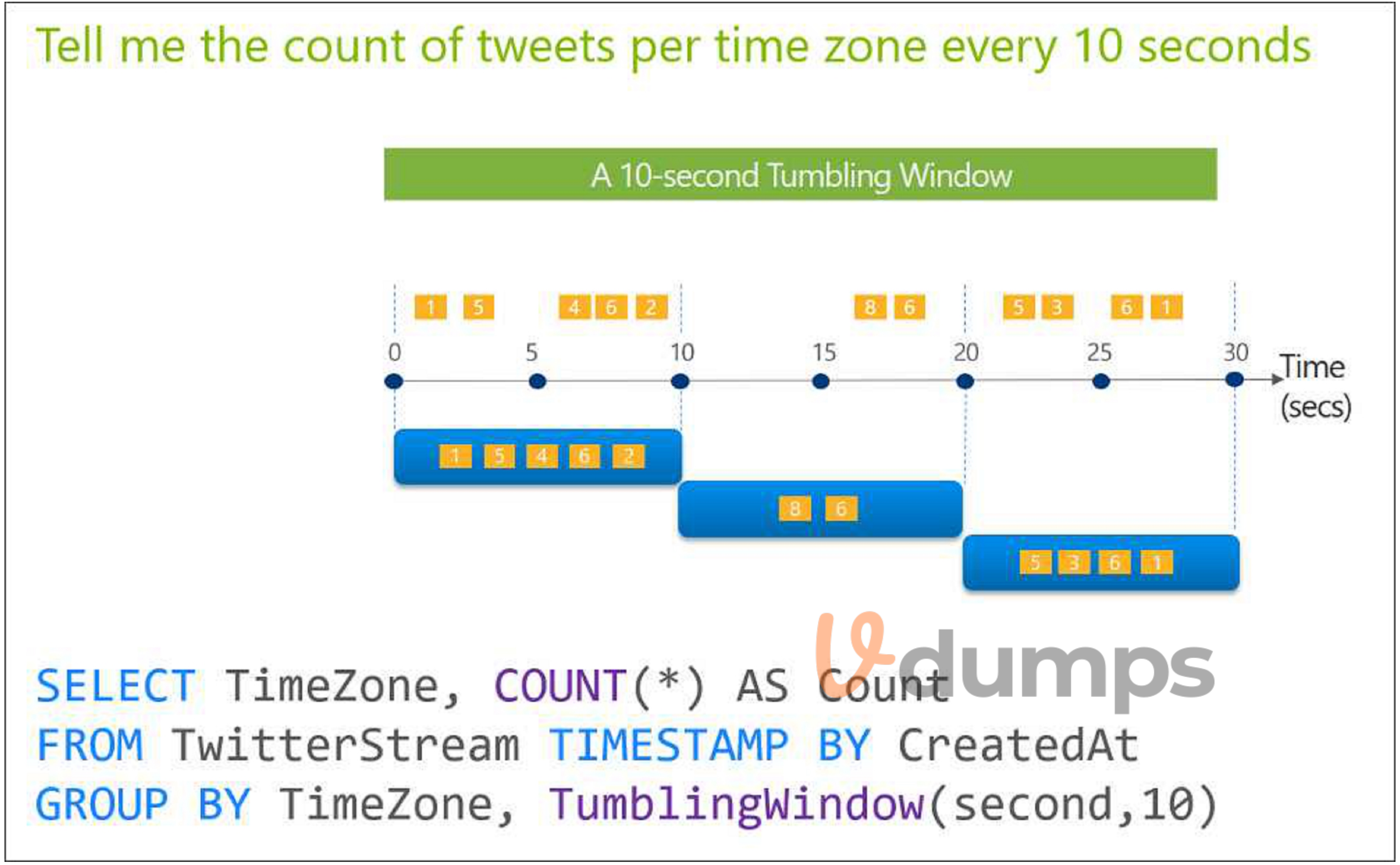
QUESTION 9
Note: This question-is part of a series of questions that present the same scenario. Each question-in the series contains a unique solution that might meet the stated goals. Some question-sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question-in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You are designing an Azure Stream Analytics solution that will analyze Twitter data. You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once. Solution: You use a tumbling window, and you set the window size to 10 seconds. Does this meet the goal?

A. Yes

B. No

Correct Answer: A
Section:
Explanation:

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



Tell me the count of tweets per time zone every 10 seconds

A 10-second Tumbling Window

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

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

QUESTION 10

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

After you answer a question-in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You are designing an Azure Stream Analytics solution that will analyze Twitter data. You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once. Solution: You use a session window that uses a timeout size of 10 seconds. Does this meet the goal?

A. Yes

B. No

Correct Answer: A
Section:

QUESTION 11

You use Azure Stream Analytics to receive data from Azure Event Hubs and to output the data to an Azure Blob Storage account. You need to output the count of records received from the last five minutes every minute. Which windowing function should you use?

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

Correct Answer: D
Section:
Explanation:

QUESTION 12
You plan to perform batch processing in Azure Databricks once daily. Which type of Databricks cluster should you use?

A. High Concurrency
B. interactive
C. automated

Correct Answer: C
Section:
Explanation:
Automated Databricks clusters are the best for jobs and automated batch processing.
Reference:
https://docs.microsoft.com/en-us/azure/databricks/clusters/create

QUESTION 13
You have the following Azure Data Factory pipelines:
Ingest Data from System1
Ingest Data from System2
Populate Dimensions
Populate Facts
Ingest Data from System1 and Ingest Data from System2 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest Data from System2. Populate Facts must execute after Populate Dimensions pipeline. All the pipelines must execute every eight hours.
What should you do to schedule the pipelines for execution?

A. Add an event trigger to all four pipelines.
B. Add a schedule trigger to all four pipelines.
C. Create a patient pipeline that contains the four pipelines and use a schedule trigger.
D. Create a patient pipeline that contains the four pipelines and use an event trigger.

Correct Answer: C
Section:
Explanation:
Schedule trigger: A trigger that invokes a pipeline on a wall-clock schedule.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers

QUESTION 14
You are monitoring an Azure Stream Analytics job by using metrics in Azure. You discover that during the last 12 hours, the average watermark delay is consistently greater than the configured late arrival tolerance. What is a possible cause of this behavior?

A. Events whose application timestamp is earlier than their arrival time by more than five minutes arrive as inputs.
B. There are errors in the input data.
C. The late arrival policy causes events to be dropped.
D. The job lacks the resources to process the volume of incoming data.

Correct Answer: D
Section:
Explanation:
Watermark Delay indicates the delay of the streaming data processing job. There are a number of resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to:
Not enough processing resources in Stream Analytics to handle the volume of input events. To scale up resources, see Understand and adjust Streaming Units. Not enough throughput within the input event brokers, so they are throttled. For possible solutions, see Automatically scale up Azure Event Hubs throughput units. Output sinks are not provisioned with enough capacity, so they are throttled. The possible solutions vary widely based on the flavor of output service being used.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling

QUESTION 15
You are designing a statistical analysis solution that will use custom proprietary Python functions on near real-time data from Azure Event Hubs. You need to recommend which Azure service to use to perform the statistical analysis. The solution must minimize latency. What should you recommend?

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

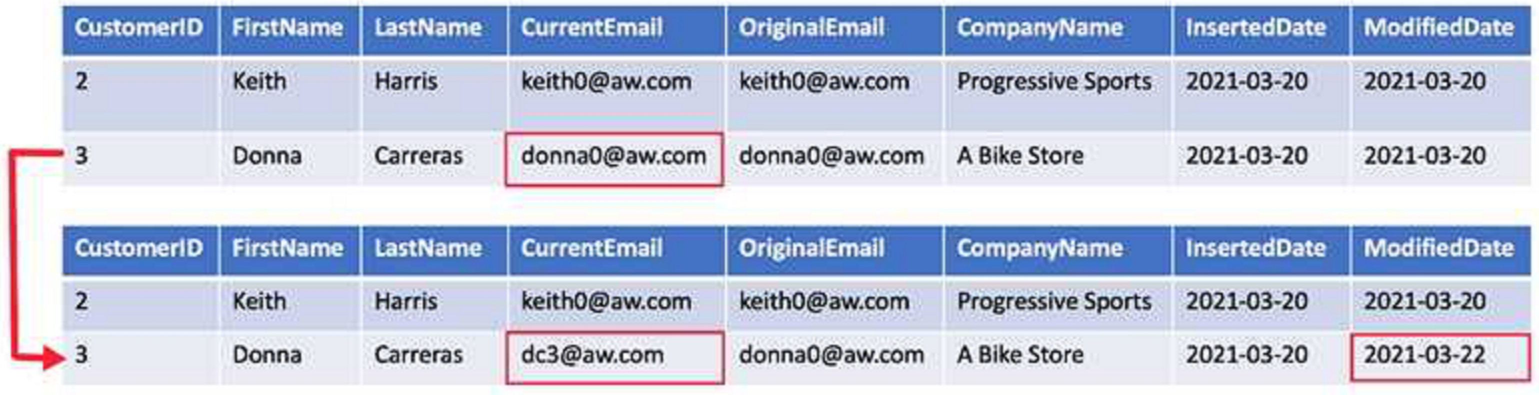
Correct Answer: C
Section:
Explanation:
Reference: https://docs.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics

QUESTION 16
You need to implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool.
You have a table that was created by using the following Transact-SQL statement. Which two columns should you add to the table? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. [EffectiveStartDate] [datetime] NOT NULL,
B. [CurrentProductCategory] [nvarchar] (100) NOT NULL,
C. [EffectiveEndDate] [datetime] NULL,
D. [ProductCategory] [nvarchar] (100) NOT NULL,
E. [OriginalProductCategory] [nvarchar] (100) NOT NULL,

Correct Answer: B, E
Section:
Explanation:
A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3

uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD.
This type of tracking may be used for one or two columns in a dimension table. It is not common to use it for many members of the same table. It is often used in combination with Type 1 or Type 2 members.



CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	donna0@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-20

CustomerID	FirstName	LastName	CurrentEmail	OriginalEmail	CompanyName	InsertedDate	ModifiedDate
2	Keith	Harris	keith0@aw.com	keith0@aw.com	Progressive Sports	2021-03-20	2021-03-20
3	Donna	Carreras	dc3@aw.com	donna0@aw.com	A Bike Store	2021-03-20	2021-03-22

Reference:
https://k21academy.com/microsoft-azure/azure-data-engineer-dp203-q-a-day-2-live-session-review/

QUESTION 17

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

A. Yes

B. No

Correct Answer: B
Section:
Explanation:
Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

QUESTION 18

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:
A workload for data engineers who will use Python and SQL.
A workload for jobs that will run notebooks that use Python, Scala, and SQL. A workload that data scientists will use to perform ad hoc analysis in Scala and R.
The enterprise architecture team at your company identifies the following standards for Databricks environments:
The data engineers must share a cluster.
The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster. All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.
You need to create the Databricks clusters for the workloads.
Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.
Does this meet the goal?

A. Yes

B. No

Correct Answer: B
Section:
Explanation:
We would need a High Concurrency cluster for the jobs.
Note:
Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL. A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.
Reference: https://docs.azuredatabricks.net/clusters/configure.html

QUESTION 19

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:
A workload for data engineers who will use Python and SQL.
A workload for jobs that will run notebooks that use Python, Scala, and SQL. A workload that data scientists will use to perform ad hoc analysis in Scala and R.
The enterprise architecture team at your company identifies the following standards for Databricks environments:
The data engineers must share a cluster.
The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster. All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.
You need to create the Databricks clusters for the workloads.
Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a High Concurrency cluster for the jobs. Does this meet the goal?

A. Yes
B. No

Correct Answer: A
Section:
Explanation:
We need a High Concurrency cluster for the data engineers and the jobs. Note: Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL. A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.
Reference: https://docs.azuredatabricks.net/clusters/configure.html

QUESTION 20

A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU).
You need to optimize performance for the Azure Stream Analytics job. Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. Implement event ordering.
B. Implement Azure Stream Analytics user-defined functions (UDF).
C. Implement query parallelization by partitioning the data output.
D. Scale the SU count for the job up.
E. Scale the SU count for the job down.
F. Implement query parallelization by partitioning the data input.

Correct Answer: D, F
Section:
Explanation:
D: Scale out the query by allowing the system to process each input partition separately. F: A Stream Analytics job definition includes inputs, a query, and output. Inputs are where the job reads the data stream from.

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

QUESTION 21
You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container. Which resource provider should you enable?

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

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

QUESTION 22
You plan to perform batch processing in Azure Databricks once daily. Which type of Databricks cluster should you use?

A. High Concurrency
B. automated
C. interactive

Correct Answer: C
Section:
Explanation:
Azure Databricks has two types of clusters: interactive and automated. You use interactive clusters to analyze data collaboratively with interactive notebooks. You use automated clusters to run fast and robust automated jobs.
Example: Scheduled batch workloads (data engineers running ETL jobs) This scenario involves running batch job JARs and notebooks on a regular cadence through the Databricks platform. The suggested best practice is to launch a new cluster for each run of critical jobs. This helps avoid any issues (failures, missing SLA, and so on) due to an existing workload (noisy neighbor) on a shared cluster.
Reference: https://docs.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-scheduled-batch-workloads-data-engineers-running-etl-jobs

QUESTION 23
Note: This question-is part of a series of questions that present the same scenario. Each question-in the series contains a unique solution that might meet the stated goals. Some question-sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question-in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You are designing an Azure Stream Analytics solution that will analyze Twitter data. You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once. Solution: You use a hopping window that uses a hop size of 5 seconds and a window size 10 seconds. Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section:
Explanation:
Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.
Reference:

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

QUESTION 24

You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytic dedicated SQL pool. The CSV file contains three columns named username, comment, and date.

The data flow already contains the following:

A source transformation.

A Derived Column transformation to set the appropriate types of data. A sink transformation to land the data in the pool. You need to ensure that the data flow meets the following requirements:

All valid rows must be written to the destination table.

Truncation errors in the comment column must be avoided proactively. Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage. Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. To the data flow, add a sink transformation to write the rows to a file in blob storage.

B. To the data flow, add a Conditional Split transformation to separate the rows that will cause truncation errors.

C. To the data flow, add a filter transformation to filter out rows that will cause truncation errors.

D. Add a select transformation to select only the rows that will cause truncation errors.

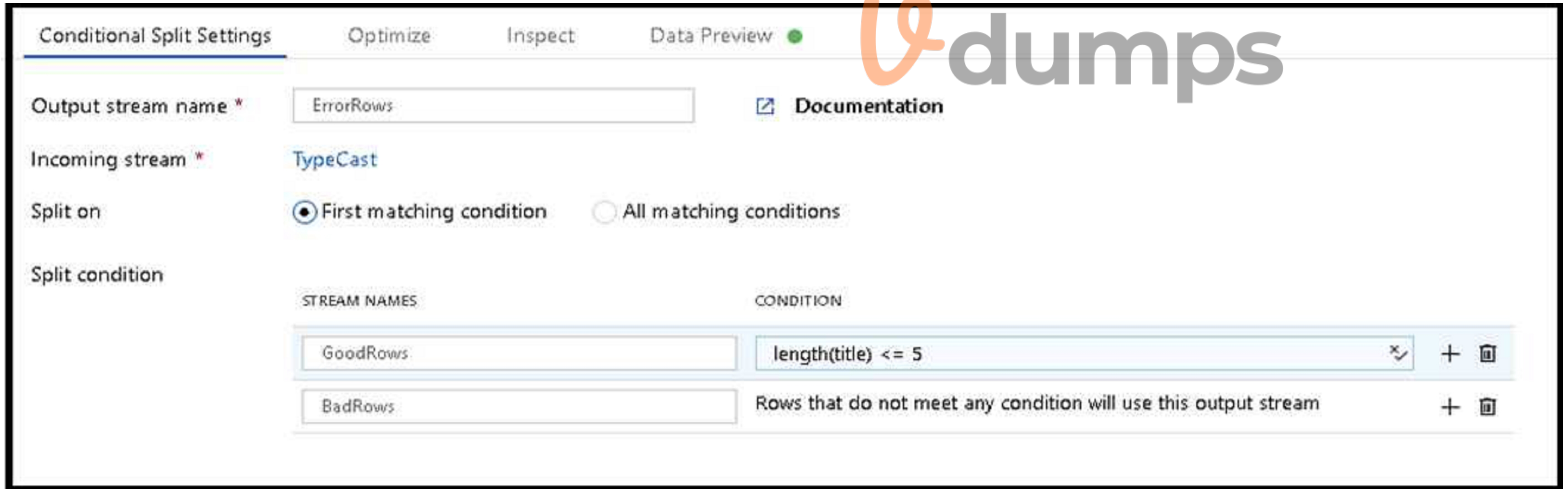
Correct Answer: A, B

Section:

Explanation:

B: Example:

1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.



Output stream name *: ErrorRows

Incoming stream *: TypeCast

Split on: ● First matching condition ○ All matching conditions

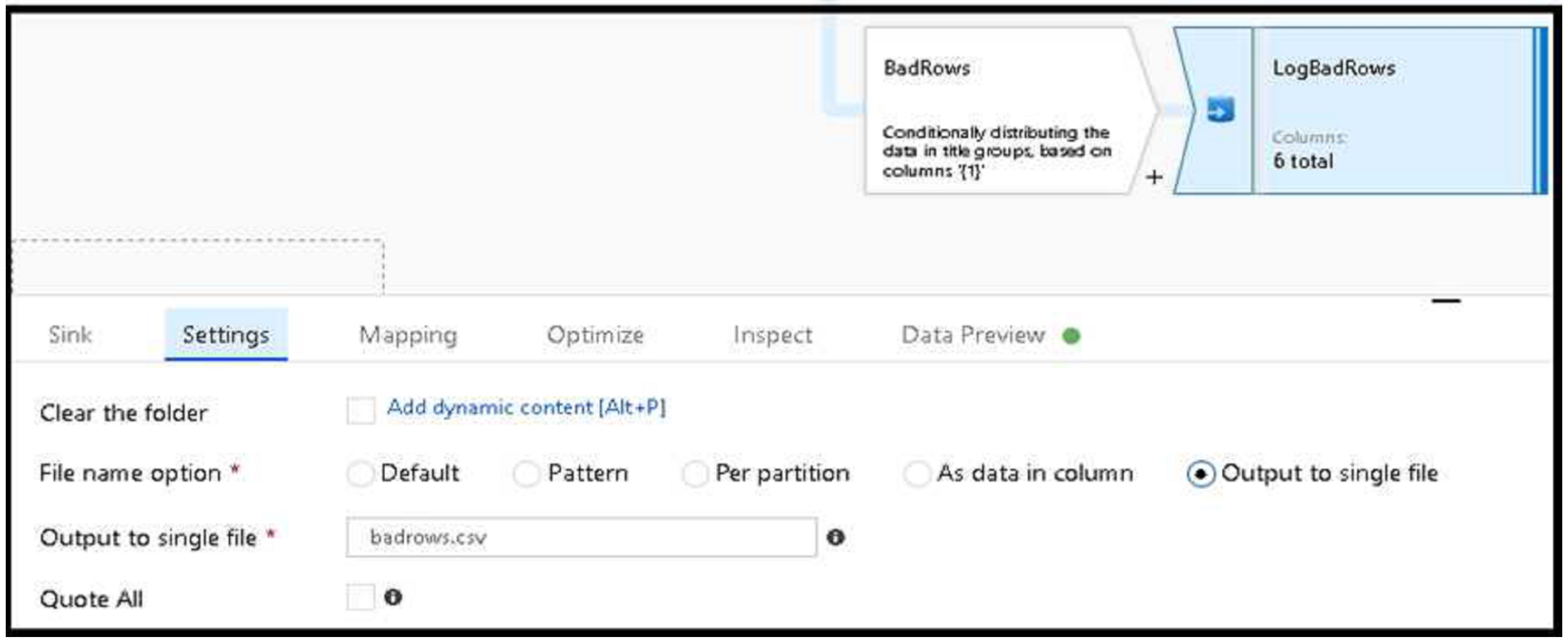
Split condition

STREAM NAMES	CONDITION
GoodRows	length(title) <= 5
BadRows	Rows that do not meet any condition will use this output stream

2. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.

A:

3. Now we need to log the rows that failed. Add a sink transformation to the BadRows stream for logging. Here, we'll "auto-map" all of the fields so that we have logging of the complete transaction record. This is a text-delimited CSV file output to a single file in Blob Storage. We'll call the log file "badrows.csv".



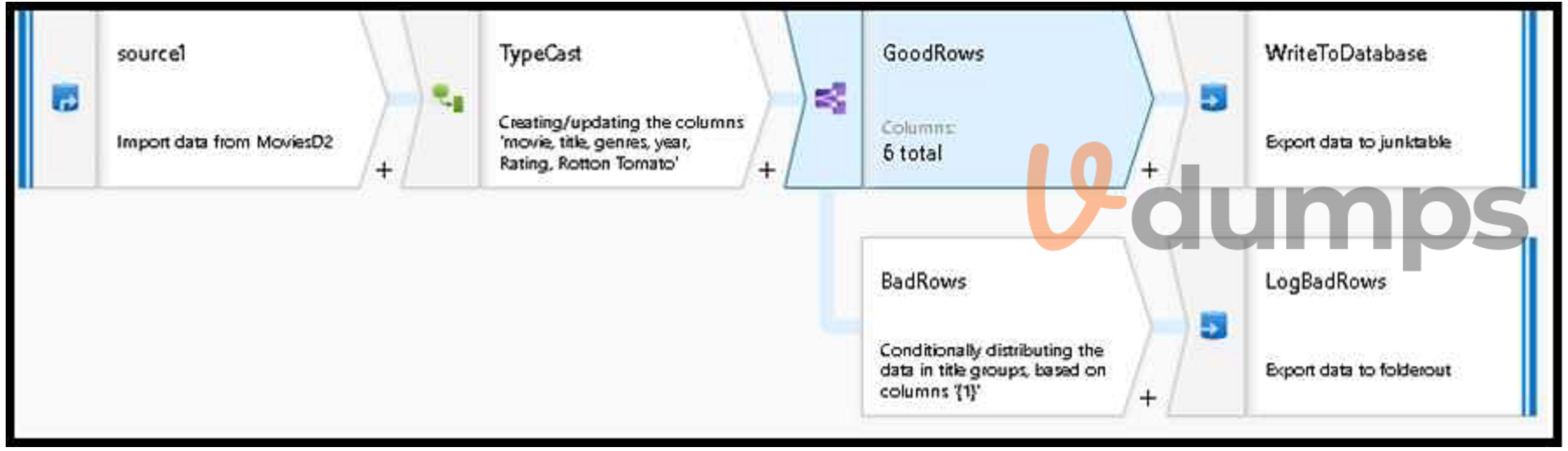
Clear the folder □ Add dynamic content [Alt+P]

File name option * ○ Default ○ Pattern ○ Per partition ○ As data in column ● Output to single file

Output to single file * badrows.csv

Quote All □

4. The completed data flow is shown below. We are now able to split off error rows to avoid the SQL truncation errors and put those entries into a log file. Meanwhile, successful rows can continue to write to our target database.


source1
Import data from MoviesD2
TypeCast
Creating/updating the columns 'movie, title, genres, year, Rating, Rotton Tomato'
GoodRows
Columns:
6 total
WriteToDatabase
Export data to junktable
BadRows
Conditionally distributing the data in title groups, based on columns '{1}'
LogBadRows
Export data to folderout

Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-data-flow-error-rows

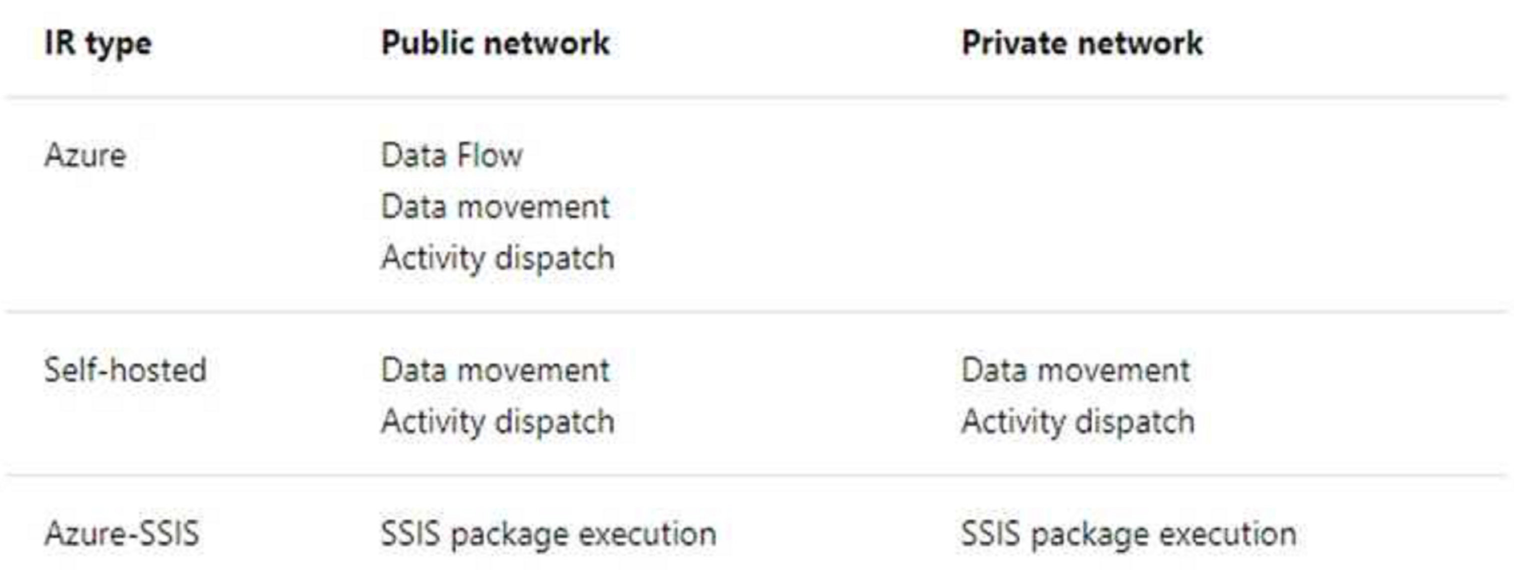
QUESTION 25

You have an Azure Storage account and a data warehouse in Azure Synapse Analytics in the UK South region. You need to copy blob data from the storage account to the data warehouse by using Azure Data Factory. The solution must meet the following requirements:

Ensure that the data remains in the UK South region at all times. Minimize administrative effort. Which type of integration runtime should you use?

A. Azure integration runtime
B. Azure-SSIS integration runtime
C. Self-hosted integration runtime

Correct Answer: A
Section:
Explanation:



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

Incorrect Answers:
C: Self-hosted integration runtime is to be used On-premises.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

QUESTION 26
You have an Azure Stream Analytics job that receives clickstream data from an Azure event hub. You need to define a query in the Stream Analytics job. The query must meet the following requirements:
Count the number of clicks within each 10-second window based on the country of a visitor. Ensure that each click is NOT counted more than once. How should you define the Query?

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

Correct Answer: B
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. Example:
Incorrect Answers:
A: Sliding windows, unlike Tumbling or Hopping windows, output events only for points in time when the content of the window actually changes. In other words, when an event enters or exits the window. Every window has at least one event, like in the case of Hopping windows, events can belong to more than one sliding window. C: Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap, so events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size. D: Session windows group events that arrive at similar times, filtering out periods of time where there is no data.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

QUESTION 27

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 copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script. Does this meet the goal?

A. Yes

B. No

Correct Answer: B

Section:

Explanation:

If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity with your own data processing logic and use the activity in the pipeline. Note: You can use data transformation activities in Azure Data Factory and Synapse pipelines to transform and process your raw data into predictions and insights at scale.

Reference:

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

QUESTION 28

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

After you answer a question-in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

A workload for data engineers who will use Python and SQL.

A workload for jobs that will run notebooks that use Python, Scala, and SQL. A workload that data scientists will use to perform ad hoc analysis in Scala and R. The enterprise architecture team at your company identifies the following standards for Databricks environments:

The data engineers must share a cluster.

The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster. All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists. You need to create the Databricks clusters for the workloads. Solution: You create a High Concurrency cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs. Does this meet the goal?

A. Yes

B. No

Correct Answer: B

Section:

Explanation:

Need a High Concurrency cluster for the jobs.

Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL. A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

Reference:

https://docs.azuredatabricks.net/clusters/configure.html

QUESTION 29

You are designing an Azure Databricks cluster that runs user-defined local processes. You need to recommend a cluster configuration that meets the following requirements:

Minimize query latency.

Maximize the number of users that can run queries on the cluster at the same time. Reduce overall costs without compromising other requirements. Which cluster type should you recommend?

A. Standard with Auto Termination

B. High Concurrency with Autoscaling
C. High Concurrency with Auto Termination
D. Standard with Autoscaling

Correct Answer: B
Section:
Explanation:
A High Concurrency cluster is a managed cloud resource. The key benefits of High Concurrency clusters are that they provide fine-grained sharing for maximum resource utilization and minimum query latencies. Databricks chooses the appropriate number of workers required to run your job. This is referred to as autoscaling. Autoscaling makes it easier to achieve high cluster utilization, because you don't need to provision the cluster to match a workload.
Incorrect Answers:
C: The cluster configuration includes an auto terminate setting whose default value depends on cluster mode:
Standard and Single Node clusters terminate automatically after 120 minutes by default. High Concurrency clusters do not terminate automatically by default.
Reference:
https://docs.microsoft.com/en-us/azure/databricks/clusters/configure

QUESTION 30
You are creating a new notebook in Azure Databricks that will support R as the primary language but will also support Scala and SQL. Which switch should you use to switch between languages?

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

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

QUESTION 31
You have an Azure Data Factory pipeline that performs an incremental load of source data to an Azure Data Lake Storage Gen2 account. Data to be loaded is identified by a column named LastUpdatedDate in the source table. You plan to execute the pipeline every four hours.
You need to ensure that the pipeline execution meets the following requirements:
Automatically retries the execution when the pipeline run fails due to concurrency or throttling limits. Supports backfilling existing data in the table. Which type of trigger should you use?

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

Correct Answer: D
Section:
Explanation:
In case of pipeline failures, tumbling window trigger can retry the execution of the referenced pipeline automatically, using the same input parameters, without the user intervention. This can be specified using the property "retryPolicy" in the trigger definition.
Reference:

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

QUESTION 32
You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account. The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/. You need to design a daily Azure Data Factory data load to minimize the data transfer between the two accounts. Which two configurations should you include in the design? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point

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

Correct Answer: A, C
Section:
Explanation:
Copy only the daily files by using filtering.
Reference: https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage

QUESTION 33
You plan to build a structured streaming solution in Azure Databricks. The solution will count new events in five-minute intervals and report only events that arrive during the interval. The output will be sent to a Delta Lake table.
Which output mode should you use?

A. update
B. complete
C. append

Correct Answer: C
Section:
Explanation:
Append Mode: Only new rows appended in the result table since the last trigger are written to external storage. This is applicable only for the queries where existing rows in the Result Table are not expected to change.
Incorrect Answers:
B: Complete Mode: The entire updated result table is written to external storage. It is up to the storage connector to decide how to handle the writing of the entire table. A: Update Mode: Only the rows that were updated in the result table since the last trigger are written to external storage. This is different from Complete Mode in that Update Mode outputs only the rows that have changed since the last trigger. If the query doesn't contain aggregations, it is equivalent to Append mode.
Reference:
https://docs.databricks.com/getting-started/spark/streaming.html

QUESTION 34
Note: This question-is part of a series of questions that present the same scenario. Each question-in the series contains a unique solution that might meet the stated goals. Some question-sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question-in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1. You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1. You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1. Solution: In an Azure Synapse Analytics pipeline, you use a data flow that contains a Derived Column transformation. Does this meet the goal?

A. Yes
B. No

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

QUESTION 35
Note: This question-is part of a series of questions that present the same scenario. Each question-in the series contains a unique solution that might meet the stated goals. Some question-sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question-in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1. You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1. You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1. Solution: You use a dedicated SQL pool to create an external table that has an additional DateTime column. Does this meet the goal?

A. Yes
B. No

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

QUESTION 36
Note: This question-is part of a series of questions that present the same scenario. Each question-in the series contains a unique solution that might meet the stated goals. Some question-sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question-in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1. You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1. You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1. Solution: You use an Azure Synapse Analytics serverless SQL pool to create an external table that has an additional DateTime column. Does this meet the goal?

A. Yes
B. No

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

QUESTION 37
Note: This question-is part of a series of questions that present the same scenario. Each question-in the series contains a unique solution that might meet the stated goals. Some question-sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question-in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen. You plan to create an Azure Databricks workspace that has a tiered structure.
The workspace will contain the following three workloads:
A workload for data engineers who will use Python and SQL.

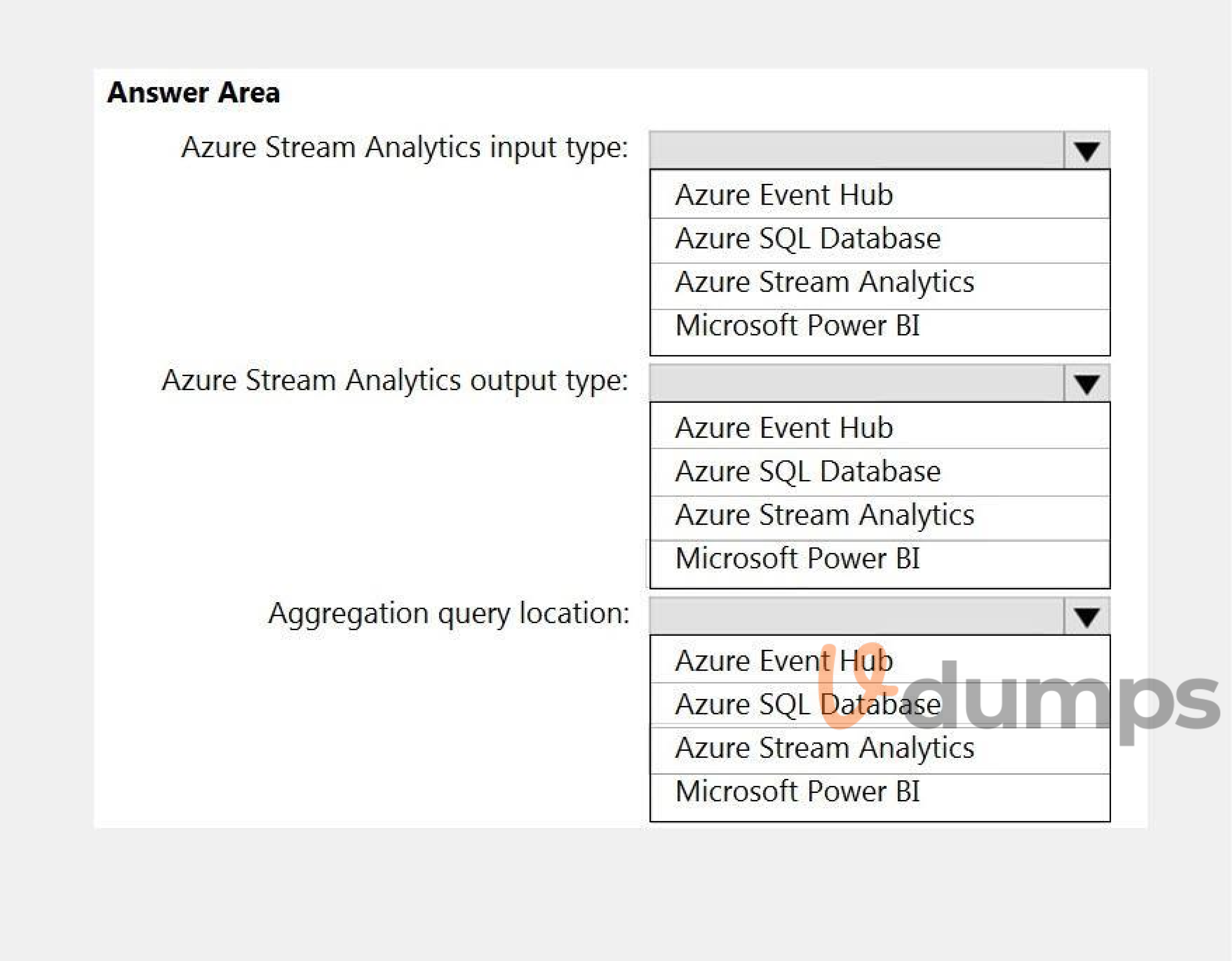
A workload for jobs that will run notebooks that use Python, Scala, and SQL. A workload that data scientists will use to perform ad hoc analysis in Scala and R. The enterprise architecture team at your company identifies the following standards for Databricks environments:
The data engineers must share a duster.
The job duster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster. All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists. You need to create the Databricks clusters for the workloads. Solution: You create a Standard cluster for each data scientist, a Standard cluster for the data engineers, and a High Concurrency cluster for the jobs. Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section:
Explanation:
We need a High Concurrency cluster for the data engineers and the jobs. Note: Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL. A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.
Reference:
https://docs.azuredatabricks.net/clusters/configure.html

QUESTION 38
HOTSPOT
You are designing a real-time dashboard solution that will visualize streaming data from remote sensors that connect to the internet. The streaming data must be aggregated to show the average value of each 10-second interval. The data will be discarded after being displayed in the dashboard.
The solution will use Azure Stream Analytics and must meet the following requirements:
Minimize latency from an Azure Event hub to the dashboard.
Minimize the required storage.
Minimize development effort.
What should you include in the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point

Hot Area:



Answer Area

Azure Stream Analytics input type:

- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

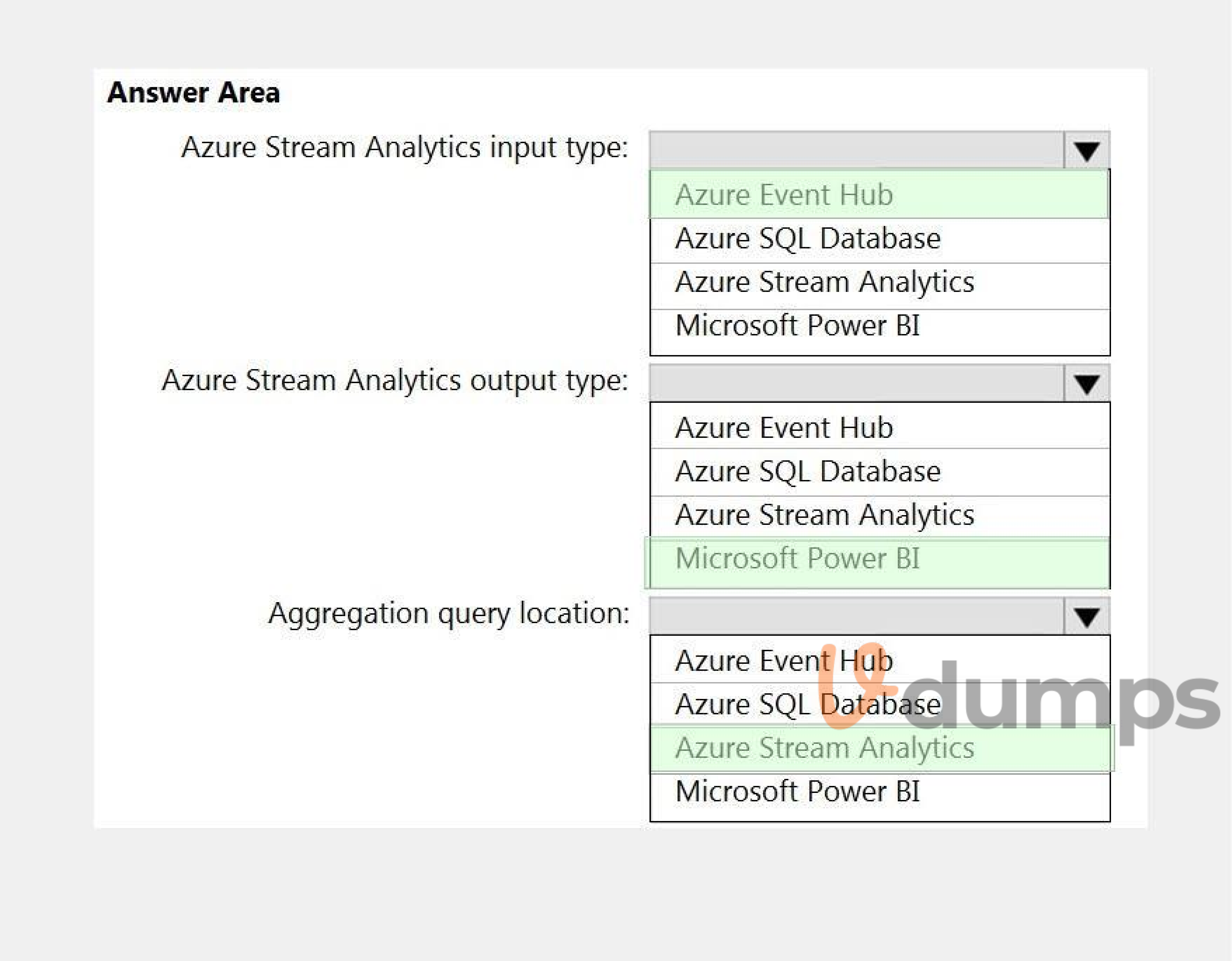
Azure Stream Analytics output type:

- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Aggregation query location:

- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Answer Area:



Answer Area

Azure Stream Analytics input type:

- [✓ Azure Event Hub]
- Azure SQL Database
- Azure Stream Analytics
- Microsoft Power BI

Azure Stream Analytics output type:

- Azure Event Hub
- Azure SQL Database
- Azure Stream Analytics
- [✓ Microsoft Power BI]

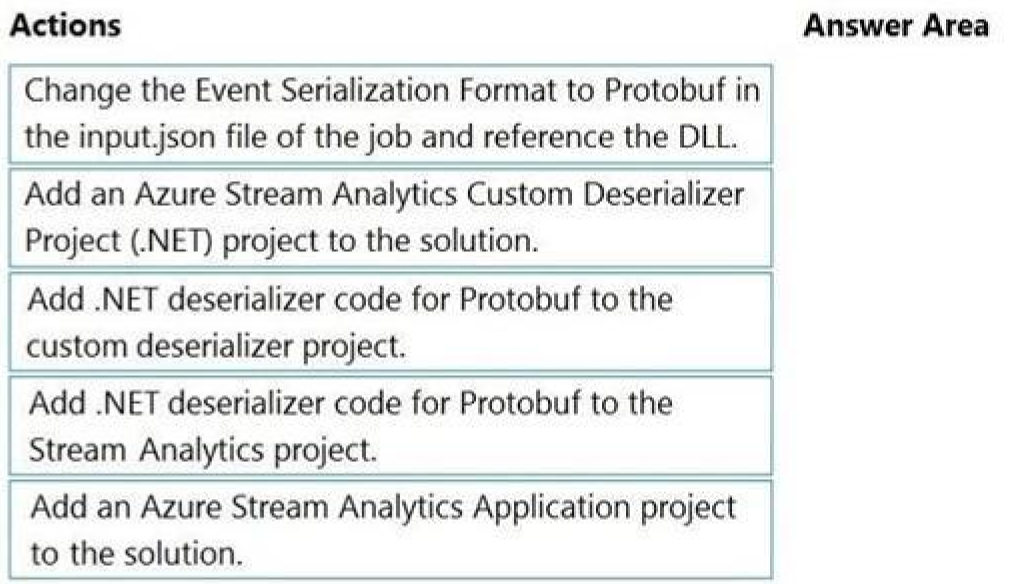
Aggregation query location:

- Azure Event Hub
- Azure SQL Database
- [✓ Azure Stream Analytics]
- Microsoft Power BI

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

QUESTION 39
DRAG DROP
You have an Azure Stream Analytics job that is a Stream Analytics project solution in Microsoft Visual Studio. The job accepts data generated by IoT devices in the JSON format.
You need to modify the job to accept data generated by the IoT devices in the Protobuf format.
Which three actions should you perform from Visual Studio on sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:



Actions

Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.

Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.

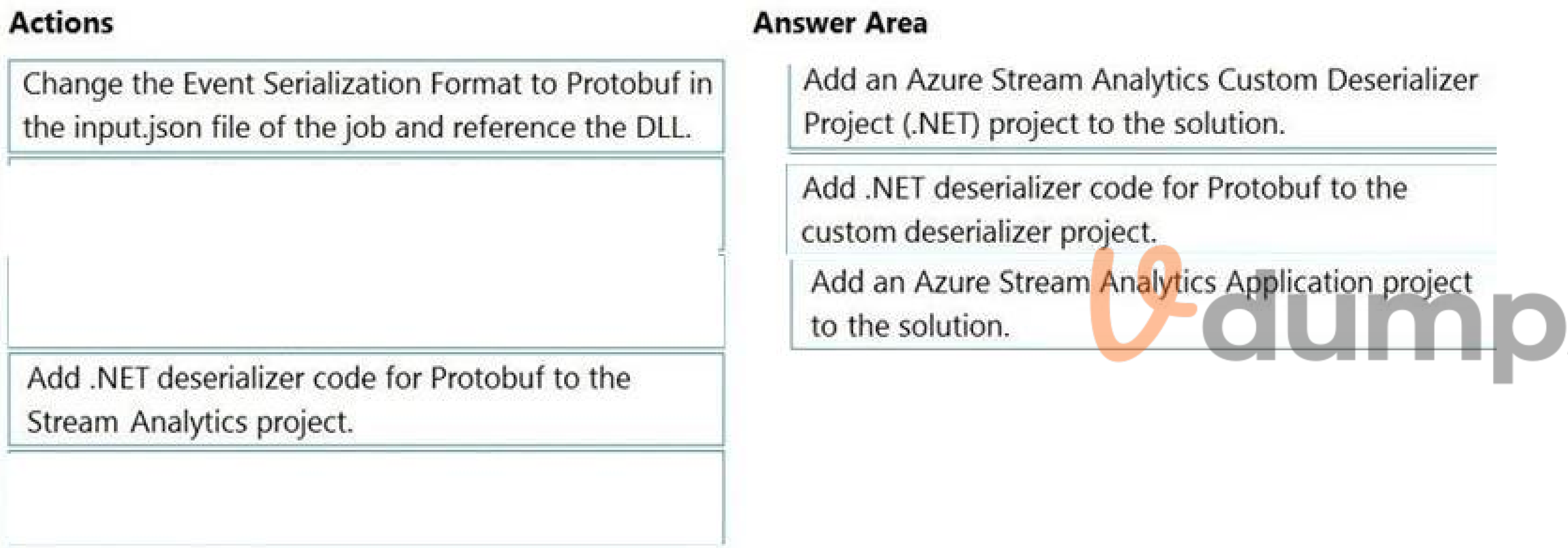
Add .NET deserializer code for Protobuf to the custom deserializer project.

Add .NET deserializer code for Protobuf to the Stream Analytics project.

Add an Azure Stream Analytics Application project to the solution.

Answer Area

Correct Answer:



Actions

Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.

Add .NET deserializer code for Protobuf to the Stream Analytics project.

Answer Area

Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.

Add .NET deserializer code for Protobuf to the custom deserializer project.

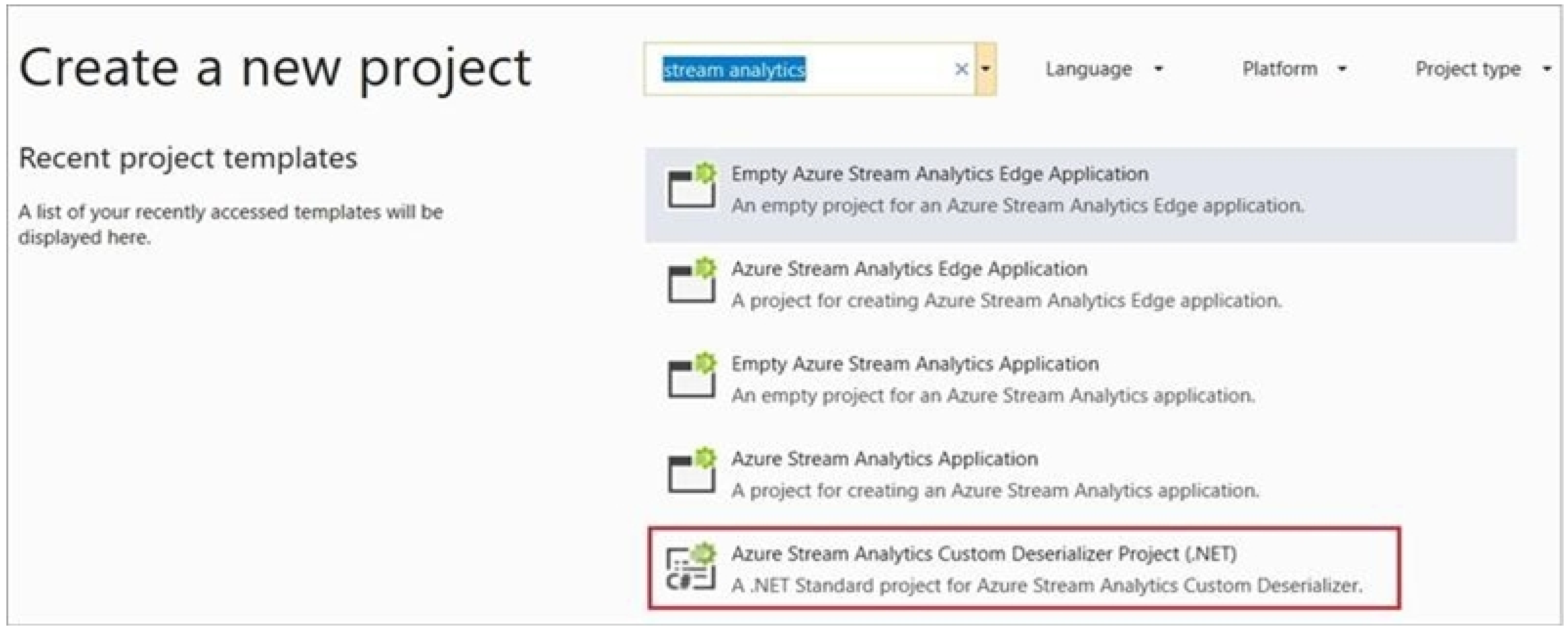
Add an Azure Stream Analytics Application project to the solution.

Section:

Explanation:

Step 1: Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution. Create a custom deserializer

1. Open Visual Studio and select File > New > Project. Search for Stream Analytics and select Azure Stream Analytics Custom Deserializer Project (.NET). Give the project a name, like Protobuf Deserializer.



Create a new project

stream analytics

Recent project templates

A list of your recently accessed templates will be displayed here.

Empty Azure Stream Analytics Edge Application
An empty project for an Azure Stream Analytics Edge application.

Azure Stream Analytics Edge Application
A project for creating Azure Stream Analytics Edge application.

Empty Azure Stream Analytics Application
An empty project for an Azure Stream Analytics application.

Azure Stream Analytics Application
A project for creating an Azure Stream Analytics application.

Azure Stream Analytics Custom Deserializer Project (.NET)
A .NET Standard project for Azure Stream Analytics Custom Deserializer.

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

Step 2: Add .NET deserializer code for Protobuf to the custom deserializer project Azure Stream Analytics has built-in support for three data formats: JSON, CSV, and Avro. With custom .NET deserializers, you can read data from other formats such as Protocol Buffer, Bond and other user defined formats for both cloud and edge jobs.

Step 3: Add an Azure Stream Analytics Application project to the solution Add an Azure Stream Analytics project

1. In Solution Explorer, right-click the Protobuf Deserializer solution and select Add > New Project. Under Azure Stream Analytics > Stream Analytics, choose Azure Stream Analytics Application. Name it ProtobufCloudDeserializer and select OK.
2. Right-click References under the ProtobufCloudDeserializer Azure Stream Analytics project. Under Projects, add Protobuf Deserializer. It should be automatically populated for you.

Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/custom-deserializer

QUESTION 40

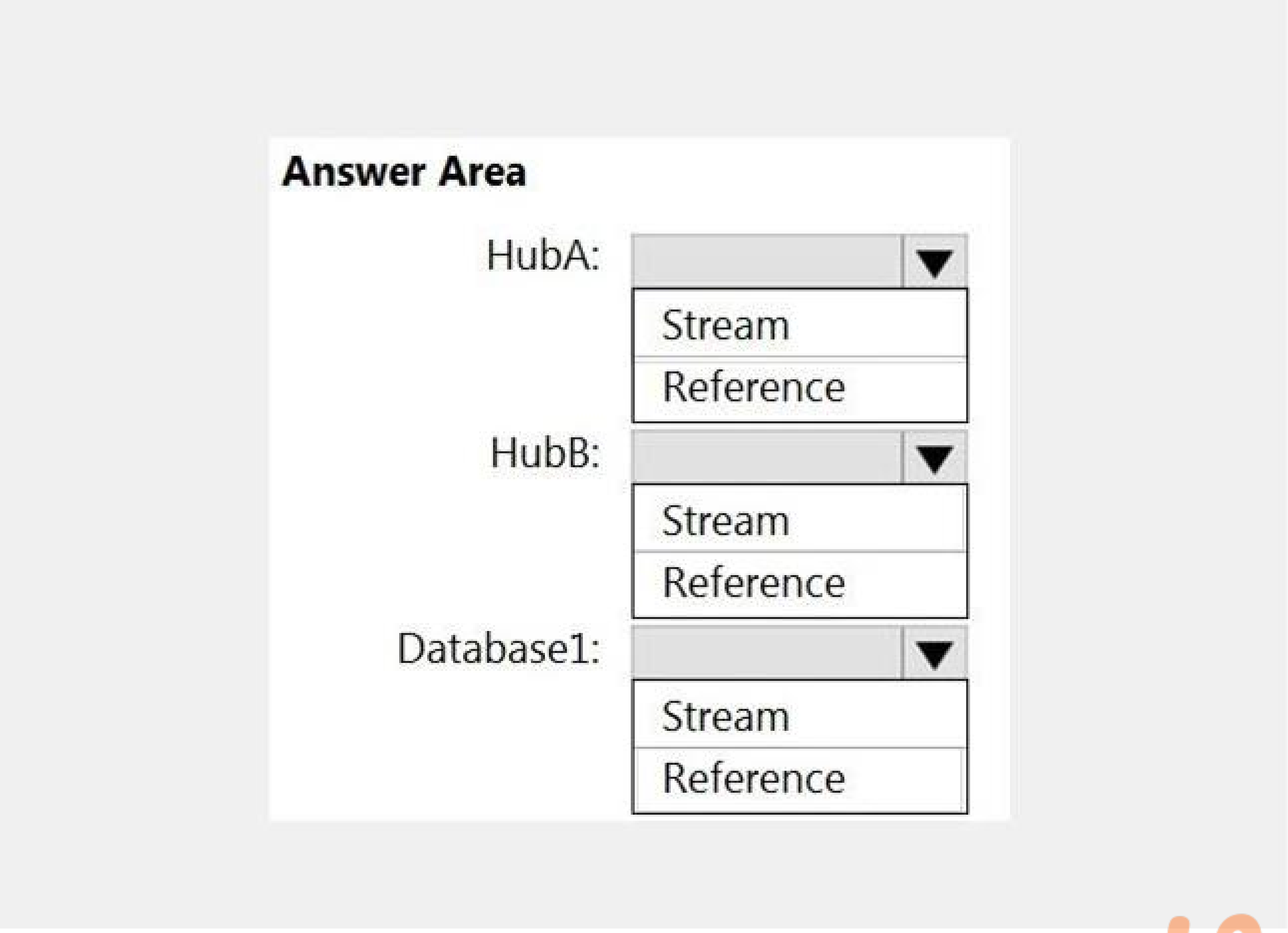
HOTSPOT

You have an Azure SQL database named Database1 and two Azure event hubs named HubA and HubB. The data consumed from each source is shown in the following table.

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

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

Hot Area:



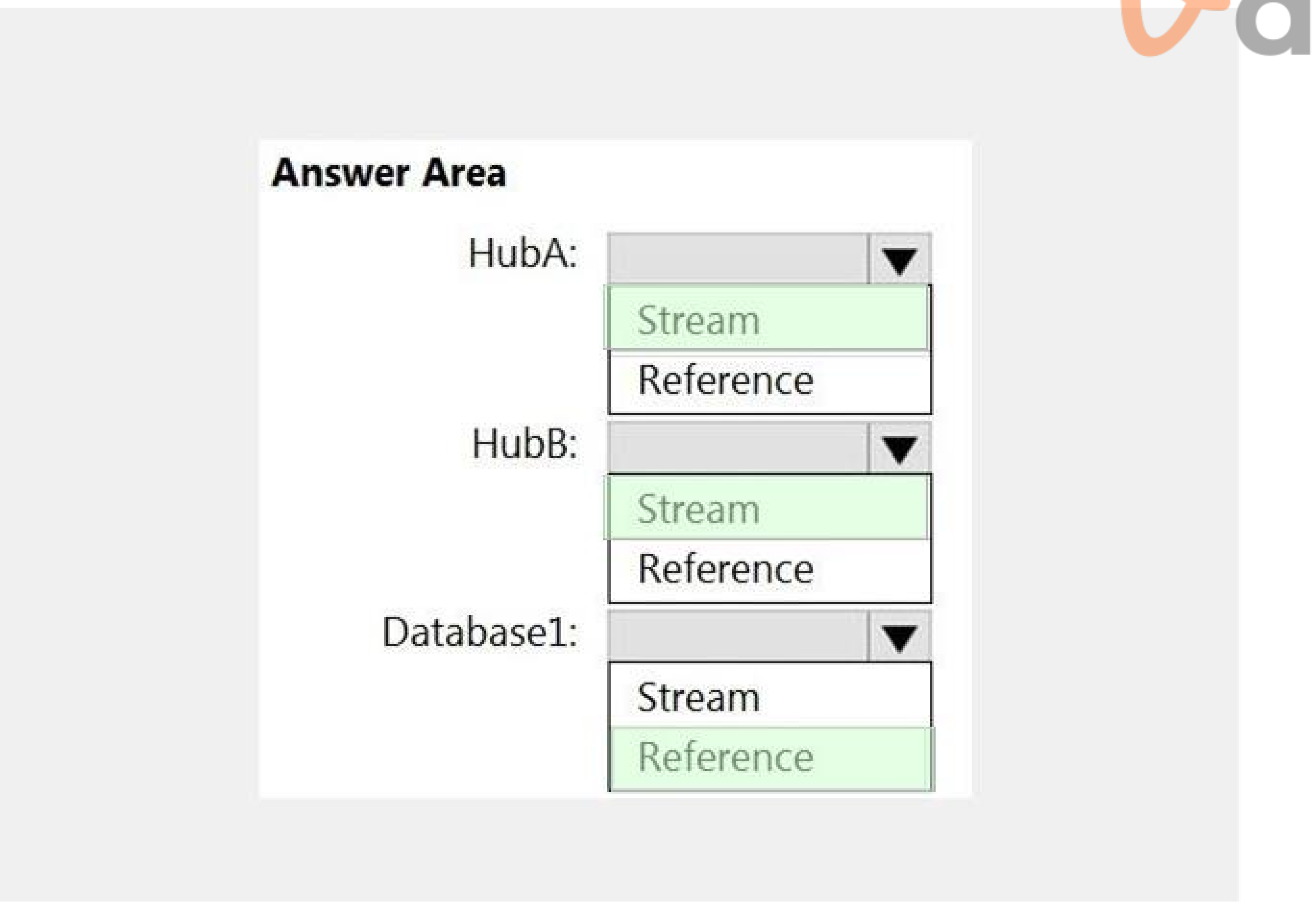
Answer Area

HubA:
- Stream
- Reference

HubB:
- Stream
- Reference

Database1:
- Stream
- Reference

Answer Area:



Answer Area

HubA:
- [✓ Stream]
- Reference

HubB:
- [✓ Stream]
- Reference

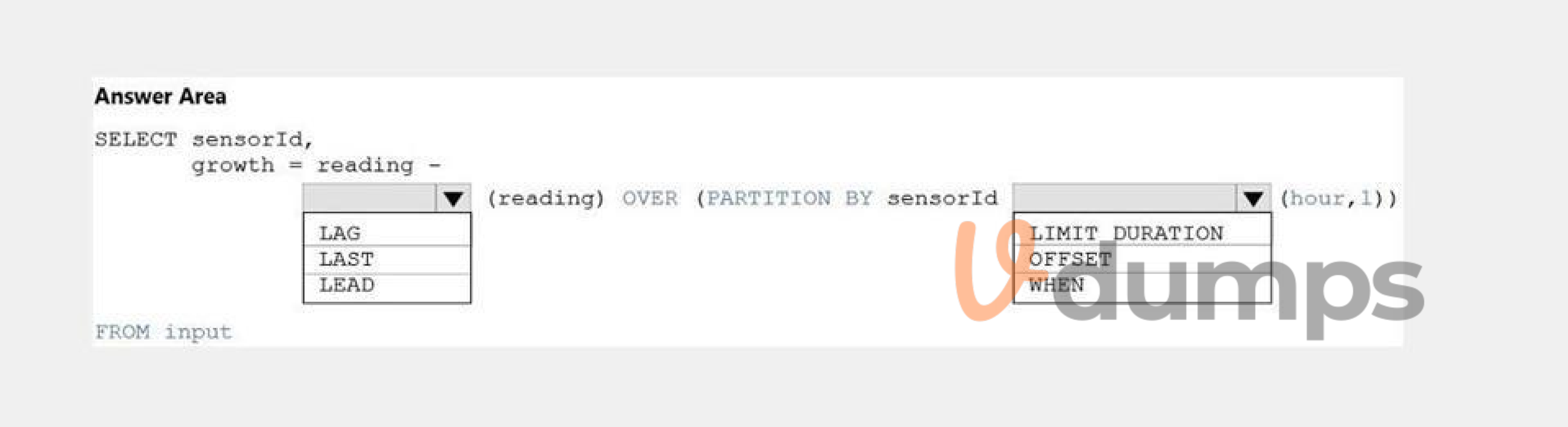
Database1:
- Stream
- [✓ Reference]

Section:

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

QUESTION 41
HOTSPOT
You are building an Azure Analytics query that will receive input data from Azure IoT Hub and write the results to Azure Blob storage.
You need to calculate the difference in readings per sensor per hour.
How should you complete the query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

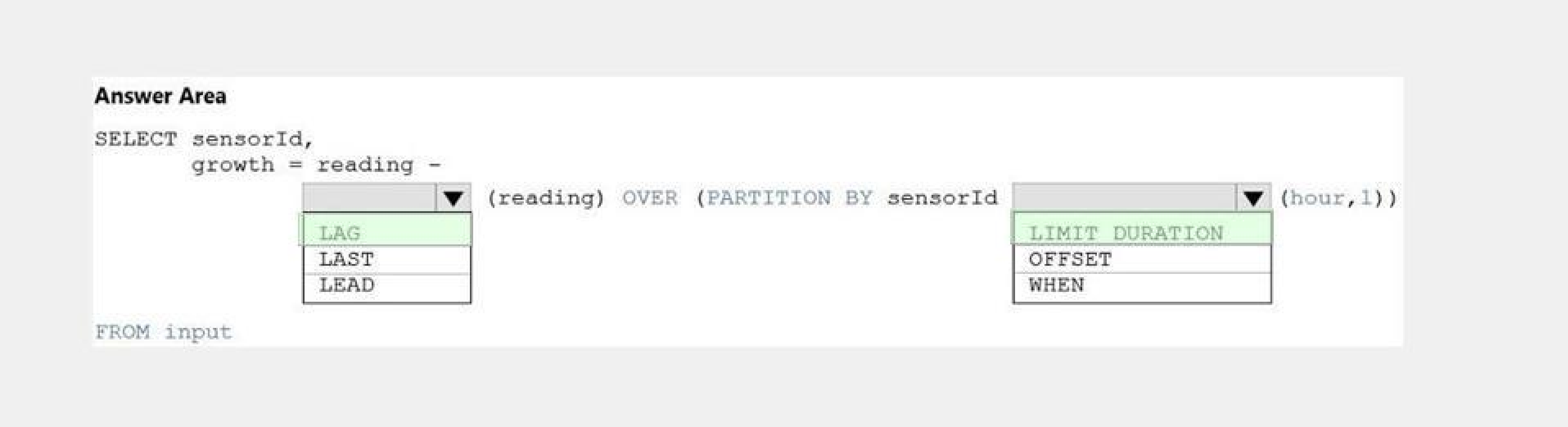
Hot Area:



Answer Area

```
SELECT sensorId,
       growth = reading -
                 [          ▼] (reading) OVER (PARTITION BY sensorId [               ▼] (hour,1))
                  LAG                                                  LIMIT DURATION
                  LAST                                                 OFFSET
                  LEAD                                                 WHEN
FROM input
```

Answer Area:



Answer Area

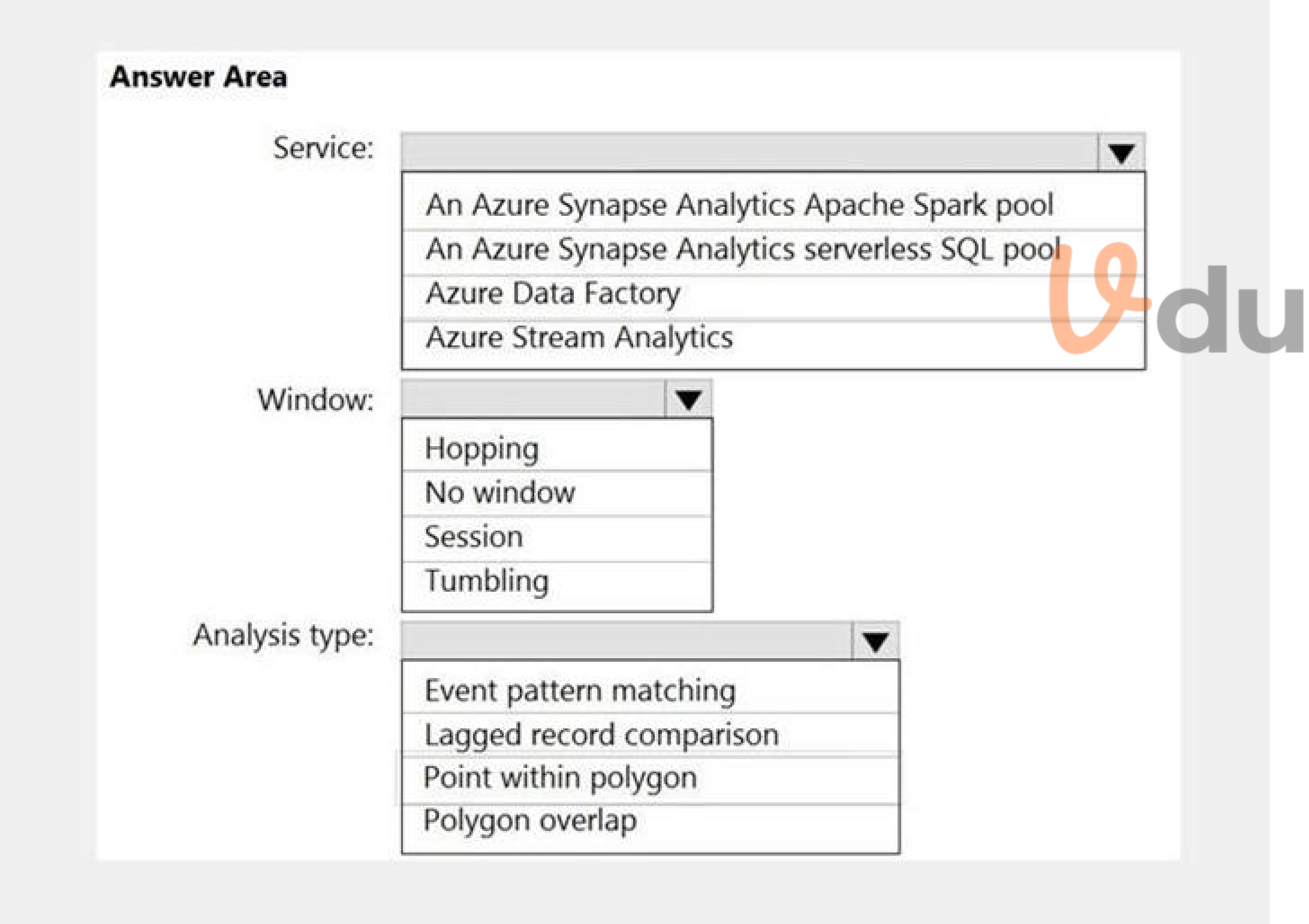
```
SELECT sensorId,
       growth = reading -
                 [          ▼] (reading) OVER (PARTITION BY sensorId [               ▼] (hour,1))
                  [✓ LAG]                                              [✓ LIMIT DURATION]
                  LAST                                                 OFFSET
                  LEAD                                                 WHEN
FROM input
```

Section:
Explanation:
Box 1: LAG
The LAG analytic operator allows one to look up a "previous" event in an event stream, within certain constraints. It is very useful for computing the rate of growth of a variable, detecting when a variable crosses a threshold, or when a condition starts or stops being true.

Box 2: LIMIT DURATION
Example: Compute the rate of growth, per sensor:
SELECT sensorId,
growth = reading -
LAG(reading) OVER (PARTITION BY sensorId LIMIT DURATION(hour, 1)) FROM input
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/lag-azure-stream-analytics

QUESTION 42
HOTSPOT
You are designing a monitoring solution for a fleet of 500 vehicles. Each vehicle has a GPS tracking device that sends data to an Azure event hub once per minute.
You have a CSV file in an Azure Data Lake Storage Gen2 container. The file maintains the expected geographical area in which each vehicle should be.
You need to ensure that when a GPS position is outside the expected area, a message is added to another event hub for processing within 30 seconds. The solution must minimize cost.
What should you include in the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:



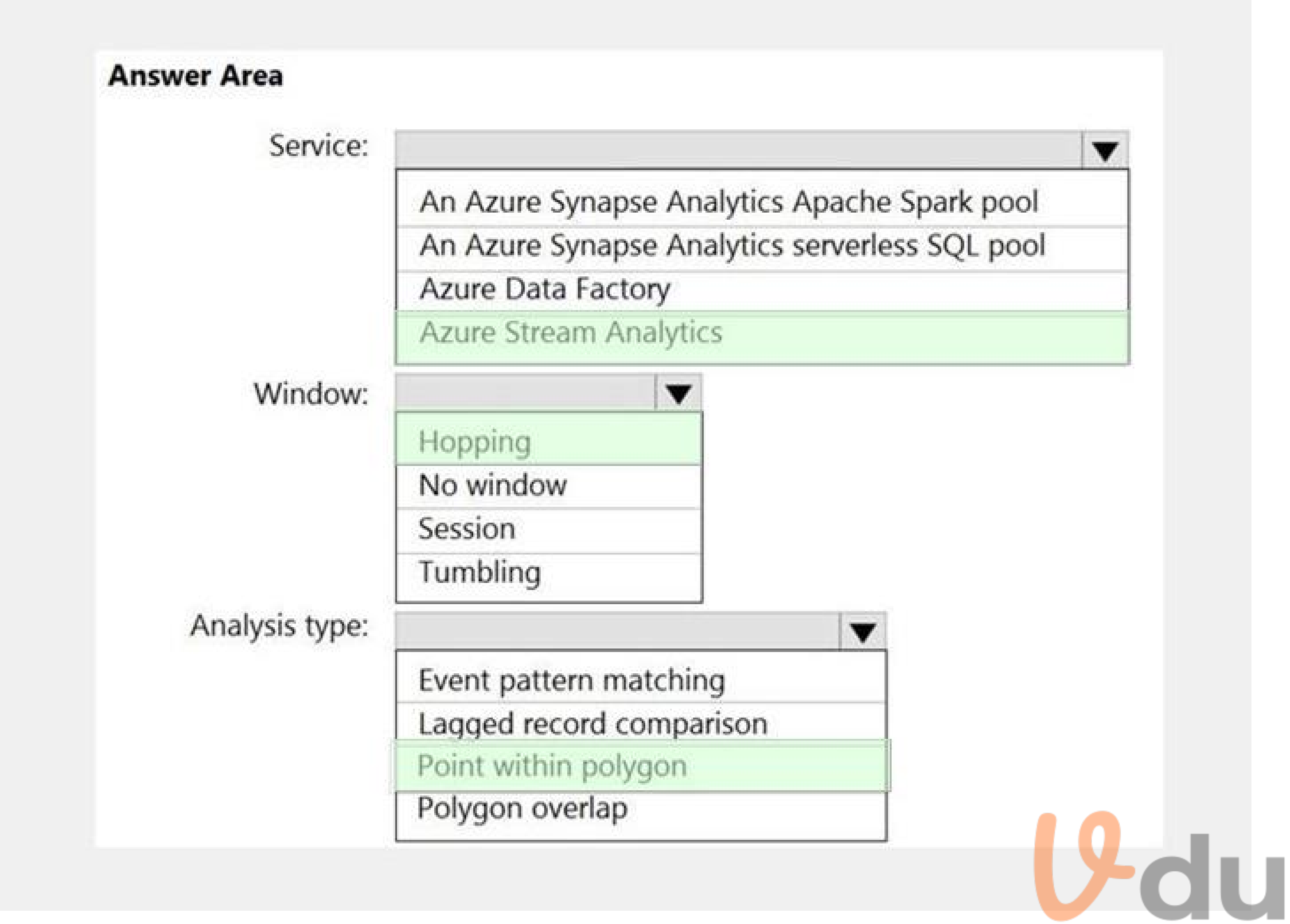
Answer Area

Service:
- An Azure Synapse Analytics Apache Spark pool
- An Azure Synapse Analytics serverless SQL pool
- Azure Data Factory
- Azure Stream Analytics

Window:
- Hopping
- No window
- Session
- Tumbling

Analysis type:
- Event pattern matching
- Lagged record comparison
- Point within polygon
- Polygon overlap

Answer Area:



Answer Area

Service:

- An Azure Synapse Analytics Apache Spark pool
- An Azure Synapse Analytics serverless SQL pool
- Azure Data Factory
- Azure Stream Analytics

Window:

- Hopping
- No window
- Session
- Tumbling

Analysis type:

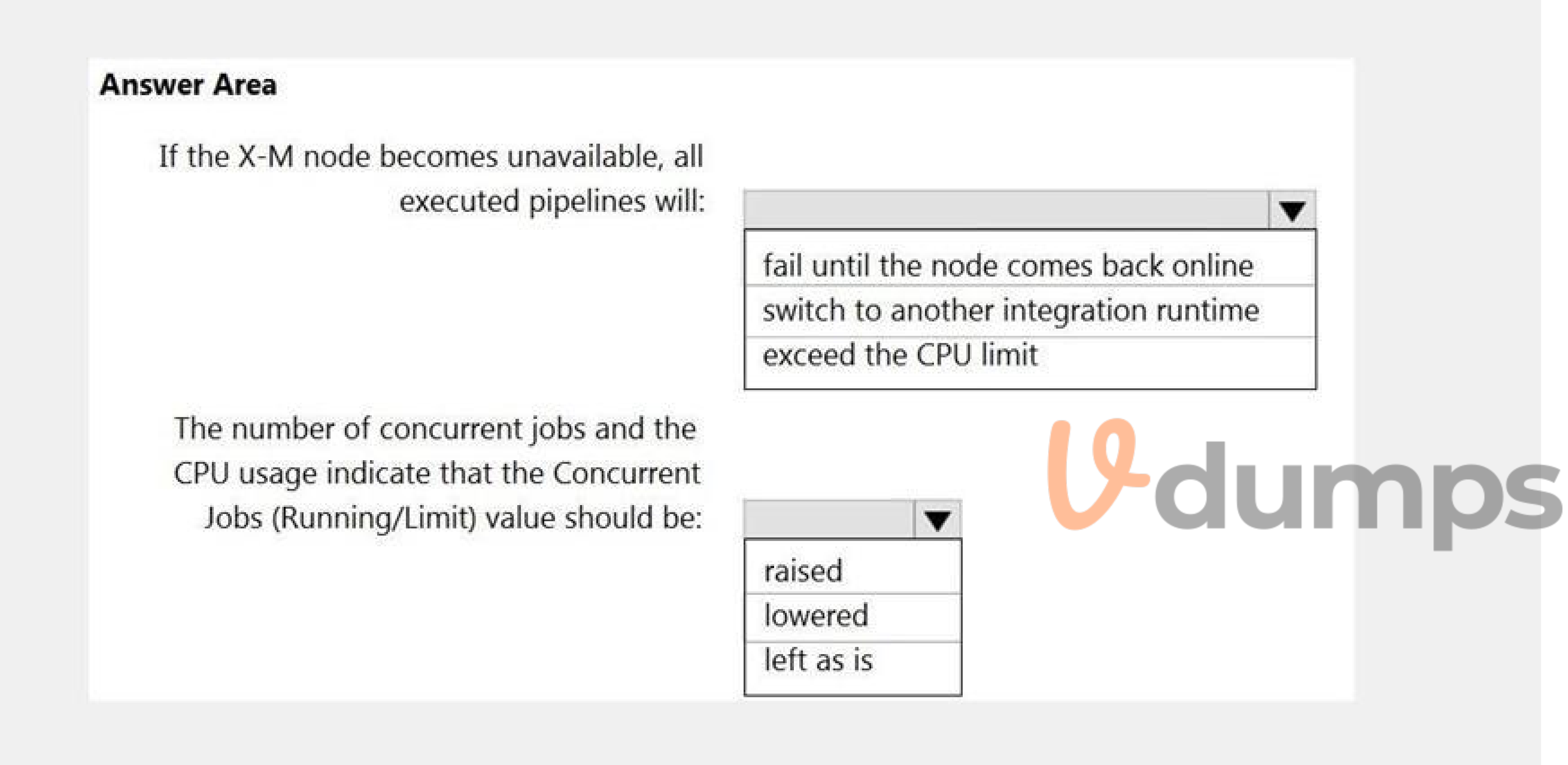
- Event pattern matching
- Lagged record comparison
- Point within polygon
- Polygon overlap

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

QUESTION 43
HOTSPOT
You have a self-hosted integration runtime in Azure Data Factory.
The current status of the integration runtime has the following configurations:
Status: Running
Type: Self-Hosted
Version: 4.4.7292.1
Running / Registered Node(s): 1/1
High Availability Enabled: False
Linked Count: 0
Queue Length: 0
Average Queue Duration. 0.00s
The integration runtime has the following node details:
Name: X-M

Status: Running
Version: 4.4.7292.1
Available Memory: 7697MB
CPU Utilization: 6%
Network (In/Out): 1.21KBps/0.83KBps
Concurrent Jobs (Running/Limit): 2/14
Role: Dispatcher/Worker
Credential Status: In Sync
Use the drop-down menus to select the answer choice that completes each statement based on the information presented.
NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

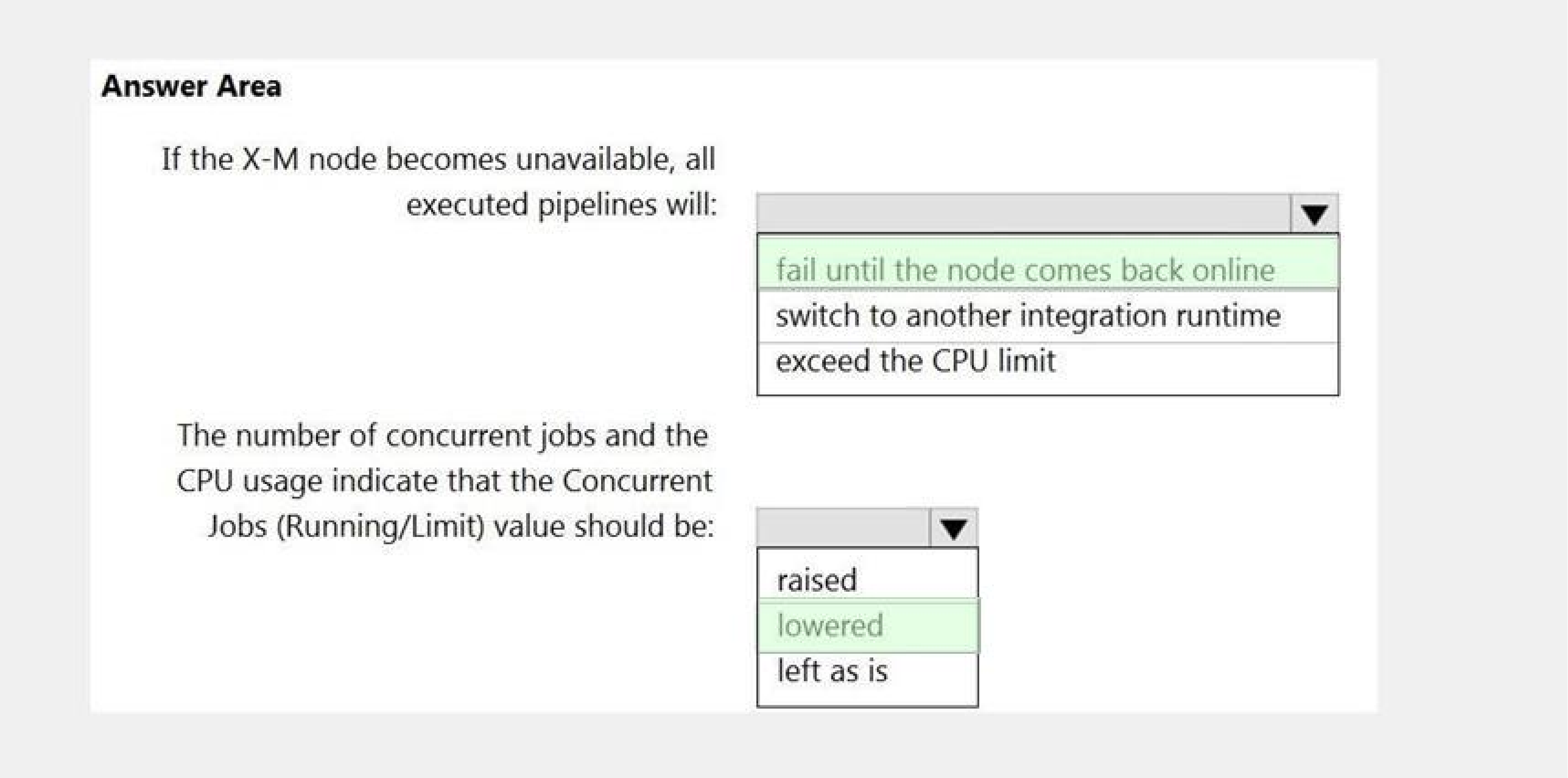
If the X-M node becomes unavailable, all executed pipelines will:

- fail until the node comes back online
- switch to another integration runtime
- exceed the CPU limit

The number of concurrent jobs and the CPU usage indicate that the Concurrent Jobs (Running/Limit) value should be:

- raised
- lowered
- left as is

Answer Area:



Answer Area

If the X-M node becomes unavailable, all executed pipelines will: ▼

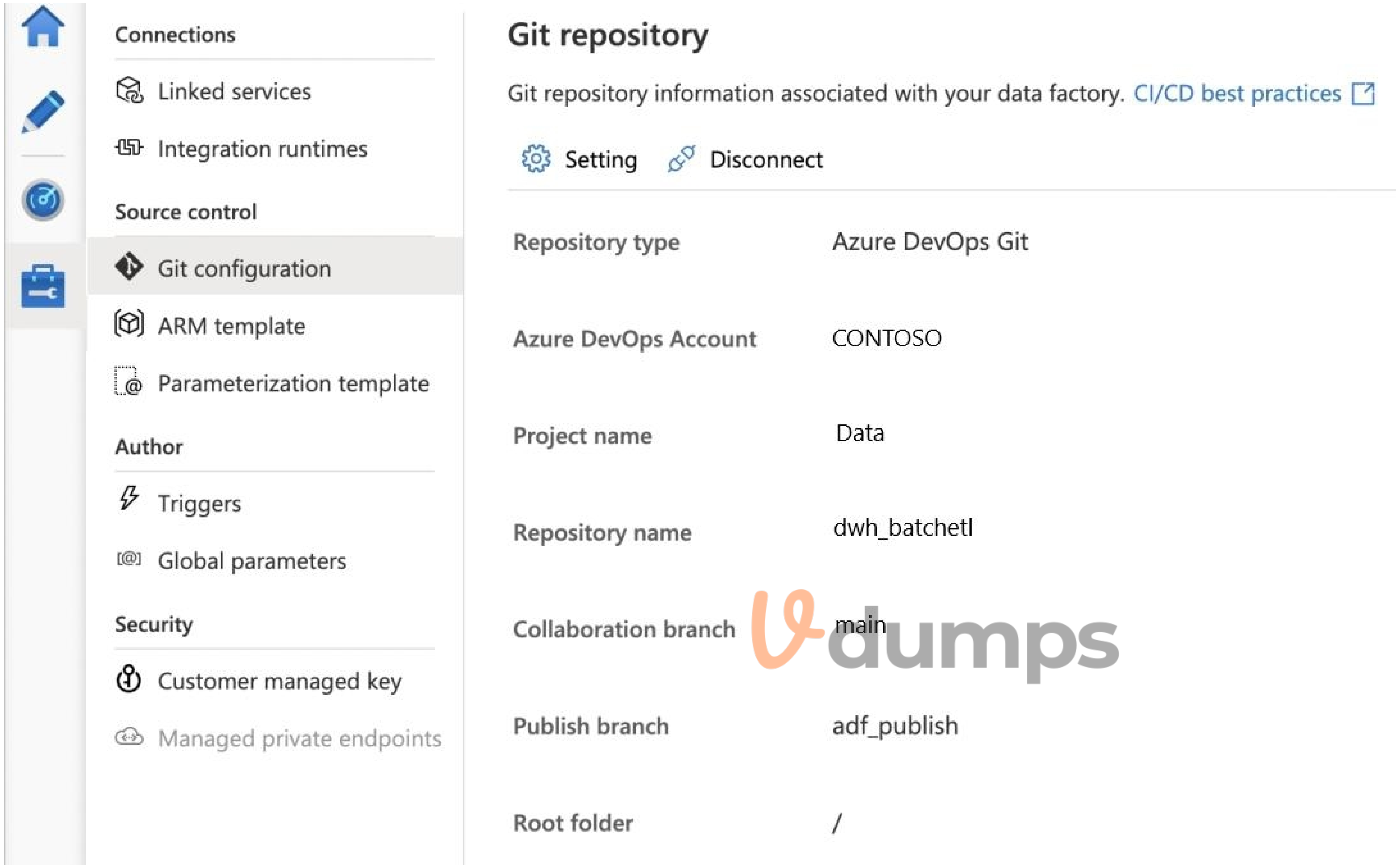
- [✓ fail until the node comes back online]
- switch to another integration runtime
- exceed the CPU limit

The number of concurrent jobs and the CPU usage indicate that the Concurrent Jobs (Running/Limit) value should be: ▼

- raised
- [✓ lowered]
- left as is

Section:
Explanation:
Box 1: fail until the node comes back online
We see: High Availability Enabled: False
Note: Higher availability of the self-hosted integration runtime so that it's no longer the single point of failure in your big data solution or cloud data integration with Data Factory.
Box 2: lowered
We see:
Concurrent Jobs (Running/Limit): 2/14
CPU Utilization: 6%
Note: When the processor and available RAM aren't well utilized, but the execution of concurrent jobs reaches a node's limits, scale up by increasing the number of concurrent jobs that a node can run
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime

QUESTION 44
HOTSPOT
You configure version control for an Azure Data Factory instance as shown in the following exhibit.



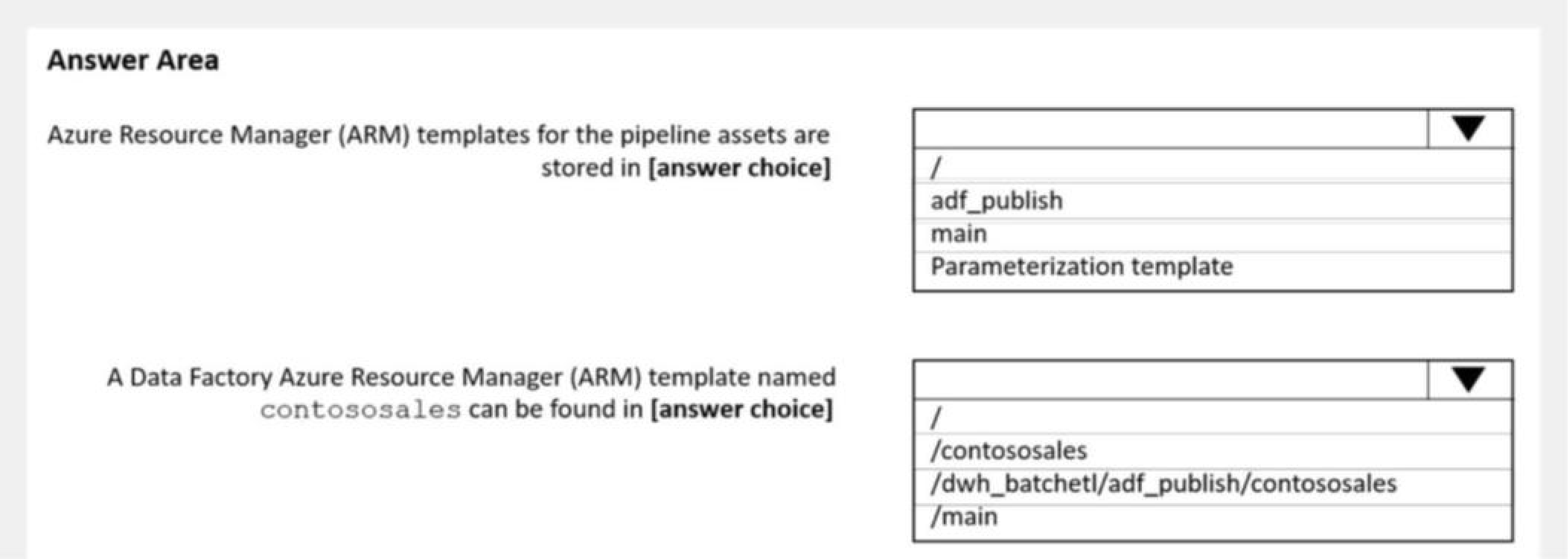
Git repository

Git repository information associated with your data factory. CI/CD best practices

Repository type	Azure DevOps Git
Azure DevOps Account	CONTOSO
Project name	Data
Repository name	dwh_batchetl
Collaboration branch	main
Publish branch	adf_publish
Root folder	/

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

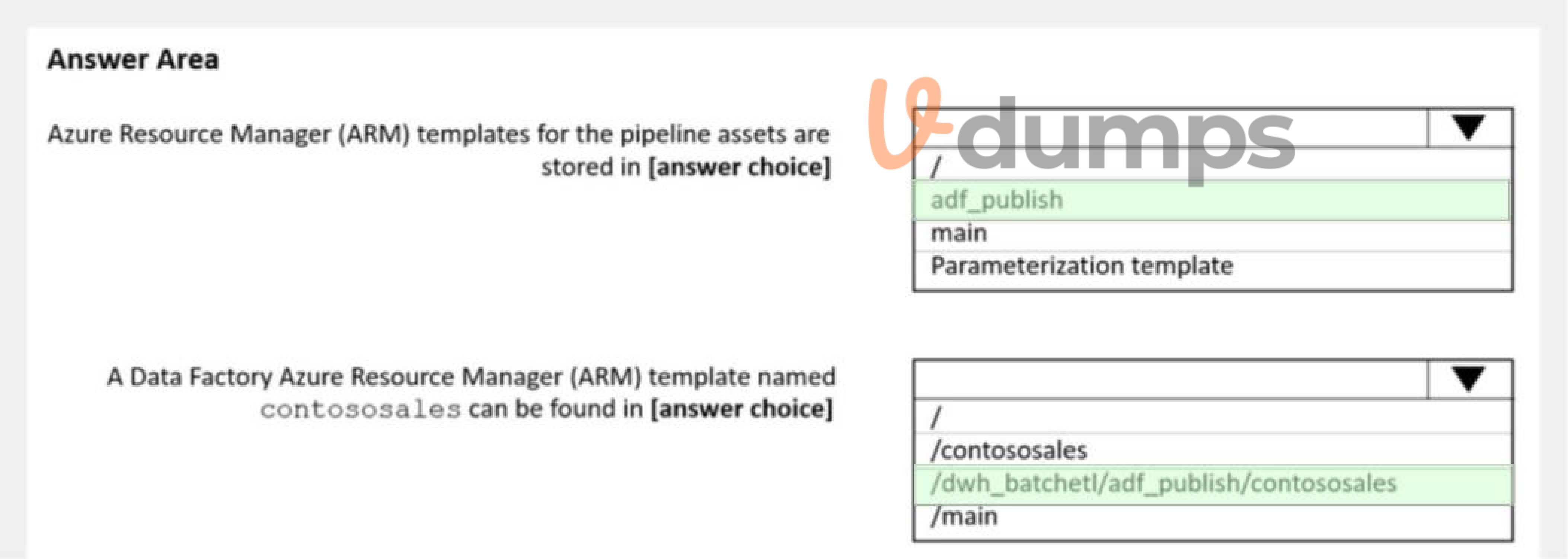
Azure Resource Manager (ARM) templates for the pipeline assets are stored in **[answer choice]**

/
adf_publish
main
Parameterization template

A Data Factory Azure Resource Manager (ARM) template named `contososales` can be found in **[answer choice]**

/
/contososales
/dwh_batchetl/adf_publish/contososales
/main

Answer Area:



Answer Area

Azure Resource Manager (ARM) templates for the pipeline assets are stored in **[answer choice]**

/
[✓ adf_publish]
main
Parameterization template

A Data Factory Azure Resource Manager (ARM) template named `contososales` can be found in **[answer choice]**

/
/contososales
[✓ /dwh_batchetl/adf_publish/contososales]
/main

Section:
Explanation:
Box 1: adf_publish
The Publish branch is the branch in your repository where publishing related ARM templates are stored and updated. By default, it's adf_publish.
Box 2: / dwh_batchetl/adf_publish/contososales
Note: RepositoryName (here dwh_batchetl): Your Azure Repos code repository name. Azure Repos projects contain Git repositories to manage your source code as your project grows. You can create a new repository or use an existing repository that's already in your project.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/source-control

QUESTION 45

HOTSPOT

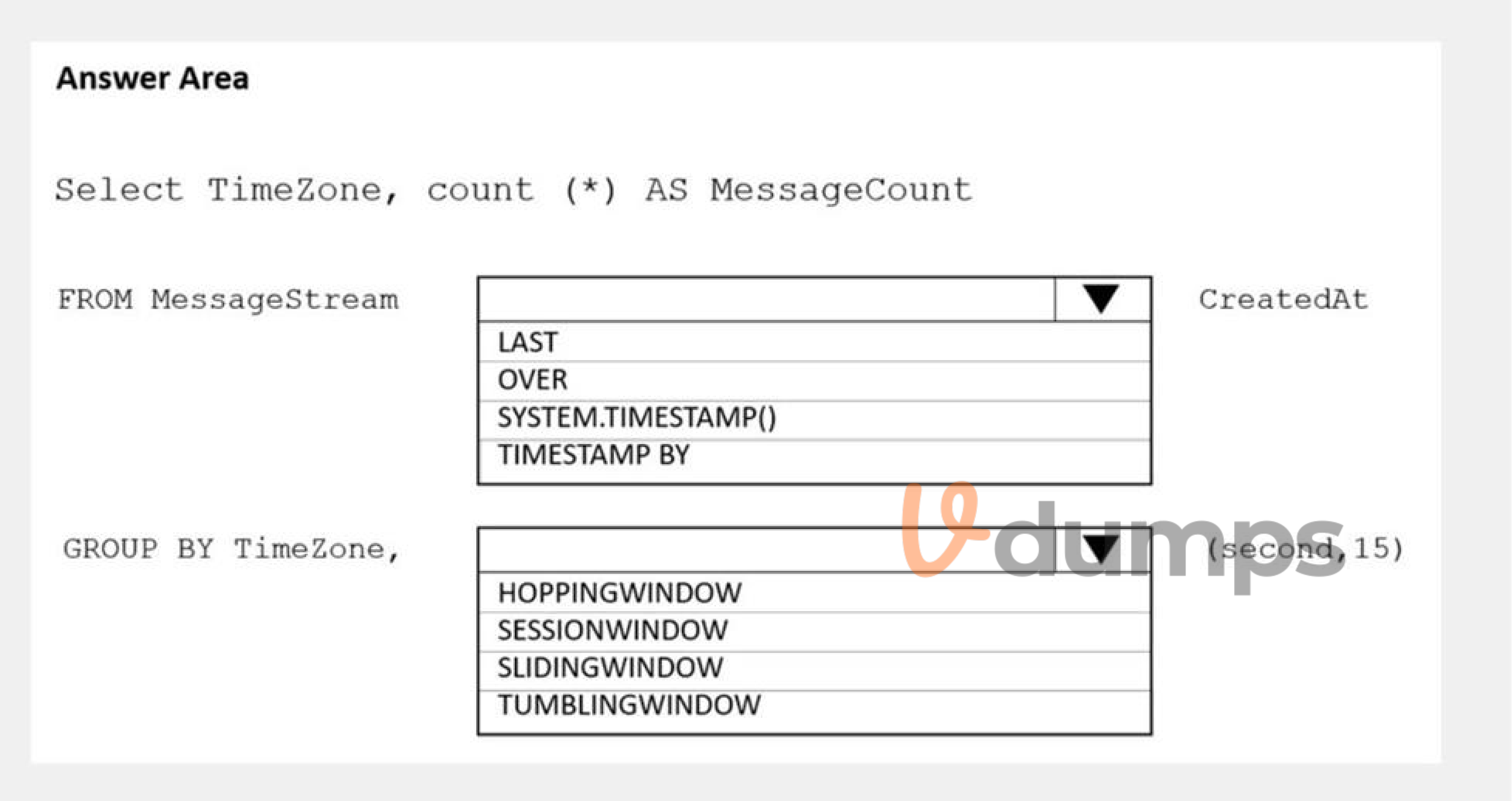
You are designing an Azure Stream Analytics solution that receives instant messaging data from an Azure Event Hub.

You need to ensure that the output from the Stream Analytics job counts the number of messages per time zone every 15 seconds.

How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



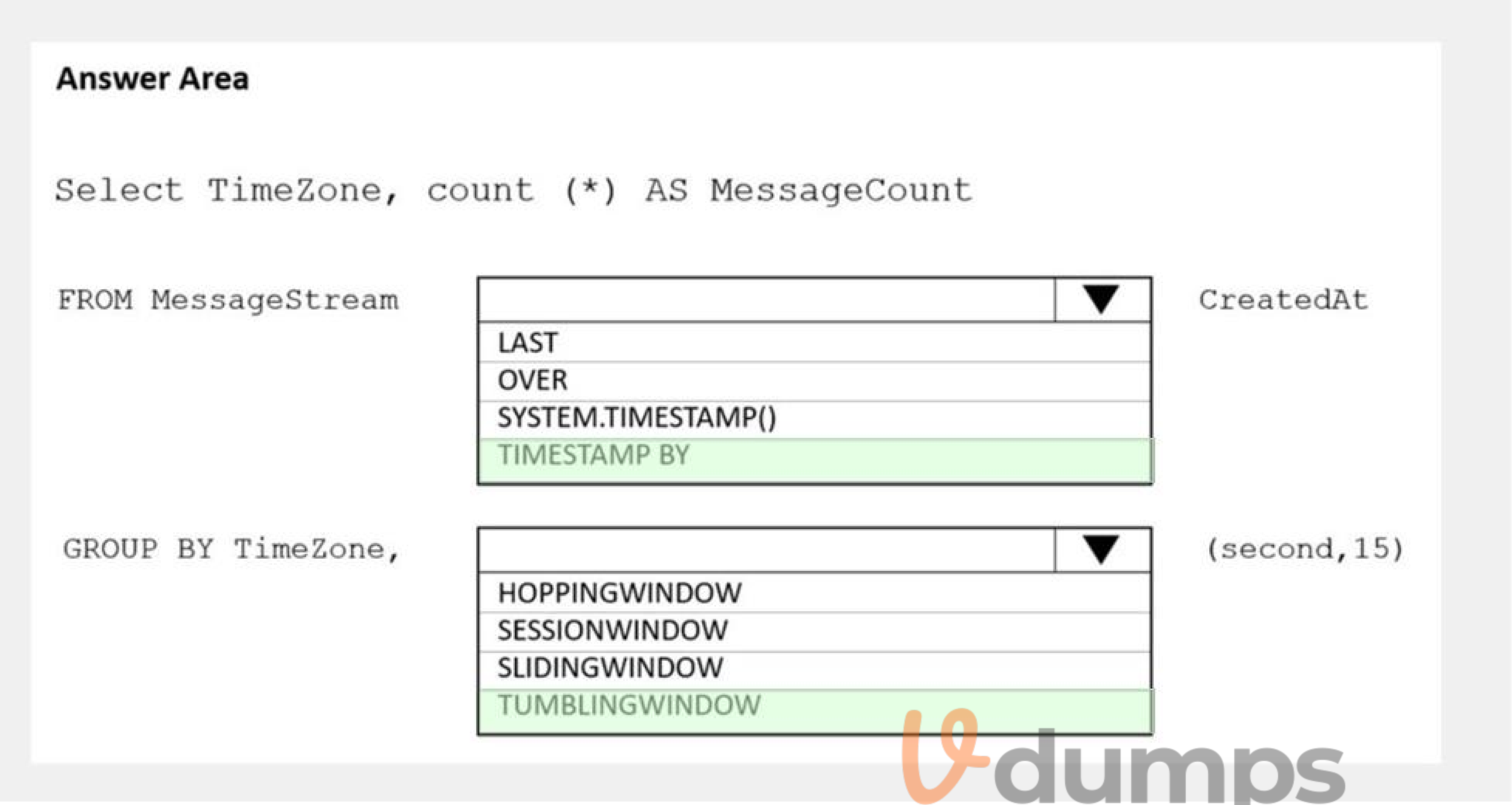
Answer Area

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

FROM MessageStream  [▼]  CreatedAt

GROUP BY TimeZone,  [▼]  (second, 15)
```

- LAST
- OVER
- SYSTEM.TIMESTAMP()
- TIMESTAMP BY

- HOPPINGWINDOW
- SESSIONWINDOW
- SLIDINGWINDOW
- TUMBLINGWINDOW

Answer Area:



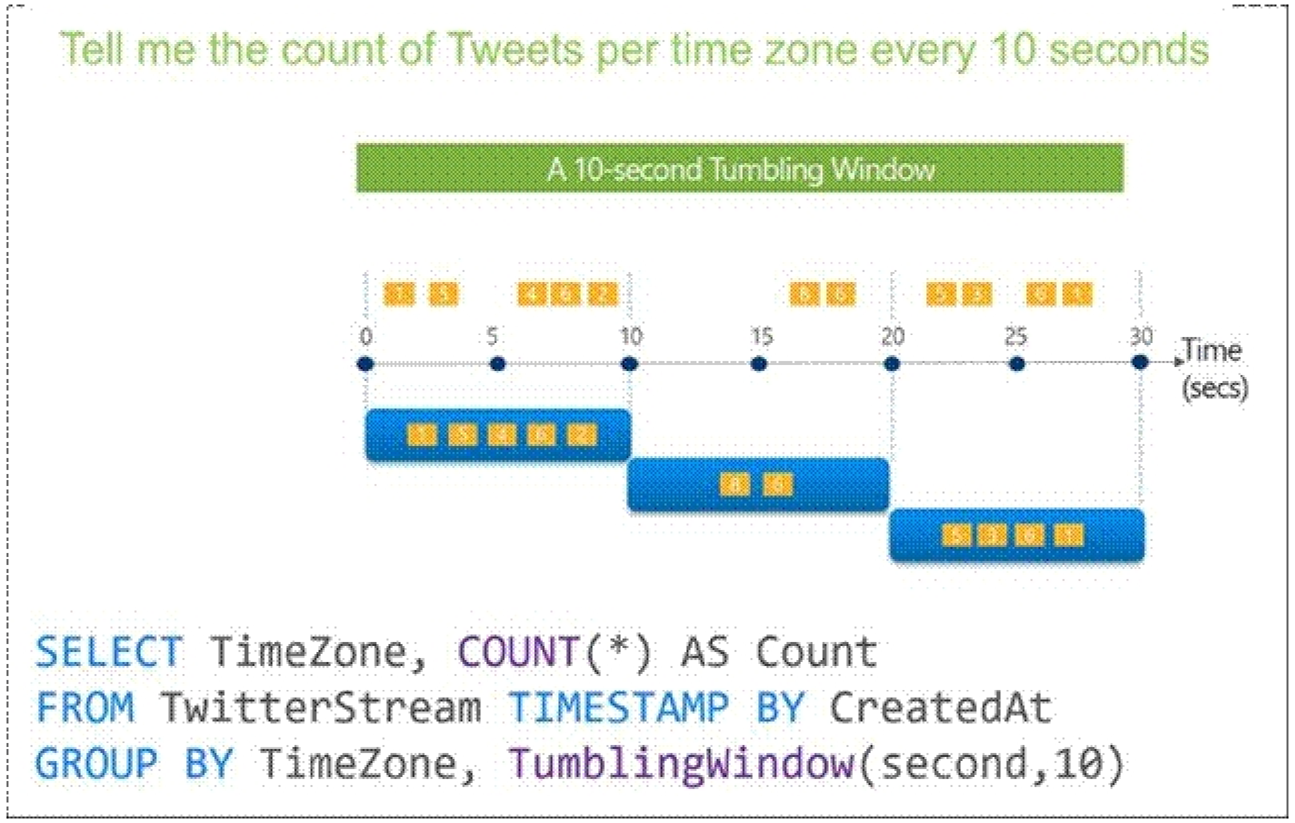
Answer Area

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

FROM MessageStream          ▼          CreatedAt

GROUP BY TimeZone,          ▼          (second,15)
```

- LAST
- OVER
- SYSTEM.TIMESTAMP()
- TIMESTAMP BY

- HOPPINGWINDOW
- SESSIONWINDOW
- SLIDINGWINDOW
- TUMBLINGWINDOW

Section:
Explanation:
Box 1: timestamp by
Box 2: TUMBLINGWINDOW
Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.



Tell me the count of Tweets per time zone every 10 seconds

A 10-second Tumbling Window

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

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

QUESTION 46

HOTSPOT

You have an Azure Data Factory instance named ADF1 and two Azure Synapse Analytics workspaces named WS1 and WS2.

ADF1 contains the following pipelines:

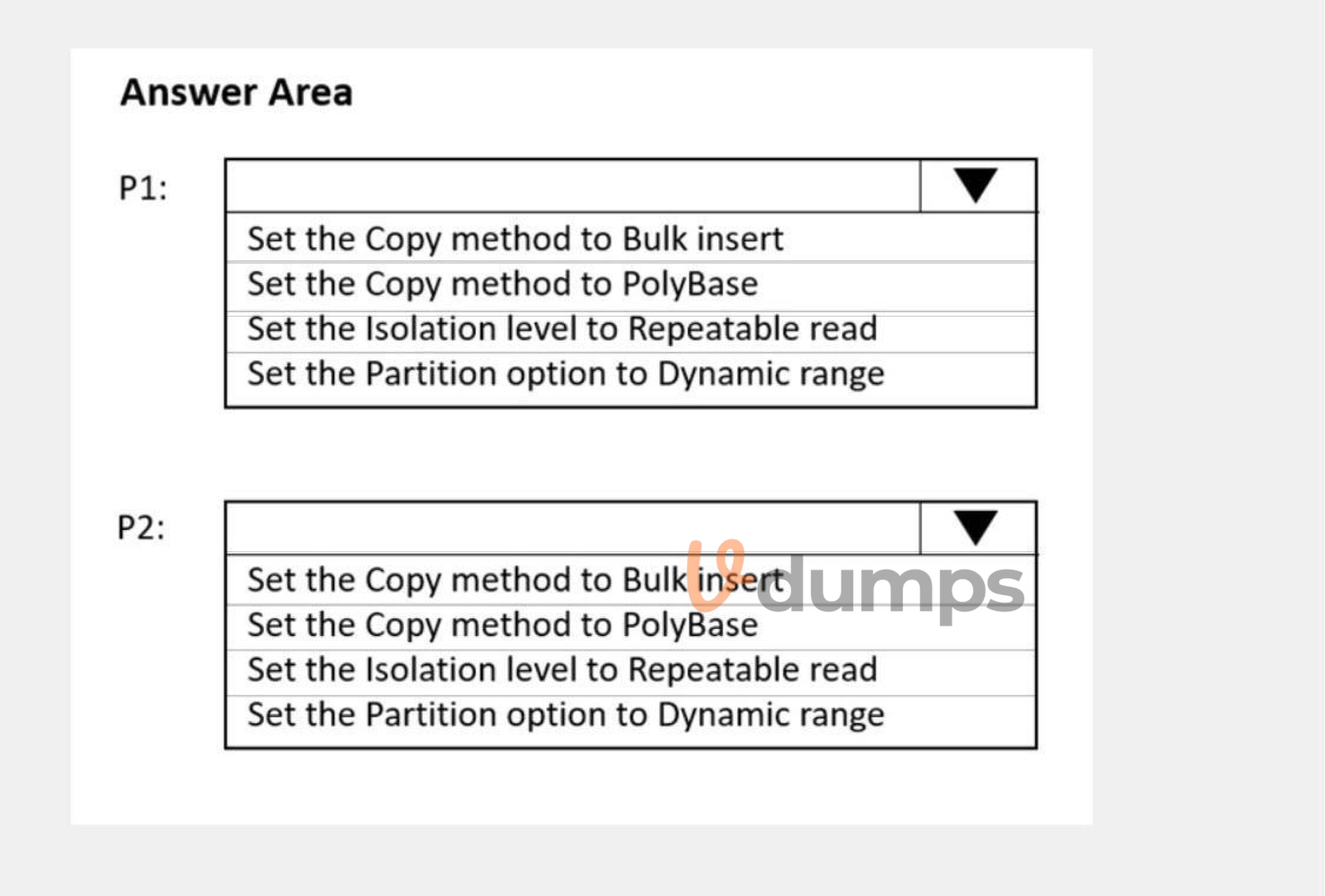
P1: Uses a copy activity to copy data from a nonpartitioned table in a dedicated SQL pool of WS1 to an Azure Data Lake Storage Gen2 account P2: Uses a copy activity to copy data from text-delimited files in an Azure Data Lake Storage Gen2 account to a nonpartitioned table in a dedicated SQL pool of WS2

You need to configure P1 and P2 to maximize parallelism and performance.

Which dataset settings should you configure for the copy activity if each pipeline? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

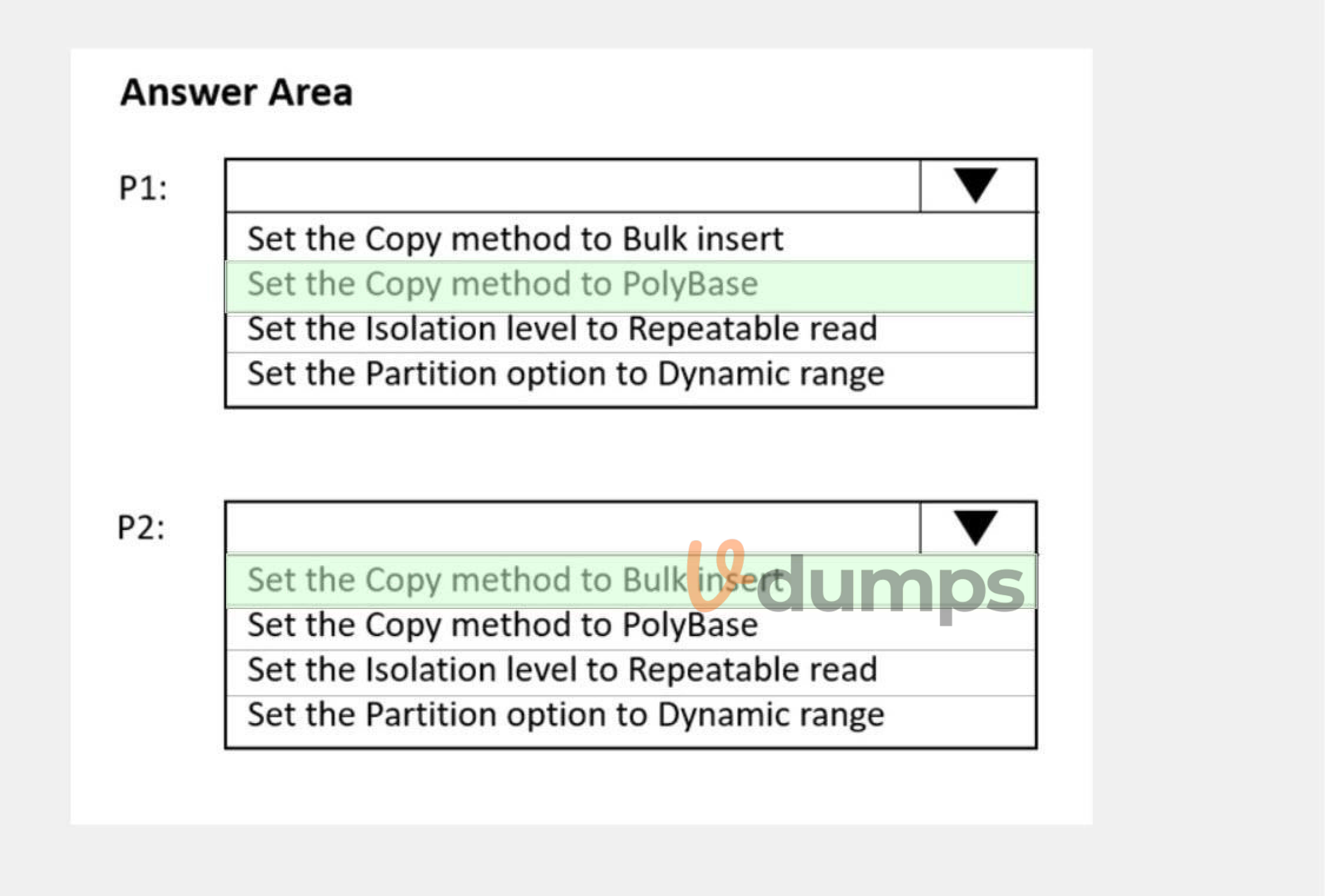
P1:

- Set the Copy method to Bulk insert
- Set the Copy method to PolyBase
- Set the Isolation level to Repeatable read
- Set the Partition option to Dynamic range

P2:

- Set the Copy method to Bulk insert
- Set the Copy method to PolyBase
- Set the Isolation level to Repeatable read
- Set the Partition option to Dynamic range

Answer Area:



Answer Area

P1:

- Set the Copy method to Bulk insert
- [✓ Set the Copy method to PolyBase]
- Set the Isolation level to Repeatable read
- Set the Partition option to Dynamic range

P2:

- [✓ Set the Copy method to Bulk insert]
- Set the Copy method to PolyBase
- Set the Isolation level to Repeatable read
- Set the Partition option to Dynamic range

Section:
Explanation:
Box 1: Set the Copy method to PolyBase
While SQL pool supports many loading methods including non-Polybase options such as BCP and SQL BulkCopy API, the fastest and most scalable way to load data is through PolyBase. PolyBase is a technology that accesses external data stored in Azure Blob storage or Azure Data Lake Store via the T-SQL language.
Box 2: Set the Copy method to Bulk insert
Polybase not possible for text files. Have to use Bulk insert.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/load-data-overview

QUESTION 47

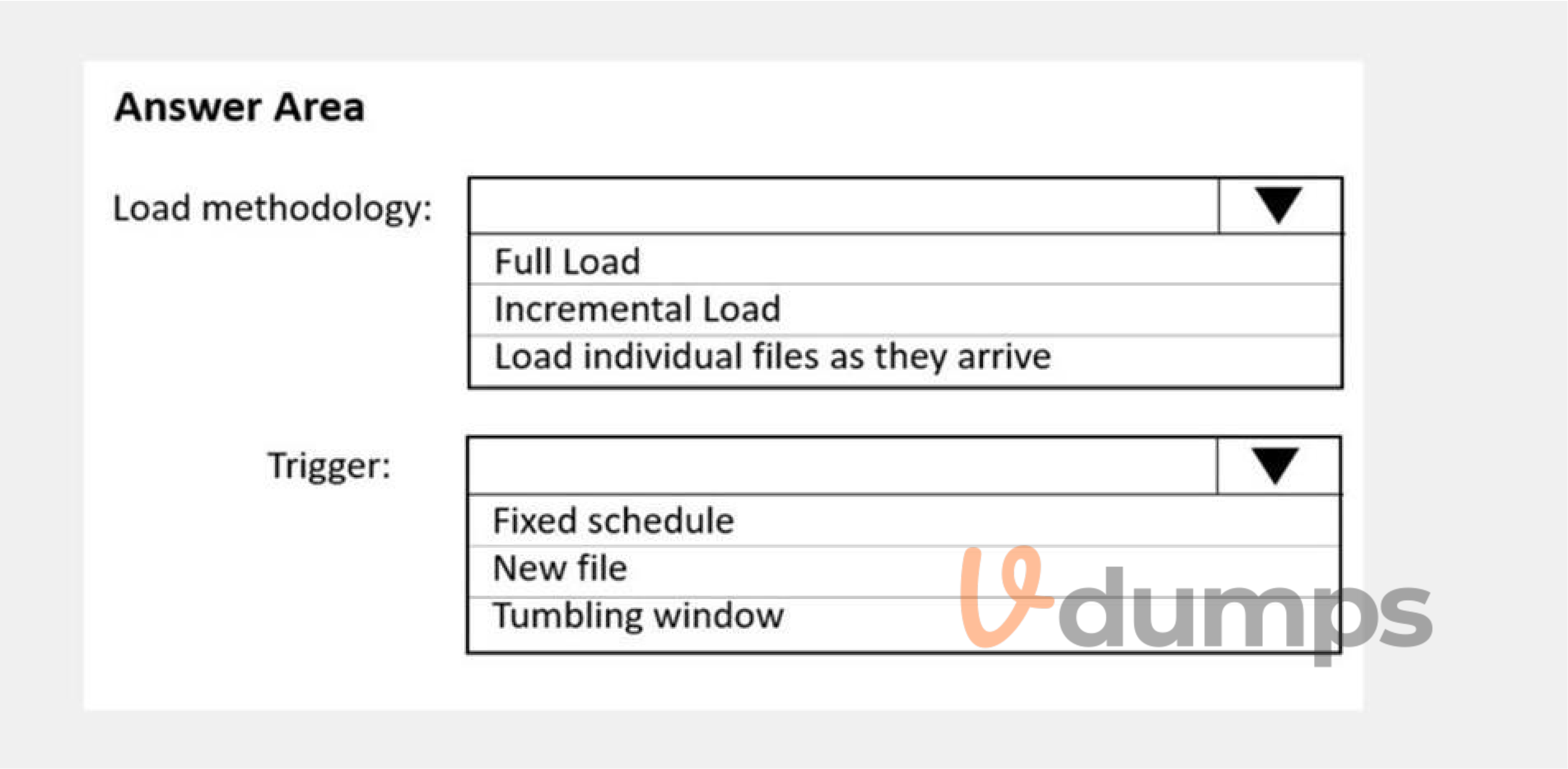
HOTSPOT

You have an Azure Storage account that generates 200,000 new files daily. The file names have a format of {YYYY}/{MM}/{DD}/{HH}/{CustomerID}.csv.

You need to design an Azure Data Factory solution that will load new data from the storage account to an Azure Data Lake once hourly. The solution must minimize load times and costs. How should you configure the solution? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

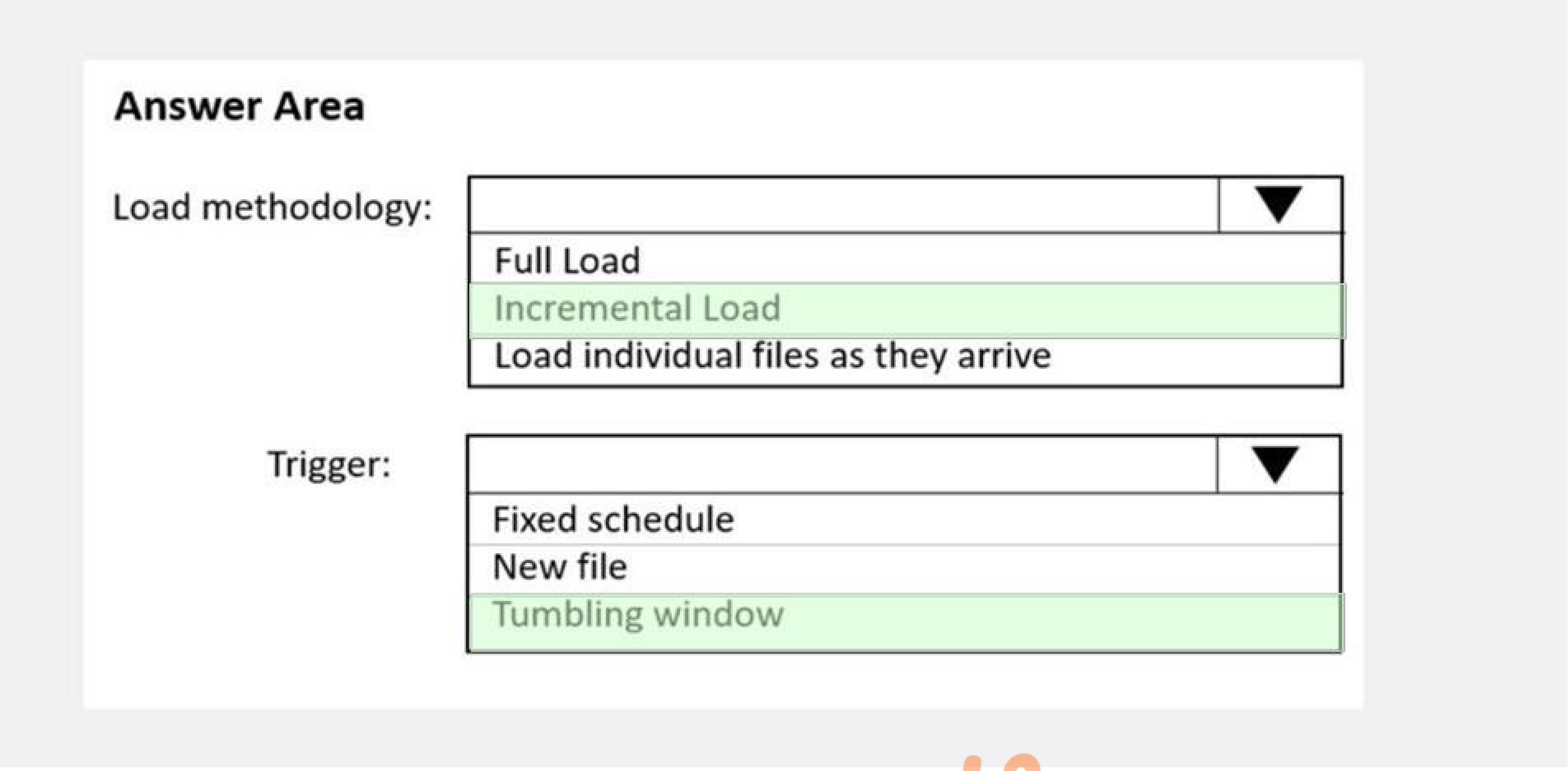


Answer Area

Load methodology: ▼
- Full Load
- Incremental Load
- Load individual files as they arrive

Trigger: ▼
- Fixed schedule
- New file
- Tumbling window

Answer Area:

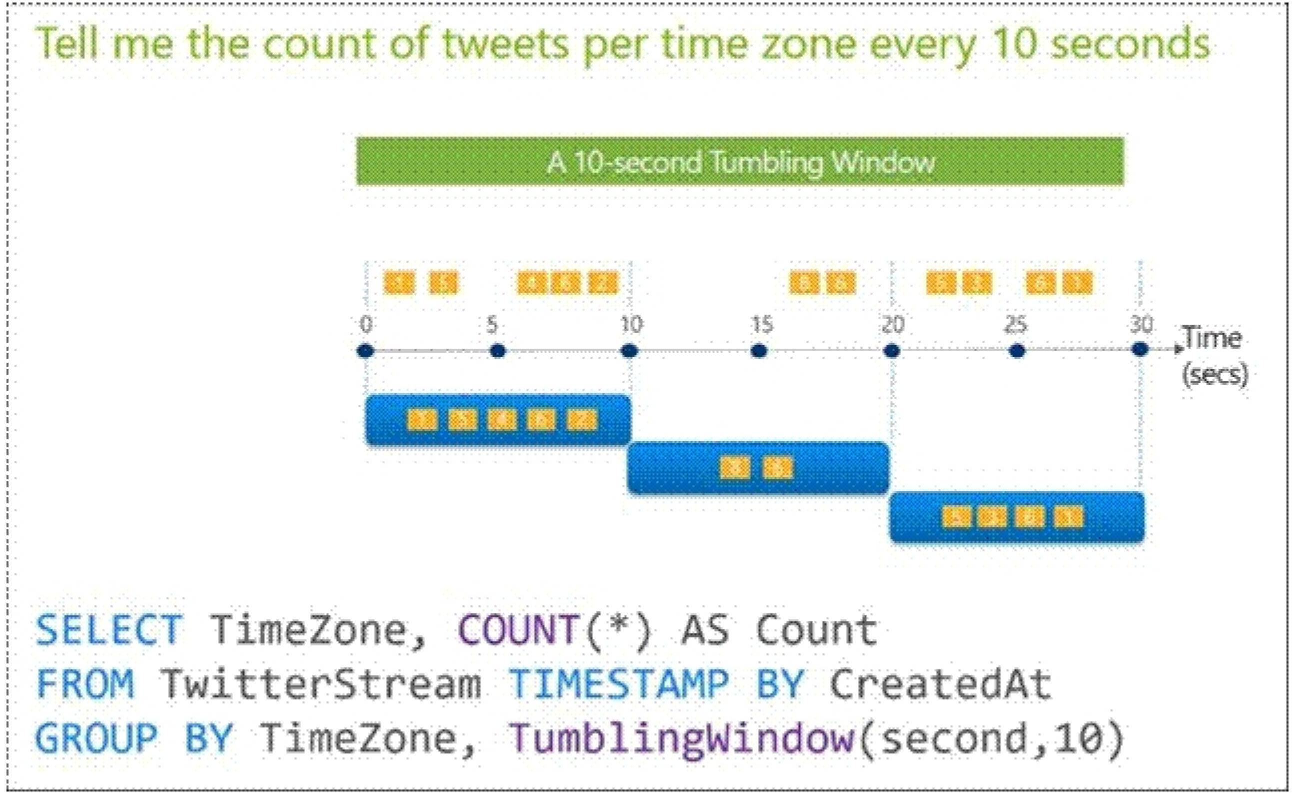


Answer Area

Load methodology:
- Full Load
- Incremental Load
- Load individual files as they arrive

Trigger:
- Fixed schedule
- New file
- Tumbling window

Section:
Explanation:
Box 1: Incremental load
Box 2: Tumbling window
Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals. The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.


Tell me the count of tweets per time zone every 10 seconds
A 10-second Tumbling Window
0
5
10
15
20
25
30
Time (secs)

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

Reference:

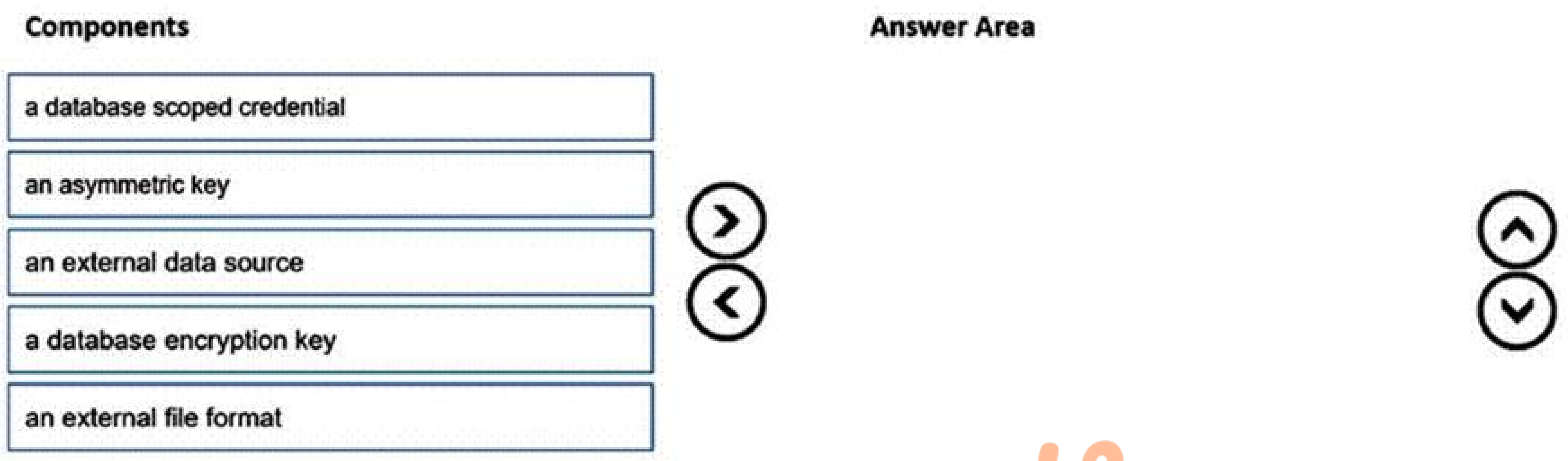
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

QUESTION 48

DRAG DROP

You are responsible for providing access to an Azure Data Lake Storage Gen2 account.
Your user account has contributor access to the storage account, and you have the application ID and access key.
You plan to use PolyBase to load data into an enterprise data warehouse in Azure Synapse Analytics.
You need to configure PolyBase to connect the data warehouse to storage account.
Which three components should you create in sequence? To answer, move the appropriate components from the list of components to the answer area and arrange them in the correct order.

Select and Place:

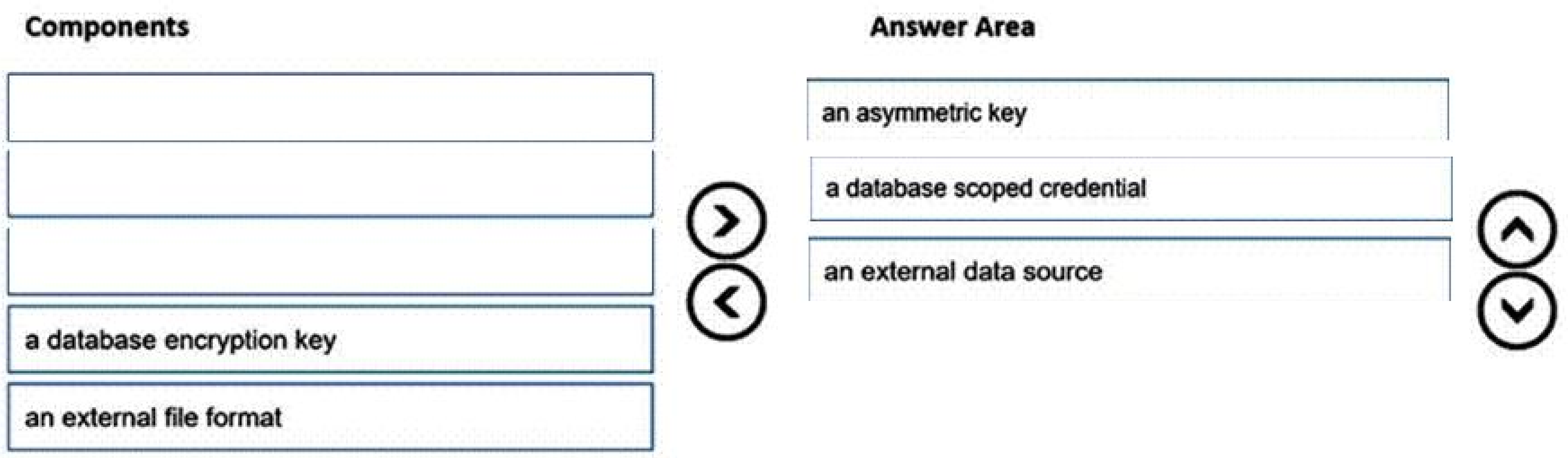


Components

- a database scoped credential
- an asymmetric key
- an external data source
- a database encryption key
- an external file format

Answer Area

Correct Answer:



Components

- a database encryption key
- an external file format

Answer Area

- an asymmetric key
- a database scoped credential
- an external data source

Section:

Explanation:

Step 1: an asymmetric key

A master key should be created only once in a database. The Database Master Key is a symmetric key used to protect the private keys of certificates and asymmetric keys in the database.

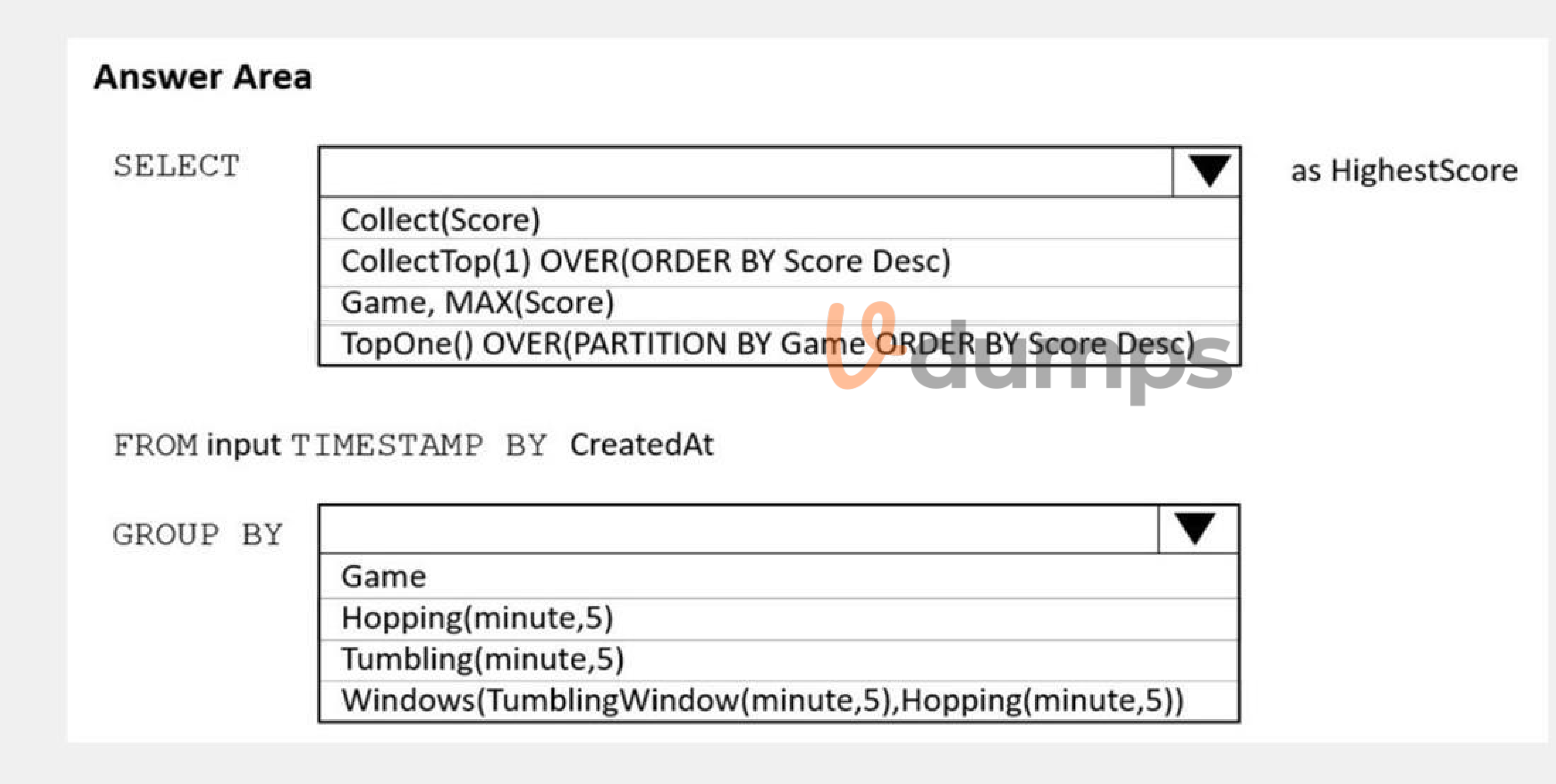
Step 2: a database scoped credential

Create a Database Scoped Credential. A Database Scoped Credential is a record that contains the authentication information required to connect an external resource. The master key needs to be created first before creating the database scoped credential.

Step 3: an external data source
Create an External Data Source. External data sources are used to establish connectivity for data loading using Polybase.
Reference:
https://www.sqlservercentral.com/articles/access-external-data-from-azure-synapse-analytics-using-polybase

QUESTION 49
HOTSPOT
You are building an Azure Stream Analytics job to retrieve game data.
You need to ensure that the job returns the highest scoring record for each five-minute time interval of each game.
How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Hot Area:



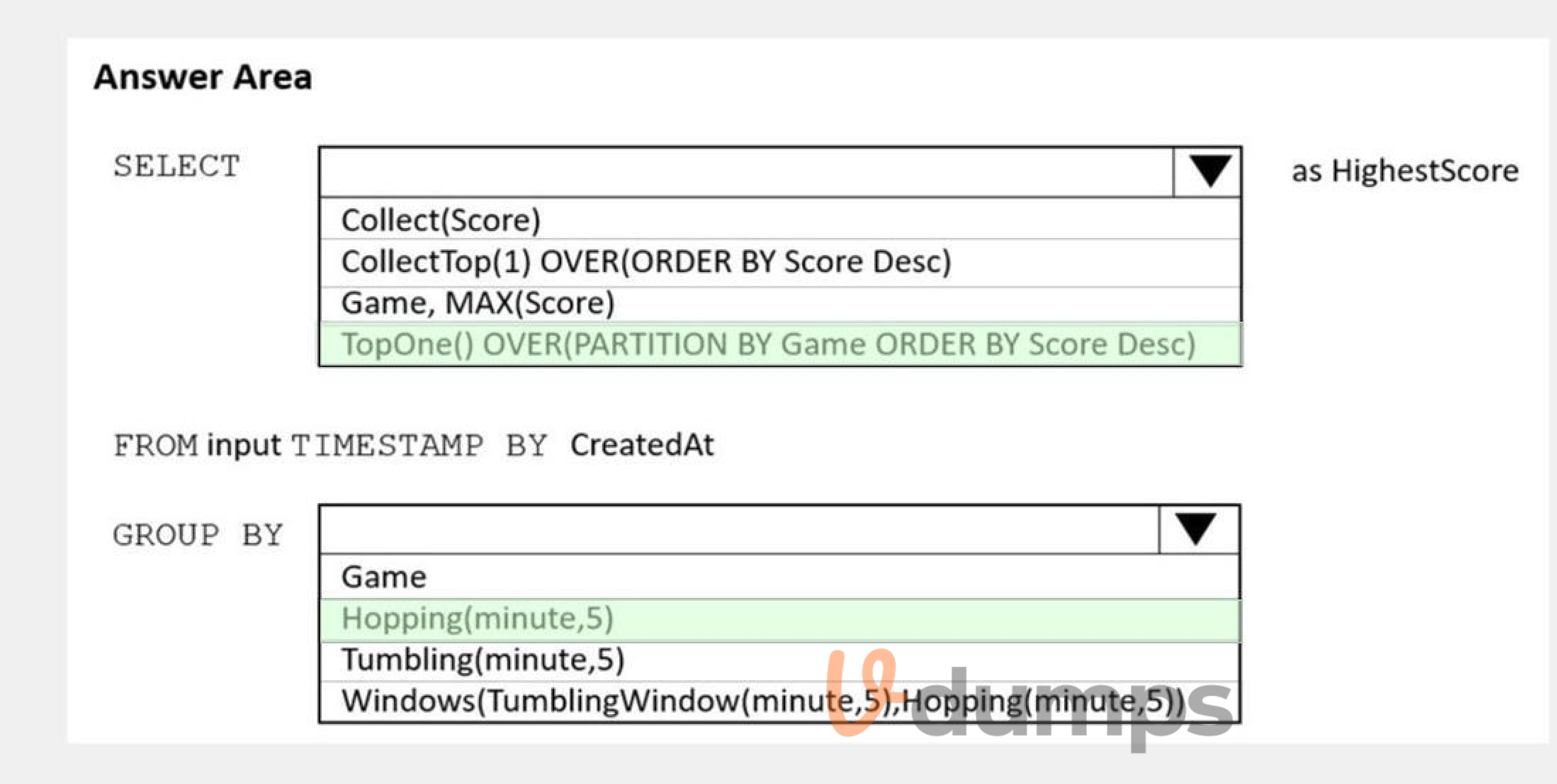
Answer Area

SELECT [▼] as HighestScore
- Collect(Score)
- CollectTop(1) OVER(ORDER BY Score Desc)
- Game, MAX(Score)
- TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)

FROM input TIMESTAMP BY CreatedAt

GROUP BY [▼]
- Game
- Hopping(minute,5)
- Tumbling(minute,5)
- Windows(TumblingWindow(minute,5),Hopping(minute,5))

Answer Area:



Answer Area

SELECT

- Collect(Score)
- CollectTop(1) OVER(ORDER BY Score Desc)
- Game, MAX(Score)
- [✓ TopOne() OVER(PARTITION BY Game ORDER BY Score Desc)]

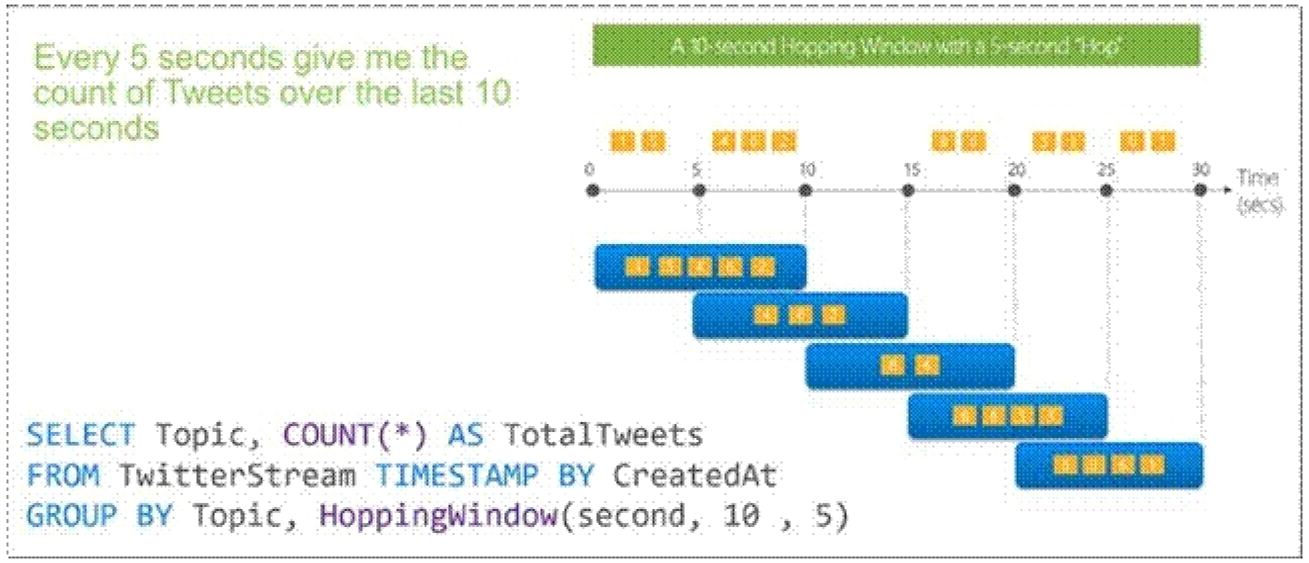
as HighestScore

FROM input TIMESTAMP BY CreatedAt

GROUP BY

- Game
- [✓ Hopping(minute,5)]
- Tumbling(minute,5)
- Windows(TumblingWindow(minute,5),Hopping(minute,5))

Section:
Explanation:
Box 1: TopOne OVER(PARTITION BY Game ORDER BY Score Desc)
TopOne returns the top-rank record, where rank defines the ranking position of the event in the window according to the specified ordering. Ordering/ranking is based on event columns and can be specified in ORDER BY clause.
Box 2: Hopping(minute,5)
Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.



Every 5 seconds give me the count of Tweets over the last 10 seconds
A 10-second Hopping Window with a 5-second 'Hop'
0
5
10
15
20
25
30
Time (secs)

```
SELECT Topic, COUNT(*) AS TotalTweets
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY Topic, HoppingWindow(second, 10 , 5)
```

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

QUESTION 50

HOTSPOT

You are building an Azure Data Factory solution to process data received from Azure Event Hubs, and then ingested into an Azure Data Lake Storage Gen2 container.

The data will be ingested every five minutes from devices into JSON files. The files have the following naming pattern.

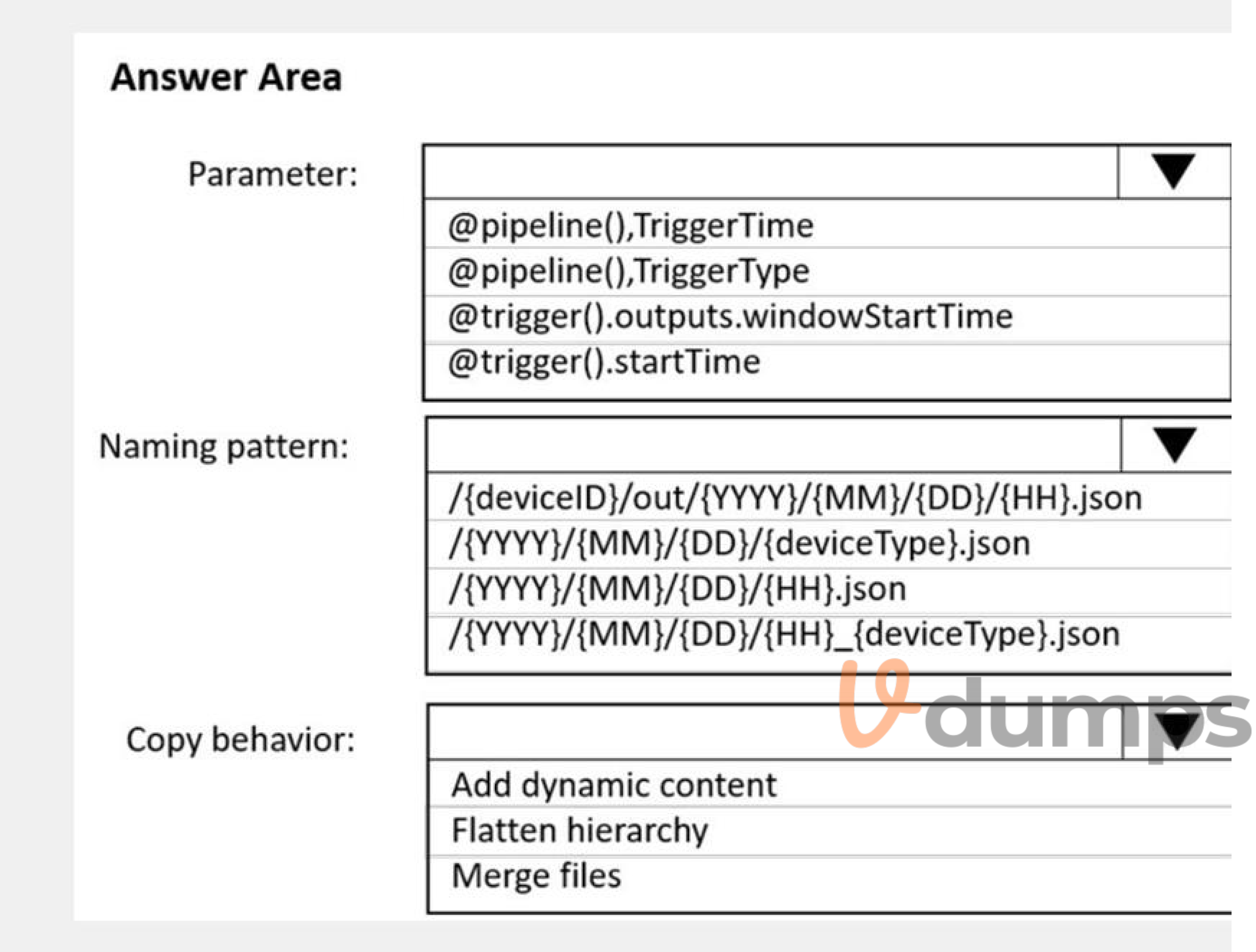
/{deviceType}/in/{YYYY}/{MM}/{DD}/{HH}/{deviceID}_{YYYY}{MM}{DD}HH}{mm}.json

You need to prepare the data for batch data processing so that there is one dataset per hour per deviceType. The solution must minimize read times.

How should you configure the sink for 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

Parameter:

- @pipeline(),TriggerTime
- @pipeline(),TriggerType
- @trigger().outputs.windowStartTime
- @trigger().startTime

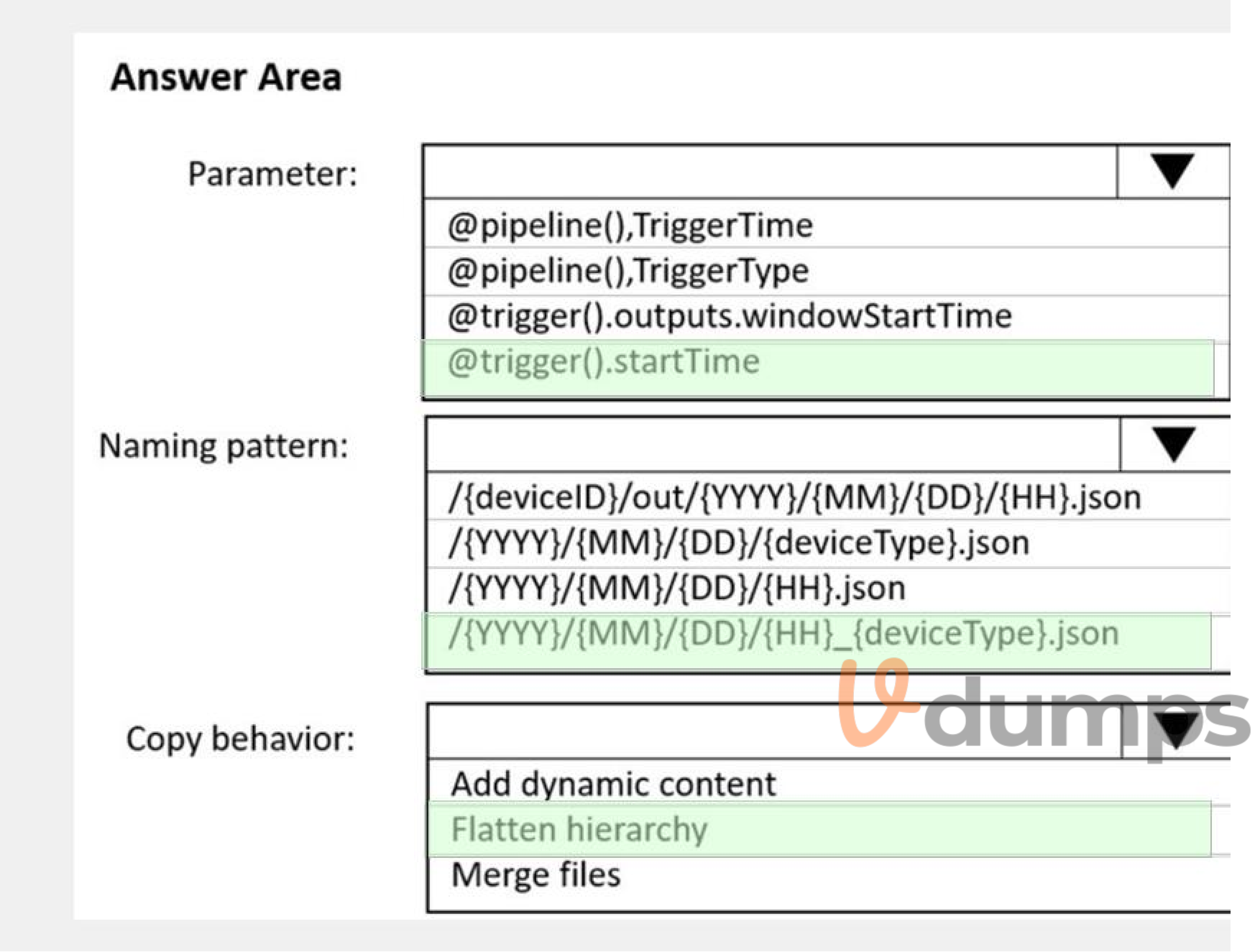
Naming pattern:

- /{deviceID}/out/{YYYY}/{MM}/{DD}/{HH}.json
- /{YYYY}/{MM}/{DD}/{deviceType}.json
- /{YYYY}/{MM}/{DD}/{HH}.json
- /{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json

Copy behavior:

- Add dynamic content
- Flatten hierarchy
- Merge files

Answer Area:



Answer Area

Parameter:

- @pipeline(),TriggerTime
- @pipeline(),TriggerType
- @trigger().outputs.windowStartTime
- [✓ @trigger().startTime]

Naming pattern:

- /{deviceID}/out/{YYYY}/{MM}/{DD}/{HH}.json
- /{YYYY}/{MM}/{DD}/{deviceType}.json
- /{YYYY}/{MM}/{DD}/{HH}.json
- [✓ /{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json]

Copy behavior:

- Add dynamic content
- [✓ Flatten hierarchy]
- Merge files

Section:
Explanation:
Box 1: @trigger().startTime
startTime: A date-time value. For basic schedules, the value of the startTime property applies to the first occurrence. For complex schedules, the trigger starts no sooner than the specified startTime value.
Box 2: /{YYYY}/{MM}/{DD}/{HH}_{deviceType}.json
One dataset per hour per deviceType.
Box 3: Flatten hierarchy
- FlattenHierarchy: All files from the source folder are in the first level of the target folder. The target files have autogenerated names.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers
https://docs.microsoft.com/en-us/azure/data-factory/connector-file-system

QUESTION 51
DRAG DROP
You are designing an Azure Data Lake Storage Gen2 structure for telemetry data from 25 million devices distributed across seven key geographical regions. Each minute, the devices will send a JSON payload of metrics to Azure Event

Hubs.

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

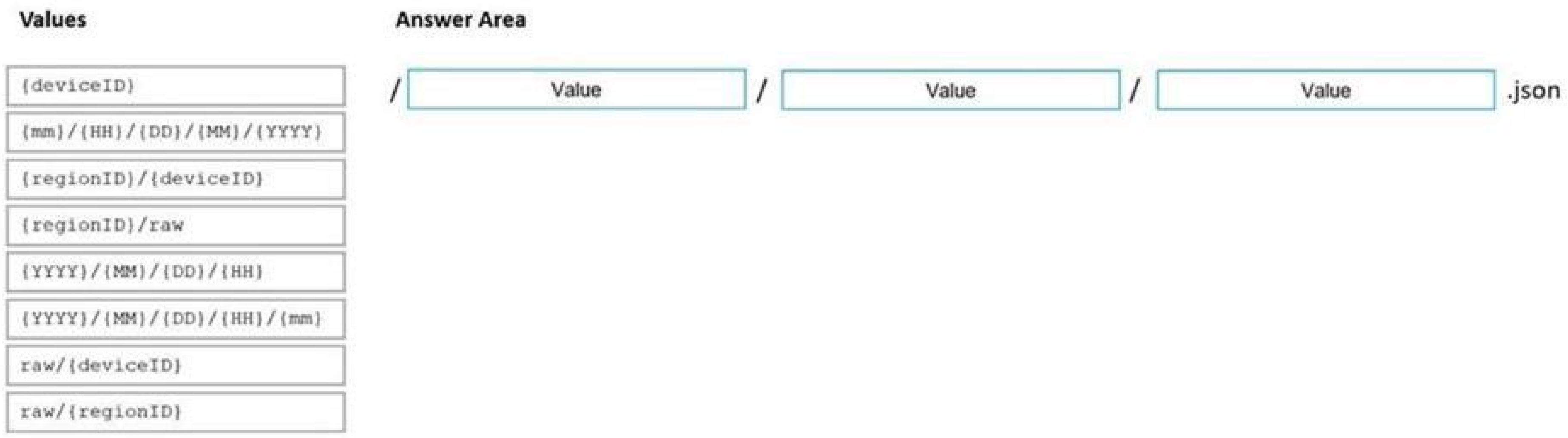
Data engineers from each region must be able to build their own pipelines for the data of their respective region only.

The data must be processed at least once every 15 minutes for inclusion in Azure Synapse Analytics serverless SQL pools.

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

NOTE: Each correct selection is worth one point.

Select and Place:



Values

{deviceID}

{mm}/{HH}/{DD}/{MM}/{YYYY}

{regionID}/{deviceID}

{regionID}/raw

{YYYY}/{MM}/{DD}/{HH}

{YYYY}/{MM}/{DD}/{HH}/{mm}

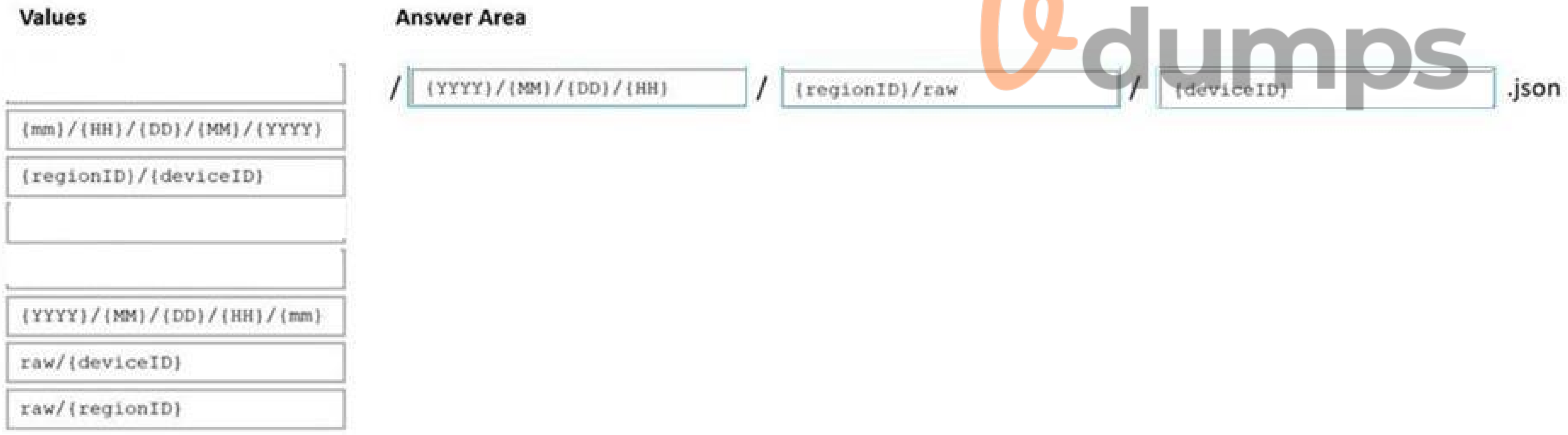
raw/{deviceID}

raw/{regionID}

Answer Area

/ Value / Value / Value .json

Correct Answer:



Values

{mm}/{HH}/{DD}/{MM}/{YYYY}

{regionID}/{deviceID}

{YYYY}/{MM}/{DD}/{HH}/{mm}

raw/{deviceID}

raw/{regionID}

Answer Area

/ {YYYY}/{MM}/{DD}/{HH} / {regionID}/raw / {deviceID} .json

Section:

Explanation:

Box 1: {YYYY}/{MM}/{DD}/{HH}

Date Format [optional]: if the date token is used in the prefix path, you can select the date format in which your files are organized. Example: YYYY/MM/DD Time Format [optional]: if the time token is used in the prefix path, specify the time format in which your files are organized. Currently the only supported value is HH.

Box 2: {regionID}/raw

Data engineers from each region must be able to build their own pipelines for the data of their respective region only.

Box 3: {deviceID}

Reference:

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

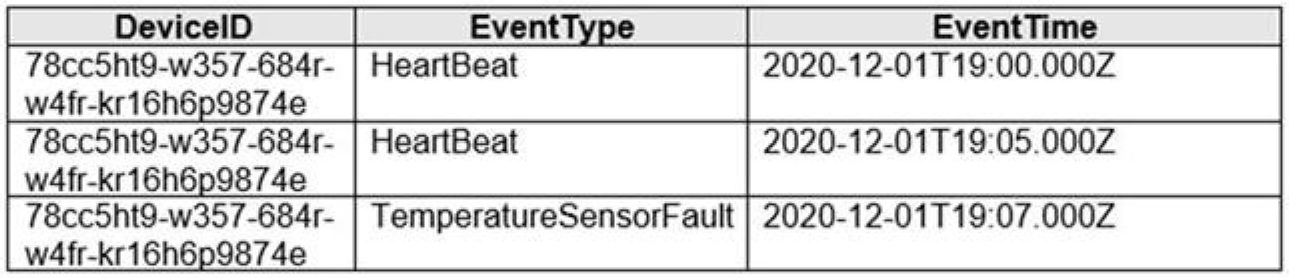
QUESTION 52

HOTSPOT

You are implementing an Azure Stream Analytics solution to process event data from devices.

The devices output events when there is a fault and emit a repeat of the event every five seconds until the fault is resolved. The devices output a heartbeat event every five seconds after a previous event if there are no faults present.

A sample of the events is shown in the following table.



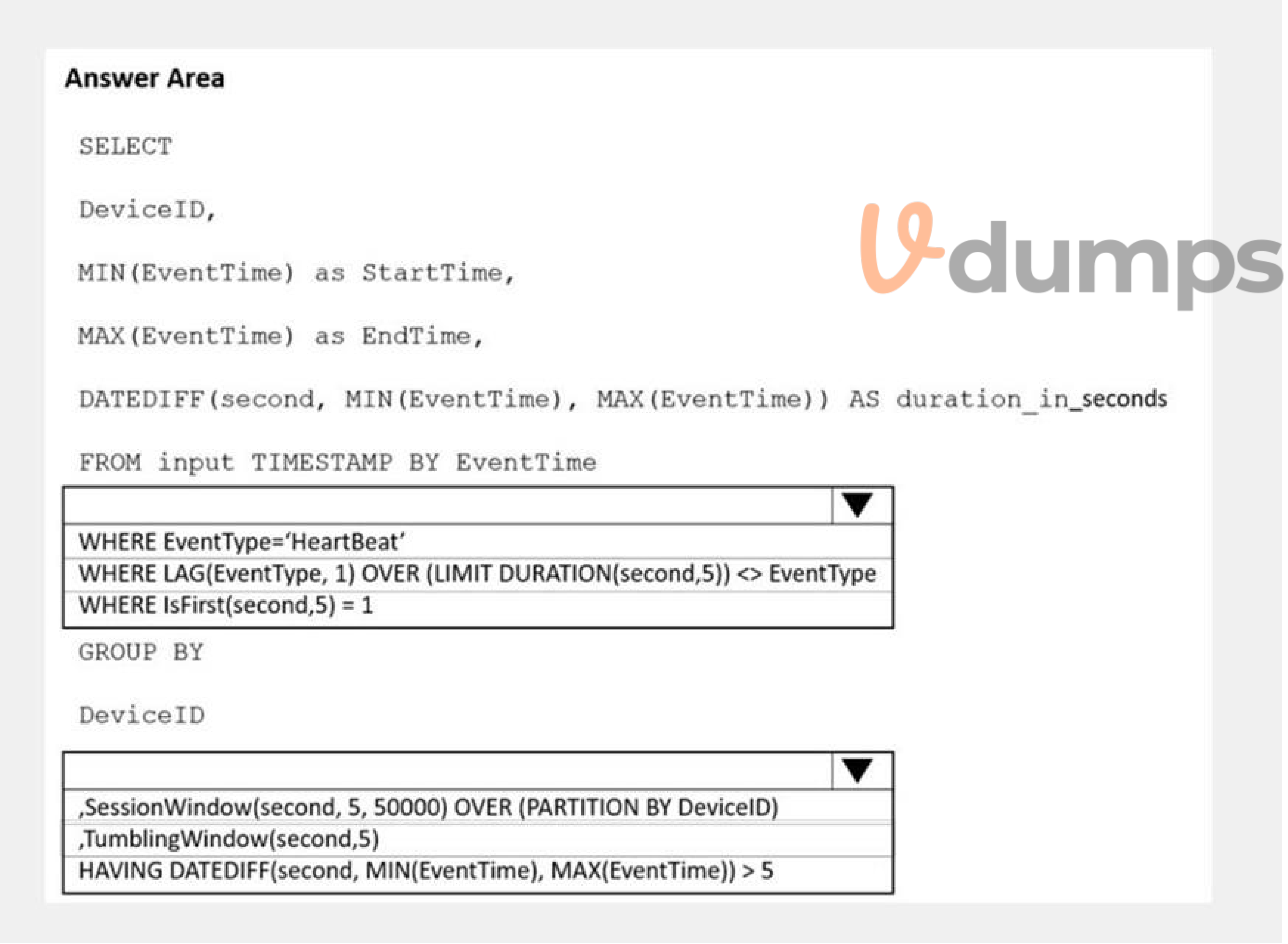
DeviceID	EventType	EventTime
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	HeartBeat	2020-12-01T19:00.000Z
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	HeartBeat	2020-12-01T19:05.000Z
78cc5ht9-w357-684r-w4fr-kr16h6p9874e	TemperatureSensorFault	2020-12-01T19:07.000Z

You need to calculate the uptime between the faults.

How should you complete the Stream Analytics SQL query? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

```
SELECT
DeviceID,
MIN(EventTime) as StartTime,
MAX(EventTime) as EndTime,
DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds
FROM input TIMESTAMP BY EventTime
```

WHERE EventType='HeartBeat'

WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType

WHERE IsFirst(second,5) = 1

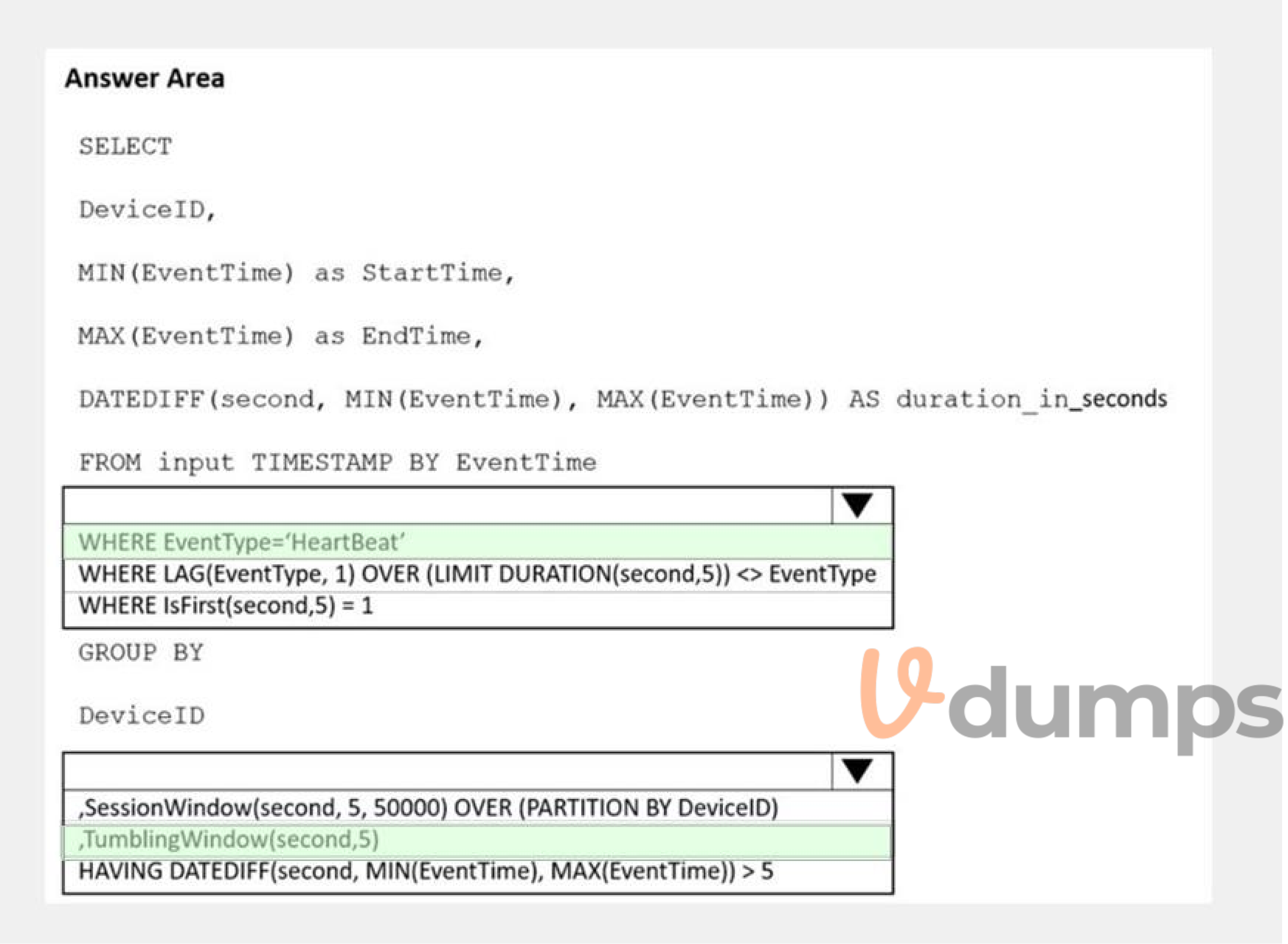
```
GROUP BY
DeviceID
```

,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID)

,TumblingWindow(second,5)

HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5

Answer Area:



Answer Area

```
SELECT

DeviceID,

MIN(EventTime) as StartTime,

MAX(EventTime) as EndTime,

DATEDIFF(second, MIN(EventTime), MAX(EventTime)) AS duration_in_seconds

FROM input TIMESTAMP BY EventTime
```

▼

[✓ WHERE EventType='HeartBeat']

WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType

WHERE IsFirst(second,5) = 1

```
GROUP BY

DeviceID
```

▼

,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID)

[✓ ,TumblingWindow(second,5)]

HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5

Section:

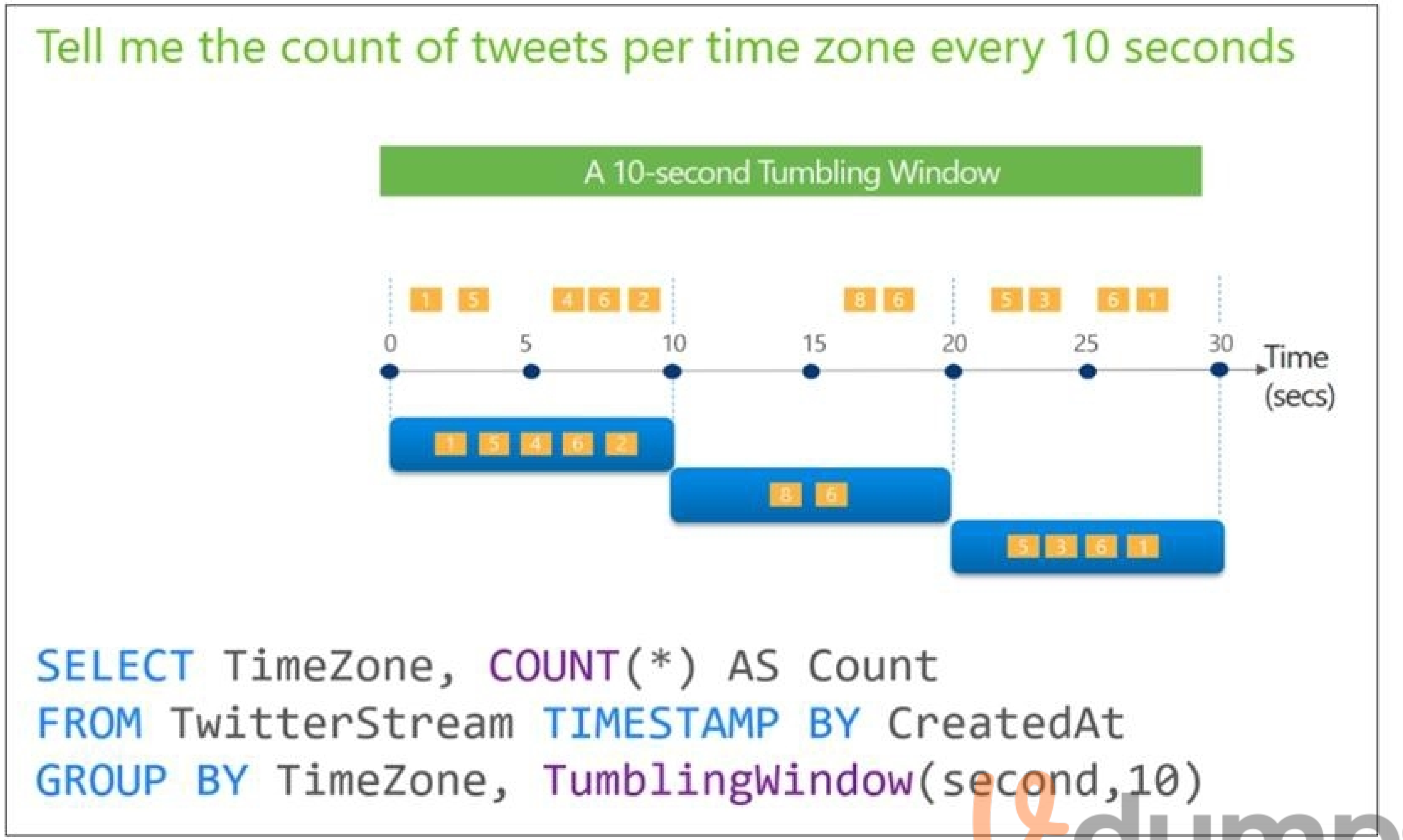
Explanation:

Box 1: WHERE EventType='HeartBeat'

Box 2: ,TumblingWindow(Second, 5)

Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.


Tell me the count of tweets per time zone every 10 seconds

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

Incorrect Answers:
,SessionWindow.. : Session windows group events that arrive at similar times, filtering out periods of time where there is no data.
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/session-window-azure-stream-analytics
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

Exam A

QUESTION 1
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 2
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 3
You have an Azure Data Factory pipeline named pipeline1 that is invoked by a tumbling window trigger named Trigger1. Trigger1 has a recurrence of 60 minutes. You need to ensure that pipeline1 will execute only if the previous execution completes successfully. How should you configure the self-dependency for Trigger1?

A. offset: "-00:01:00" size: "00:01:00"
B. offset: "01:00:00" size: "-01:00:00"
C. offset: "01:00:00" size: "01:00:00"
D. offset: "-01:00:00" size: "01:00:00"

Correct Answer: D
Section:
Explanation:

Tumbling window self-dependency properties
In scenarios where the trigger shouldn't proceed to the next window until the preceding window is successfully completed, build a self-dependency. A self-dependency trigger that's dependent on the success of earlier runs of itself within the preceding hour will have the properties indicated in the following code.
Example code:

```
"name": "DemoSelfDependency",
"properties": {
"runtimeState": "Started",
"pipeline": {
"pipelineReference": {
"referenceName": "Demo",
"type": "PipelineReference"
}
},
"type": "TumblingWindowTrigger",
"typeProperties": {
"frequency": "Hour",
"interval": 1,
"startTime": "2018-10-04T00:00:00Z",
"delay": "00:01:00",
"maxConcurrency": 50,
"retryPolicy": {
"intervalInSeconds": 30
},
"dependsOn": [
{
"type": "SelfDependencyTumblingWindowTriggerReference",
"size": "01:00:00",
"offset": "-01:00:00"
}
]
}
}
}
```

Reference:
https://docs.microsoft.com/en-us/azure/data-factory/tumbling-window-trigger-dependency

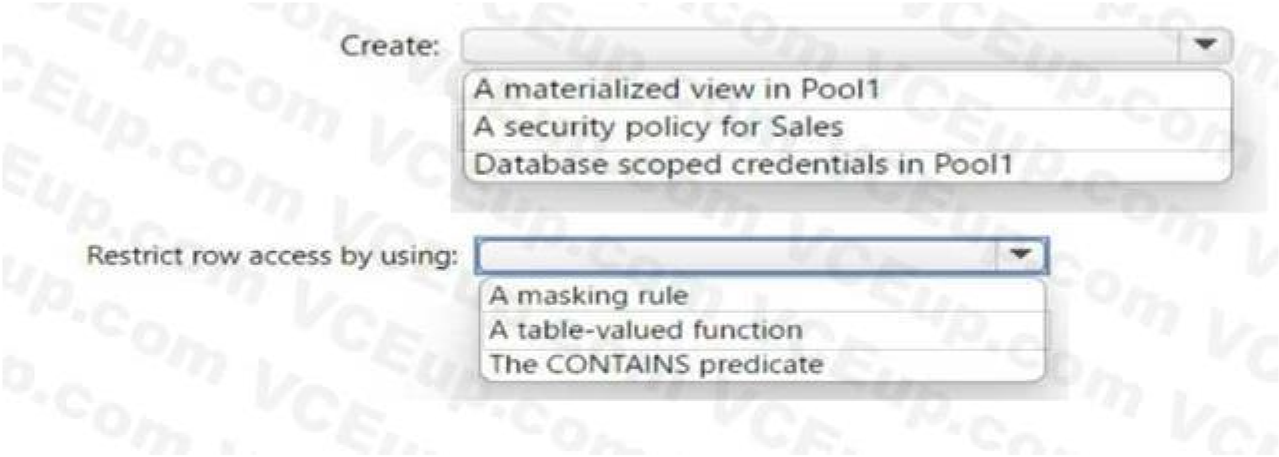
QUESTION 4
You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named SQLPool1. SQLPool1 is currently paused.
You need to restore the current state of SQLPool1 to a new SQL pool. What should you do first?

A. Create a workspace.
B. Create a user-defined restore point.
C. Resume SQLPool1.
D. Create a new SQL pool.

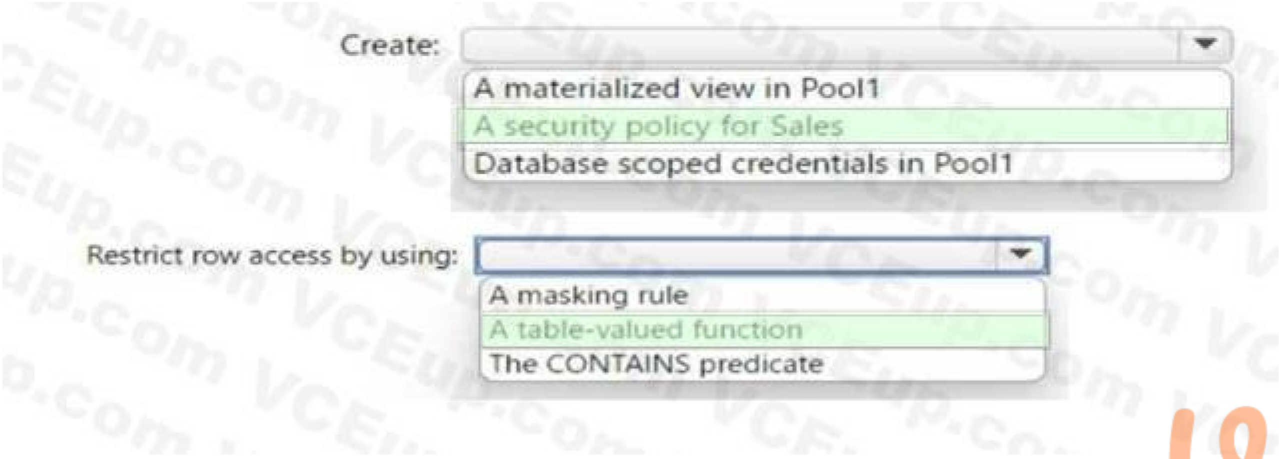
Correct Answer: B
Section:
Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouserestore-active-paused-dw

QUESTION 5
HOTSPOT
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains an external table named Sales. Sales contains sales data. Each row in Sales contains data on a single sale, including the name of the salesperson. You need to implement row-level security (RLS). The solution must ensure that the salespeople can access only their respective sales.
What should you do? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Hot Area:


Create:
A materialized view in Pool1
A security policy for Sales
Database scoped credentials in Pool1
Restrict row access by using:
A masking rule
A table-valued function
The CONTAINS predicate

Answer Area:


Create:
A materialized view in Pool1
A security policy for Sales
Database scoped credentials in Pool1
Restrict row access by using:
A masking rule
A table-valued function
The CONTAINS predicate

Section:
Explanation:
Box 1: A security policy for sale
Here are the steps to create a security policy for Sales:
Create a user-defined function that returns the name of the current user:

```
CREATE FUNCTION dbo.GetCurrentUser()
RETURNS NVARCHAR(128)
AS
BEGIN
RETURN SUSER_SNAME();
END;
```

Create a security predicate function that filters the Sales table based on the current user:

```
CREATE FUNCTION dbo.SalesPredicate(@salesperson NVARCHAR(128)) RETURNS TABLE
WITH SCHEMABINDING
AS
RETURN SELECT 1 AS access_result
WHERE @salesperson = SalespersonName;
```

Create a security policy on the Sales table that uses the SalesPredicate function to filter the data:

```
CREATE SECURITY POLICY SalesFilter
ADD FILTER PREDICATE dbo.SalesPredicate(dbo.GetCurrentUser()) ON dbo.Sales WITH (STATE = ON);
```

By creating a security policy for the Sales table, you ensure that each salesperson can only access their own sales data. The security policy uses a user-defined function to get the name of the current user and a security predicate function to filter the Sales table based on the current user. Box 2: table-value function
to restrict row access by using row-level security, you need to create a table-valued function that returns a table of values that represent the rows that a user can access. You then use this function in a security policy that applies a predicate on the table.

QUESTION 6

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

A. Yes

B. No

Correct Answer: B
Section:
Explanation:

QUESTION 7

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

A. Data 10 percentage

B. CPU percentage

C. DWU used

D. DWU percentage

Correct Answer: C
Section:

QUESTION 8

You have two Azure Blob Storage accounts named account1 and account2?

You plan to create an Azure Data Factory pipeline that will use scheduled intervals to replicate newly created or modified blobs from account1 to account?

You need to recommend a solution to implement the pipeline. The solution must meet the following requirements:

* Ensure that the pipeline only copies blobs that were created of modified since the most recent replication event.

* Minimize the effort to create the pipeline.

What should you recommend?

A. Create a pipeline that contains a flowlet.

B. Create a pipeline that contains a Data Flow activity.

C. Run the Copy Data tool and select Metadata-driven copy task.

D. Run the Copy Data tool and select Built-in copy task.

Correct Answer: A
Section:

QUESTION 9

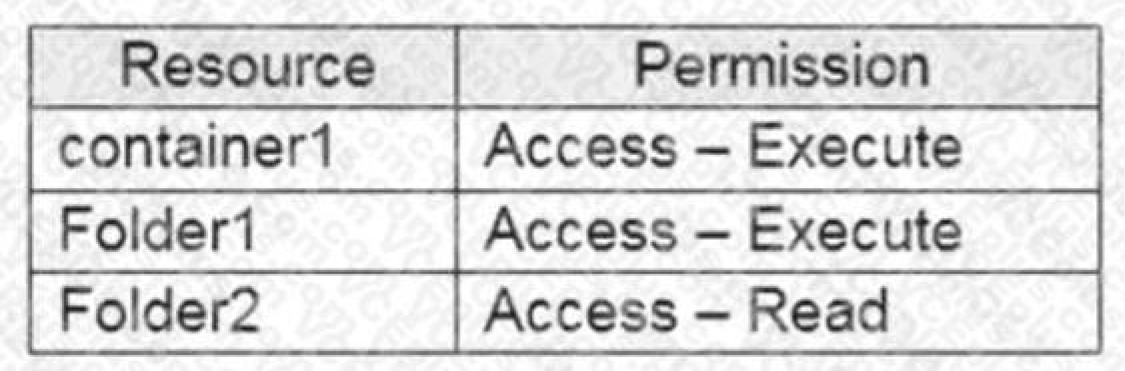
You are designing database for an Azure Synapse Analytics dedicated SQL pool to support workloads for detecting ecommerce transaction fraud. Data will be combined from multiple ecommerce sites and can include sensitive financial information such as credit card numbers. You need to recommend a solution 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 10
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

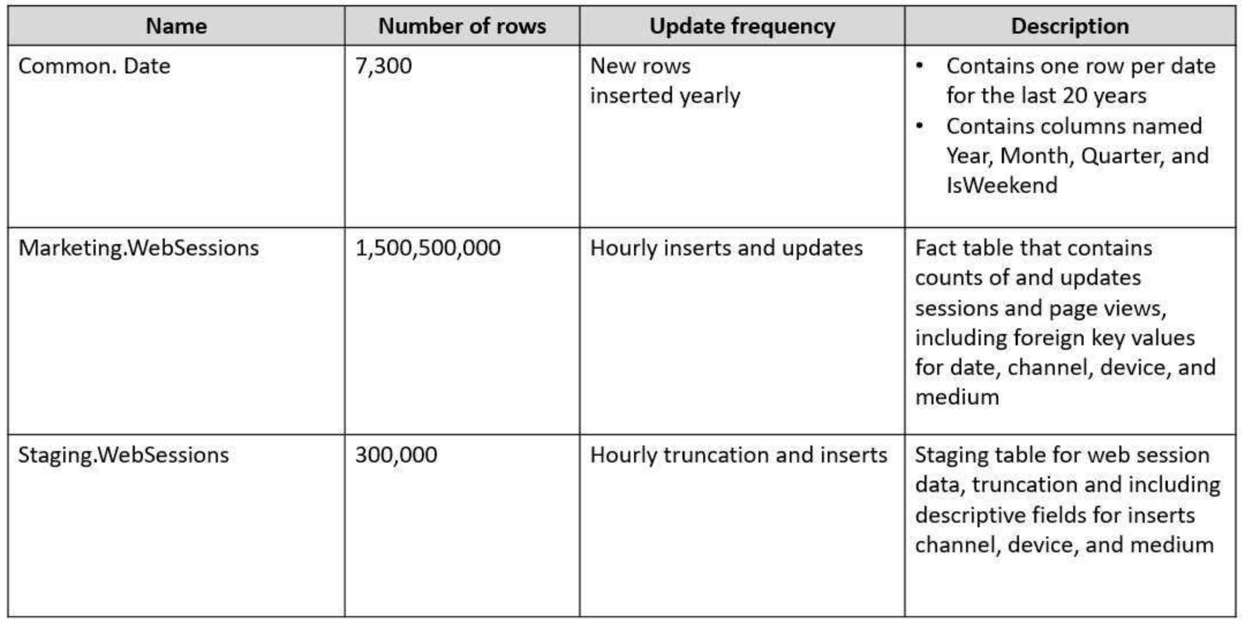
You need to ensure that ServicePrincipal1 can perform the following actions:
Traverse child items that are created in Folder2. Read files that are created in Folder2. The solution must use the principle of least privilege.
Which two permissions should you grant to ServicePrincipal1 for Folder2? Each 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 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:

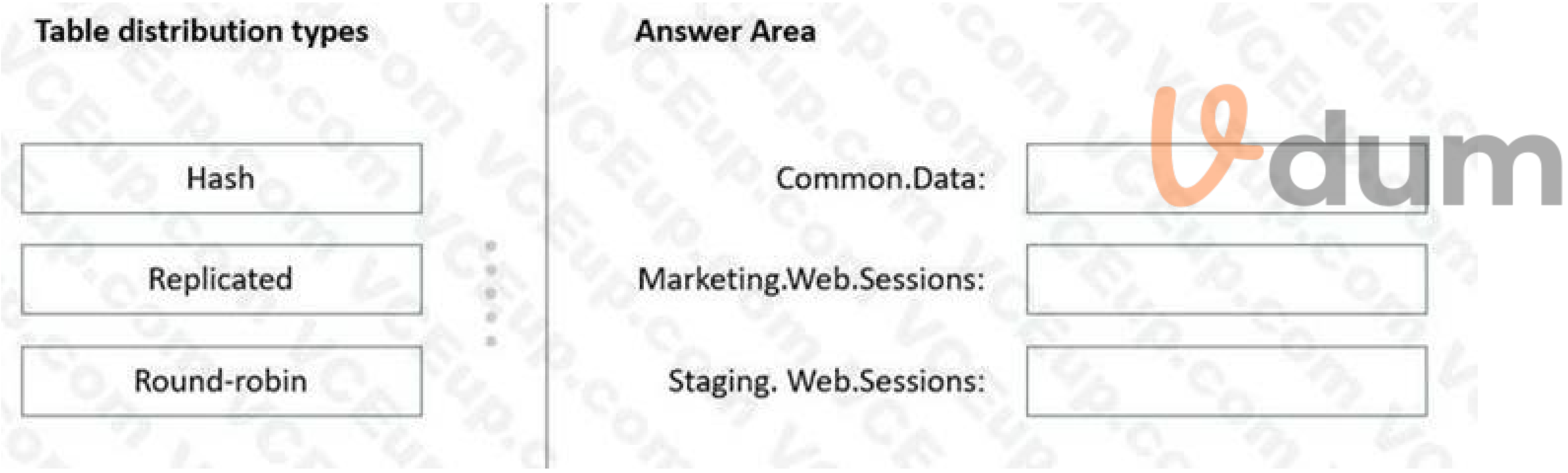


Table distribution types

Hash

Replicated

Round-robin

Answer Area

Common.Data:

Marketing.Web.Sessions:

Staging. Web.Sessions:

Correct Answer:

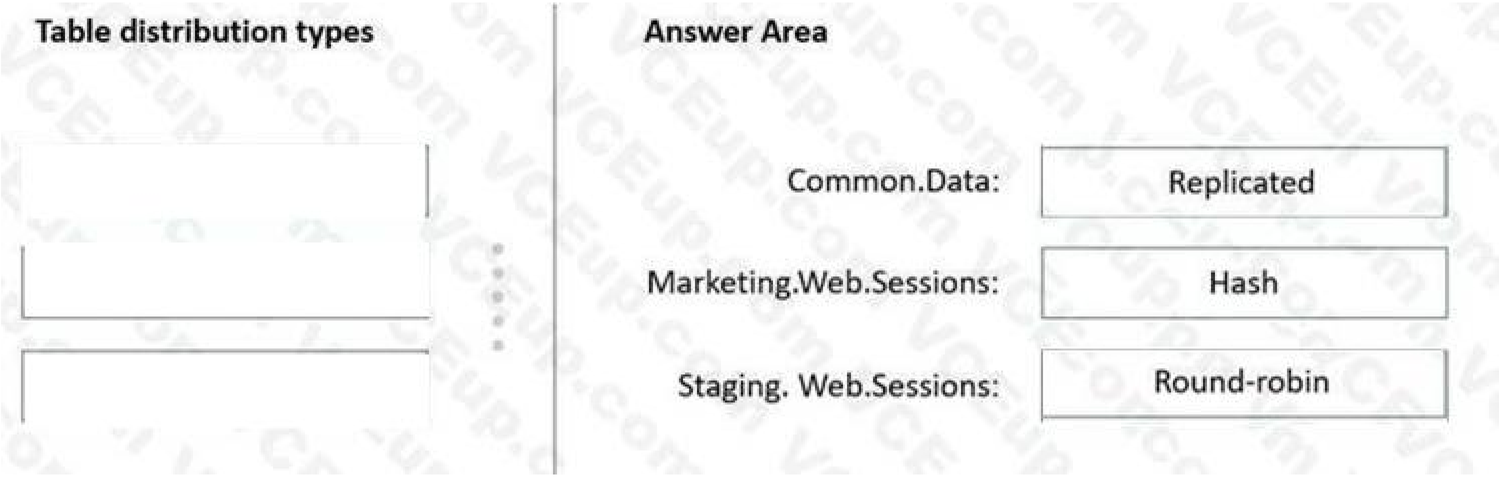


Table distribution types

Answer Area

Common.Data: Replicated

Marketing.Web.Sessions: Hash

Staging. Web.Sessions: Round-robin

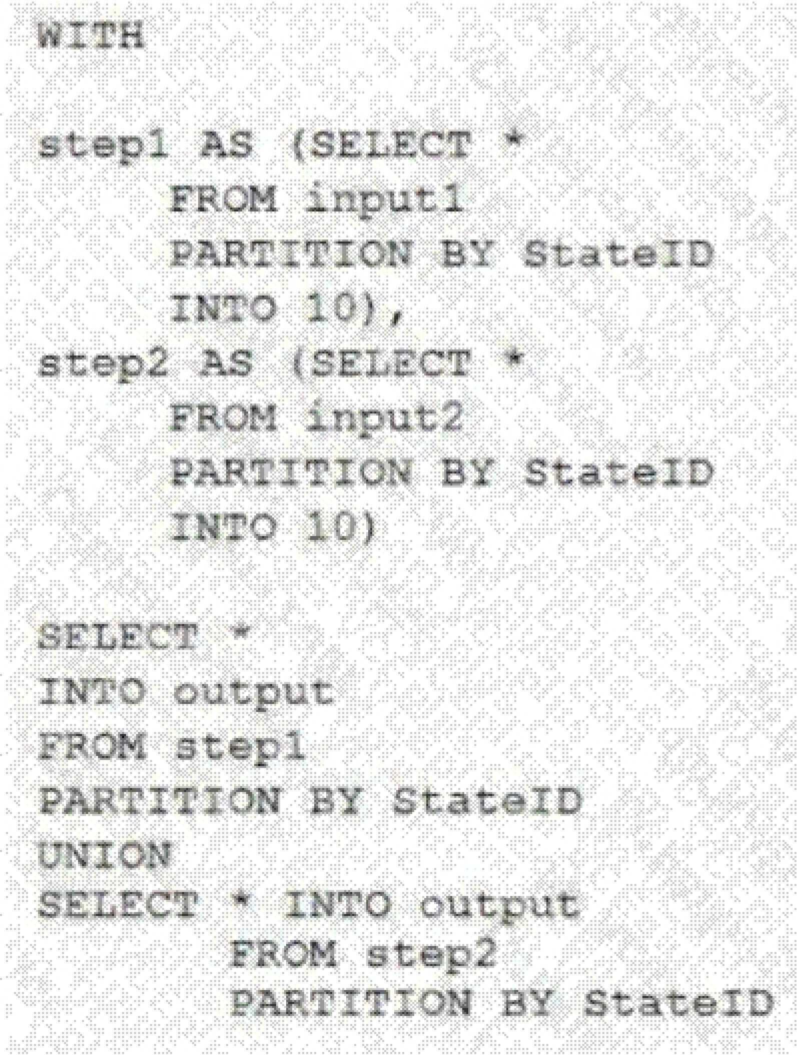
Section:
Explanation:
Box 1: Replicated

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

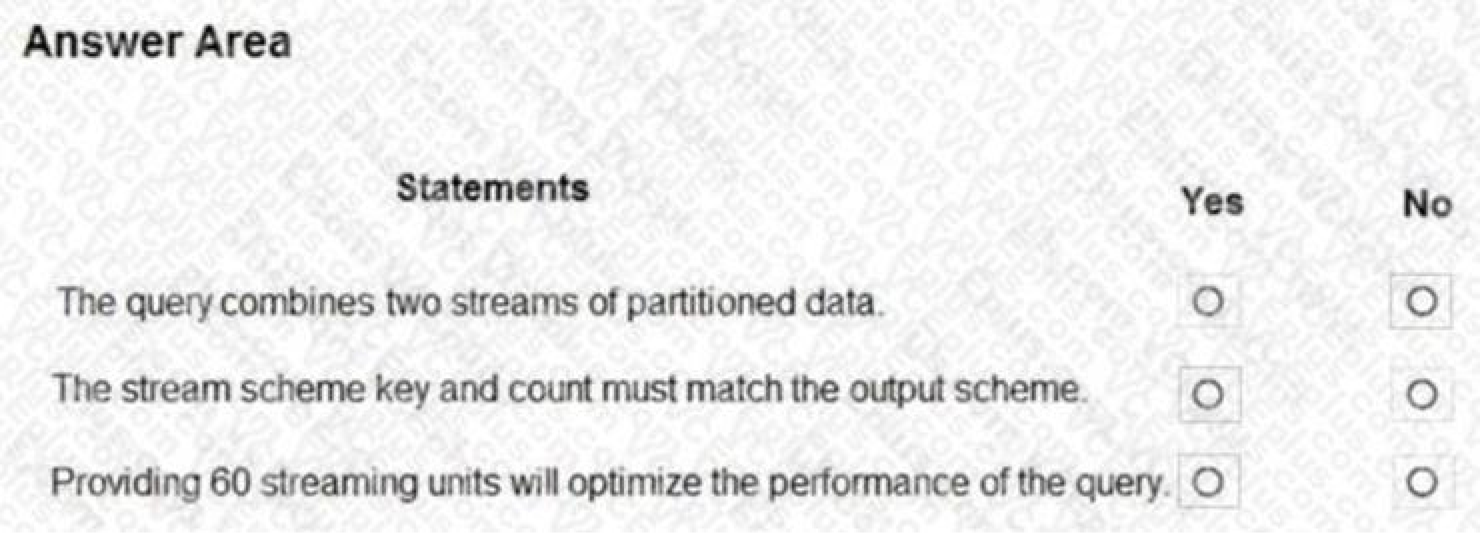
You have the following Azure Stream Analytics query.



```
WITH

step1 AS (SELECT *
     FROM input1
     PARTITION BY StateID
     INTO 10),
step2 AS (SELECT *
     FROM input2
     PARTITION BY StateID
     INTO 10)

SELECT *
INTO output
FROM step1
PARTITION BY StateID
UNION
SELECT * INTO output
       FROM step2
       PARTITION BY StateID
```

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

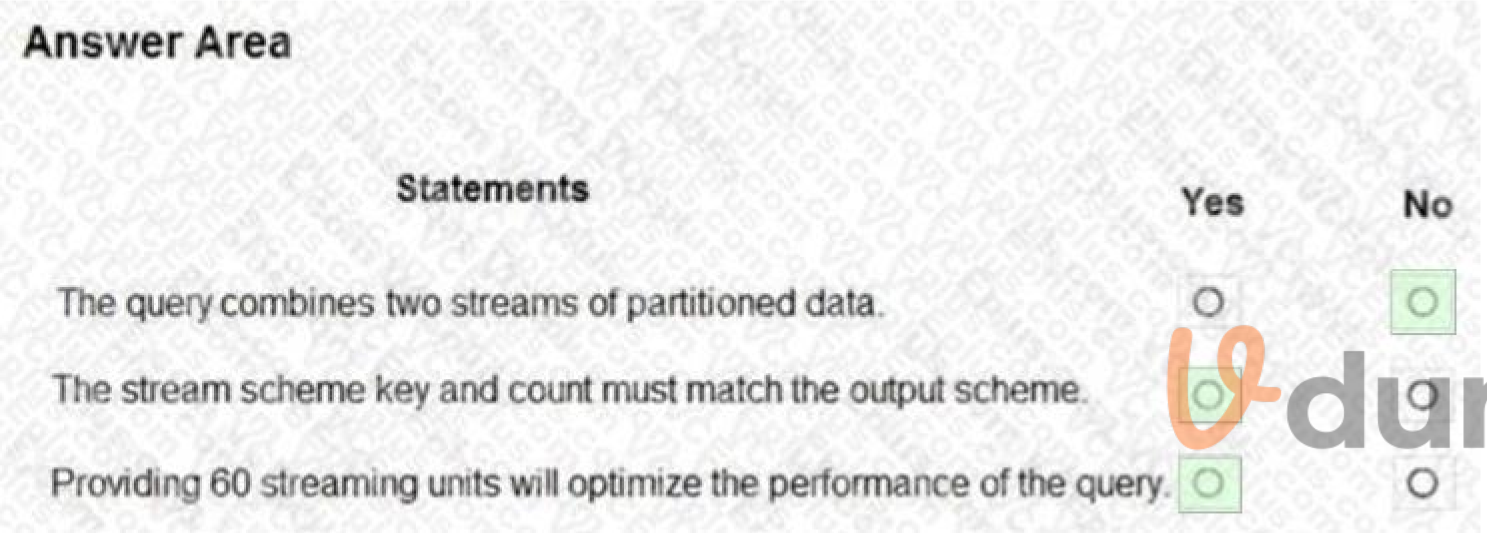
Hot Area:



Answer Area

Statements	Yes	No
The query combines two streams of partitioned data.	○	○
The stream scheme key and count must match the output scheme.	○	○
Providing 60 streaming units will optimize the performance of the query.	○	○

Answer Area:



Answer Area

Statements	Yes	No
The query combines two streams of partitioned data.	○	[✓ ○]
The stream scheme key and count must match the output scheme.	[✓ ○]	○
Providing 60 streaming units will optimize the performance of the query.	[✓ ○]	○

Section:

Explanation:

Box 1: No

Note: You can now use a new extension of Azure Stream Analytics SQL to specify the number of partitions of a stream when reshuffling the data.

The outcome is a stream that has the same partition scheme. Please see below for an example:

WITH step1 AS (SELECT * FROM [input1] PARTITION BY DeviceID INTO 10), step2 AS (SELECT * FROM [input2] PARTITION BY DeviceID INTO 10)

SELECT * INTO [output] FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceID Note: The new extension of Azure Stream Analytics SQL includes a keyword INTO that allows you to specify the number of partitions for a stream when performing reshuffling using a PARTITION BY statement. Box 2: Yes

When joining two streams of data explicitly repartitioned, these streams must have the same partition key and partition count.

Box 3: Yes

Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job. In general, the best practice is to start with 6 SUs for queries that don't use PARTITION BY. Here there are 10 partitions, so 6x10 = 60 SUs is good.

Note: Remember, Streaming Unit (SU) count, which is the unit of scale for Azure Stream Analytics, must be adjusted so the number of physical resources available to the job can fit the partitioned flow. In general, six SUs is a good number to assign to each partition. In case there are insufficient resources assigned to the job, the system will only apply the repartition if it benefits the job.

Reference:

https://azure.microsoft.com/en-in/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/ https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption

QUESTION 13

HOTSPOT

You are building a database in an Azure Synapse Analytics serverless SQL pool. You have data stored in Parquet files in an Azure Data Lake Storege Gen2 container. Records are structured as shown in the following sample.

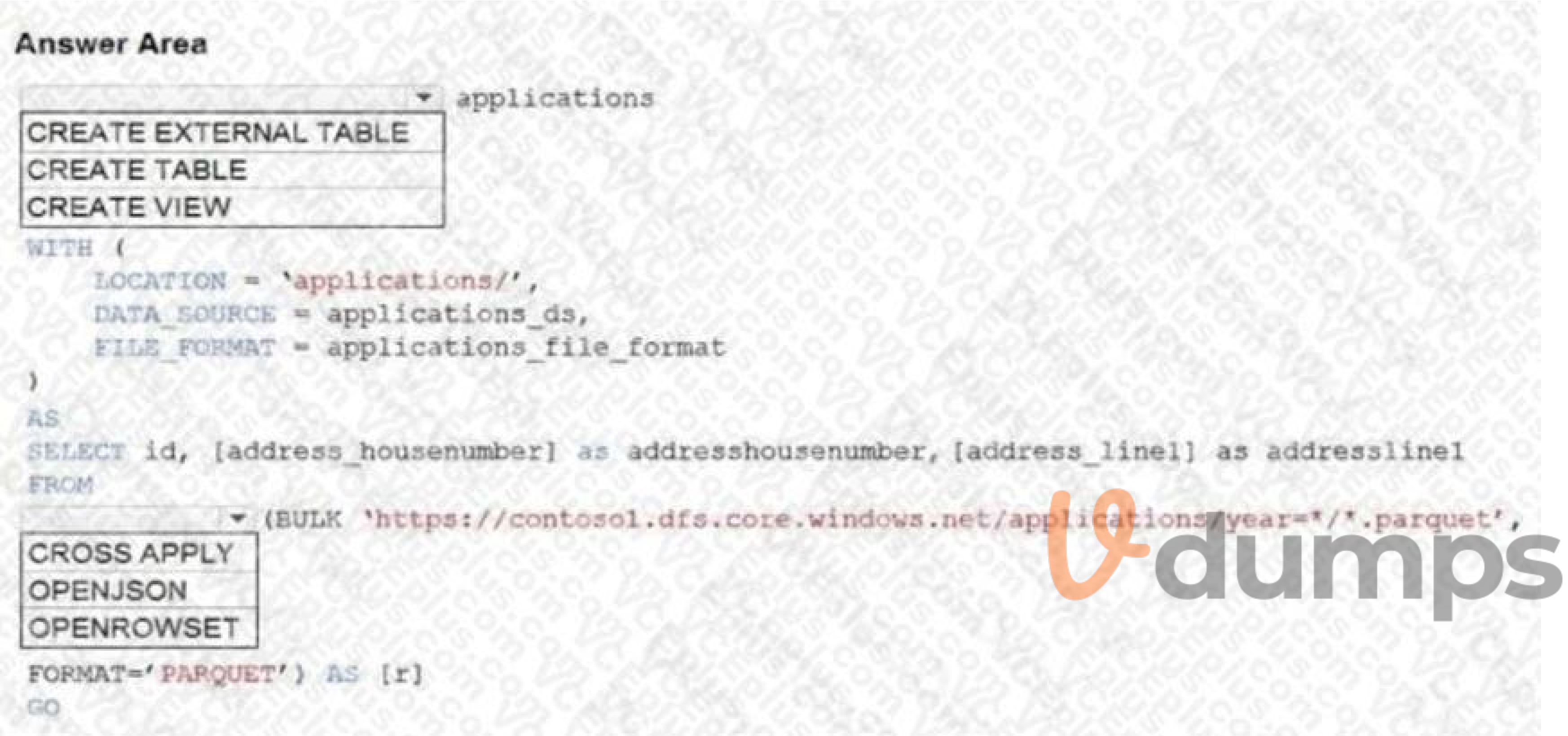
{

"id": 123,
"address_housenumber": "19c",
"address_line": "Memory Lane",
"applicant1_name": "Jane",
"applicant2_name": "Dev" }

The records contain two applicants at most.

You need to build a table that includes only the address fields. 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.

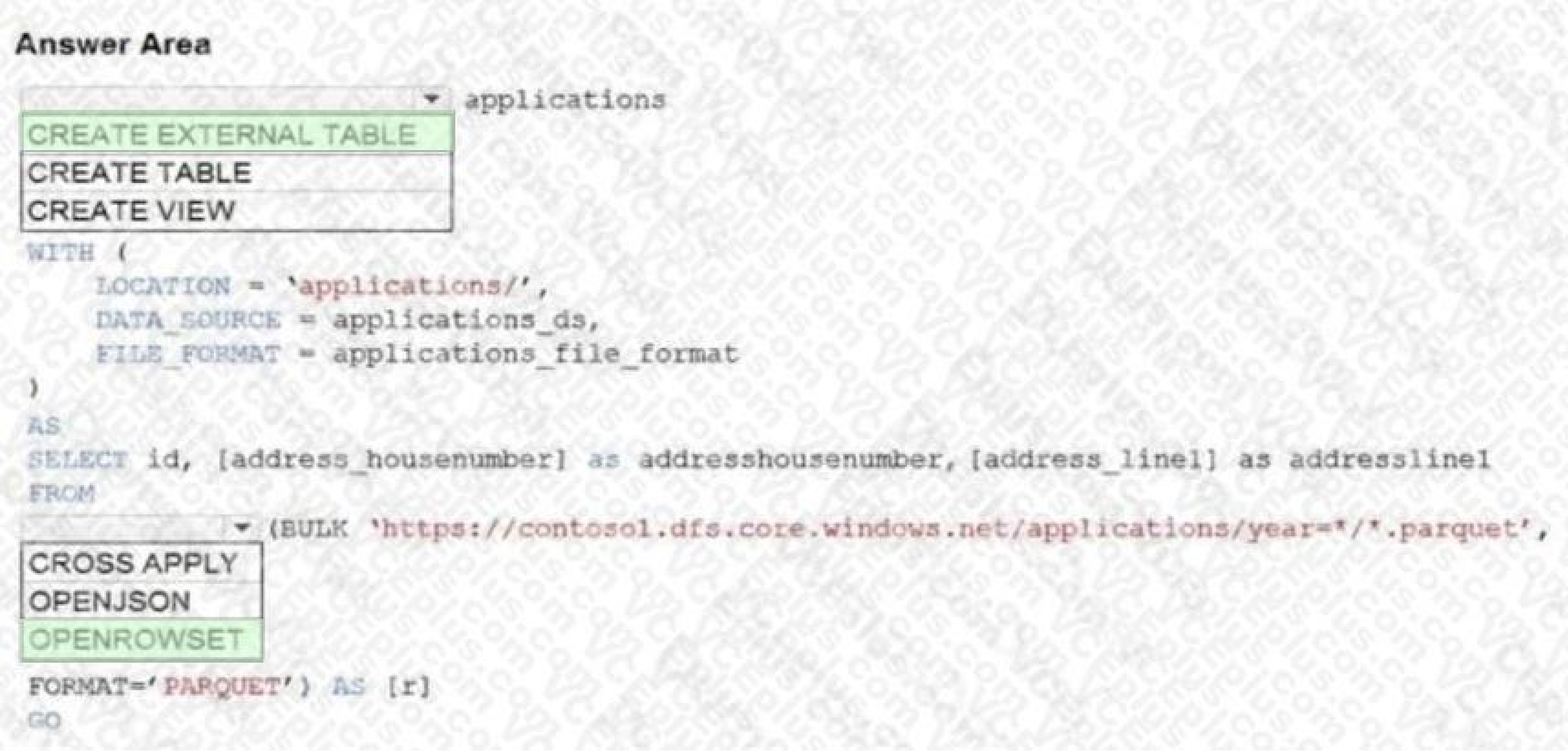
Hot Area:



Answer Area

```
[                    ▼] applications
CREATE EXTERNAL TABLE
CREATE TABLE
CREATE VIEW
WITH (
     LOCATION = 'applications/',
     DATA_SOURCE = applications_ds,
     FILE_FORMAT = applications_file_format
)
AS
SELECT id, [address_housenumber] as addresshousenumber, [address_line1] as addressline1
FROM
[           ▼] (BULK 'https://contoso1.dfs.core.windows.net/applications/year=*/*.parquet',
CROSS APPLY
OPENJSON
OPENROWSET
FORMAT='PARQUET') AS [r]
GO
```

Answer Area:



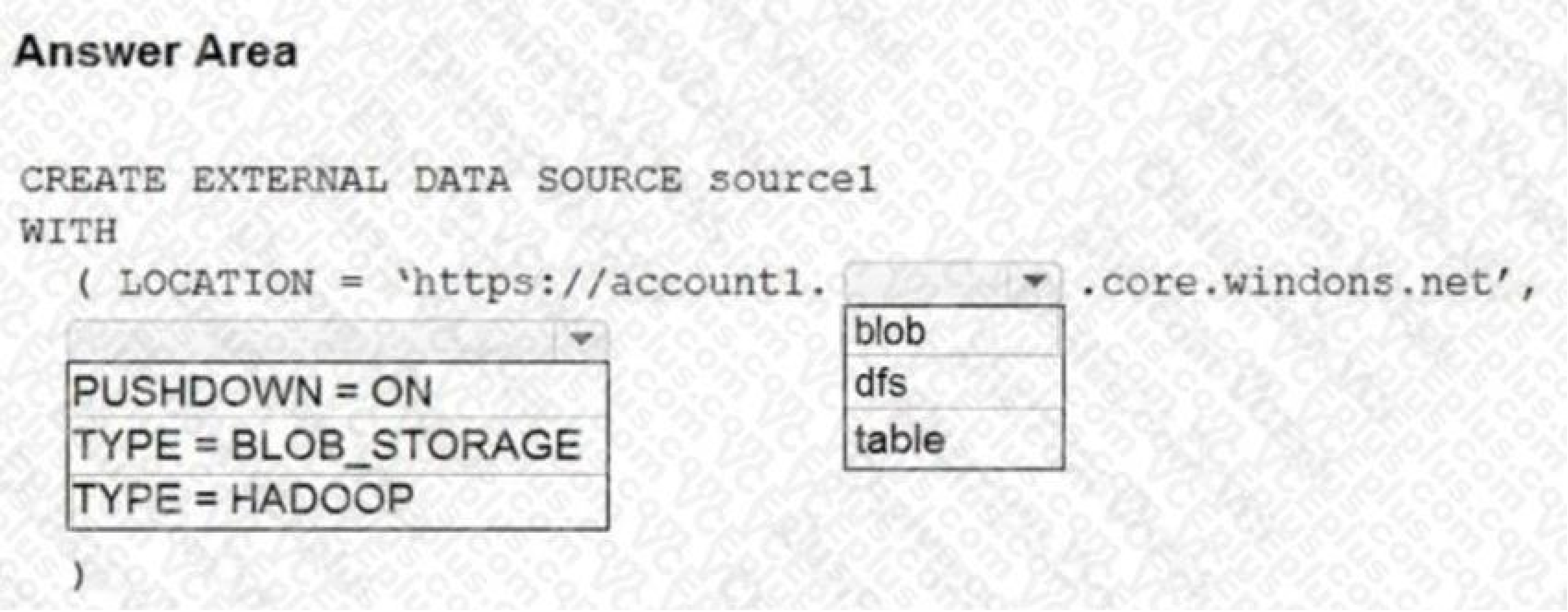
Answer Area

```
[                      ▾] applications
[✓ CREATE EXTERNAL TABLE]
CREATE TABLE
CREATE VIEW
WITH (
    LOCATION = 'applications/',
    DATA_SOURCE = applications_ds,
    FILE_FORMAT = applications_file_format
)
AS
SELECT id, [address_housenumber] as addresshousenumber, [address_line1] as addressline1
FROM
[            ▾] (BULK 'https://contoso1.dfs.core.windows.net/applications/year=*/*.parquet',
CROSS APPLY
OPENJSON
[✓ OPENROWSET]
FORMAT='PARQUET') AS [r]
GO
```

Section:
Explanation:
Box 1: CREATE EXTERNAL TABLE
An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool. Syntax:
CREATE EXTERNAL TABLE { database_name.schema_name.table_name | schema_name.table_name | table_name } ([,...n])
WITH (
LOCATION = 'folder_or_filepath',
DATA_SOURCE = external_data_source_name,
FILE_FORMAT = external_file_format_name
Box 2. OPENROWSET
When using serverless SQL pool, CETAS is used to create an external table and export query results to Azure Storage Blob or Azure Data Lake Storage Gen2.
Example:
AS
SELECT decennialTime, stateName, SUM(population) AS population FROM
OPENROWSET(BULK
'https://azureopendatastorage.blob.core.windows.net/censusdatacontainer/release/us_population_county/year=*/*.parquet', FORMAT='PARQUET') AS [r]
GROUP BY decennialTime, stateName GO
Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

QUESTION 14
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.

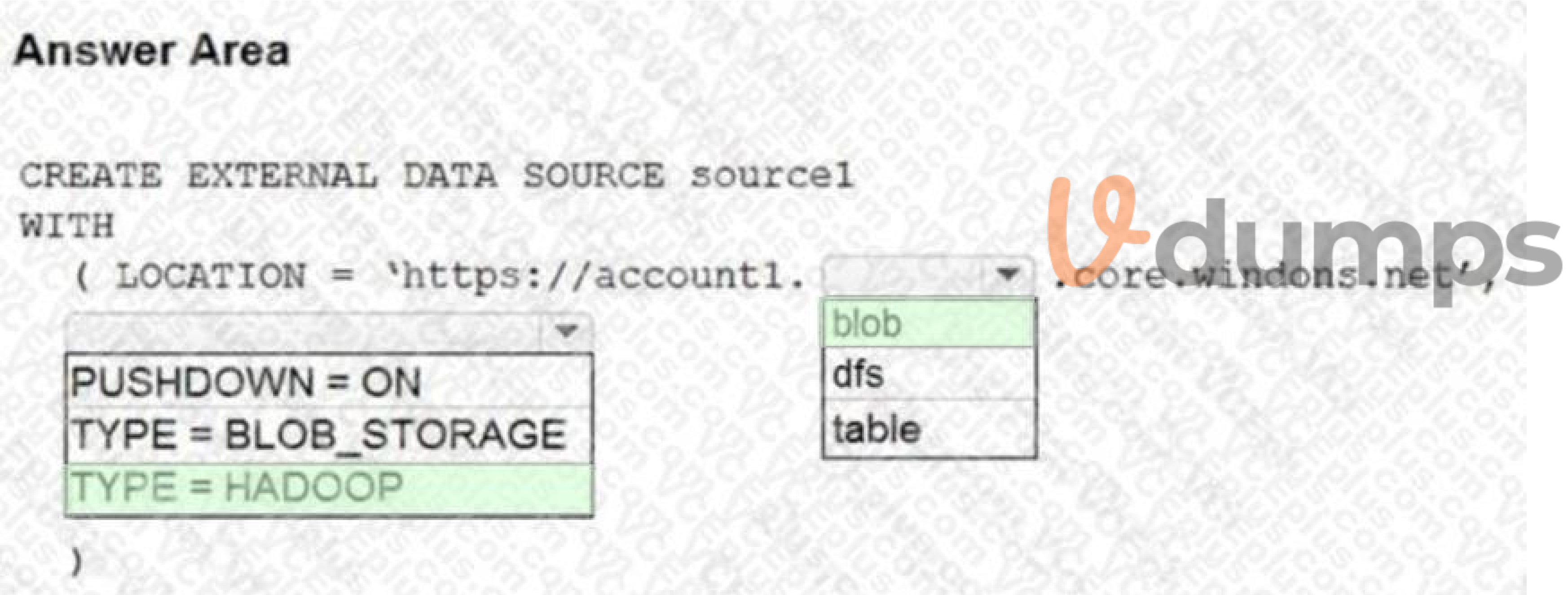
Hot Area:



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

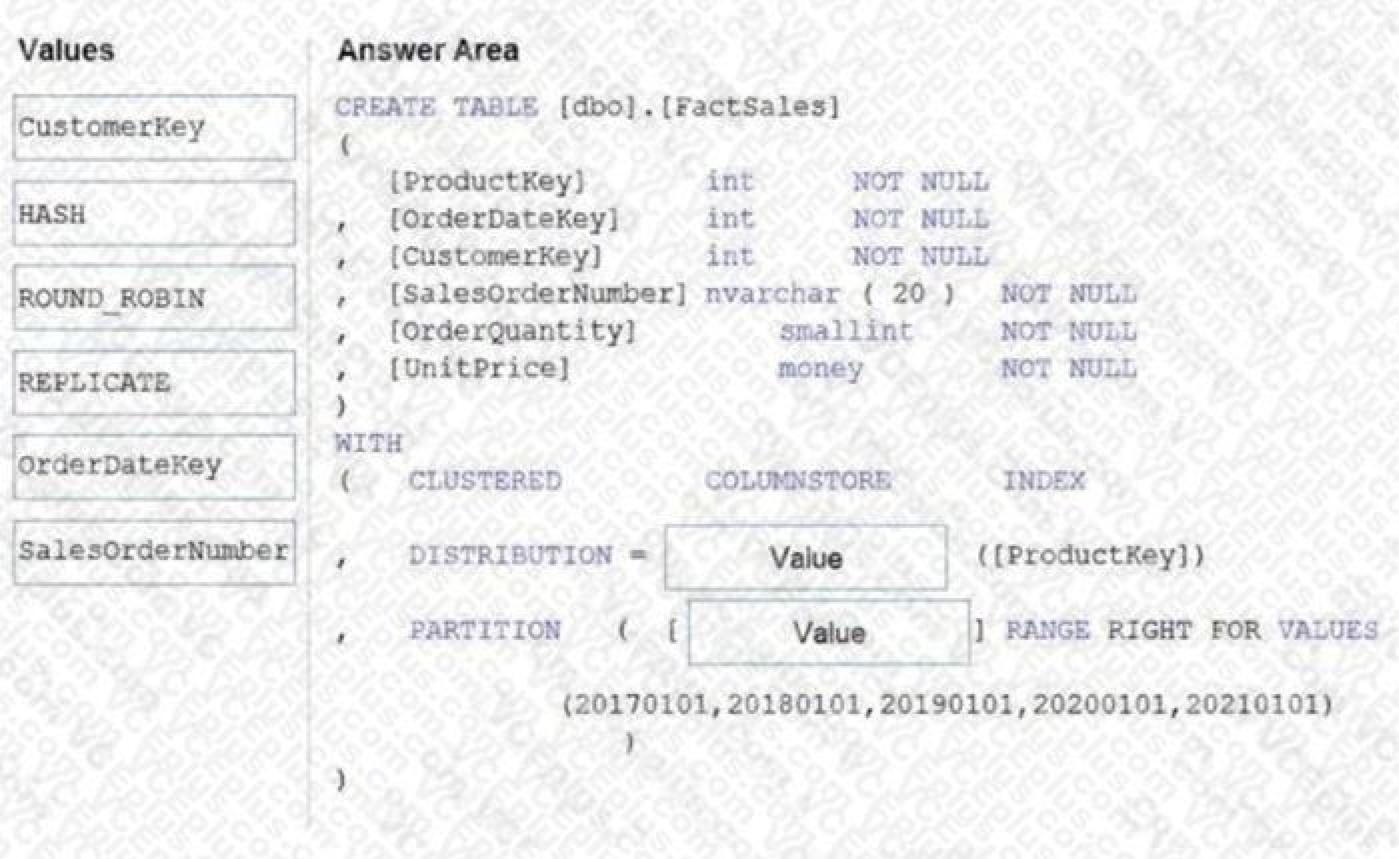
QUESTION 15

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:



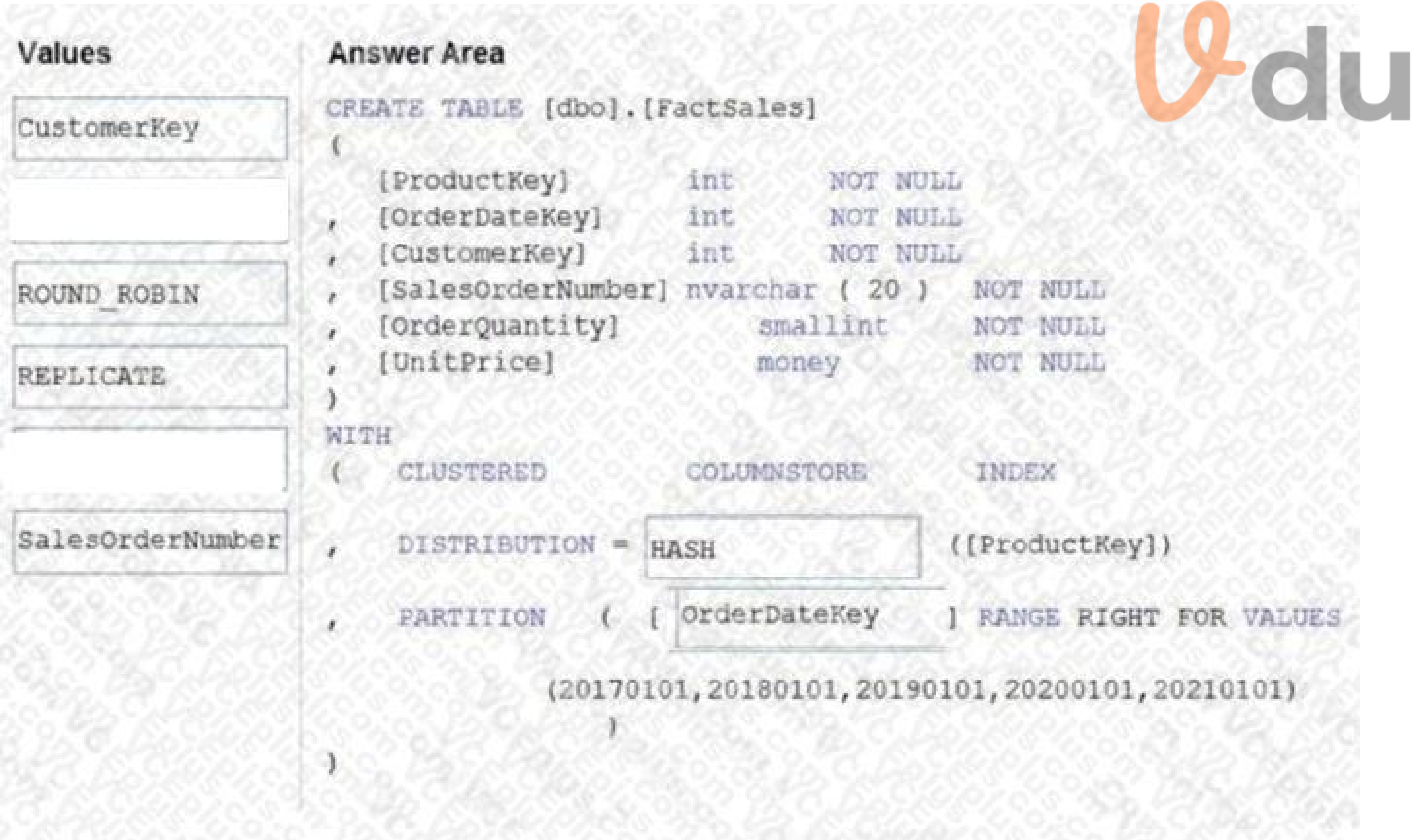
Values

- 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:

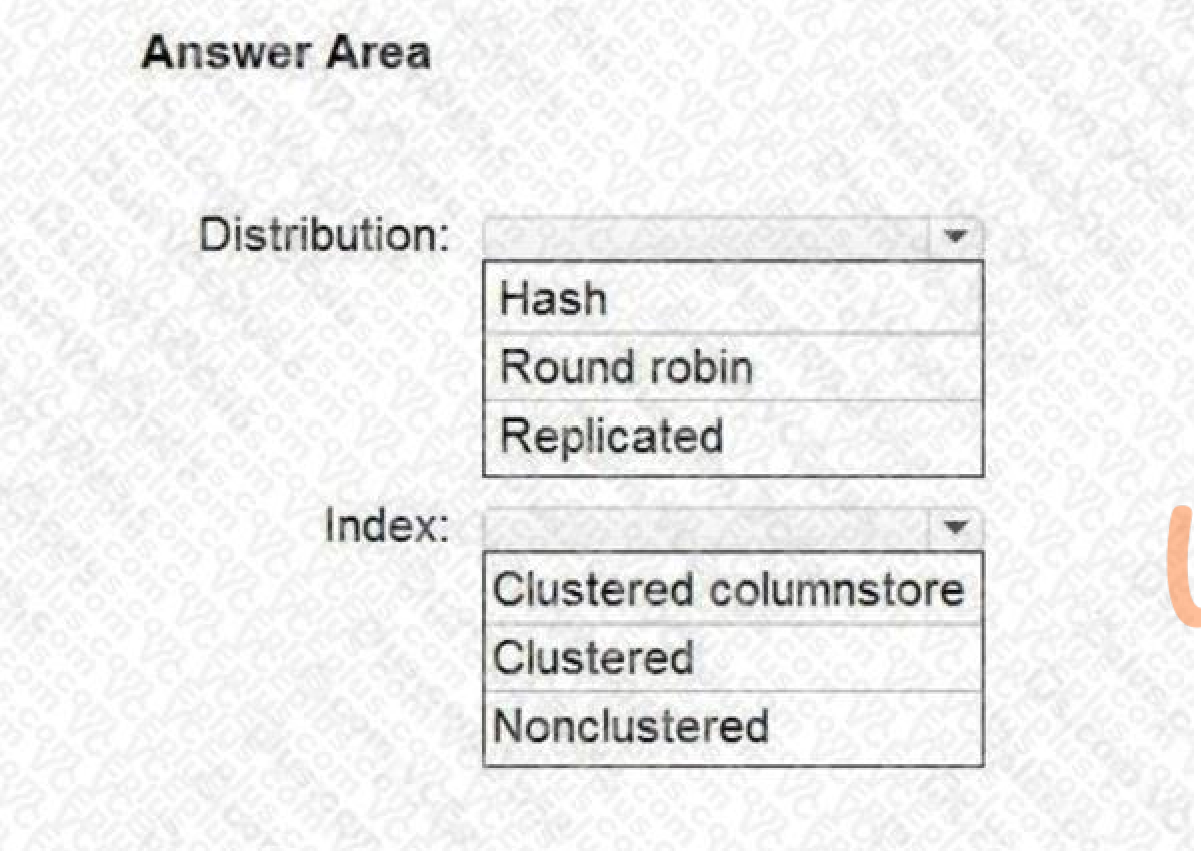
Box 1: HASH

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 16
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:

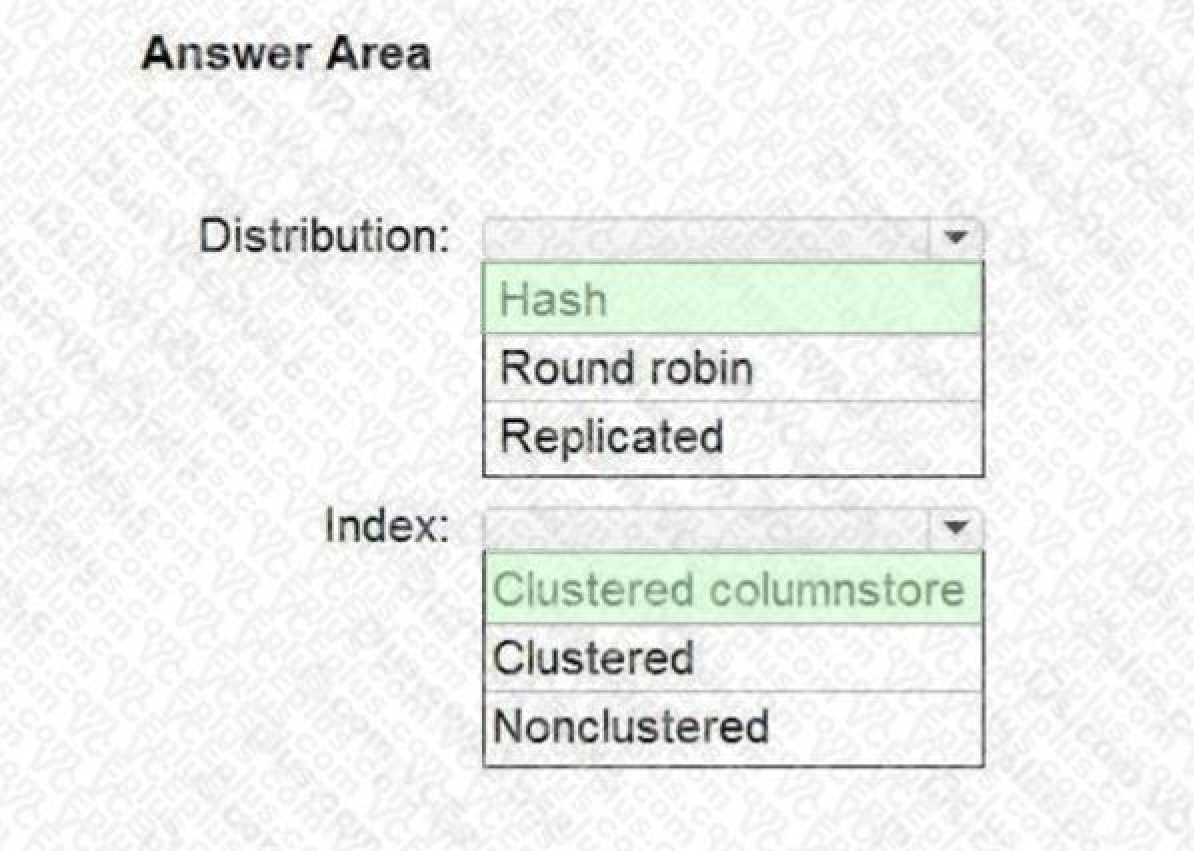


Answer Area

Distribution:
- Hash
- Round robin
- Replicated

Index:
- Clustered columnstore
- Clustered
- Nonclustered

Answer Area:



Answer Area

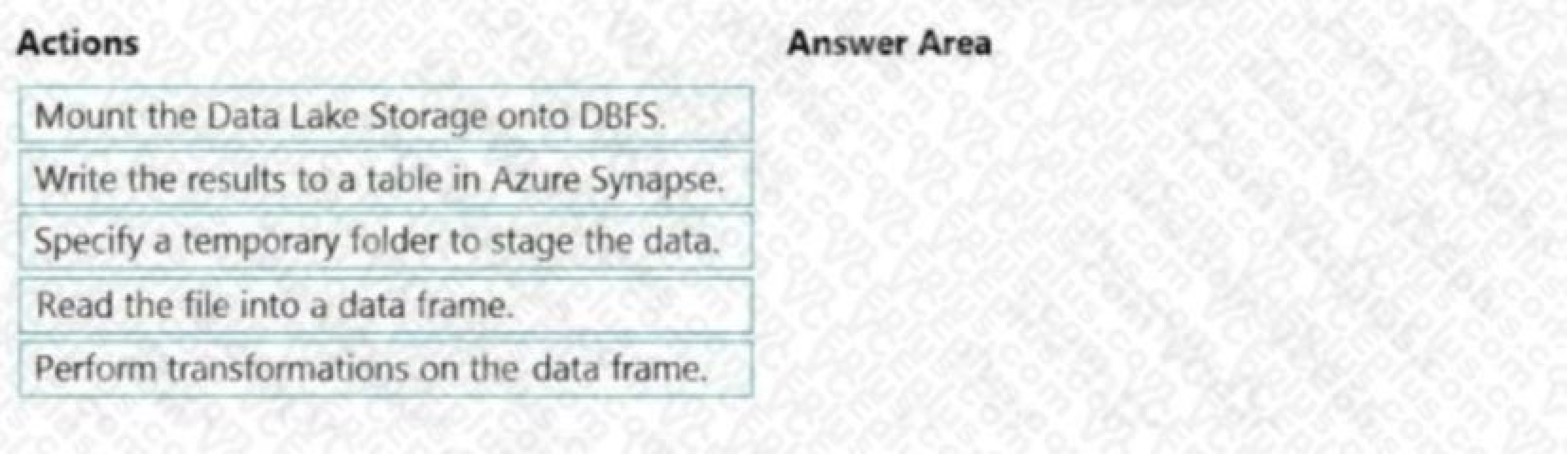
Distribution:
- Hash
- Round robin
- Replicated

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 17
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:

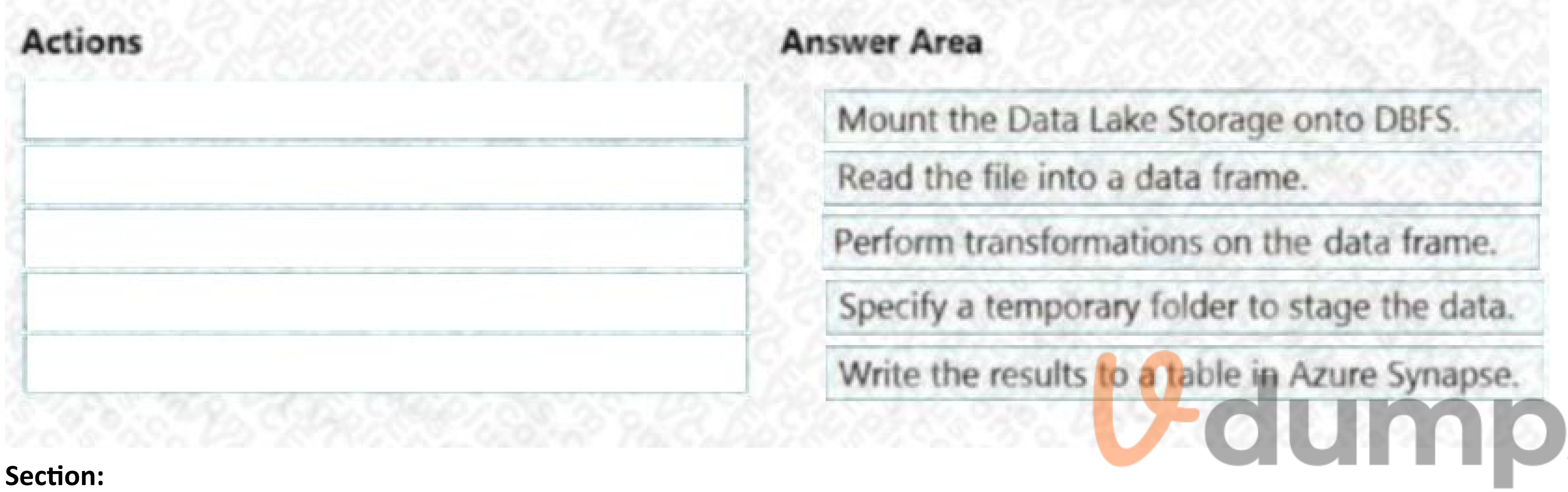


Actions

- 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

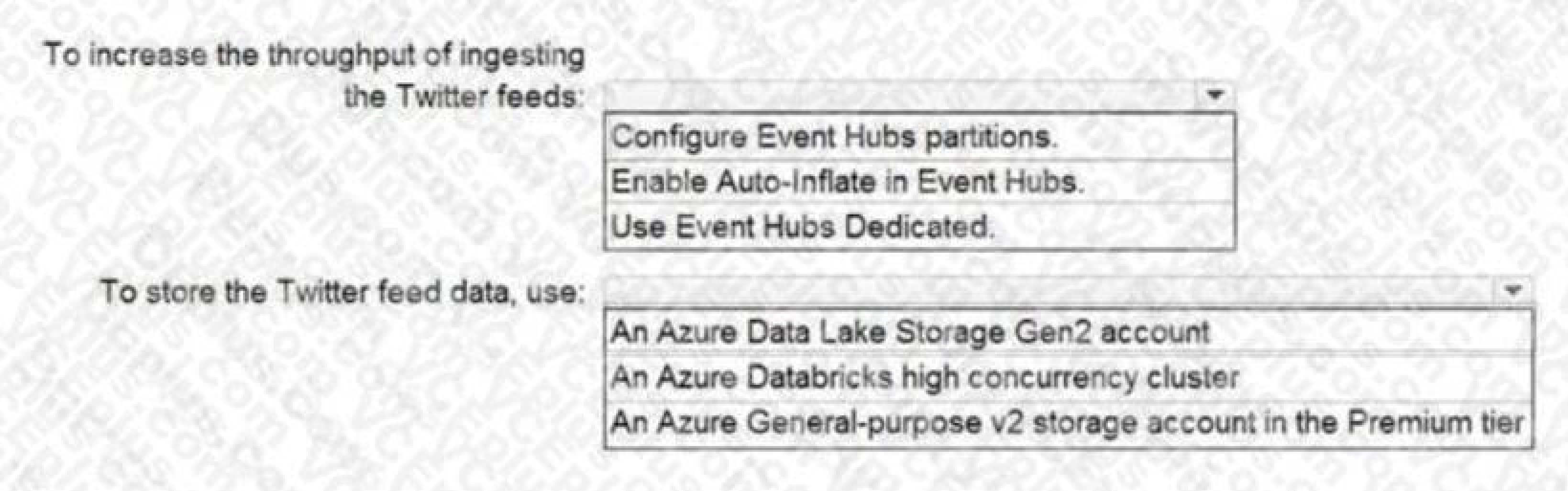
Answer Area

- 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 18
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:



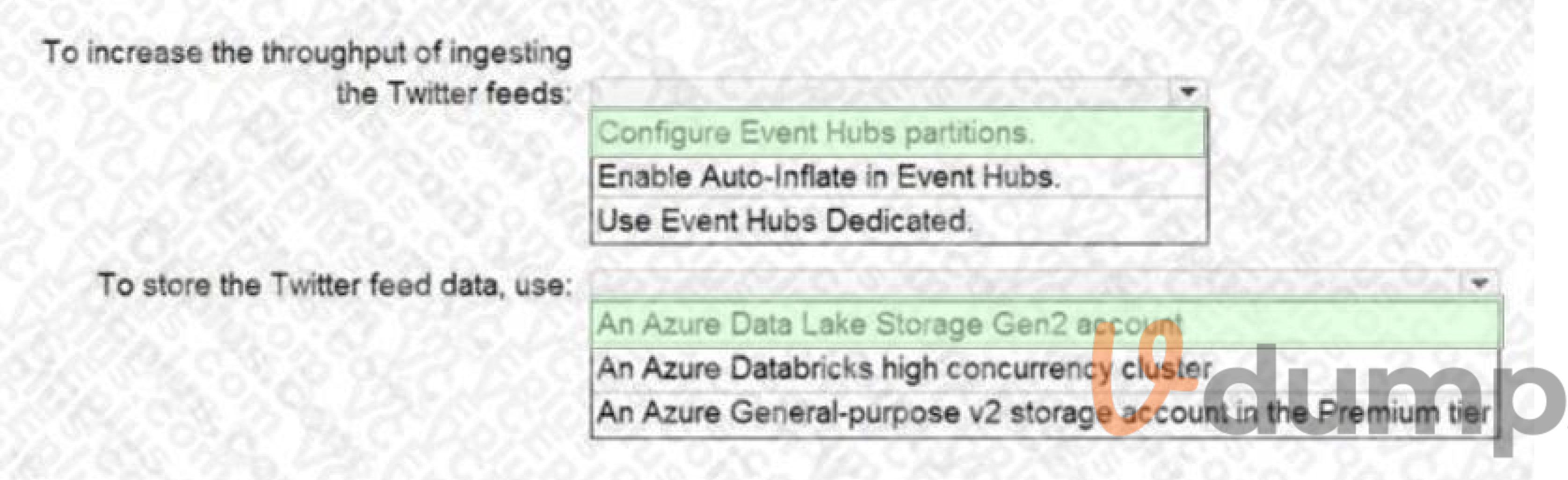
To increase the throughput of ingesting the Twitter feeds:

- 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.

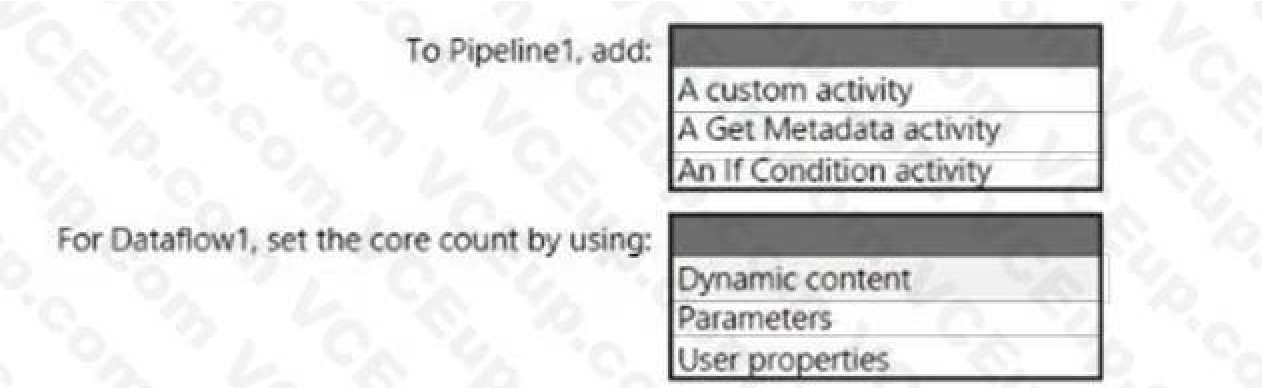
Reference: https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control

QUESTION 19

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.

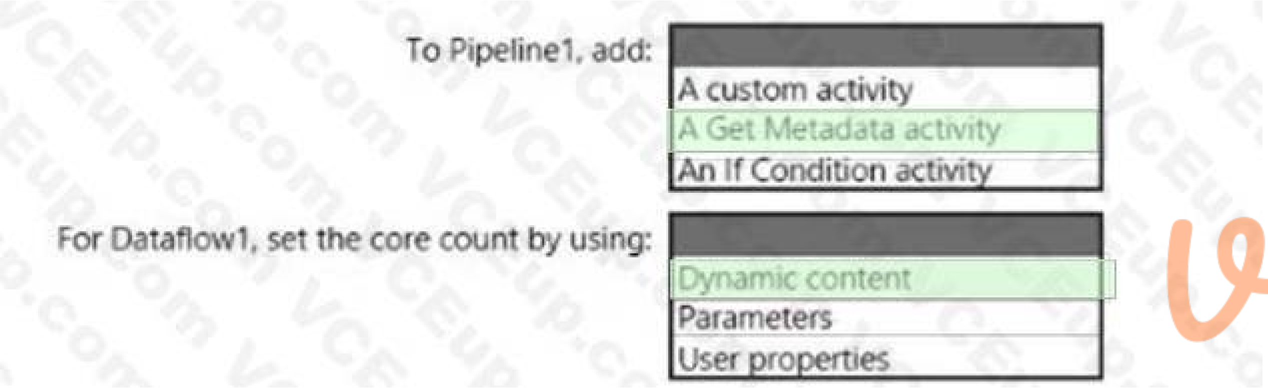
Hot Area:



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

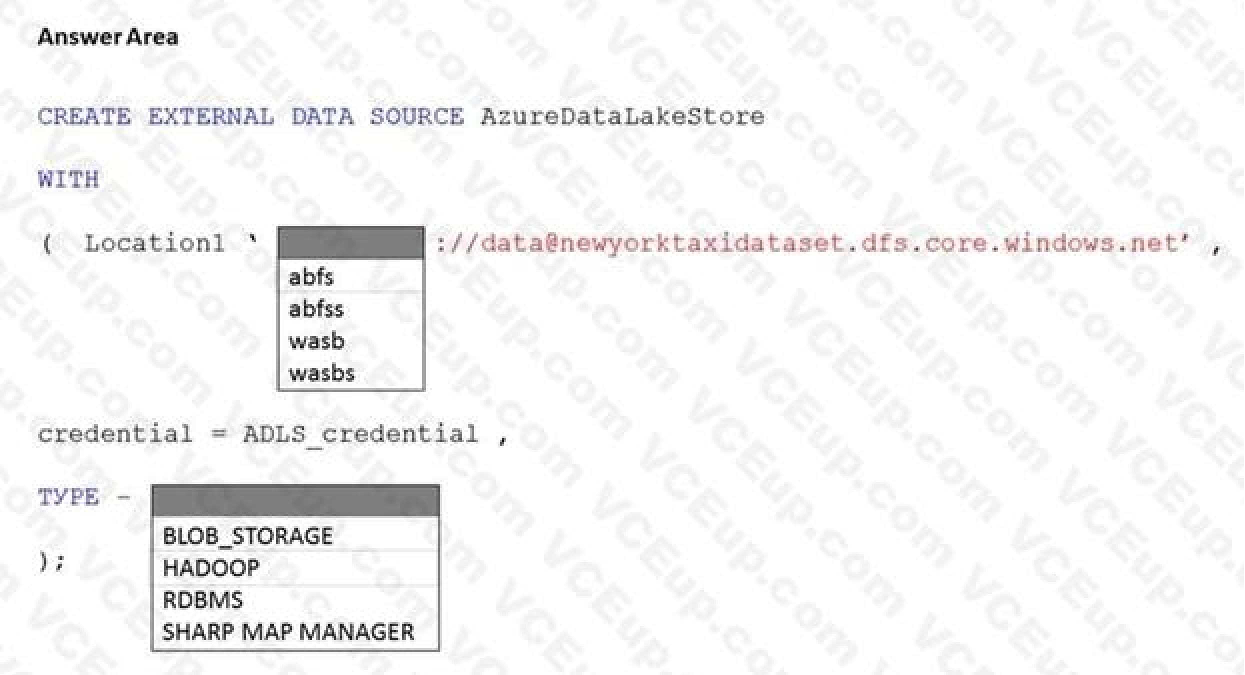
Reference: https://docs.microsoft.com/en-us/azure/data-factory/control-flow-execute-data-flowactivity

QUESTION 20

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:



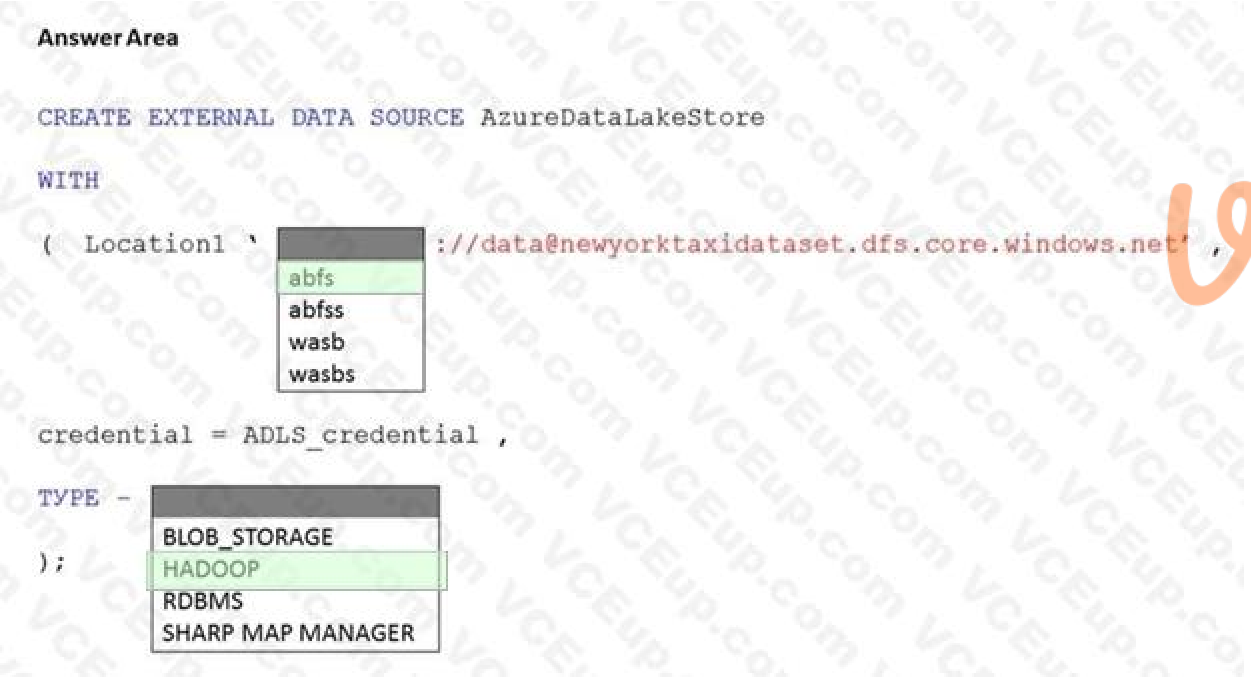
Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore

WITH

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

credential = ADLS_credential ,

TYPE =

);
```

- abfs
- abfss
- wasb
- wasbs

- BLOB_STORAGE
- HADOOP
- RDBMS
- SHARP MAP MANAGER

Answer Area:



Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore

WITH

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

credential = ADLS_credential ,

TYPE =

);
```

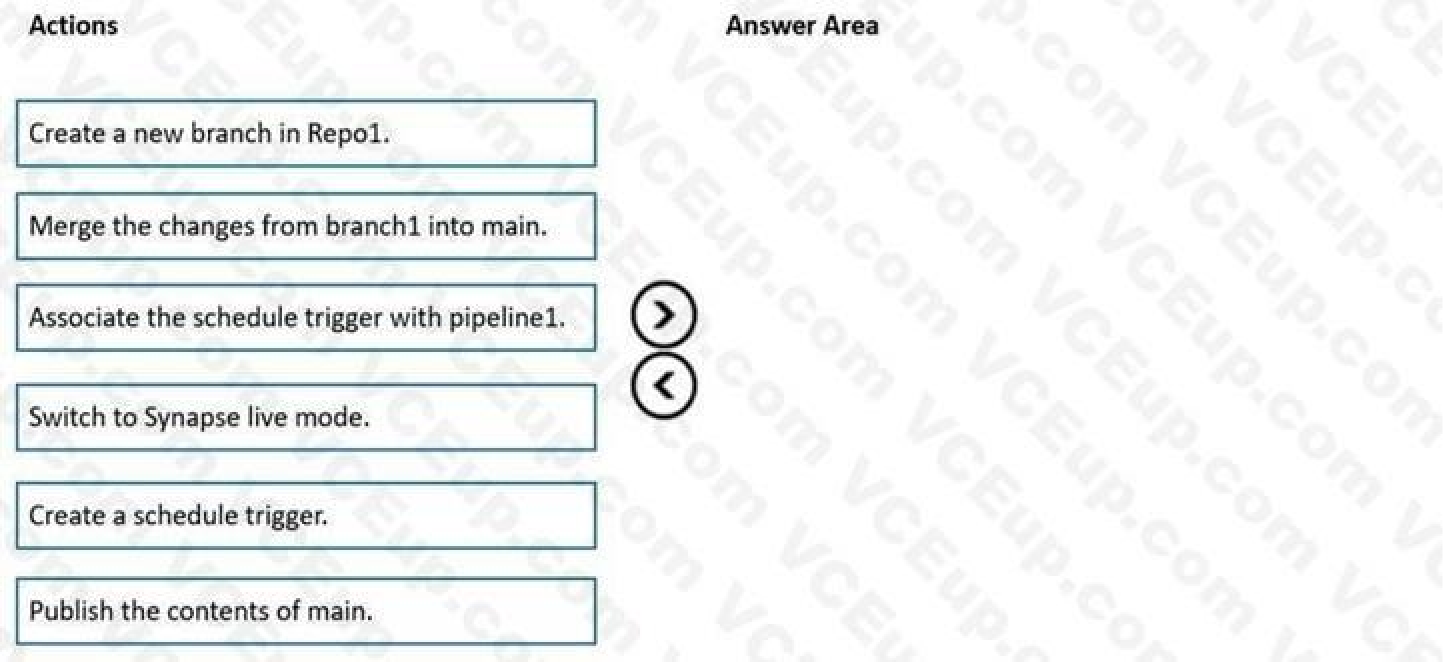
- [✓ abfs]
- abfss
- wasb
- wasbs

- 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 21
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:

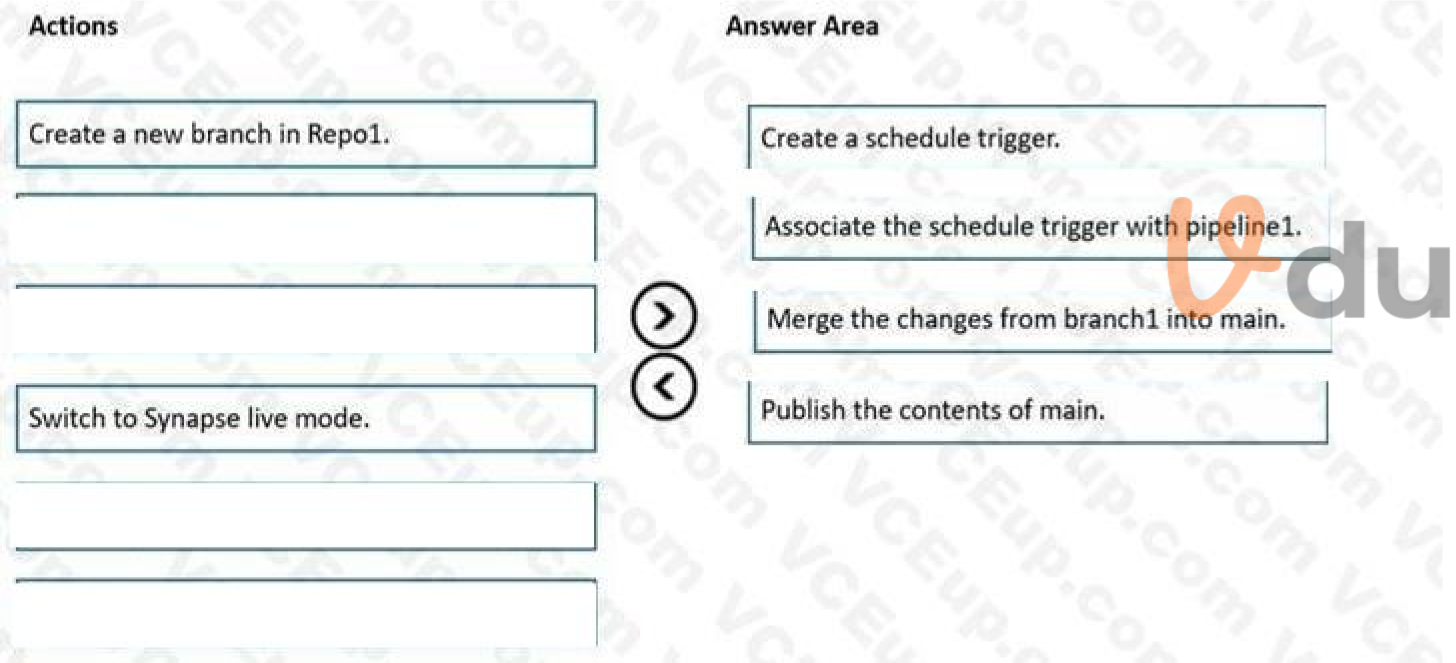


Actions

- 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.

Answer Area

Correct Answer:



Actions

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

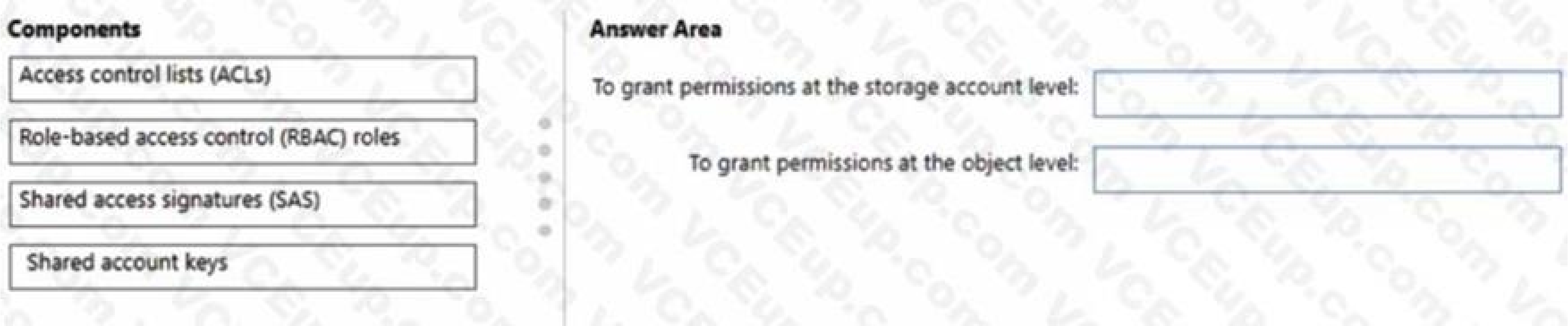
Answer Area

- 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 22
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

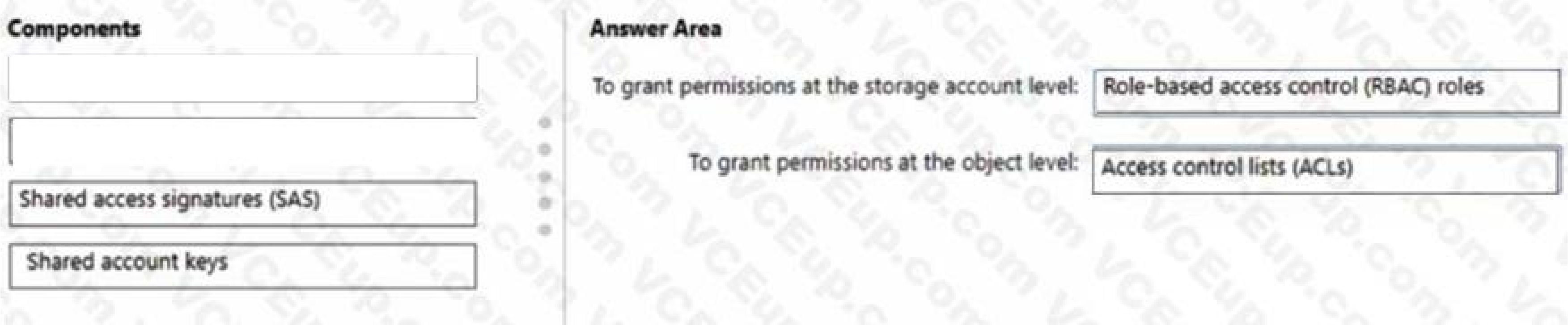
- Access control lists (ACLs)
- Role-based access control (RBAC) roles
- Shared access signatures (SAS)
- Shared account keys

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

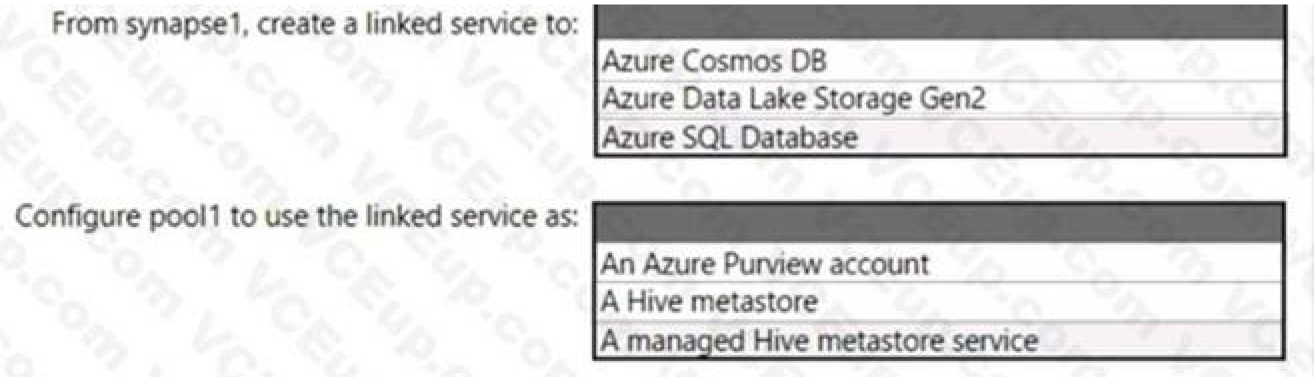
To grant permissions at the storage account level: Role-based access control (RBAC) roles

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 23
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.

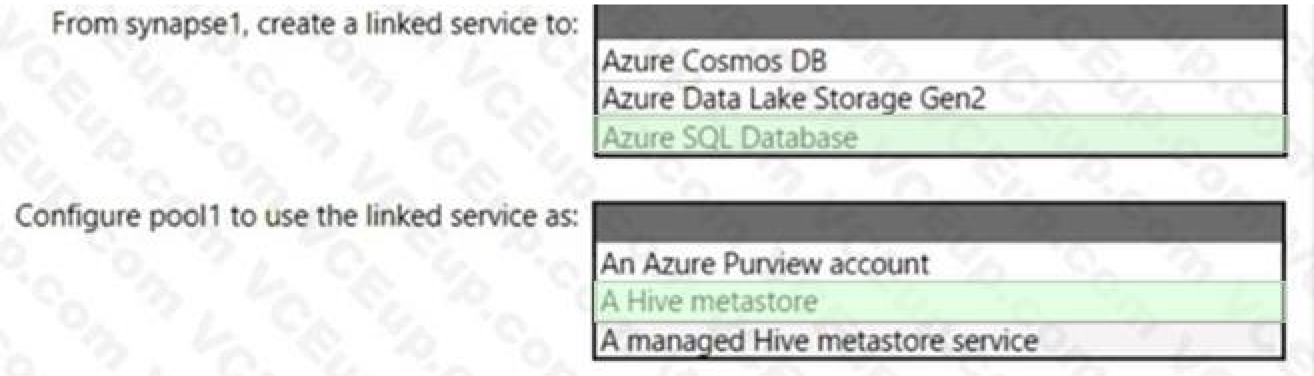
Hot Area:



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]

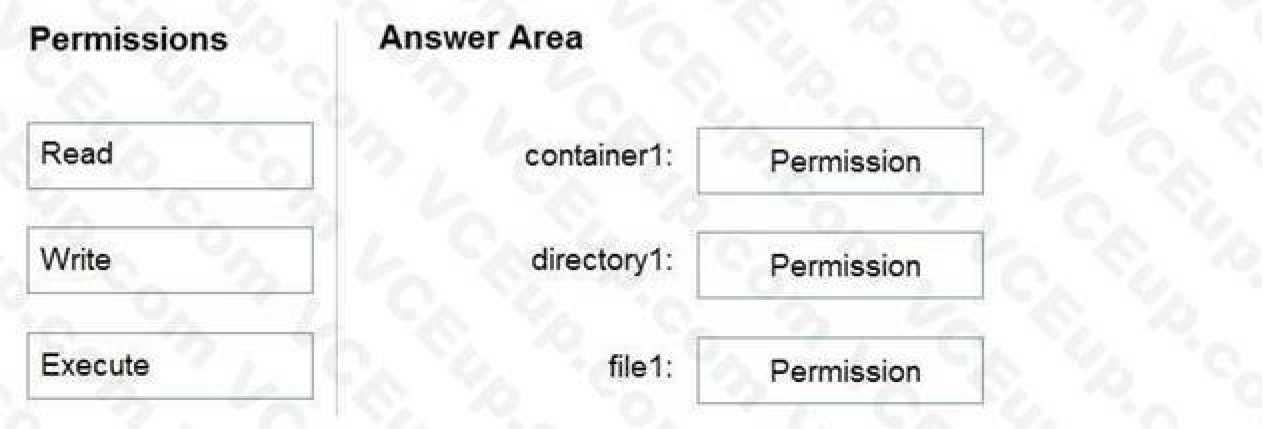
Configure pool1 to use the linked service as:

- 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 24
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

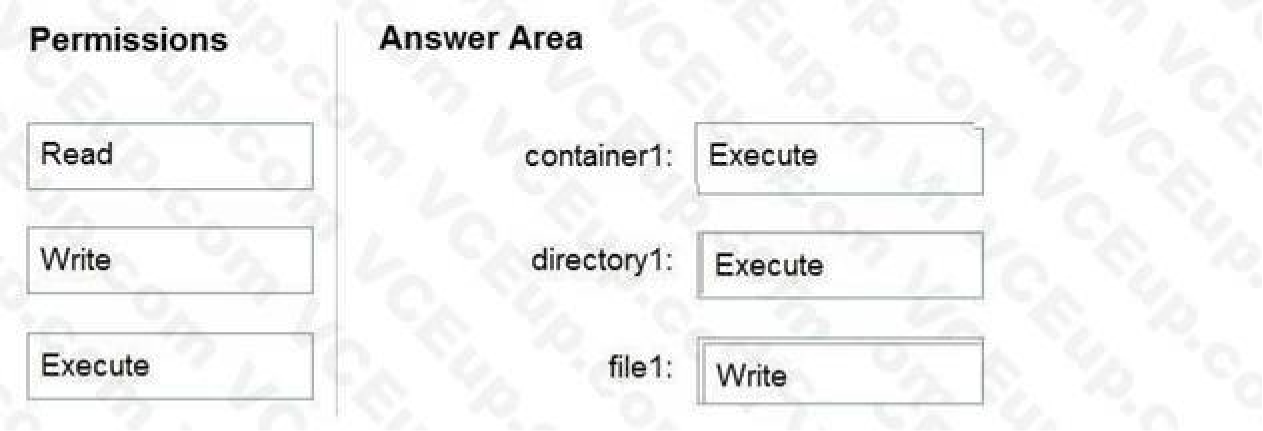
Answer Area

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 25
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

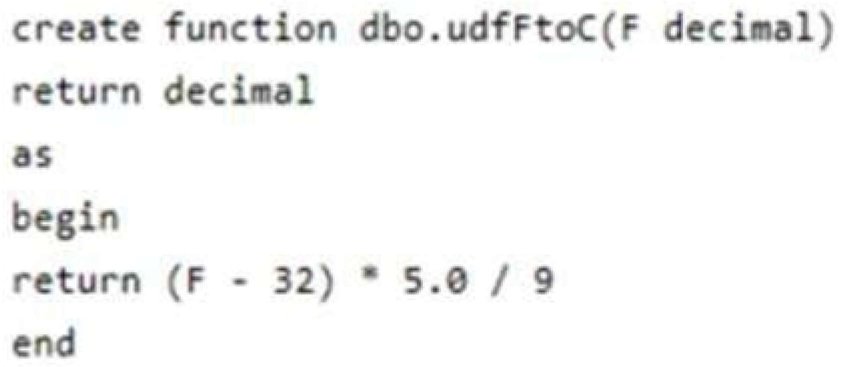
Correct Answer: B
Section:
Explanation:
Need a minimum 1 million rows per distribution. Each table is 60 distributions. 30 millions rows is added each month. Need 2 months to get a minimum of 1 million rows per distribution in a new partition. Note: When creating partitions on clustered columnstore tables, 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 26
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

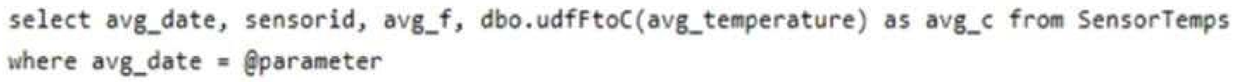
Correct Answer: A
Section:
Explanation:
Convert nested JSON to a flattened DataFrame
You can to flatten nested JSON, using only $"column.*" and explode methods. Note: Extract and flatten
Use $"column.*" and explode methods to flatten the struct and array types before displaying the flattened DataFrame. Scala
display(DF.select($"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 27
You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 receives new data once every 24 hours. You have the following function.



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

You have the following query.



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

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

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

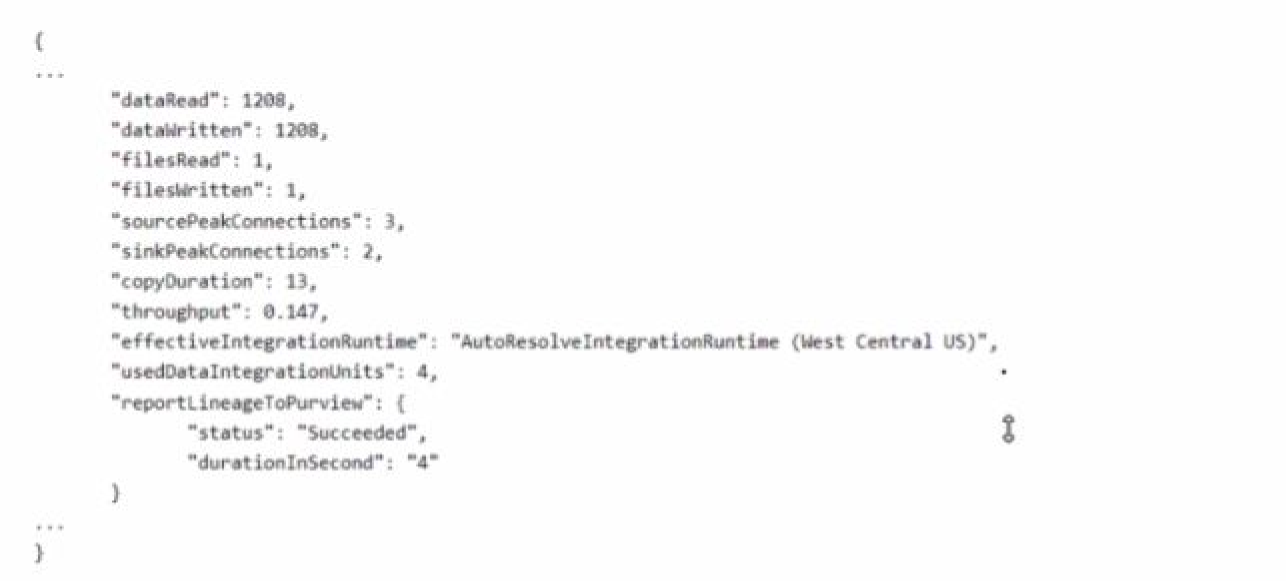
Correct Answer: B, D
Section:

QUESTION 28

HOTSPOT

You have an Azure data factory.

You execute a pipeline that contains an activity named Activity1. Activity1 produces the following output.

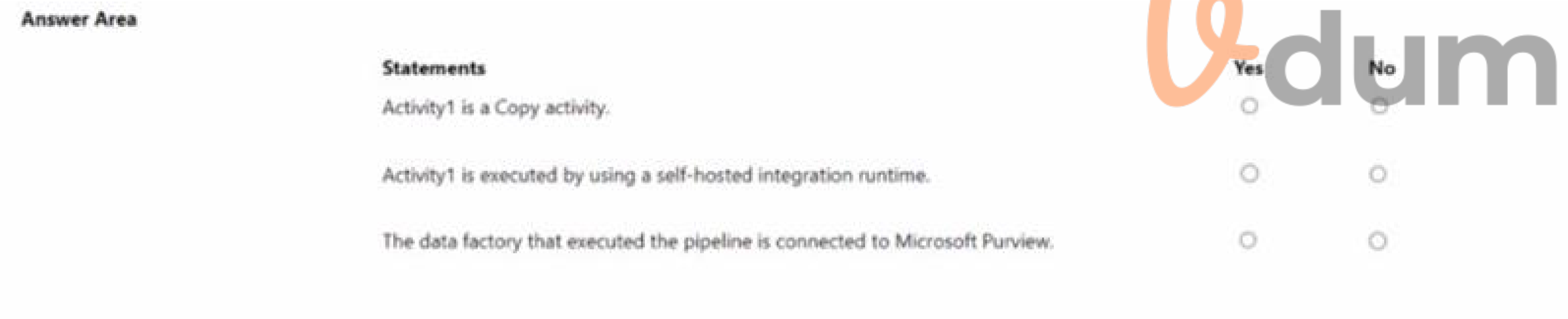


```
{
...
       "dataRead": 1208,
       "dataWritten": 1208,
       "filesRead": 1,
       "filesWritten": 1,
       "sourcePeakConnections": 3,
       "sinkPeakConnections": 2,
       "copyDuration": 13,
       "throughput": 0.147,
       "effectiveIntegrationRuntime": "AutoResolveIntegrationRuntime (West Central US)",
       "usedDataIntegrationUnits": 4,
       "reportLineageToPurview": {
              "status": "Succeeded",
              "durationInSecond": "4"
       }
...
}
```

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

NOTE: Each correct selection is worth one point.

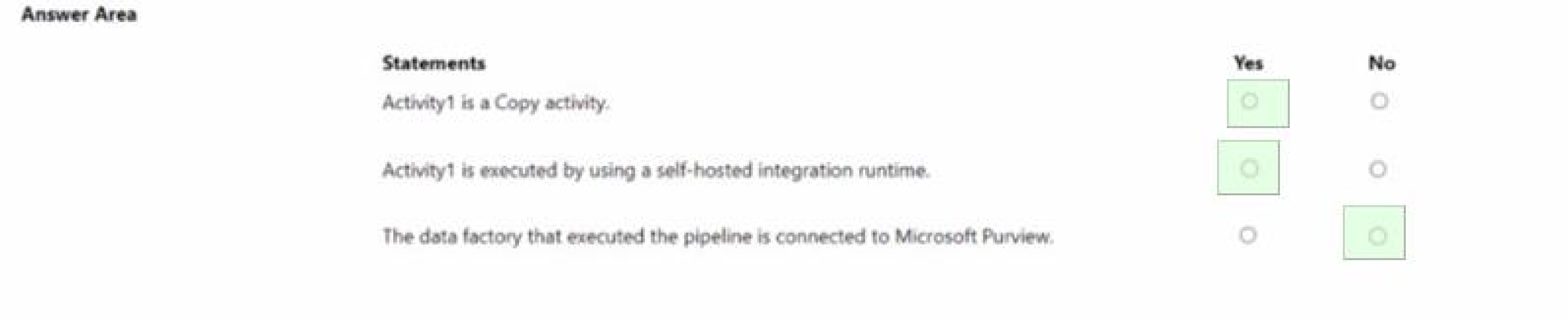
Hot Area:



Answer Area

Statements	Yes	No
Activity1 is a Copy activity.	○	○
Activity1 is executed by using a self-hosted integration runtime.	○	○
The data factory that executed the pipeline is connected to Microsoft Purview.	○	○

Answer Area:



Answer Area

Statements	Yes	No
Activity1 is a Copy activity.	[✓ ○]	○
Activity1 is executed by using a self-hosted integration runtime.	[✓ ○]	○
The data factory that executed the pipeline is connected to Microsoft Purview.	○	[✓ ○]

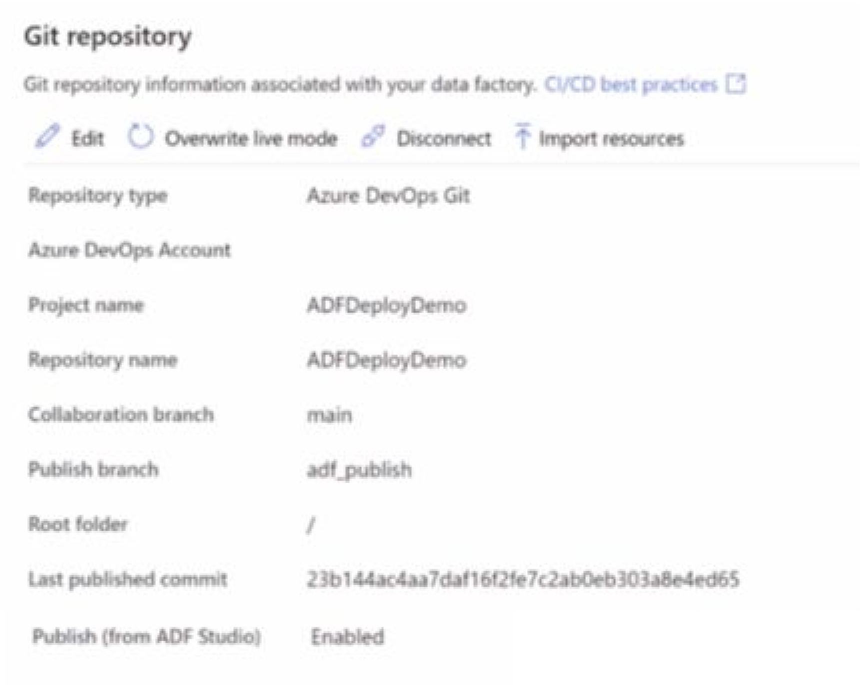
Section:

Explanation:

QUESTION 29

HOTSPOT

You have an Azure data factory that has the Git repository settings shown in the following exhibit.



Git repository

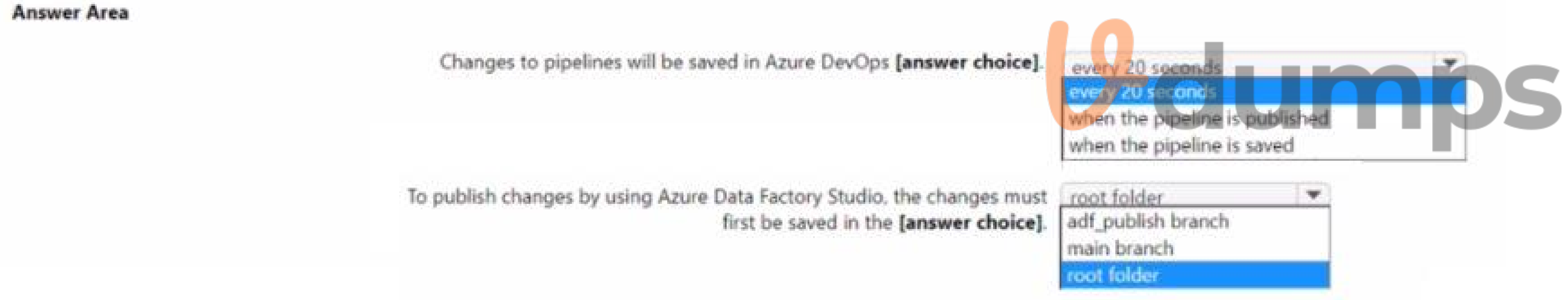
Git repository information associated with your data factory. CI/CD best practices

Repository type	Azure DevOps Git
Azure DevOps Account	
Project name	ADFDeployDemo
Repository name	ADFDeployDemo
Collaboration branch	main
Publish branch	adf_publish
Root folder	/
Last published commit	23b144ac4aa7daf16f2fe7c2ab0eb303a8e4ed65
Publish (from ADF Studio)	Enabled

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

NOTE: Each correct answer is worth one point.

Hot Area:



Answer Area

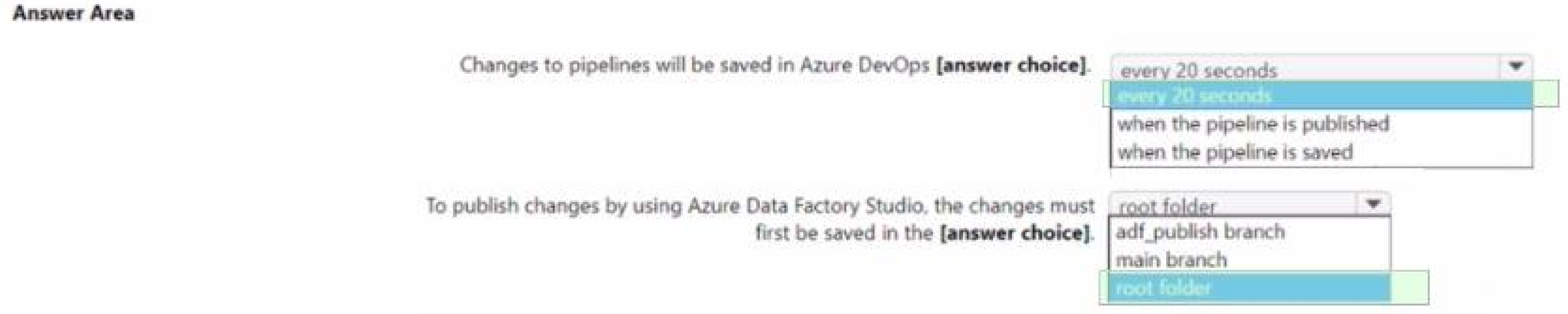
Changes to pipelines will be saved in Azure DevOps **[answer choice]**.

every 20 seconds
every 20 seconds
when the pipeline is published
when the pipeline is saved

To publish changes by using Azure Data Factory Studio, the changes must first be saved in the **[answer choice]**.

root folder
adf_publish branch
main branch
root folder

Answer Area:



Answer Area

Changes to pipelines will be saved in Azure DevOps **[answer choice]**.

every 20 seconds
every 20 seconds
when the pipeline is published
when the pipeline is saved

To publish changes by using Azure Data Factory Studio, the changes must first be saved in the **[answer choice]**.

root folder
adf_publish branch
main branch
root folder

Section:
Explanation:

QUESTION 30
You have an Azure Synapse Analytics dedicated SQL pool.
You plan to create a fact table named Table1 that will contain a clustered columnstore index.
You need to optimize data compression and query performance for Table1.
What is the minimum number of rows that Table1 should contain before you create partitions?

A. 100.000

B. 600,000

C. 1 million

D. 60 million

Correct Answer: A
Section:

QUESTION 31

You have an Azure subscription that contains an Azure Data Factory data pipeline named Pipeline1, a Log Analytics workspace named LA1, and a storage account named account1.
You need to retain pipeline-run data for 90 days. The solution must meet the following requirements:
* The pipeline-run data must be removed automatically after 90 days.
* Ongoing costs must be minimized.
Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. Configure Pipeline1 to send logs to LA1.

B. From the Diagnostic settings (classic) settings of account1. set the retention period to 90 days.

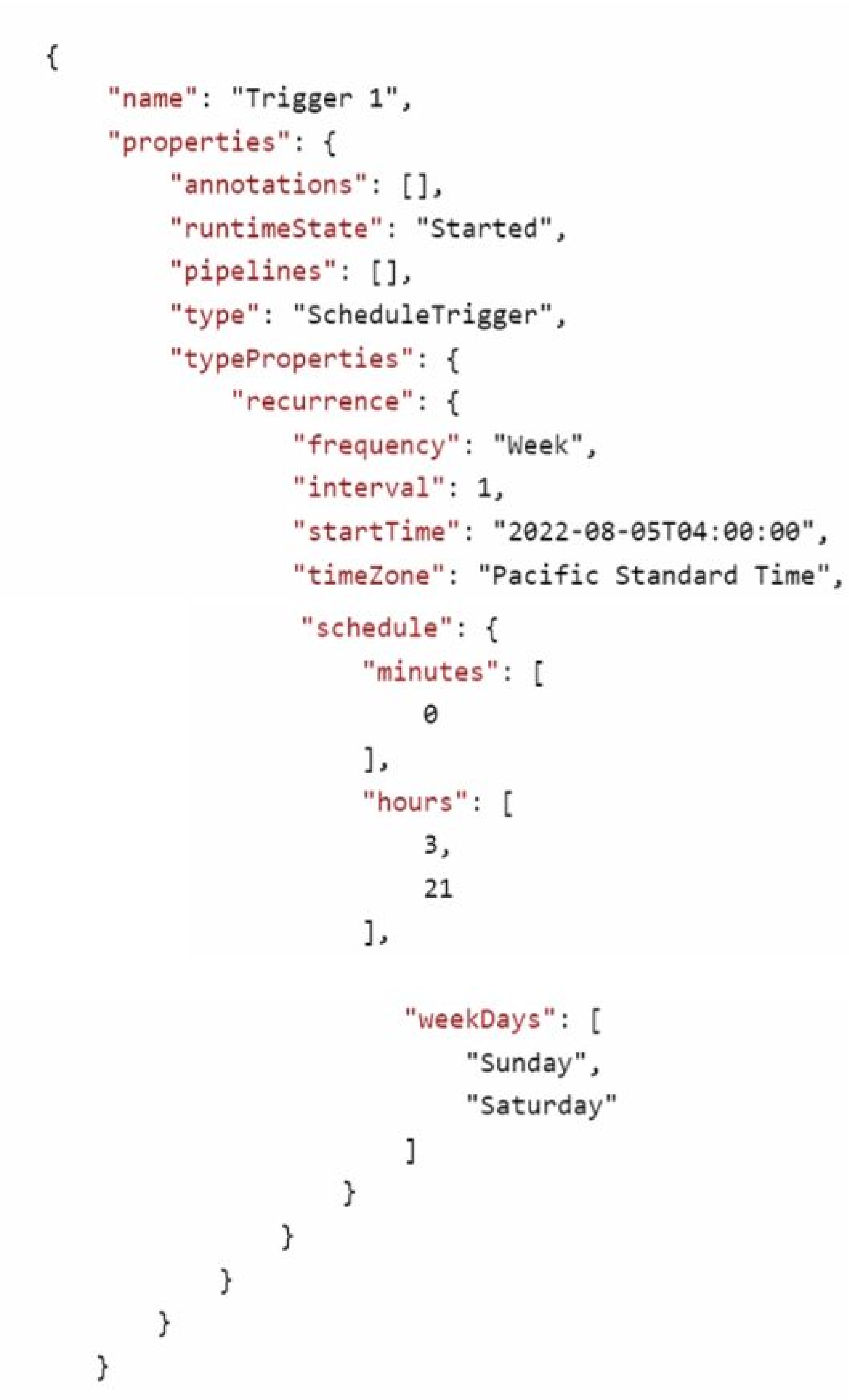
C. Configure Pipeline1 to send logs to account1.

D. From the Data Retention settings of LA1, set the data retention period to 90 days.

Correct Answer: A, B
Section:

QUESTION 32

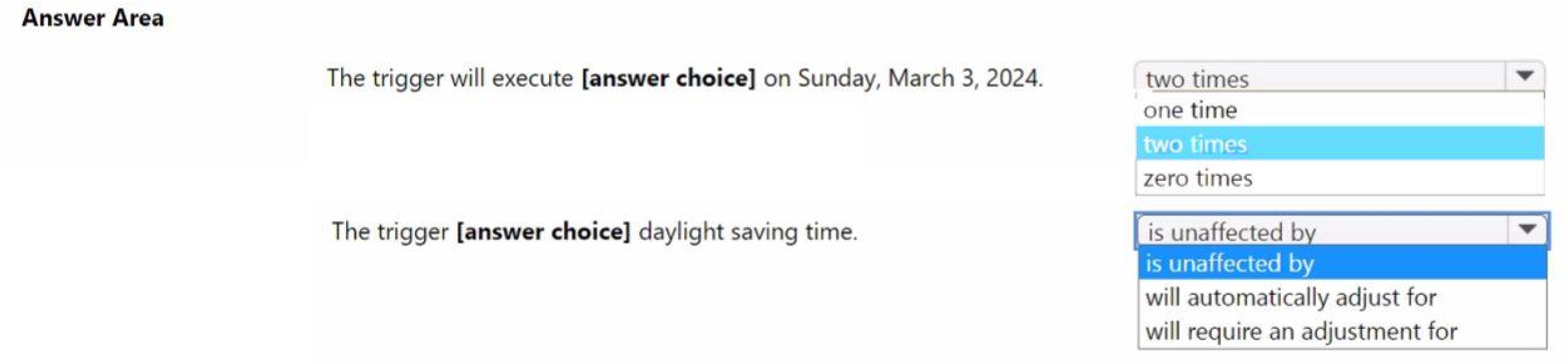
HOTSPOT
In Azure Data Factory, you have a schedule trigger that is scheduled in Pacific Time.
Pacific Time observes daylight saving time.
The trigger has the following JSON file.



```
{
    "name": "Trigger 1",
    "properties": {
        "annotations": [],
        "runtimeState": "Started",
        "pipelines": [],
        "type": "ScheduleTrigger",
        "typeProperties": {
            "recurrence": {
                "frequency": "Week",
                "interval": 1,
                "startTime": "2022-08-05T04:00:00",
                "timeZone": "Pacific Standard Time",
                 "schedule": {
                     "minutes": [
                         0
                     ],
                     "hours": [
                         3,
                         21
                     ],

                       "weekDays": [
                           "Sunday",
                           "Saturday"
                       ]
                   }
               }
           }
       }
   }
```

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

Hot Area:

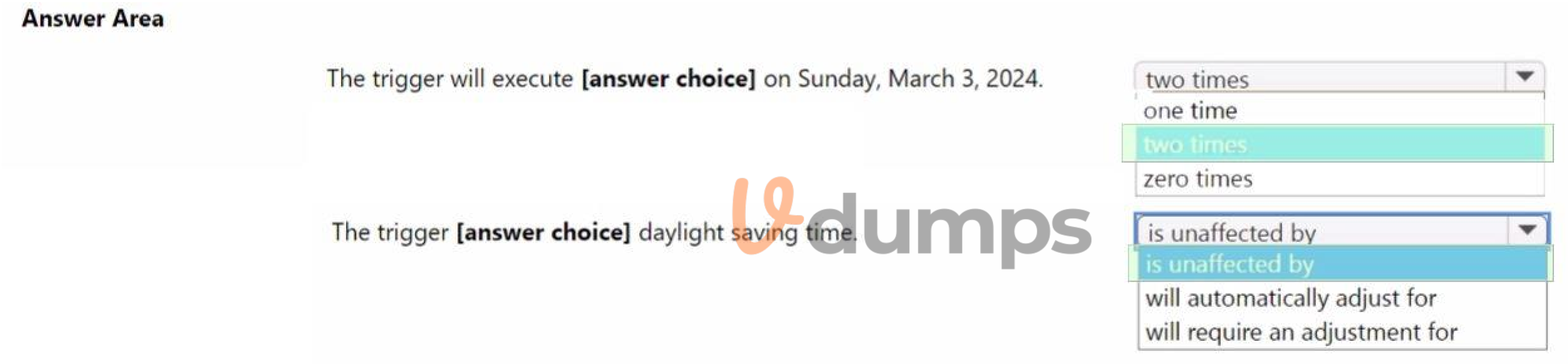


Answer Area

The trigger will execute [answer choice] on Sunday, March 3, 2024.

The trigger [answer choice] daylight saving time.

Answer Area:



Answer Area

The trigger will execute [answer choice] on Sunday, March 3, 2024.

The trigger [answer choice] daylight saving time.

Section:
Explanation: