Microsof.VCEup.com.DP-203.2022-July-10.107q

Number: DP-203
Passing Score: 800
Time Limit: 120 min

Exam Code: DP-203
Exam Name: Data Engineering on Microsoft Azure
Certification Provider: Microsoft
Corresponding Certification: Microsoft Certified: Azure Data Engineer Associate
Website: https://VCEup.com/
Free Exam: https://vceup.com/exam-dp-203/

01 - Design and implement data storage

QUESTION 1
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.

A.

Correct Answer:
Section: (none)
Explanation

Explanation/Reference:

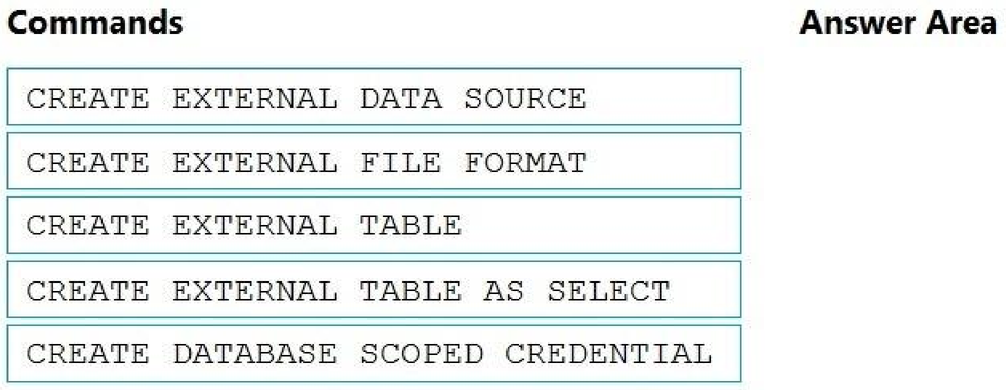
QUESTION 2
DRAG DROP

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:

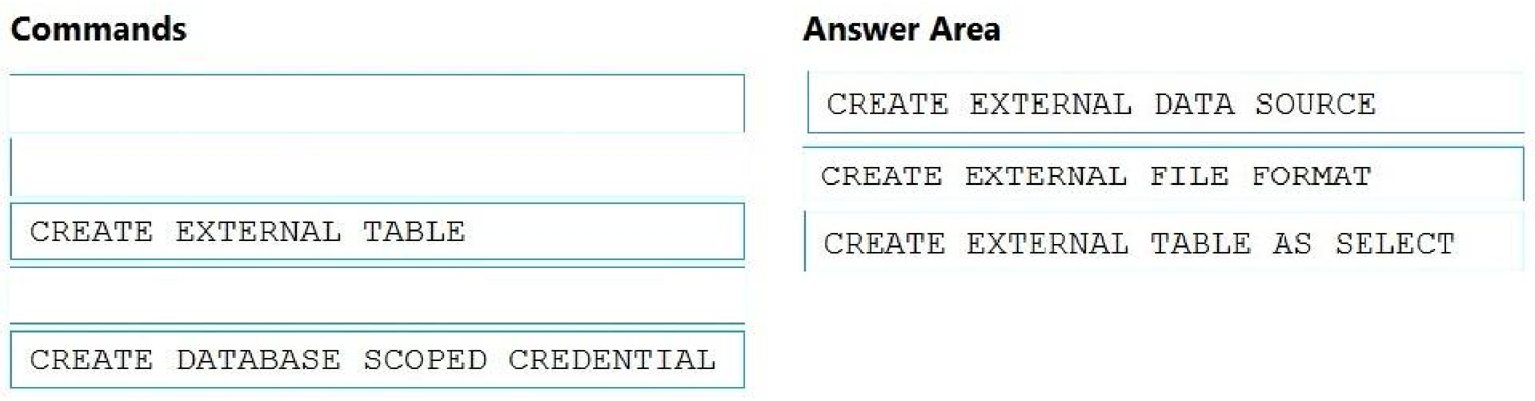


Commands

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

Answer Area

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: (none)
Explanation

Explanation/Reference:
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

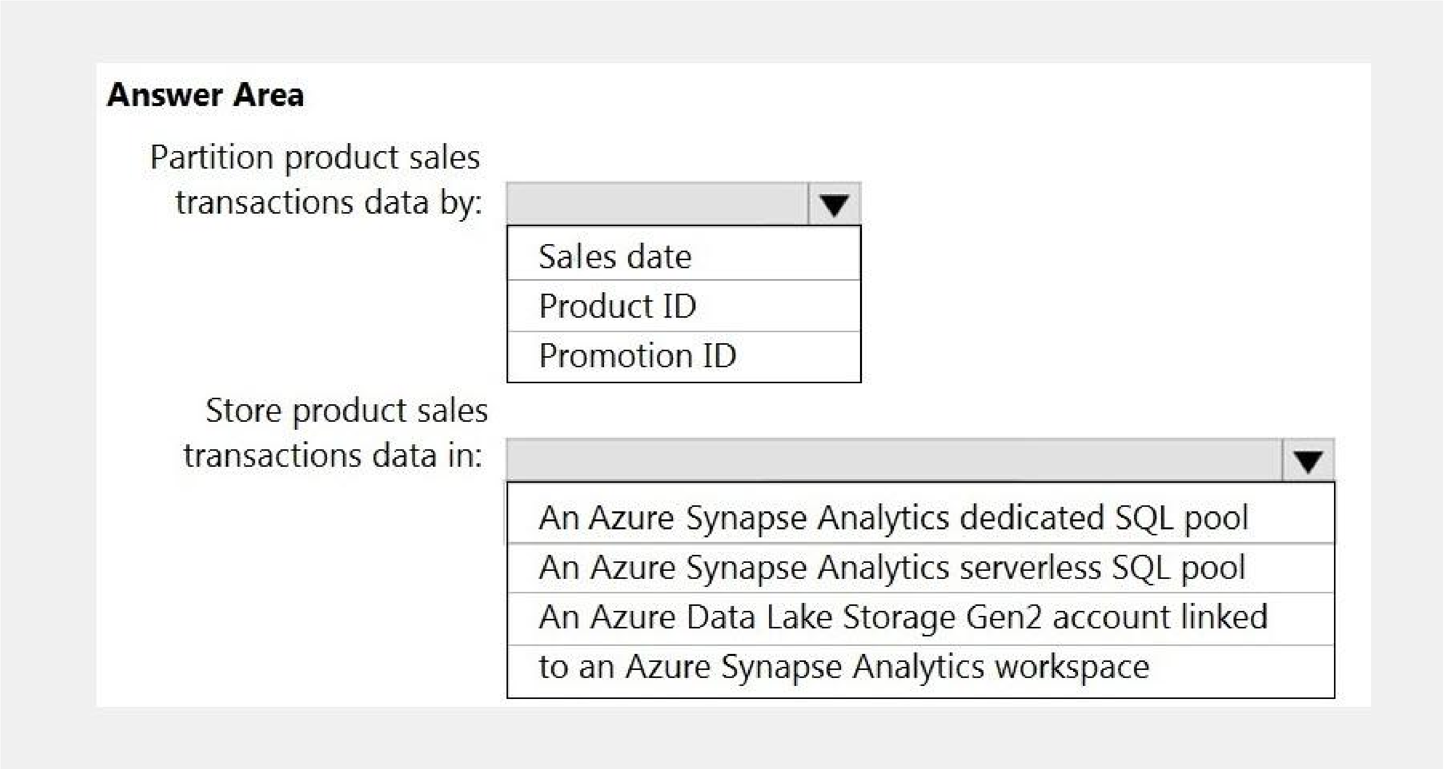
QUESTION 3
HOTSPOT

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

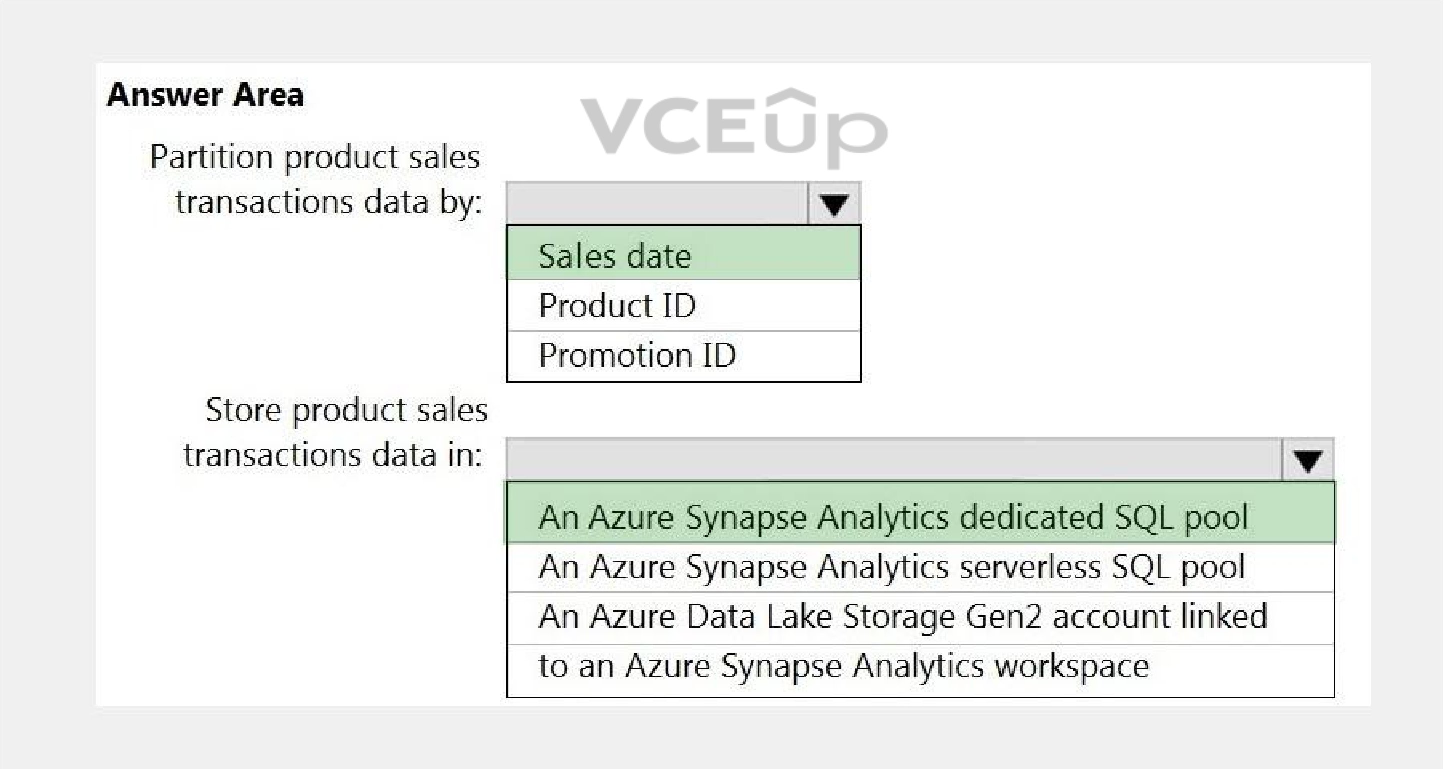
Partition product sales transactions data by:

- 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

Correct Answer:



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: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: Sales date
Scenario: Contoso requirements for data integration include:

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

Box 2: An Azure Synapse Analytics Dedicated SQL pool
Scenario: Contoso requirements for data integration include:

- Ensure that data storage costs and performance are predictable.

The size of a dedicated SQL pool (formerly SQL DW) is determined by Data Warehousing Units (DWU).
Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage. This format significantly reduces the data storage costs, and improves query performance.
Synapse analytics dedicated sql pool

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

QUESTION 4
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: (none)
Explanation

Explanation/Reference:
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

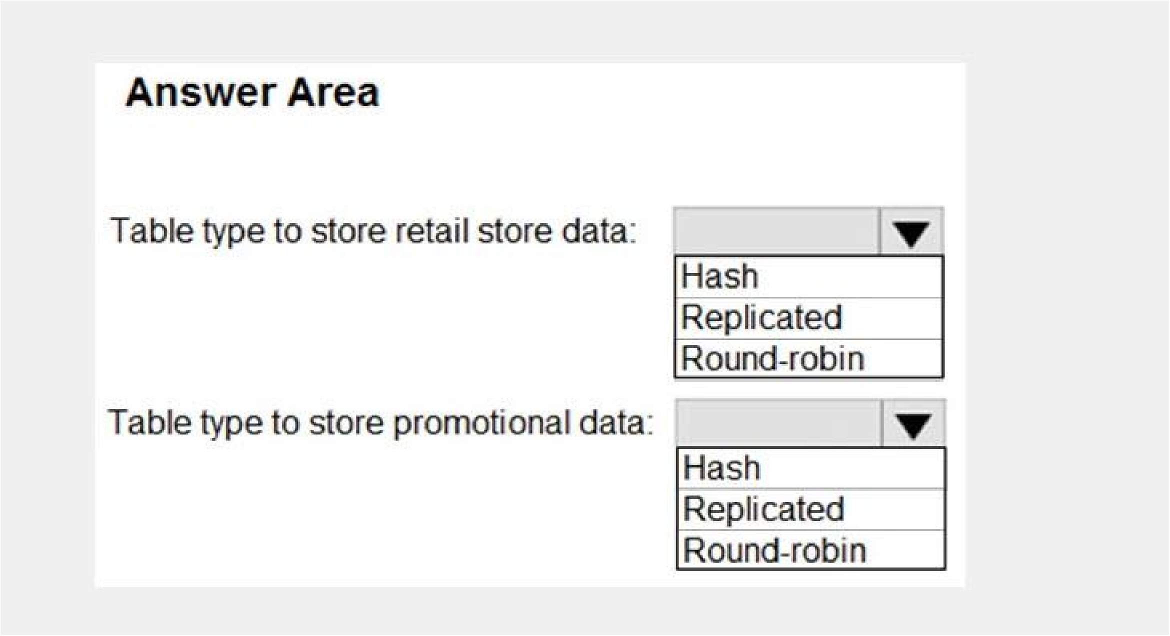
QUESTION 5
HOTSPOT

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:

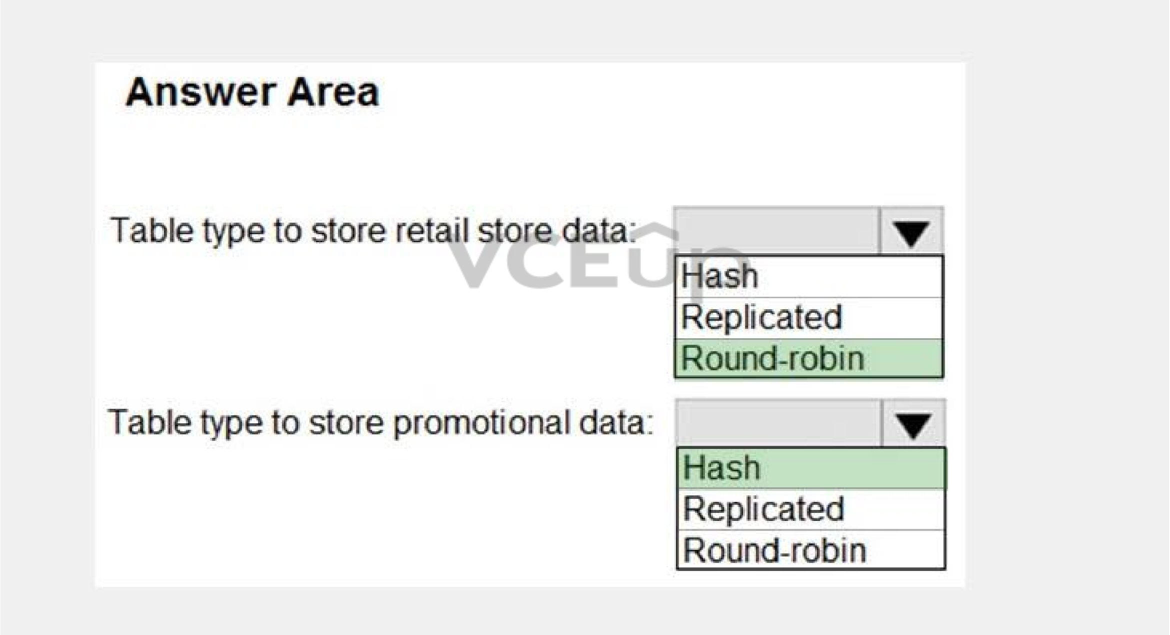


Answer Area

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

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

Correct Answer:



Answer Area

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

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

Section: (none)
Explanation

Explanation/Reference:
Explanation:
Box 1: Round-robin
Round-robin tables are useful for improving loading speed.

Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month.

Box 2: Hash
Hash-distributed tables improve query performance on large fact tables.

Scenario:
- 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.
- Ensure that queries joining and filtering sales transaction records based on product ID complete as quickly as possible.

Reference:

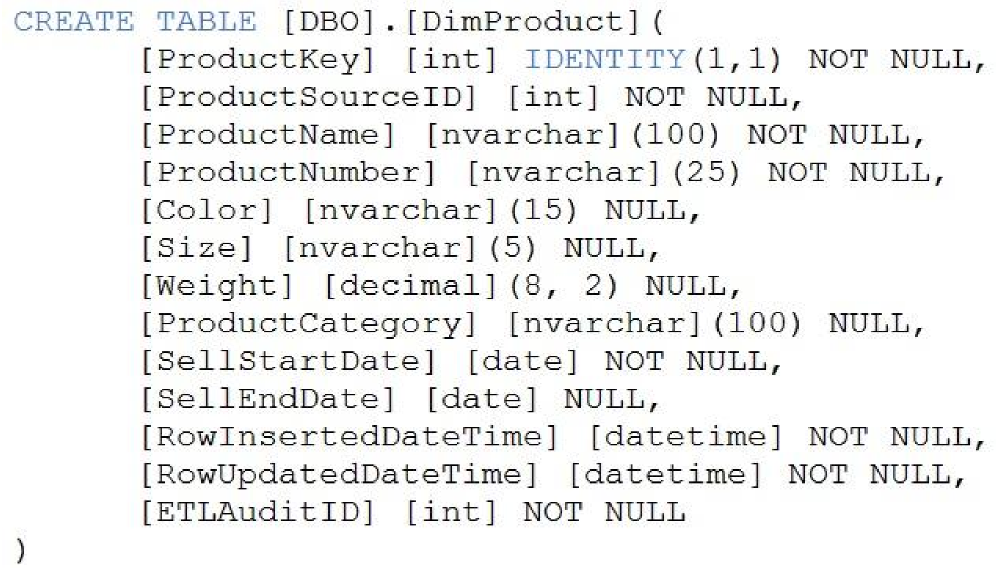
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

02 - Design and implement data storage

QUESTION 1
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.

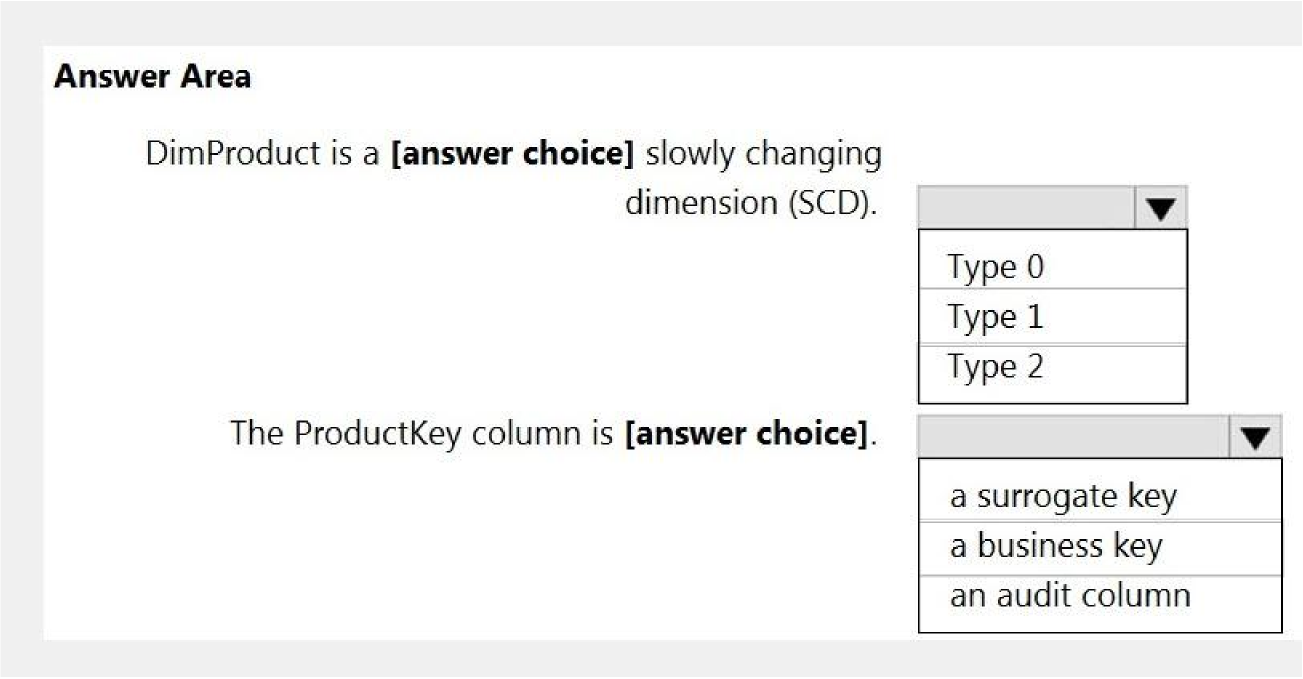


```
CREATE TABLE [DBO].[DimProduct](
      [ProductKey] [int] IDENTITY(1,1) NOT NULL,
      [ProductSourceID] [int] NOT NULL,
      [ProductName] [nvarchar](100) NOT NULL,
      [ProductNumber] [nvarchar](25) NOT NULL,
      [Color] [nvarchar](15) NULL,
      [Size] [nvarchar](5) NULL,
      [Weight] [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:

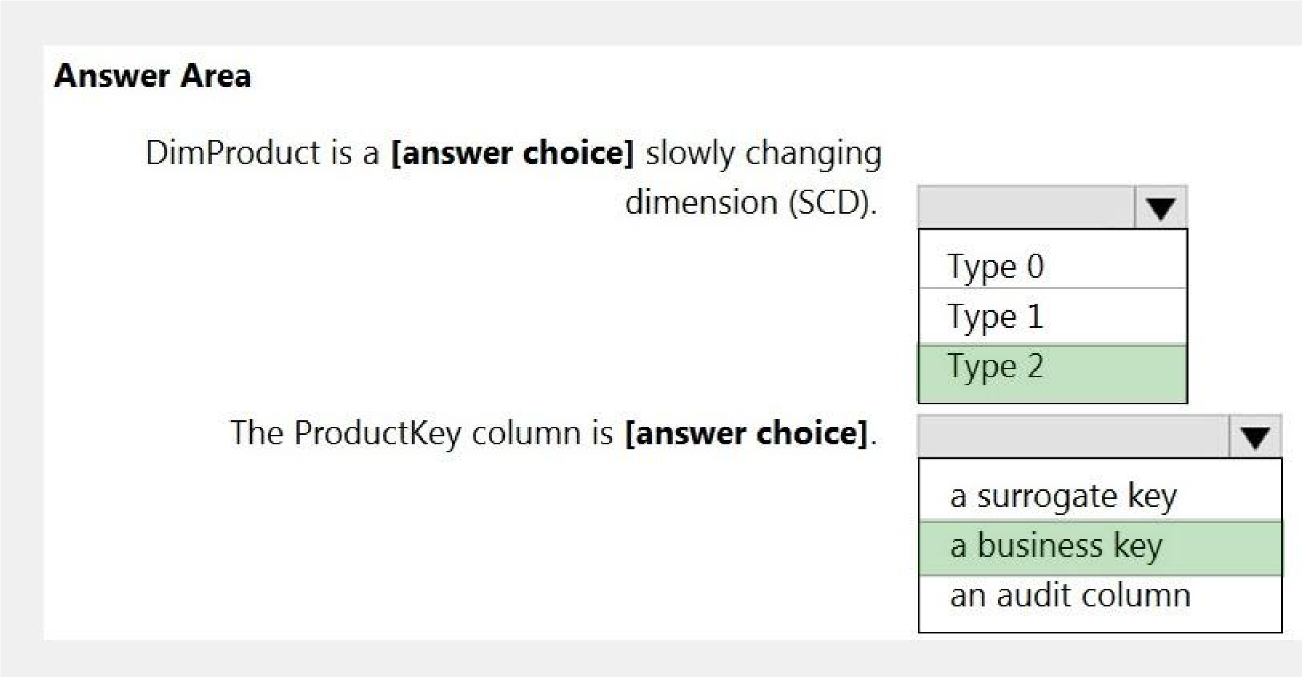


Answer Area

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

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

Correct Answer:



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: (none)
Explanation

Explanation/Reference:
Explanation:

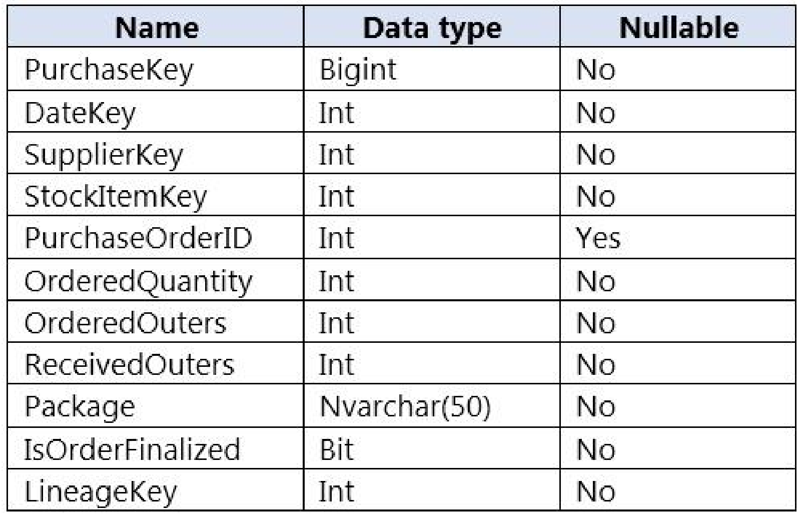
Box 1: Type 2
A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members.

Incorrect Answers:
A Type 1 SCD always reflects the latest values, and when changes in source data are detected, the dimension table data is overwritten.

Box 2: a business key
A business key or natural key is an index which identifies uniqueness of a row based on columns that exist naturally in a table according to business rules. For example business keys are customer code in a customer table, composite of sales order header number and sales order item line number within a sales order details table.

Reference:
https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types

QUESTION 2
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

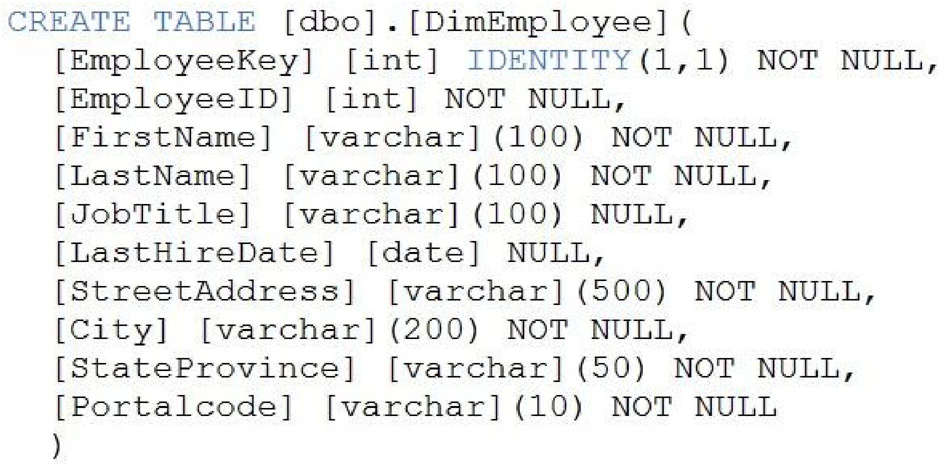
Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
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 3
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] [int] NULL`
B. `[ManagerEmployeeID] [smallint] NULL`
C. `[ManagerEmployeeKey] [int] NULL`
D. `[ManagerName] [varchar](200) NULL`

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Use the same definition as the EmployeeID column.

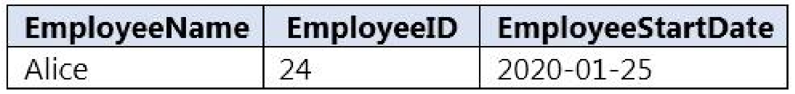
Reference:
https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular

QUESTION 4
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: (none)
Explanation

Explanation/Reference:
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 5
DRAG DROP

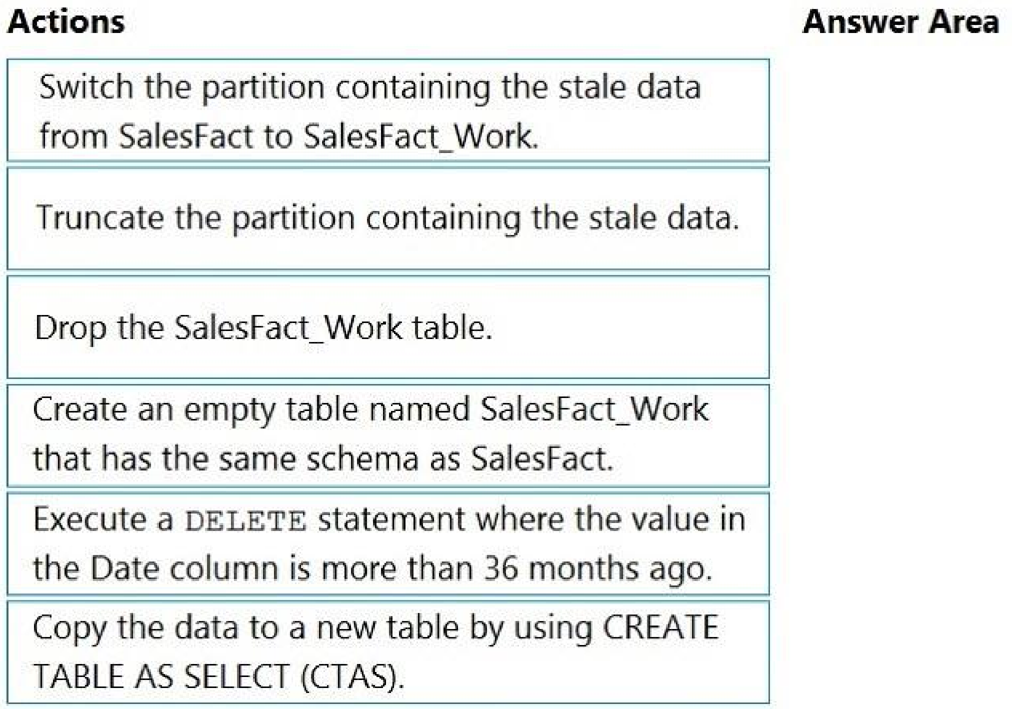
You have a table named SalesFact in an enterprise data warehouse in Azure Synapse Analytics. SalesFact contains sales data from the past 36 months and has the following characteristics:

- 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

Answer Area

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

Truncate the partition containing the stale data.

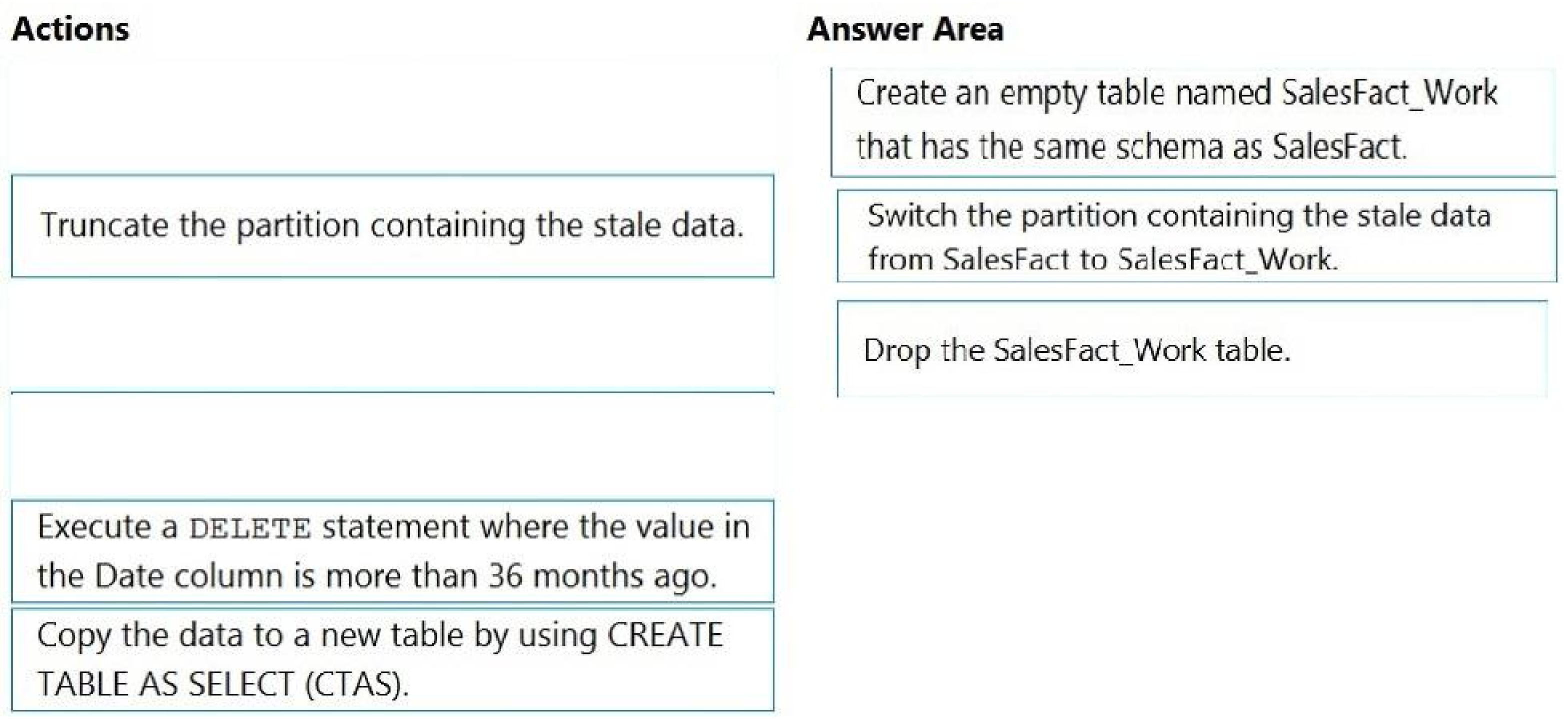
Drop the SalesFact_Work table.

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).

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: (none)
Explanation

Explanation/Reference:
Explanation:

Step 1: Create an empty table named SalesFact_work that has the same schema as SalesFact.

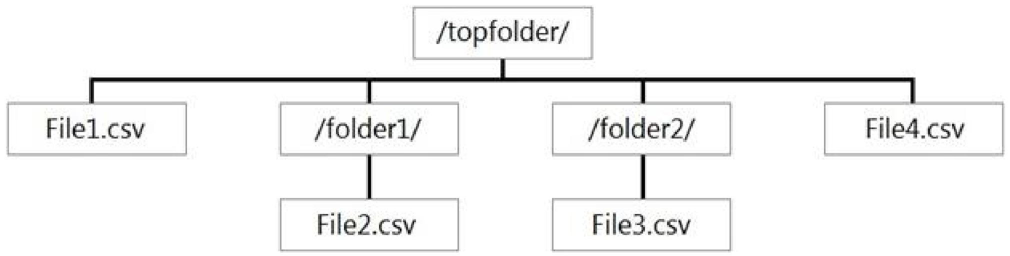
Step 2: Switch the partition containing the stale data from SalesFact to SalesFact_Work.
SQL Data Warehouse supports partition splitting, merging, and switching. To switch partitions between two tables, you must ensure that the partitions align on their respective boundaries and that the table definitions 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 6
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: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
To run a T-SQL query over a set of files within a folder or set of folders while treating them as a single entity or rowset, provide a path to a folder or a pattern (using wildcards) over a set of files or folders.

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

QUESTION 7
HOTSPOT

You are planning the deployment of Azure Data Lake Storage Gen2.

You have the following two reports that will access the data lake:

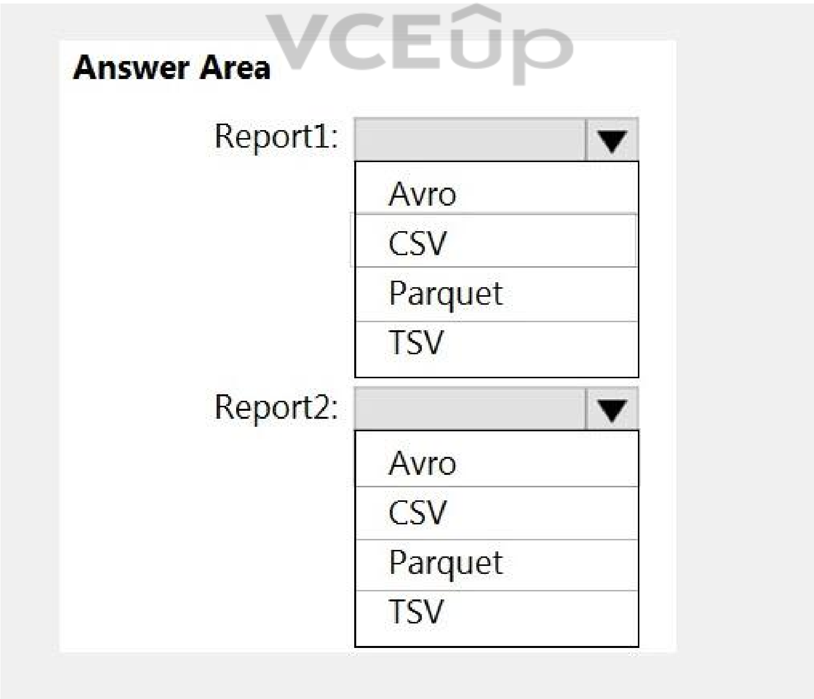
- Report1: Reads three columns from a file that contains 50 columns.
- Report2: Queries a single record based on a timestamp.

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.

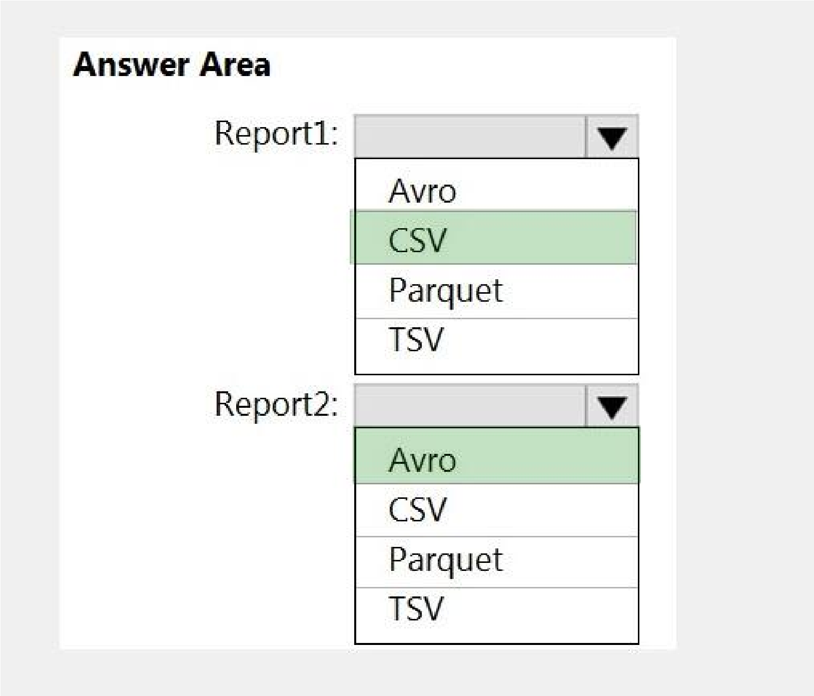
What should you recommend for each report? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:



Answer Area

Report1:
- Avro
- [✓ CSV]
- Parquet
- TSV

Report2:
- [✓ Avro]
- CSV
- Parquet
- TSV

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Report1: CSV
CSV: The destination writes records as delimited data.

Report2: AVRO
AVRO supports timestamps.
Not Parquet, TSV: Not options for Azure Data Lake Storage Gen2.

Reference:
https://streamsets.com/documentation/datacollector/latest/help/datacollector/UserGuide/Destinations/ADLS-G2-D.html

QUESTION 8
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: (none)
Explanation

Explanation/Reference:
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}/

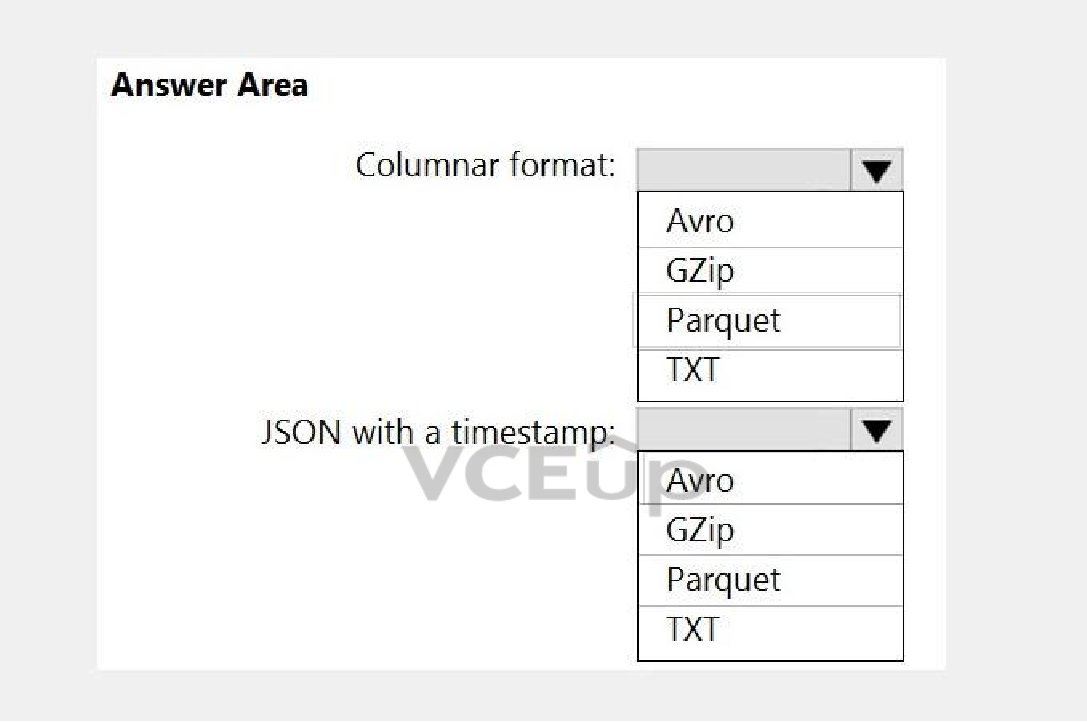
QUESTION 9
HOTSPOT

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:

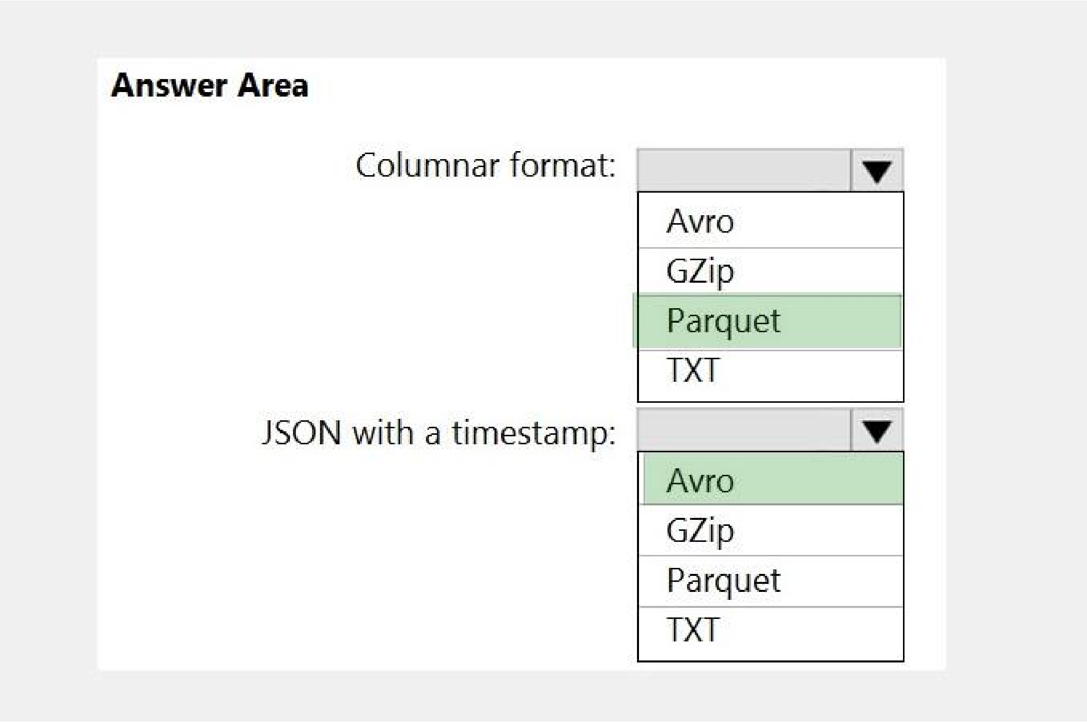


Answer Area

Columnar format: ▼
- Avro
- GZip
- Parquet
- TXT

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

Correct Answer:



Answer Area

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

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

Section: (none)
Explanation

Explanation/Reference:
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 10
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.

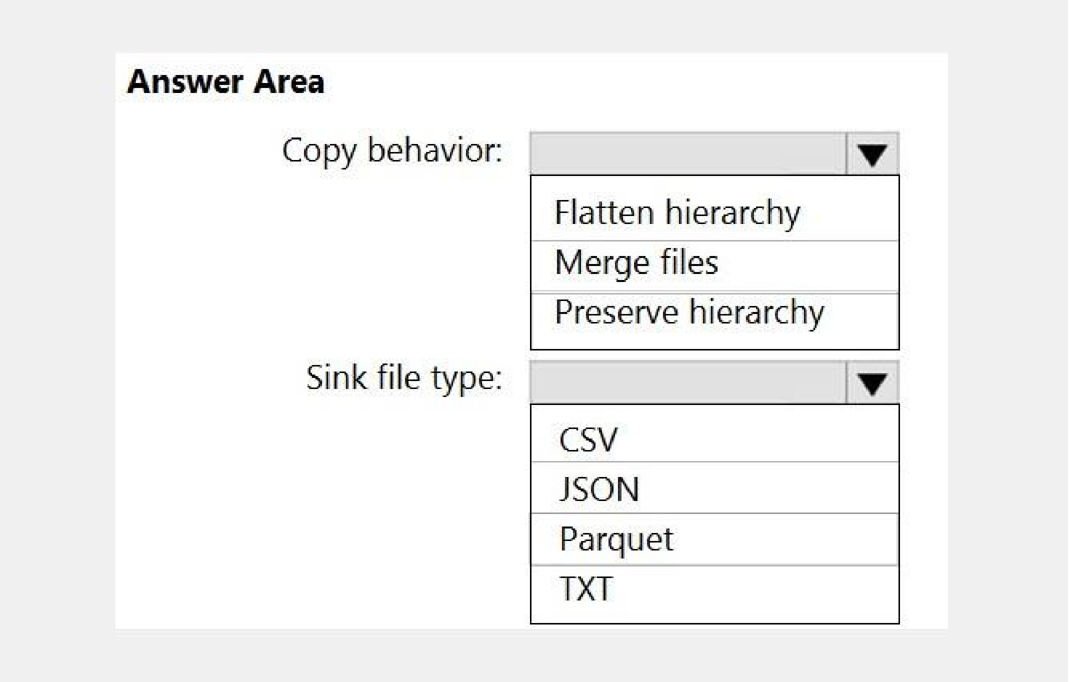
You need to move the files to a different folder and transform the data to meet the following requirements:

- 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

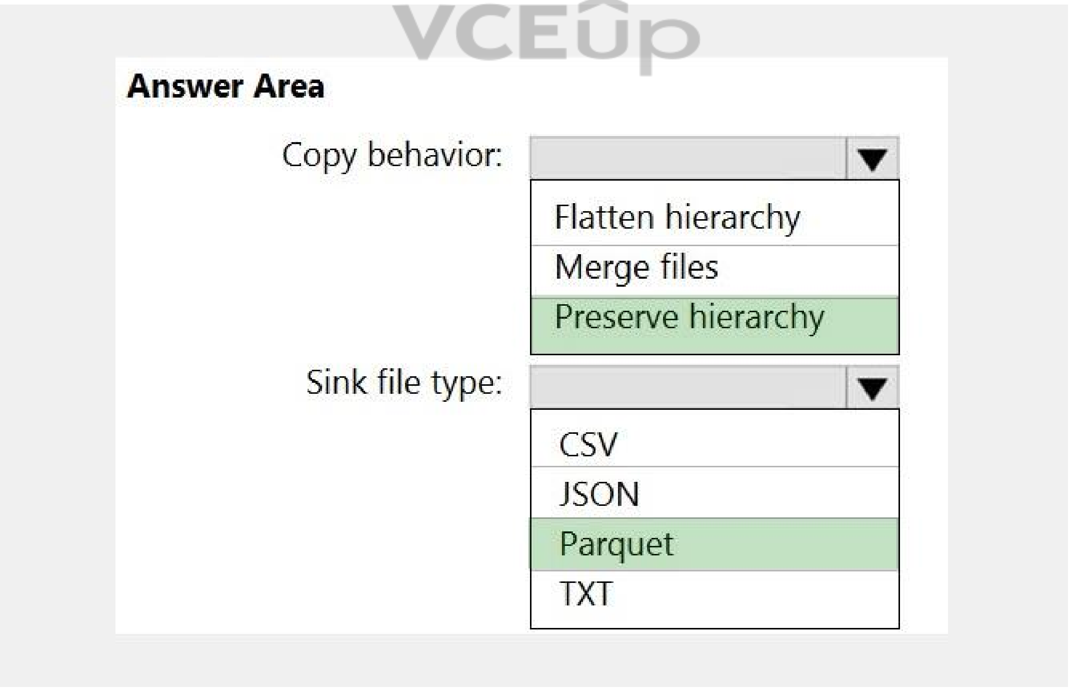
Copy behavior:

- Flatten hierarchy
- Merge files
- Preserve hierarchy

Sink file type:

- CSV
- JSON
- Parquet
- TXT

Correct Answer:



Answer Area

Copy behavior:

- Flatten hierarchy
- Merge files
- [✓ Preserve hierarchy]

Sink file type:

- CSV
- JSON
- [✓ Parquet]
- TXT

Section: (none)
Explanation

Explanation/Reference:
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.

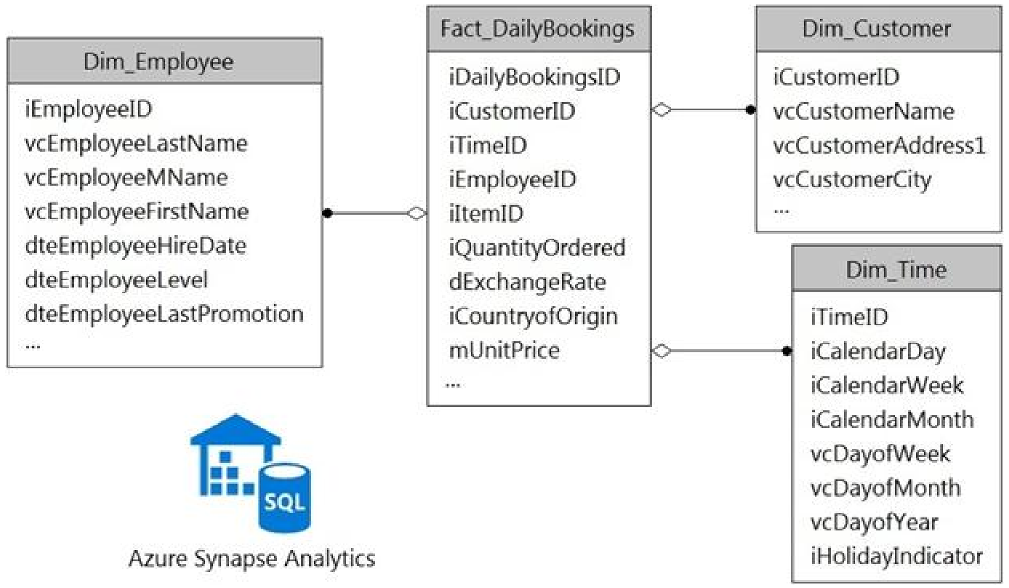
Box 2: Parquet
Azure Data Factory parquet format is supported for Azure Data Lake Storage Gen2.
Parquet supports the schema property.

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 11
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.

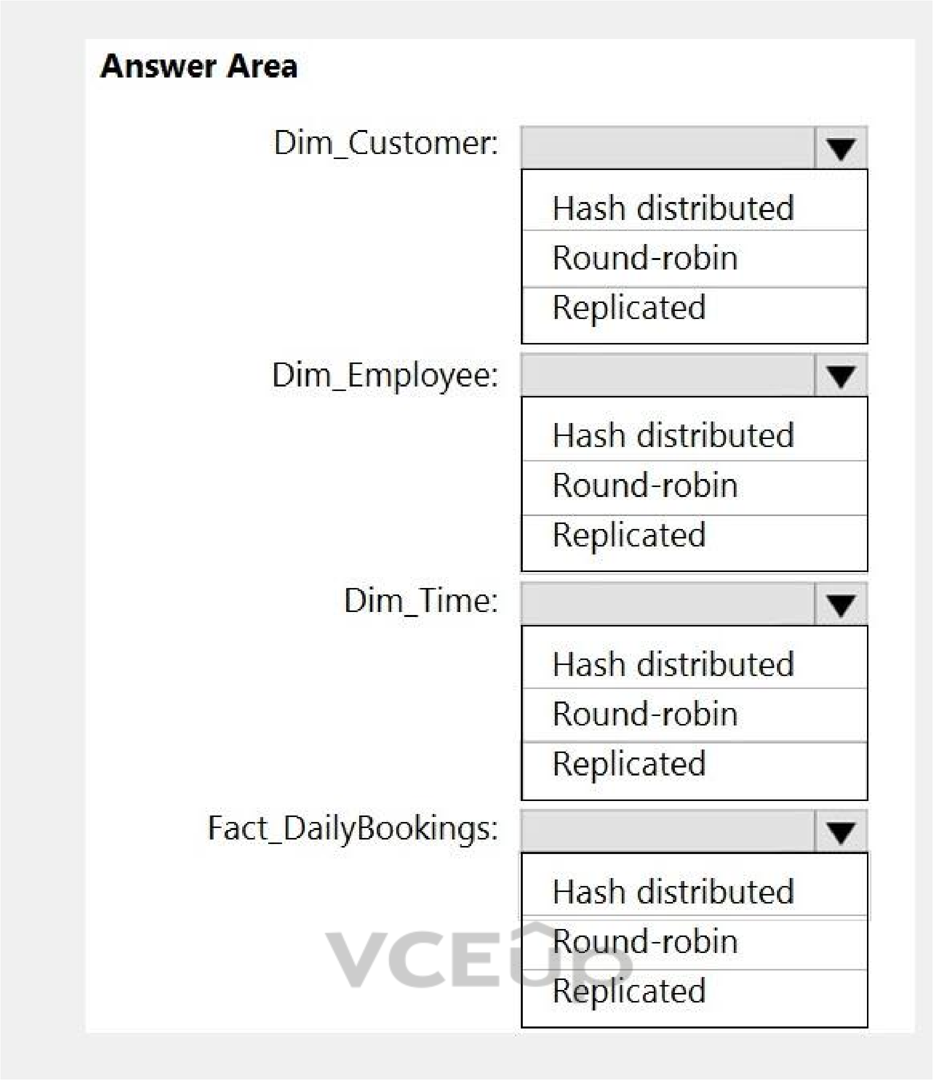


All the dimension tables will be less than 2 GB after compression, and the fact table will be approximately 6 TB. The dimension tables will be relatively static with very few data inserts and updates.

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

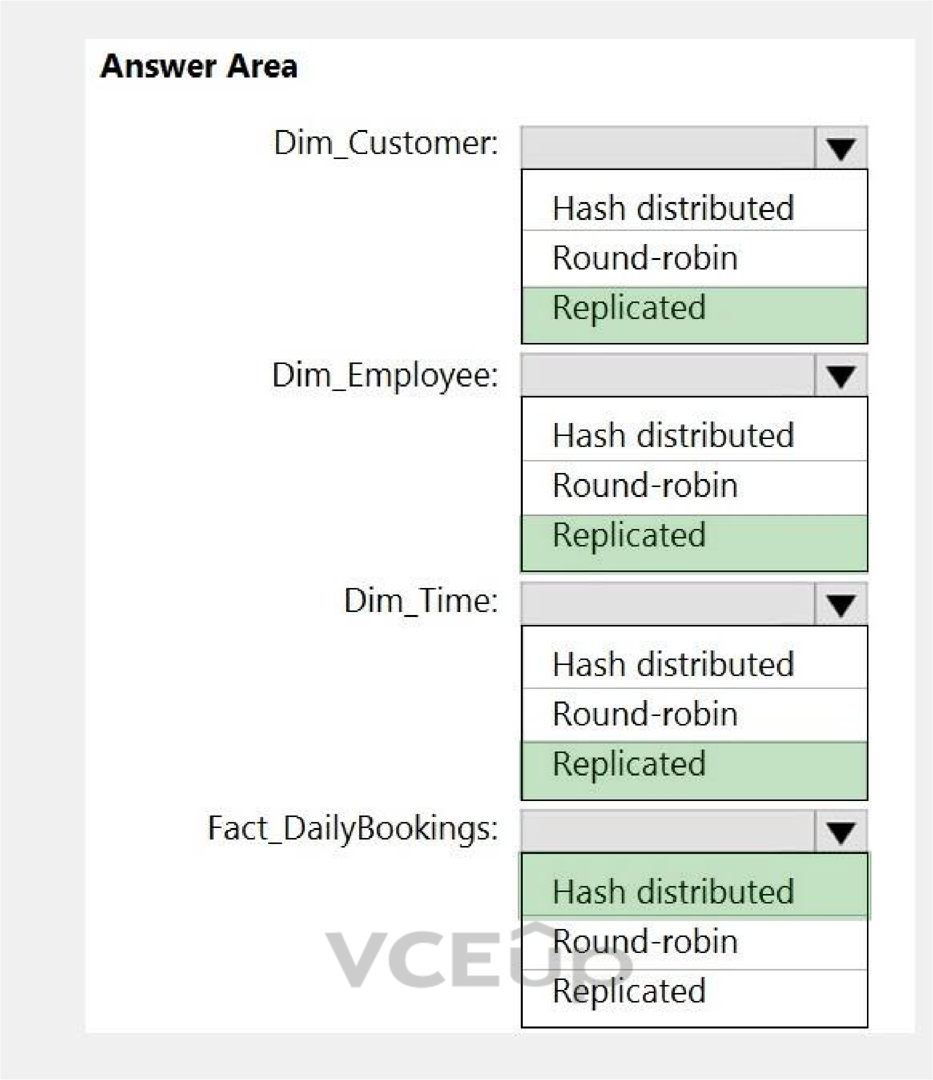
Dim_Customer:
- Hash distributed
- Round-robin
- Replicated

Dim_Employee:
- Hash distributed
- Round-robin
- Replicated

Dim_Time:
- Hash distributed
- Round-robin
- Replicated

Fact_DailyBookings:
- Hash distributed
- Round-robin
- Replicated

Correct Answer:



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: (none)
Explanation

Explanation/Reference:
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 12
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.

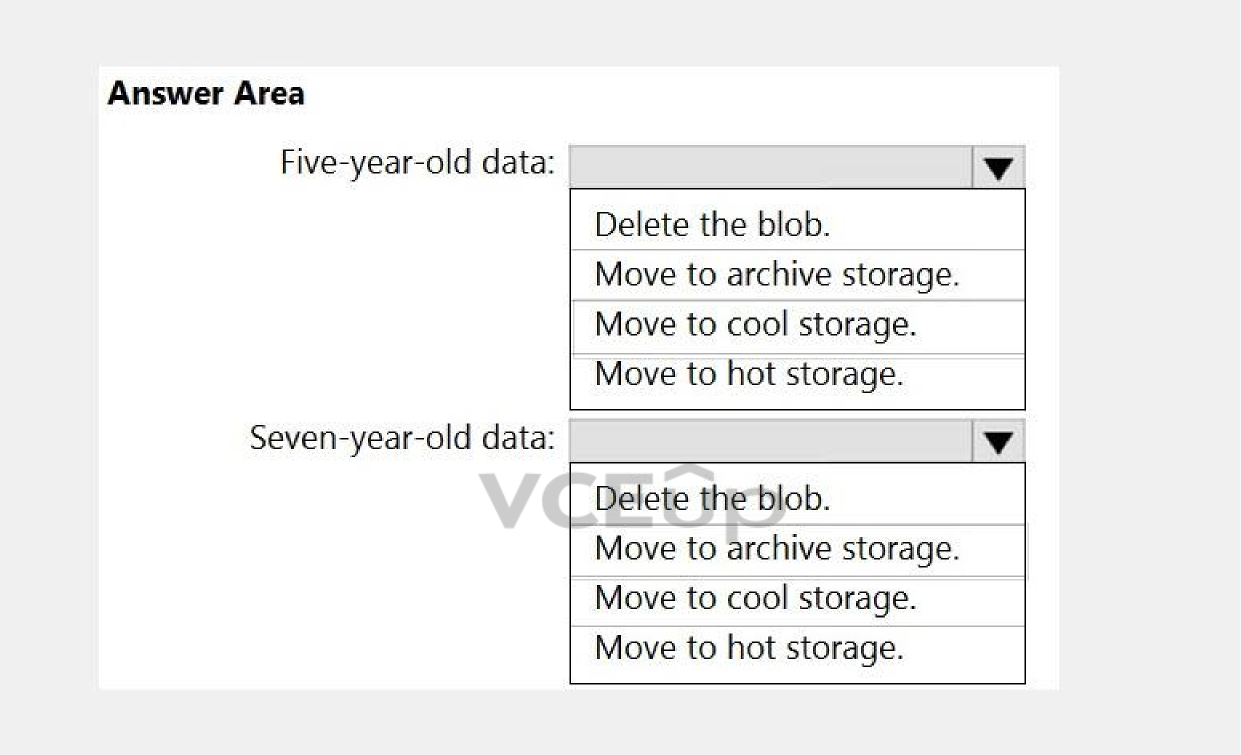
You need to design a data archiving solution that meets the following requirements:

- 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

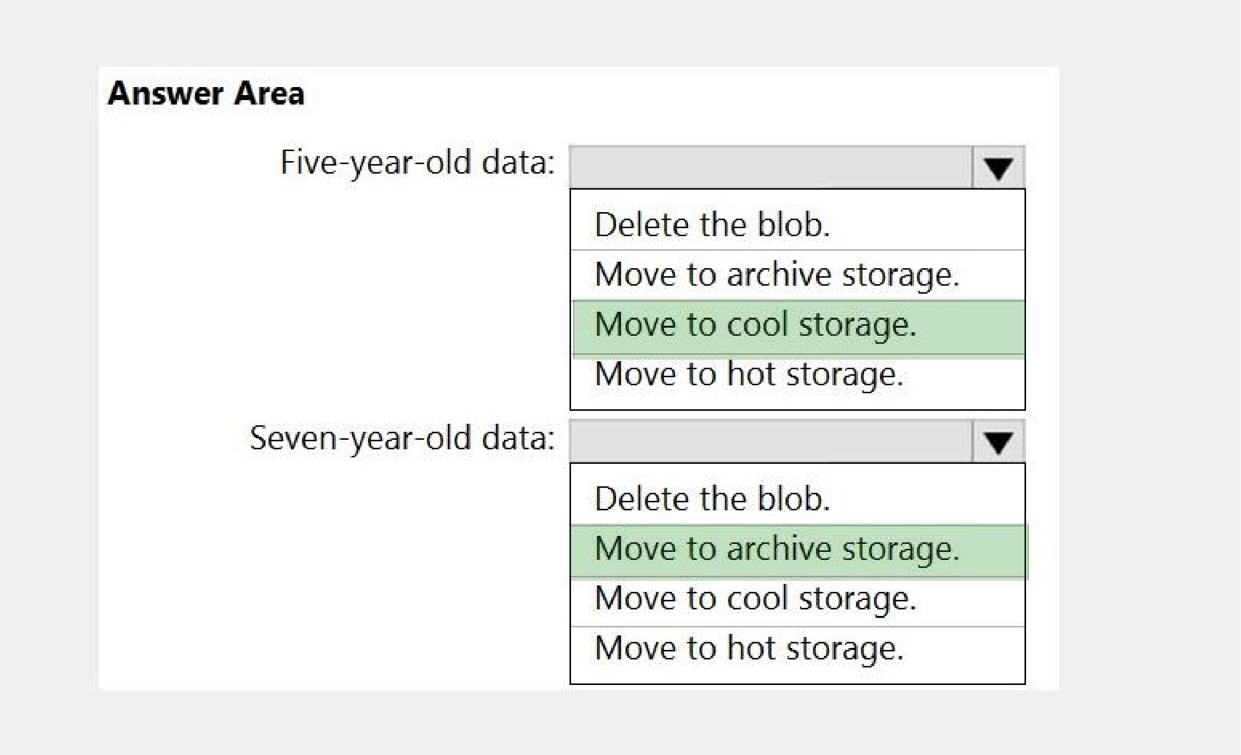
Five-year-old data: ▼

- 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.

Correct Answer:



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: (none)
Explanation

Explanation/Reference:
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

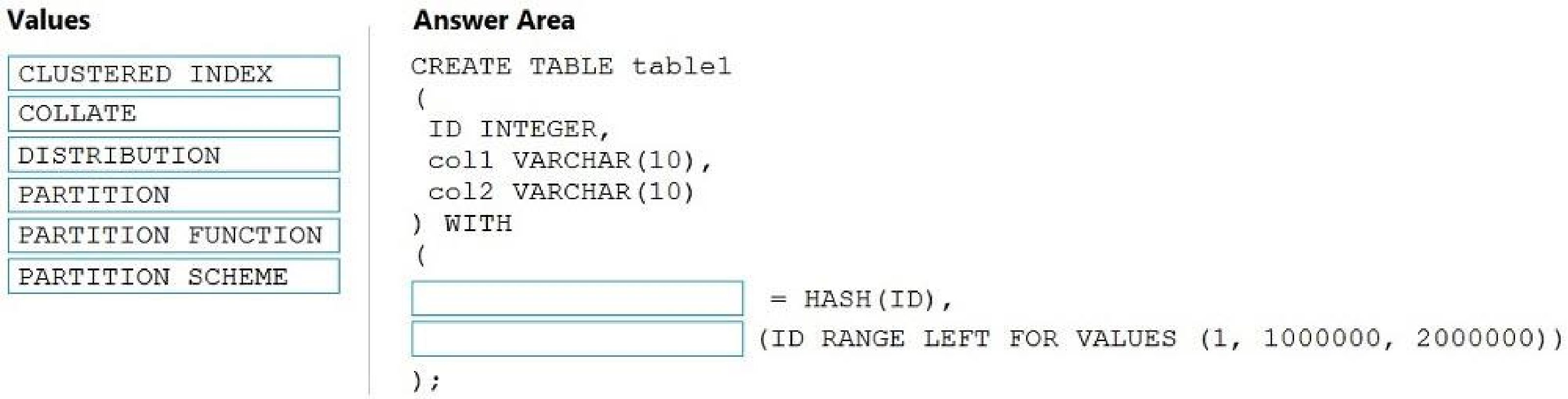
QUESTION 13
DRAG DROP

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:



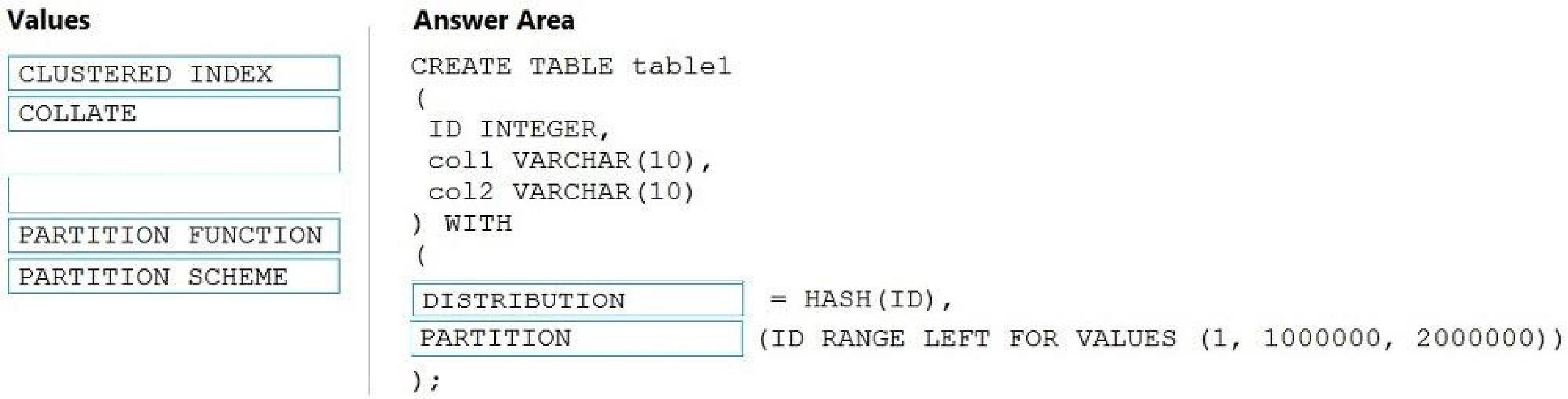
Values

- 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: (none)
Explanation

Explanation/Reference:
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 14
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: (none)

Explanation

Explanation/Reference:
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 15
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 snared access signature (SAS) as the credentials for the data load process.
F. Create a managed identity.

Correct Answer: ADF
Section: (none)
Explanation

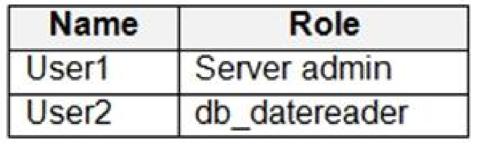
Explanation/Reference:
Explanation:
The managed identity grants permissions to the dedicated SQL pools in the workspace.

Note: Managed identity for Azure resources is a feature of Azure Active Directory. The feature provides Azure services with an automatically managed identity in Azure AD

Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-identity

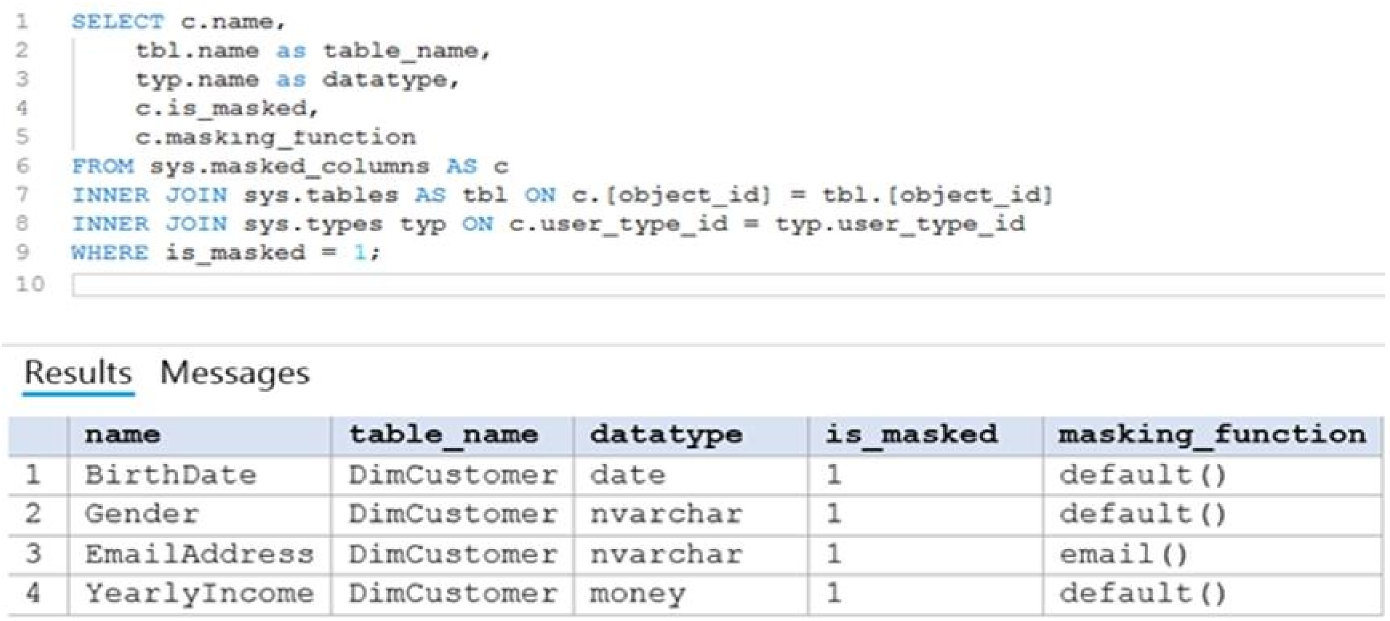
QUESTION 16
HOTSPOT

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.



```
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;
```

Results Messages

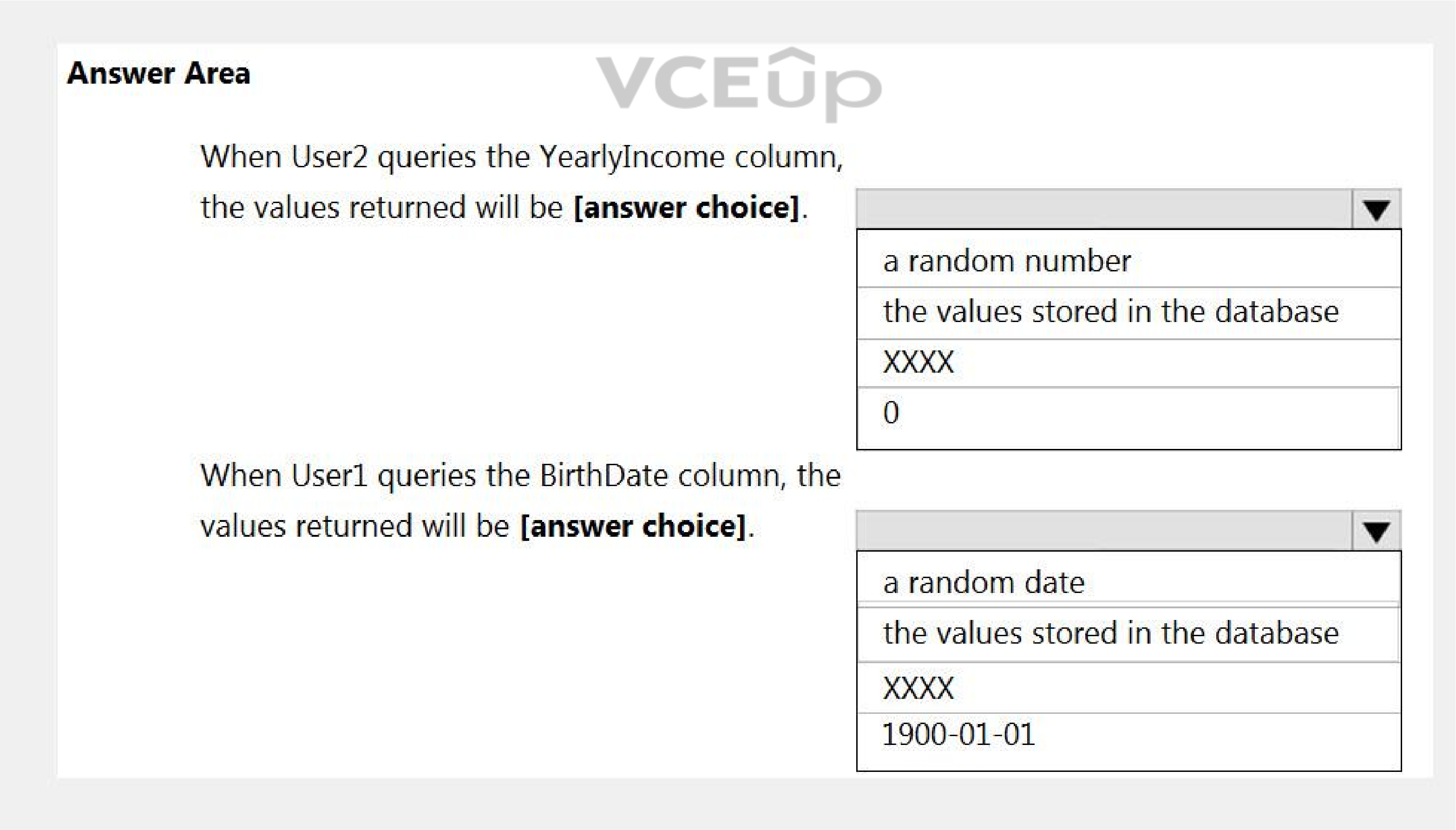
	name	table_name	datatype	is_masked	masking_function
1	BirthDate	DimCustomer	date	1	default()
2	Gender	DimCustomer	nvarchar	1	default()
3	EmailAddress	DimCustomer	nvarchar	1	email()
4	YearlyIncome	DimCustomer	money	1	default()

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

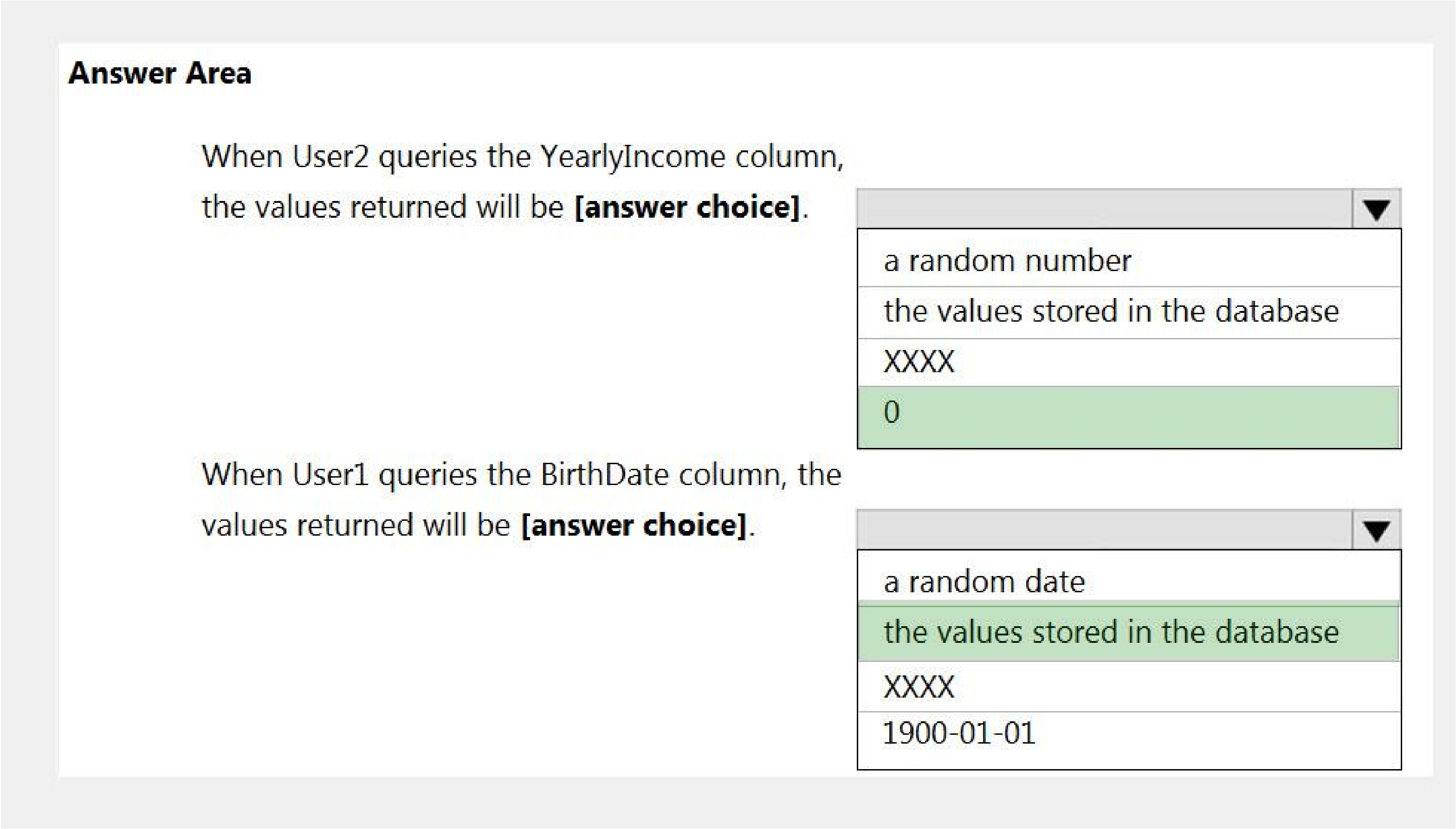
When User2 queries the YearlyIncome column, the values returned will be **[answer choice]**.

- 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

Correct Answer:



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: (none)
Explanation

Explanation/Reference:
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 17
You have an enterprise data warehouse in Azure Synapse Analytics.

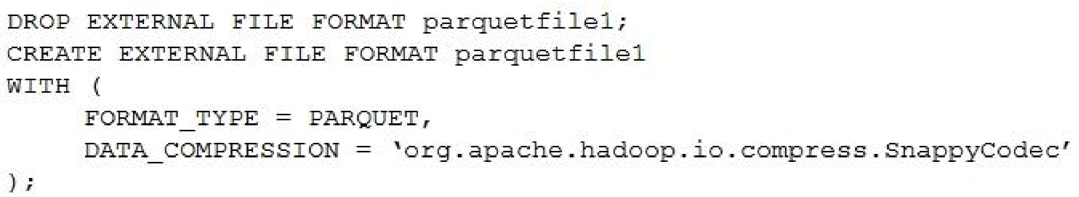
Using PolyBase, you create an external table named [Ext].[Items] to query Parquet files stored in Azure Data Lake Storage Gen2 without importing the data to the data warehouse.

The external table has three columns.

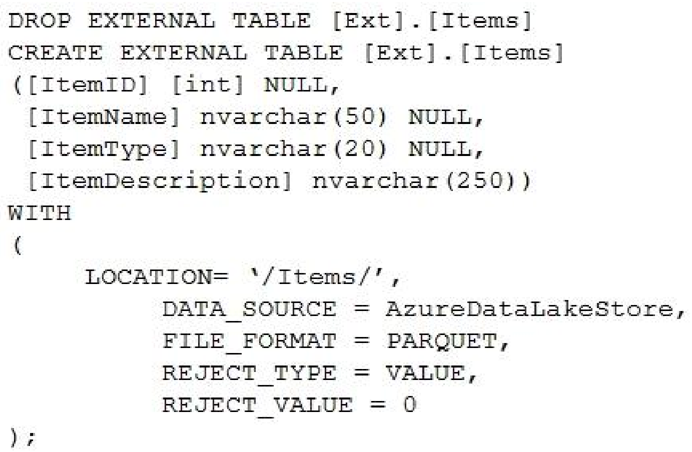
You discover that the Parquet files have a fourth column named ItemID.

Which command should you run to add the ItemID column to the external table?

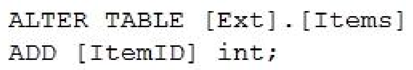
A.
```
ALTER EXTERNAL TABLE [Ext].[Items]
   ADD [ItemID] int;
```

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;
```

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
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 18
HOTSPOT

You have two Azure Storage accounts named Storage1 and Storage2. Each account holds one container and has the hierarchical namespace enabled. The system has files that contain data stored in the Apache Parquet format.

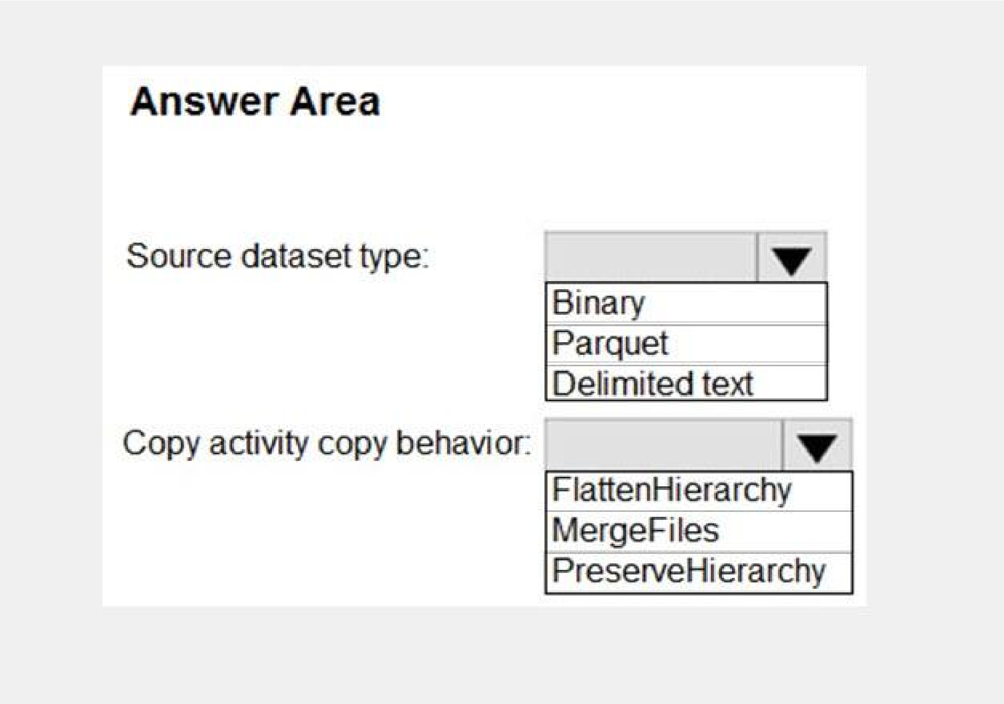
You need to copy folders and files from Storage1 to Storage2 by using a Data Factory copy activity. The solution must meet the following requirements:

- No transformations must be performed.
- The original folder structure must be retained.
- Minimize time required to perform the copy activity.

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

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

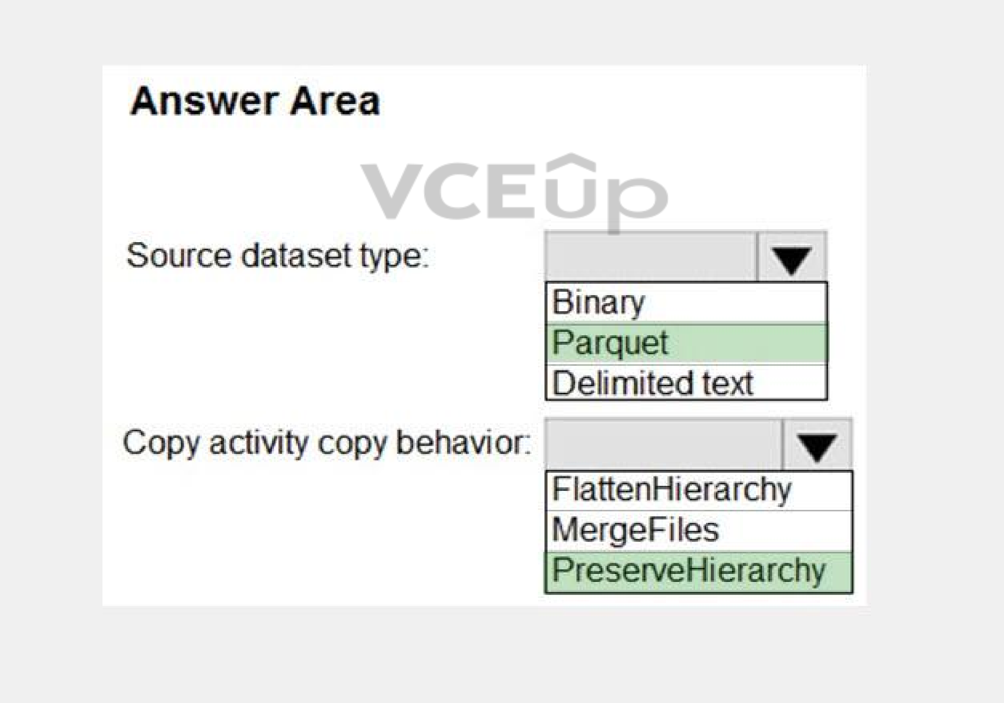
Source dataset type:

- Binary
- Parquet
- Delimited text

Copy activity copy behavior:

- FlattenHierarchy
- MergeFiles
- PreserveHierarchy

Correct Answer:



Answer Area

Source dataset type:

- Binary
- [✓ Parquet]
- Delimited text

Copy activity copy behavior:

- FlattenHierarchy
- MergeFiles
- [✓ PreserveHierarchy]

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: Parquet
For Parquet datasets, the type property of the copy activity source must be set to ParquetSource.

Box 2: PreserveHierarchy
PreserveHierarchy (default): Preserves the file hierarchy in the target folder. The relative path of the source file to the source folder is identical to the relative path of the target file to the target folder.

Incorrect Answers:

- FlattenHierarchy: All files from the source folder are in the first level of the target folder. The target files have autogenerated names.
- MergeFiles: Merges all files from the source folder to one file. If the file name is specified, the merged file name is the specified name. Otherwise, it's an autogenerated file name.

Reference:
https://docs.microsoft.com/en-us/azure/data-factory/format-parquet

https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage

QUESTION 19
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: (none)
Explanation

Explanation/Reference:
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 20
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: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Locally redundant storage (LRS) copies your data synchronously three times within a single physical location in the primary region. LRS is the least expensive replication option

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

QUESTION 21
HOTSPOT

You have a SQL pool in Azure Synapse.

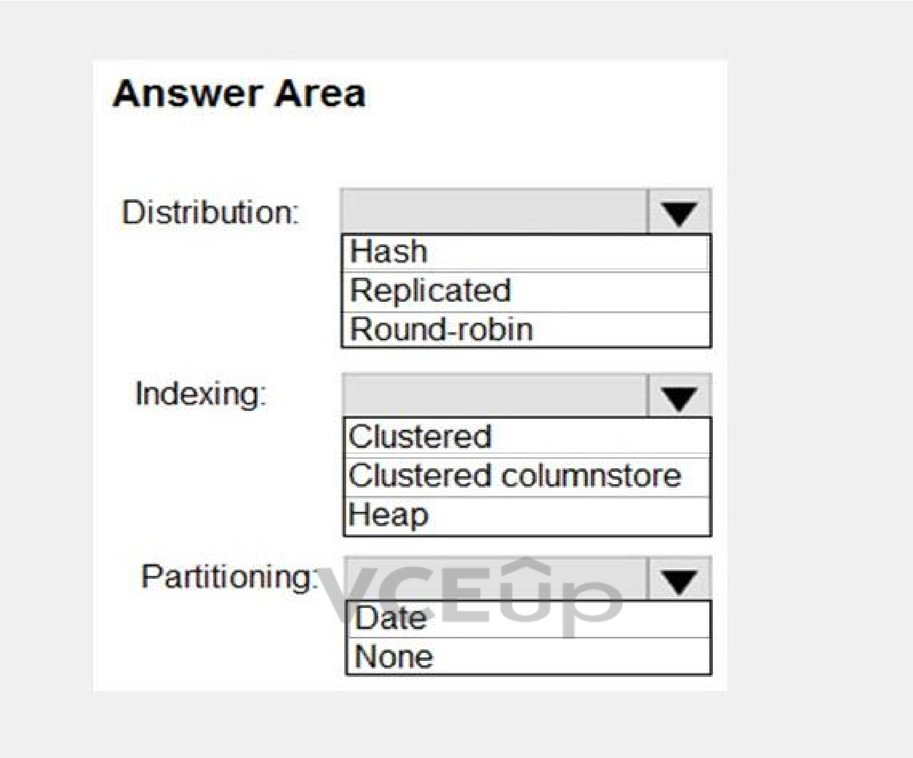
You plan to load data from Azure Blob storage to a staging table. Approximately 1 million rows of data will be loaded daily. The table will be truncated before each daily load.

You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.

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

NOTE: Each correct selection is worth one point.

Hot Area:



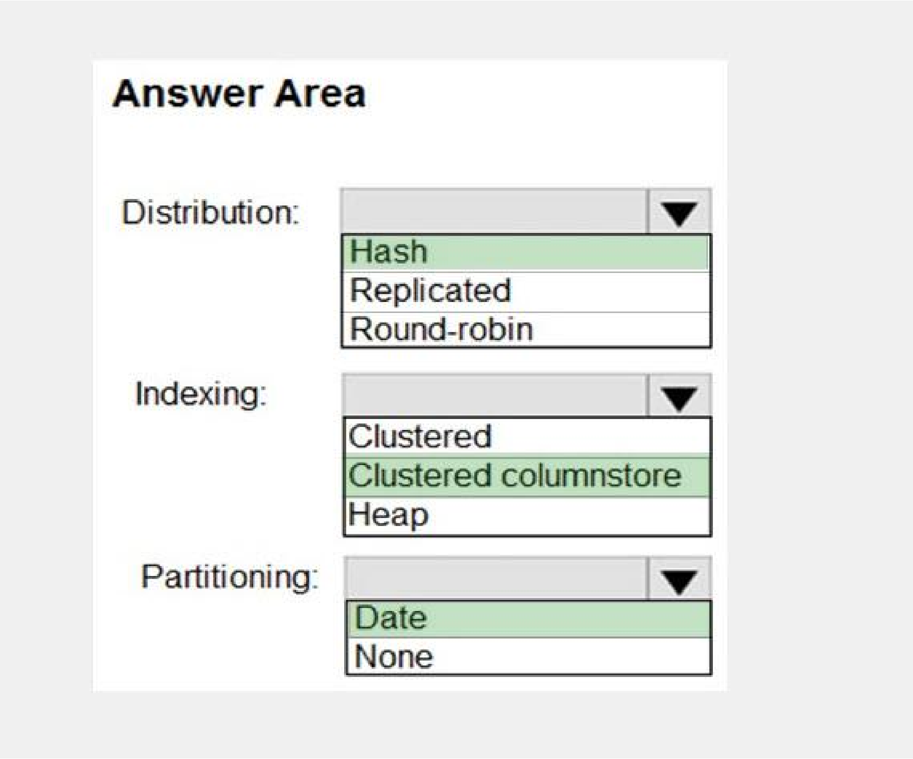
Answer Area

Distribution:
- Hash
- Replicated
- Round-robin

Indexing:
- Clustered
- Clustered columnstore
- Heap

Partitioning:
- Date
- None

Correct Answer:



Answer Area

Distribution:
[✓ Hash]
Replicated
Round-robin

Indexing:
Clustered
[✓ Clustered columnstore]
Heap

Partitioning:
[✓ Date]
None

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: Hash
Hash-distributed tables improve query performance on large fact tables. They can have very large numbers of rows and still achieve high performance.

Incorrect Answers:
Round-robin tables are useful for improving loading speed.

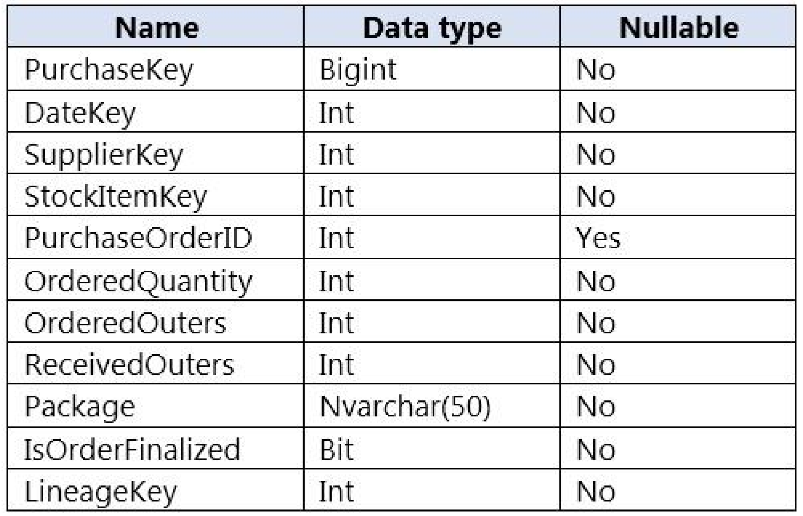
Box 2: Clustered columnstore
When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed.

Box 3: Date
Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.
Partition switching can be used to quickly remove or replace a section of a table.

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

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

QUESTION 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: (none)
Explanation

Explanation/Reference:
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.

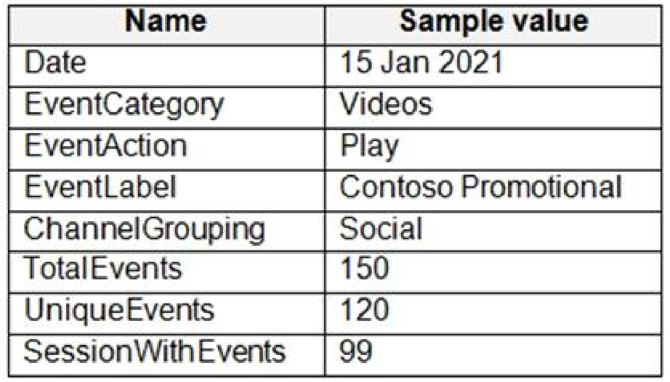
Incorrect Answers:
C: Round-robin tables are useful for improving loading speed.

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

QUESTION 23
HOTSPOT

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

The data contains the following columns.



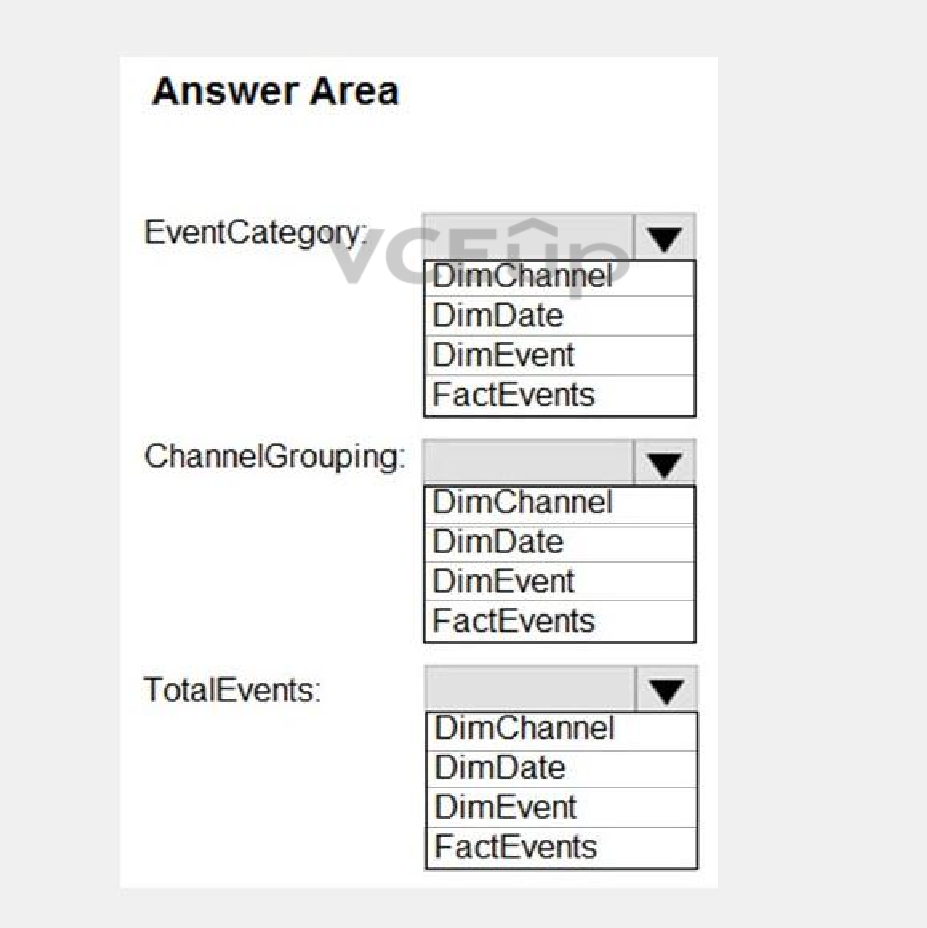
Name	Sample value
Date	15 Jan 2021
EventCategory	Videos
EventAction	Play
EventLabel	Contoso Promotional
ChannelGrouping	Social
TotalEvents	150
UniqueEvents	120
SessionWithEvents	99

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

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

NOTE: Each correct selection is worth one point.

Hot Area:



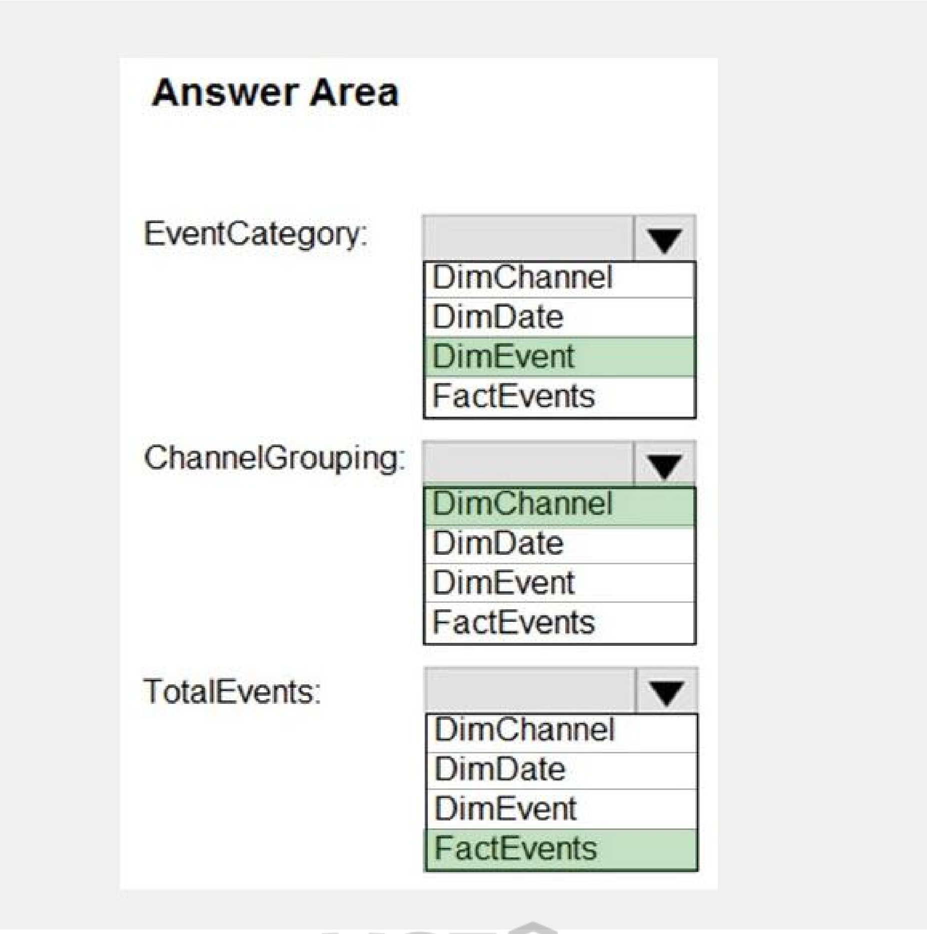
Answer Area

EventCategory: ▼
DimChannel
DimDate
DimEvent
FactEvents

ChannelGrouping: ▼
DimChannel
DimDate
DimEvent
FactEvents

TotalEvents: ▼
DimChannel
DimDate
DimEvent
FactEvents

Correct Answer:



Answer Area

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

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

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

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: DimEvent

Box 2: DimChannel

Box 3: FactEvents
Fact tables store observations or events, and can be sales orders, stock balances, exchange rates, temperatures, etc

Reference:
https://docs.microsoft.com/en-us/power-bi/guidance/star-schema

QUESTION 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 convert the files to compressed delimited text files.

Does this meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
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 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 copy the files to a table that has a columnstore index.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
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 26
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: A
Section: (none)
Explanation

Explanation/Reference:
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

01 - Design and develop data processing

QUESTION 1
Case study

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study
To display the first question in this case study, click the **Next** button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an **All Information** tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the **Question** button to return to the question.

Overview

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements

Business Goals

Litware wants to create a new analytics environment in Azure to meet the following requirements:

- See inventory levels across the stores. Data must be updated as close to real time as possible.
- Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.
- Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements

Litware identifies the following technical requirements:

- Minimize the number of different Azure services needed to achieve the business goals.
- Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.
- Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.
- Use Azure Active Directory (Azure AD) authentication whenever possible.
- Use the principle of least privilege when designing security.
- Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.
- Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.
- Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment

Litware plans to implement the following environment:

- The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.
- Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
- Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
- Daily inventory data comes from a Microsoft SQL server located on a private network.
- Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.
- Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.
- Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

A.

Correct Answer:
Section: (none)
Explanation

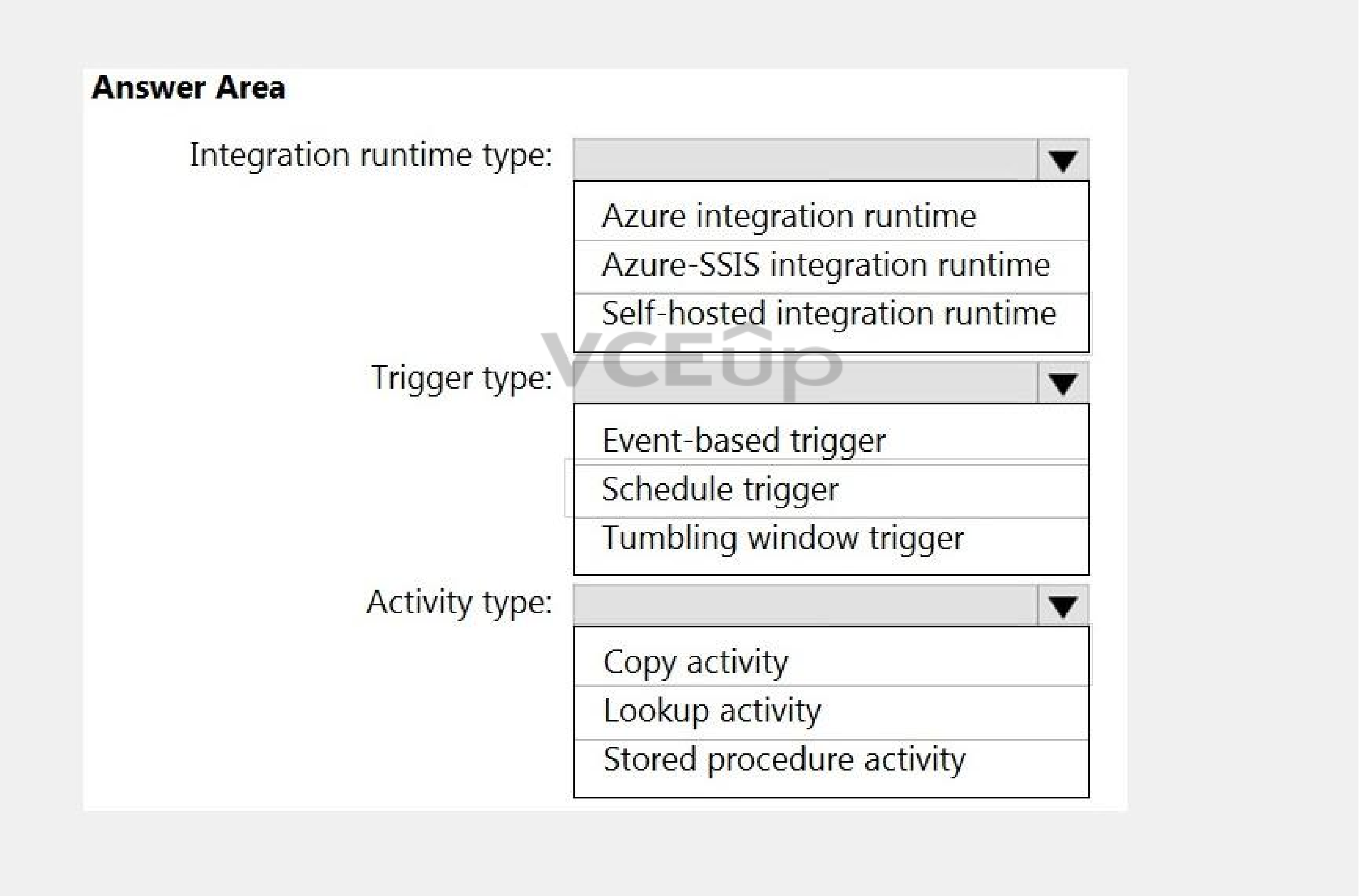
Explanation/Reference:

QUESTION 2
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:



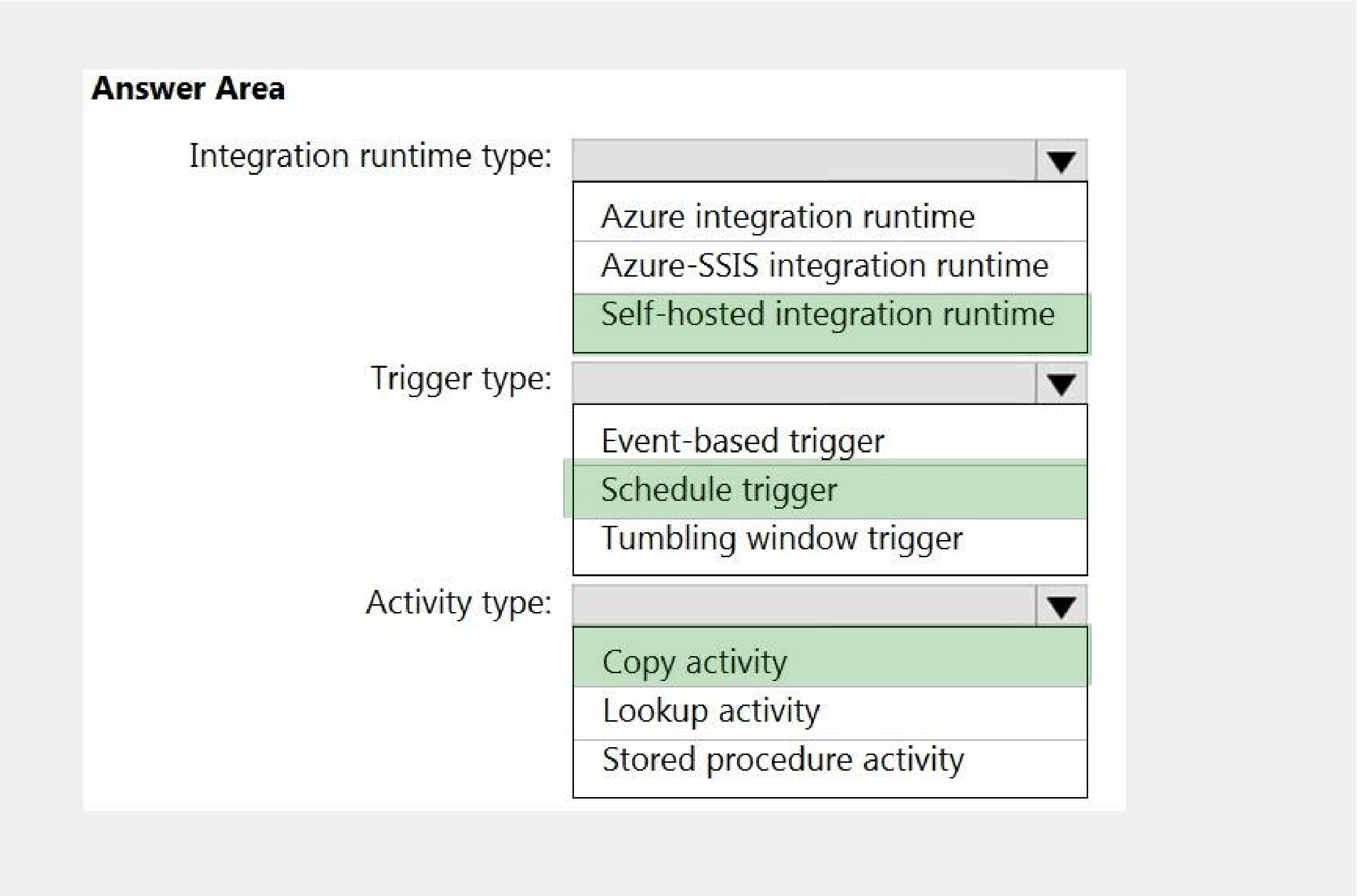
Answer Area

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

Correct Answer:



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: (none)
Explanation

Explanation/Reference:
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.

02 -Design and develop data processing

QUESTION 1
HOTSPOT

You are building an Azure Stream Analytics job to identify how much time a user spends interacting with a feature on a webpage.

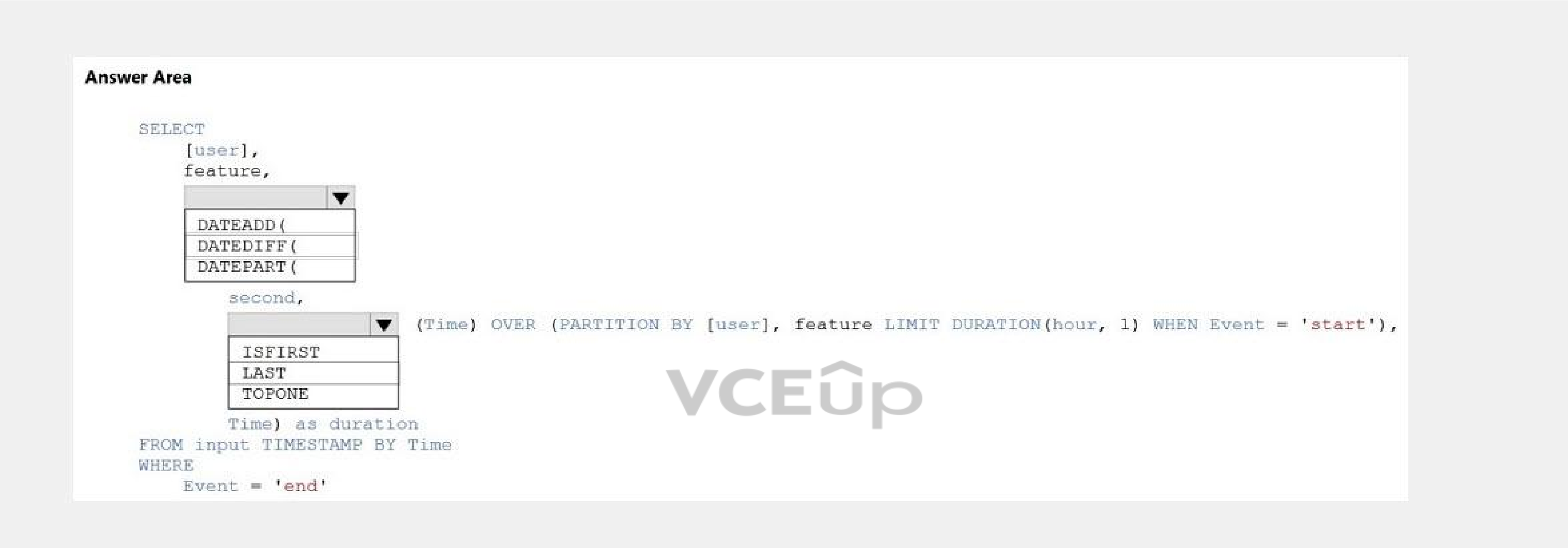
The job receives events based on user actions on the webpage. Each row of data represents an event. Each event has a type of either `'start'` or `'end'`.

You need to calculate the duration between start and end events.

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

NOTE: Each correct selection is worth one point.

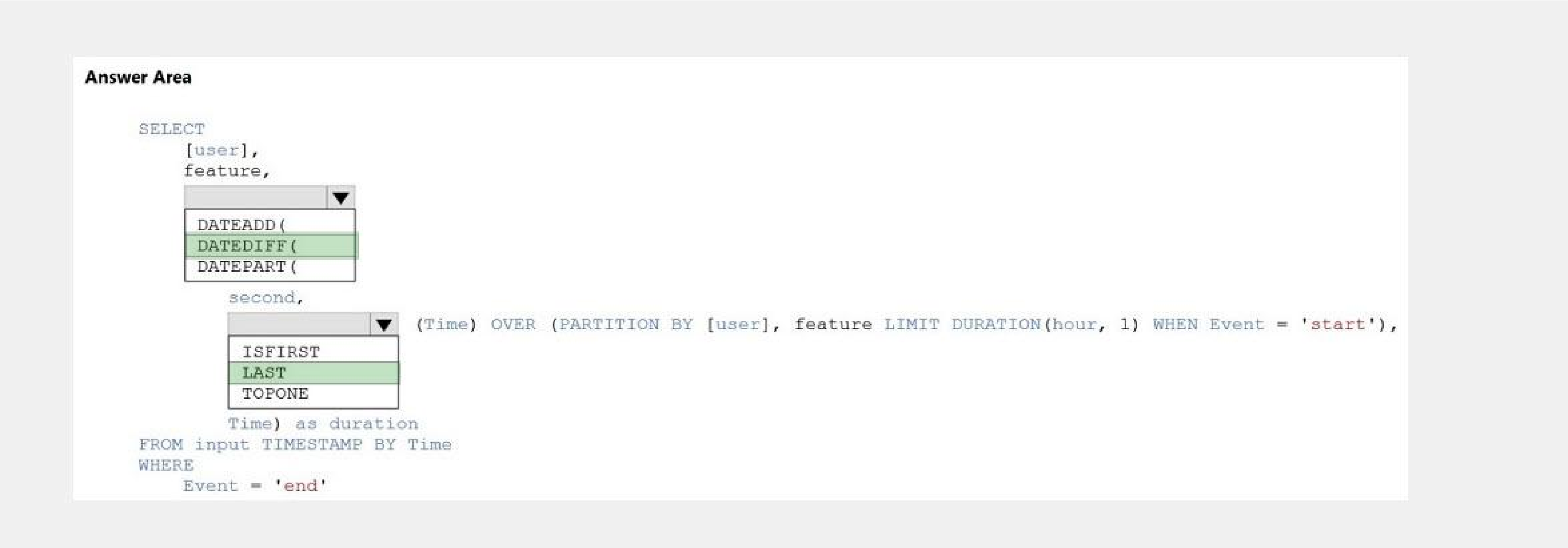
Hot Area:



Answer Area

```
SELECT
    [user],
    feature,
    [ ▼ ]
    DATEADD(
    DATEDIFF(
    DATEPART(
        second,
        [ ▼ ]  (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
        ISFIRST
        LAST
        TOPONE
        Time) as duration
FROM input TIMESTAMP BY Time
WHERE
    Event = 'end'
```

Correct Answer:



Answer Area

```
SELECT
    [user],
    feature,
    [ ▼ ]
    DATEADD(
    [✓ DATEDIFF(]
    DATEPART(
        second,
        [ ▼ ]  (Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),
        ISFIRST
        [✓ LAST]
        TOPONE
        Time) as duration
FROM input TIMESTAMP BY Time
WHERE
    Event = 'end'
```

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: DATEDIFF
DATEDIFF function returns the count (as a signed integer value) of the specified datepart boundaries crossed between the specified startdate and enddate.
Syntax: DATEDIFF (datepart , startdate, enddate)

Box 2: LAST
The LAST function can be used to retrieve the last event within a specific condition. In this example, the condition is an event of type Start, partitioning the search by PARTITION BY user and feature. This way, every user and feature is treated independently when searching for the Start event. LIMIT DURATION limits the search back in time to 1 hour between the End and Start events.

Example:

```
SELECT
    [user],
    feature,
    DATEDIFF(
        second,
        LAST(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour,
            1) WHEN Event = 'start'),
        Time) as duration
FROM input TIMESTAMP BY Time
WHERE
    Event = 'end'
```

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

QUESTION 2
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.

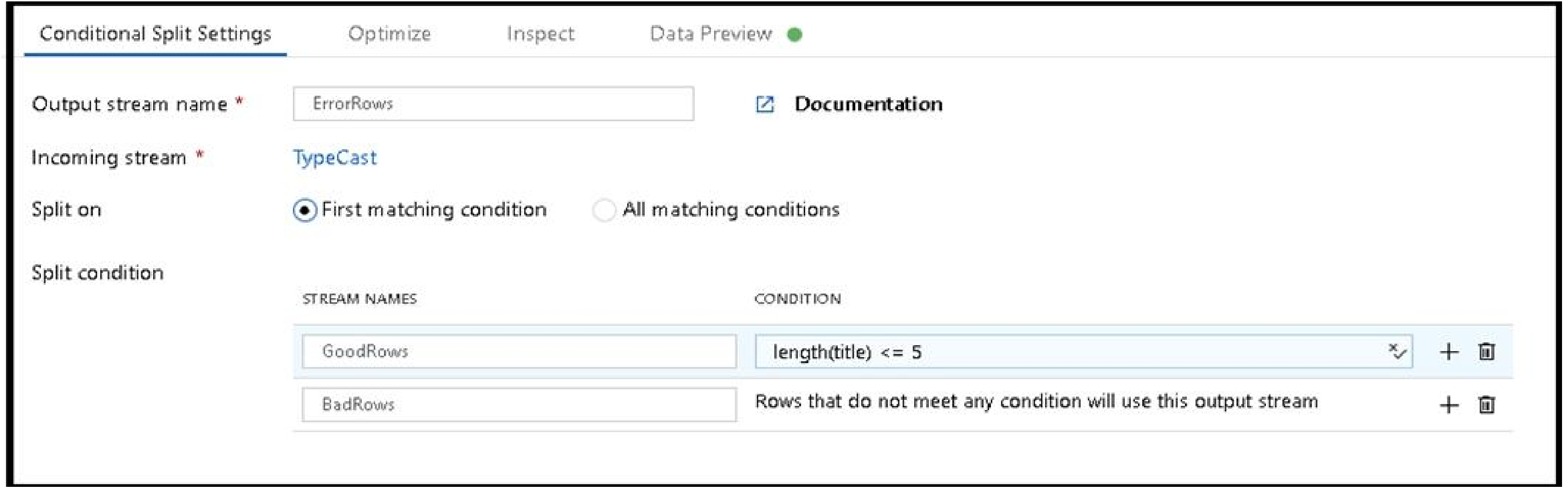
NOTE: Each correct selection is worth one point.

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: AB
Section: (none)
Explanation

Explanation/Reference:
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

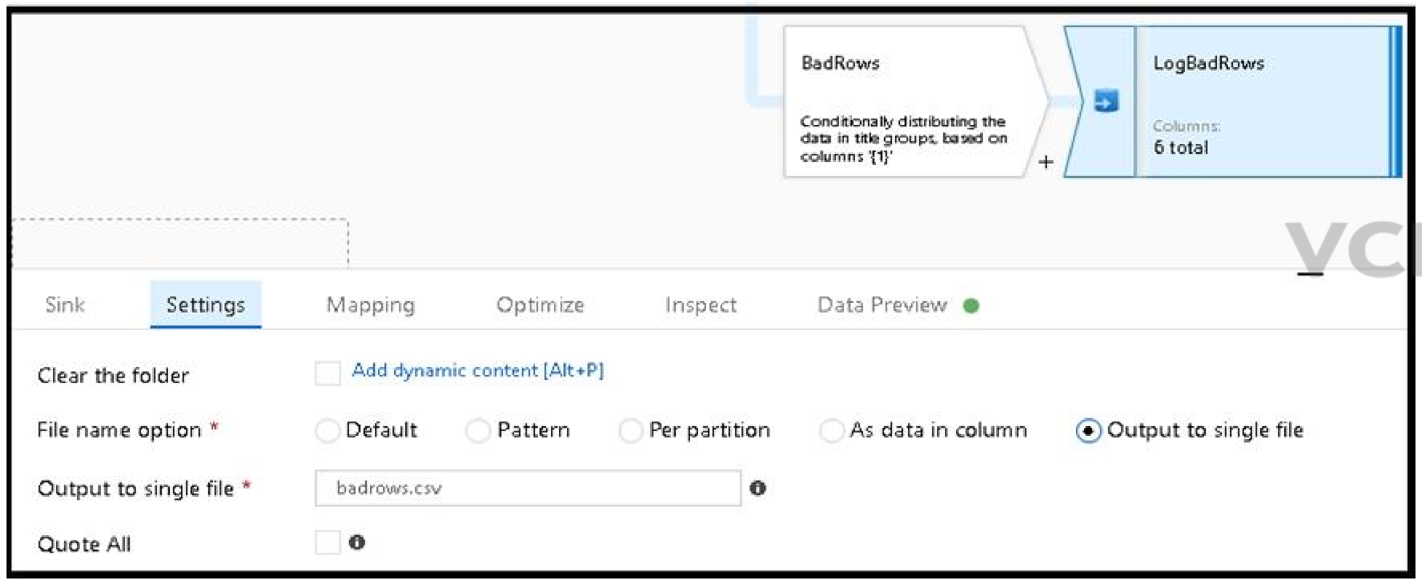
Split condition

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".



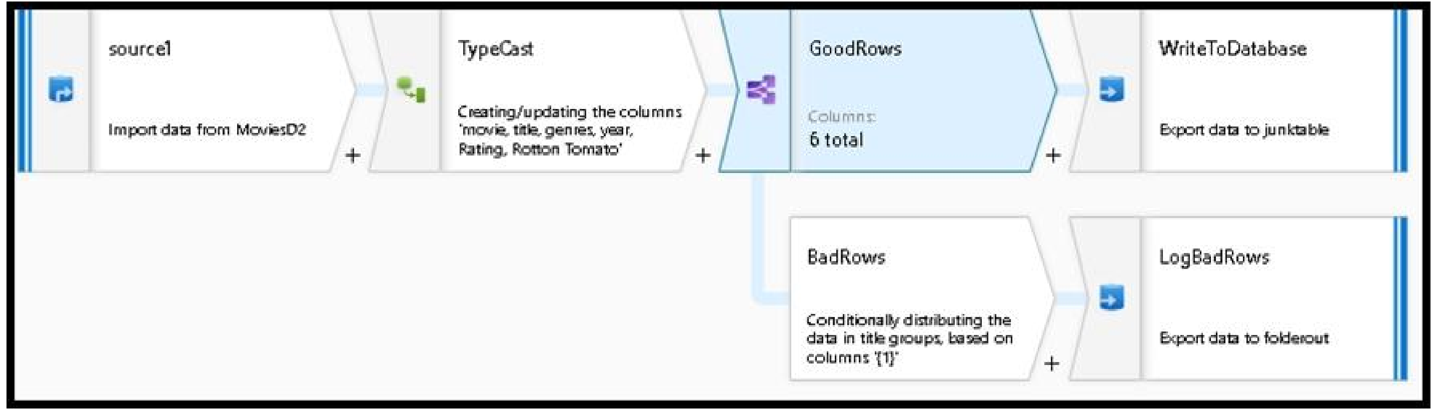
Clear the folder □ Add dynamic content [Alt+P]

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

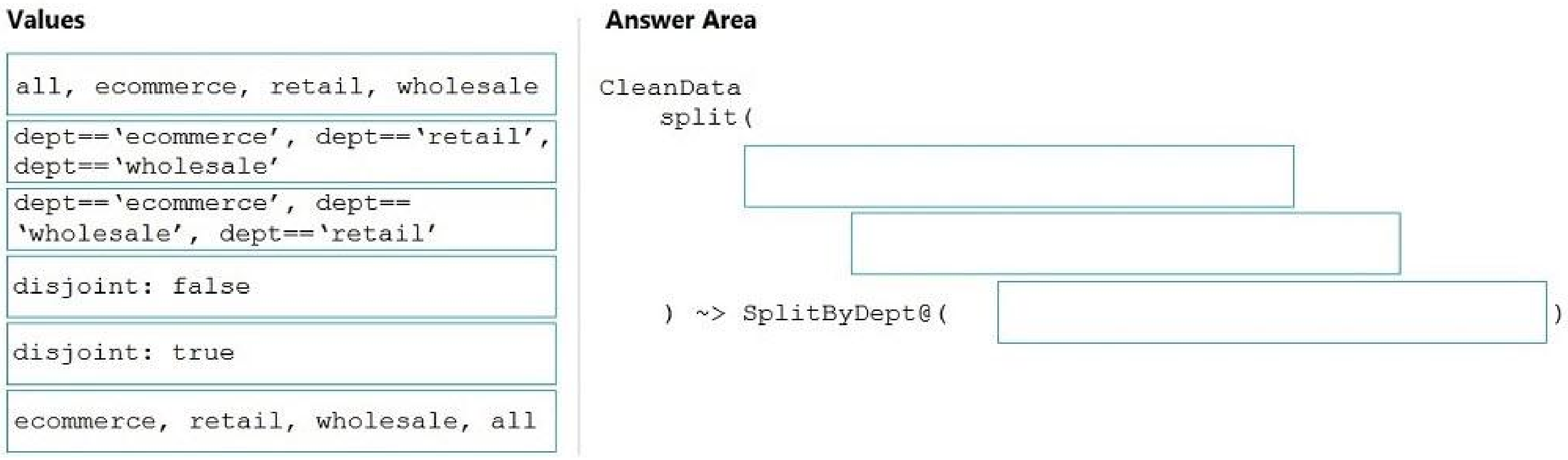
QUESTION 3
DRAG DROP

You need to create an Azure Data Factory pipeline to process data for the following three departments at your company: Ecommerce, retail, and wholesale. The solution must ensure that data can also be processed for the entire company.

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

NOTE: Each correct selection is worth one point.

Select and Place:



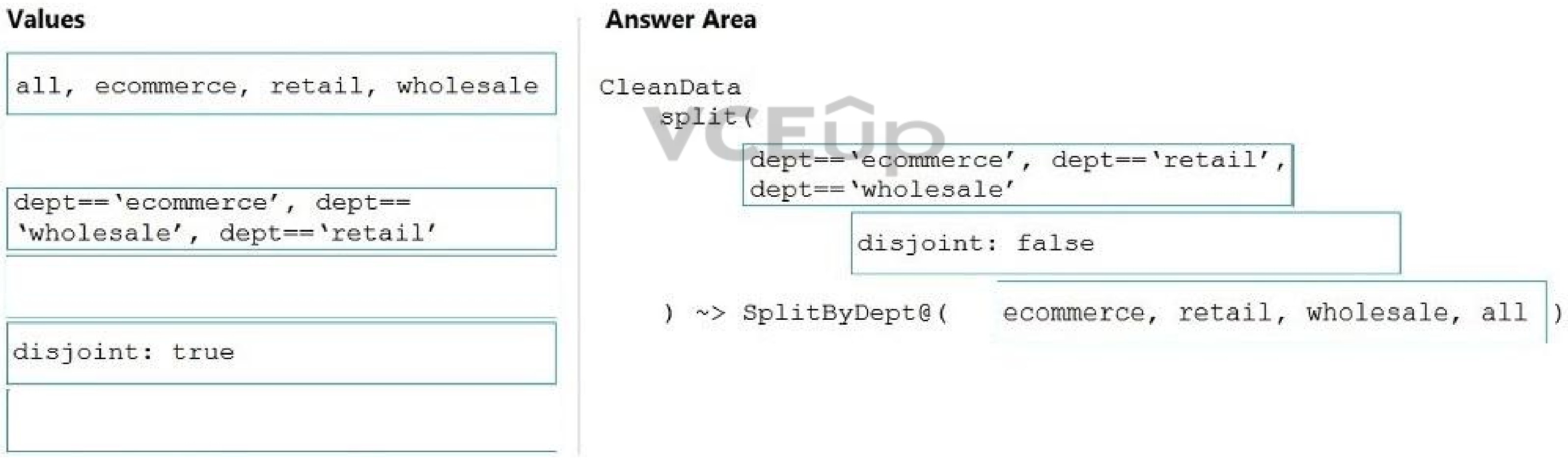
Values

- all, ecommerce, retail, wholesale
- dept==‘ecommerce’, dept==‘retail’, dept==‘wholesale’
- dept==‘ecommerce’, dept== ‘wholesale’, dept==‘retail’
- disjoint: false
- disjoint: true
- ecommerce, retail, wholesale, all

Answer Area

```
CleanData
    split(
          [          ]
                 [          ]
    ) ~> SplitByDept@( [          ] )
```

Correct Answer:



Values

- all, ecommerce, retail, wholesale
- dept==‘ecommerce’, dept== ‘wholesale’, dept==‘retail’
- disjoint: true

Answer Area

```
CleanData
    split(
          dept==‘ecommerce’, dept==‘retail’,
          dept==‘wholesale’
                 disjoint: false
    ) ~> SplitByDept@( ecommerce, retail, wholesale, all )
```

Section: (none)
Explanation

Explanation/Reference:
Explanation:

The conditional split transformation routes data rows to different streams based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language. The transformation evaluates expressions, and based on the results, directs the data row to the specified stream.

Box 1: dept=='ecommerce', dept=='retail', dept=='wholesale'
First we put the condition. The order must match the stream labeling we define in Box 3.

Syntax:

```
<incomingStream>
  split(
      <conditionalExpression1>
      <conditionalExpression2>
      ...
      disjoint: {true | false}
```

) ~> <splitTx>@(stream1, stream2, ..., <defaultStream>)

Box 2: discount : false
disjoint is false because the data goes to the first matching condition. All remaining rows matching the third condition go to output stream all.

Box 3: ecommerce, retail, wholesale, all
Label the streams

Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-conditional-split

QUESTION 4
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`.

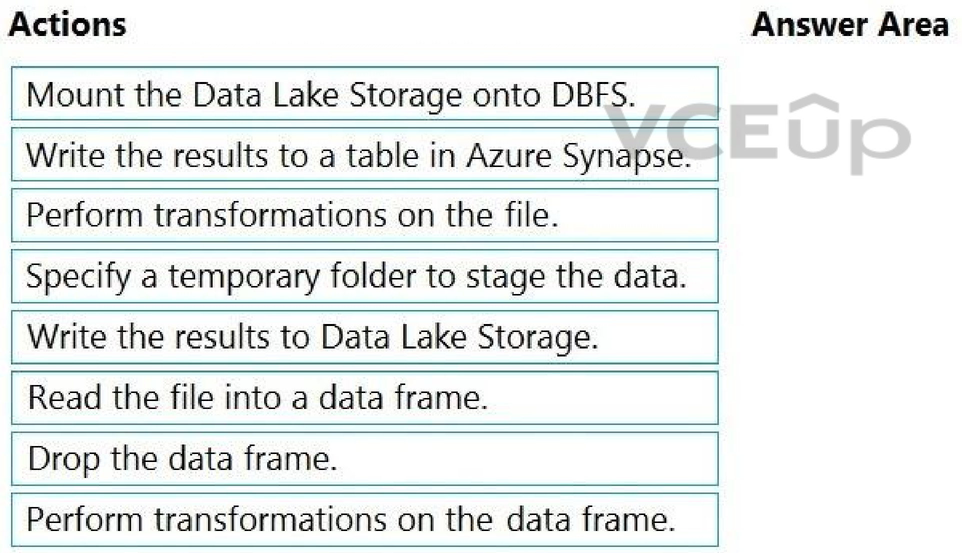
You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the `FirstName` and `LastName` values.

You create the following components:

- A destination table in Azure Synapse
- An Azure Blob storage container
- A service principal

Which five actions should you perform in sequence next in is Databricks notebook? 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

Answer Area

Mount the Data Lake Storage onto DBFS.

Write the results to a table in Azure Synapse.

Perform transformations on the file.

Specify a temporary folder to stage the data.

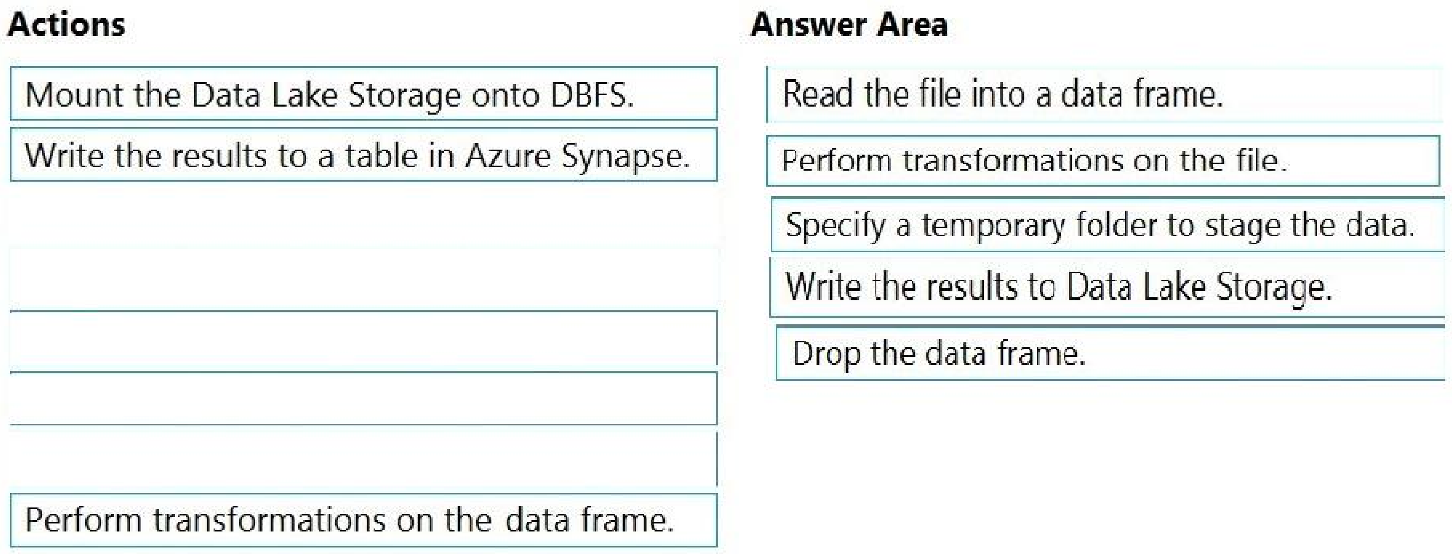
Write the results to Data Lake Storage.

Read the file into a data frame.

Drop the data frame.

Perform transformations on the data frame.

Correct Answer:



Actions

Mount the Data Lake Storage onto DBFS.

Write the results to a table in Azure Synapse.

Perform transformations on the data frame.

Answer Area

Read the file into a data frame.

Perform transformations on the file.

Specify a temporary folder to stage the data.

Write the results to Data Lake Storage.

Drop the data frame.

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Step 1: Read the file into a data frame.
You can load the json files as a data frame in Azure Databricks.

Step 2: Perform transformations on the data frame.

Step 3:Specify a temporary folder to stage the data
Specify a temporary folder to use while moving data between Azure Databricks and Azure Synapse.

Step 4: 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.

Step 5: Drop the data frame
Clean up resources. You can terminate the cluster. From the Azure Databricks workspace, select Clusters on the left. For the cluster to terminate, under Actions, point to the ellipsis (...) and select the Terminate icon.

Reference:
https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-extract-load-sql-data-warehouse

QUESTION 5
HOTSPOT

You build an Azure Data Factory pipeline to move data from an Azure Data Lake Storage Gen2 container to a database in an Azure Synapse Analytics dedicated SQL pool.

Data in the container is stored in the following folder structure.

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

The earliest folder is `/in/2021/01/01/00/00`. The latest folder is `/in/2021/01/15/01/45`.

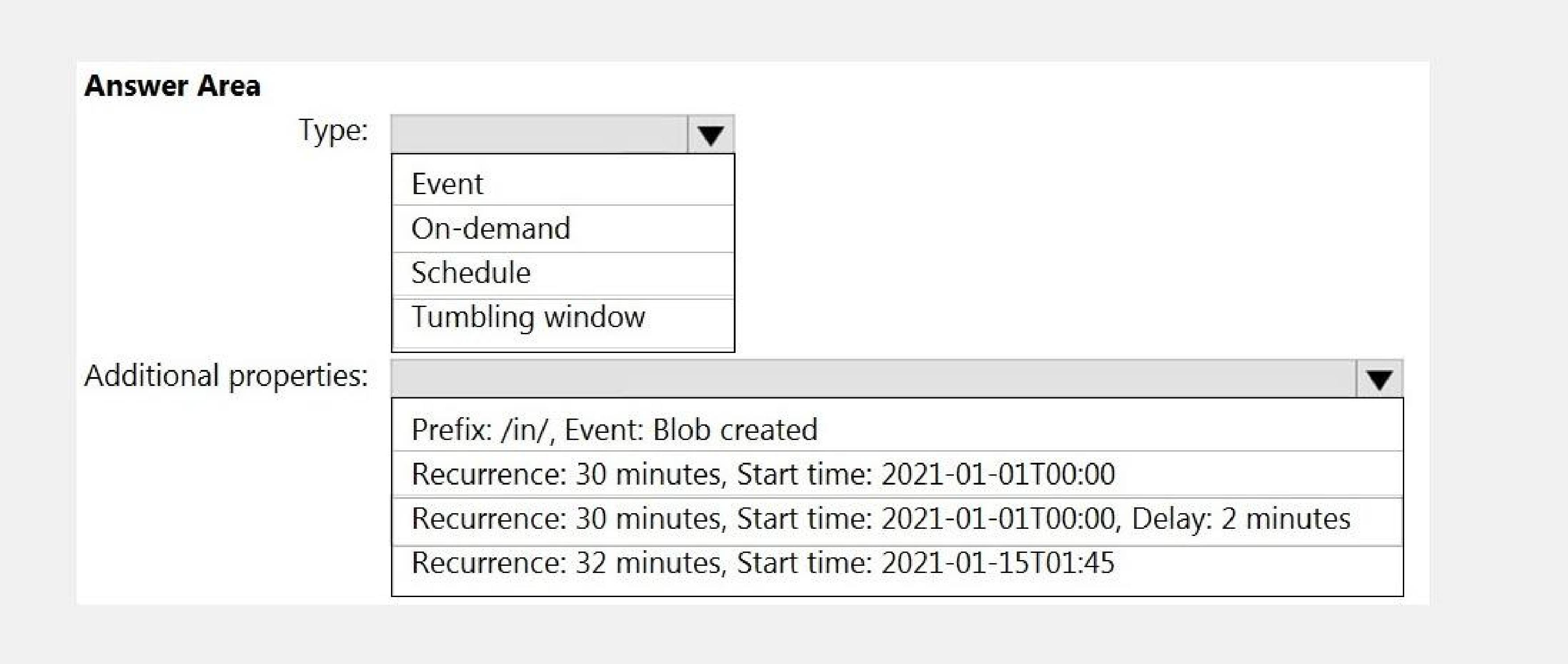
You need to configure a pipeline trigger to meet the following requirements:

- Existing data must be loaded.
- Data must be loaded every 30 minutes.
- Late-arriving data of up to two minutes must he included in the load for the time at which the data should have arrived.

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

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

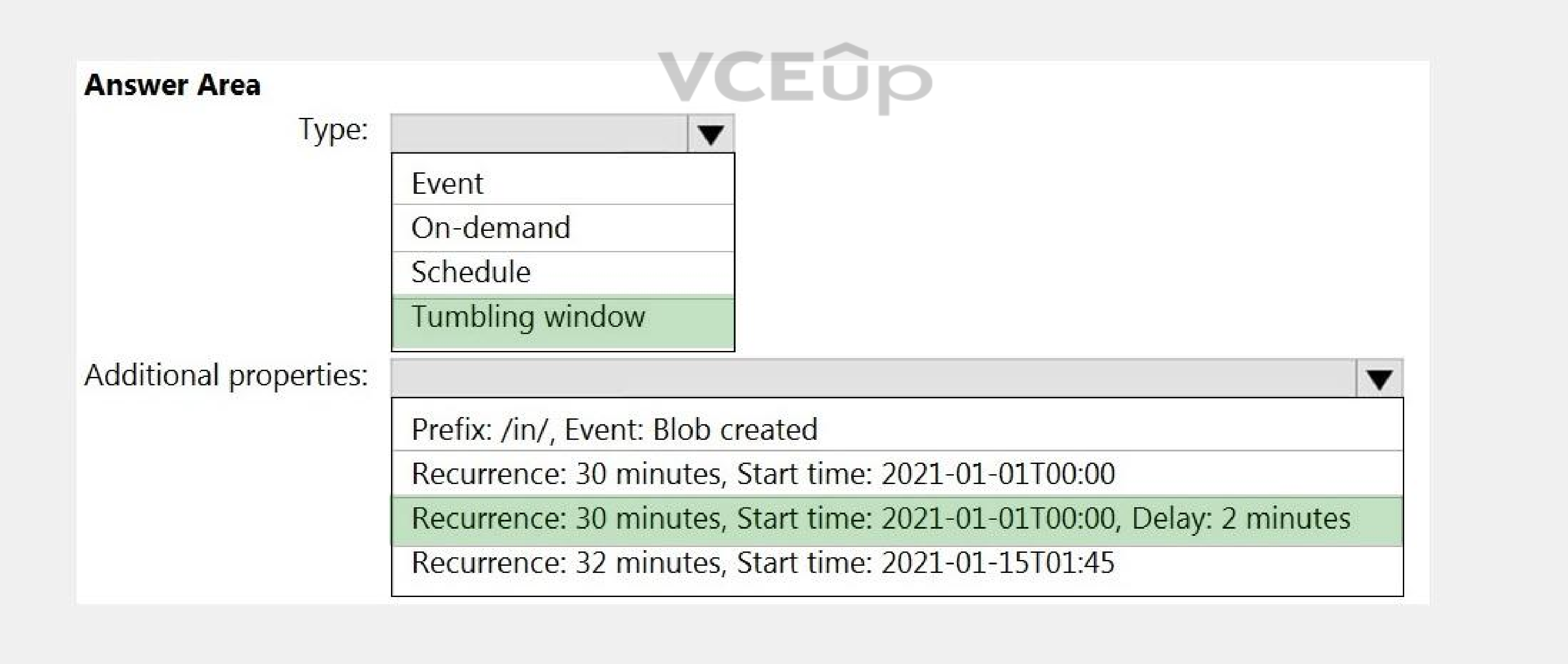
Type:

- Event
- On-demand
- Schedule
- Tumbling window

Additional properties:

- Prefix: /in/, Event: Blob created
- Recurrence: 30 minutes, Start time: 2021-01-01T00:00
- Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes
- Recurrence: 32 minutes, Start time: 2021-01-15T01:45

Correct Answer:



Answer Area

Type:

- Event
- On-demand
- Schedule
- [✓ Tumbling window]

Additional properties:

- Prefix: /in/, Event: Blob created
- Recurrence: 30 minutes, Start time: 2021-01-01T00:00
- [✓ Recurrence: 30 minutes, Start time: 2021-01-01T00:00, Delay: 2 minutes]
- Recurrence: 32 minutes, Start time: 2021-01-15T01:45

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: Tumbling window

To be able to use the Delay parameter we select Tumbling window.

Box 2:
Recurrence: 30 minutes, not 32 minutes
Delay: 2 minutes.
The amount of time to delay the start of data processing for the window. The pipeline run is started after the expected execution time plus the amount of delay. The delay defines how long the trigger waits past the due time before triggering a new run. The delay doesn't alter the window startTime.

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

QUESTION 6
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.

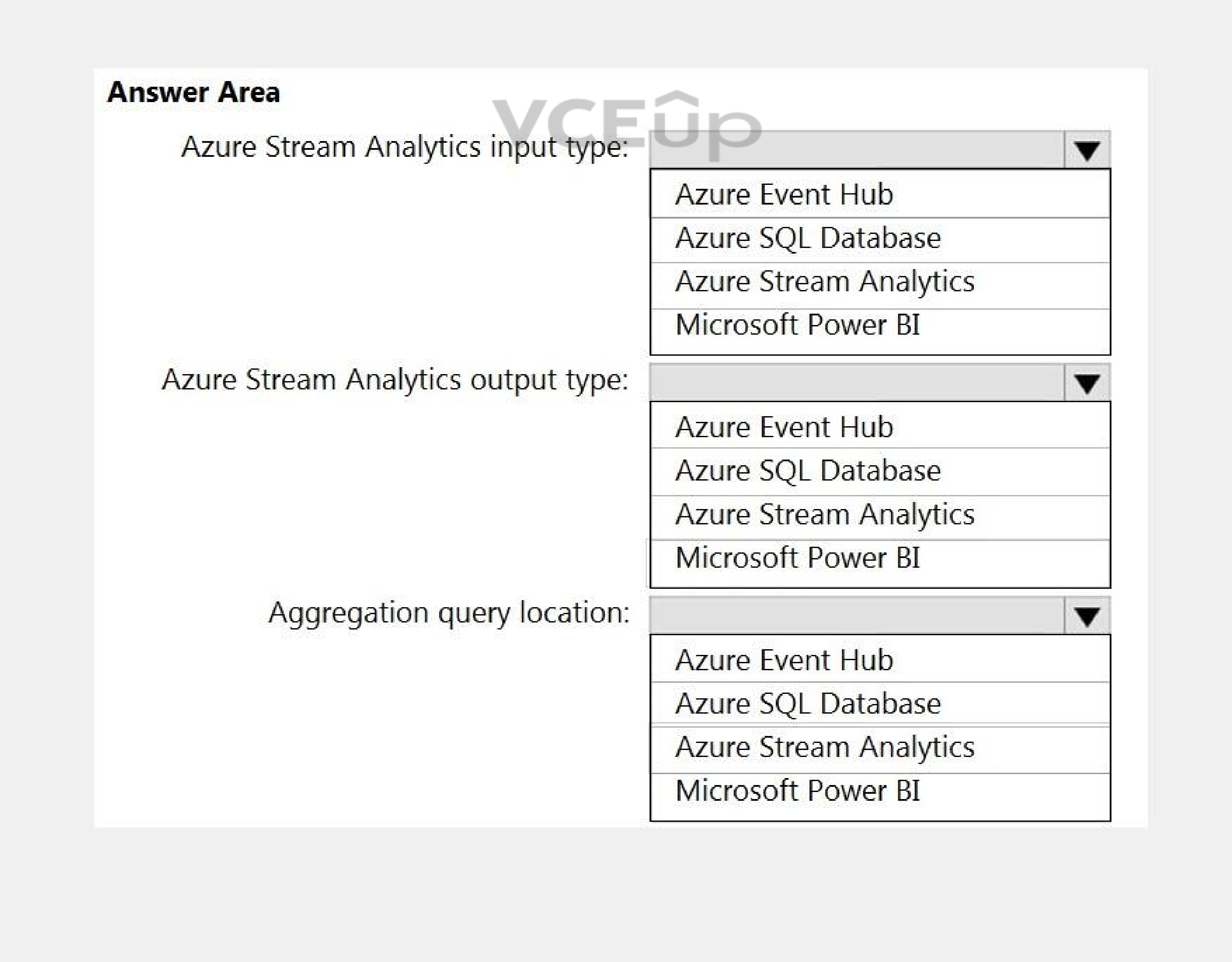
The solution will use Azure Stream Analytics and must meet the following requirements:

- 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:



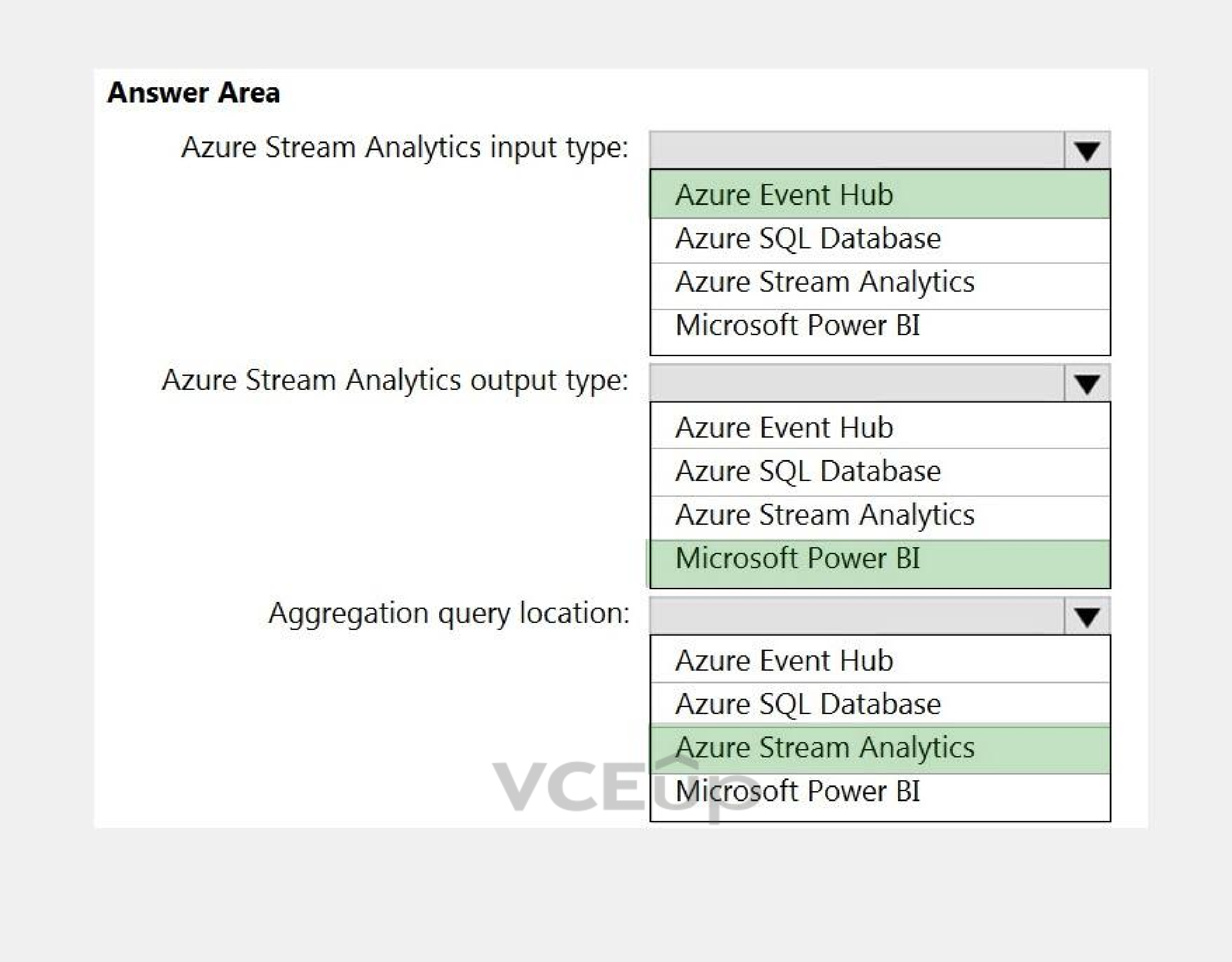
Answer Area

Azure Stream Analytics input type:

Azure Stream Analytics output type:

Aggregation query location:

Correct Answer:



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: (none)
Explanation

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

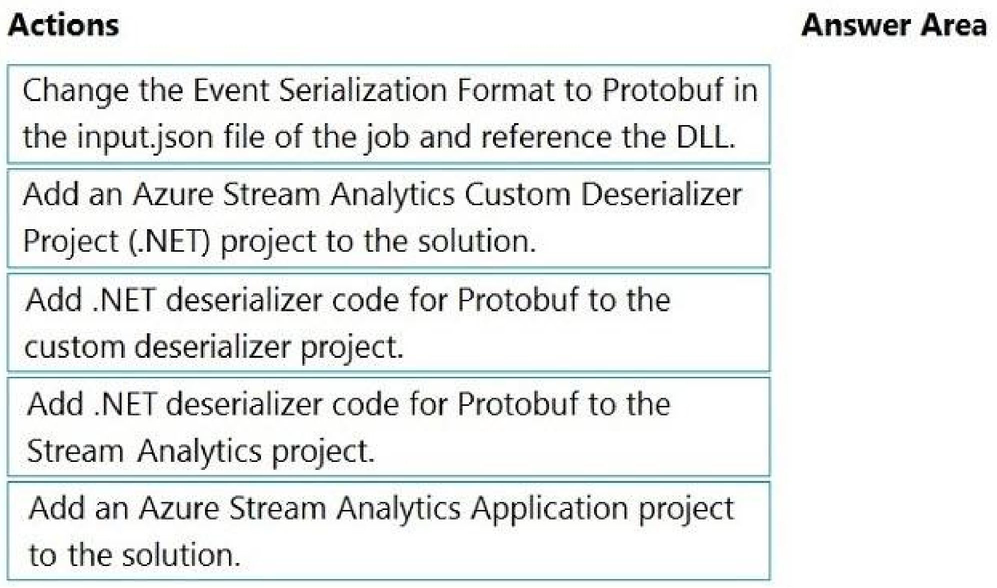
QUESTION 7
DRAG DROP

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:

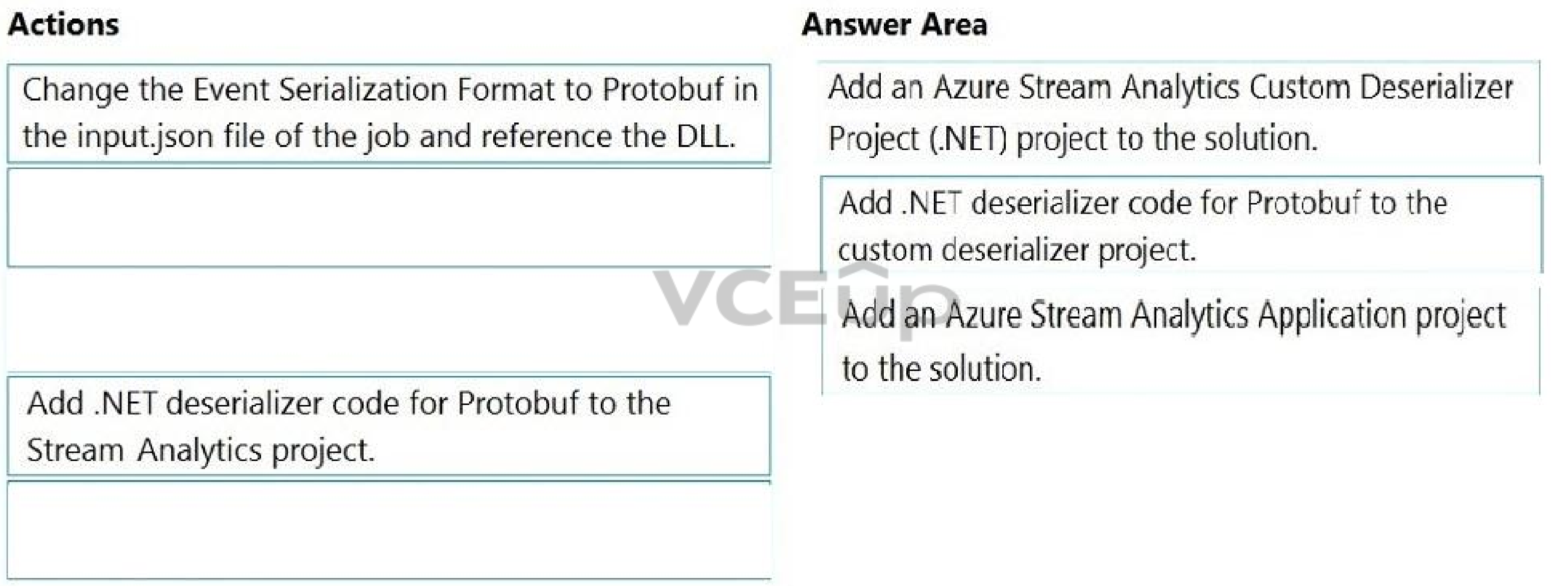


Actions

- 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.

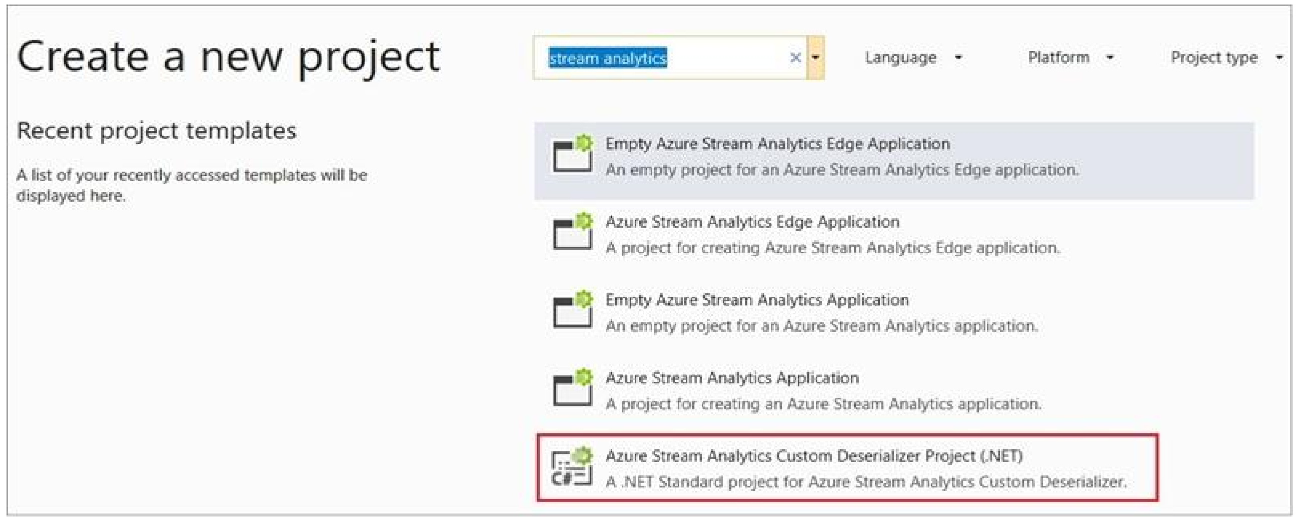
Answer Area

- 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: (none)
Explanation

Explanation/Reference:
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

Language

Platform

Project type

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 8
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:

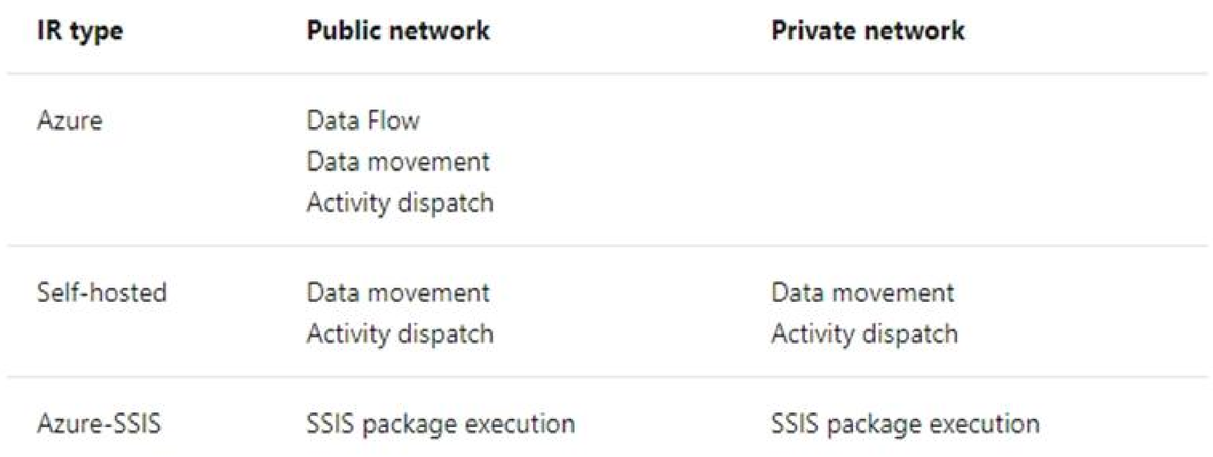
- Ensure that the data remains in the UK South region at all times.
- Minimize administrative effort.

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: (none)
Explanation

Explanation/Reference:
Explanation:



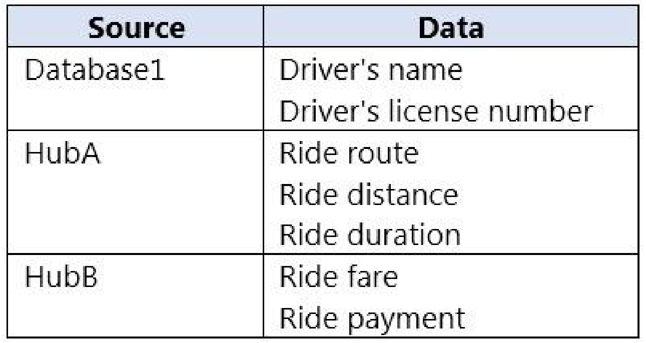
IR type	Public network	Private network
Azure	Data Flow Data movement Activity dispatch	
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

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 9
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.



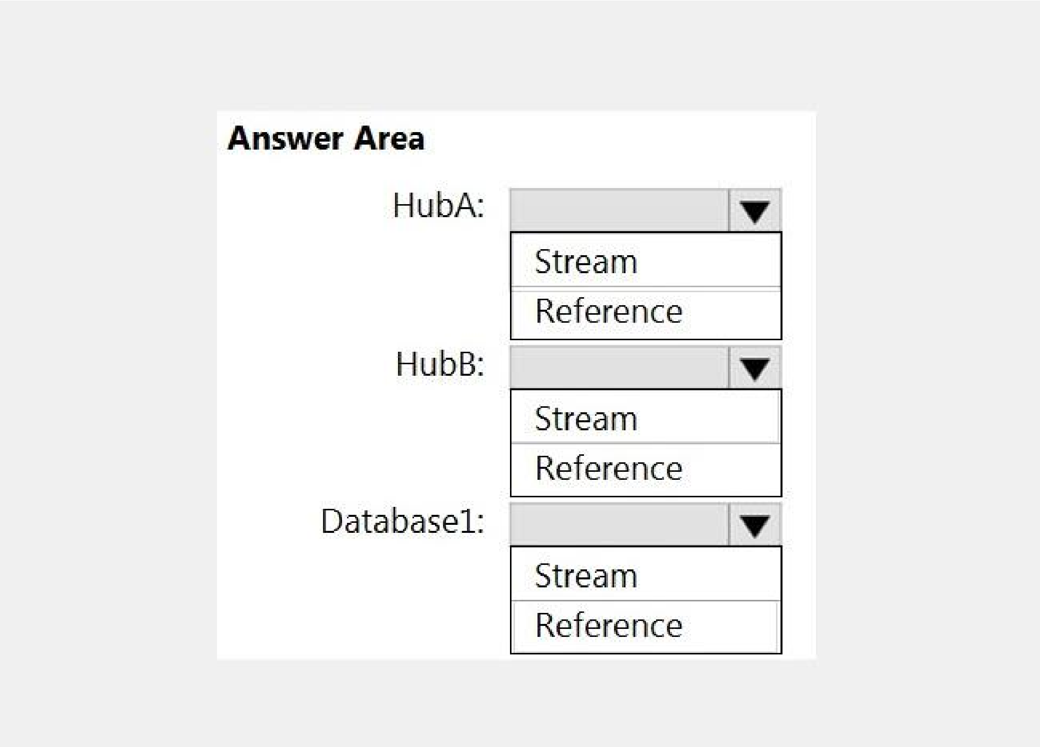
Source	Data
Database1	Driver's name Driver's license number
HubA	Ride route Ride distance Ride duration
HubB	Ride fare Ride payment

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:



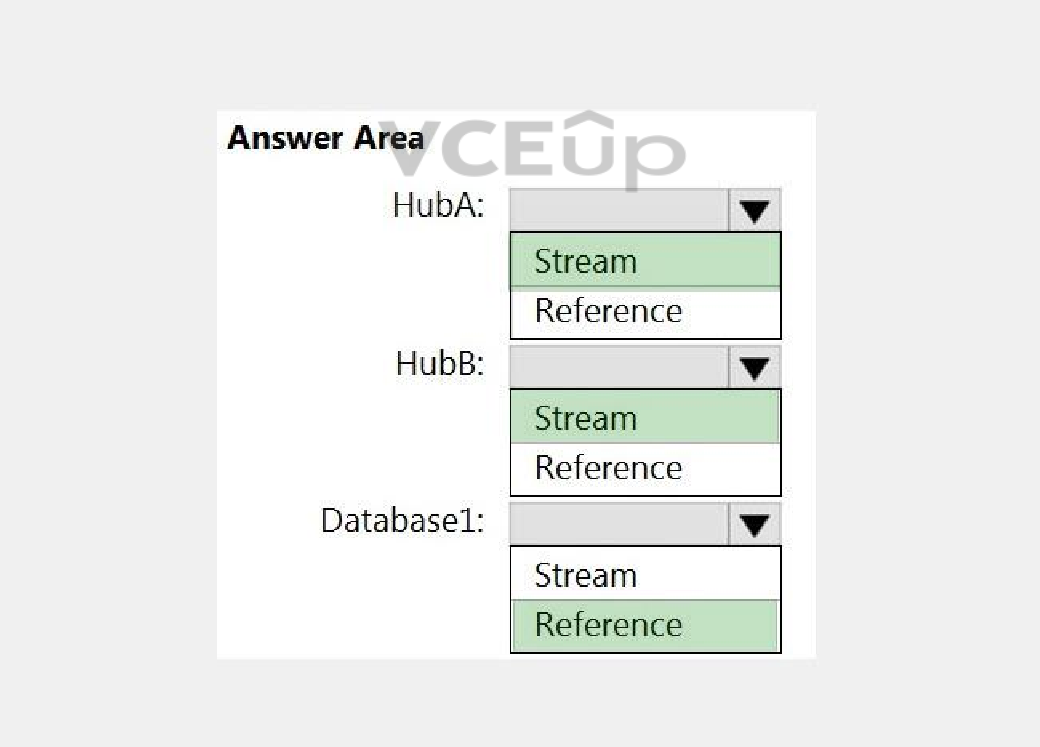
Answer Area

HubA: ▼
- Stream
- Reference

HubB: ▼
- Stream
- Reference

Database1: ▼
- Stream
- Reference

Correct Answer:



Answer Area

HubA: ▼
- [✓ Stream]
- Reference

HubB: ▼
- [✓ Stream]
- Reference

Database1: ▼
- Stream
- [✓ Reference]

Section: (none)
Explanation

Explanation/Reference:
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 10
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: (none)
Explanation

Explanation/Reference:
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 11
HOTSPOT

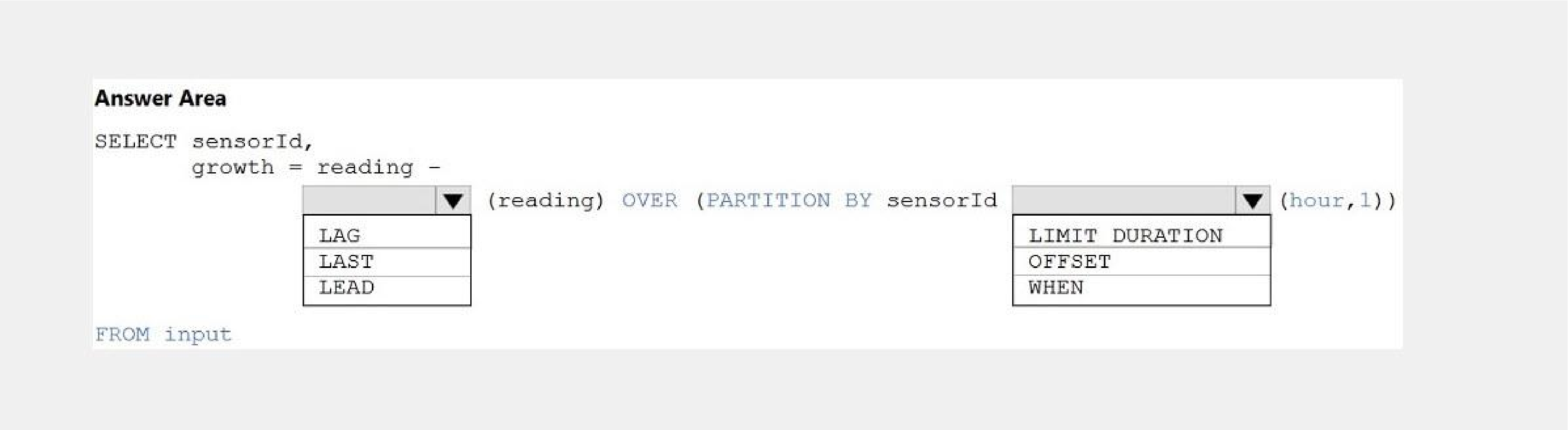
You are building an Azure Analytics query that will receive input data from Azure IoT Hub and write the results to Azure Blob storage.

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.

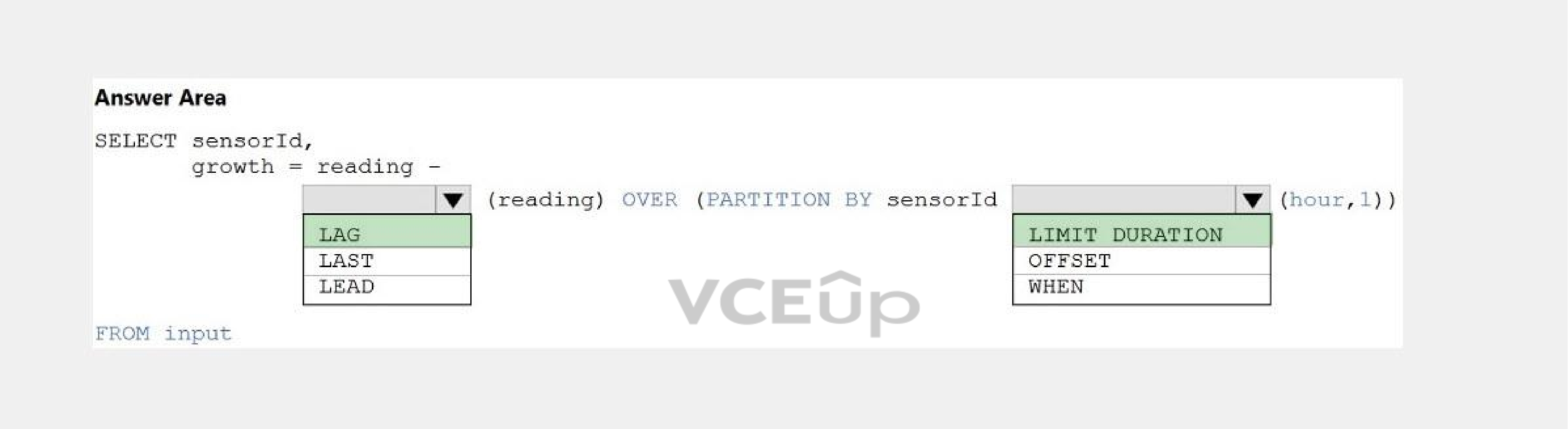
Hot Area:



Answer Area

```
SELECT sensorId,
       growth = reading -
                [          ▼] (reading) OVER (PARTITION BY sensorId [          ▼] (hour,1))
FROM input
```

Correct Answer:



Answer Area

```
SELECT sensorId,
       growth = reading -
                [          ▼] (reading) OVER (PARTITION BY sensorId [          ▼] (hour,1))
FROM input
```

Section: (none)
Explanation

Explanation/Reference:
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 12
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: (none)
Explanation

Explanation/Reference:
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 13
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: (none)
Explanation

Explanation/Reference:
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.

1. In Azure DevOps, open the project that's configured with your data factory.
2. On the left side of the page, select Pipelines, and then select Releases.
3. Select New pipeline, or, if you have existing pipelines, select New and then New release pipeline.
4. In the Stage name box, enter the name of your environment.
5. 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.
6. Select the Empty job template.

Reference:
https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment

QUESTION 14
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: (none)
Explanation

Explanation/Reference:
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 15
You are designing an Azure Stream Analytics job to process incoming events from sensors in retail environments.

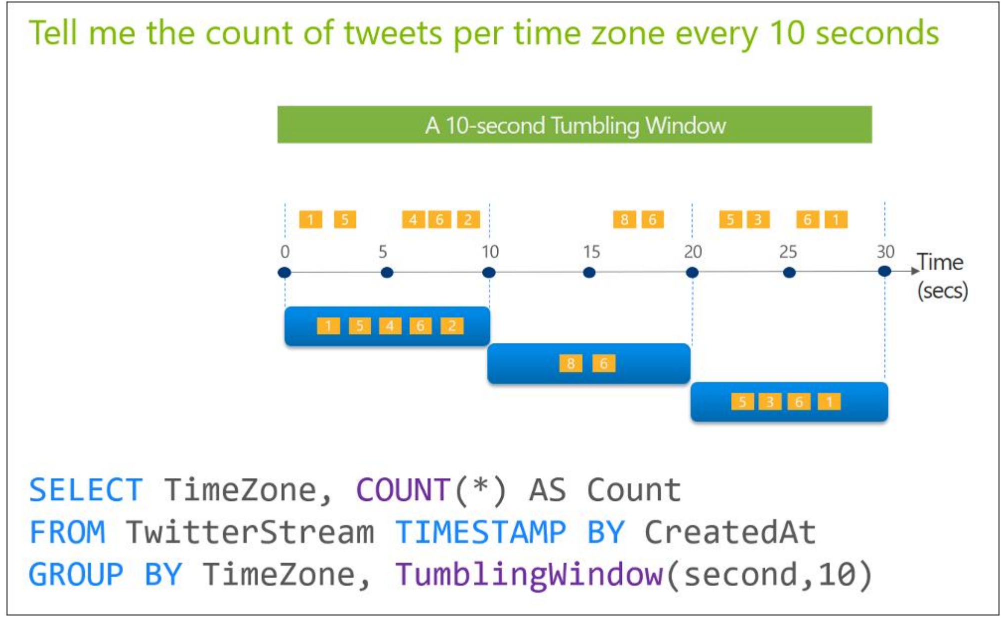
You need to process the events to produce a running average of shopper counts during the previous 15 minutes, calculated at five-minute intervals.

Which type of window should you use?

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

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
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 16
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.

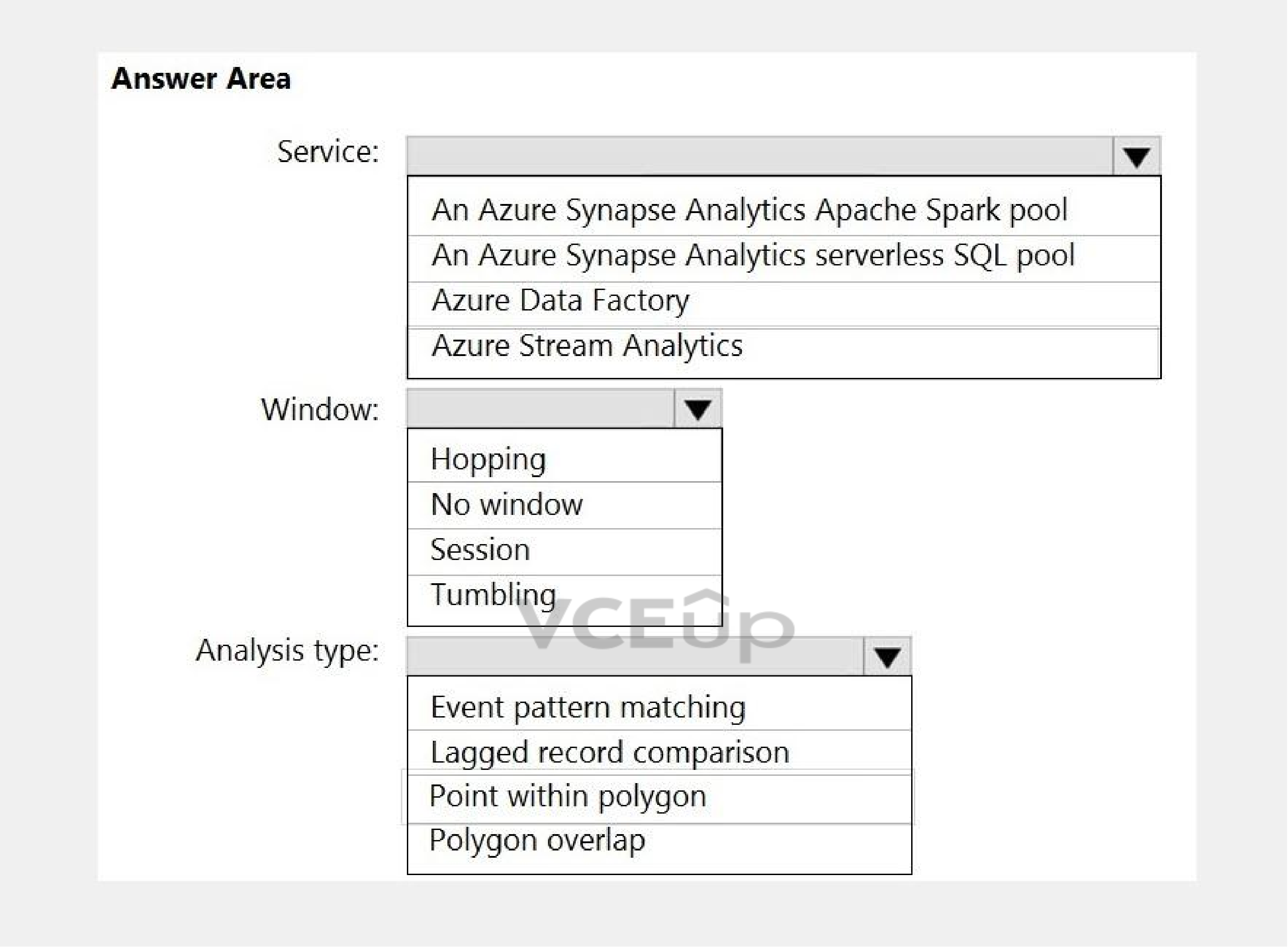
You have a CSV file in an Azure Data Lake Storage Gen2 container. The file maintains the expected geographical area in which each vehicle should be.

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:



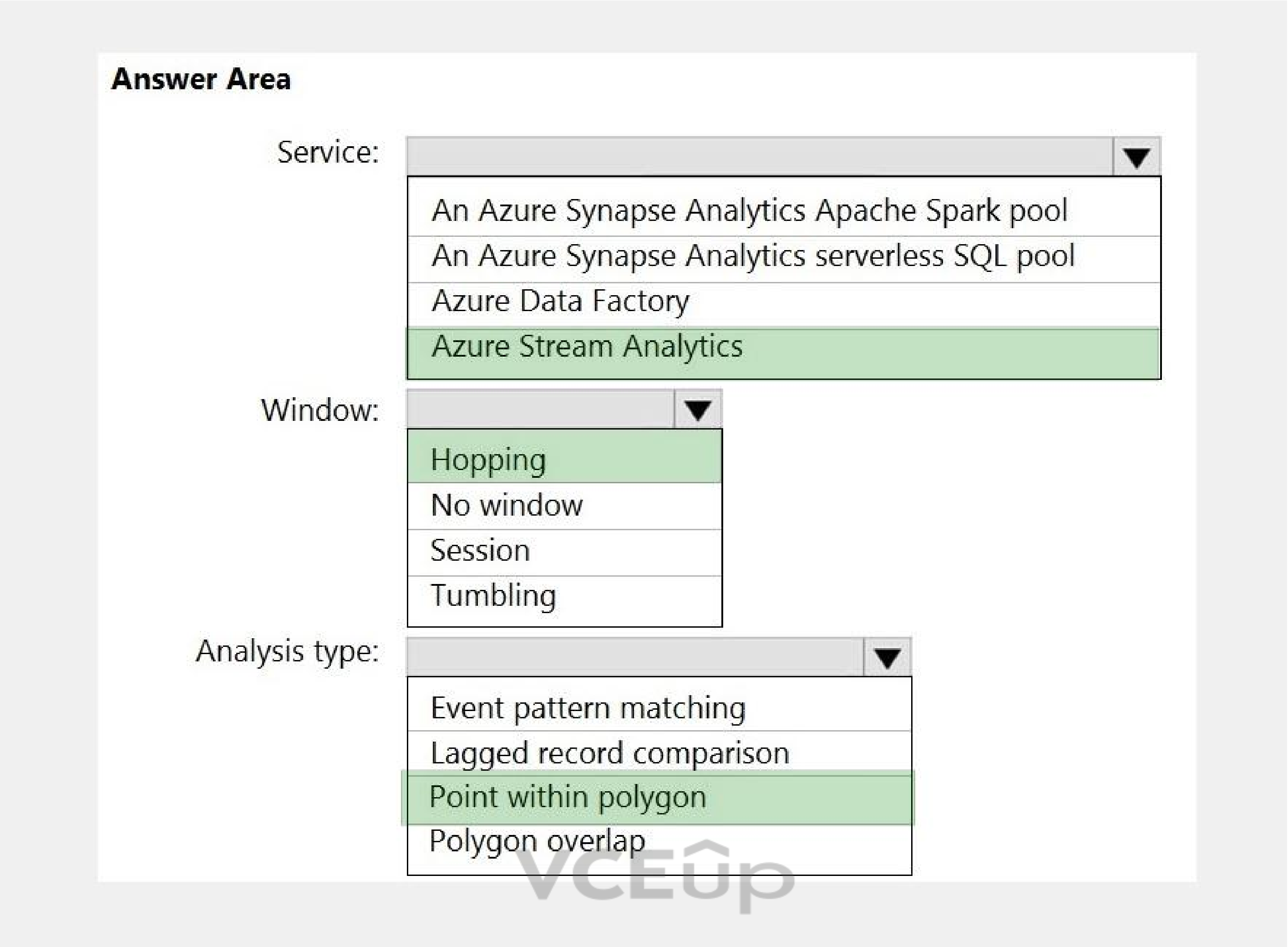
Answer Area

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

Correct Answer:



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: (none)
Explanation

Explanation/Reference:
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 17
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: (none)
Explanation

Explanation/Reference:
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 18
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

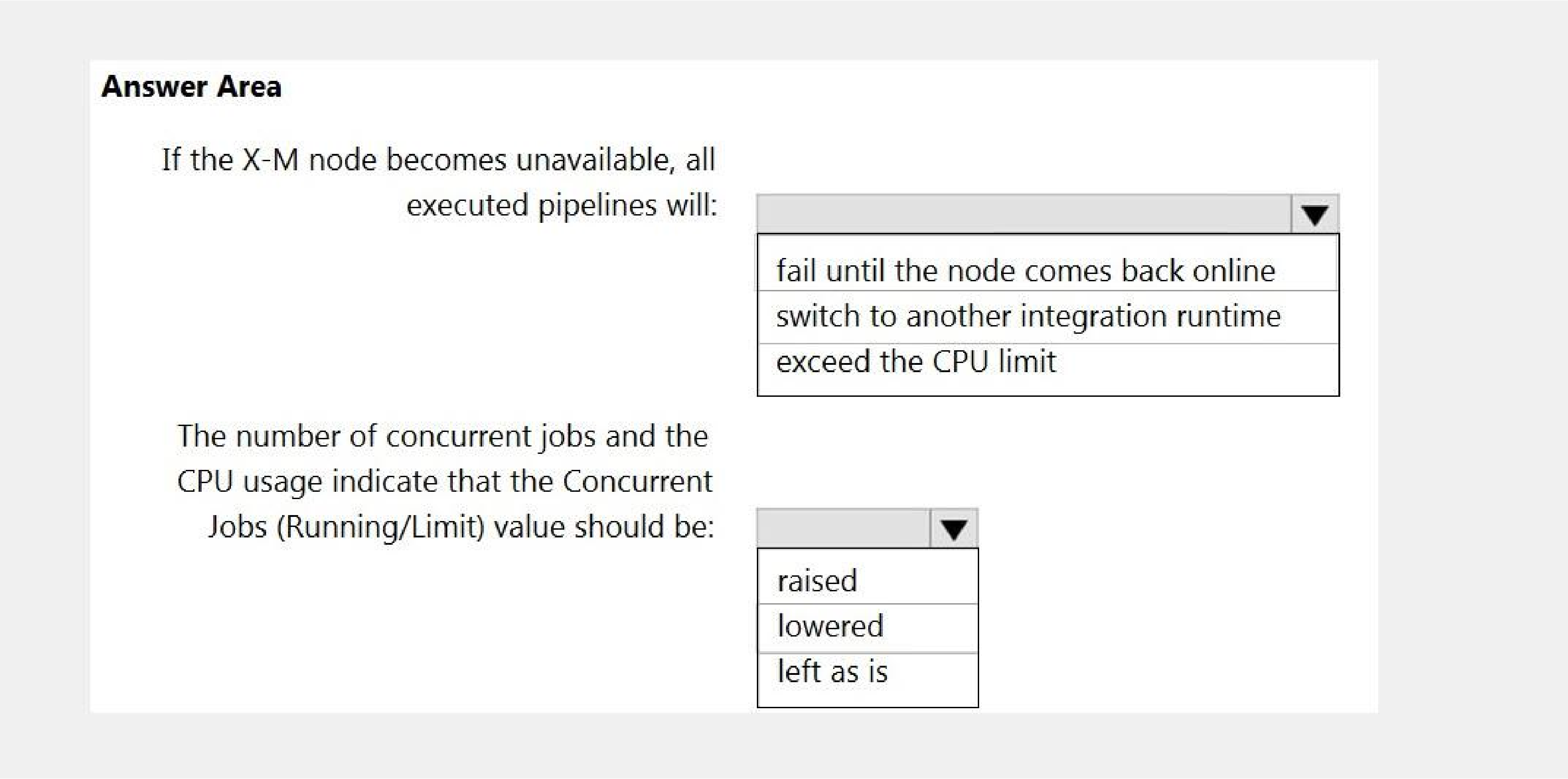
The integration runtime has the following node details:

- 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

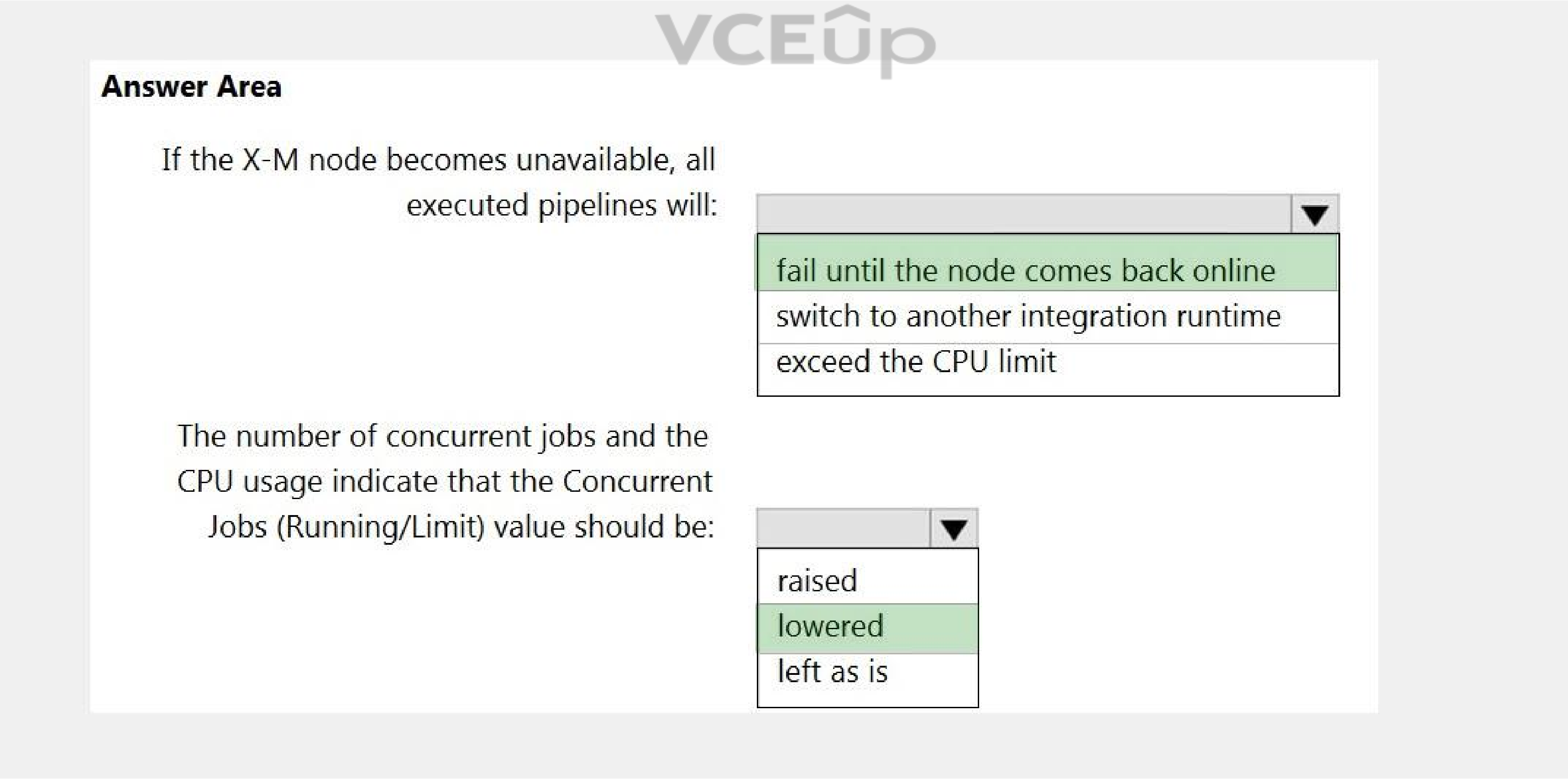
If the X-M node becomes unavailable, all executed pipelines will:

- 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

Correct Answer:



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: (none)
Explanation

Explanation/Reference:
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 19
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: (none)
Explanation

Explanation/Reference:
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 20
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.

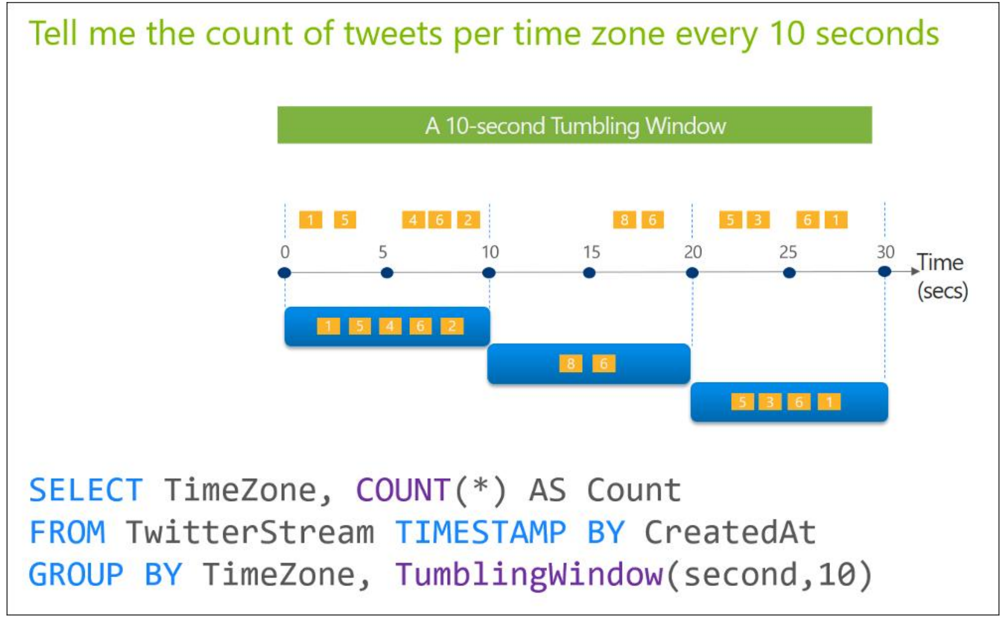
Solution: You use a tumbling window, and you set the window size to 10 seconds.

Does this meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
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 21
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: B

Section: (none)
Explanation

Explanation/Reference:
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 22
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: (none)
Explanation

Explanation/Reference:
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 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 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: (none)
Explanation

Explanation/Reference:
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 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 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: (none)
Explanation

Explanation/Reference:
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 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 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: (none)
Explanation

Explanation/Reference:
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 26
HOTSPOT

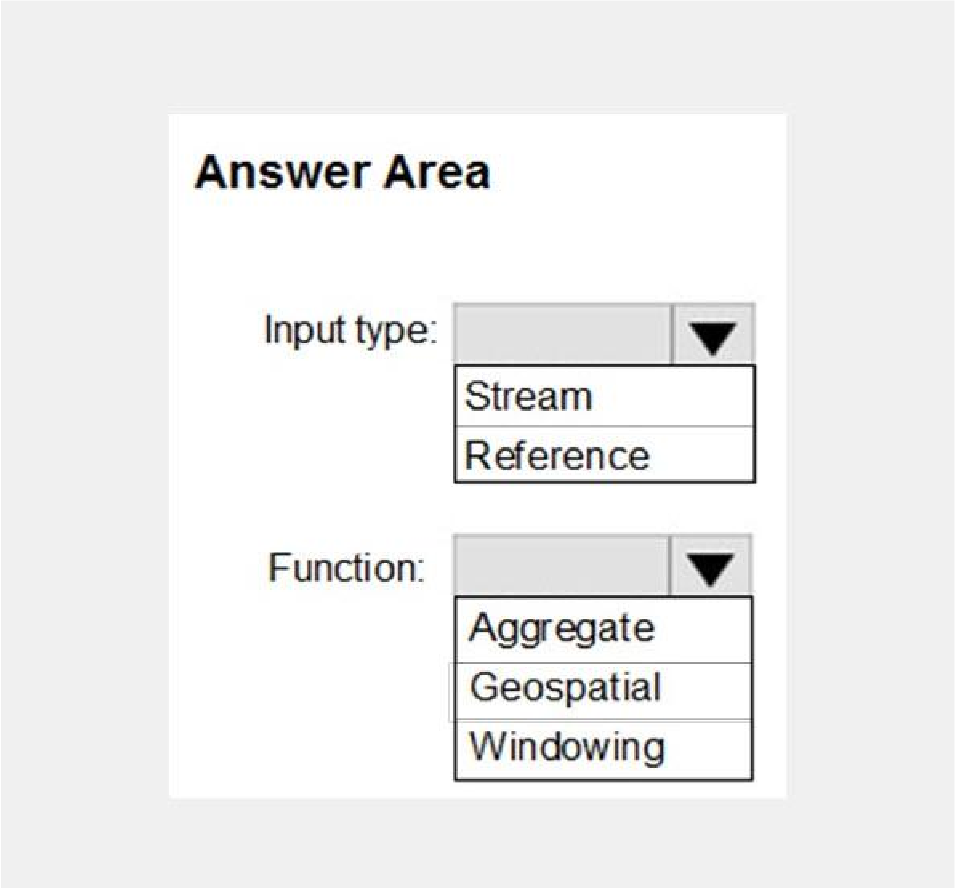
You plan to create a real-time monitoring app that alerts users when a device travels more than 200 meters away from a designated location.

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

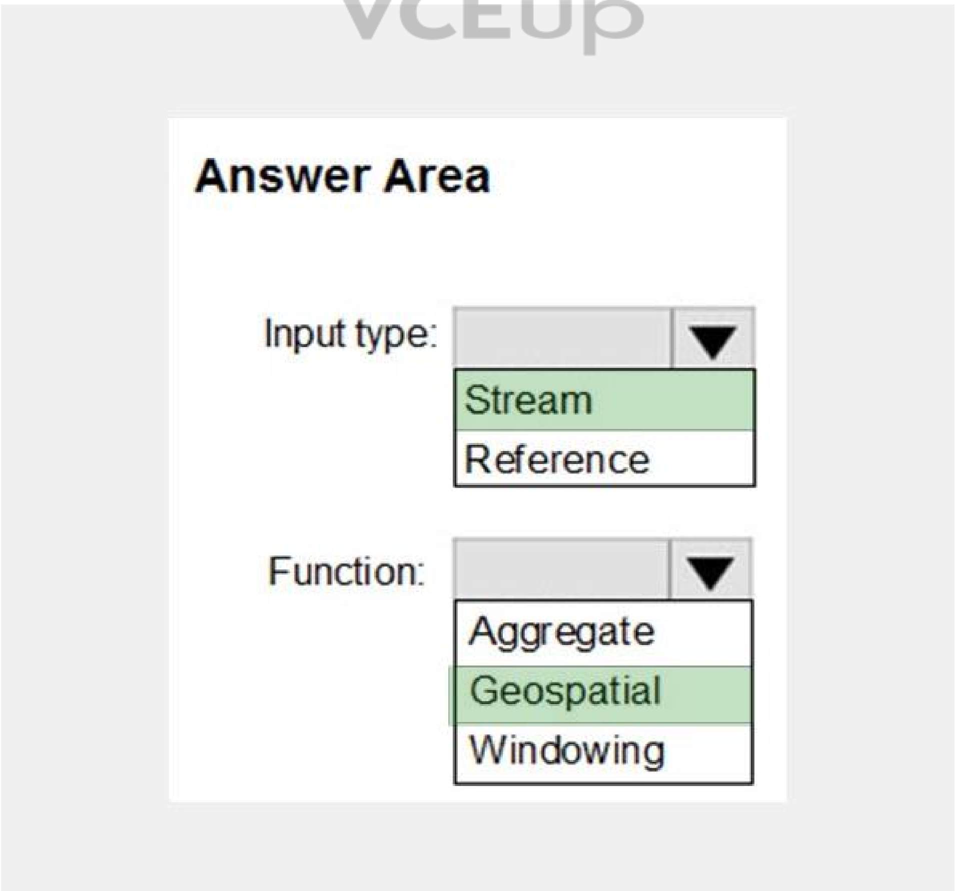


Answer Area

Input type:
- Stream
- Reference

Function:
- Aggregate
- Geospatial
- Windowing

Correct Answer:



Answer Area

Input type:
- [✓ Stream]
- Reference

Function:
- Aggregate
- [✓ Geospatial]
- Windowing

Section: (none)

Explanation

Explanation/Reference:
Explanation:

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

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

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

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

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

QUESTION 27
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: DF
Section: (none)
Explanation

Explanation/Reference:
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 28
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: (none)
Explanation

Explanation/Reference:
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 29
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: B
Section: (none)
Explanation

Explanation/Reference:
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 30
HOTSPOT

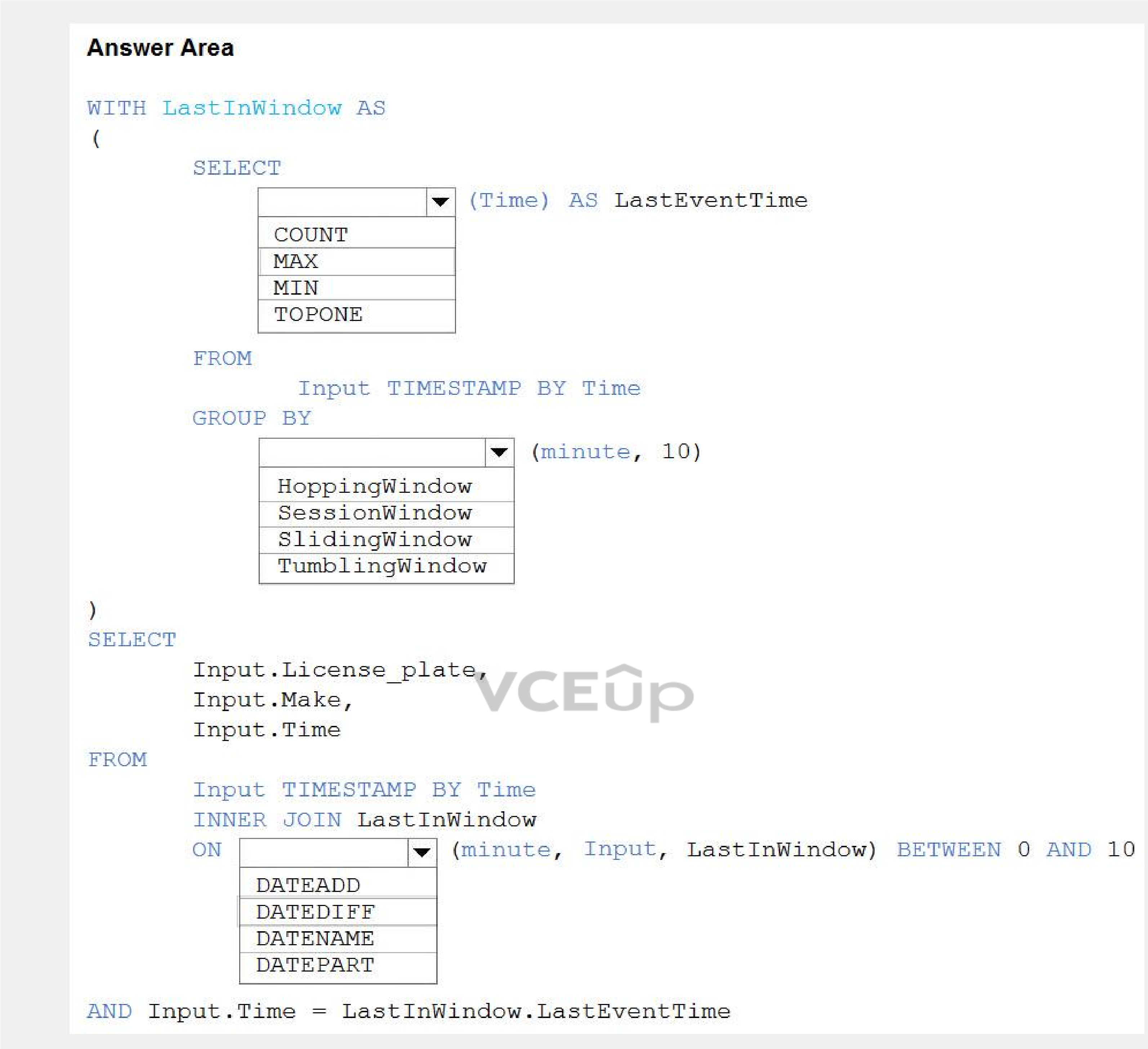
You are processing streaming data from vehicles that pass through a toll booth.

You need to use Azure Stream Analytics to return the license plate, vehicle make, and hour the last vehicle passed during each 10-minute window.

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

NOTE: Each correct selection is worth one point.

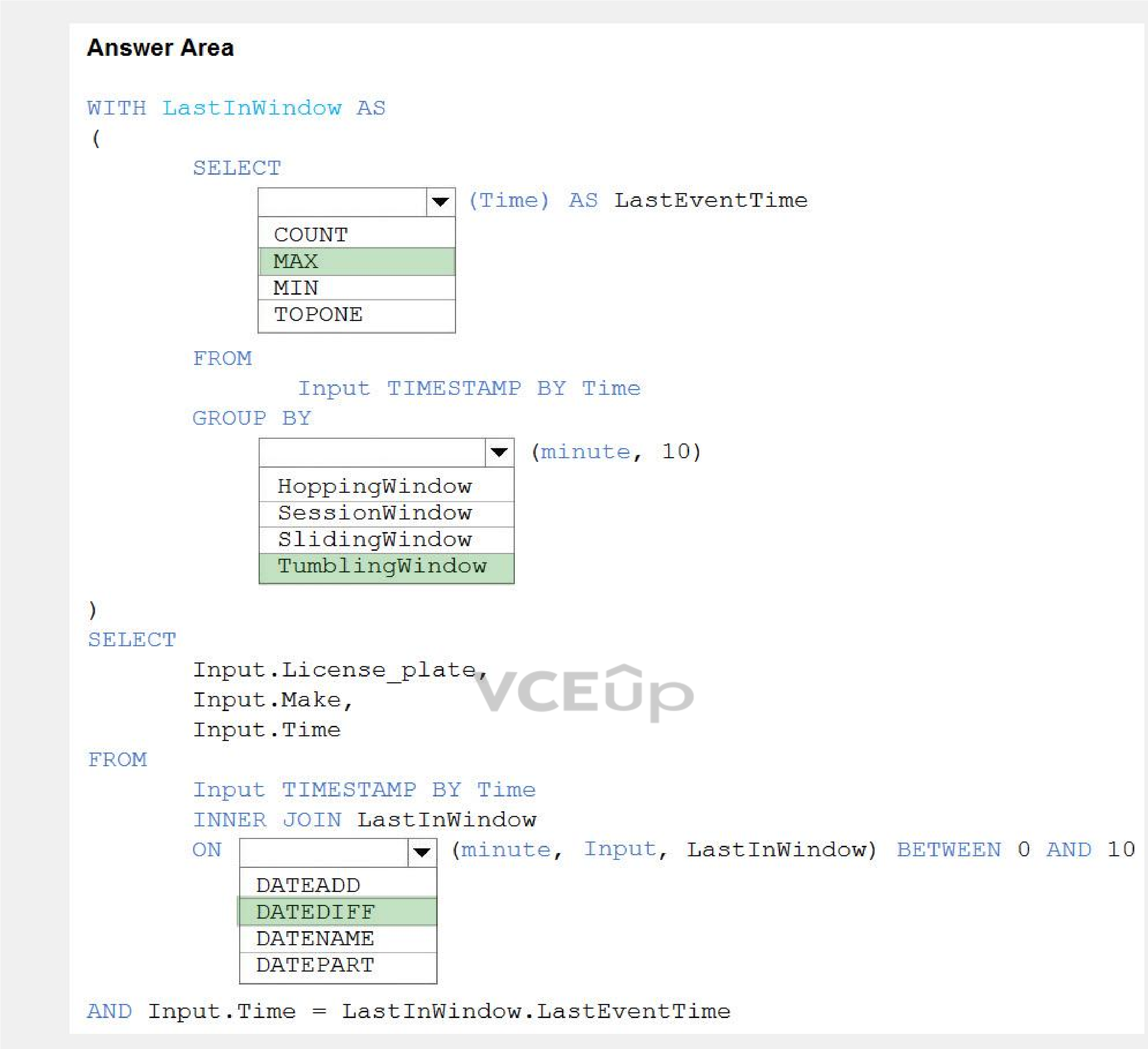
Hot Area:



Answer Area

```
WITH LastInWindow AS
(
        SELECT
                [          ▼] (Time) AS LastEventTime
                 COUNT
                 MAX
                 MIN
                 TOPONE
        FROM
                Input TIMESTAMP BY Time
        GROUP BY
                [              ▼] (minute, 10)
                 HoppingWindow
                 SessionWindow
                 SlidingWindow
                 TumblingWindow
)
SELECT
        Input.License_plate,
        Input.Make,
        Input.Time
FROM
        Input TIMESTAMP BY Time
        INNER JOIN LastInWindow
        ON [          ▼] (minute, Input, LastInWindow) BETWEEN 0 AND 10
            DATEADD
            DATEDIFF
            DATENAME
            DATEPART
AND Input.Time = LastInWindow.LastEventTime
```

Correct Answer:



Answer Area

```
WITH LastInWindow AS
(
        SELECT
             [          ▼] (Time) AS LastEventTime
              COUNT
              MAX
              MIN
              TOPONE
        FROM
                Input TIMESTAMP BY Time
        GROUP BY
             [              ▼] (minute, 10)
              HoppingWindow
              SessionWindow
              SlidingWindow
              TumblingWindow
)
SELECT
        Input.License_plate,
        Input.Make,
        Input.Time
FROM
        Input TIMESTAMP BY Time
        INNER JOIN LastInWindow
        ON [          ▼] (minute, Input, LastInWindow) BETWEEN 0 AND 10
             DATEADD
             DATEDIFF
             DATENAME
             DATEPART
AND Input.Time = LastInWindow.LastEventTime
```

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: MAX
The first step on the query finds the maximum time stamp in 10-minute windows, that is the time stamp of the last event for that window. The second step joins the results of the first query with the original stream to find the event that match the last time stamps in each window.

Query:
WITH LastInWindow AS
(
  SELECT

```
    MAX(Time) AS LastEventTime
  FROM
    Input TIMESTAMP BY Time
  GROUP BY
    TumblingWindow(minute, 10)
)

SELECT
  Input.License_plate,
  Input.Make,
  Input.Time
FROM
  Input TIMESTAMP BY Time
  INNER JOIN LastInWindow
  ON DATEDIFF(minute, Input, LastInWindow) BETWEEN 0 AND 10
  AND Input.Time = LastInWindow.LastEventTime
```

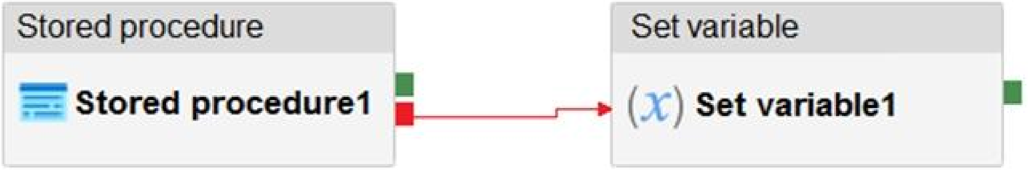
Box 2: TumblingWindow
Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.

Box 3: DATEDIFF
DATEDIFF is a date-specific function that compares and returns the time difference between two DateTime fields, for more information, refer to date functions.

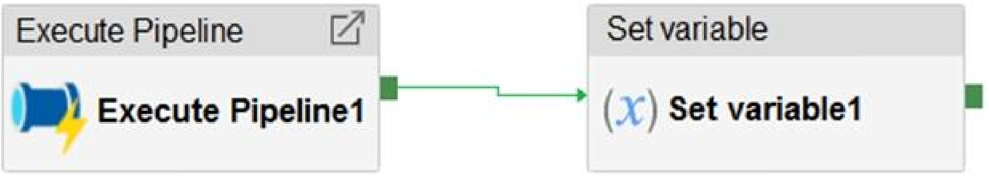
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

QUESTION 31
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?

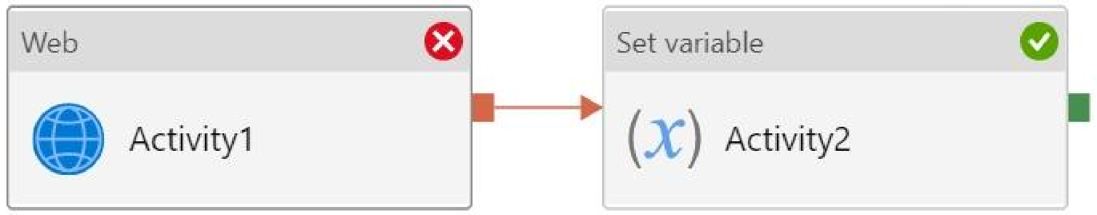
A. Pipeline1 and Pipeline2 succeeded.
B. Pipeline1 and Pipeline2 failed.
C. Pipeline1 succeeded and Pipeline2 failed.
D. Pipeline1 failed and Pipeline2 succeeded.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

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 32
HOTSPOT

A company plans to use Platform-as-a-Service (PaaS) to create the new data pipeline process. The process must meet the following requirements:

Ingest:

- Access multiple data sources.
- Provide the ability to orchestrate workflow.
- Provide the capability to run SQL Server Integration Services packages.

Store:

- Optimize storage for big data workloads.
- Provide encryption of data at rest.
- Operate with no size limits.

Prepare and Train:

- Provide a fully-managed and interactive workspace for exploration and visualization.
- Provide the ability to program in R, SQL, Python, Scala, and Java.
- Provide seamless user authentication with Azure Active Directory.

Model & Serve:

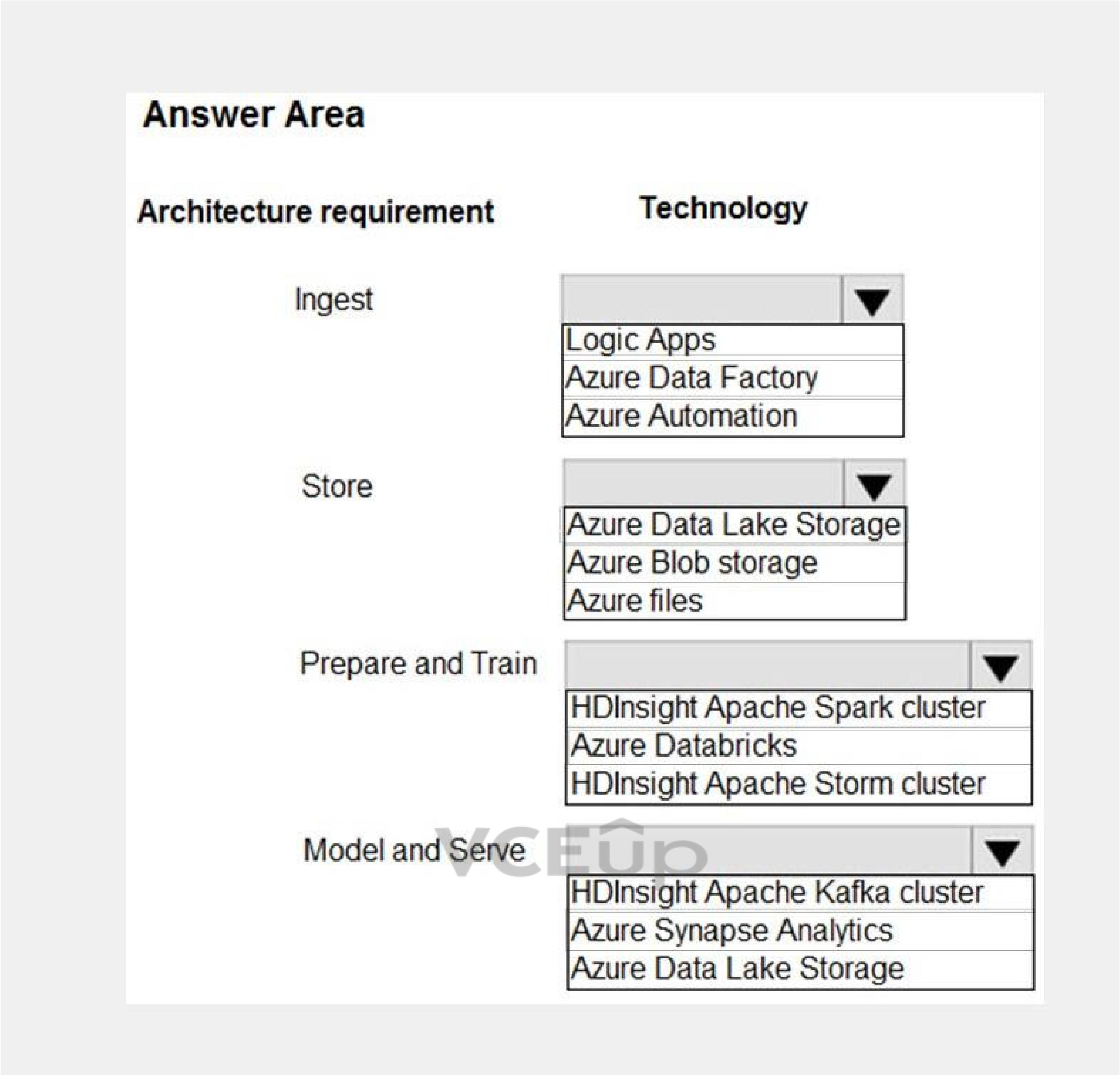
- Implement native columnar storage.
- Support for the SQL language
- Provide support for structured streaming.

You need to build the data integration pipeline.

Which technologies should you use? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

Architecture requirement | **Technology**

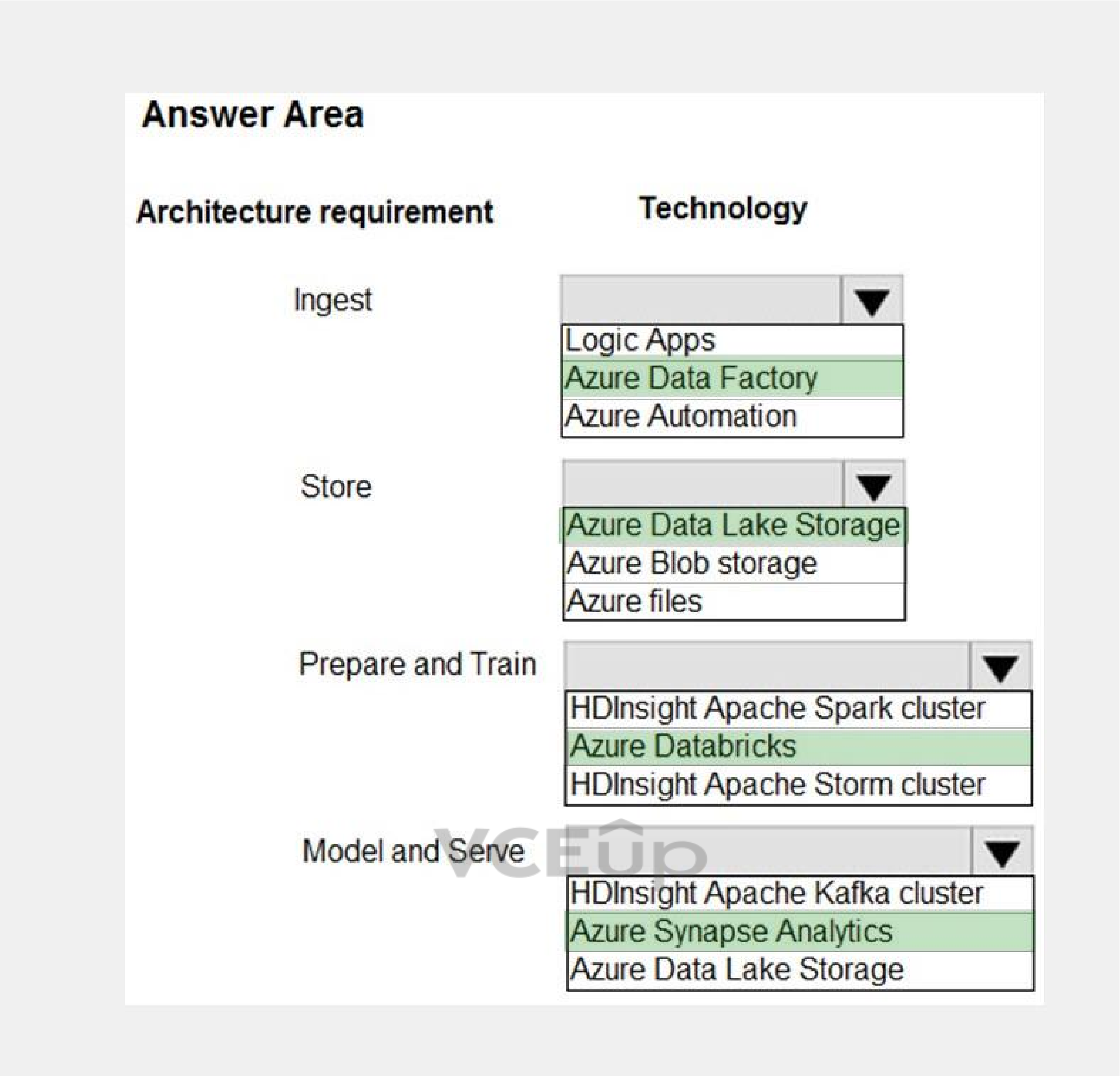
Ingest
- Logic Apps
- Azure Data Factory
- Azure Automation

Store
- Azure Data Lake Storage
- Azure Blob storage
- Azure files

Prepare and Train
- HDInsight Apache Spark cluster
- Azure Databricks
- HDInsight Apache Storm cluster

Model and Serve
- HDInsight Apache Kafka cluster
- Azure Synapse Analytics
- Azure Data Lake Storage

Correct Answer:



Answer Area

Architecture requirement	Technology
Ingest	Logic Apps [✓ Azure Data Factory] Azure Automation
Store	[✓ Azure Data Lake Storage] Azure Blob storage Azure files
Prepare and Train	HDInsight Apache Spark cluster [✓ Azure Databricks] HDInsight Apache Storm cluster
Model and Serve	HDInsight Apache Kafka cluster [✓ Azure Synapse Analytics] Azure Data Lake Storage

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Ingest: Azure Data Factory
Azure Data Factory pipelines can execute SSIS packages.
In Azure, the following services and tools will meet the core requirements for pipeline orchestration, control flow, and data movement: Azure Data Factory, Oozie on HDInsight, and SQL Server Integration Services (SSIS).

Store: Data Lake Storage
Data Lake Storage Gen1 provides unlimited storage.

Note: Data at rest includes information that resides in persistent storage on physical media, in any digital format. Microsoft Azure offers a variety of data storage solutions to meet different needs, including file, disk, blob, and table storage. Microsoft also provides encryption to protect Azure SQL Database, Azure Cosmos DB, and Azure Data Lake.

Prepare and Train: Azure Databricks
Azure Databricks provides enterprise-grade Azure security, including Azure Active Directory integration.
With Azure Databricks, you can set up your Apache Spark environment in minutes, autoscale and collaborate on shared projects in an interactive workspace. Azure Databricks supports Python, Scala, R, Java and SQL, as well as data science frameworks and libraries including TensorFlow, PyTorch and scikit-learn.

Model and Serve: Azure Synapse Analytics

Azure Synapse Analytics/ SQL Data Warehouse stores data into relational tables with columnar storage.
Azure SQL Data Warehouse connector now offers efficient and scalable structured streaming write support for SQL Data Warehouse. Access SQL Data Warehouse from Azure Databricks using the SQL Data Warehouse connector.

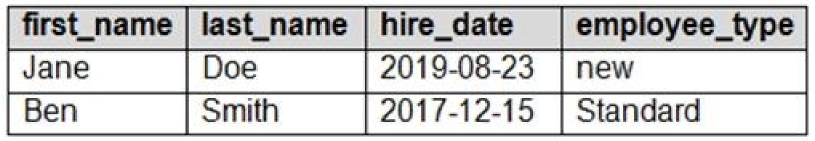
Note: Note: As of November 2019, Azure SQL Data Warehouse is now Azure Synapse Analytics.

Reference:
https://docs.microsoft.com/bs-latn-ba/azure/architecture/data-guide/technology-choices/pipeline-orchestration-data-movement

https://docs.microsoft.com/en-us/azure/azure-databricks/what-is-azure-databricks

QUESTION 33
DRAG DROP

You have the following table named Employees.



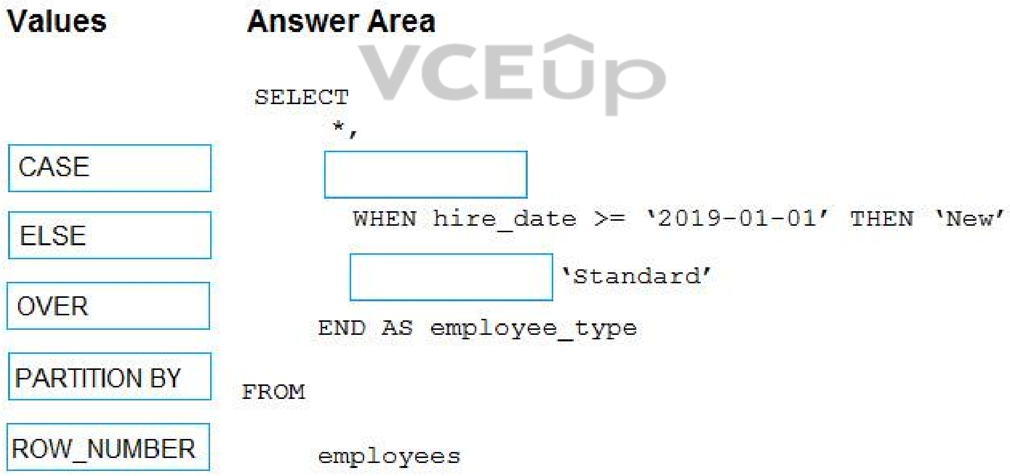
first_name	last_name	hire_date	employee_type
Jane	Doe	2019-08-23	new
Ben	Smith	2017-12-15	Standard

You need to calculate the employee_type value based on the hire_date value.

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

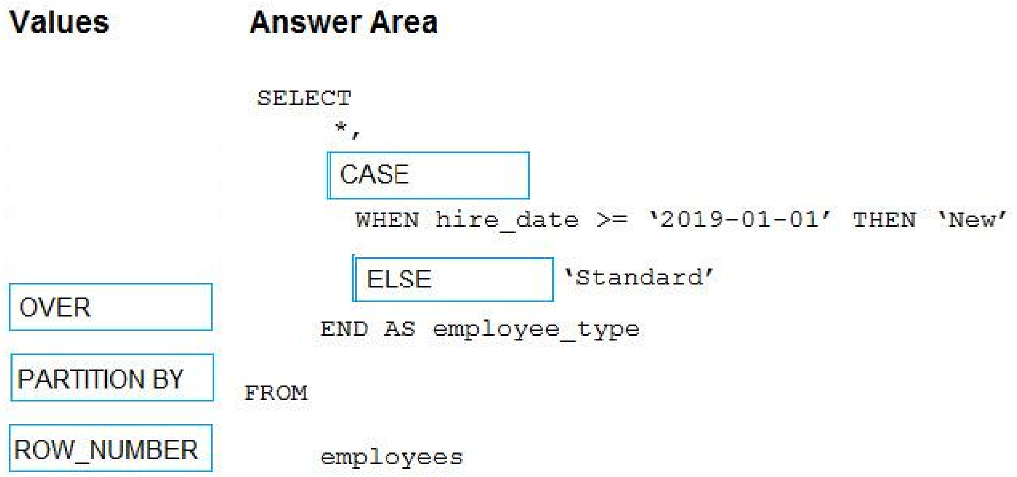
NOTE: Each correct selection is worth one point.

Select and Place:



```
SELECT
    *,
    [          ]
      WHEN hire_date >= '2019-01-01' THEN 'New'
      [          ] 'Standard'
    END AS employee_type
FROM
    employees
```

Correct Answer:



Values

OVER

PARTITION BY

ROW_NUMBER

Answer Area

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

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: CASE
CASE evaluates a list of conditions and returns one of multiple possible result expressions.

CASE can be used in any statement or clause that allows a valid expression. For example, you can use CASE in statements such as SELECT, UPDATE, DELETE and SET, and in clauses such as select_list, IN, WHERE, ORDER BY, and HAVING.

Syntax: Simple CASE expression:

```
CASE input_expression
    WHEN when_expression THEN result_expression [ ...n ]
    [ ELSE else_result_expression ]
END
```

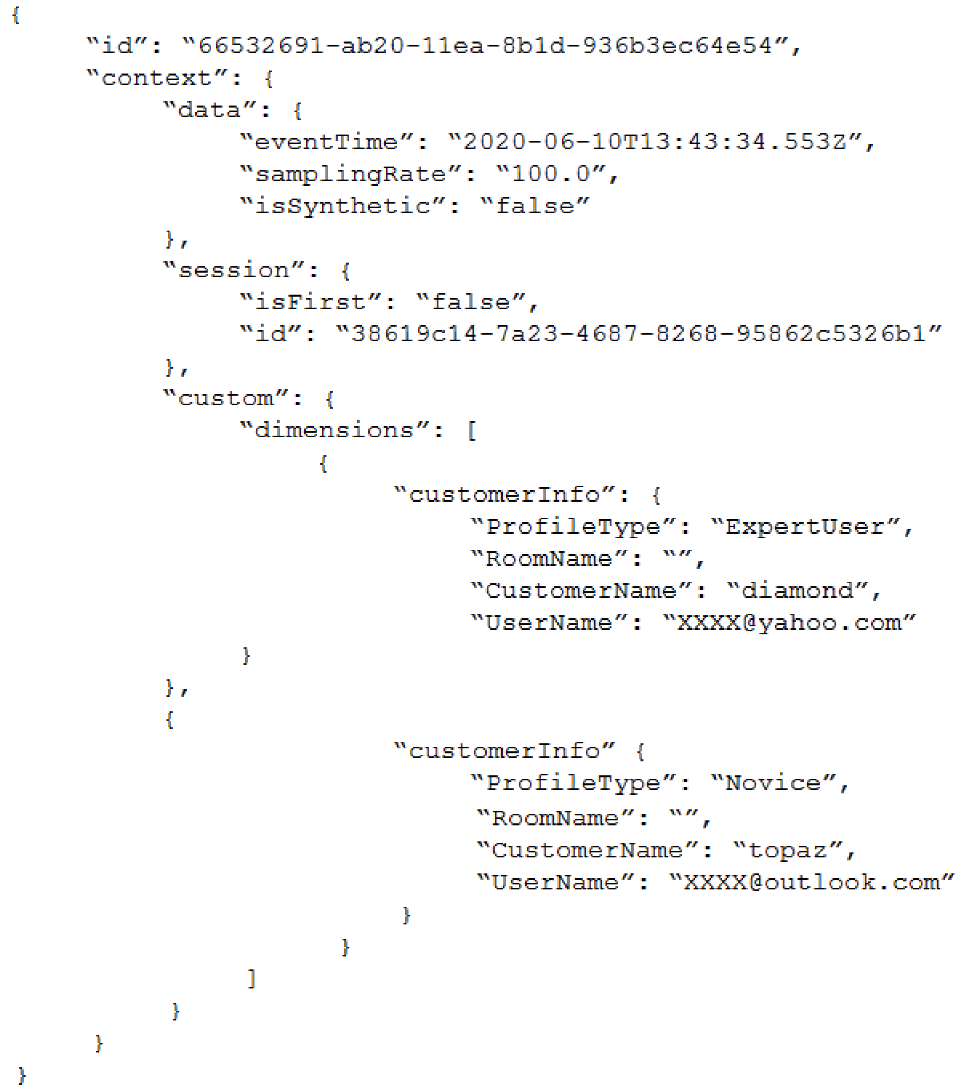
Box 2: ELSE

Reference:
https://docs.microsoft.com/en-us/sql/t-sql/language-elements/case-transact-sql

QUESTION 34
DRAG DROP

You have an Azure Synapse Analytics workspace named WS1.

You have an Azure Data Lake Storage Gen2 container that contains JSON-formatted files in the following format.



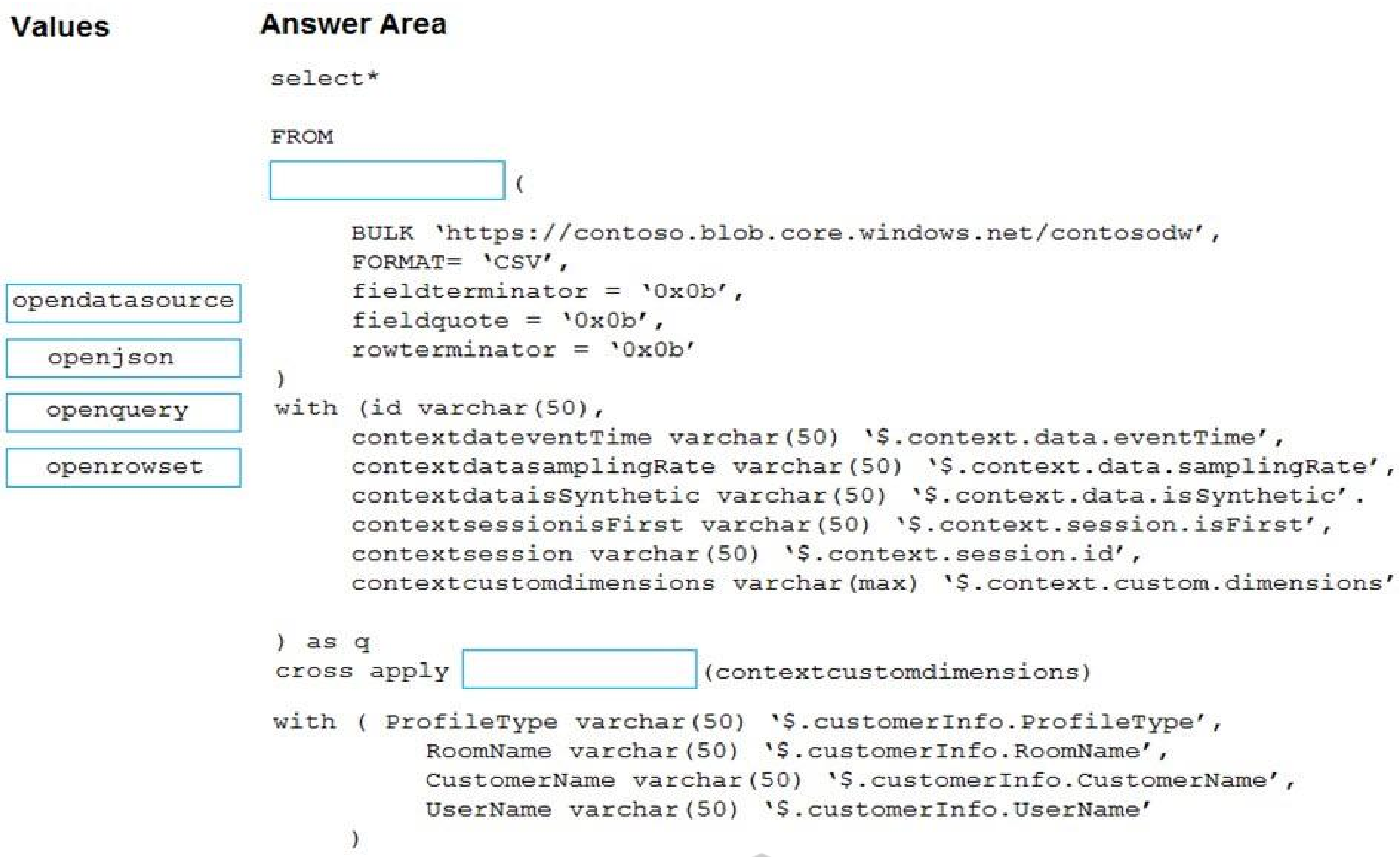
```
{
     “id”: “66532691-ab20-11ea-8b1d-936b3ec64e54”,
     “context”: {
          “data”: {
               “eventTime”: “2020-06-10T13:43:34.553Z”,
               “samplingRate”: “100.0”,
               “isSynthetic”: “false”
          },
          “session”: {
               “isFirst”: “false”,
               “id”: “38619c14-7a23-4687-8268-95862c5326b1”
          },
          “custom”: {
               “dimensions”: [
                    {
                         “customerInfo”: {
                              “ProfileType”: “ExpertUser”,
                              “RoomName”: “”,
                              “CustomerName”: “diamond”,
                              “UserName”: “XXXX@yahoo.com”
                    }
          },
          {
                         “customerInfo” {
                              “ProfileType”: “Novice”,
                              “RoomName”: “”,
                              “CustomerName”: “topaz”,
                              “UserName”: “XXXX@outlook.com”
                         }
                    }
               ]
          }
     }
}
```

You need to use the serverless SQL pool in WS1 to read the files.

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:



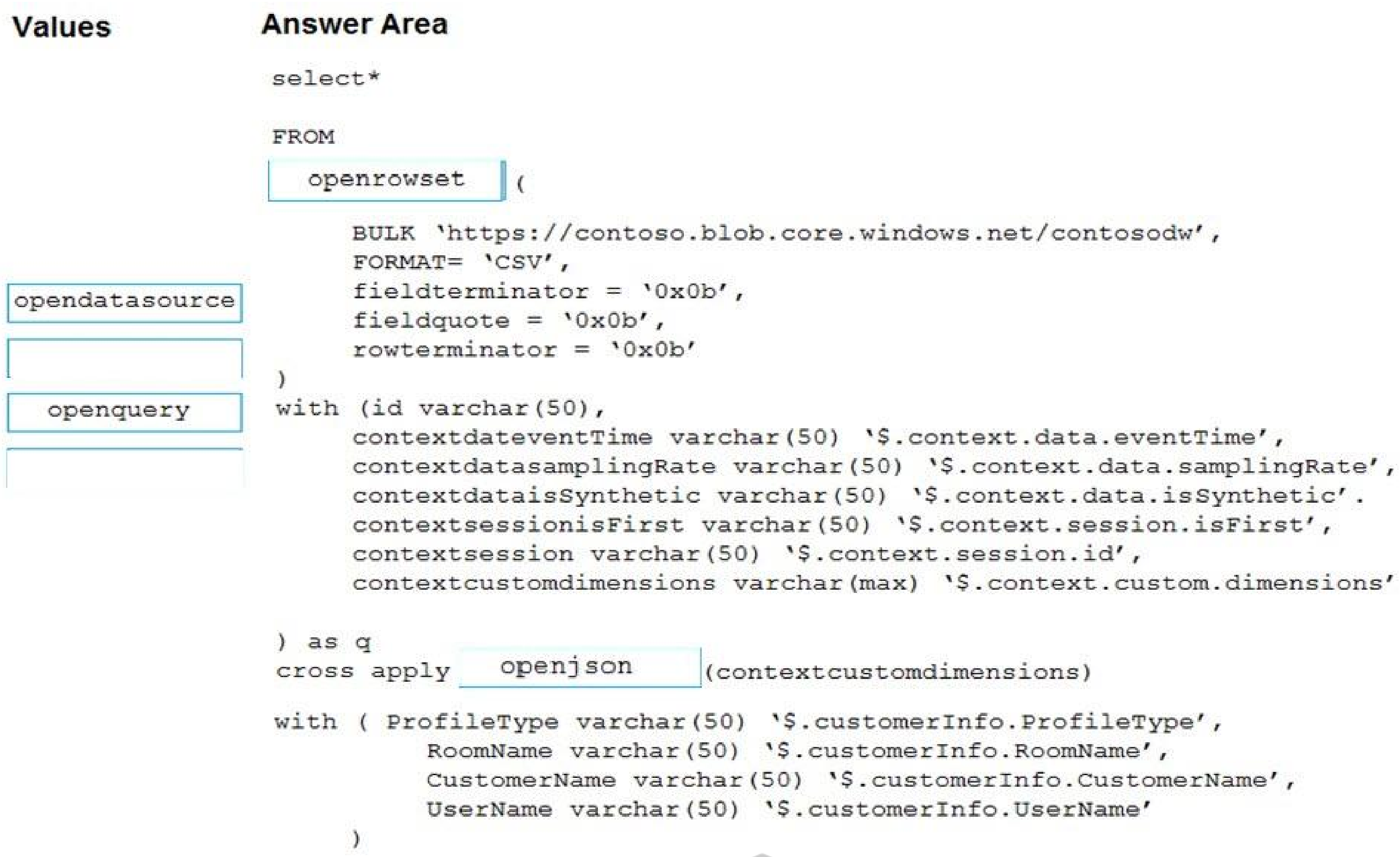
Values

- opendatasource
- openjson
- openquery
- openrowset

Answer Area

```
select*

FROM
                  (

     BULK ‘https://contoso.blob.core.windows.net/contosodw’,
     FORMAT= ‘CSV’,
     fieldterminator = ‘0x0b’,
     fieldquote = ‘0x0b’,
     rowterminator = ‘0x0b’
)
with (id varchar(50),
     contextdateventTime varchar(50) ‘$.context.data.eventTime’,
     contextdatasamplingRate varchar(50) ‘$.context.data.samplingRate’,
     contextdataisSynthetic varchar(50) ‘$.context.data.isSynthetic’.
     contextsessionisFirst varchar(50) ‘$.context.session.isFirst’,
     contextsession varchar(50) ‘$.context.session.id’,
     contextcustomdimensions varchar(max) ‘$.context.custom.dimensions’

) as q
cross apply                  (contextcustomdimensions)

with ( ProfileType varchar(50) ‘$.customerInfo.ProfileType’,
          RoomName varchar(50) ‘$.customerInfo.RoomName’,
          CustomerName varchar(50) ‘$.customerInfo.CustomerName’,
          UserName varchar(50) ‘$.customerInfo.UserName’
     )
```

Correct Answer:



```
select*

FROM
openrowset (

     BULK 'https://contoso.blob.core.windows.net/contosodw',
     FORMAT= 'CSV',
     fieldterminator = '0x0b',
     fieldquote = '0x0b',
     rowterminator = '0x0b'
)
with (id varchar(50),
     contextdateventTime varchar(50) '$.context.data.eventTime',
     contextdatasamplingRate varchar(50) '$.context.data.samplingRate',
     contextdataisSynthetic varchar(50) '$.context.data.isSynthetic'.
     contextsessionisFirst varchar(50) '$.context.session.isFirst',
     contextsession varchar(50) '$.context.session.id',
     contextcustomdimensions varchar(max) '$.context.custom.dimensions'

) as q
cross apply openjson (contextcustomdimensions)

with ( ProfileType varchar(50) '$.customerInfo.ProfileType',
          RoomName varchar(50) '$.customerInfo.RoomName',
          CustomerName varchar(50) '$.customerInfo.CustomerName',
          UserName varchar(50) '$.customerInfo.UserName'
     )
```

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: openrowset
The easiest way to see to the content of your CSV file is to provide file URL to OPENROWSET function, specify csv FORMAT.

Example:

```
SELECT *
FROM OPENROWSET(
      BULK 'csv/population/population.csv',
      DATA_SOURCE = 'SqlOnDemandDemo',
      FORMAT = 'CSV', PARSER_VERSION = '2.0',
      FIELDTERMINATOR =',',
      ROWTERMINATOR = '\n'
```

Box 2: openjson
You can access your JSON files from the Azure File Storage share by using the mapped drive, as shown in the following example:

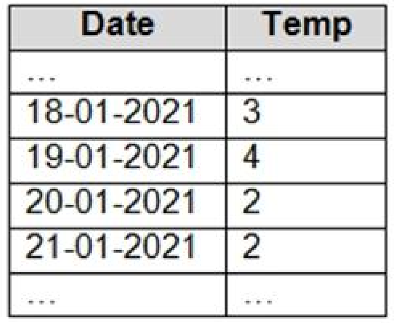
```
SELECT book.* FROM
 OPENROWSET(BULK N't:\books\books.json', SINGLE_CLOB) AS json
 CROSS APPLY OPENJSON(BulkColumn)
 WITH( id nvarchar(100), name nvarchar(100), price float,
    pages_i int, author nvarchar(100)) AS book
```

Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/query-single-csv-file

https://docs.microsoft.com/en-us/sql/relational-databases/json/import-json-documents-into-sql-server

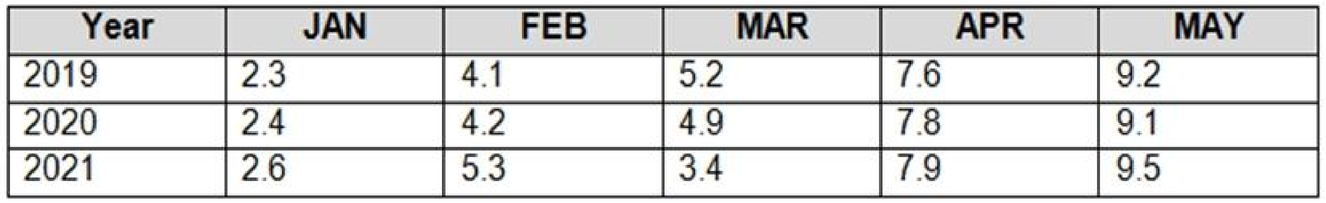
QUESTION 35
HOTSPOT

You have an Apache Spark DataFrame named temperatures. A sample of the data is shown in the following table.



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

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

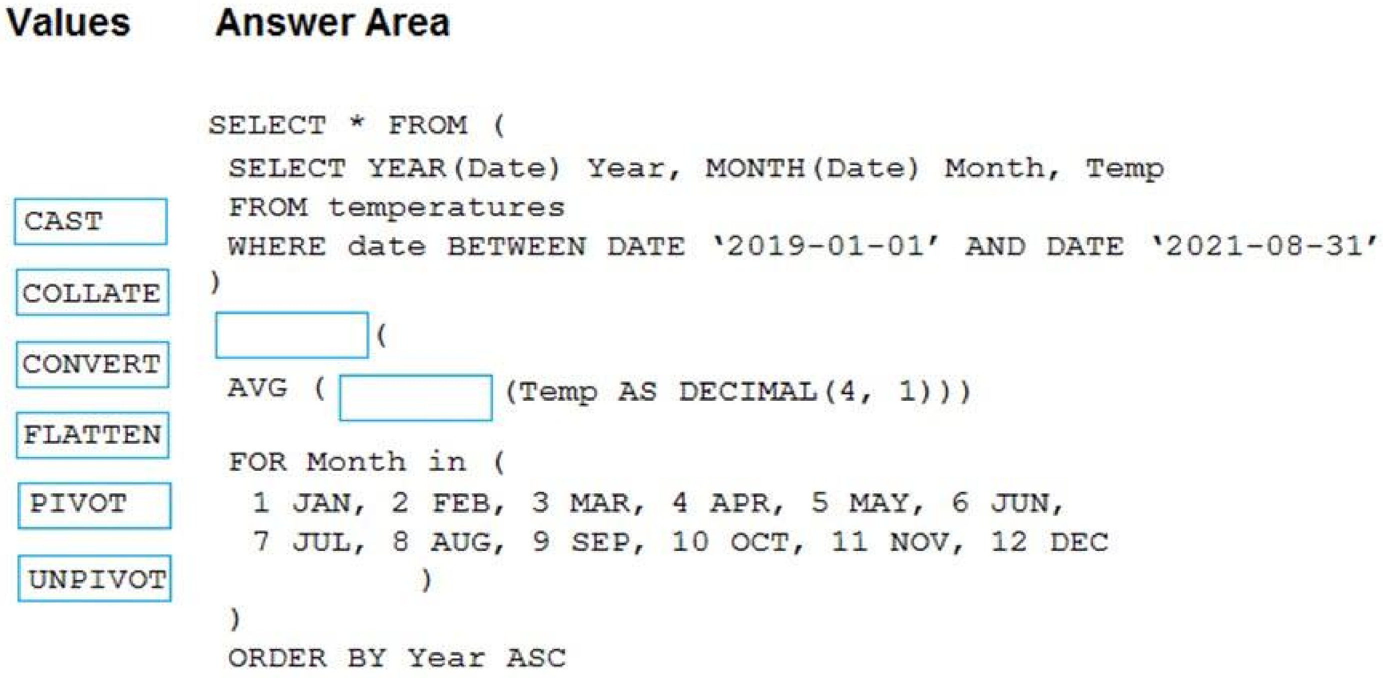


Year	JAN	FEB	MAR	APR	MAY
2019	2.3	4.1	5.2	7.6	9.2
2020	2.4	4.2	4.9	7.8	9.1
2021	2.6	5.3	3.4	7.9	9.5

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

NOTE: Each correct selection is worth one point.

Select and Place:



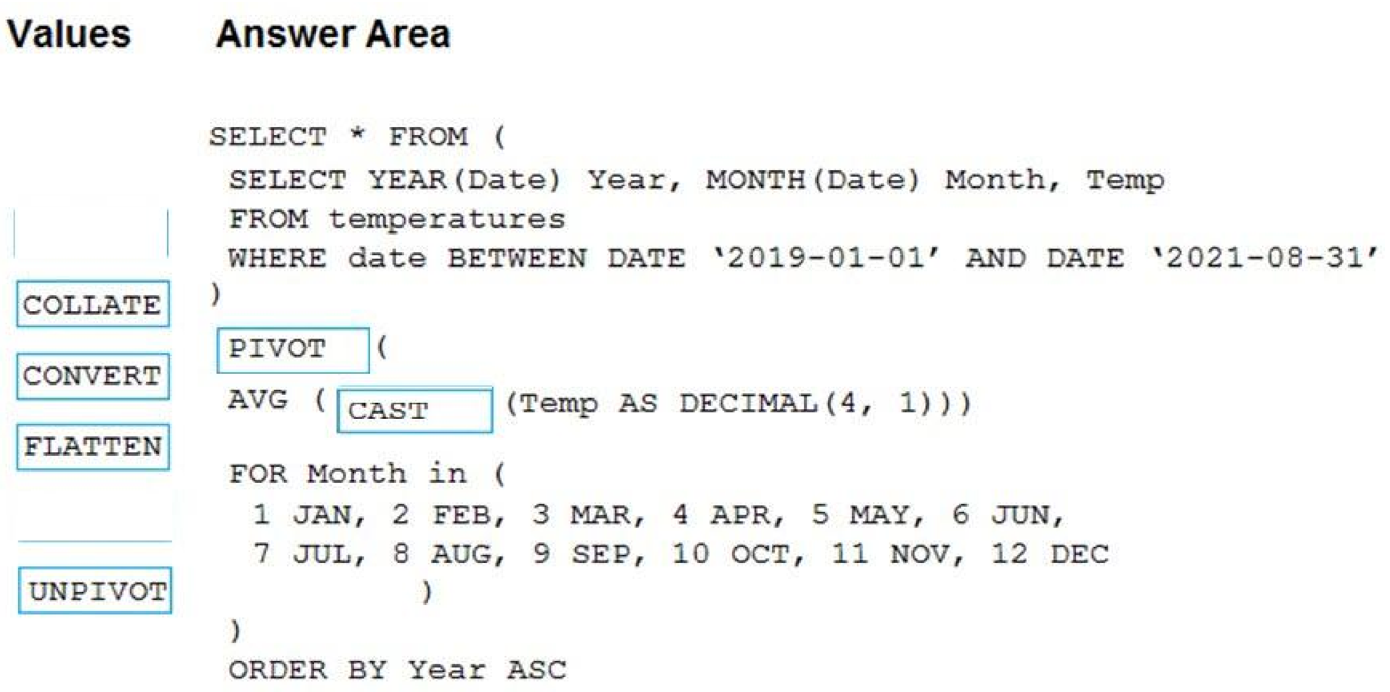
Values

- CAST
- COLLATE
- CONVERT
- FLATTEN
- PIVOT
- UNPIVOT

Answer Area

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

Correct Answer:



Values

COLLATE

CONVERT

FLATTEN

UNPIVOT

Answer Area

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

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: PIVOT
PIVOT rotates a table-valued expression by turning the unique values from one column in the expression into multiple columns in the output. And PIVOT runs aggregations where they're required on any remaining column values that are wanted in the final output.

Incorrect Answers:
UNPIVOT carries out the opposite operation to PIVOT by rotating columns of a table-valued expression into column values.

Box 2: CAST
If you want to convert an integer value to a DECIMAL data type in SQL Server use the CAST() function.

Example:

```
SELECT
  CAST(12 AS DECIMAL(7,2) ) AS decimal_value;
```

Here is the result:

```
decimal_value
12.00
```

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

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

QUESTION 36
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

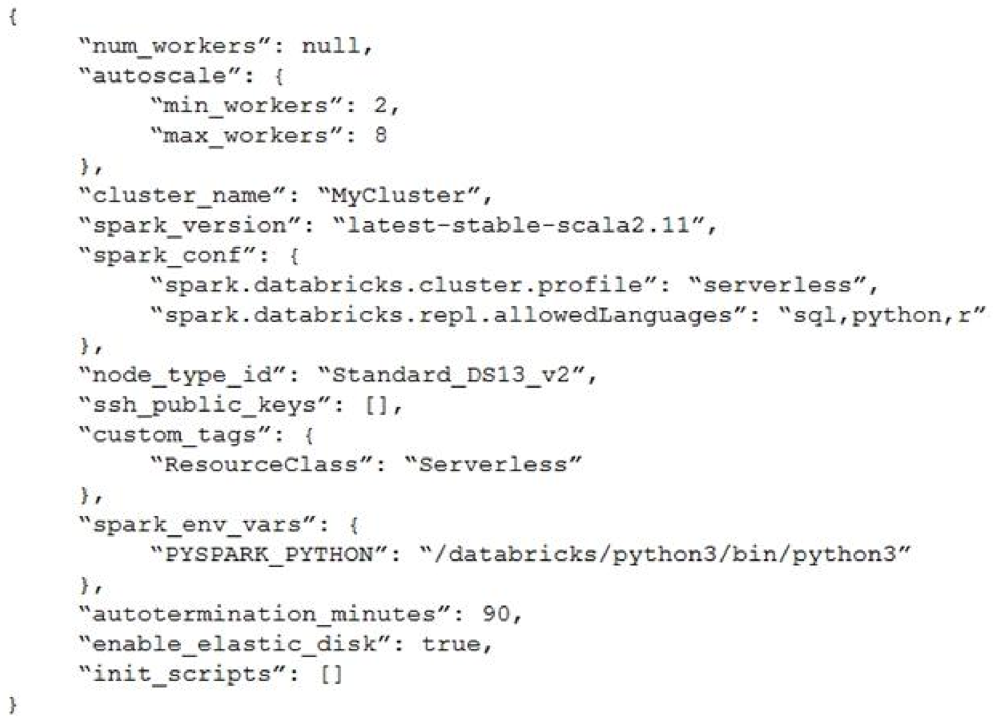
Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
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 37
HOTSPOT

The following code segment is used to create an Azure Databricks cluster.

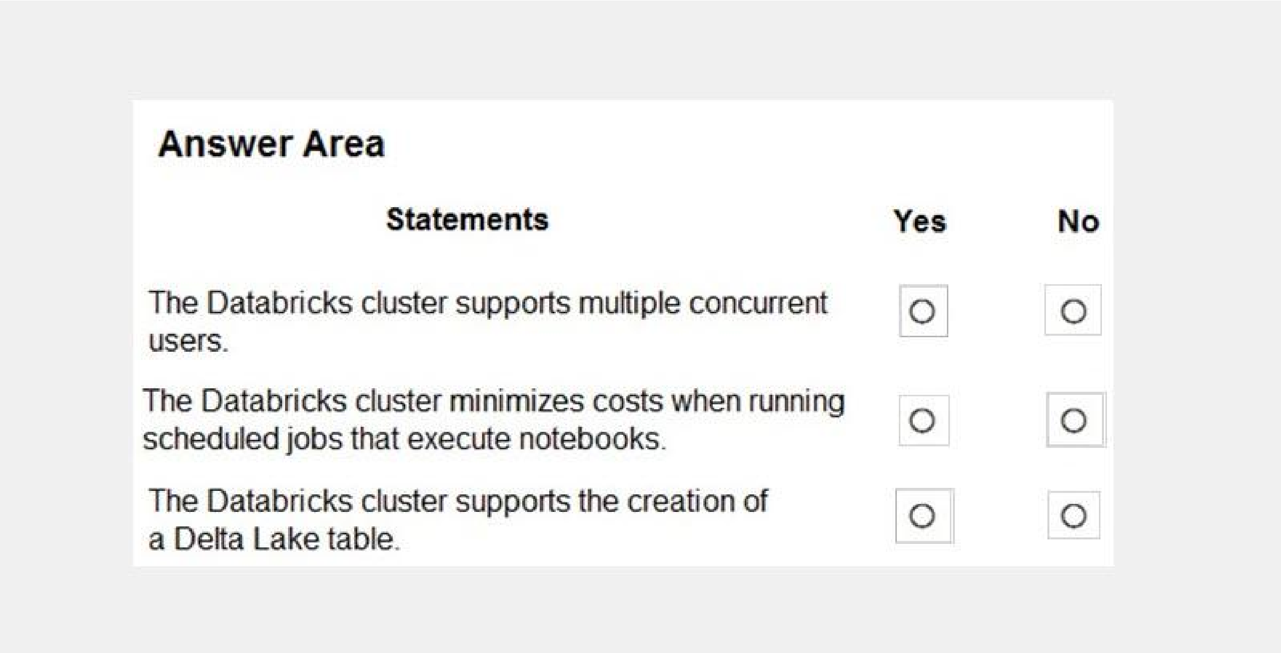


```
{
     “num_workers”: null,
     “autoscale”: {
          “min_workers”: 2,
          “max_workers”: 8
     },
     “cluster_name”: “MyCluster”,
     “spark_version”: “latest-stable-scala2.11”,
     “spark_conf”: {
          “spark.databricks.cluster.profile”: “serverless”,
          “spark.databricks.repl.allowedLanguages”: “sql,python,r”
     },
     “node_type_id”: “Standard_DS13_v2”,
     “ssh_public_keys”: [],
     “custom_tags”: {
          “ResourceClass”: “Serverless”
     },
     “spark_env_vars”: {
          “PYSPARK_PYTHON”: “/databricks/python3/bin/python3”
     },
     “autotermination_minutes”: 90,
     “enable_elastic_disk”: true,
     “init_scripts”: []
}
```

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

NOTE: Each correct selection is worth one point.

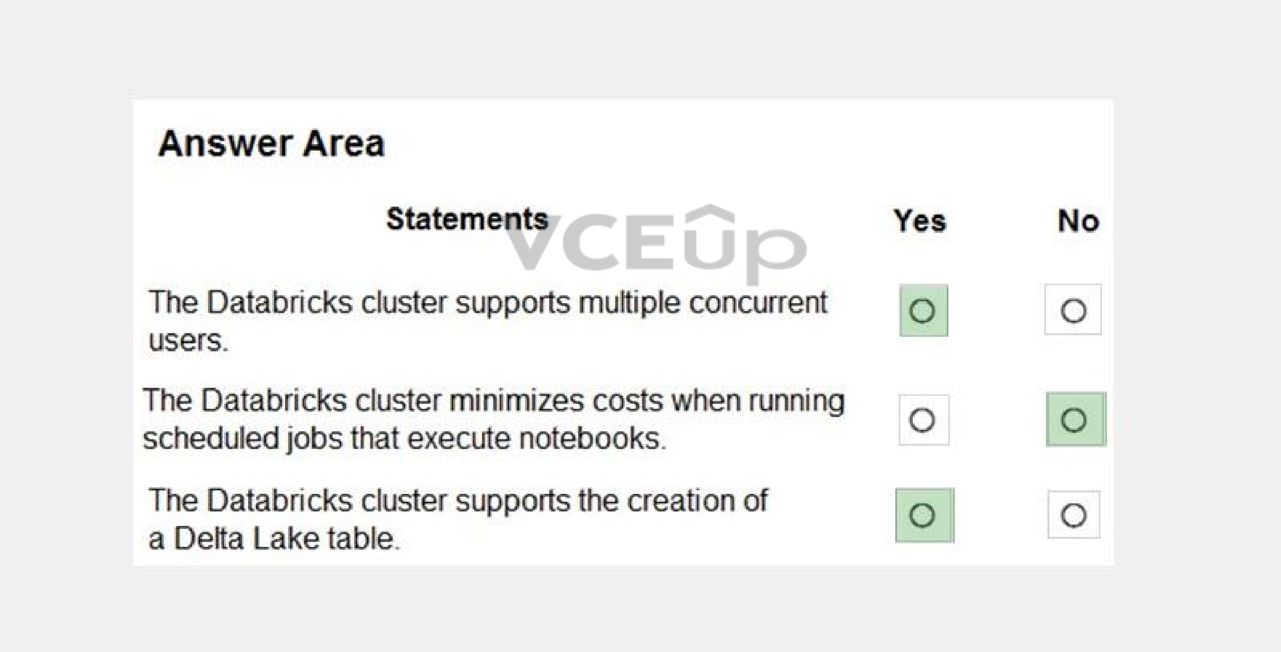
Hot Area:



Answer Area

Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	○	○
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	○	○
The Databricks cluster supports the creation of a Delta Lake table.	○	○

Correct Answer:



Answer Area

Statements	Yes	No
The Databricks cluster supports multiple concurrent users.	[✓ ○]	○
The Databricks cluster minimizes costs when running scheduled jobs that execute notebooks.	○	[✓ ○]
The Databricks cluster supports the creation of a Delta Lake table.	[✓ ○]	○

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: Yes
A cluster mode of ‘High Concurrency’ is selected, unlike all the others which are ‘Standard’. This results in a worker type of Standard_DS13_v2.

Box 2: No
When you run a job on a new cluster, the job is treated as a data engineering (job) workload subject to the job workload pricing. When you run a job on an existing cluster, the job is treated as a data analytics (all-purpose) workload subject to all-purpose workload pricing.

Box 3: Yes
Delta Lake on Databricks allows you to configure Delta Lake based on your workload patterns.

Reference:

https://adatis.co.uk/databricks-cluster-sizing/

https://docs.microsoft.com/en-us/azure/databricks/jobs

https://docs.databricks.com/administration-guide/capacity-planning/cmbp.html

https://docs.databricks.com/delta/index.html

QUESTION 38
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: (none)
Explanation

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

QUESTION 39
HOTSPOT

You have an enterprise data warehouse in Azure Synapse Analytics that contains a table named FactOnlineSales. The table contains data from the start of 2009 to the end of 2012.

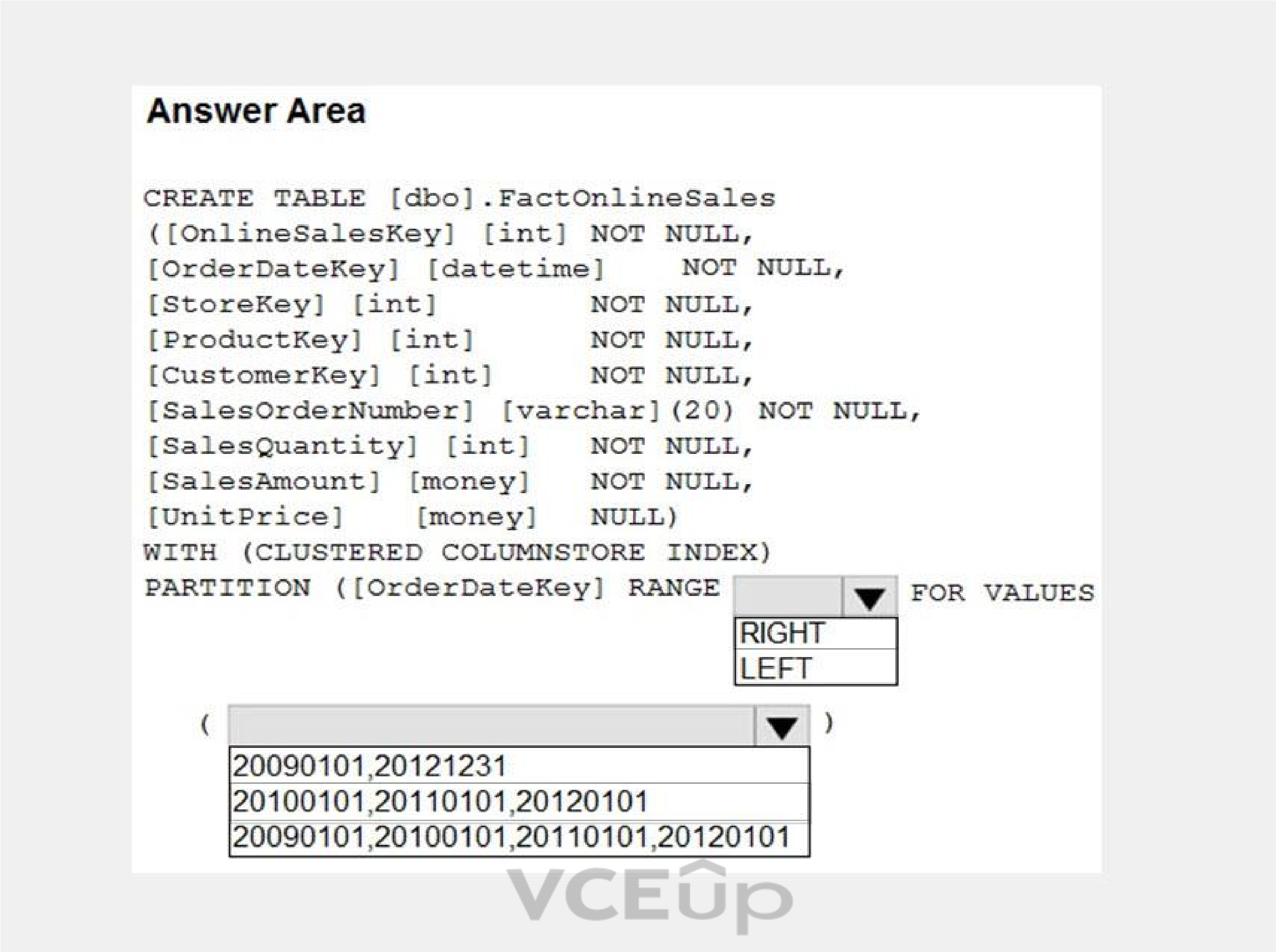
You need to improve the performance of queries against FactOnlineSales by using table partitions. The solution must meet the following requirements:

- Create four partitions based on the order date.
- Ensure that each partition contains all the orders places during a given calendar year.

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

NOTE: Each correct selection is worth one point.

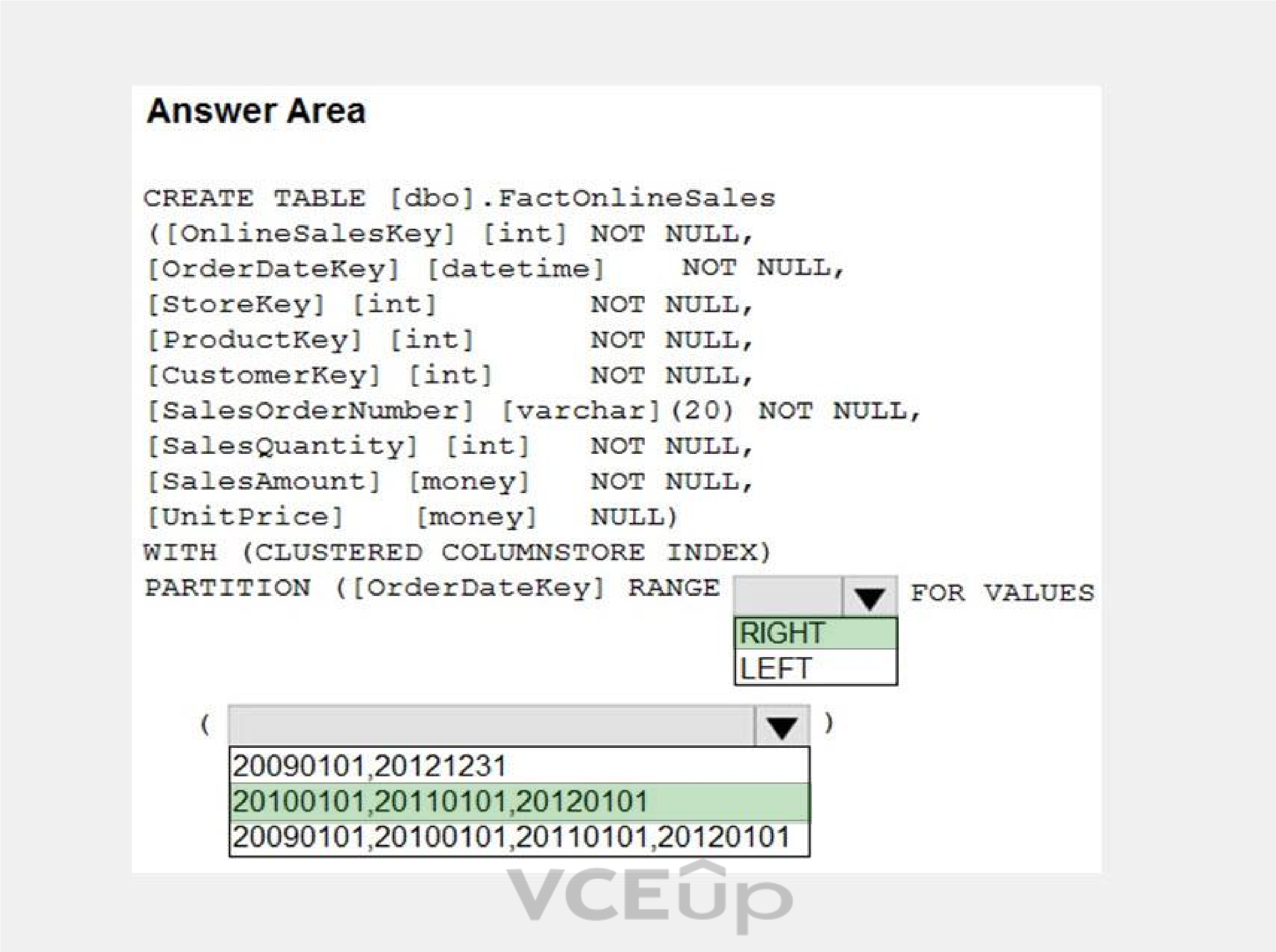
Hot Area:



Answer Area

```
CREATE TABLE [dbo].FactOnlineSales
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime]    NOT NULL,
[StoreKey] [int]         NOT NULL,
[ProductKey] [int]       NOT NULL,
[CustomerKey] [int]      NOT NULL,
[SalesOrderNumber] [varchar](20) NOT NULL,
[SalesQuantity] [int]    NOT NULL,
[SalesAmount] [money]    NOT NULL,
[UnitPrice]    [money]   NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey] RANGE ▼ FOR VALUES
                                RIGHT
                                LEFT
   ( ▼ )
     20090101,20121231
     20100101,20110101,20120101
     20090101,20100101,20110101,20120101
```

Correct Answer:



Answer Area

```
CREATE TABLE [dbo].FactOnlineSales
([OnlineSalesKey] [int] NOT NULL,
[OrderDateKey] [datetime]    NOT NULL,
[StoreKey] [int]        NOT NULL,
[ProductKey] [int]      NOT NULL,
[CustomerKey] [int]     NOT NULL,
[SalesOrderNumber] [varchar](20) NOT NULL,
[SalesQuantity] [int]   NOT NULL,
[SalesAmount] [money]   NOT NULL,
[UnitPrice]    [money]  NULL)
WITH (CLUSTERED COLUMNSTORE INDEX)
PARTITION ([OrderDateKey] RANGE [▼] FOR VALUES
    ( [▼] )
```

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Range Left or Right, both are creating similar partition but there is difference in comparison

For example: in this scenario, when you use LEFT and 20100101,20110101,20120101
Partition will be, datecol<=20100101, datecol>20100101 and datecol<=20110101, datecol>20110101 and datecol<=20120101, datecol>20120101

But if you use range RIGHT and 20100101,20110101,20120101
Partition will be, datecol<20100101, datecol>=20100101 and datecol<20110101, datecol>=20110101 and datecol<20120101, datecol>=20120101

In this example, Range RIGHT will be suitable for calendar comparison Jan 1st to Dec 31st

Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-partition-function-transact-sql?view=sql-server-ver15

QUESTION 40
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.

```
CREATE TABLE [DBO].[DimProduct] (
[ProductKey] [int] IDENTITY(1,1) NOT NULL,
[ProductSourceID] [int] NOT NULL,
[ProductName] [nvarchar] (100) NULL,
[Color] [nvarchar] (15) NULL,
[SellStartDate] [date] NOT NULL,
[SellEndDate] [date] NULL,
[RowInsertedDateTime] [datetime] NOT NULL,
[RowUpdatedDateTime] [datetime] NOT NULL,
[ETLAuditID] [int] NOT NULL
)
```

Which two columns should you add to the table? Each correct answer presents part of the solution.

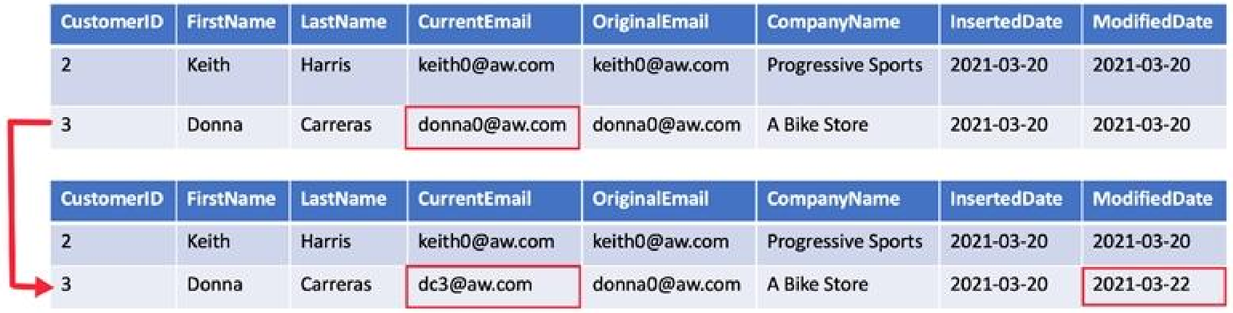
NOTE: Each correct selection is worth one point.

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: BE
Section: (none)
Explanation

Explanation/Reference:
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 41
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: (none)
Explanation

Explanation/Reference:
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

01 - Design and implement data security

QUESTION 1
Case study

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study
To display the first question in this case study, click the **Next** button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an **All Information** tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the **Question** button to return to the question.

Overview

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements

Business Goals

Litware wants to create a new analytics environment in Azure to meet the following requirements:

- See inventory levels across the stores. Data must be updated as close to real time as possible.
- Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.
- Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements

Litware identifies the following technical requirements:

- Minimize the number of different Azure services needed to achieve the business goals.
- Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.
- Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.
- Use Azure Active Directory (Azure AD) authentication whenever possible.
- Use the principle of least privilege when designing security.
- Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.
- Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.
- Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment

Litware plans to implement the following environment:

- The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.
- Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
- Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
- Daily inventory data comes from a Microsoft SQL server located on a private network.
- Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.
- Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.
- Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

A.

Correct Answer:
Section: (none)
Explanation

Explanation/Reference:

QUESTION 2
What should you recommend to prevent users outside the Litware on-premises network from accessing the analytical data store?

A. a server-level virtual network rule
B. a database-level virtual network rule
C. a server-level firewall IP rule
D. a database-level firewall IP rule

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Scenario: Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.

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

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

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

QUESTION 3
What should you recommend using to secure sensitive customer contact information?

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

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Scenario: Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.

Labeling: You can apply sensitivity-classification labels persistently to columns by using new metadata attributes that have been added to the SQL Server database engine. This metadata can then be used for advanced, sensitivity-based auditing and protection scenarios.

Incorrect Answers:
A: Transparent Data Encryption (TDE) encrypts SQL Server, Azure SQL Database, and Azure Synapse Analytics data files, known as encrypting data at rest. TDE does not provide encryption across communication channels.

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

https://docs.microsoft.com/en-us/azure/sql-database/sql-database-security-overview

02 - Design and implement data security

QUESTION 1
DRAG DROP

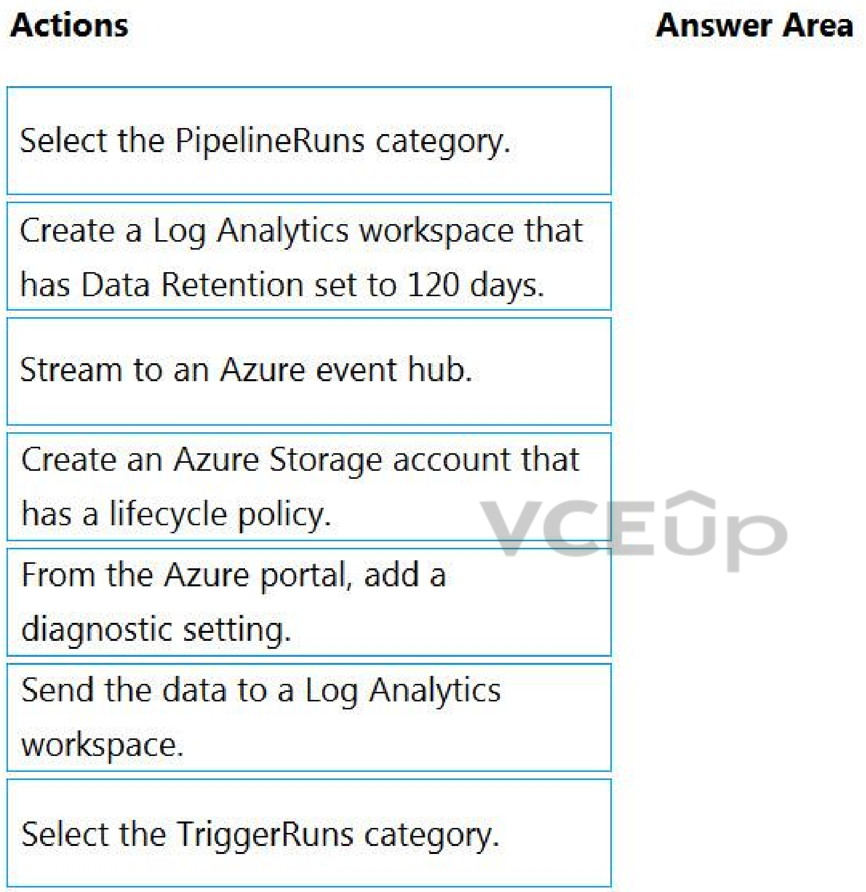
You have an Azure data factory.

You need to ensure that pipeline-run data is retained for 120 days. The solution must ensure that you can query the data by using the Kusto query language.

Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Select and Place:

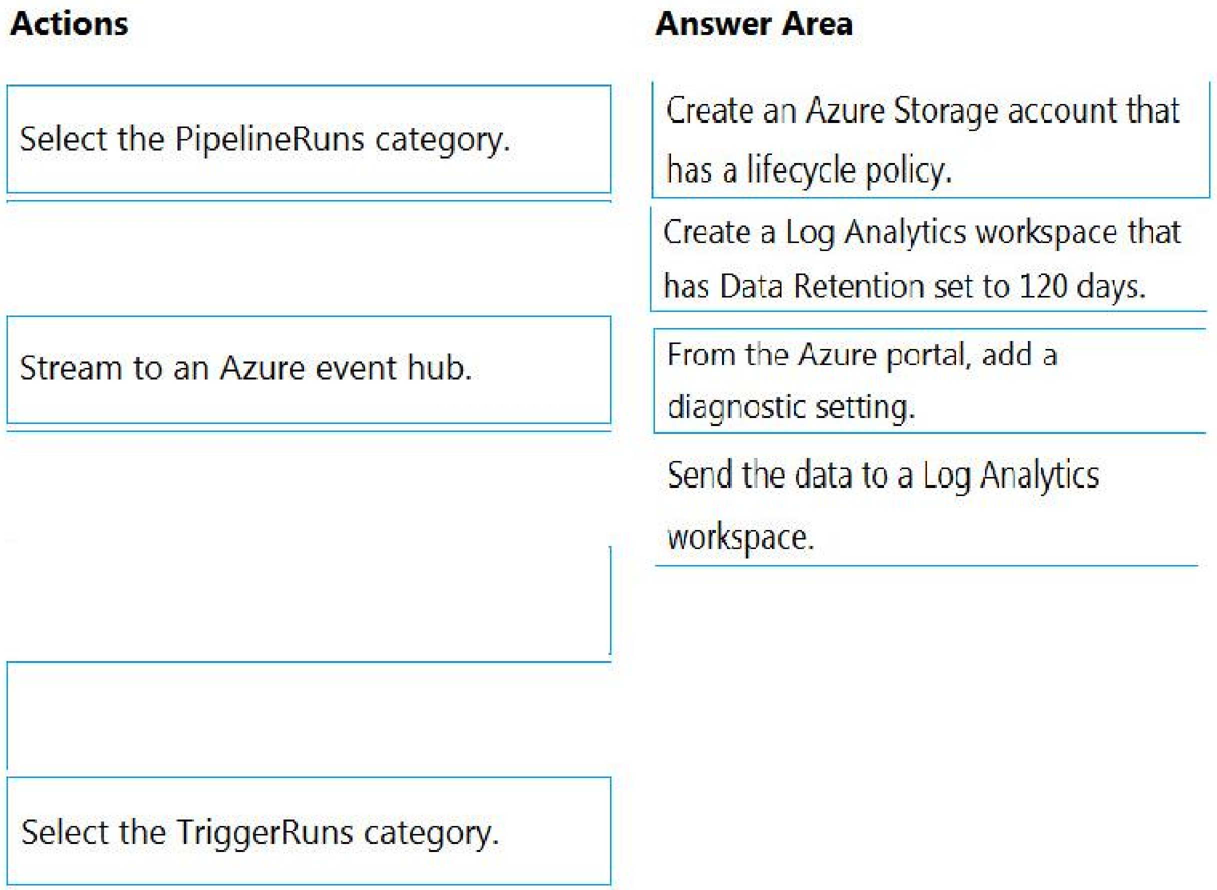


Actions

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

Answer Area

Correct Answer:



Actions

Select the PipelineRuns category.

Stream to an Azure event hub.

Select the TriggerRuns category.

Answer Area

Create an Azure Storage account that has a lifecycle policy.

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

From the Azure portal, add a diagnostic setting.

Send the data to a Log Analytics workspace.

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Step 1: Create an Azure Storage account that has a lifecycle policy
To automate common data management tasks, Microsoft created a solution based on Azure Data Factory. The service, Data Lifecycle Management, makes frequently accessed data available and archives or purges other data according to retention policies. Teams across the company use the service to reduce storage costs, improve app performance, and comply with data retention policies.

Step 2: Create a Log Analytics workspace that has Data Retention set to 120 days.
Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time. With Monitor, you can route diagnostic logs for analysis to multiple different targets, such as a Storage Account: Save your diagnostic logs to a storage account for auditing or manual inspection. You can use the diagnostic settings to specify the retention time in days.

Step 3: From Azure Portal, add a diagnostic setting.

Step 4: Send the data to a log Analytics workspace,
Event Hub: A pipeline that transfers events from services to Azure Data Explorer.

Keeping Azure Data Factory metrics and pipeline-run data.

Configure diagnostic settings and workspace.
Create or add diagnostic settings for your data factory.

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

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

QUESTION 2
You have an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that data in the pool is encrypted at rest. The solution must **NOT** require modifying applications that query the data.

What should you do?

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

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Transparent Data Encryption (TDE) helps protect against the threat of malicious activity by encrypting and decrypting your data at rest. When you encrypt your database, associated backups and transaction log files are encrypted without requiring any changes to your applications. TDE encrypts the storage of an entire database by using a symmetric key called the database encryption key.

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

QUESTION 3
DRAG DROP

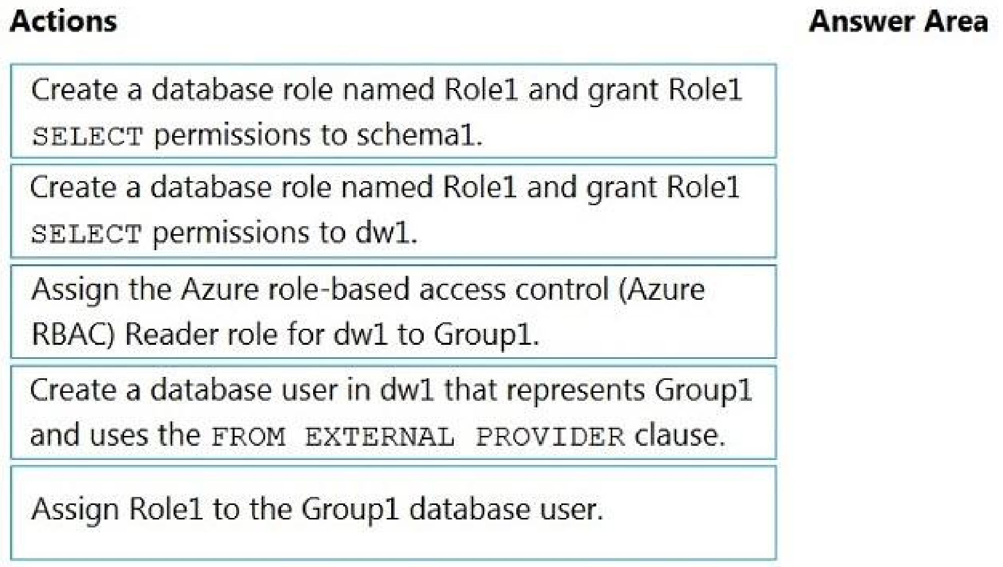
You have an Azure Active Directory (Azure AD) tenant that contains a security group named Group1. You have an Azure Synapse Analytics dedicated SQL pool named dw1 that contains a schema named schema1.

You need to grant Group1 read-only permissions to all the tables and views in schema1. The solution must use the principle of least privilege.

Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Select and Place:

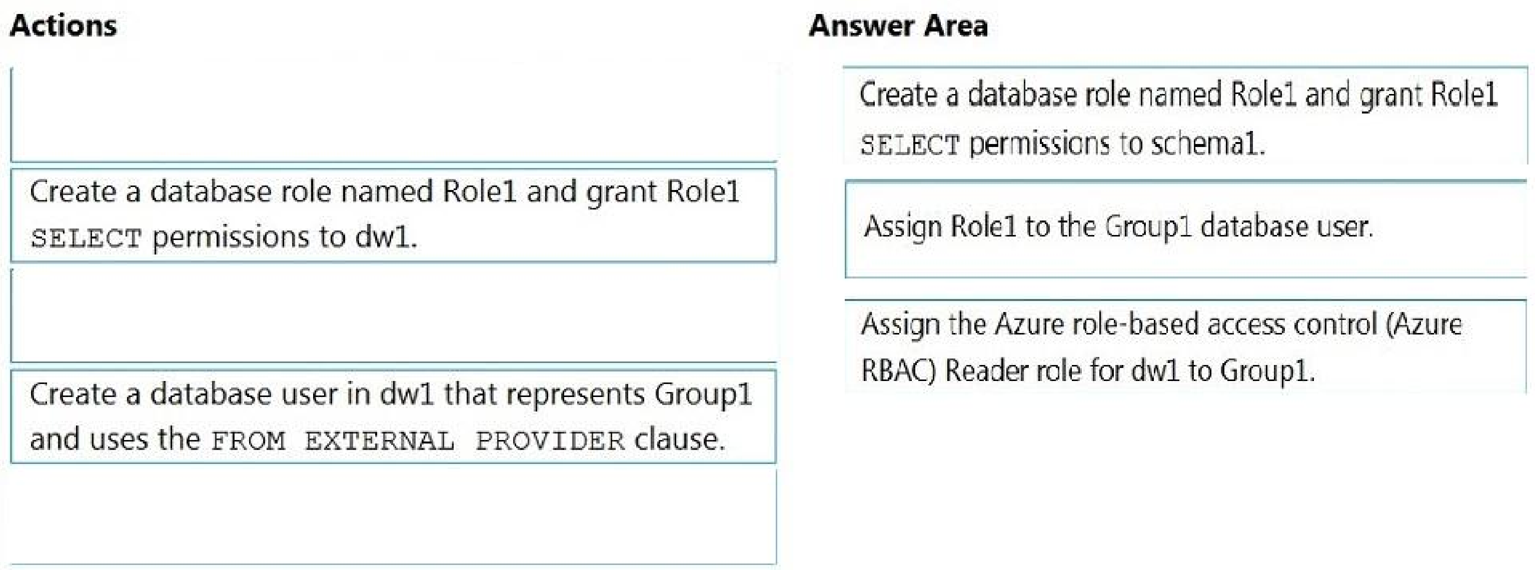


Actions

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

Answer Area

Correct Answer:



Actions

Create a database role named Role1 and grant Role1 SELECT permissions to dw1.

Create a database user in dw1 that represents Group1 and uses the FROM EXTERNAL PROVIDER clause.

Answer Area

Create a database role named Role1 and grant Role1 SELECT permissions to schema1.

Assign Role1 to the Group1 database user.

Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1.

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Step 1: Create a database role named Role1 and grant Role1 SELECT permissions to schema
You need to grant Group1 read-only permissions to all the tables and views in schema1.
Place one or more database users into a database role and then assign permissions to the database role.

Step 2: Assign Rol1 to the Group database user

Step 3: Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1

Reference:
https://docs.microsoft.com/en-us/azure/data-share/how-to-share-from-sql

QUESTION 4
HOTSPOT

You have an Azure subscription that contains a logical Microsoft SQL server named Server1. Server1 hosts an Azure Synapse Analytics SQL dedicated pool named Pool1.

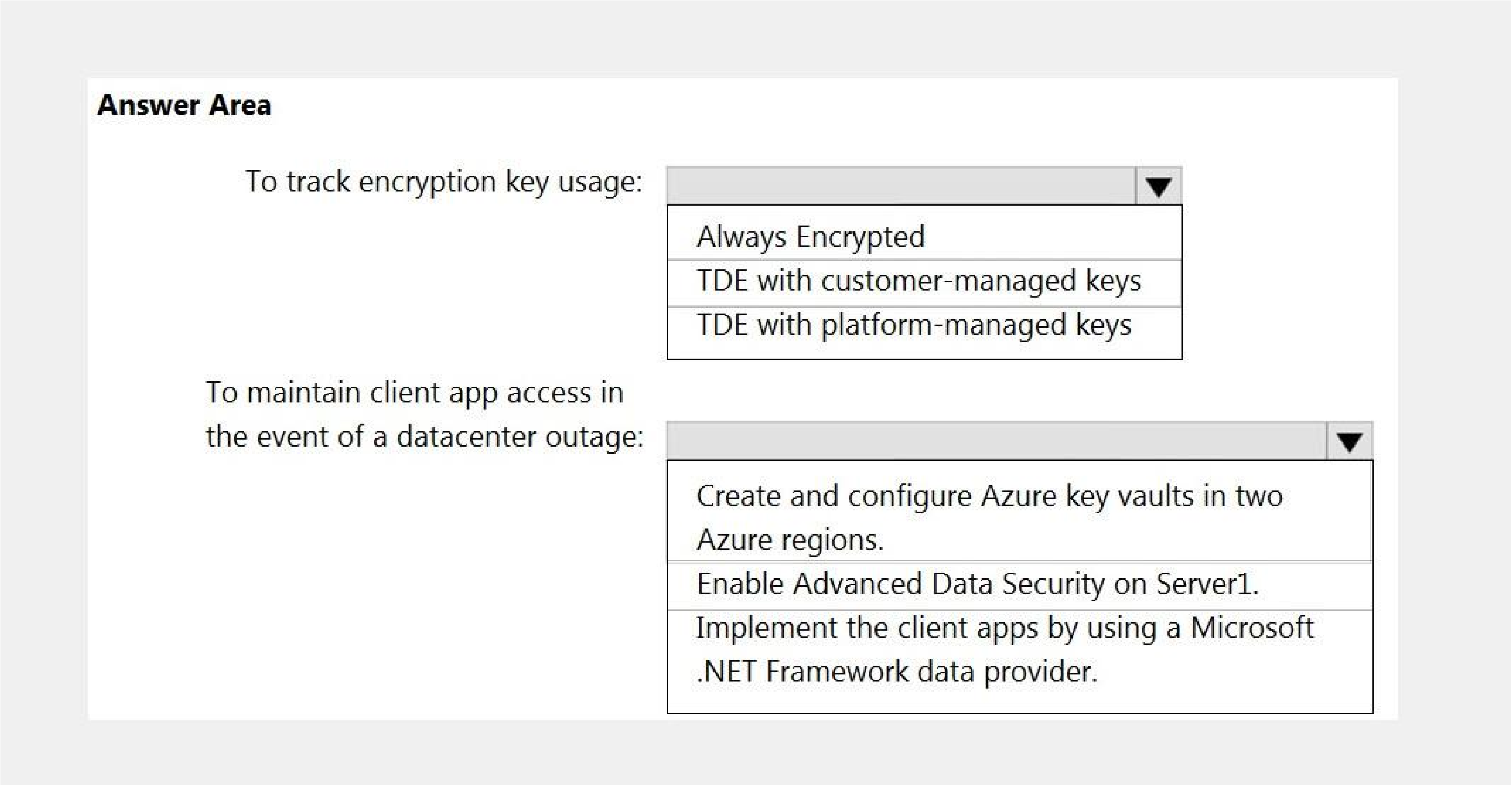
You need to recommend a Transparent Data Encryption (TDE) solution for Server1. The solution must meet the following requirements:

- Track the usage of encryption keys.
- Maintain the access of client apps to Pool1 in the event of an Azure datacenter outage that affects the availability of the encryption keys.

What should you include in the recommendation? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

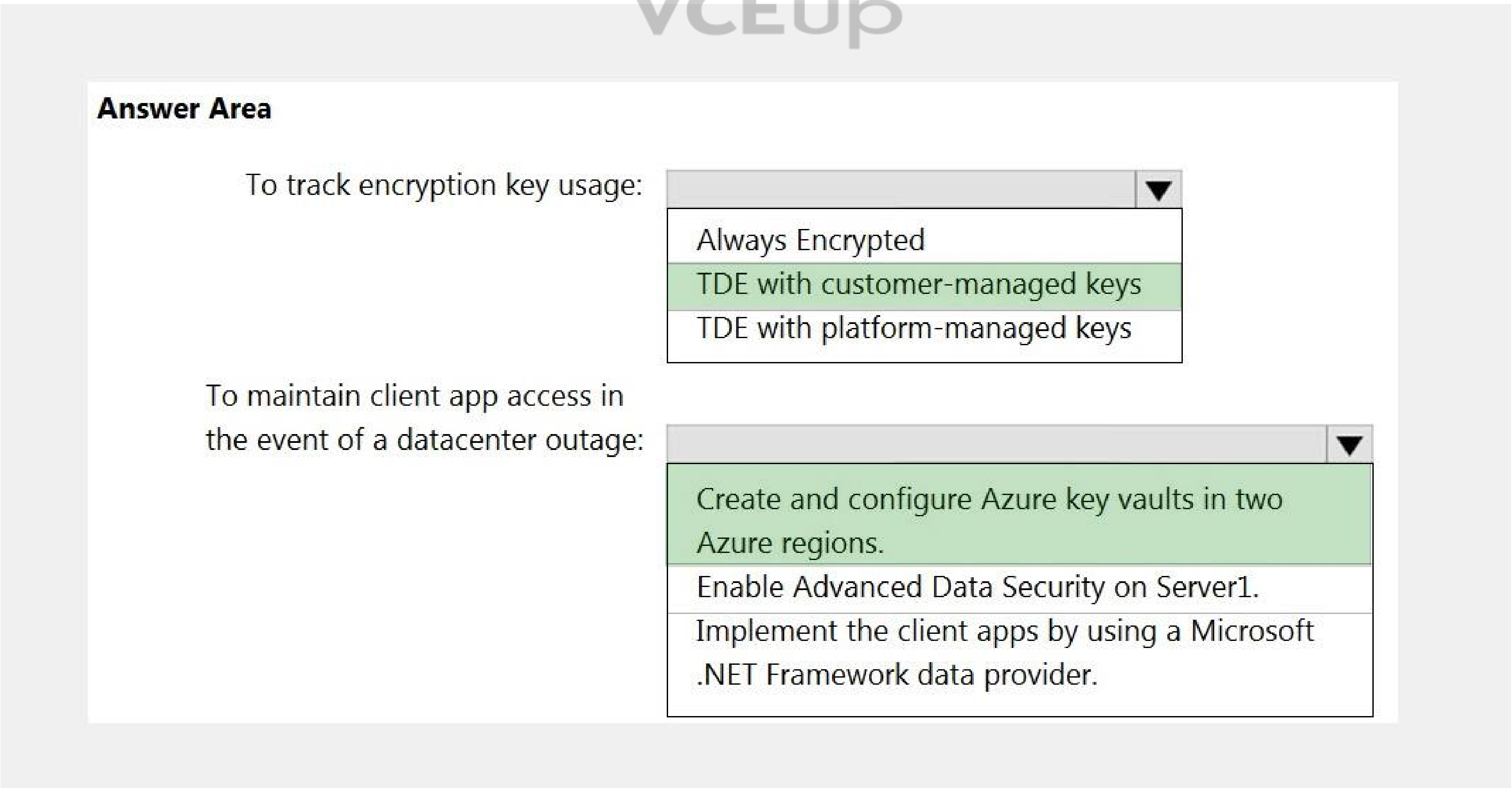
To track encryption key usage: ▼

- Always Encrypted
- TDE with customer-managed keys
- TDE with platform-managed keys

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

- Create and configure Azure key vaults in two Azure regions.
- Enable Advanced Data Security on Server1.
- Implement the client apps by using a Microsoft .NET Framework data provider.

Correct Answer:



Answer Area

To track encryption key usage: ▼

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

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

- [✓ Create and configure Azure key vaults in two Azure regions.]
- Enable Advanced Data Security on Server1.
- Implement the client apps by using a Microsoft .NET Framework data provider.

Section: (none)

Explanation

Explanation/Reference:
Explanation:

Box 1: TDE with customer-managed keys
Customer-managed keys are stored in the Azure Key Vault. You can monitor how and when your key vaults are accessed, and by whom. You can do this by enabling logging for Azure Key Vault, which saves information in an Azure storage account that you provide.

Box 2: Create and configure Azure key vaults in two Azure regions
The contents of your key vault are replicated within the region and to a secondary region at least 150 miles away, but within the same geography to maintain high durability of your keys and secrets.

Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption

https://docs.microsoft.com/en-us/azure/key-vault/general/logging

QUESTION 5
You plan to create an Azure Synapse Analytics dedicated SQL pool.

You need to minimize the time it takes to identify queries that return confidential information as defined by the company's data privacy regulations and the users who executed the queues.

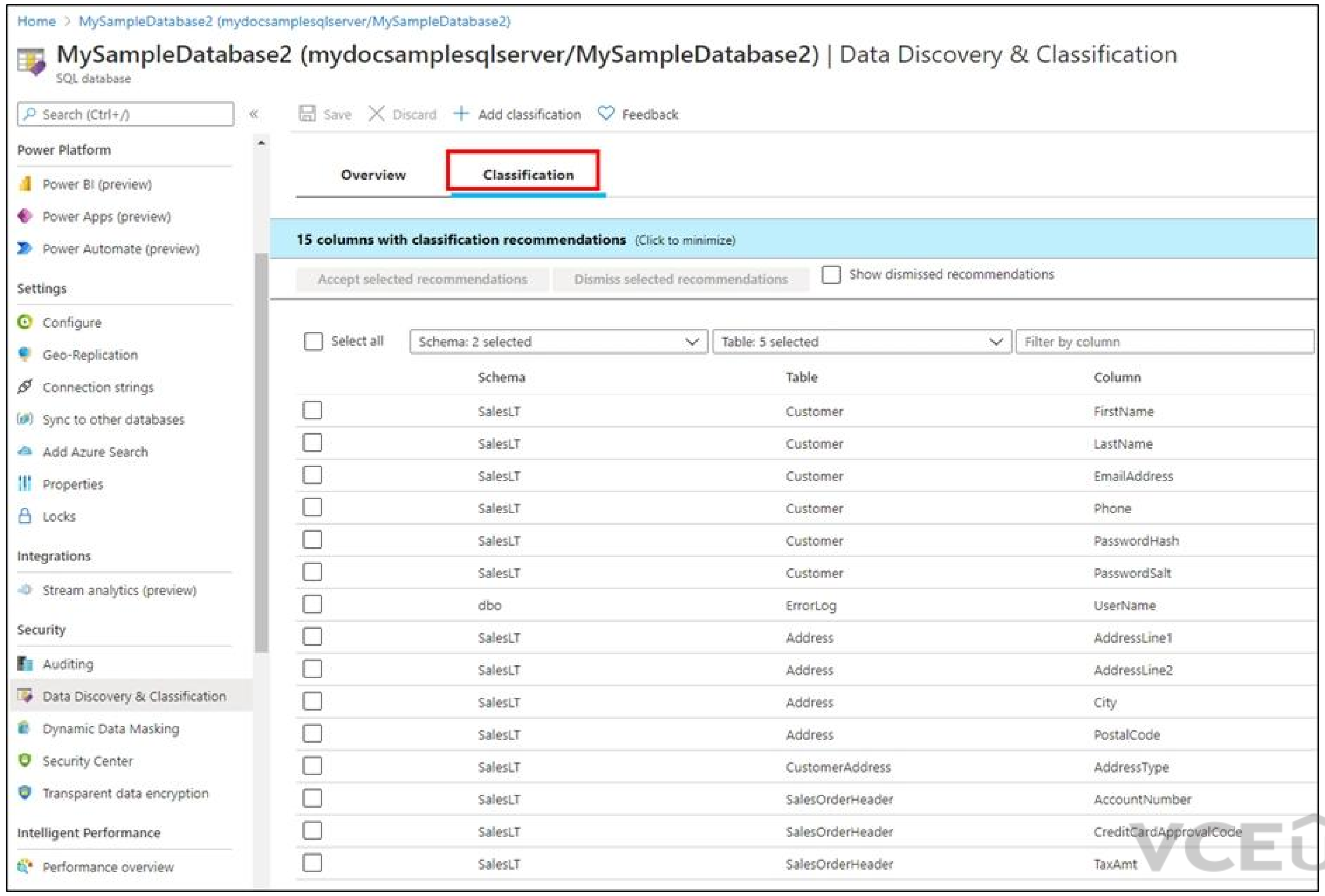
Which two components should you include in the solution? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. sensitivity-classification labels applied to columns that contain confidential information
B. resource tags for databases that contain confidential information
C. audit logs sent to a Log Analytics workspace
D. dynamic data masking for columns that contain confidential information

Correct Answer: AC
Section: (none)
Explanation

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



Home > MySampleDatabase2 (mydocsamplesqlserver/MySampleDatabase2)

MySampleDatabase2 (mydocsamplesqlserver/MySampleDatabase2) | Data Discovery & Classification

SQL database

15 columns with classification recommendations (Click to minimize)

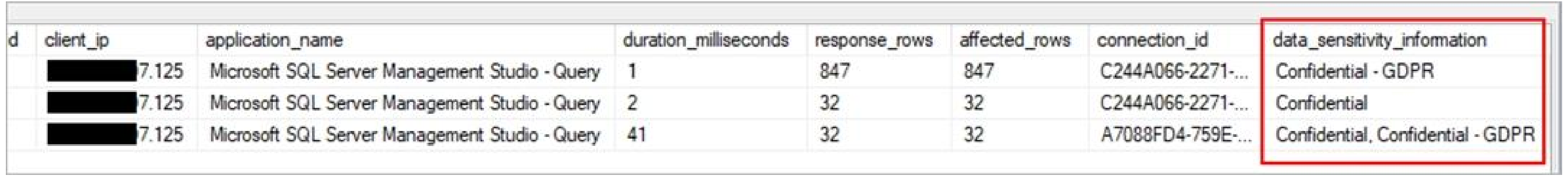
Accept selected recommendations | Dismiss selected recommendations | □ Show dismissed recommendations

□ Select all | Schema: 2 selected | Table: 5 selected | Filter by column

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

1. Select Add classification in the top menu of the pane.
2. In the context window that opens, select the schema, table, and column that you want to classify, and the information type and sensitivity label.
3. Select Add classification at the bottom of the context window.

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



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

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

QUESTION 6

You are designing an enterprise data warehouse in Azure Synapse Analytics that will contain a table named Customers. Customers will contain credit card information.

You need to recommend a solution to provide salespeople with the ability to view all the entries in Customers. The solution must prevent all the salespeople from viewing or inferring the credit card information.

What should you include in the recommendation?

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

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
SQL Database dynamic data masking limits sensitive data exposure by masking it to non-privileged users.

The Credit card masking method exposes the last four digits of the designated fields and adds a constant string as a prefix in the form of a credit card.

Example: XXXX-XXXX-XXXX-1234

Reference:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-dynamic-data-masking-get-started

QUESTION 7
You develop data engineering solutions for a company.

A project requires the deployment of data to Azure Data Lake Storage.

You need to implement role-based access control (RBAC) so that project members can manage the Azure Data Lake Storage resources.

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

NOTE: Each correct selection is worth one point.

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

Correct Answer: ACE
Section: (none)
Explanation

Explanation/Reference:
Explanation:
AC: Create security groups in Azure Active Directory. Assign users or security groups to Data Lake Storage Gen1 accounts.

E: Assign users or security groups as ACLs to the Data Lake Storage Gen1 file system

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

QUESTION 8
You have an Azure Data Factory version 2 (V2) resource named Df1. Df1 contains a linked service.

You have an Azure Key vault named vault1 that contains an encryption key named key1.

You need to encrypt Df1 by using key1.

What should you do first?

A. Add a private endpoint connection to vaul1.
B. Enable Azure role-based access control on vault1.
C. Remove the linked service from Df1.
D. Create a self-hosted integration runtime.

Correct Answer: C
Section: (none)

Explanation

Explanation/Reference:
Explanation:
Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources.

Incorrect Answers:
D: A self-hosted integration runtime copies data between an on-premises store and cloud storage.

Reference:
https://docs.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key

https://docs.microsoft.com/en-us/azure/data-factory/concepts-linked-services

https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime

QUESTION 9
You are designing an Azure Synapse Analytics dedicated SQL pool.

You need to ensure that you can audit access to Personally Identifiable Information (PII).

What should you include in the solution?

A. column-level security
B. dynamic data masking
C. row-level security (RLS)
D. sensitivity classifications

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Data Discovery & Classification is built into Azure SQL Database, Azure SQL Managed Instance, and Azure Synapse Analytics. It provides basic capabilities for discovering, classifying, labeling, and reporting the sensitive data in your databases.

Your most sensitive data might include business, financial, healthcare, or personal information. Discovering and classifying this data can play a pivotal role in your organization's information-protection approach. It can serve as infrastructure for:

- Helping to meet standards for data privacy and requirements for regulatory compliance.
- Various security scenarios, such as monitoring (auditing) access to sensitive data.
- Controlling access to and hardening the security of databases that contain highly sensitive data.

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

QUESTION 10
HOTSPOT

You have an Azure subscription that contains an Azure Data Lake Storage account. The storage account contains a data lake named DataLake1.

You plan to use an Azure data factory to ingest data from a folder in DataLake1, transform the data, and land the data in another folder.

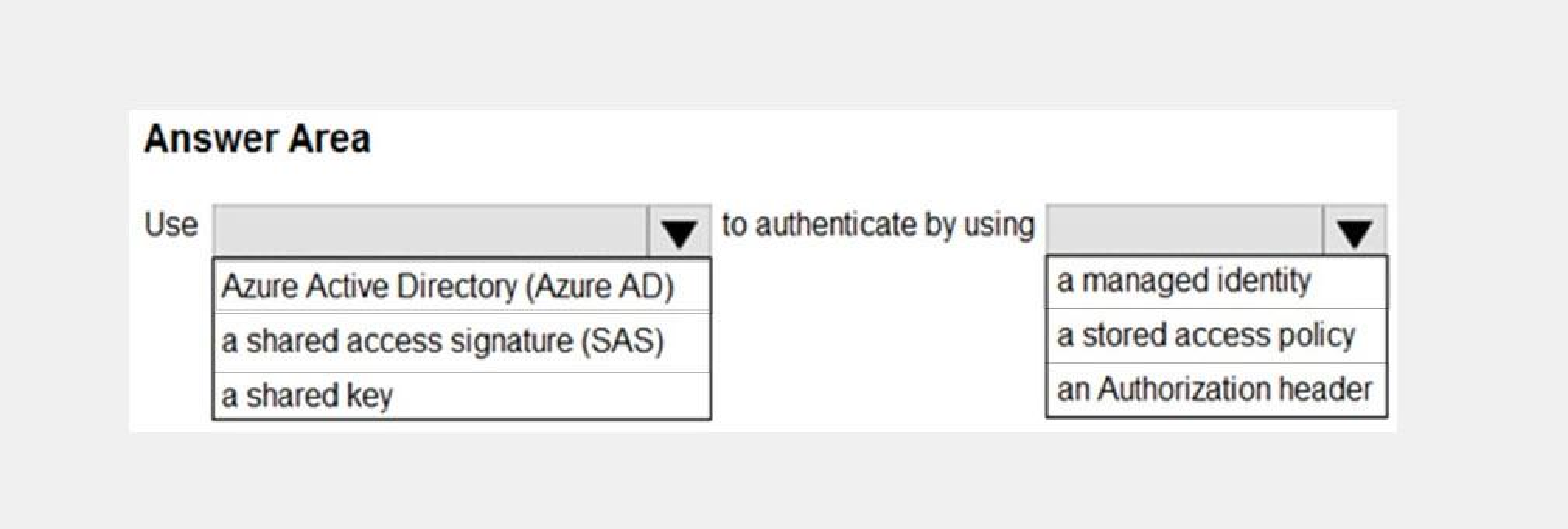
You need to ensure that the data factory can read and write data from any folder in the DataLake1 file system. The solution must meet the following requirements:

- Minimize the risk of unauthorized user access.
- Use the principle of least privilege.
- Minimize maintenance effort.

How should you configure access to the storage account for the data factory? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

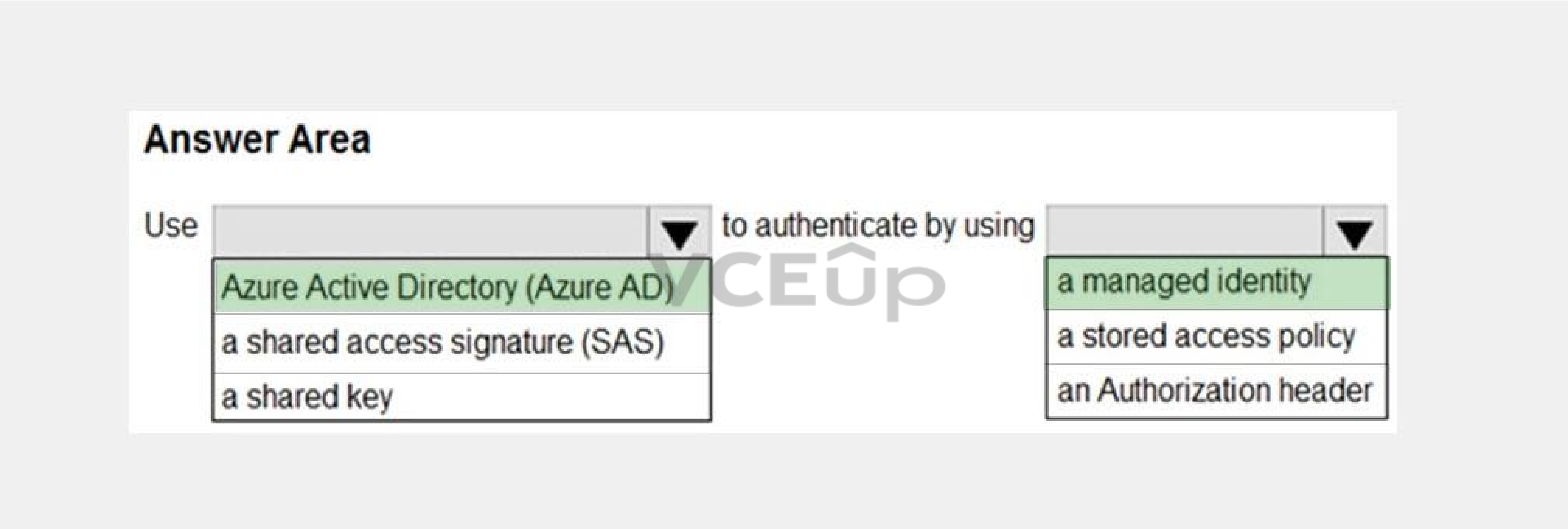


Answer Area

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

to authenticate by using
- a managed identity
- a stored access policy
- an Authorization header

Correct Answer:



Answer Area

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

to authenticate by using
- [✓ a managed identity]
- a stored access policy
- an Authorization header

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: Azure Active Directory (Azure AD)
On Azure, managed identities eliminate the need for developers having to manage credentials by providing an identity for the Azure resource in Azure AD and using it to obtain Azure Active Directory (Azure AD) tokens.

Box 2: a managed identity
A data factory can be associated with a managed identity for Azure resources, which represents this specific data factory. You can directly use this managed identity for Data Lake Storage Gen2 authentication, similar to using your own service principal. It allows this designated factory to access and copy data to or from your Data Lake Storage Gen2.

Note: The Azure Data Lake Storage Gen2 connector supports the following authentication types.
- Account key authentication
- Service principal authentication
- Managed identities for Azure resources authentication

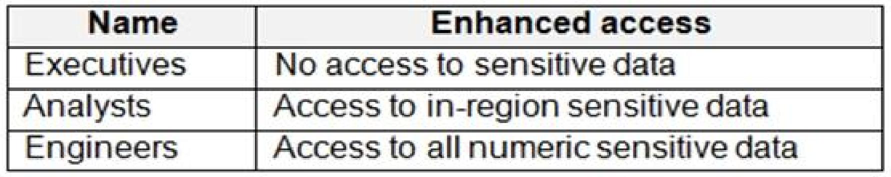
Reference:
https://docs.microsoft.com/en-us/azure/active-directory/managed-identities-azure-resources/overview

https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage

QUESTION 11
HOTSPOT

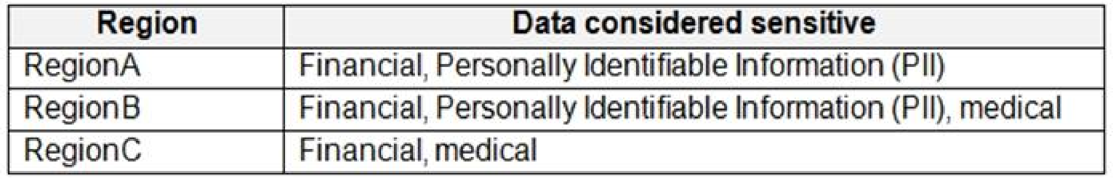
You are designing an Azure Synapse Analytics dedicated SQL pool.

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



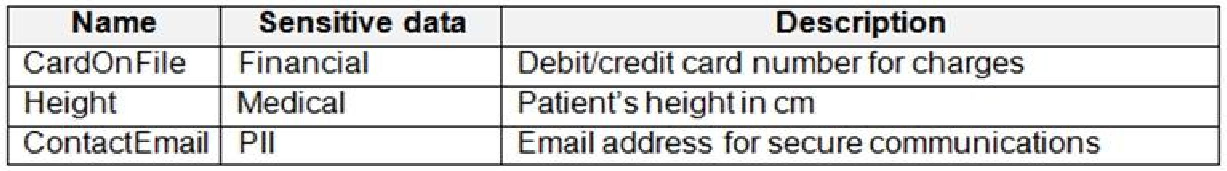
Name	Enhanced access
Executives	No access to sensitive data
Analysts	Access to in-region sensitive data
Engineers	Access to all numeric sensitive data

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



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

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



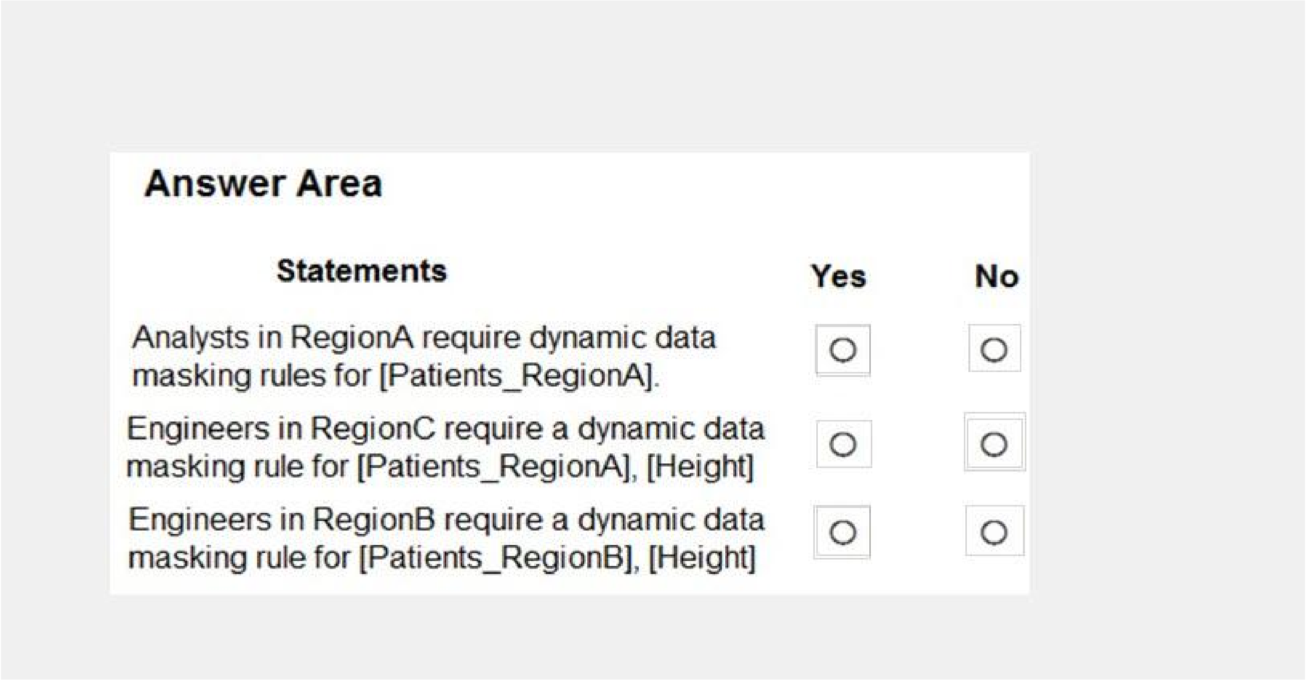
Name	Sensitive data	Description
CardOnFile	Financial	Debit/credit card number for charges
Height	Medical	Patient's height in cm
ContactEmail	PII	Email address for secure communications

You are designing dynamic data masking to maintain compliance.

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

NOTE: Each correct selection is worth one point.

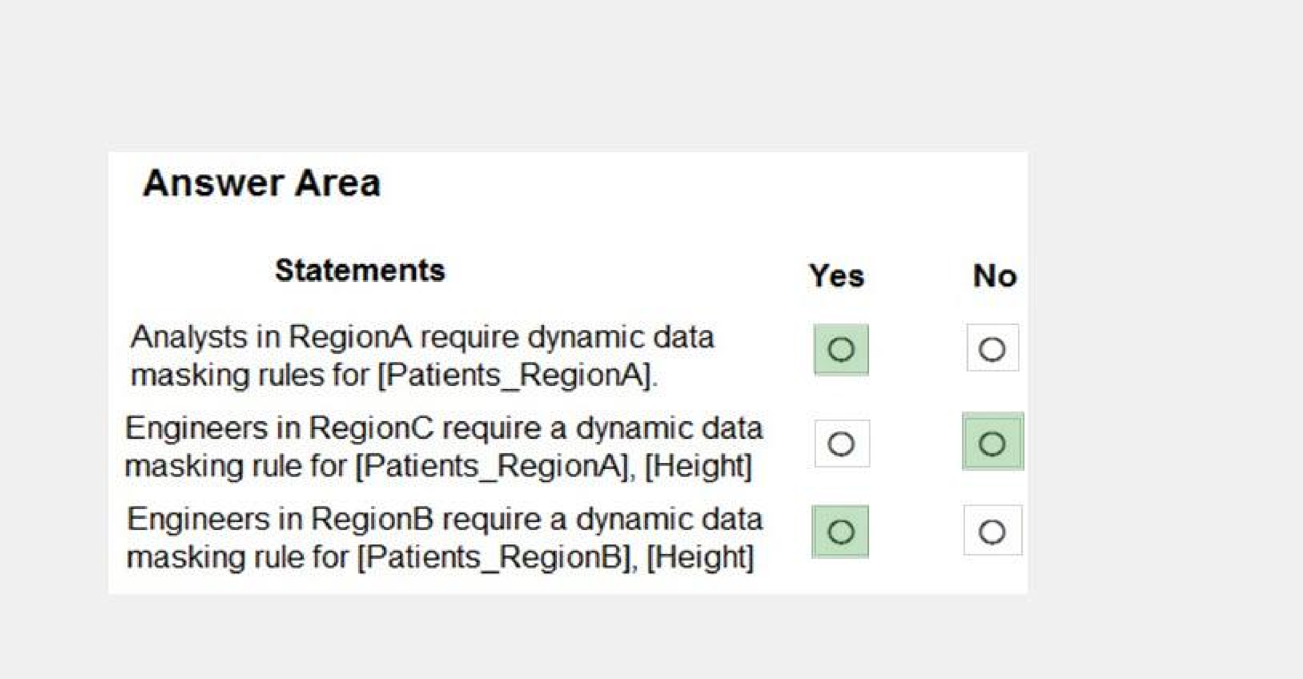
Hot Area:



Answer Area

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

Correct Answer:



Answer Area

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

Section: (none)
Explanation

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

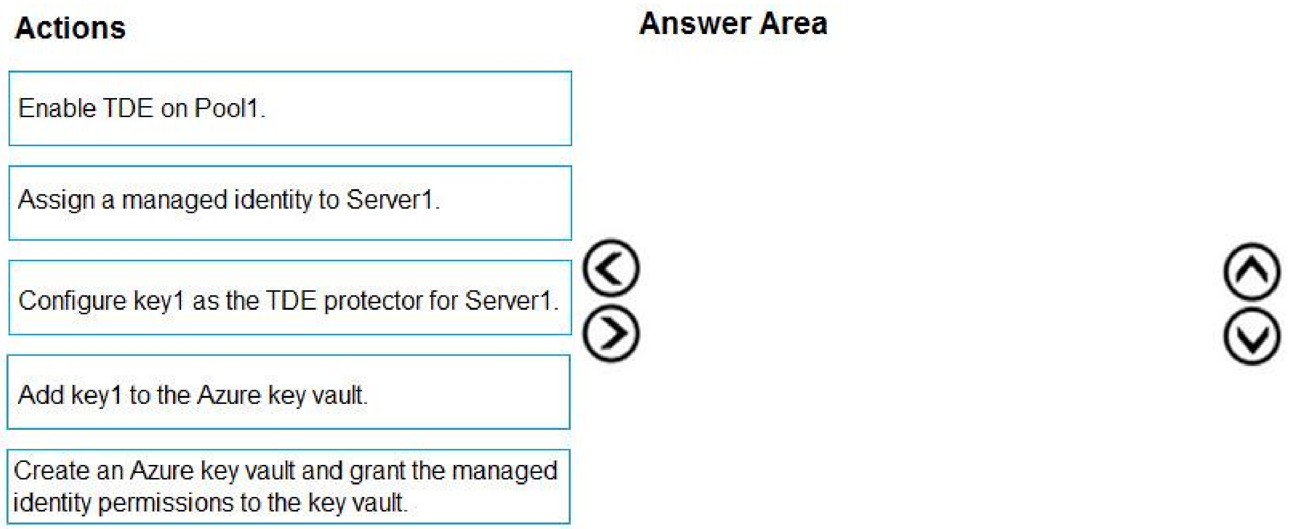
QUESTION 12
DRAG DROP

You have an Azure Synapse Analytics SQL pool named Pool1 on a logical Microsoft SQL server named Server1.

You need to implement Transparent Data Encryption (TDE) on Pool1 by using a custom key named key1.

Which five actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

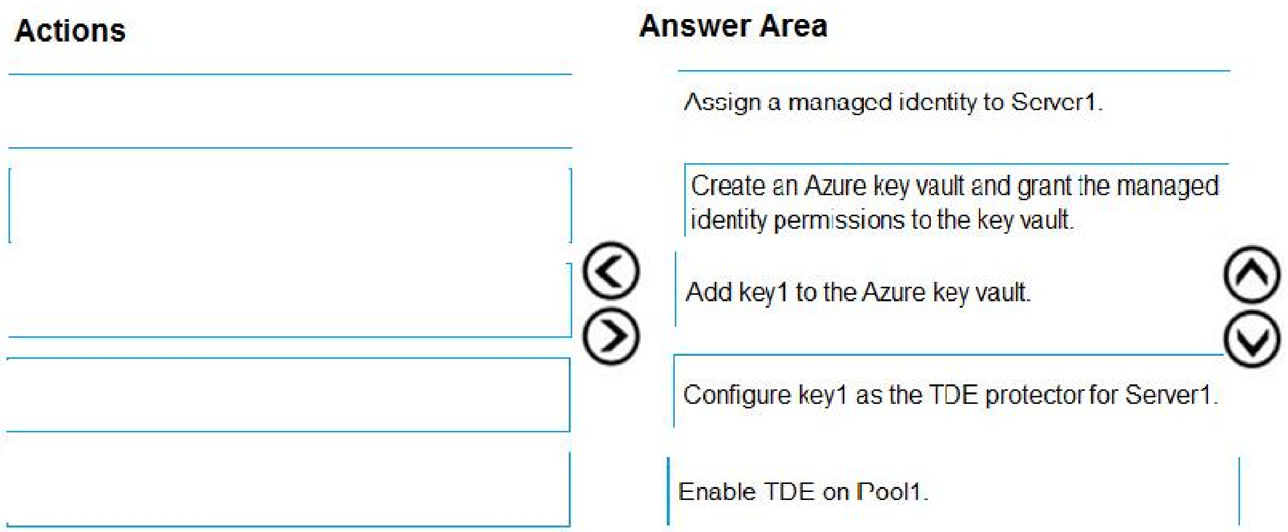


Actions

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

Answer Area

Correct Answer:



Actions

Answer Area

Assign a managed identity to Server1.

Create an Azure key vault and grant the managed identity permissions to the key vault.

Add key1 to the Azure key vault.

Configure key1 as the TDE protector for Server1.

Enable TDE on Pool1.

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Step 1: Assign a managed identity to Server1
You will need an existing Managed Instance as a prerequisite.

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

Step 3: Add key1 to the Azure key vault
The recommended way is to import an existing key from a .pfx file or get an existing key from the vault. Alternatively, generate a new key directly in Azure Key Vault.

Step 4: Configure key1 as the TDE protector for Server1
Provide TDE Protector key

Step 5: Enable TDE on Pool1

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

QUESTION 13
You have a data warehouse in Azure Synapse Analytics.

You need to ensure that the data in the data warehouse is encrypted at rest.

What should you enable?

A. Advanced Data Security for this database
B. Transparent Data Encryption (TDE)
C. Secure transfer required
D. Dynamic Data Masking

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Azure SQL Database currently supports encryption at rest for Microsoft-managed service side and client-side encryption scenarios.

- Support for server encryption is currently provided through the SQL feature called Transparent Data Encryption.

- Client-side encryption of Azure SQL Database data is supported through the Always Encrypted feature.

Reference:
https://docs.microsoft.com/en-us/azure/security/fundamentals/encryption-atrest

QUESTION 14
You are designing a streaming data solution that will ingest variable volumes of data.

You need to ensure that you can change the partition count after creation.

Which service should you use to ingest the data?

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

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
You can't change the partition count for an event hub after its creation except for the event hub in a dedicated cluster.

Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features

QUESTION 15
You are designing a date dimension table in an Azure Synapse Analytics dedicated SQL pool. The date dimension table will be used by all the fact tables.

Which distribution type should you recommend to minimize data movement?

A. `HASH`
B. `REPLICATE`
C. `ROUND_ROBIN`

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
A replicated table has a full copy of the table available on every Compute node. Queries run fast on replicated tables since joins on replicated tables don't require data movement. Replication requires extra storage, though, and isn't practical for large tables.

Incorrect Answers:
A: A hash distributed table is designed to achieve high performance for queries on large tables.
C: A round-robin table distributes table rows evenly across all distributions. The rows are distributed randomly. Loading data into a round-robin table is fast. Keep in mind that queries can require more data movement than the other distribution methods.

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

QUESTION 16
HOTSPOT

You develop a dataset named DBTBL1 by using Azure Databricks.

DBTBL1 contains the following columns:

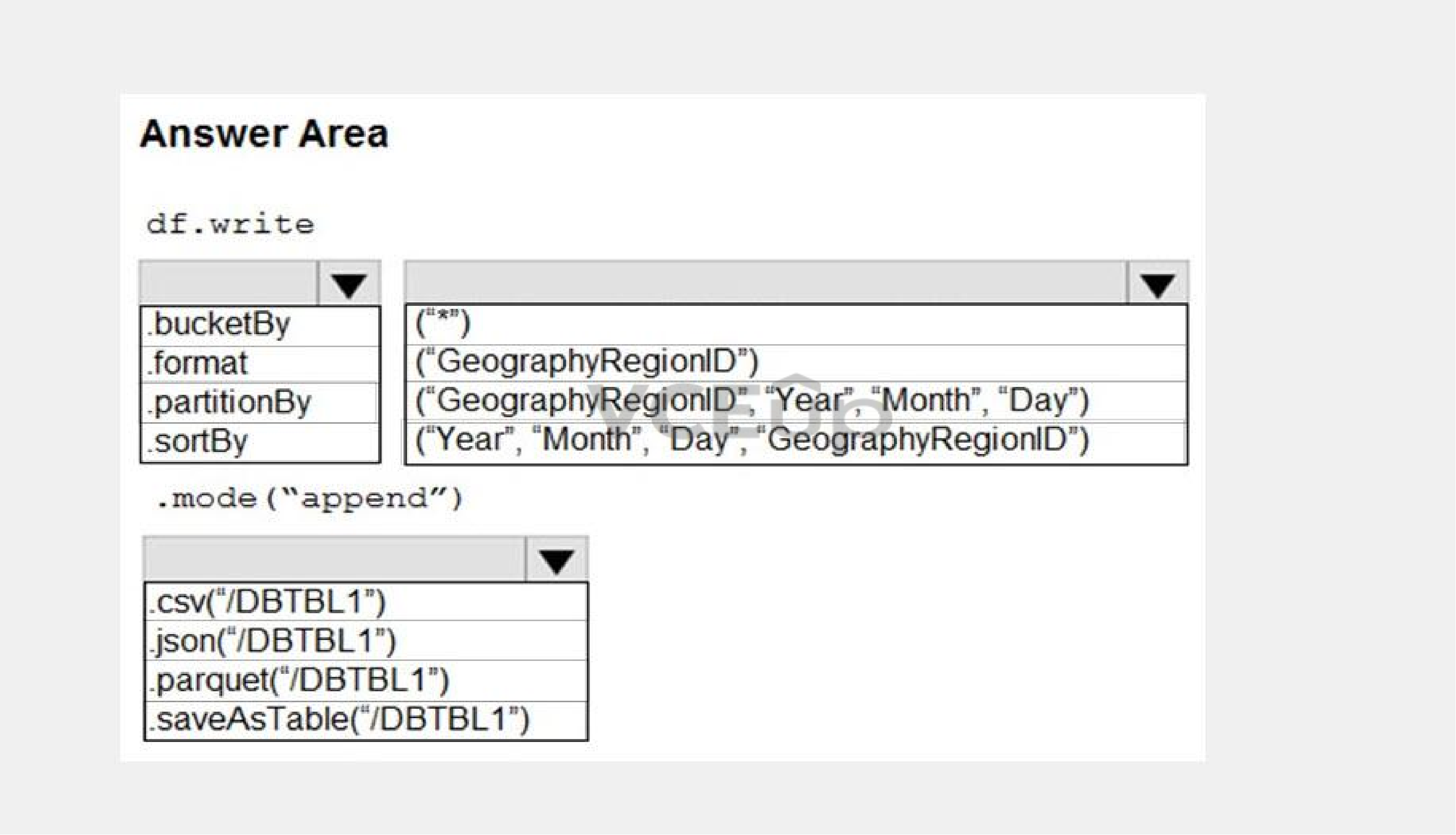
- SensorTypeID
- GeographyRegionID
- Year
- Month
- Day
- Hour
- Minute
- Temperature
- WindSpeed
- Other

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

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

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

```
df.write
```

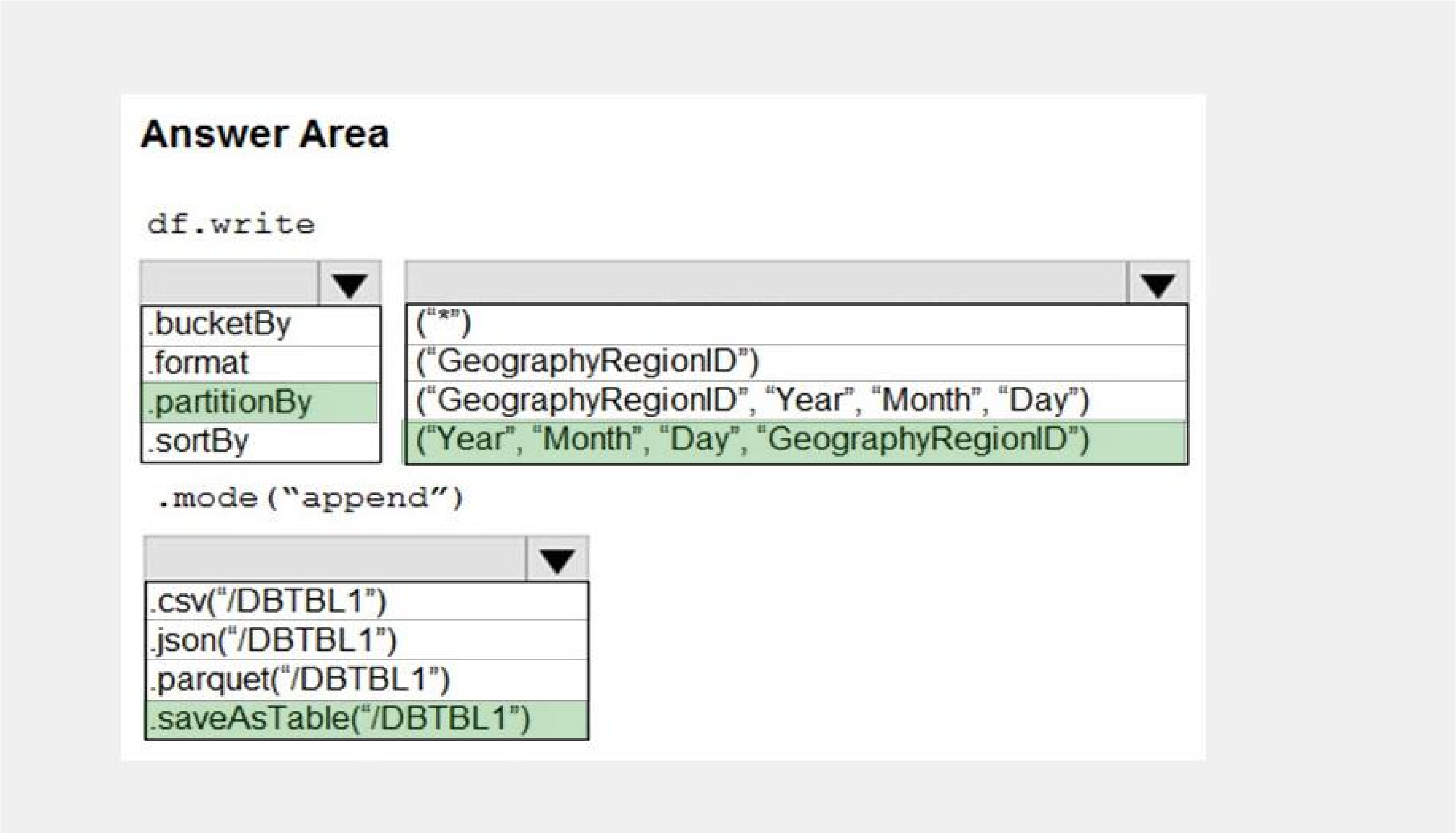
.bucketBy
.format
.partitionBy
.sortBy

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

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

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

Correct Answer:



Answer Area

```
df.write
```

- .bucketBy
- .format
- [✓ .partitionBy]
- .sortBy

- (“*”)
- (“GeographyRegionID”)
- (“GeographyRegionID”, “Year”, “Month”, “Day”)
- [✓ (“Year”, “Month”, “Day”, “GeographyRegionID”)]

```
.mode(“append”)
```

- .csv(“/DBTBL1”)
- .json(“/DBTBL1”)
- .parquet(“/DBTBL1”)
- [✓ .saveAsTable(“/DBTBL1”)]

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: .partitionBy

Incorrect Answers:
- .format:

Method: format():
Arguments: "parquet", "csv", "txt", "json", "jdbc", "orc", "avro", etc.

- .bucketBy:

Method: bucketBy()
Arguments: (numBuckets, col, col..., coln)
The number of buckets and names of columns to bucket by. Uses Hive’s bucketing scheme on a filesystem.

Box 2: ("Year", "Month", "Day","GeographyRegionID")
Specify the columns on which to do the partition. Use the date columns followed by the GeographyRegionID column.

Box 3: .saveAsTable("/DBTBL1")
Method: saveAsTable()
Argument: "table_name"
The table to save to.

Reference:
https://www.oreilly.com/library/view/learning-spark-2nd/9781492050032/ch04.html

https://docs.microsoft.com/en-us/azure/databricks/delta/delta-batch

01 - Monitor and optimize data storage and data processing

QUESTION 1
Case study

This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study
To display the first question in this case study, click the **Next** button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an **All Information** tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the **Question** button to return to the question.

Overview

Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.

Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.

Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.

Requirements

Business Goals

Litware wants to create a new analytics environment in Azure to meet the following requirements:

- See inventory levels across the stores. Data must be updated as close to real time as possible.
- Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.
- Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.

Technical Requirements

Litware identifies the following technical requirements:

- Minimize the number of different Azure services needed to achieve the business goals.
- Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.
- Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.
- Use Azure Active Directory (Azure AD) authentication whenever possible.
- Use the principle of least privilege when designing security.
- Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data Lake Storage once the data is no longer in use. Files that have a modified date that is older than 14 days must be removed.
- Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.
- Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.

Planned Environment

Litware plans to implement the following environment:

- The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.
- Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
- Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
- Daily inventory data comes from a Microsoft SQL server located on a private network.
- Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.
- Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.
- Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

A.

Correct Answer:
Section: (none)
Explanation

Explanation/Reference:

QUESTION 2
What should you do to improve high availability of the real-time data processing solution?

A. Deploy a High Concurrency Databricks cluster.
B. Deploy an Azure Stream Analytics job and use an Azure Automation runbook to check the status of the job and to start the job if it stops.
C. Set Data Lake Storage to use geo-redundant storage (GRS).
D. Deploy identical Azure Stream Analytics jobs to paired regions in Azure.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Guarantee Stream Analytics job reliability during service updates
Part of being a fully managed service is the capability to introduce new service functionality and improvements at a rapid pace. As a result, Stream Analytics can have a service update deploy on a weekly (or more frequent) basis. No matter how much testing is done there is still a risk that an existing, running job may break due to the introduction of a bug. If you are running mission critical jobs, these risks need to be avoided. You can reduce this risk by following Azure's paired region model.

Scenario: 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

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

02 - Monitor and optimize data storage and data processing

QUESTION 1
HOTSPOT

You have an Azure event hub named retailhub that has 16 partitions. Transactions are posted to retailhub. Each transaction includes the transaction ID, the individual line items, and the payment details. The transaction ID is used as the partition key.

You are designing an Azure Stream Analytics job to identify potentially fraudulent transactions at a retail store. The job will use retailhub as the input. The job will output the transaction ID, the individual line items, the payment details, a fraud score, and a fraud indicator.

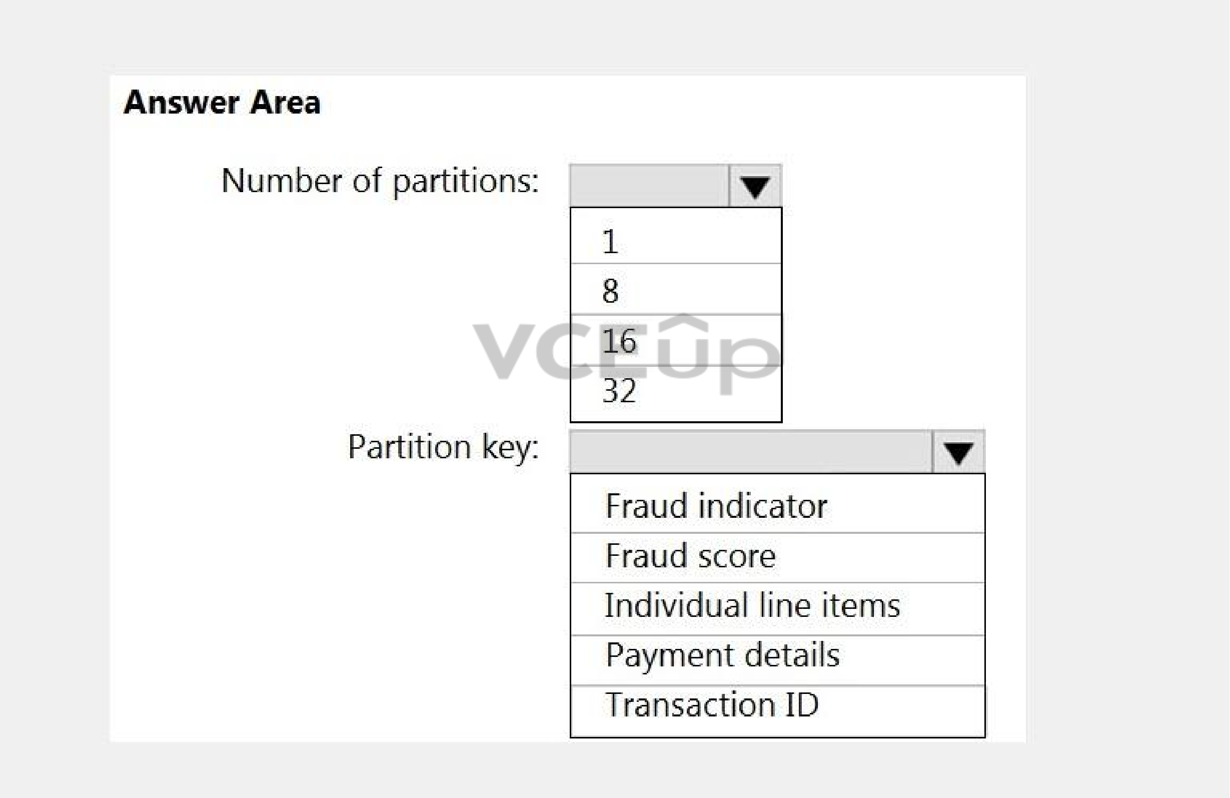
You plan to send the output to an Azure event hub named fraudhub.

You need to ensure that the fraud detection solution is highly scalable and processes transactions as quickly as possible.

How should you structure the output of the Stream Analytics job? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Answer Area

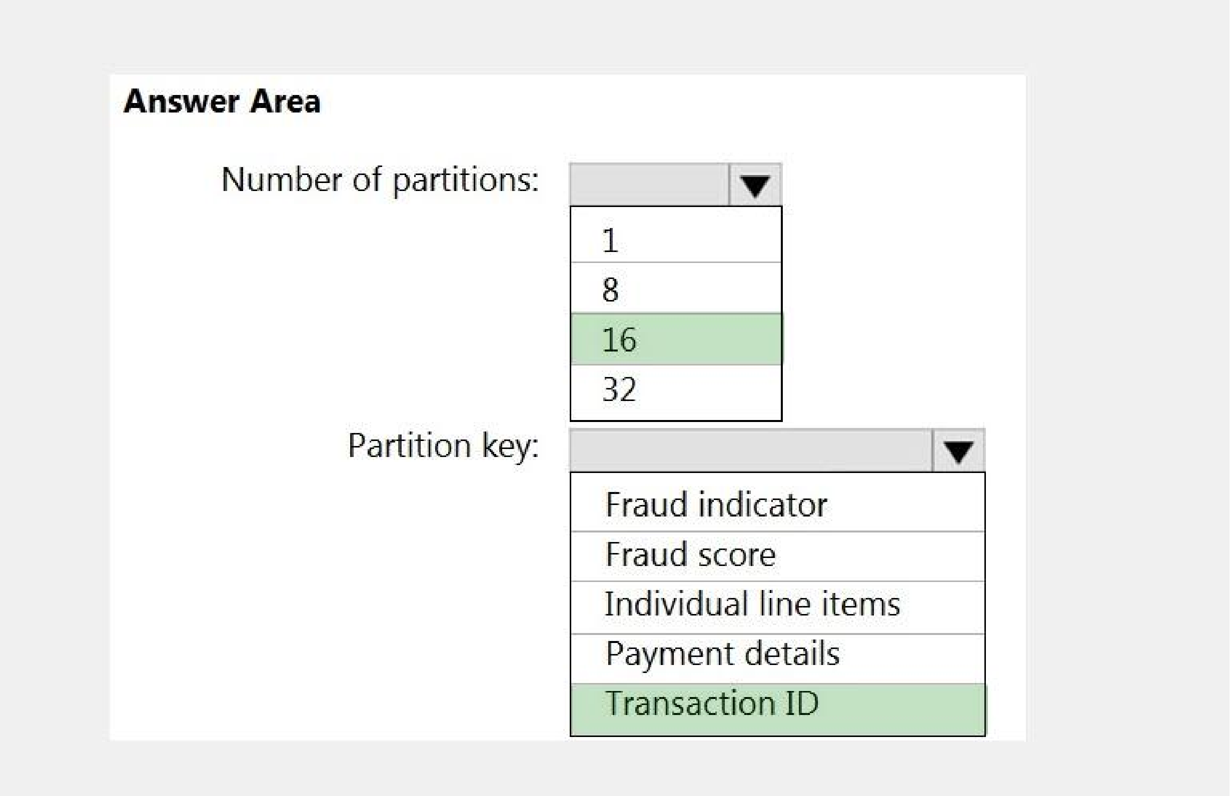
Number of partitions:

- 1
- 8
- 16
- 32

Partition key:

- Fraud indicator
- Fraud score
- Individual line items
- Payment details
- Transaction ID

Correct Answer:



Answer Area

Number of partitions:
- 1
- 8
- 16 🟩
- 32

Partition key:
- Fraud indicator
- Fraud score
- Individual line items
- Payment details
- Transaction ID 🟩

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: 16
For Event Hubs you need to set the partition key explicitly.

An embarrassingly parallel job is the most scalable scenario in Azure Stream Analytics. It connects one partition of the input to one instance of the query to one partition of the output.

Box 2: Transaction ID

Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features#partitions

QUESTION 2
HOTSPOT

You have an on-premises data warehouse that includes the following fact tables. Both tables have the following columns: DateKey, ProductKey, RegionKey. There are 120 unique product keys and 65 unique region keys.

Table	Comments
Sales	The table is 600 GB in size. DateKey is used extensively in the WHERE clause in queries. ProductKey is used extensively in join operations. RegionKey is used for grouping. Severity-five percent of records relate to one of 40 regions.
Invoice	The table is 6 GB in size. DateKey and ProductKey are used extensively in the WHERE clause in queries. RegionKey is used for grouping.

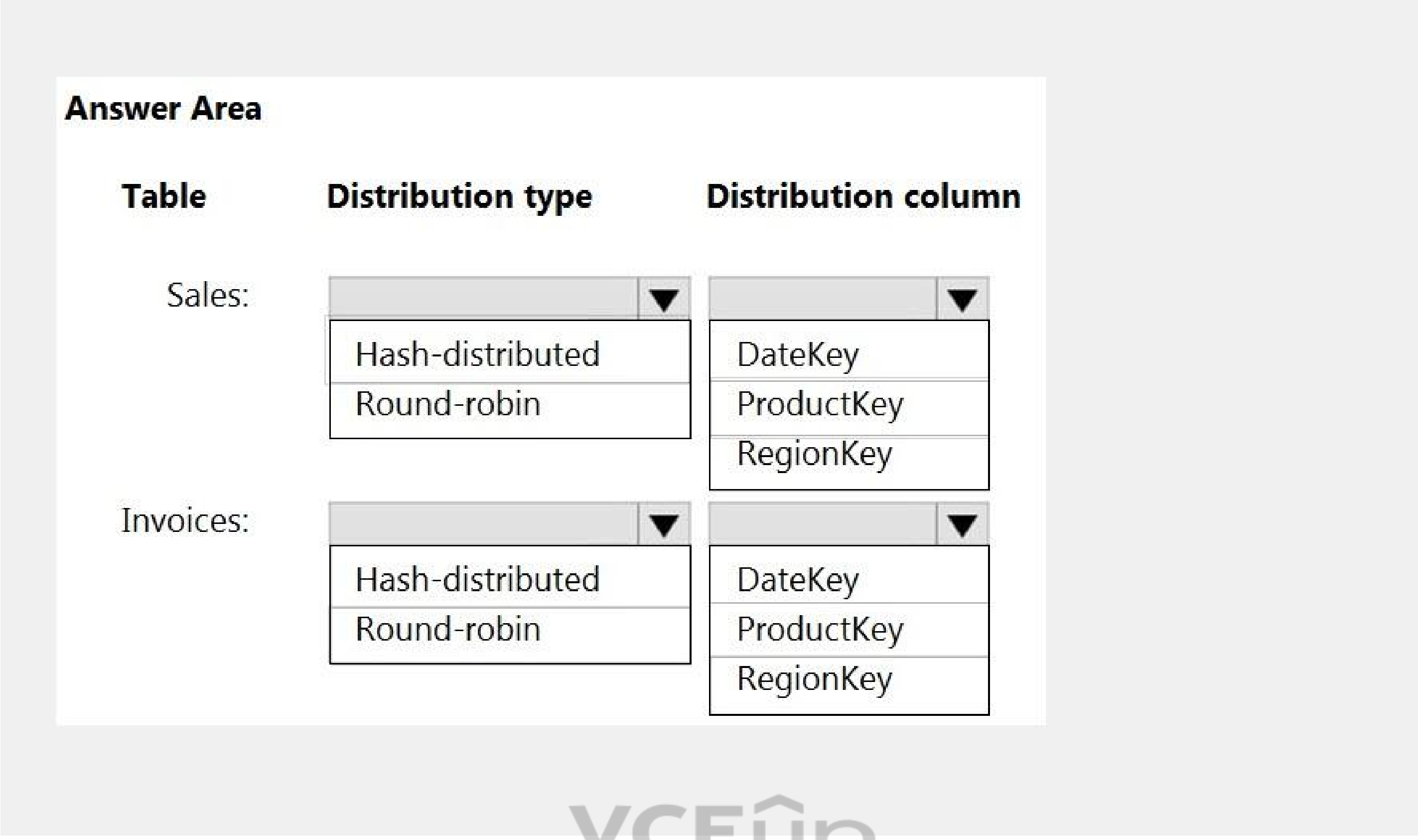
Queries that use the data warehouse take a long time to complete.

You plan to migrate the solution to use Azure Synapse Analytics. You need to ensure that the Azure-based solution optimizes query performance and minimizes processing skew.

What should you recommend? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point

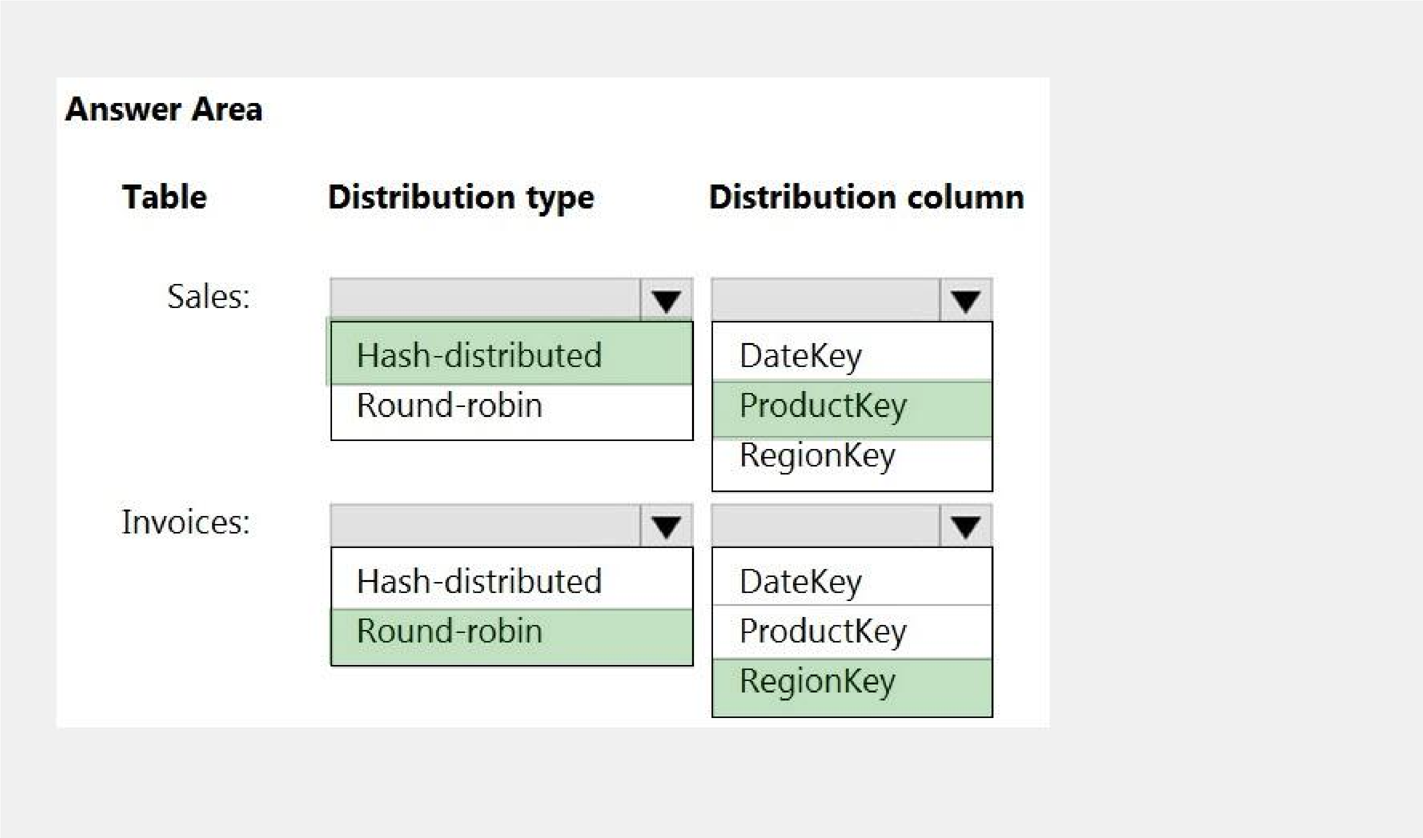
Hot Area:



Answer Area

Table	Distribution type	Distribution column
Sales:	Hash-distributed Round-robin	DateKey ProductKey RegionKey
Invoices:	Hash-distributed Round-robin	DateKey ProductKey RegionKey

Correct Answer:



Answer Area

Table	Distribution type	Distribution column
Sales:	[✓ Hash-distributed] Round-robin	DateKey [✓ ProductKey] RegionKey
Invoices:	Hash-distributed [✓ Round-robin]	DateKey ProductKey [✓ RegionKey]

Section: (none)
Explanation

Explanation/Reference:
Explanation:

Box 1: Hash-distributed

Box 2: ProductKey
ProductKey is used extensively in joins.
Hash-distributed tables improve query performance on large fact tables.

Box 3: Round-robin

Box 4: RegionKey
Round-robin tables are useful for improving loading speed.

Consider using the round-robin distribution for your table in the following scenarios:

- When getting started as a simple starting point since it is the default
- If there is no obvious joining key
- If there is not good candidate column for hash distributing the table
- If the table does not share a common join key with other tables
- If the join is less significant than other joins in the query
- When the table is a temporary staging table

Note: A distributed table appears as a single table, but the rows are actually stored across 60 distributions. The rows are distributed with a hash or round-robin algorithm.

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

QUESTION 3
You have a partitioned table in an Azure Synapse Analytics dedicated SQL pool.

You need to design queries to maximize the benefits of partition elimination.

What should you include in the Transact-SQL queries?

A. `JOIN`
B. `WHERE`
C. `DISTINCT`
D. `GROUP BY`

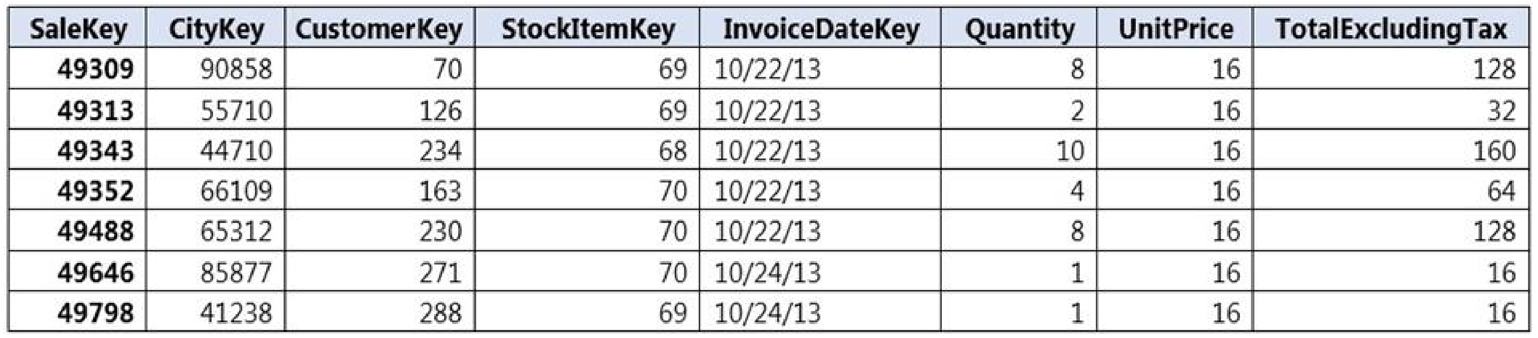
Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 4
You implement an enterprise data warehouse in Azure Synapse Analytics.

You have a large fact table that is 10 terabytes (TB) in size.

Incoming queries use the primary key SaleKey column to retrieve data as displayed in the following table:



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

You need to distribute the large fact table across multiple nodes to optimize performance of the table.

Which technology should you use?

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

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Hash-distributed tables improve query performance on large fact tables.

Columnstore indexes can achieve up to 100x better performance on analytics and data warehousing workloads and up to 10x better data compression than traditional rowstore indexes.

Incorrect Answers:
C, D: Round-robin tables are useful for improving loading speed.

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

https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-query-performance

QUESTION 5
You have an Azure Synapse Analytics dedicated SQL pool that contains a large fact table. The table contains 50 columns and 5 billion rows and is a heap.

Most queries against the table aggregate values from approximately 100 million rows and return only two columns.

You discover that the queries against the fact table are very slow.

Which type of index should you add to provide the fastest query times?

A. nonclustered columnstore
B. clustered columnstore
C. nonclustered
D. clustered

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Clustered columnstore indexes are one of the most efficient ways you can store your data in dedicated SQL pool.

Columnstore tables won't benefit a query unless the table has more than 60 million rows.

Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

QUESTION 6
You create an Azure Databricks cluster and specify an additional library to install.

When you attempt to load the library to a notebook, the library in not found.

You need to identify the cause of the issue.

What should you review?

A. notebook logs
B. cluster event logs
C. global init scripts logs
D. workspace logs

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Cluster-scoped Init Scripts: Init scripts are shell scripts that run during the startup of each cluster node before the Spark driver or worker JVM starts. Databricks customers use init scripts for various purposes such as installing custom libraries, launching background processes, or applying enterprise security policies.

Logs for Cluster-scoped init scripts are now more consistent with Cluster Log Delivery and can be found in the same root folder as driver and executor logs for the cluster.

Reference:
https://databricks.com/blog/2018/08/30/introducing-cluster-scoped-init-scripts.html

QUESTION 7
You have an Azure data factory.

You need to examine the pipeline failures from the last 60 days.

What should you use?

A. the Activity log blade for the Data Factory resource
B. the Monitor & Manage app in Data Factory
C. the Resource health blade for the Data Factory resource
D. Azure Monitor

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.

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

QUESTION 8
You are monitoring an Azure Stream Analytics job.

The Backlogged Input Events count has been 20 for the last hour.

You need to reduce the Backlogged Input Events count.

What should you do?

A. Drop late arriving events from the job.
B. Add an Azure Storage account to the job.
C. Increase the streaming units for the job.
D. Stop the job.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
General symptoms of the job hitting system resource limits include:
- If the backlog event metric keeps increasing, it's an indicator that the system resource is constrained (either because of output sink throttling, or high CPU).

Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job: adjust Streaming Units.

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

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

QUESTION 9
You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination.

You need to use that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs.

What should you do?

A. Pin the cluster.
B. Create an Azure runbook that starts the cluster every 90 days.
C. Terminate the cluster manually when processing completes.
D. Clone the cluster after it is terminated.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Azure Databricks retains cluster configuration information for up to 70 all-purpose clusters terminated in the last 30 days and up to 30 job clusters recently terminated by the job scheduler. To keep an all-purpose cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.

Reference:
https://docs.microsoft.com/en-us/azure/databricks/clusters/

QUESTION 10
You have an Azure data solution that contains an enterprise data warehouse in Azure Synapse Analytics named DW1.

Several users execute ad hoc queries to DW1 concurrently.

You regularly perform automated data loads to DW1.

You need to ensure that the automated data loads have enough memory available to complete quickly and successfully when the adhoc queries run.

What should you do?

A. Hash distribute the large fact tables in DW1 before performing the automated data loads.
B. Assign a smaller resource class to the automated data load queries.
C. Assign a larger resource class to the automated data load queries.
D. Create sampled statistics for every column in each table of DW1.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
The performance capacity of a query is determined by the user's resource class. Resource classes are pre-determined resource limits in Synapse SQL pool that govern compute resources and concurrency for query execution.
Resource classes can help you configure resources for your queries by setting limits on the number of queries that run concurrently and on the compute-resources assigned to each query. There's a trade-off between memory and concurrency.

Smaller resource classes reduce the maximum memory per query, but increase concurrency.
Larger resource classes increase the maximum memory per query, but reduce concurrency.

Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/resource-classes-for-workload-management

QUESTION 11
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.

You need to identify the extent of the data skew in Table1.

What should you do in Synapse Studio?

A. Connect to the built-in pool and run `DBCC PDW_SHOWSPACEUSED`.
B. Connect to the built-in pool and run `DBCC CHECKALLOC`.
C. Connect to Pool1 and query `sys.dm_pdw_node_status`.
D. Connect to Pool1 and query `sys.dm_pdw_nodes_db_partition_stats`.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
A quick way to check for data skew is to use DBCC PDW_SHOWSPACEUSED. The following SQL code returns the number of table rows that are stored in each of the 60 distributions. For balanced performance, the rows in your distributed table should be spread evenly across all the distributions.

DBCC PDW_SHOWSPACEUSED('dbo.FactInternetSales');

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

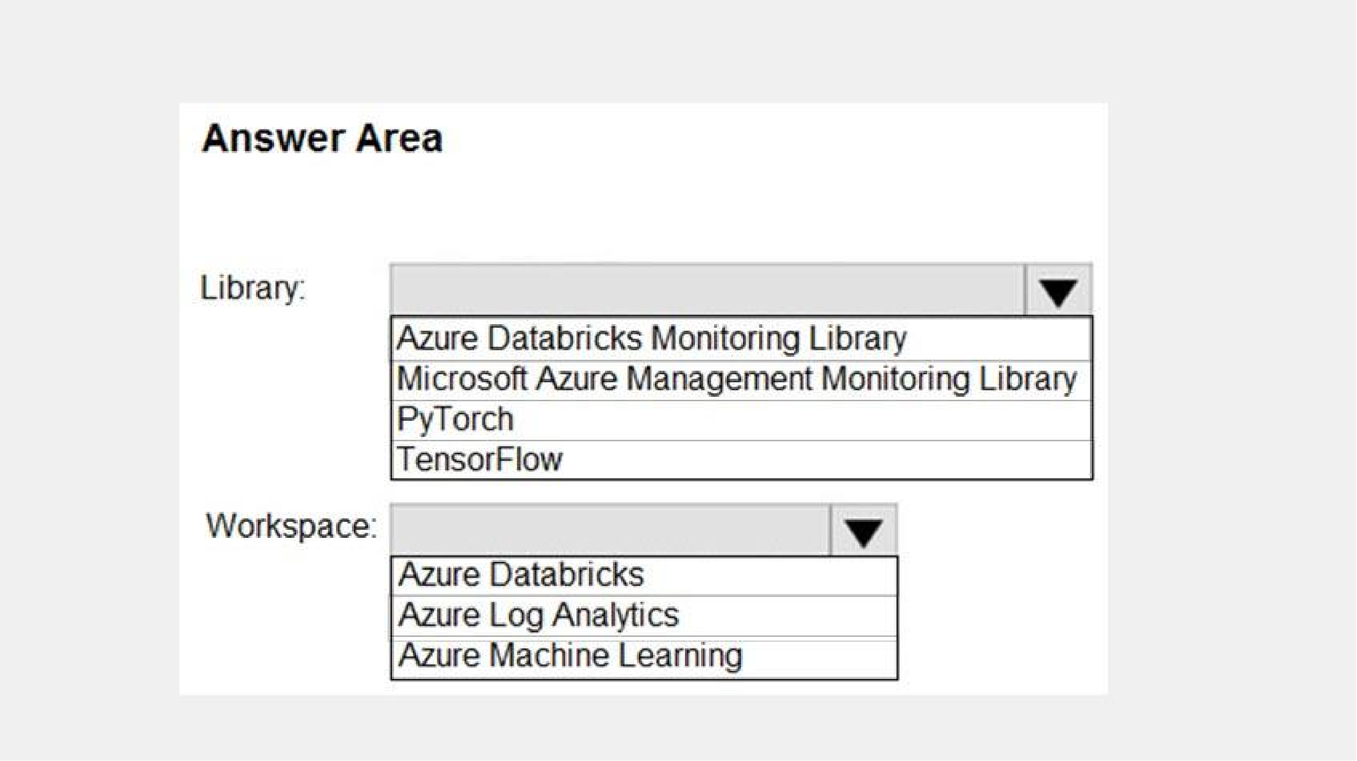
QUESTION 12
HOTSPOT

You need to collect application metrics, streaming query events, and application log messages for an Azure Databrick cluster.

Which type of library and workspace should you implement? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

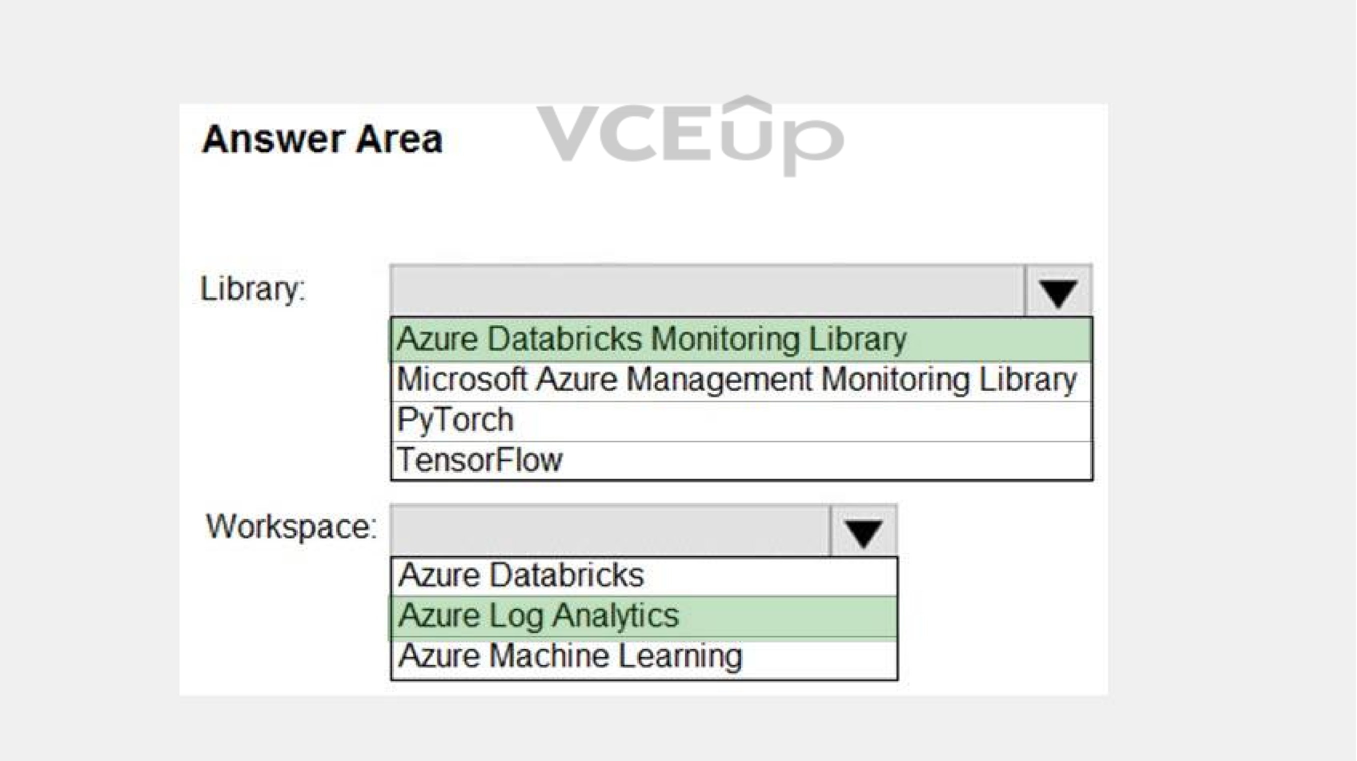


Answer Area

Library:
- Azure Databricks Monitoring Library
- Microsoft Azure Management Monitoring Library
- PyTorch
- TensorFlow

Workspace:
- Azure Databricks
- Azure Log Analytics
- Azure Machine Learning

Correct Answer:



Answer Area

Library:
- [✓ Azure Databricks Monitoring Library]
- Microsoft Azure Management Monitoring Library
- PyTorch
- TensorFlow

Workspace:
- Azure Databricks
- [✓ Azure Log Analytics]
- Azure Machine Learning

Section: (none)
Explanation

Explanation/Reference:
Explanation:

You can send application logs and metrics from Azure Databricks to a Log Analytics workspace. It uses the Azure Databricks Monitoring Library, which is available on GitHub.

Reference:
https://docs.microsoft.com/en-us/azure/architecture/databricks-monitoring/application-logs

QUESTION 13
You have a SQL pool in Azure Synapse.

You discover that some queries fail or take a long time to complete.

You need to monitor for transactions that have rolled back.

Which dynamic management view should you query?

A. `sys.dm_pdw_request_steps`
B. `sys.dm_pdw_nodes_tran_database_transactions`
C. `sys.dm_pdw_waits`
D. `sys.dm_pdw_exec_sessions`

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
You can use Dynamic Management Views (DMVs) to monitor your workload including investigating query execution in SQL pool.

If your queries are failing or taking a long time to proceed, you can check and monitor if you have any transactions rolling back.

Example:

```
-- Monitor rollback
SELECT
    SUM(CASE WHEN t.database_transaction_next_undo_lsn IS NOT NULL THEN 1 ELSE 0 END),
    t.pdw_node_id,
    nod.[type]
FROM sys.dm_pdw_nodes_tran_database_transactions t
JOIN sys.dm_pdw_nodes nod ON t.pdw_node_id = nod.pdw_node_id
GROUP BY t.pdw_node_id, nod.[type]
```

Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-manage-monitor#monitor-transaction-log-rollback

QUESTION 14
You are monitoring an Azure Stream Analytics job.

You discover that the Backlogged Input Events metric is increasing slowly and is consistently non-zero.

You need to ensure that the job can handle all the events.

What should you do?

A. Change the compatibility level of the Stream Analytics job.
B. Increase the number of streaming units (SUs).
C. Remove any named consumer groups from the connection and use $default.
D. Create an additional output stream for the existing input stream.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job. You should increase the Streaming Units.

Note: 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.

Reference:
https://docs.microsoft.com/bs-cyrl-ba/azure/stream-analytics/stream-analytics-monitoring

QUESTION 15
You are designing an inventory updates table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:

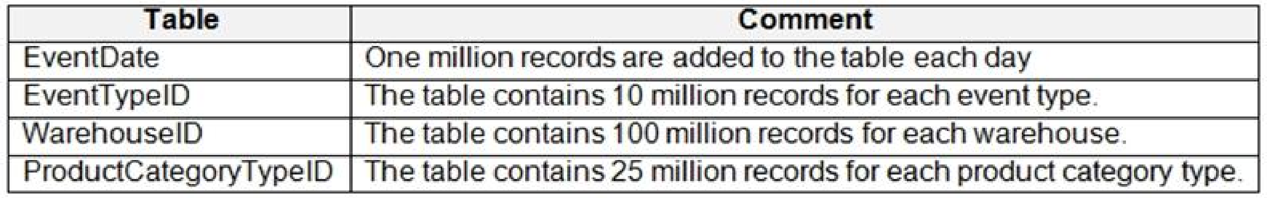


Table	Comment
EventDate	One million records are added to the table each day
EventTypeID	The table contains 10 million records for each event type.
WarehouseID	The table contains 100 million records for each warehouse.
ProductCategoryTypeID	The table contains 25 million records for each product category type.

You identify the following usage patterns:

- Analysts will most commonly analyze transactions for a warehouse.
- Queries will summarize by product category type, date, and/or inventory event type.

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

On which column should you partition the table?

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

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
The number of records for each warehouse is big enough for a good partitioning.

Note: Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.

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 distributed databases.

QUESTION 16
You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date.

You need to ensure that the design provides the fastest query times of the records when querying for arbitrary date ranges and aggregating by fiscal calendar attributes.

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

NOTE: Each correct selection is worth one point.

A. Create a date dimension table that has a DateTime key.
B. Use built-in SQL functions to extract date attributes.
C. Create a date dimension table that has an integer key in the format of `YYYYMMDD`.
D. In the fact table, use integer columns for the date fields.
E. Use DateTime columns for the date fields.

Correct Answer: BD
Section: (none)
Explanation

Explanation/Reference: