# Microsoft.DP-203.v2023-08-12.q133

| | |
|---|---|
| **Exam Code:** | DP-203 |
| **Exam Name:** | Data Engineering on Microsoft Azure |
| **Certification Provider:** | Microsoft |
| **Free Question Number:** | 133 |
| **Version:** | v2023-08-12 |
| **# of views:** | 1403 |
| **# of Questions views:** | 1330 |
| https://www.freepdfdumps.com/Microsoft.DP-203.v2023-08-12.q133.html | |

**NEW QUESTION: 1**

You have an Azure subscription.

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

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


**Answer:**


Explanation

**NEW QUESTION: 2**
You need to integrate the on-premises data sources and Azure Synapse Analytics. The solution must meet the data integration requirements.
Which type of integration runtime should you use?
**A.** Azure-SSIS integration runtime
**B.** self-hosted integration runtime
**C.** Azure integration runtime
**Answer: C (LEAVE A REPLY)**

**NEW QUESTION: 3**
You have an Azure Synapse Analytics dedicated SQL pool mat contains a table named dbo.Users.
You need to prevent a group of users from reading user email addresses from dbo.Users. What should you use?
**A.** row-level security
**B.** Dynamic data masking
**C.** column-level security
**D.** Transparent Data Encryption (TDD
**Answer: (SHOW ANSWER)**

**NEW QUESTION: 4**
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.** ProductCategoryTypeID
**B.** EventDate
**C.** WarehouseID
**D.** EventTypeID
**Answer: C (LEAVE A REPLY)**
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.

**NEW QUESTION: 5**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values.
75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly.
Solution: You convert the files to compressed delimited text files.
Does this meet the goal?
**A.** Yes
**B.** No
**Answer: (SHOW ANSWER)**
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

**NEW QUESTION: 6**
You are designing the folder structure for an Azure Data Lake Storage Gen2 account.
You identify the following usage patterns:
* Users will query data by using Azure Synapse Analytics serverless SQL pools and Azure Synapse Analytics serverless Apache Spark pods.
* Most queries will include a filter on the current year or week.
* Data will be secured by data source.
You need to recommend a folder structure that meets the following requirements:
* Supports the usage patterns
* Simplifies folder security
* Minimizes query times
Which folder structure should you recommend?

A. `\YYYYY\WW\DataSource\SubjectArea\FileData_YYYY_MM_DD.parquet`

B. `DataSource\SubjectArea\WW\YYYY\FileData_YYYY_MM_DD.parquet`

C. `\DataSource\SubjectArea\YYYY\WW\FileData_YYYY_MM_DD.parquet`

D. `\DataSource\SubjectArea\YYYY-WW\FileData_YYYY_MM_DD.parquet`

E. `WW\YYYY\SubjectArea\DataSource\FileData_YYYY_MM_DD.parquet`

**Answer: C (LEAVE A REPLY)**

Explanation

Data will be secured by data source. -> Use DataSource as top folder.

Most queries will include a filter on the current year or week -> Use \YYYY\WW\ as subfolders.

Common Use Cases

A common use case is to filter data stored in a date (and possibly time) folder structure such as /YYYY/MM/DD/ or /YYYY/MM/YYYY-MM-DD/. As new data is generated/sent/copied/moved to the storage account, a new folder is created for each specific time period. This strategy organises data into a maintainable folder structure.

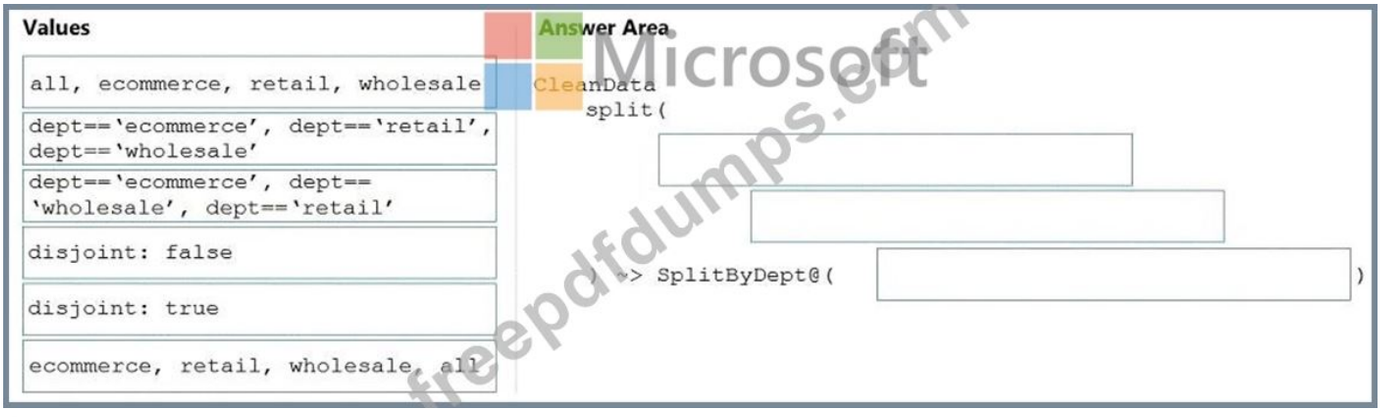
Reference: https://www.serverlesssql.com/optimisation/azurestoragefilteringusingfilepath/

**NEW QUESTION: 7**

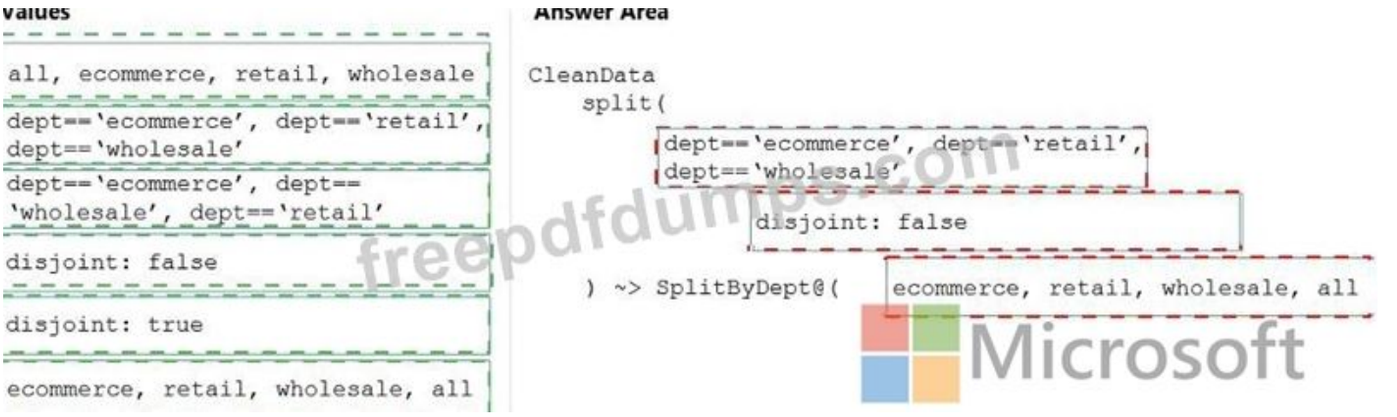
You need to create an Azure Data Factory pipeline to process data for the following three departments at your company: Ecommerce, retail, and wholesale. The solution must ensure that data can also be processed for the entire company.

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.




**Answer:**




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

**NEW QUESTION: 8**

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values.
75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an Azure SQL data warehouse.
You need to prepare the files to ensure that the data copies quickly.
Solution: You modify the files to ensure that each row is less than 1 MB.
Does this meet the goal?
**A.** Yes
**B.** No
**Answer: A (LEAVE A REPLY)**
Explanation
When exporting data into an ORC File Format, you might get Java out-of-memory errors when there are large text columns. To work around this limitation, export only a subset of the columns.
References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

**NEW QUESTION: 9**
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 query sysdm_pdw_sys_info.
**B.** Connect to Pool1 and run DBCC CHECKALLOC.
**C.** Connect to the built-in pool and run DBCC CHECKALLOC.
**D.** Connect to Pool! and query sys.dm_pdw_nodes_db_partition_stats.
**Answer: D (LEAVE A REPLY)**
Explanation
Microsoft recommends use of sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet

**NEW QUESTION: 10**
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.

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

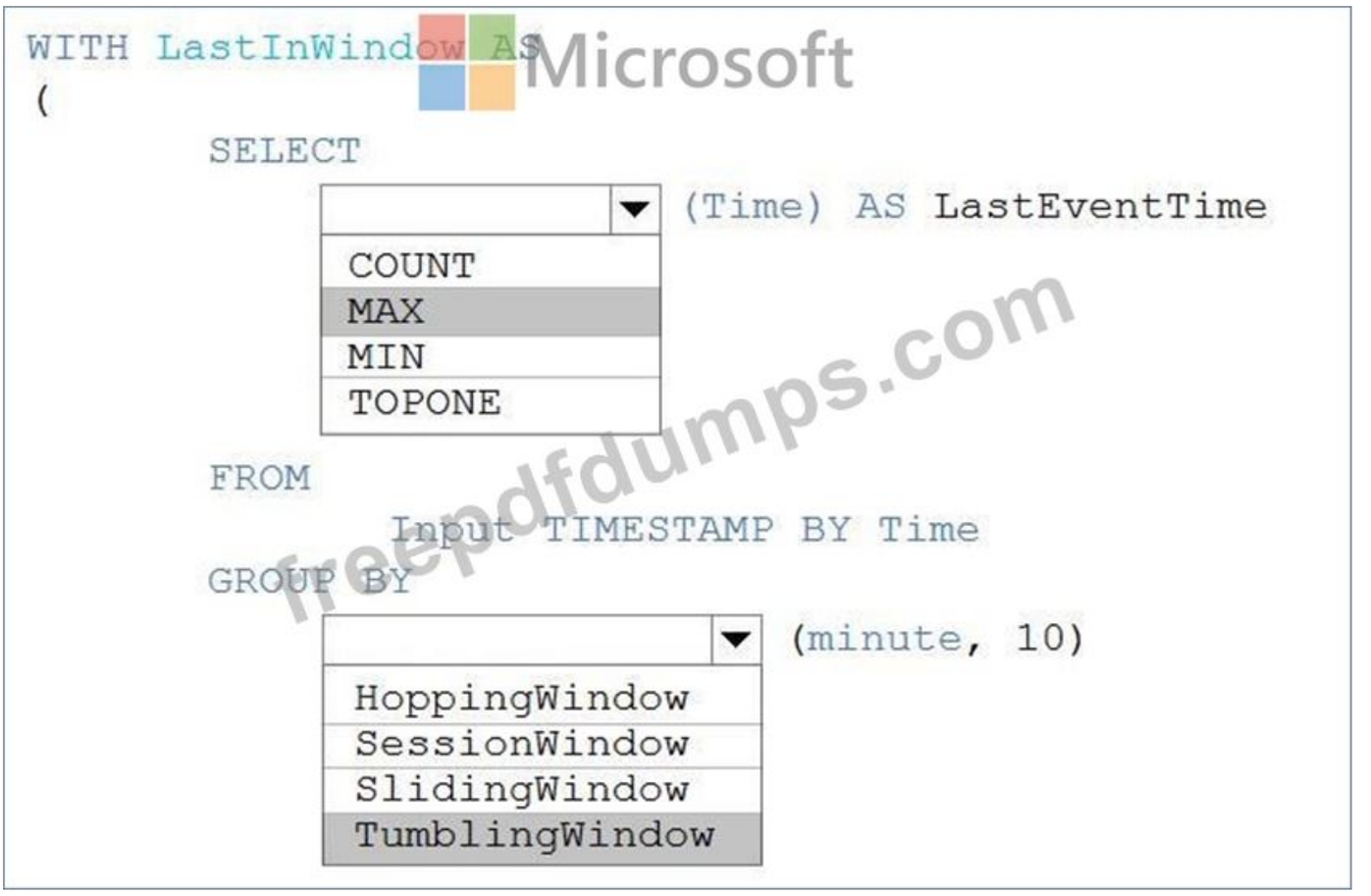


**Answer:**

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



Explanation




Graphical user interface, text, application Description automatically generated

```
)
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
```

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

**NEW QUESTION: 11**
You plan to implement an Azure Data Lake Gen2 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.** zone-redundant storage (ZRS)
**C.** locally-redundant storage (LRS)
**D.** geo-zone-redundant storage (GZRS)
**Answer: C (LEAVE A REPLY)**
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

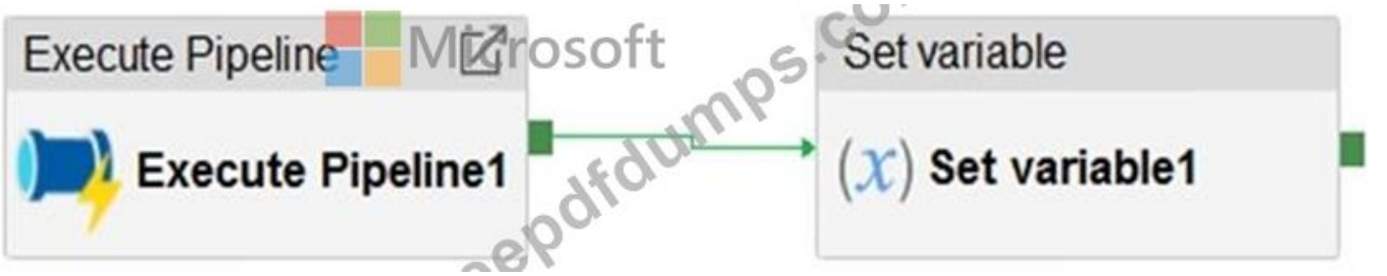
**NEW QUESTION: 12**
You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date.

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 DateTime columns for the date fields.
**C.** Use built-in SQL functions to extract date attributes.
**D.** In the fact table, use integer columns for the date fields.
**E.** Create a date dimension table that has an integer key in the format of yyyymmdd.
**Answer: (SHOW ANSWER)**

**NEW QUESTION: 13**
You are designing an anomaly detection solution for streaming data from an Azure IoT hub. The solution must meet the following requirements:
* Send the output to Azure Synapse.
* Identify spikes and dips in time series data.
* Minimize development and configuration effort.
Which should you include in the solution?
**A.** Azure Databricks
**B.** Azure Stream Analytics
**C.** Azure SQL Database
**Answer: (SHOW ANSWER)**
Explanation
You can identify anomalies by routing data via IoT Hub to a built-in ML model in Azure Stream Analytics.
Reference:
https://docs.microsoft.com/en-us/learn/modules/data-anomaly-detection-using-azure-iot-hub/

**NEW QUESTION: 14**
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.

**Answer: A (LEAVE A REPLY)**

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.

Waterfall chart Description automatically generated with medium confidence


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/

**NEW QUESTION: 15**

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_scacus.
**D.** Connect to Pool1 and query sys.dm_pdw_nodes_db_partition_scacs.
**Answer: A (LEAVE A REPLY)**
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-distribu

**NEW QUESTION: 16**
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;
```



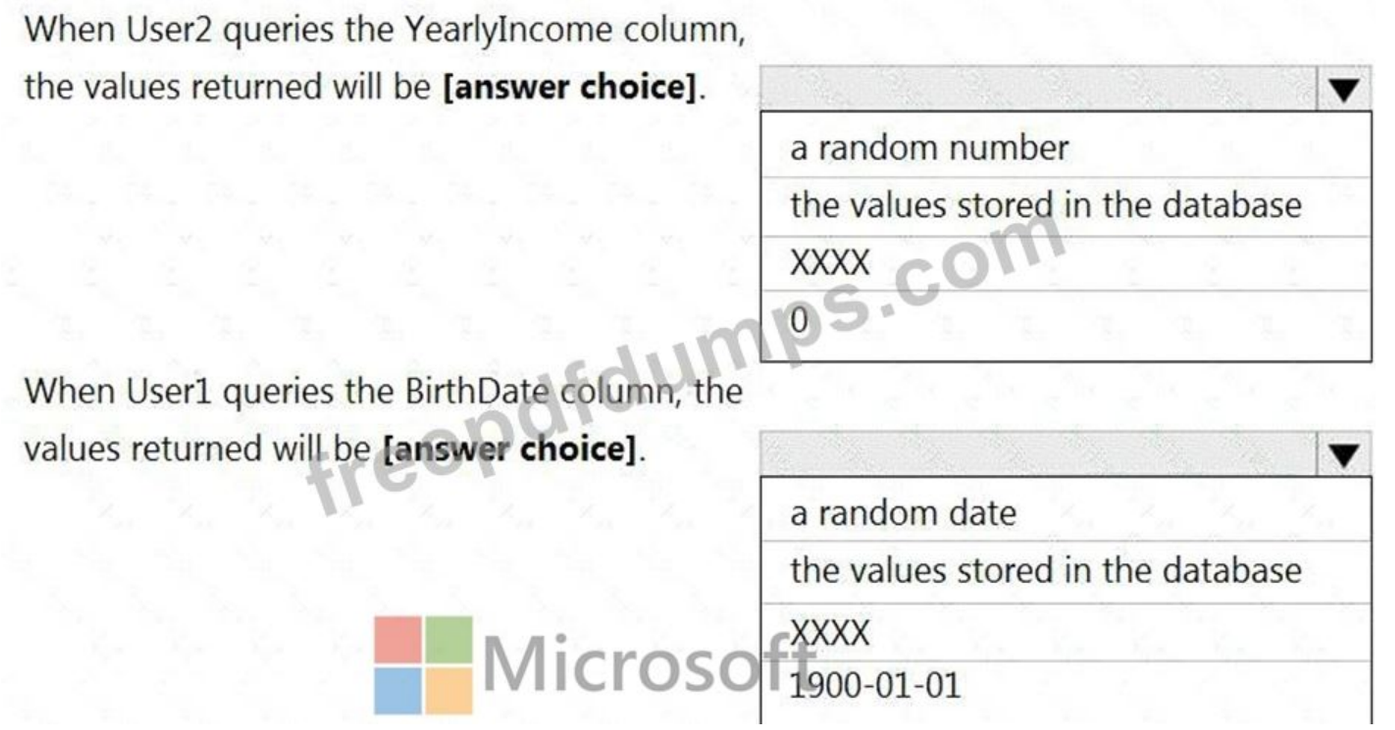
Results Messages

| | 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() |

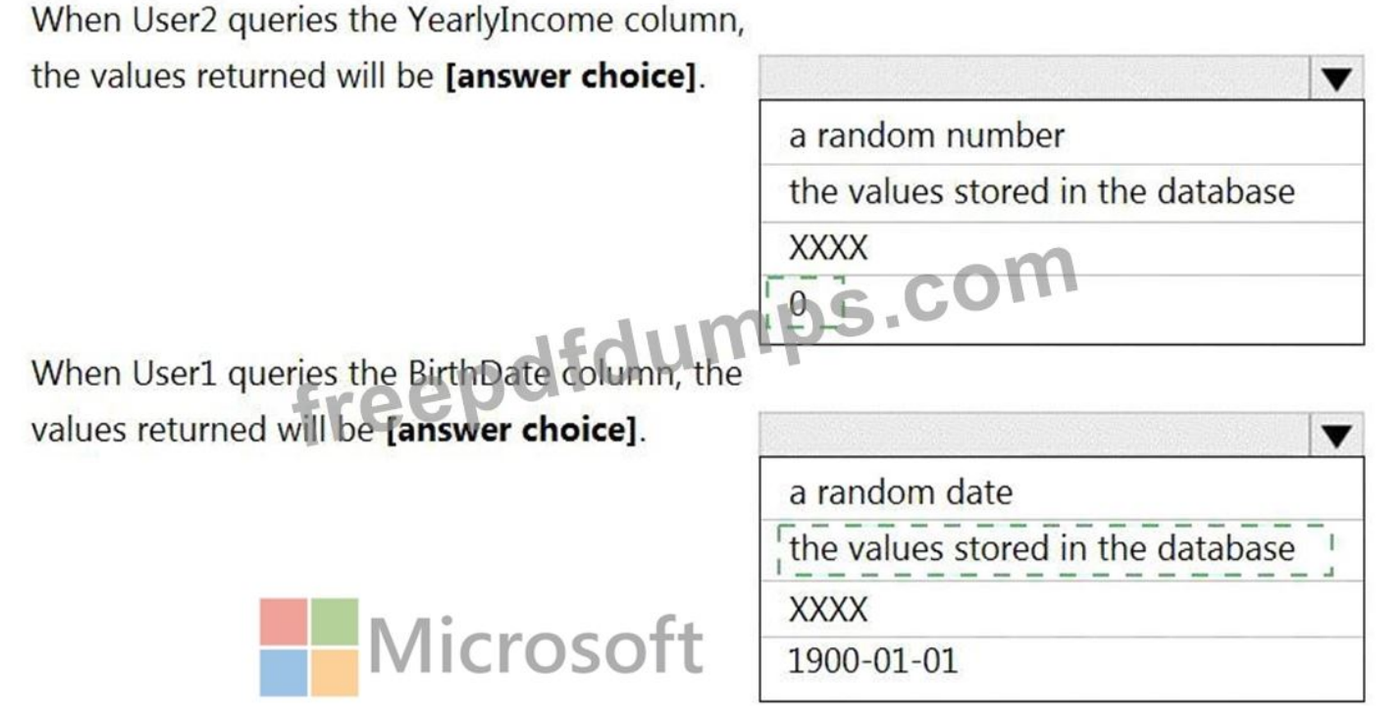


User1 is the only user who has access to the unmasked data.

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.

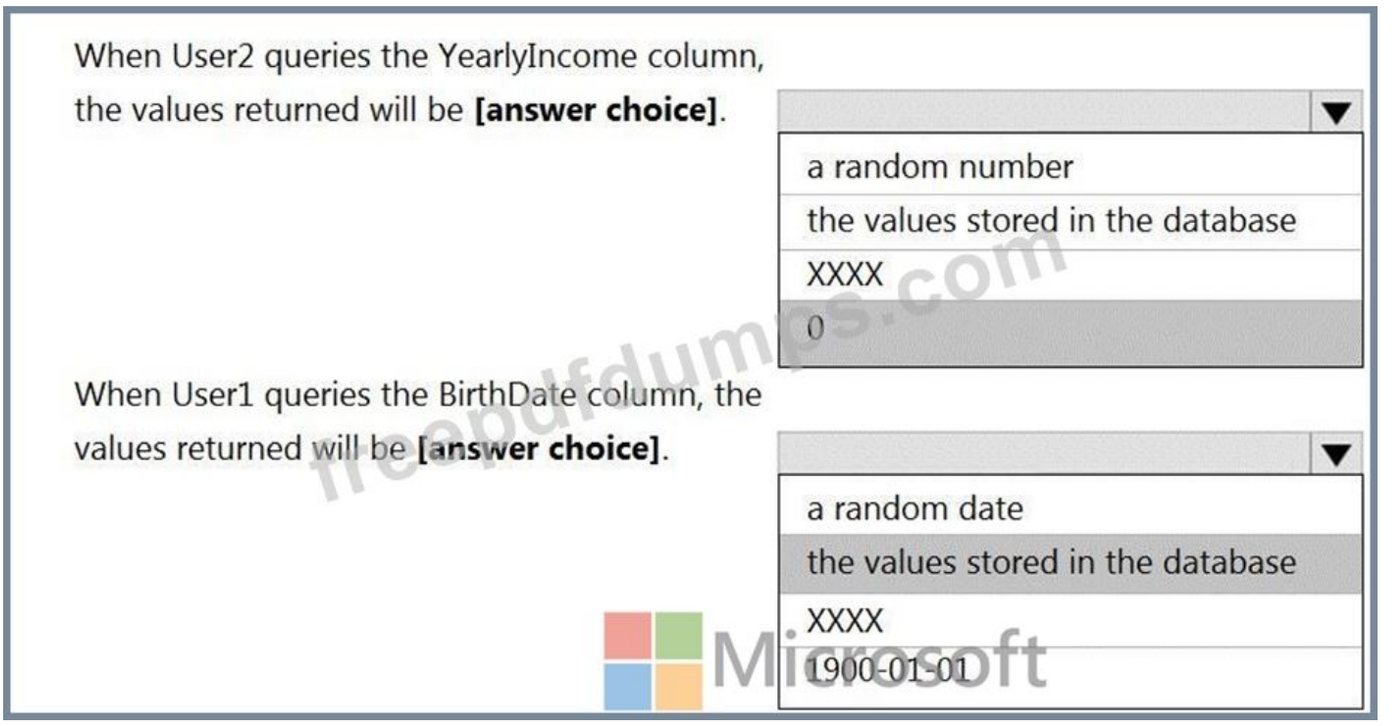



**Answer:**




Explanation

Graphical user interface, text, application, email Description automatically generated




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



**NEW QUESTION: 17**

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.
**Answer: (SHOW ANSWER)**
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

**NEW QUESTION: 18**
You build a data warehouse in an Azure Synapse Analytics dedicated SQL pool.
Analysts write a complex SELECT query that contains multiple JOIN and CASE statements to transform data for use in inventory reports. The inventory reports will use the data and additional WHERE parameters depending on the report. The reports will be produced once daily.
You need to implement a solution to make the dataset available for the reports. The solution must minimize query times.
What should you implement?
**A.** a materialized view
**B.** a replicated table
**C.** in ordered clustered columnstore index
**D.** result set chaching
**Answer: A (LEAVE A REPLY)**
Explanation
Materialized views for dedicated SQL pools in Azure Synapse provide a low maintenance method for complex analytical queries to get fast performance without any query change.
Note: When result set caching is enabled, dedicated SQL pool automatically caches query results in the user database for repetitive use. This allows subsequent query executions to get results directly from the persisted cache so recomputation is not needed. Result set caching improves query performance and reduces compute resource usage. In addition, queries using cached

results set do not use any concurrency slots and thus do not count against existing concurrency limits.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-materialized-v
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-cach

**NEW QUESTION: 19**
You are implementing a star schema in an Azure Synapse Analytics dedicated SQL pool.
You plan to create a table named DimProduct.
DimProduct must be a Type 3 slowly changing dimension (SCO) table that meets the following requirements:
* The values in two columns named ProductKey and ProductSourceID will remain the same.
* The values in three columns named ProductName, ProductDescription, and Color can change.
You need to add additional columns to complete the following table definition.

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



**A.** `[IsCurrentRow] [bit] NOT NULL`

**B.** `[OriginalProductName] NVARCHAR(100) NULL`

**C.** `[EffectiveEndDate] [datetime] NOT NULL`

**D.** `[OriginalProductDescription] NVARCHAR(2000) NOT NULL`

**E.** `[EffectiveStartDate] [datetime] NOT NULL`

**F.** `[OriginalColor] NVARCHAR(50) NOT NULL`

**Answer: A,D,E (LEAVE A REPLY)**

**NEW QUESTION: 20**
You have an Azure Synapse Analytics dedicated SQL pool mat contains a table named dbo.Users.
You need to prevent a group of users from reading user email addresses from dbo.Users. What should you use?
**A.** row-level security
**B.** Dynamic data masking
**C.** column-level security
**D.** Transparent Data Encryption (TDD
**Answer: C (LEAVE A REPLY)**

**NEW QUESTION: 21**
You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account.
You need to output the count of tweets during the last five minutes every five minutes. Each tweet must only be counted once.
Which windowing function should you use?
**A.** a five-minute Session window
**B.** a five-minute Sliding window
**C.** a five-minute Tumbling window
**D.** a five-minute Hopping window that has one-minute hop
**Answer: (SHOW ANSWER)**
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.
References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

**NEW QUESTION: 22**
You have an Azure Synapse Analytics dedicated SQL pool named SA1 that contains a table named Table1.
You need to identify tables that have a high percentage of deleted rows. What should you run?
**A.** `sys.pdw_nodes_column_store_row_groups`
**B.** `sys.dm_db_column_store_row_group_physical_stats`
**C.** `sys.dm_db_column_store_row_group_operational_stats`
**D.** `sys.pdw_nodes_column_store_segments`
**Answer: C (LEAVE A REPLY)**

**NEW QUESTION: 23**
You are performing exploratory analysis of the bus fare data in an Azure Data Lake Storage Gen2 account by using an Azure Synapse Analytics serverless SQL pool.
You execute the Transact-SQL query shown in the following exhibit.

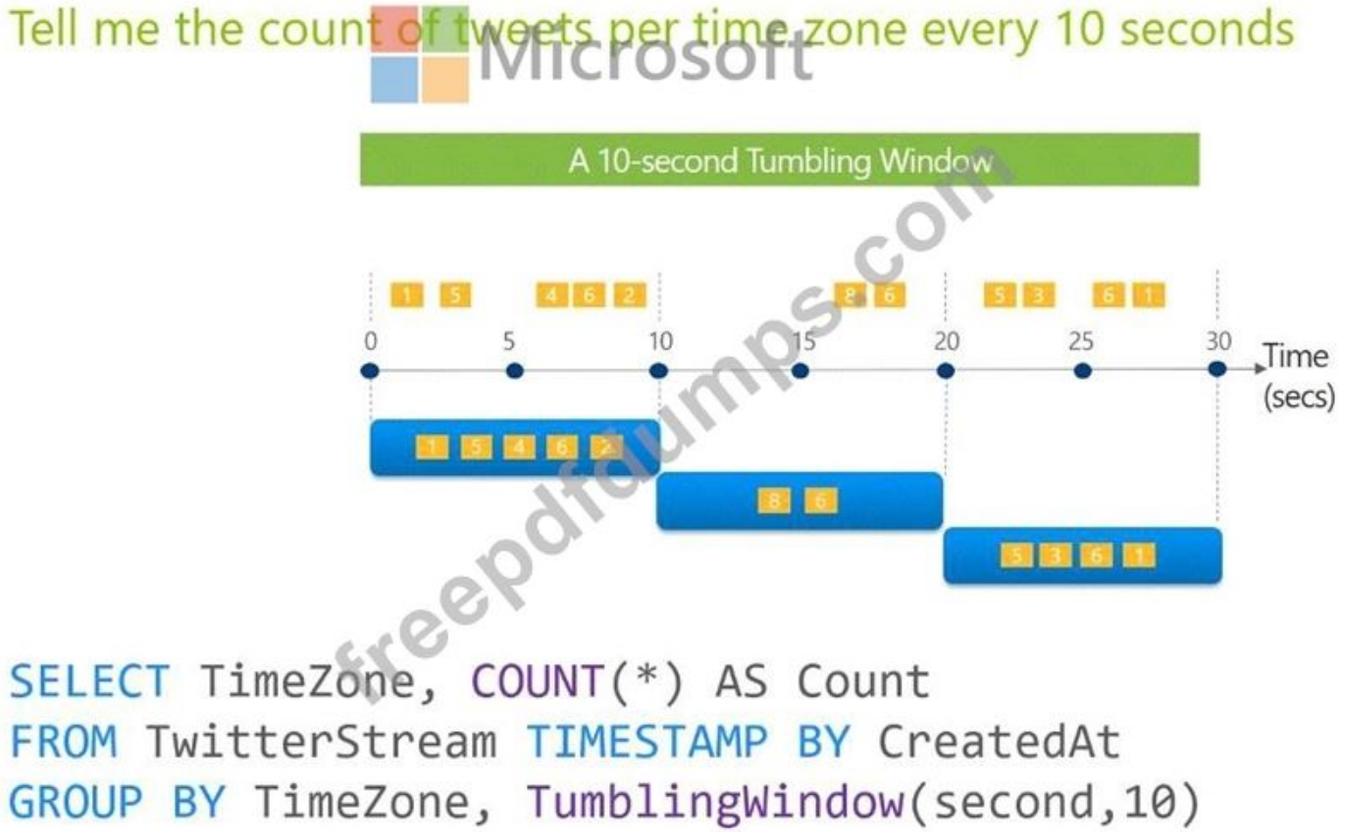
```
SELECT
    payment_type,
    SUM(fare_amount) AS fare_total
FROM OPENROWSET(
            BULK 'csv/busfare/tripdata_2020*.csv',
            DATA_SOURCE = 'BusData',
            FORMAT = 'CSV', PARSER_VERSION = '2.0',
            FIRSTROW = 2
    )
    WITH (
            payment_type INT 10,
    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



What do the query results include?
**A.** Only CSV that have file names that beginning with "tripdata_2020".
**B.** All files that have file names that beginning with "tripdata_2020".
**C.** Only CSV files in the tripdata_2020 subfolder.
**D.** All CSV files that have file names that contain "tripdata_2020".
**Answer: (SHOW ANSWER)**

**NEW QUESTION: 24**
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
**Answer: (SHOW ANSWER)**
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.





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

**NEW QUESTION: 25**
You have an Azure Synapse Analytics dedicated SQL Pool1. Pool1 contains a partitioned fact table named dbo.Sales and a staging table named stg.Sales that has the matching table and partition definitions.
You need to overwrite the content of the first partition in dbo.Sales with the content of the same partition in stg.Sales. The solution must minimize load times.
What should you do?
**A.** Insert the data from stg.Sales into dbo.Sales.
**B.** Switch the first partition from stg.Sales to dbo. Sales.
**C.** Switch the first partition from dbo.Sales to stg.Sales.
**D.** Update dbo.Sales from stg.Sales.
**Answer: C (LEAVE A REPLY)**

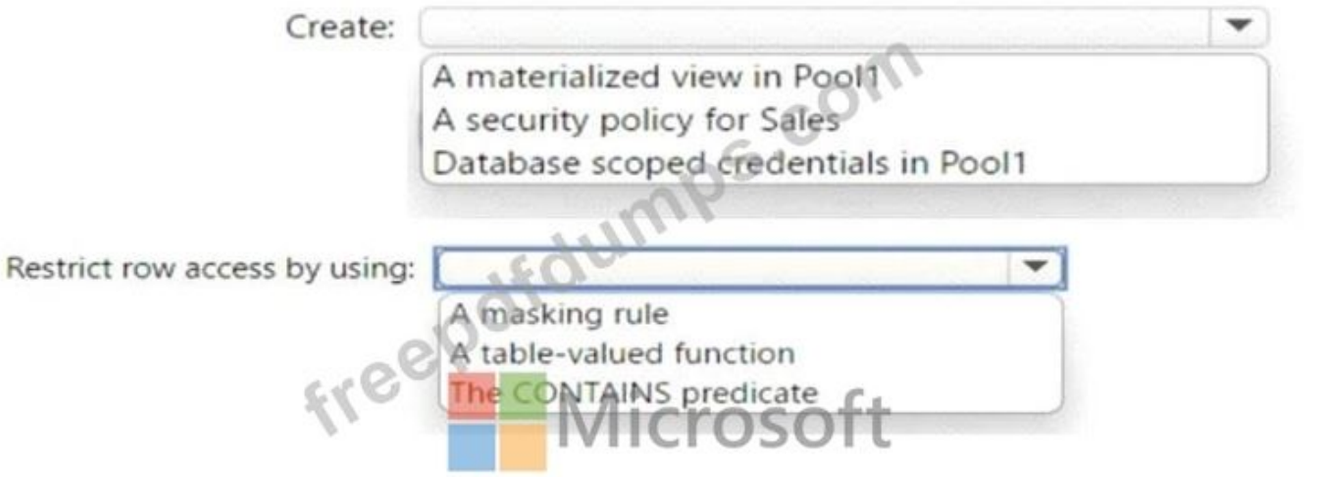
**NEW QUESTION: 26**
You have an Azure Data Factory pipeline named pipeline1 that is invoked by a tumbling window trigger named Trigger1. Trigger1 has a recurrence of 60 minutes.

You need to ensure that pipeline1 will execute only if the previous execution completes successfully.
How should you configure the self-dependency for Trigger1?
**A.** offset: "-00:01:00" size: "00:01:00"
**B.** offset: "01:00:00" size: "-01:00:00"
**C.** offset: "01:00:00" size: "01:00:00"
**D.** offset: "-01:00:00" size: "01:00:00"
**Answer: D (LEAVE A REPLY)**
Explanation
Tumbling window self-dependency properties
In scenarios where the trigger shouldn't proceed to the next window until the preceding window is successfully completed, build a self-dependency. A self-dependency trigger that's dependent on the success of earlier runs of itself within the preceding hour will have the properties indicated in the following code.
Example code:

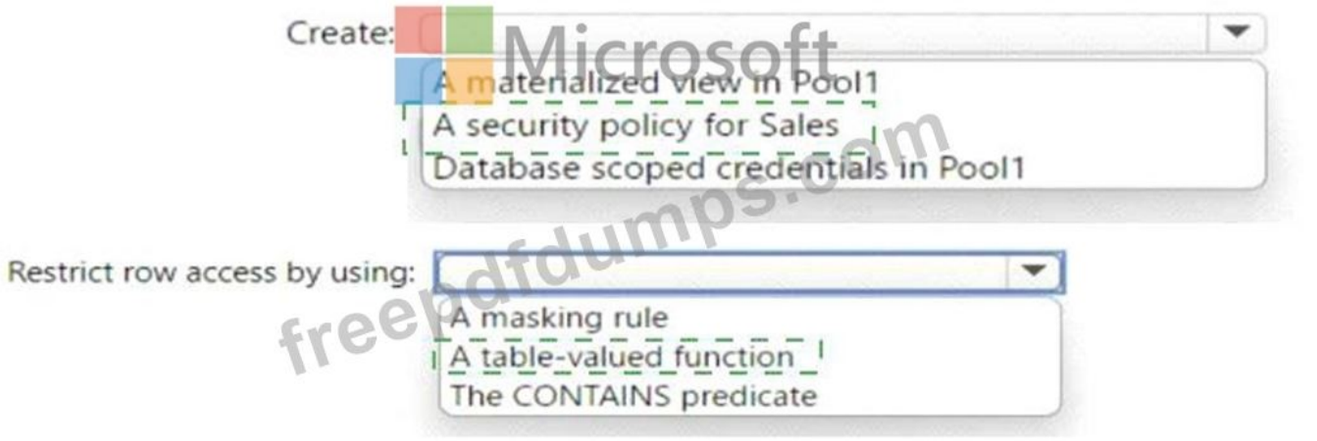
```
"name": "DemoSelfDependency",
"properties": {
"runtimeState": "Started",
"pipeline": {
"pipelineReference": {
"referenceName": "Demo",
"type": "PipelineReference"
}
},
"type": "TumblingWindowTrigger",
"typeProperties": {
"frequency": "Hour",
"interval": 1,
"startTime": "2018-10-04T00:00:00Z",
"delay":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

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

**NEW QUESTION: 27**
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains an external table named Sales. Sales contains sales data. Each row in Sales contains data on a single sale, including the name of the salesperson.
You need to implement row-level security (RLS). The solution must ensure that the salespeople can access only their respective sales.
What should you do? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



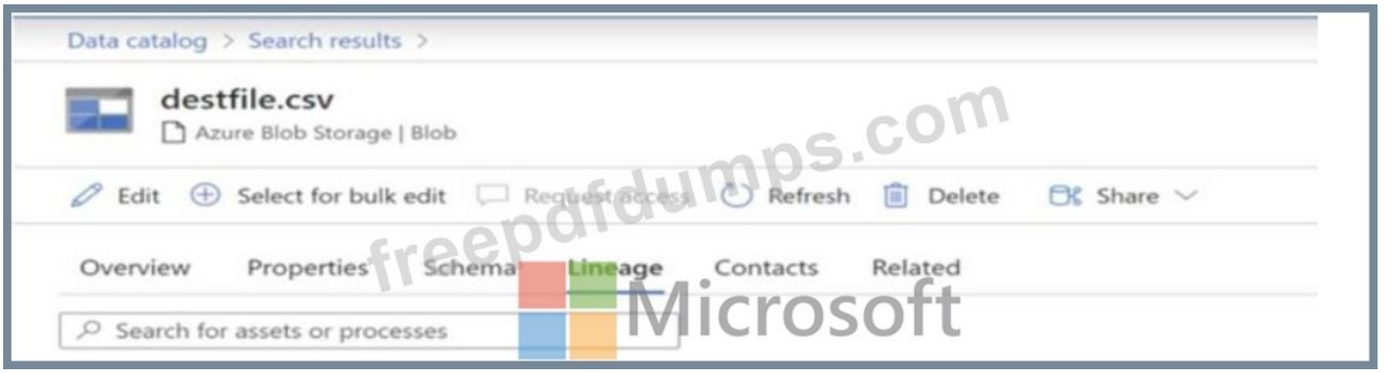

**Answer:**




Explanation
Box 1: A security policy for sale
Here are the steps to create a security policy for Sales:
Create a user-defined function that returns the name of the current user:
CREATE FUNCTION dbo.GetCurrentUser()

RETURNS NVARCHAR(128)
AS
BEGIN
RETURN SUSER_SNAME();
END;
Create a security predicate function that filters the Sales table based on the current user:
CREATE FUNCTION dbo.SalesPredicate(@salesperson NVARCHAR(128))
RETURNS TABLE
WITH SCHEMABINDING
AS
RETURN SELECT 1 AS access_result
WHERE @salesperson = SalespersonName;
Create a security policy on the Sales table that uses the SalesPredicate function to filter the data:
CREATE SECURITY POLICY SalesFilter
ADD FILTER PREDICATE dbo.SalesPredicate(dbo.GetCurrentUser()) ON dbo.Sales WITH (STATE = ON); By creating a security policy for the Sales table, you ensure that each salesperson can only access their own sales data. The security policy uses a user-defined function to get the name of the current user and a security predicate function to filter the Sales table based on the current user.
Box 2: table-value function
to restrict row access by using row-level security, you need to create a table-valued function that returns a table of values that represent the rows that a user can access. You then use this function in a security policy that applies a predicate on the table.

**NEW QUESTION: 28**
You have a Microsoft Purview account. The Lineage view of a CSV file is shown in the following exhibit.





How is the data for the lineage populated?
**A.** manually
**B.** by scanning data stores
**C.** by executing a Data Factory pipeline
**Answer: (SHOW ANSWER)**
Explanation

According to Microsoft Purview Data Catalog lineage user guide , data lineage in Microsoft Purview is a core platform capability that populates the Microsoft Purview Data Map with data movement and transformations across systems2. Lineage is captured as it flows in the enterprise and stitched without gaps irrespective of its source2.

**NEW QUESTION: 29**
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
**Answer: B (LEAVE A REPLY)**
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

**NEW QUESTION: 30**
You have an enterprise data warehouse in Azure Synapse Analytics.
Using PolyBase, you create an external table named [Ext].[Items] to query Parquet files stored in Azure Data Lake Storage Gen2 without importing the data to the data warehouse.
The external table has three columns.
You discover that the Parquet files have a fourth column named ItemID.
Which command should you run to add the ItemID column to the external table?

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

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

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

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



**A.** Option A
**B.** Option B
**C.** Option C
**D.** Option D
**Answer: C (LEAVE A REPLY)**
Explanation
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql

**NEW QUESTION: 31**
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

**Answer: (SHOW ANSWER)**

Explanation

Use the same definition as the EmployeeID column.

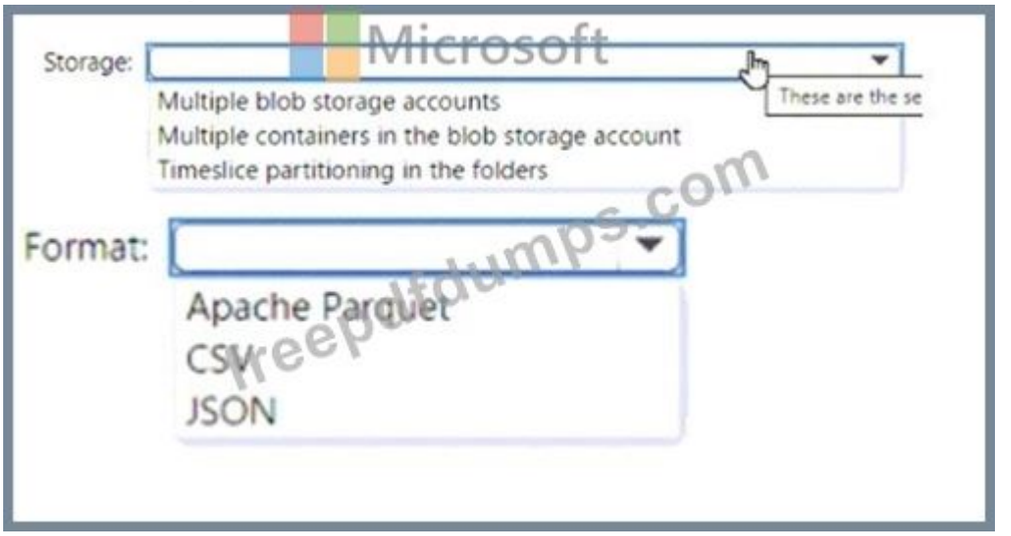
Reference:

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

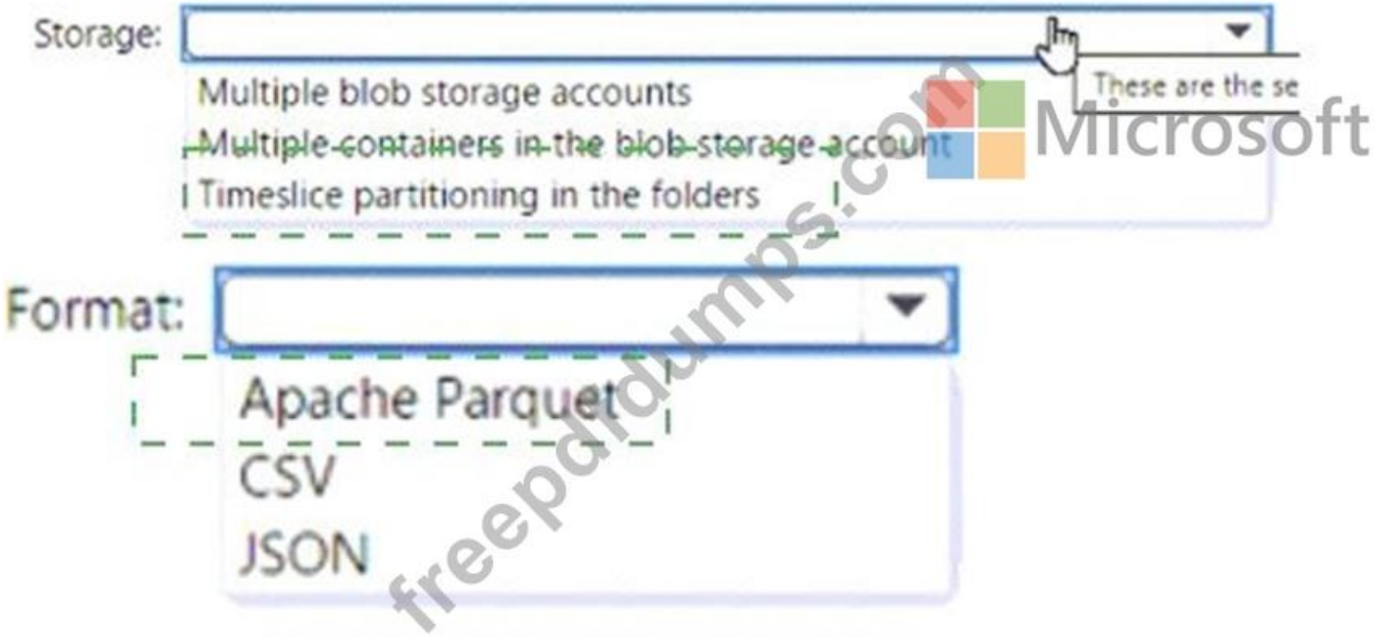


**NEW QUESTION: 32**

You have an Azure Blob storage account that contains a folder. The folder contains 120,000 files. Each file contains 62 columns.
Each day, 1,500 new files are added to the folder.
You plan to incrementally load five data columns from each new file into an Azure Synapse Analytics workspace.
You need to minimize how long it takes to perform the incremental loads.
What should you use to store the files and format?




**Answer:**




Explanation
Box 1 = timeslice partitioning in the foldersThis means that you should organize your files into folders based on a time attribute, such as year, month, day, or hour. For example, you can have a folder structure like
/yyyy/mm/dd/file.csv. This way, you can easily identify and load only the new files that are added each day by using a time filter in your Azure Synapse pipeline . Timeslice partitioning can also

improve the performance of data loading and querying by reducing the number of files that need to be scanned Box = 2 Apache Parquet This is because Parquet is a columnar file format that can efficiently store and compress data with many columns. Parquet files can also be partitioned by a time attribute, which can improve the performance of incremental loading and querying by reducing the number of files that need to be scanned1
23. Parquet files are supported by both dedicated SQL pool and serverless SQL pool in Azure Synapse Analytics2.

**NEW QUESTION: 33**
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
**Answer: B (LEAVE A REPLY)**
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-monit

**NEW QUESTION: 34**
You have an Azure Synapse Analytics job that uses Scala.
You need to view the status of the job.
What should you do?
**A.** From Azure Monitor, run a Kusto query against the AzureDiagnostics table.
**B.** From Azure Monitor, run a Kusto query against the SparkLogying1 Event.CL table.
**C.** From Synapse Studio, select the workspace. From Monitor, select Apache Sparks applications.

**D.** From Synapse Studio, select the workspace. From Monitor, select SQL requests.

**Answer: C (LEAVE A REPLY)**

Explanation

Use Synapse Studio to monitor your Apache Spark applications. To monitor running Apache Spark application Open Monitor, then select Apache Spark applications. To view the details about the Apache Spark applications that are running, select the submitting Apache Spark application and view the details. If the Apache Spark application is still running, you can monitor the progress.

Reference:

https://docs.microsoft.com/en-us/azure/synapse-analytics/monitoring/apache-spark-applications

**NEW QUESTION: 35**

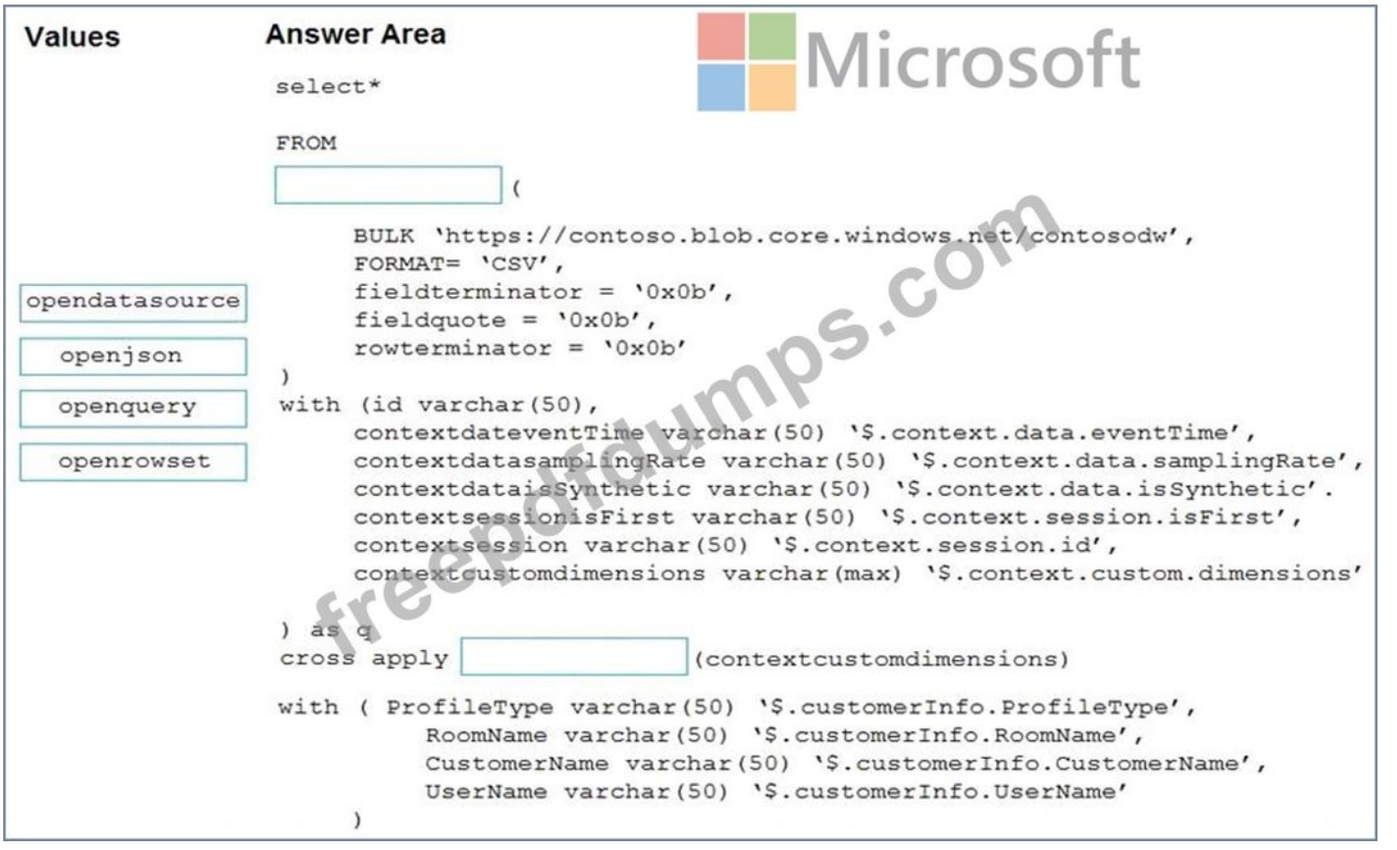
You have an Azure Synapse Analytics workspace named WS1.

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

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.






**Answer:**

**Values**

opendatasource

openjson

openquery

openrowset

**Answer Area**

```
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'
      )
```



Explanation

Graphical user interface, text, application, email Description automatically generated

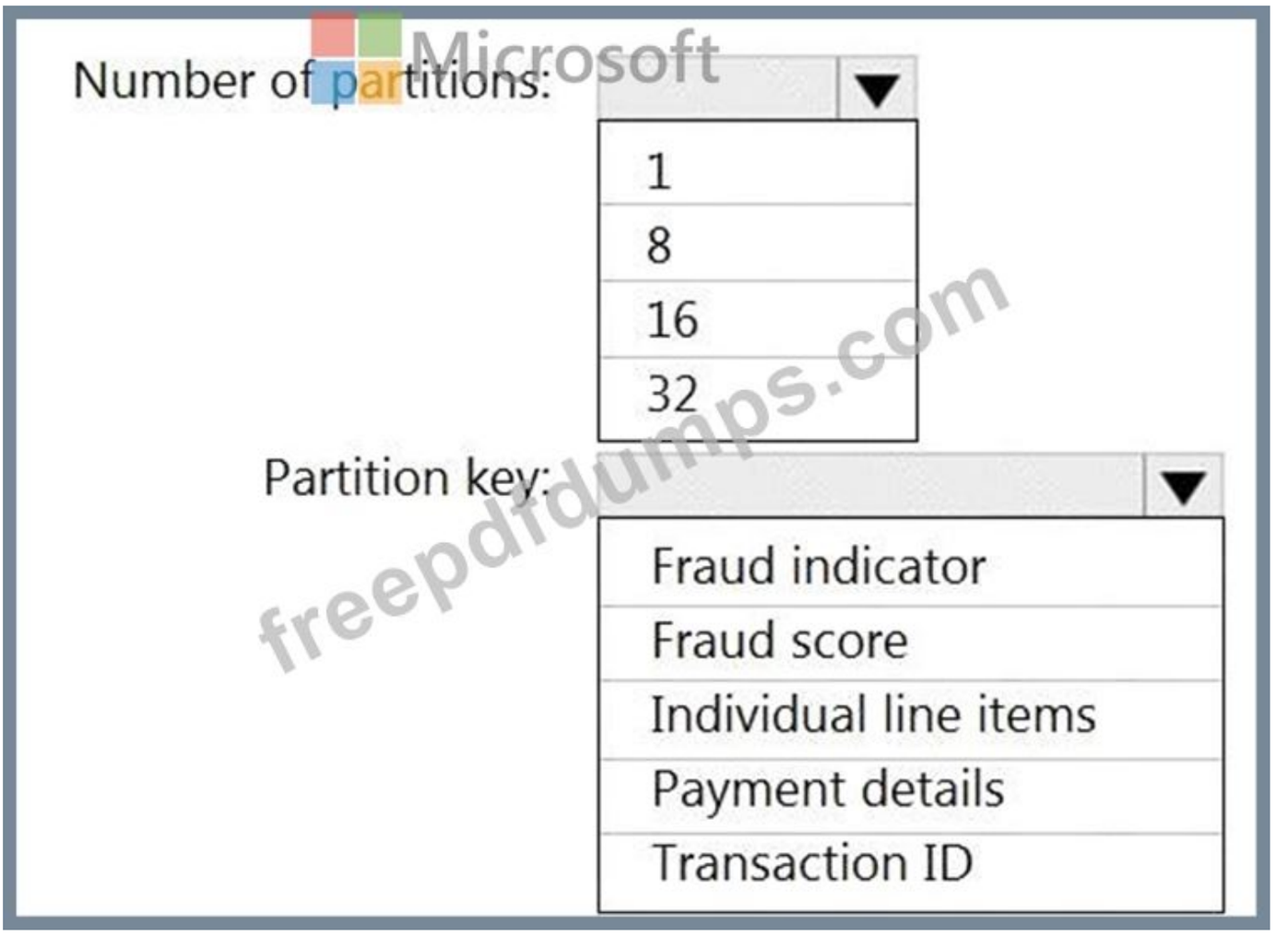



Box 1: openrowset

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

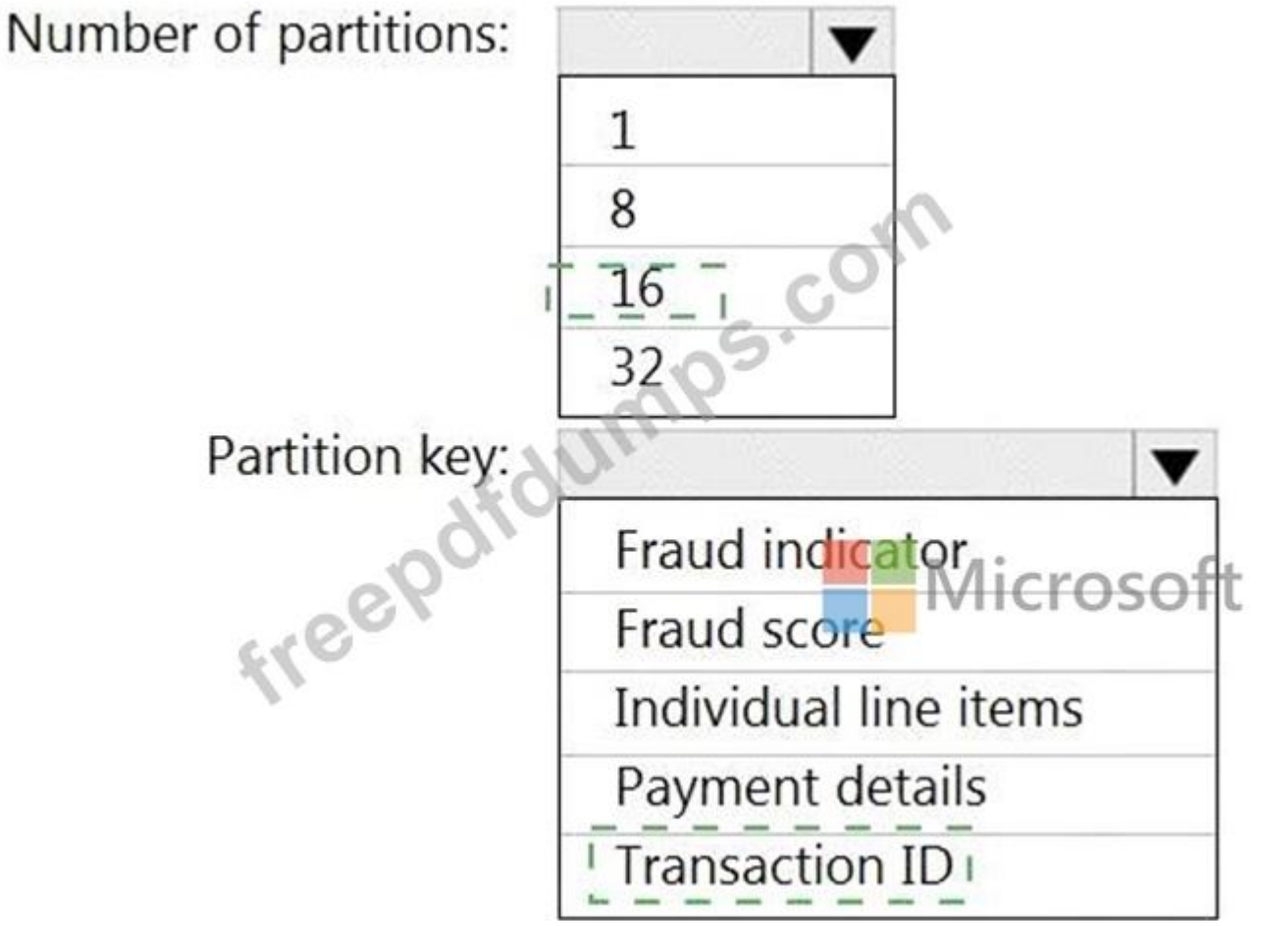
Example:
SELECT *
FROM OPENROWSET(
BULK 'csv/population/population.csv',
DATA_SOURCE = 'SqlOnDemandDemo',
FORMAT = 'CSV', PARSER_VERSION = '2.0',
FIELDTERMINATOR =',',
ROWTERMINATOR = '\n'
Box 2: openjson
You can access your JSON files from the Azure File Storage share by using the mapped drive, as shown 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

**NEW QUESTION: 36**
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.



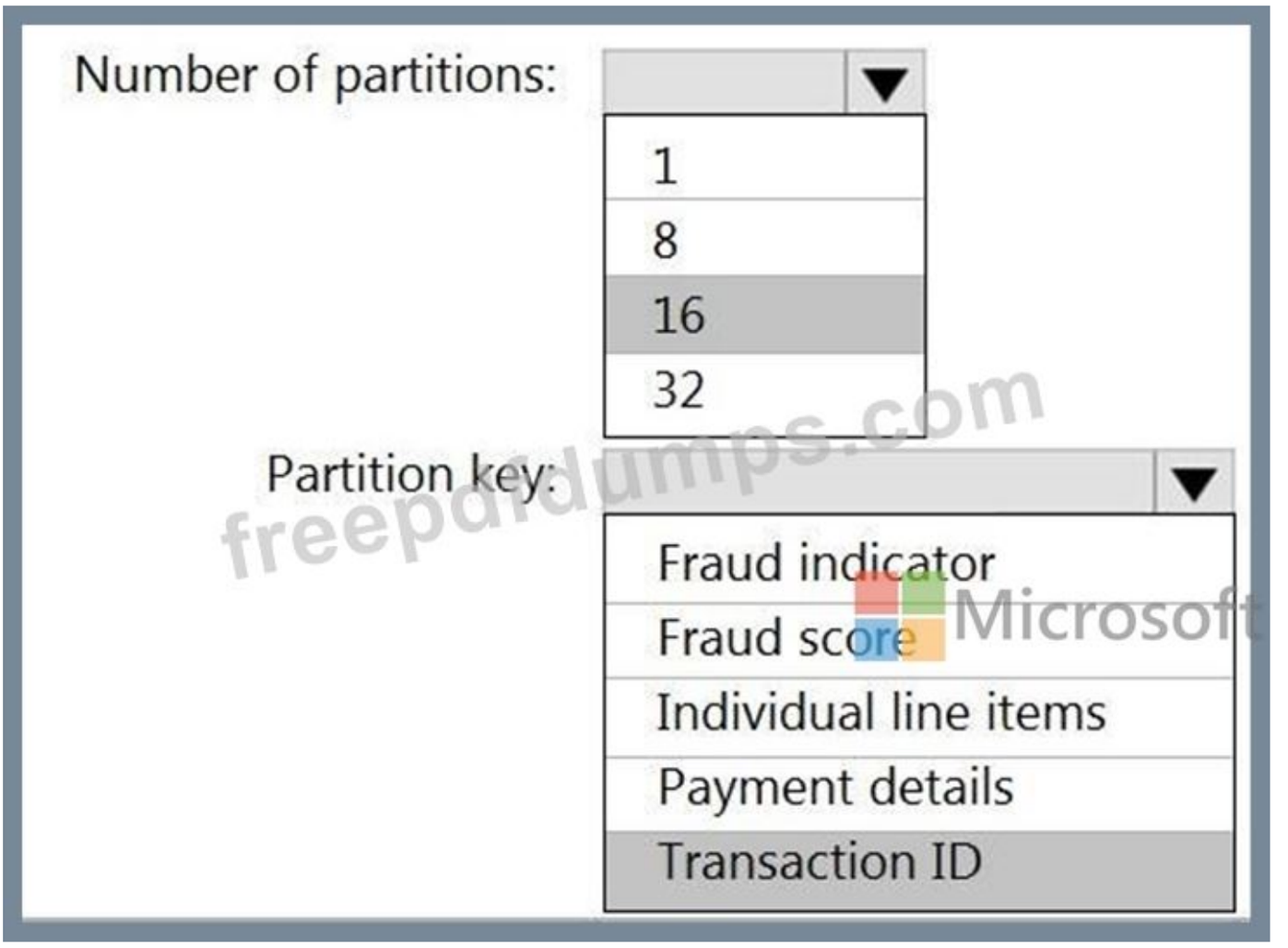



**Answer:**






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

**NEW QUESTION: 37**

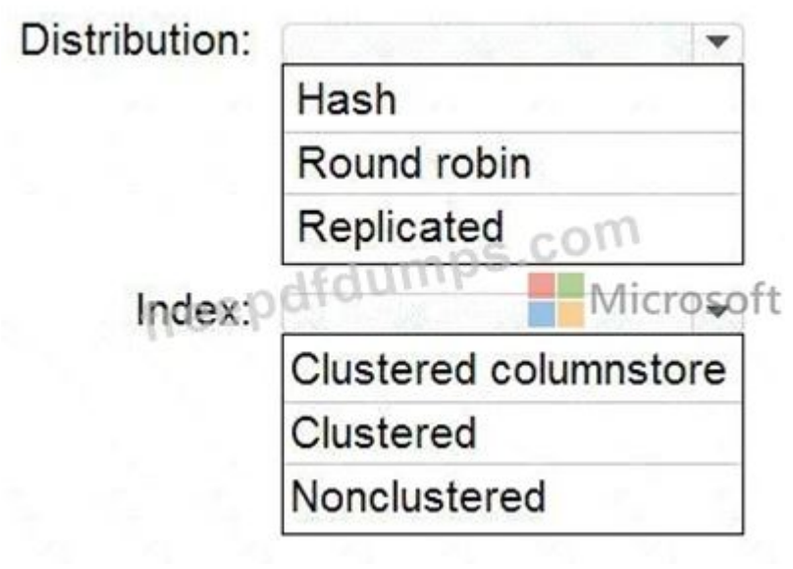
You are designing an enterprise data warehouse in Azure Synapse Analytics that will store website traffic analytics in a star schema.

You plan to have a fact table for website visits. The table will be approximately 5 GB.

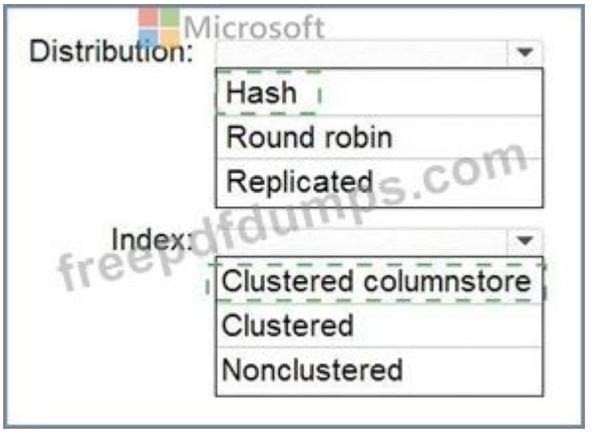
You need to recommend which distribution type and index type to use for the table. The solution must provide the fastest query performance.

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

NOTE: Each correct selection is worth one point.




**Answer:**




Explanation

Box 1: Hash

Consider using a hash-distributed table when:

The table size on disk is more than 2 GB.

The table has frequent insert, update, and delete operations.

Box 2: Clustered columnstore

Clustered columnstore tables offer both the highest level of data compression and the best overall query performance.

Reference:

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

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

**NEW QUESTION: 38**

You have an Azure Data Lake Storage account that has a virtual network service endpoint configured.

You plan to use Azure Data Factory to extract data from the Data Lake Storage account. The data will then be loaded to a data warehouse in Azure Synapse Analytics by using PolyBase.

Which authentication method should you use to access Data Lake Storage?
**A.** shared access key authentication
**B.** managed identity authentication
**C.** account key authentication
**D.** service principal authentication
**Answer: B (LEAVE A REPLY)**
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-sql-data-warehouse#use-polybase-to-load-d

**NEW QUESTION: 39**
You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account.
You need to output the count of tweets from the last five minutes every minute.
Which windowing function should you use?
**A.** Sliding
**B.** Session
**C.** Tumbling
**D.** Hopping
**Answer: D (LEAVE A REPLY)**
Explanation
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.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

**NEW QUESTION: 40**
You have an Azure Factory instance named DF1 that contains a pipeline named PL1.PL1 includes a tumbling window trigger.
You create five clones of PL1. You configure each clone pipeline to use a different data source.
You need to ensure that the execution schedules of the clone pipeline match the execution schedule of PL1.
What should you do?
**A.** Associate each cloned pipeline to an existing trigger.
**B.** Create a tumbling window trigger dependency for the trigger of PL1.
**C.** Add a new trigger to each cloned pipeline
**D.** Modify the Concurrency setting of each pipeline.
**Answer: (SHOW ANSWER)**

**NEW QUESTION: 41**
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 |

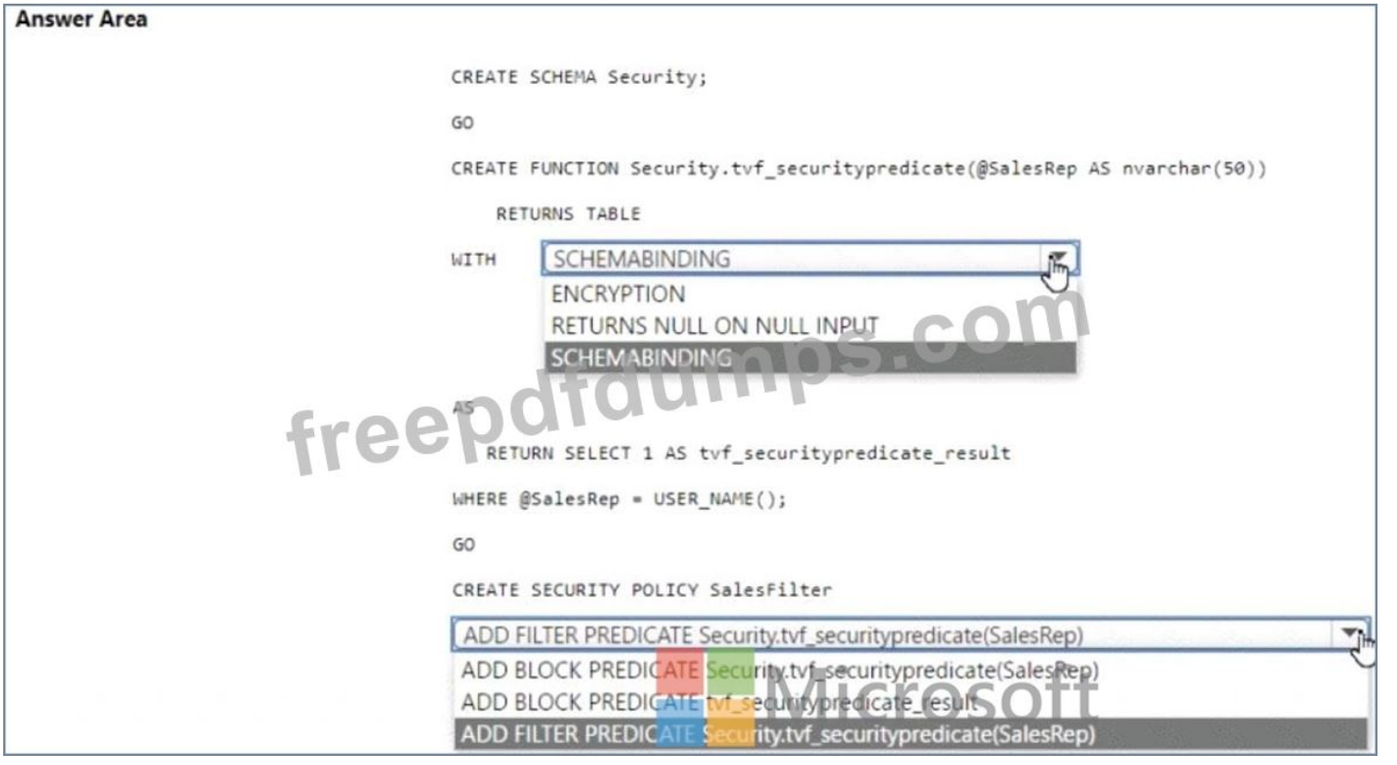
One minute later, you execute the following query from a serverless SQL pool in MyWorkspace.
SELECT EmployeeID
FROM mytestdb.dbo.myParquetTable
WHERE name = 'Alice';
What will be returned by the query?
**A.** 24
**B.** an error
**C.** a null value
**Answer: B (LEAVE A REPLY)**
Explanation
Once a database 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

**NEW QUESTION: 42**
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Sales.Orders.
Sales.Orders contains a column named SalesRep.
You plan to implement row-level security (RLS) for Sales.Orders.

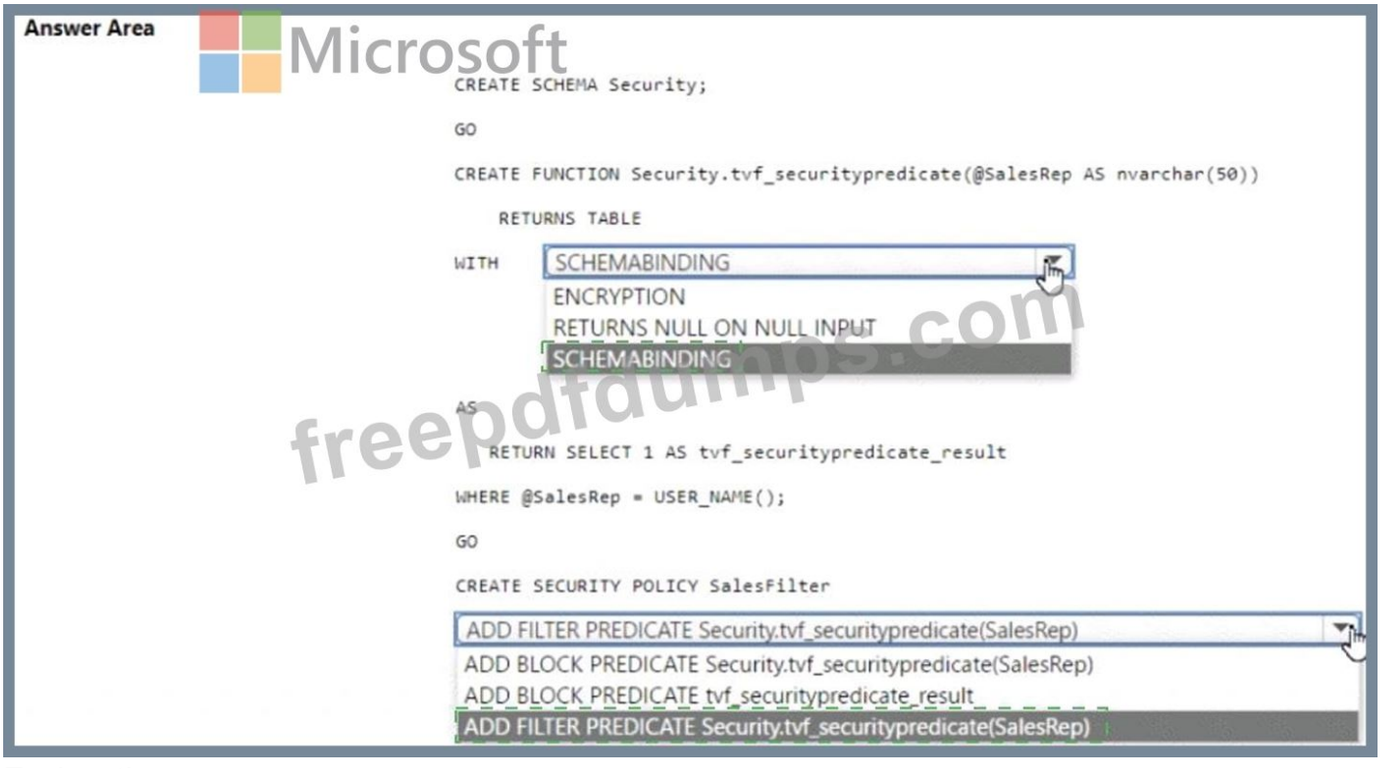
You need to create the security policy that will be used to implement RLS. The solution must ensure that sales representatives only see rows for which the value of the SalesRep column matches their username.

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

NOTE: Each correct selection is worth one point.




**Answer:**




Explanation

Answer Area

```
CREATE SCHEMA Security;
GO
CREATE FUNCTION Security.tvf_securitypredicate(@SalesRep AS nvarchar(50))
    RETURNS TABLE
WITH    SCHEMABINDING
AS
    RETURN SELECT 1 AS tvf_securitypredicate_result
WHERE @SalesRep = USER_NAME();
GO
CREATE SECURITY POLICY SalesFilter
    ADD FILTER PREDICATE Security.tvf_securitypredicate(SalesRep)
ON Sales.Orders
WITH (STATE = ON);
```




**NEW QUESTION: 43**

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

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

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

NOTE: Each correct selection is worth one point.

**A.** sensitivity-classification labels applied to columns that contain confidential information

**B.** resource tags for databases that contain confidential information

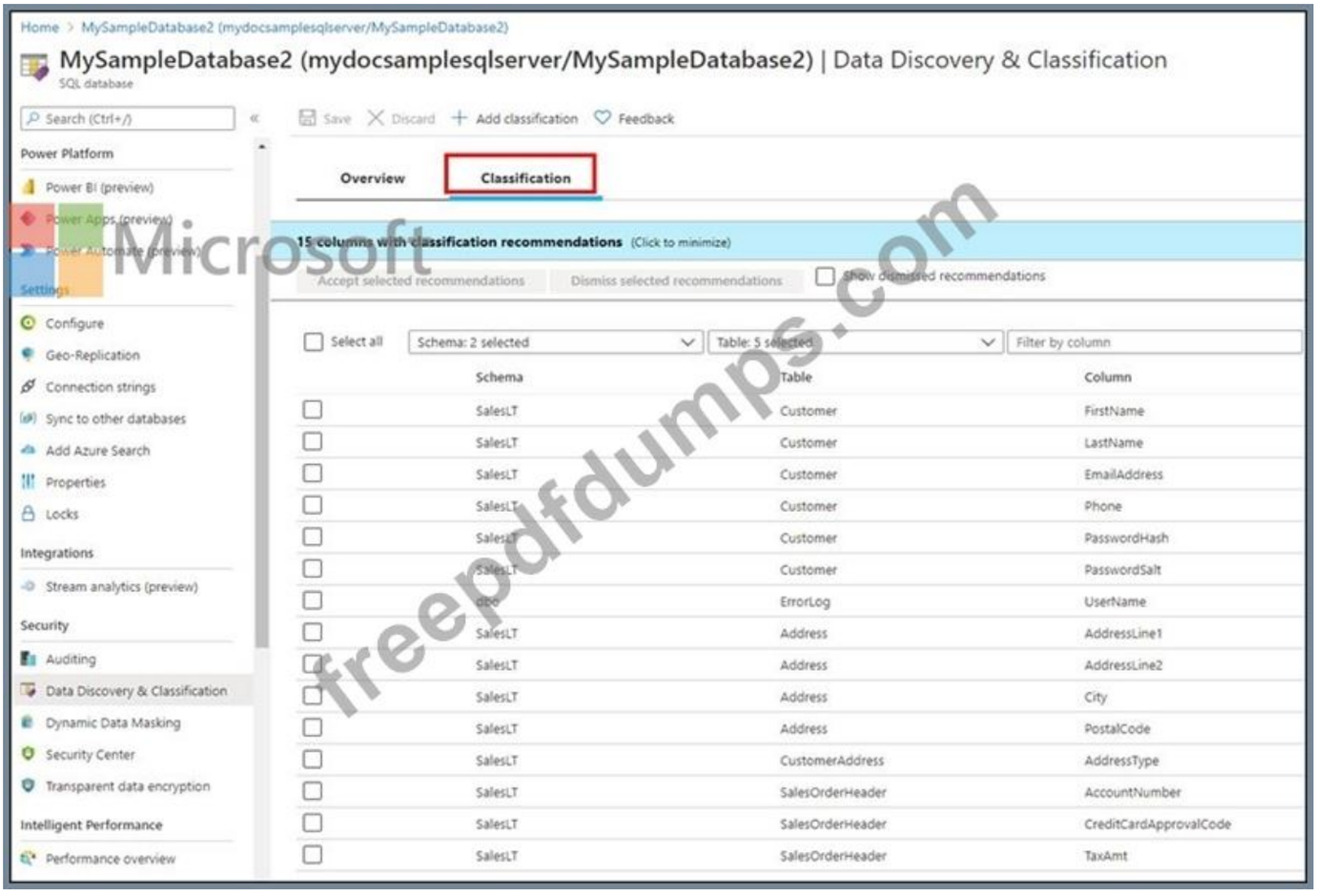
**C.** audit logs sent to a Log Analytics workspace

**D.** dynamic data masking for columns that contain confidential information

**Answer: (SHOW ANSWER)**

Explanation

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




Select Add classification in the top menu of the pane.

In the context window that opens, select the schema, table, and column that you want to classify, and the information type and sensitivity label.

Select Add classification at the bottom of the context window.

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

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

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

**NEW QUESTION: 44**

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.

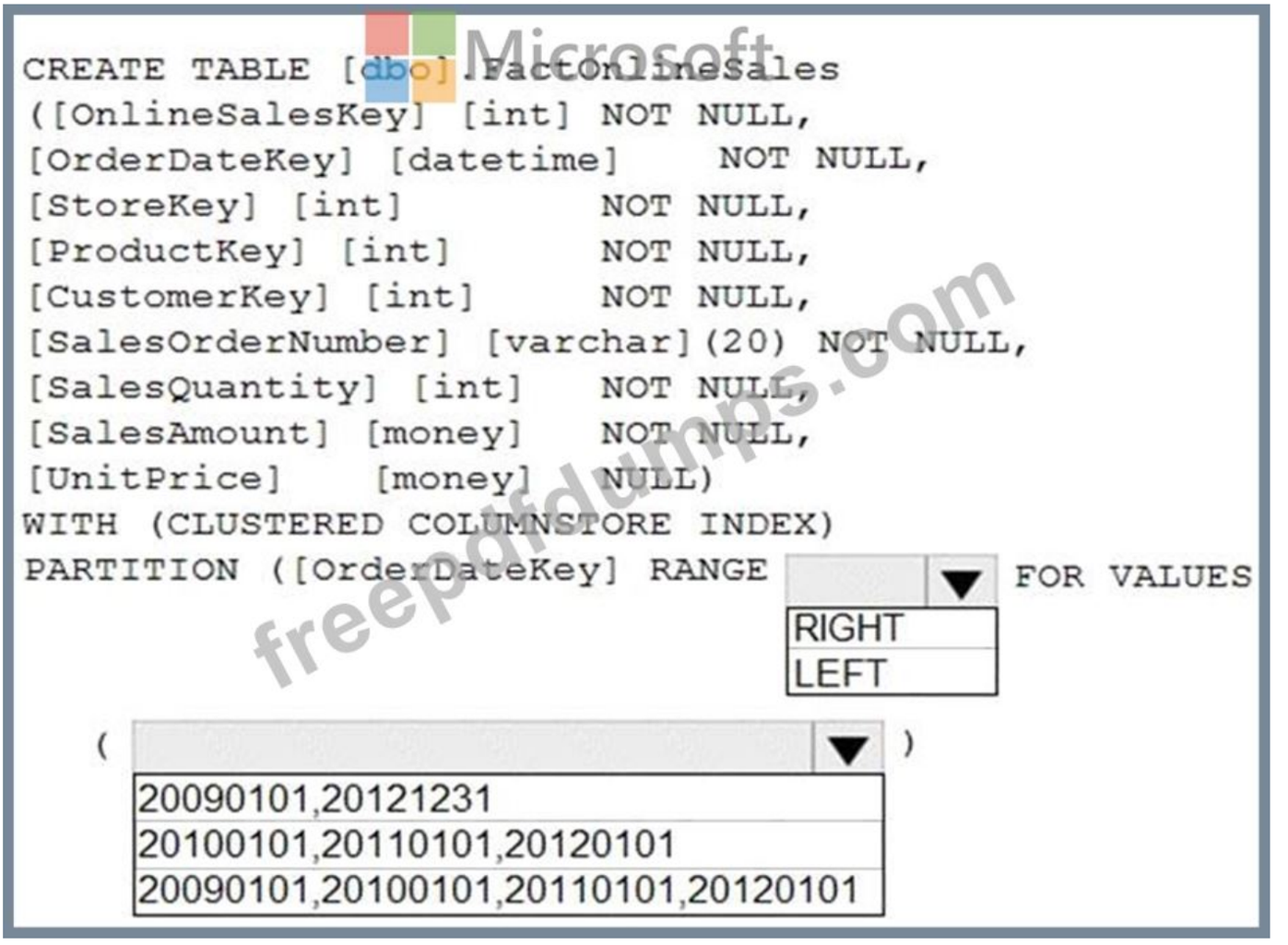
You need to improve the performance of queries against FactOnlineSales by using table partitions. The solution must meet the following requirements:

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.




**Answer:**

```
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 [ RIGHT | LEFT ] FOR VALUES
(
  [ 20090101,20121231 | 20100101,20110101,20120101 | 20090101,20100101,20110101,20120101 ]
)
```

Explanation

Text Description automatically generated

```
[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)
ITH (CLUSTERED COLUMNSTORE INDEX)
ARTITION ([OrderDateKey] RANGE [ RIGHT | LEFT ] FOR VALUE
(
  [ 20090101,20121231 | 20100101,20110101,20120101 | 20090101,20100101,20110101,20120101 ]
```

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-ver1

**NEW QUESTION: 45**

You have an Azure subscription that contains an Azure Databricks workspace. The workspace contains a notebook named Notebook1. In Notebook1, you create an Apache Spark DataFrame named df_sales that contains the following columns:

* Customer
* Salesperson
* Region
* Amount

You need to identify the three top performing salespersons by amount for a region named HQ. 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.


**Answer:**

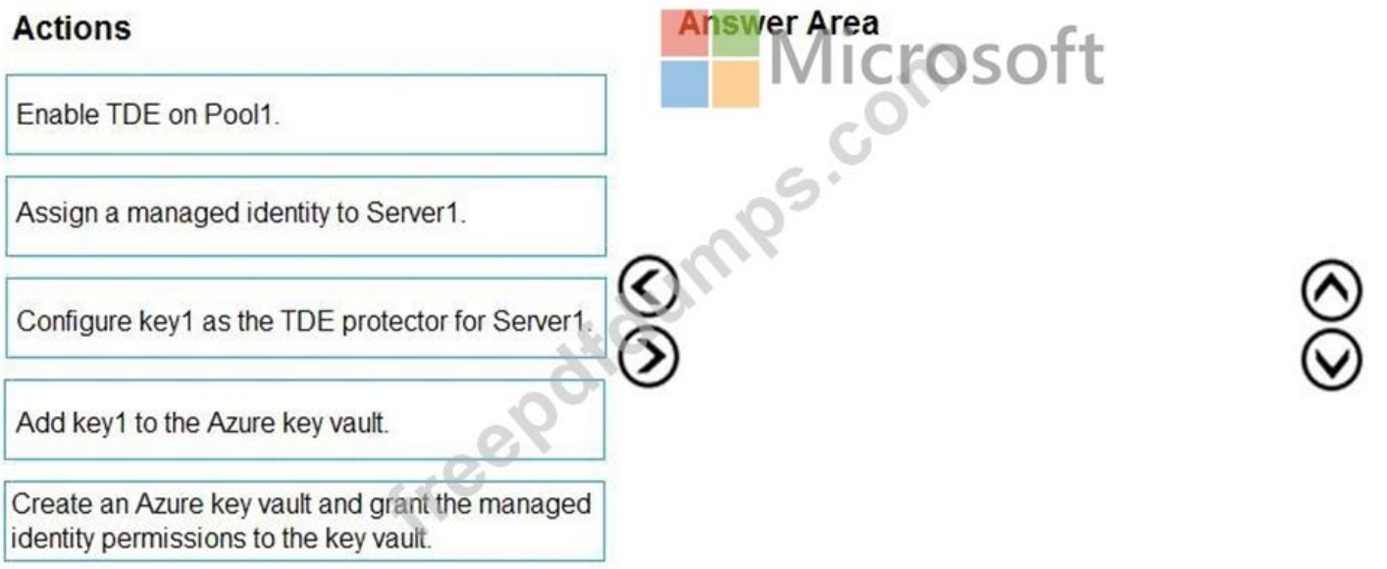

Explanation


**NEW QUESTION: 46**

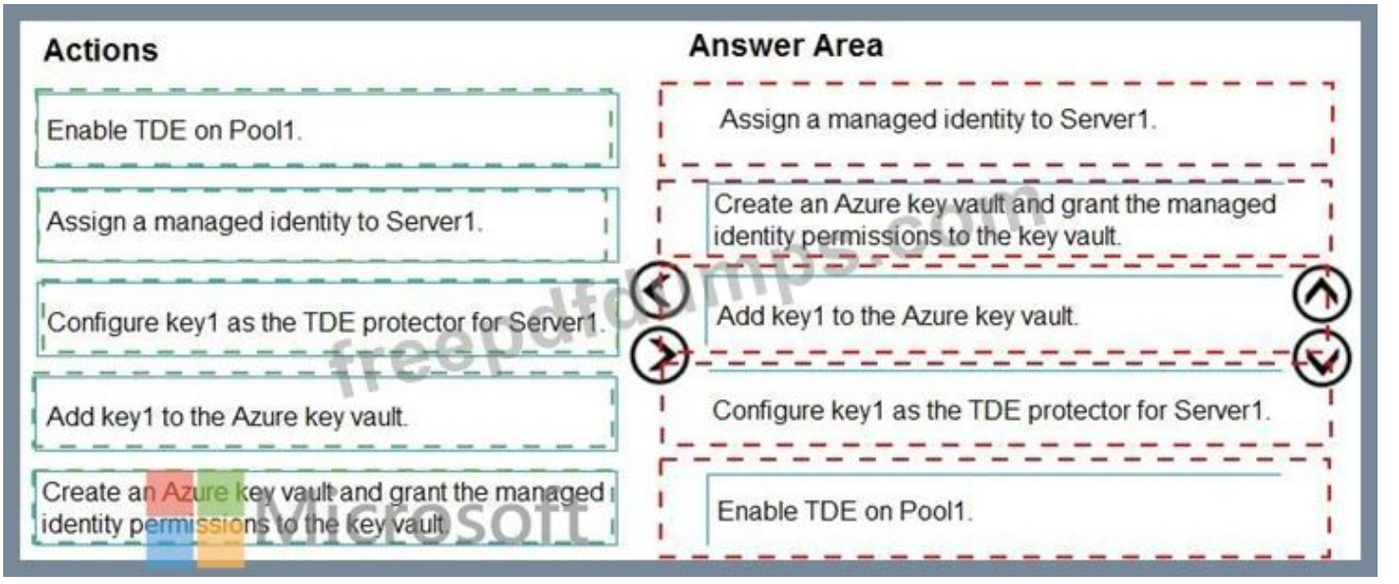
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.
You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.
You plan to insert data from the files into Table1 and azure Data Lake Storage Gen2 container named container1.
You plan to insert data from the files into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.
You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.
Solution: In an Azure Synapse Analytics pipeline, you use a data flow that contains a Derived Column transformation.
**A.** Yes
**B.** No
**Answer: A (LEAVE A REPLY)**
Explanation
Use the derived column transformation to generate new columns in your data flow or to modify existing fields.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column



**NEW QUESTION: 47**
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.

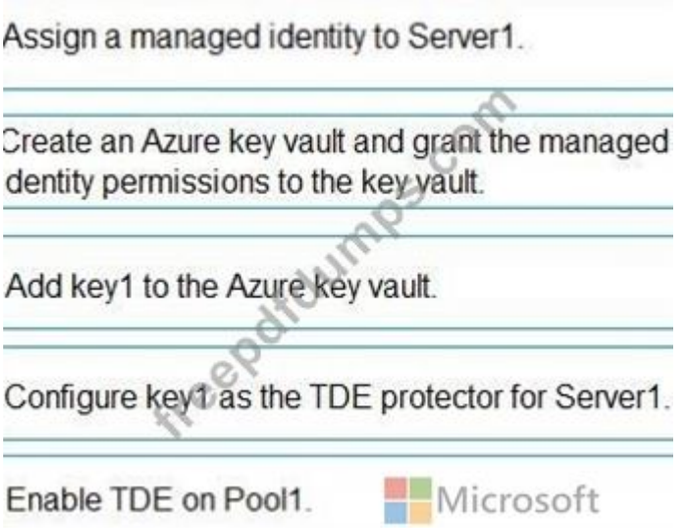



**Answer:**




Explanation

Graphical user interface, text, application Description automatically generated

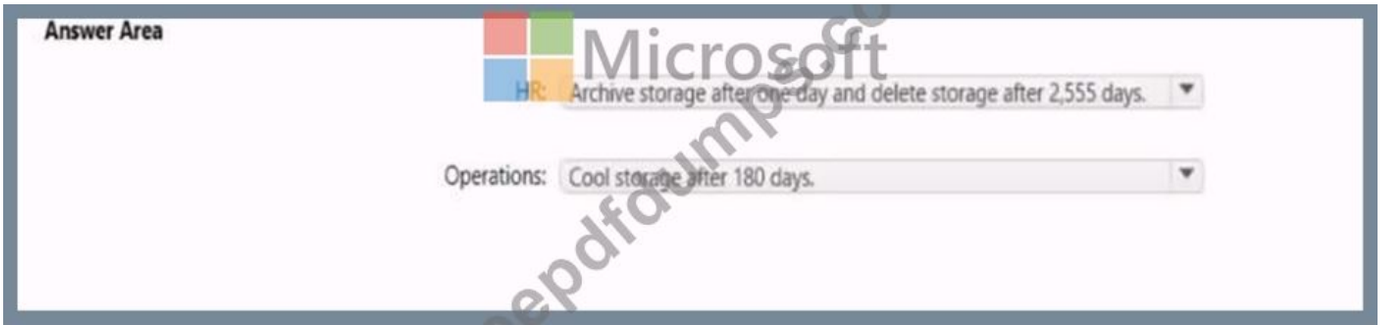



Step 1: Assign a managed identity to Server1

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-pow

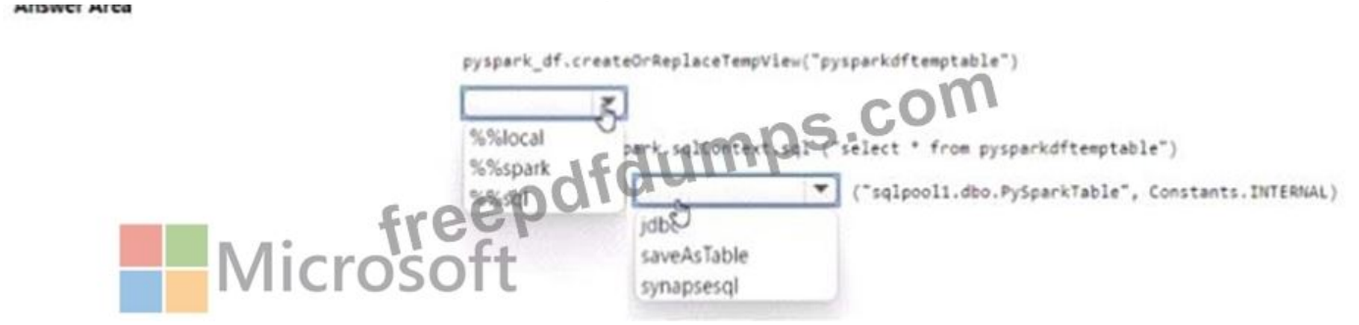
**NEW QUESTION: 48**
You are designing an Azure Data Lake Storage Gen2 container to store data for the human resources (HR) department and the operations department at your company. You have the following data access requirements:
* After initial processing, the HR department data will be retained for seven years.
* The operations department data will be accessed frequently for the first six months, and then accessed once per month.
You need to design a data retention solution to meet the access requirements. The solution must minimize storage costs.
**Answer:**
See the answer in explanation.
Explanation
answer is below



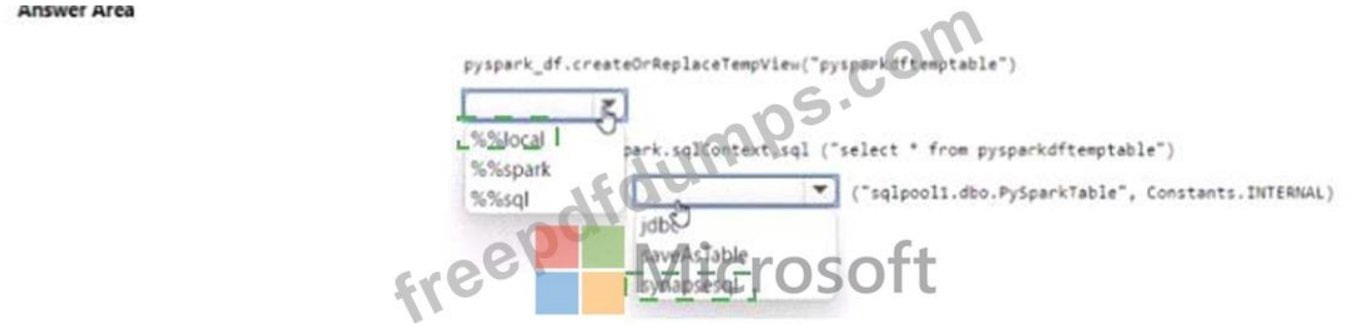

**NEW QUESTION: 49**
You use Azure Data Lake Storage Gen2.
You need to ensure that workloads can use filter predicates and column projections to filter data at the time the data is read from disk.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
**A.** Create a storage policy that is scoped to a container prefix filter.

**B.** Register the query acceleration feature.
**C.** Reregister the Azure Storage resource provider.
**D.** Create a storage policy that is scoped to a container.
**E.** Reregister the Microsoft Data Lake Store resource provider.
**Answer: B,C (LEAVE A REPLY)**

**NEW QUESTION: 50**
You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1.
Workspace1 contains a dedicated SQL pool named SQL Pool and an Apache Spark pool named sparkpool.
Sparkpool1 contains a DataFrame named pyspark.df.
You need to write the contents of pyspark_df to a tabte in SQLPooM by using a PySpark notebook.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.




**Answer:**

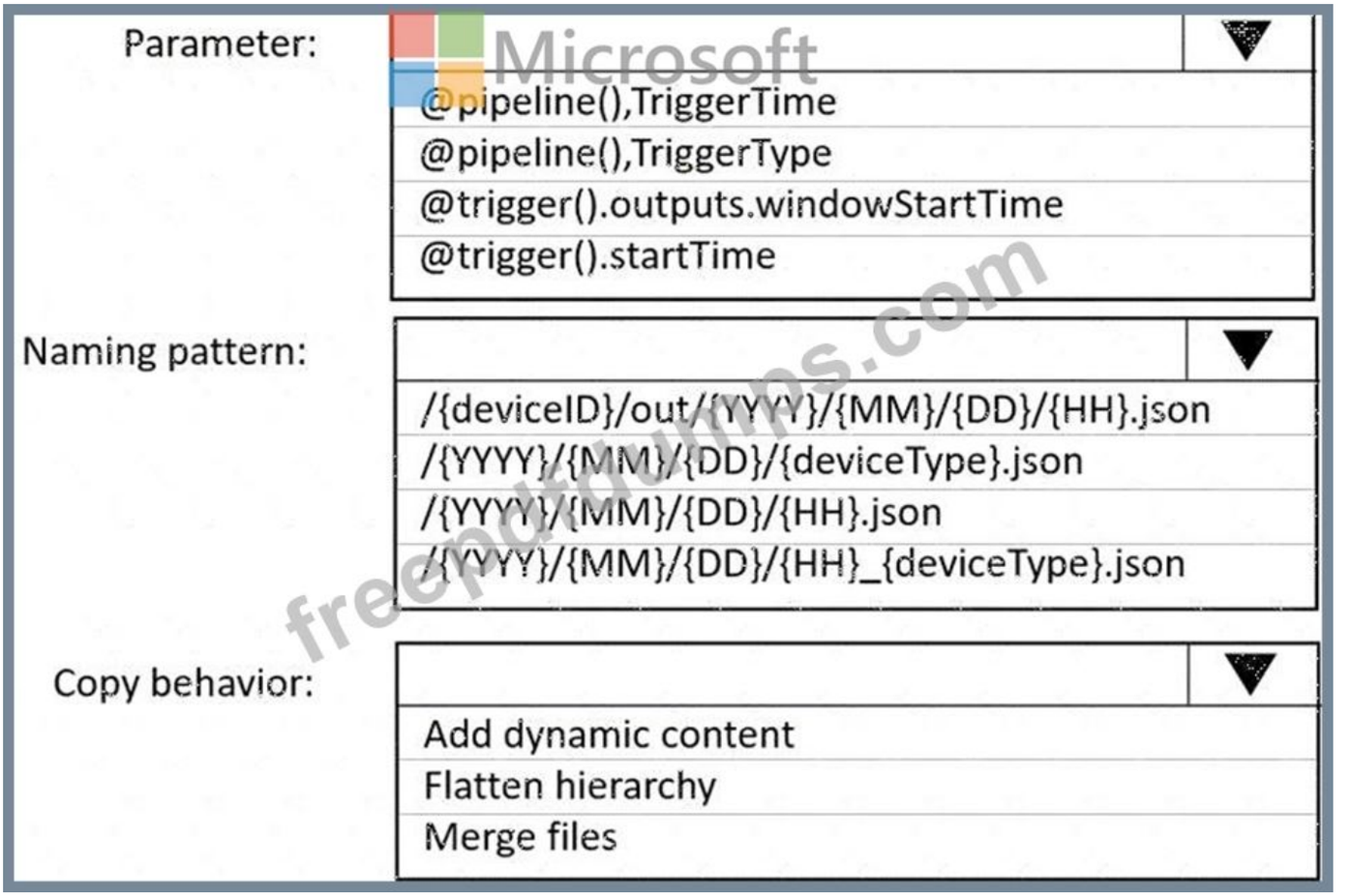



Explanation

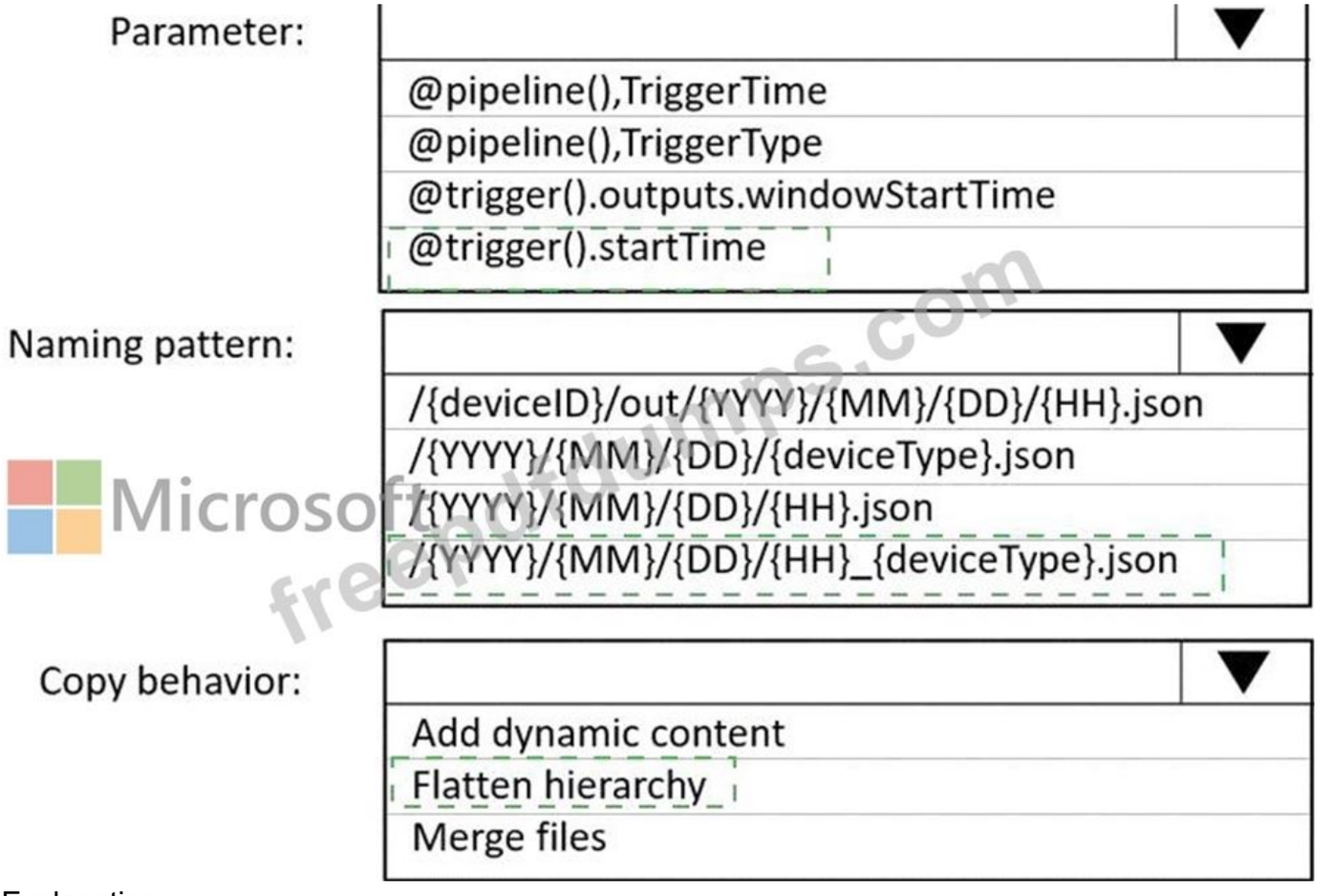

**NEW QUESTION: 51**

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
**Answer: (SHOW ANSWER)**
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

**NEW QUESTION: 52**
You are building an Azure Data Factory solution to process data received from Azure Event Hubs, and then ingested into an Azure Data Lake Storage Gen2 container.
The data will be ingested every five minutes from devices into JSON files. The files have the following naming pattern.
/{deviceType}/in/{YYYY}/{MM}/{DD}/{HH}/{deviceID}_{YYYY}{MM}{DD}HH}{mm}.json You need to prepare the data for batch data processing so that there is one dataset per hour per deviceType.
The solution must minimize read times.
How should you configure the sink for the copy activity? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



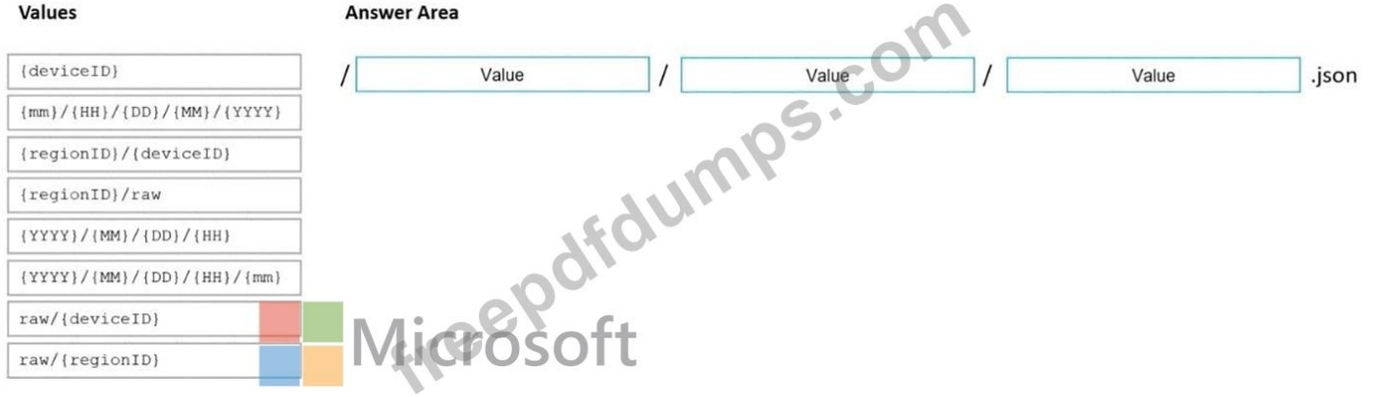

**Answer:**




Explanation

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

**NEW QUESTION: 53**
You are designing an Azure Data Lake Storage Gen2 structure for telemetry data from 25 million devices distributed across seven key geographical regions. Each minute, the devices will send a JSON payload of metrics to Azure Event Hubs.
You need to recommend a folder structure for the data. The solution must meet the following requirements:
Data engineers from each region must be able to build their own pipelines for the data of their respective region only.
The data must be processed at least once every 15 minutes for inclusion in Azure Synapse Analytics serverless SQL pools.
How should you recommend completing the structure? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.




**Answer:**



Explanation

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

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

Box 2: {regionID}/raw

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

Box 3: {deviceID}

Reference:

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

**NEW QUESTION: 54**

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.

**Answer: A (LEAVE A REPLY)**

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:

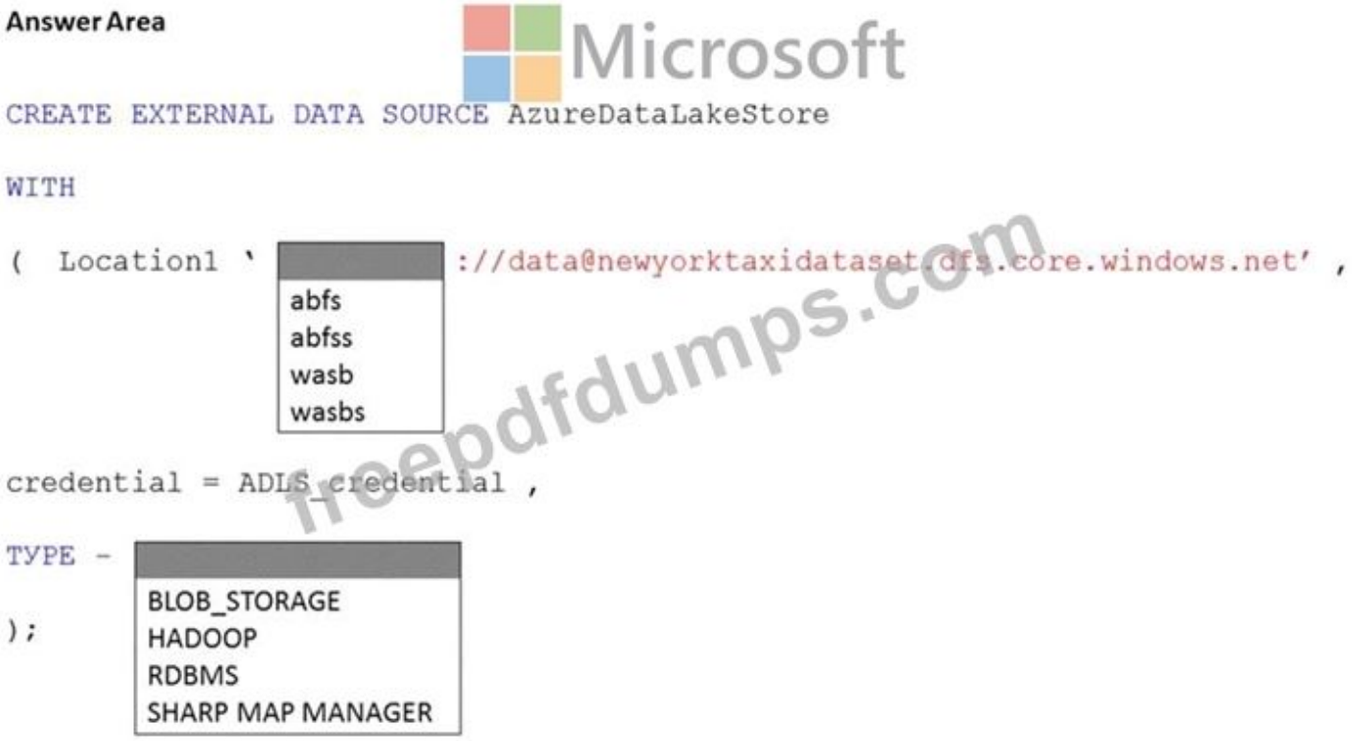
* Lower latency: no need to list nested directory structures on ABS, which is slow and resource intensive.

* 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

**NEW QUESTION: 55**
You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers.
You need to create an external data source in Pool1 that will be used to read .orc files in storage1.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

**Answer Area**






**Answer:**

Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore

WITH

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

credential = ADLS_credential ,

TYPE - [ BLOB_STORAGE | HADOOP | RDBMS | SHARP MAP MANAGER ]

);
```

Explanation

Graphical user interface, text, application, email Description automatically generated

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore

WITH

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

credential = ADLS_credential ,

TYPE - [ BLOB_STORAGE | HADOOP | RDBMS | SHARP MAP MANAGER ]

);
```

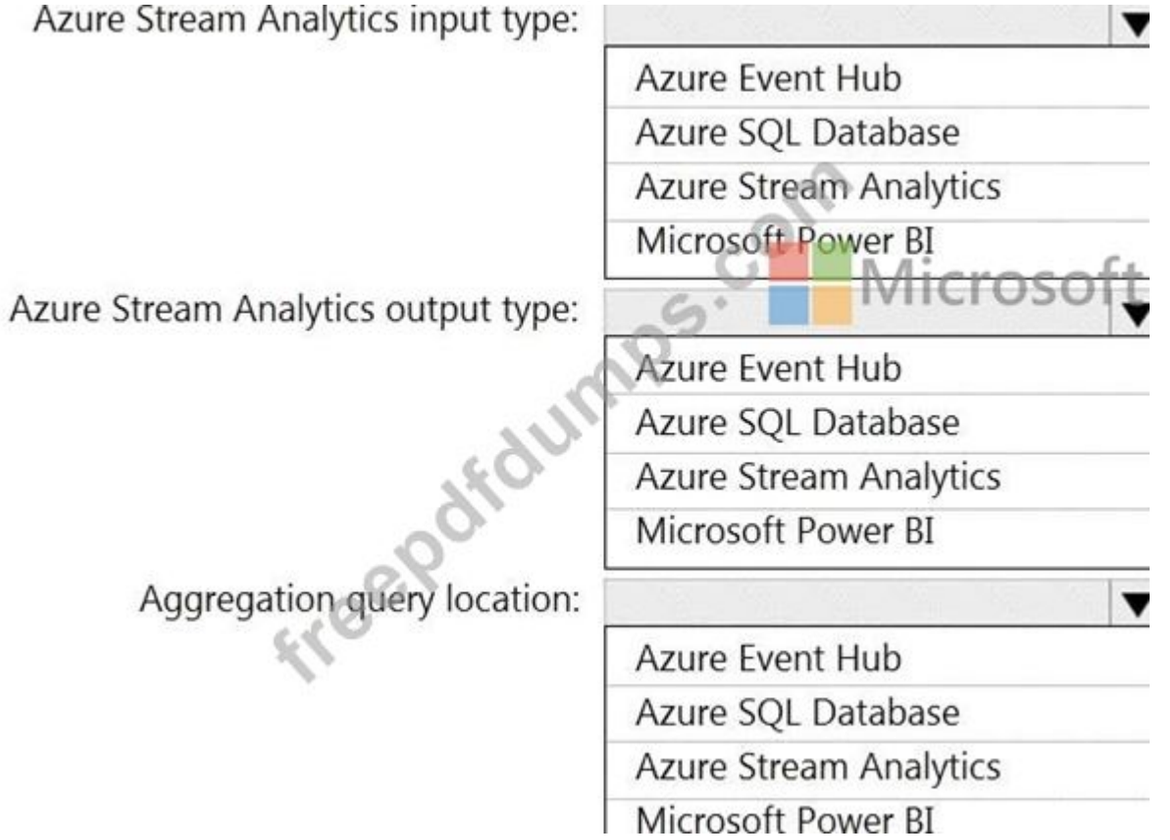
Reference:

https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw

**NEW QUESTION: 56**

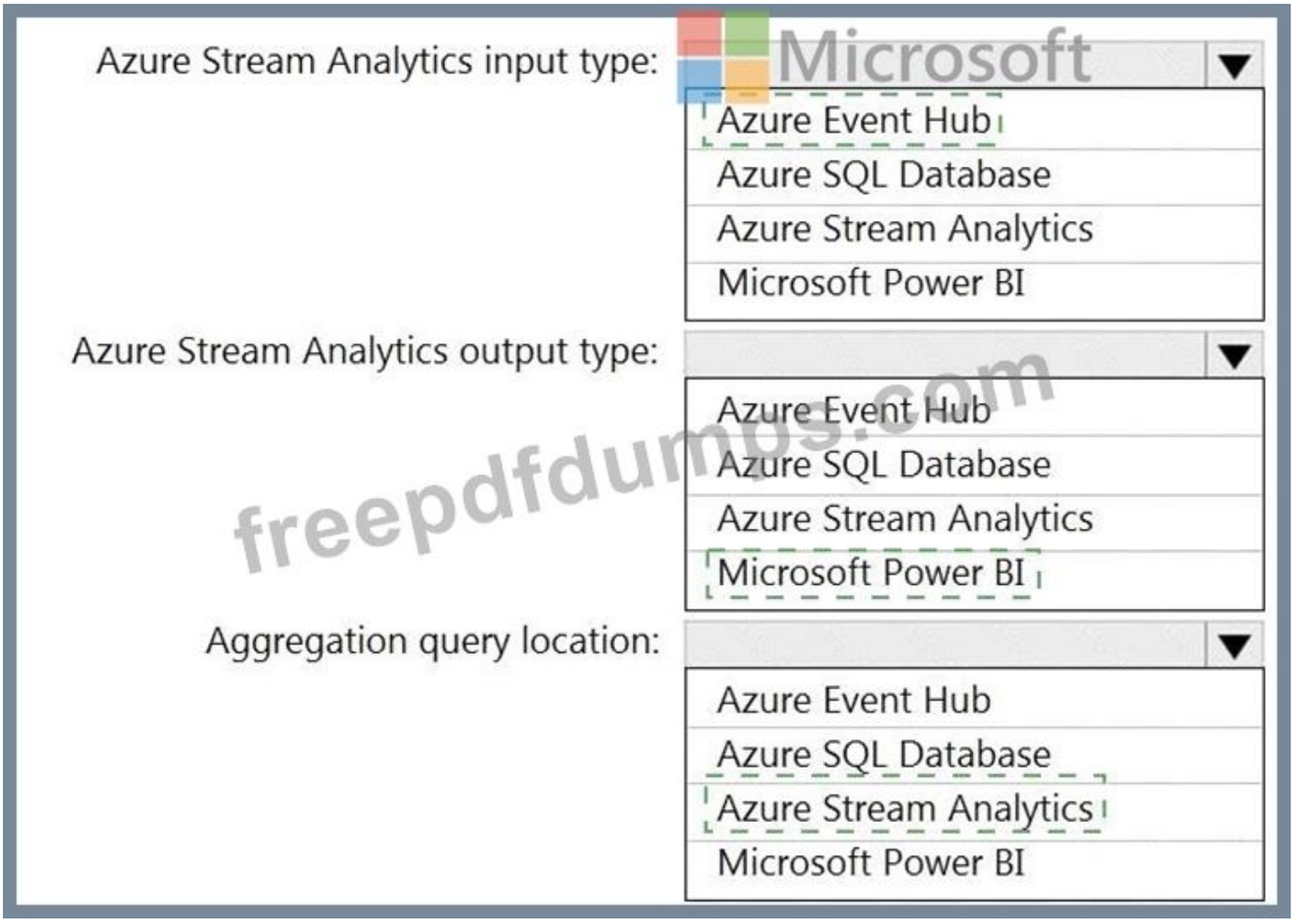
You are designing a real-time dashboard solution that will visualize streaming data from remote sensors that connect to the internet. The streaming data must be aggregated to show the average value of each 10-second interval. The data will be discarded after being displayed in the dashboard.

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



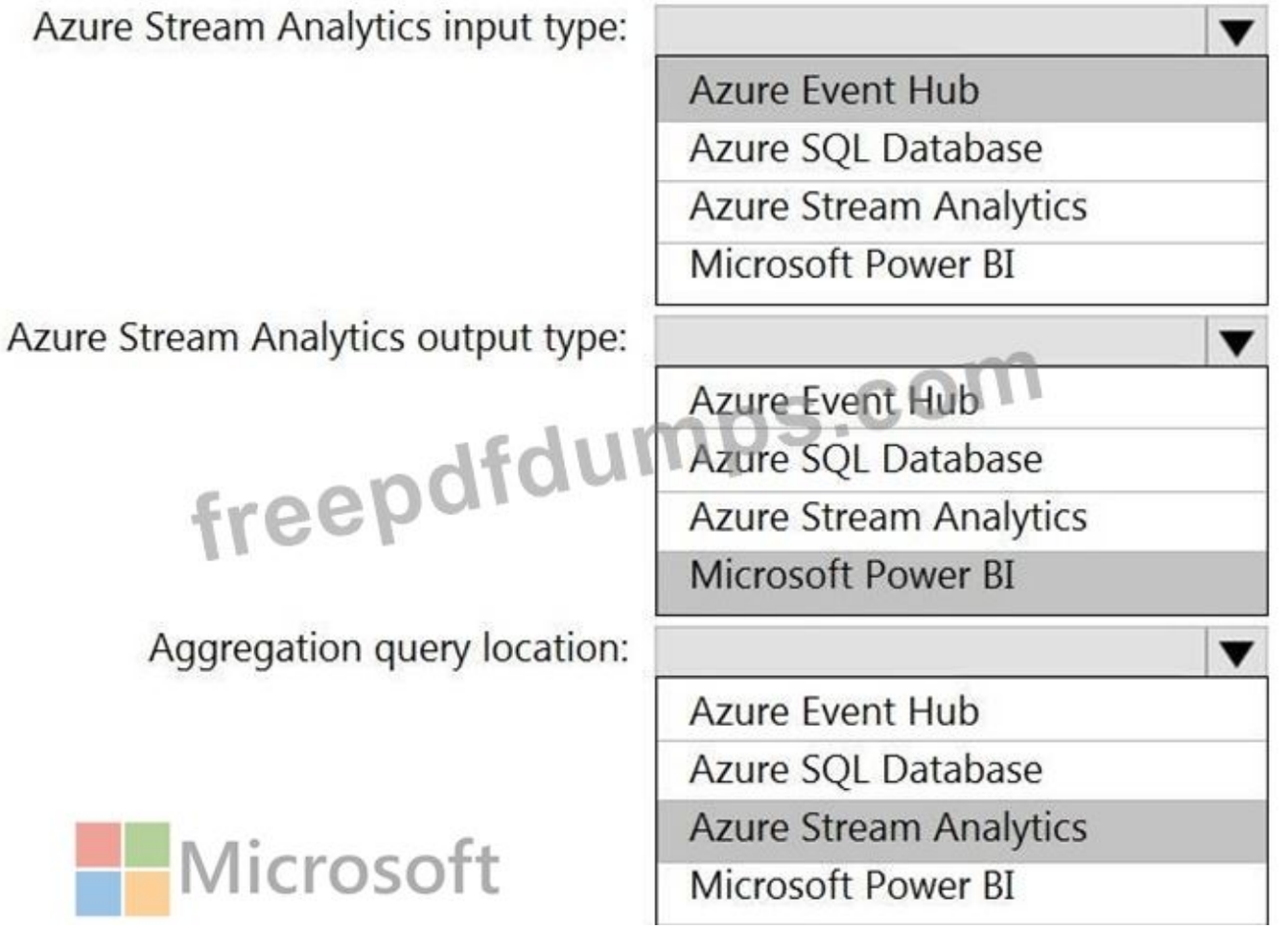



**Answer:**





Explanation





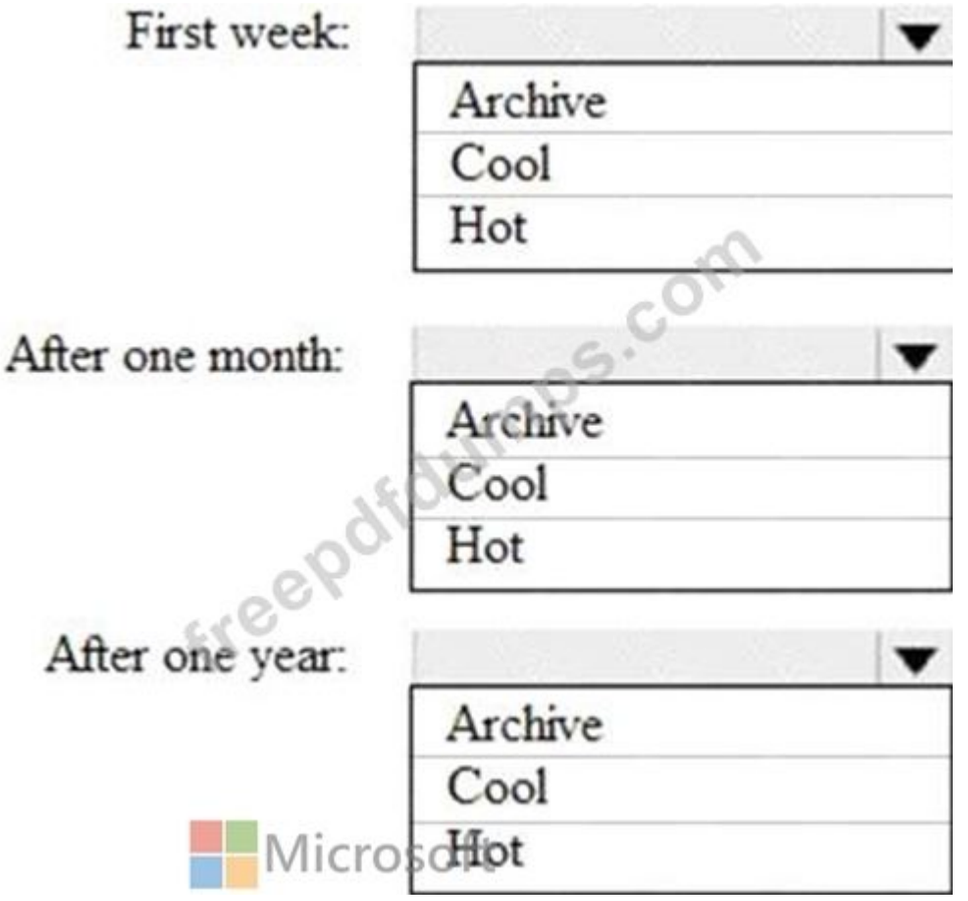
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-power-bi-dashboard

**NEW QUESTION: 57**
A company uses Azure Stream Analytics to monitor devices.
The company plans to double the number of devices that are monitored.
You need to monitor a Stream Analytics job to ensure that there are enough processing resources to handle the additional load.
Which metric should you monitor?
**A.** Early Input Events
**B.** Late Input Events
**C.** Watermark delay
**D.** Input Deserialization Errors
**Answer: A (LEAVE A REPLY)**
Explanation
There are a number of resource constraints that can cause the streaming pipeline to slow down.
The watermark delay metric can rise due to:
* Not enough processing resources in Stream Analytics to handle the volume of input events.
* Not enough throughput within the input event brokers, so they are throttled.

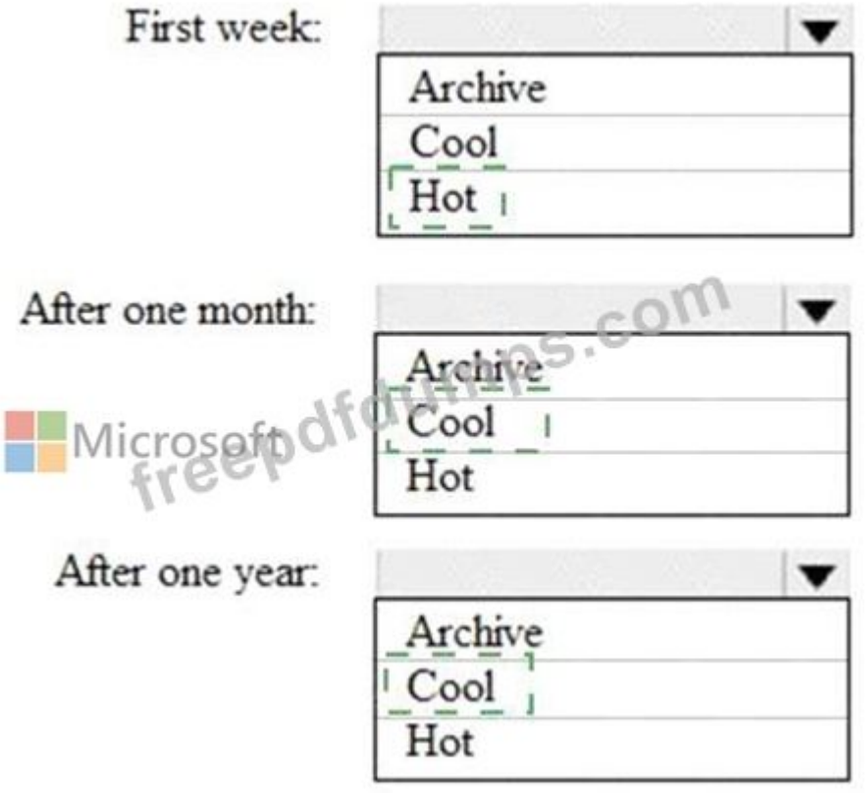
* Output sinks are not provisioned with enough capacity, so they are throttled. The possible solutions vary widely based on the flavor of output service being used.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling

**NEW QUESTION: 58**
You have an Azure Data lake Storage account that contains a staging zone.
You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes an Azure Databricks notebook, and then inserts the data into the data warehouse.
Dow this meet the goal?
**A.** Yes
**B.** No
**Answer: B (LEAVE A REPLY)**
Explanation
If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity, not an Azure Databricks notebook, with your own data processing logic and use the activity in the pipeline.
You can create a custom activity to run R scripts on your HDInsight cluster with R installed.
Reference:
https://docs.microsoft.com/en-US/azure/data-factory/transform-data

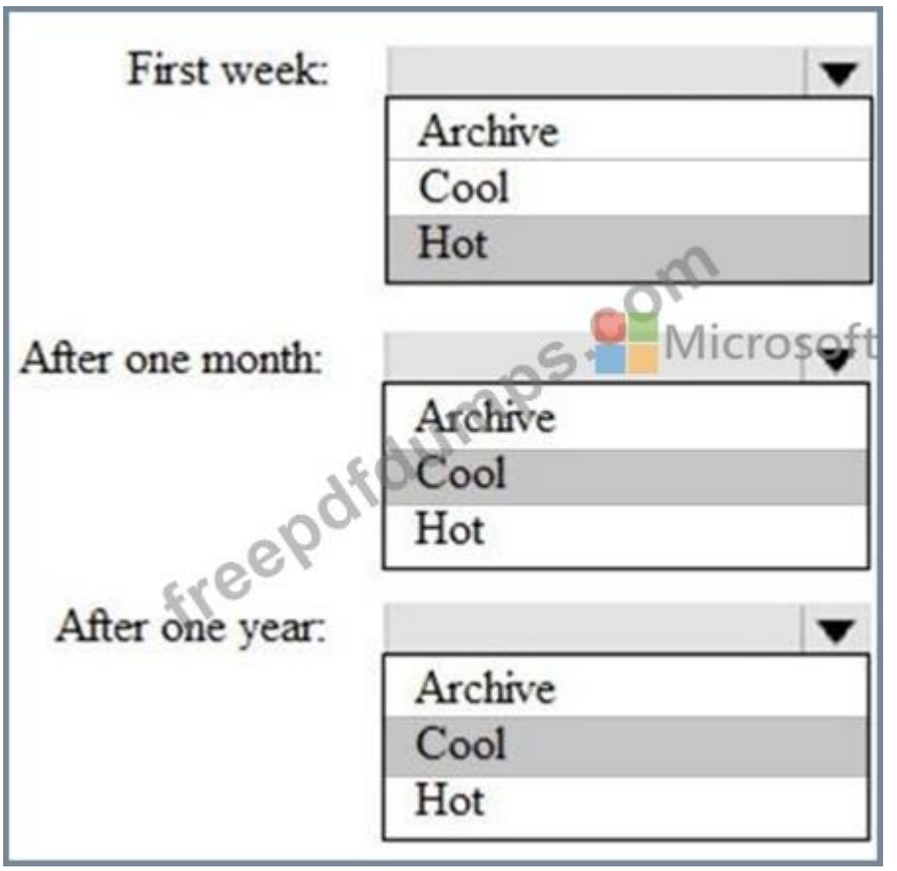
**NEW QUESTION: 59**
You are designing an application that will store petabytes of medical imaging data When the data is first created, the data will be accessed frequently during the first week. After one month, the data must be accessible within 30 seconds, but files will be accessed infrequently. After one year, the data will be accessed infrequently but must be accessible within five minutes.
You need to select a storage strategy for the data. The solution must minimize costs.
Which storage tier should you use for each time frame? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.




**Answer:**




Explanation




First week: Hot

Hot - Optimized for storing data that is accessed frequently.

After one month: Cool

Cool - Optimized for storing data that is infrequently accessed and stored for at least 30 days.

After one year: Cool

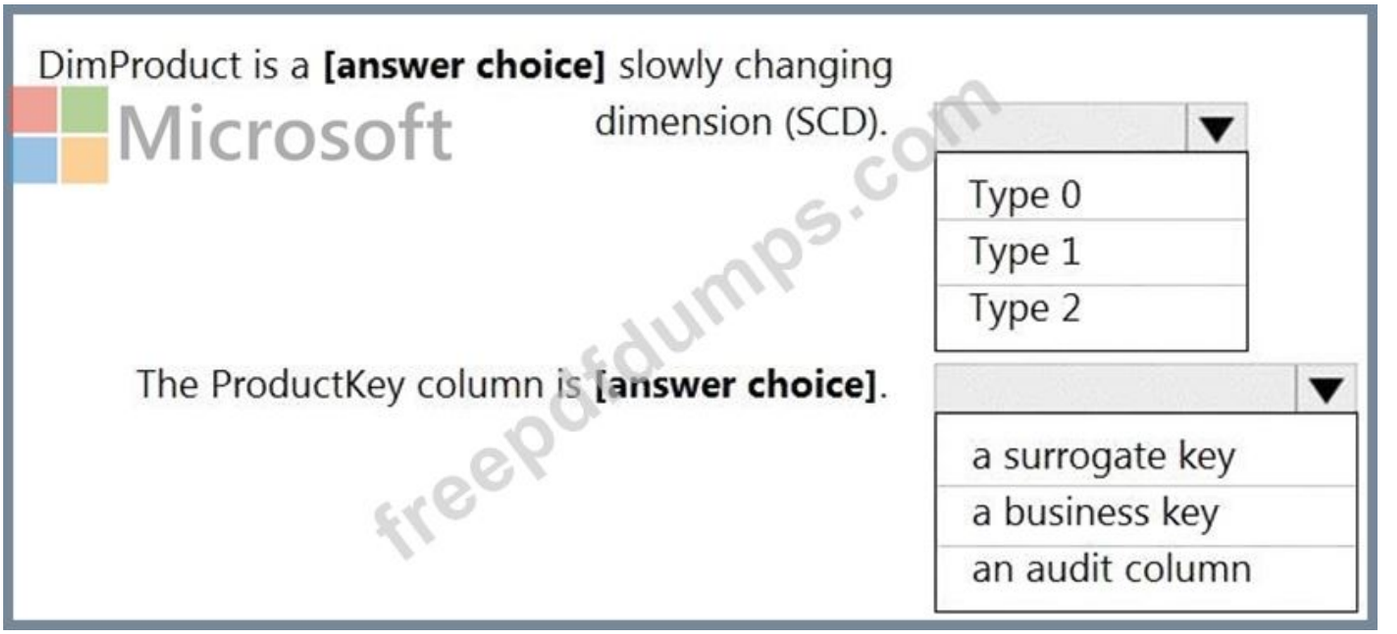
**NEW QUESTION: 60**

You are creating dimensions for a data warehouse in an Azure Synapse Analytics dedicated SQL pool.

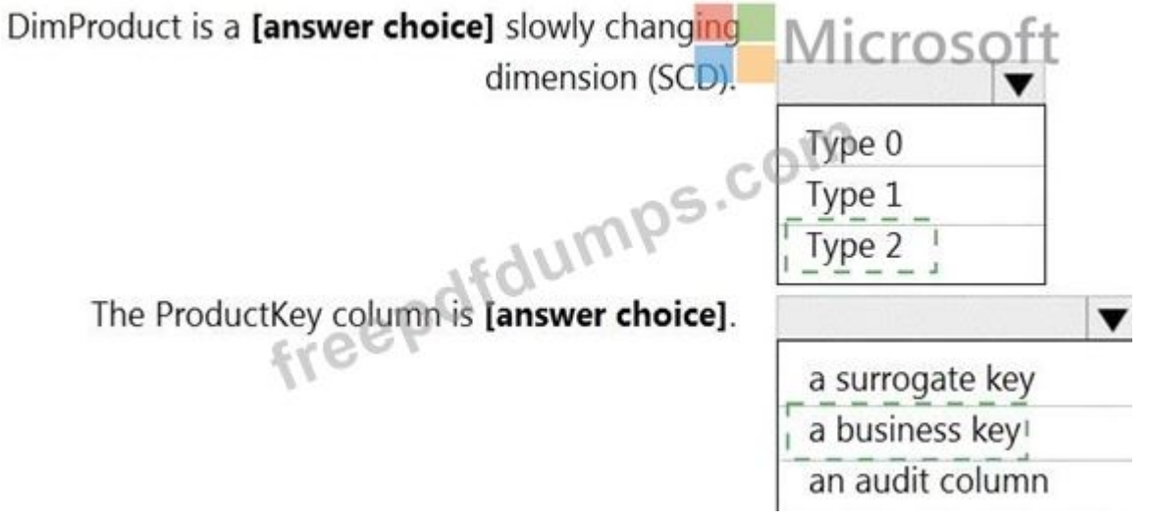
You create a table by using the Transact-SQL statement shown in the following exhibit.

```
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
```

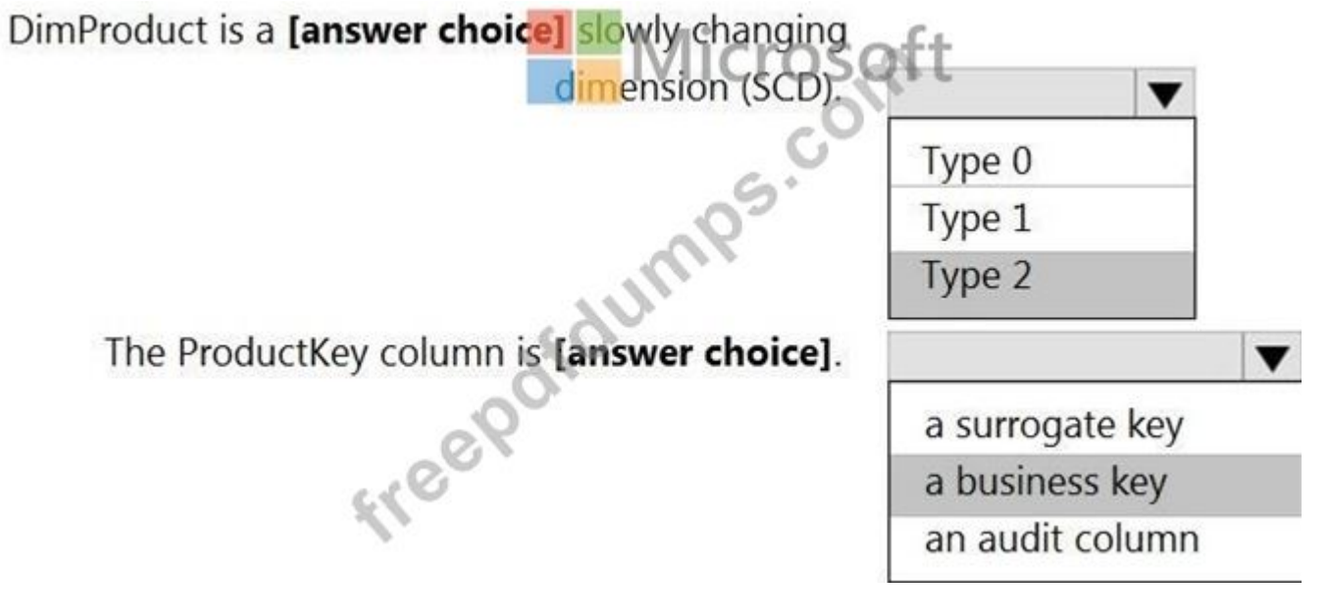
Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.




**Answer:**




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.

Reference:

https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-p

**NEW QUESTION: 61**

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.

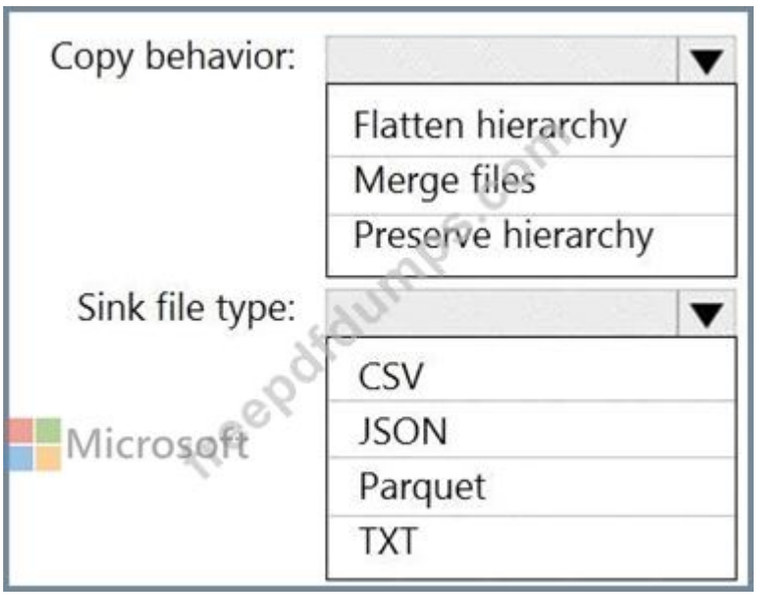
You need to move the files to a different folder and transform the data to meet the following requirements:

Provide the fastest possible query times.

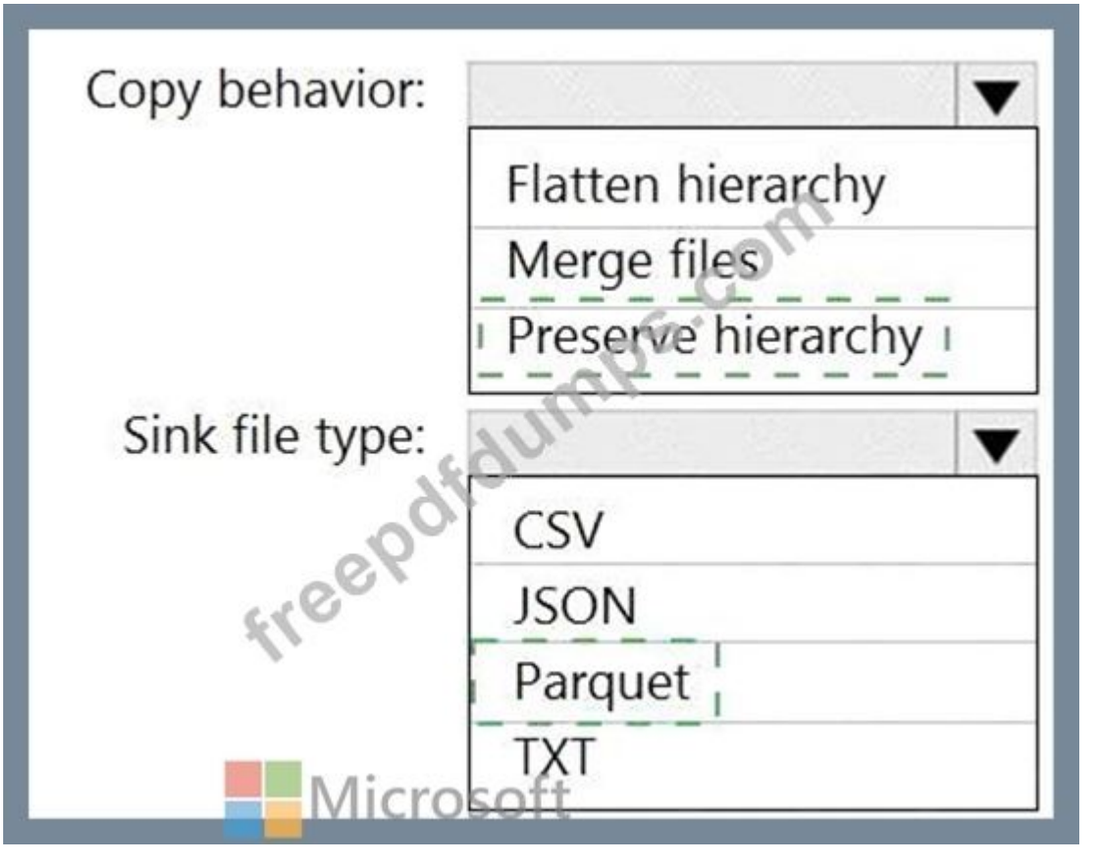
Automatically infer the schema from the underlying files.

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

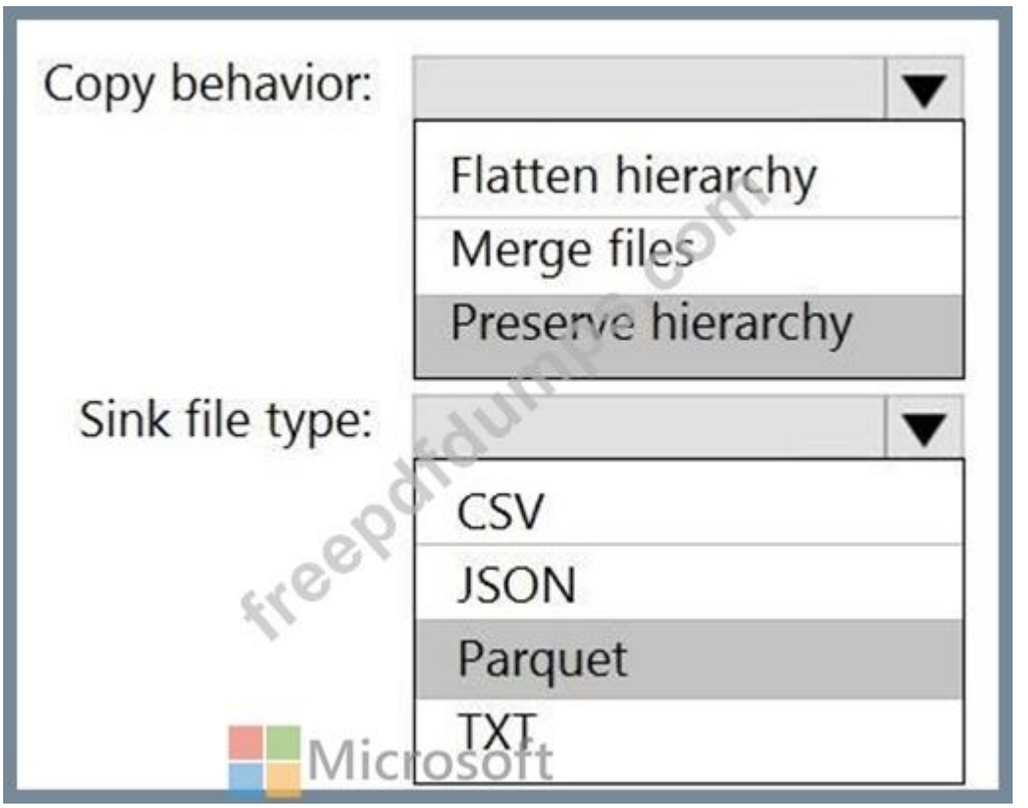
NOTE: Each correct selection is worth one point.




**Answer:**




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



**NEW QUESTION: 62**

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.


**Answer:**


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

**NEW QUESTION: 63**

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a table named FactInternetSales that will be a large fact table in a dimensional model.

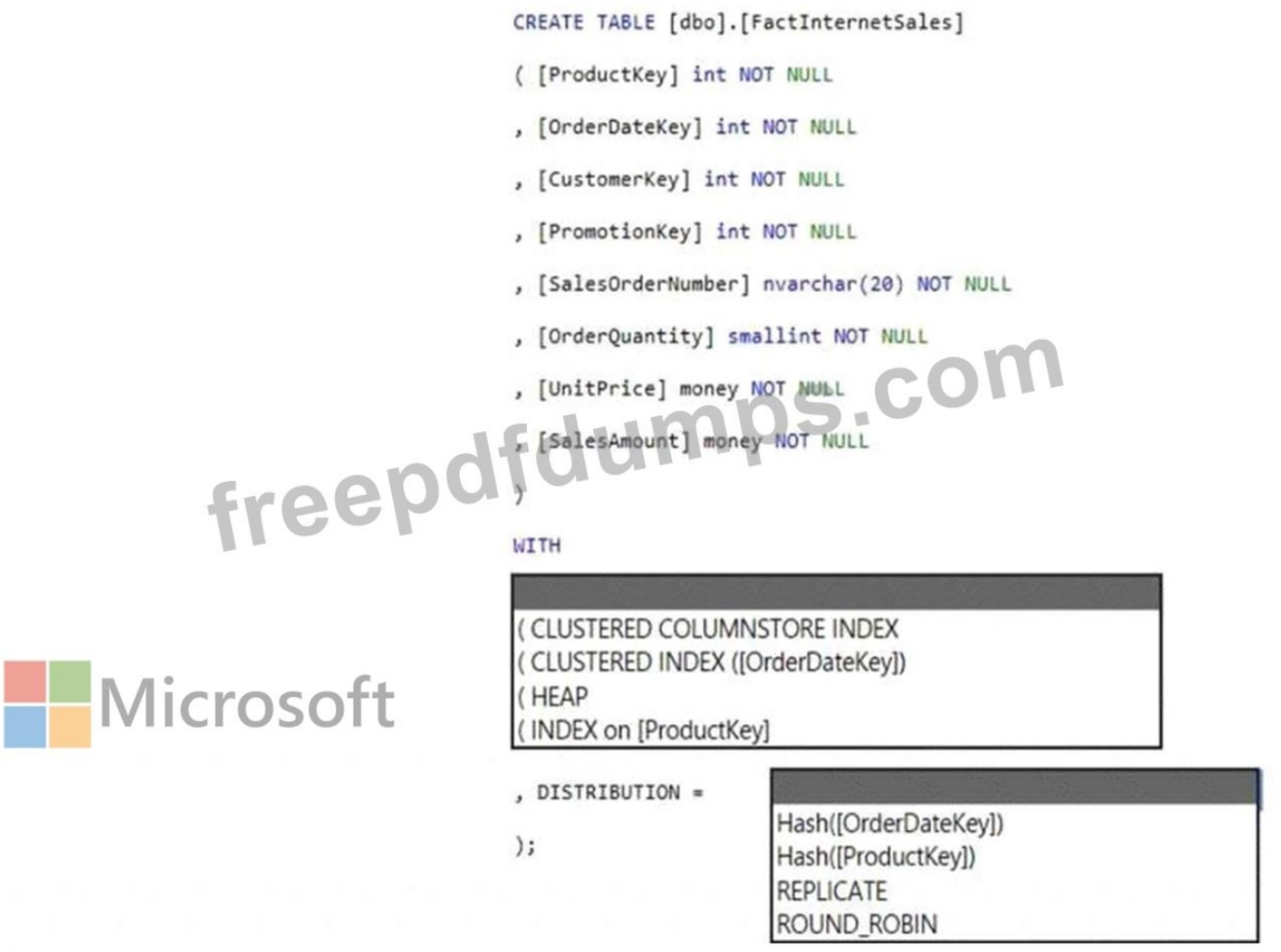
FactInternetSales will contain 100 million rows and two columns named SalesAmount and OrderQuantity.

Queries executed on FactInternetSales will aggregate the values in SalesAmount and OrderQuantity from the last year for a specific product. The solution must minimize the data size and query execution time.

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

NOTE: Each correct selection is worth one point.

**Answer Area**






**Answer:**

**Answer Area**

```
CREATE TABLE [dbo].[FactInternetSales]
( [ProductKey] int NOT NULL
, [OrderDateKey] int NOT NULL
, [CustomerKey] int NOT NULL
, [PromotionKey] int NOT NULL
, [SalesOrderNumber] nvarchar(20) NOT NULL
, [OrderQuantity] smallint NOT NULL
, [UnitPrice] money NOT NULL
, [SalesAmount] money NOT NULL
)
WITH
( CLUSTERED COLUMNSTORE INDEX
( CLUSTERED INDEX ([OrderDateKey])
( HEAP
( INDEX on [ProductKey]
, DISTRIBUTION =
Hash([OrderDateKey])
Hash([ProductKey])
REPLICATE
ROUND_ROBIN
);
```





Explanation

Box 1: (CLUSTERED COLUMNSTORE INDEX

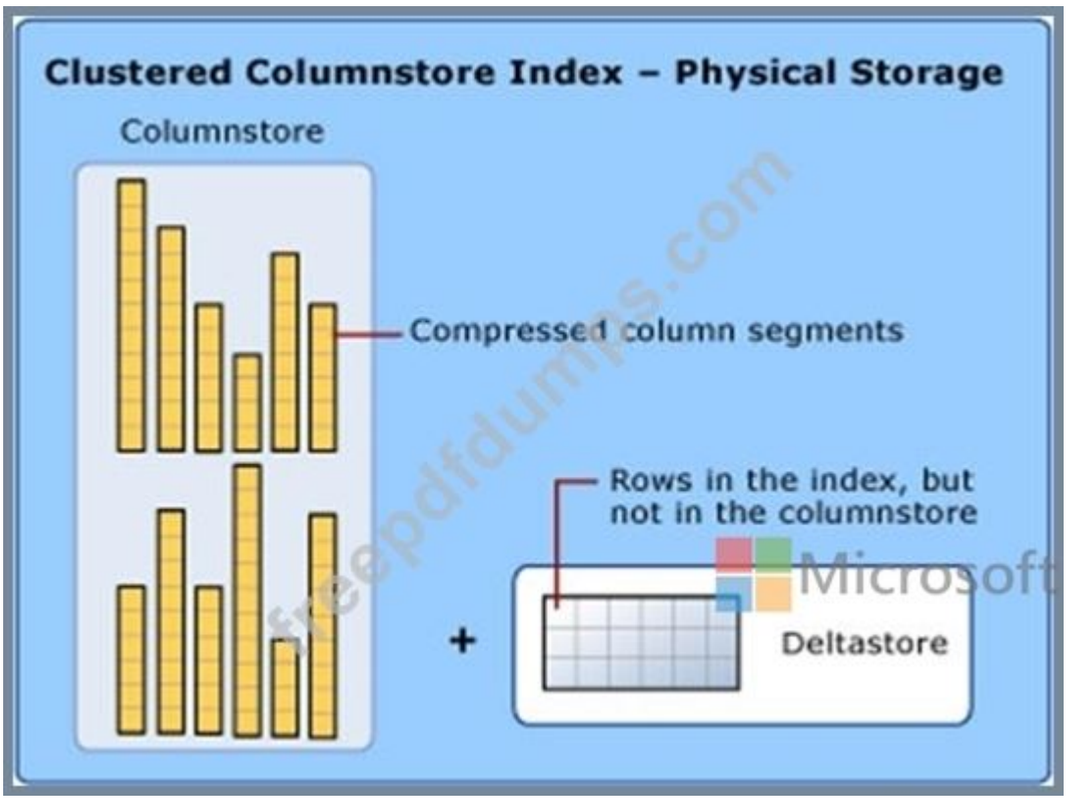
CLUSTERED COLUMNSTORE INDEX

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

Note: Clustered columnstore index

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

Diagram Description automatically generated

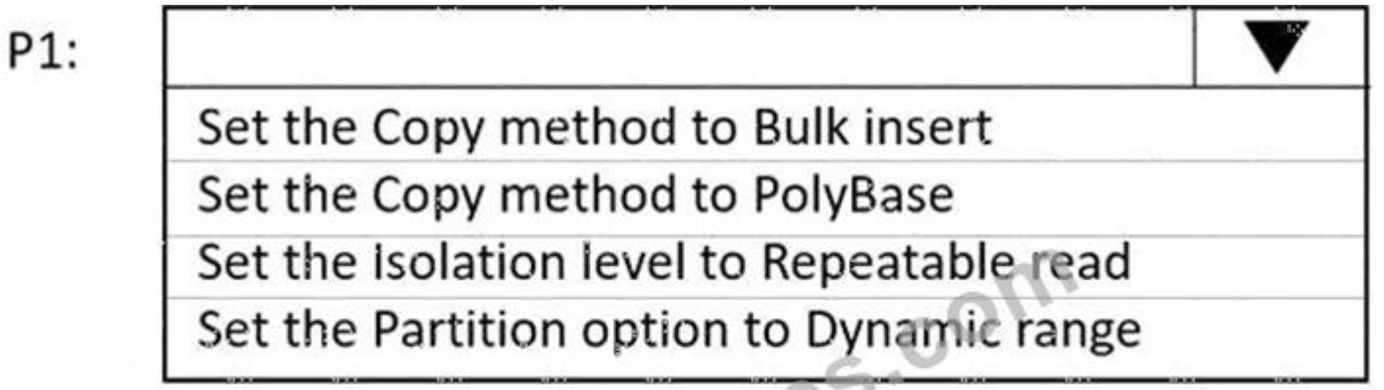


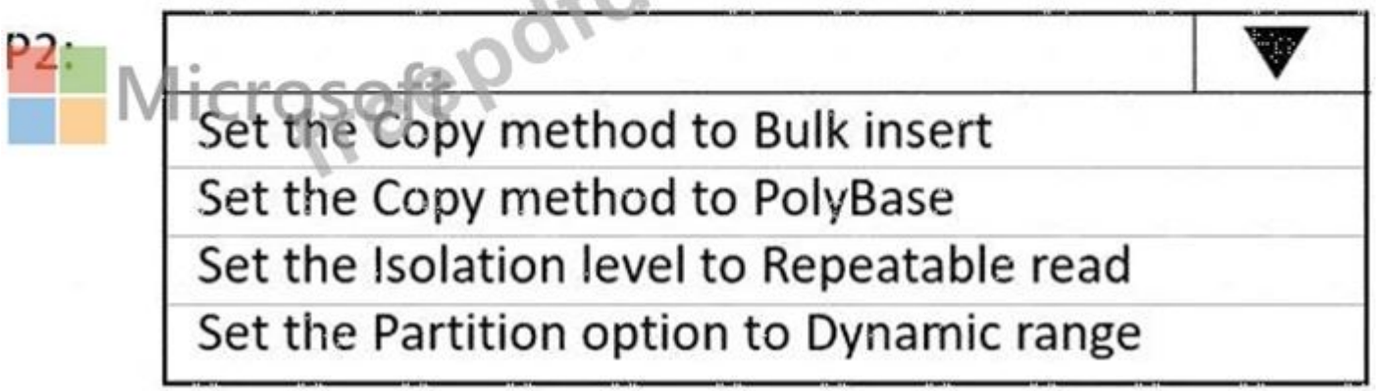
To reduce fragmentation of the column segments and improve performance, the columnstore index might store some data temporarily into a clustered index called a deltastore and a B-tree list of IDs for deleted rows. The deltastore operations are handled behind the scenes. To return the correct query results, the clustered columnstore index combines query results from both the columnstore and the deltastore.
Box 2: HASH([ProductKey])
A hash distributed table distributes rows based on the value in the distribution column. A hash distributed table is designed to achieve high performance for queries on large tables.
Choose a distribution column with data that distributes evenly
Reference: https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-overview
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overview
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

**NEW QUESTION: 64**
You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination.
You need to ensure that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs.
What should you do?
**A.** Clone the cluster after it is terminated.
**B.** Terminate the cluster manually when processing completes.

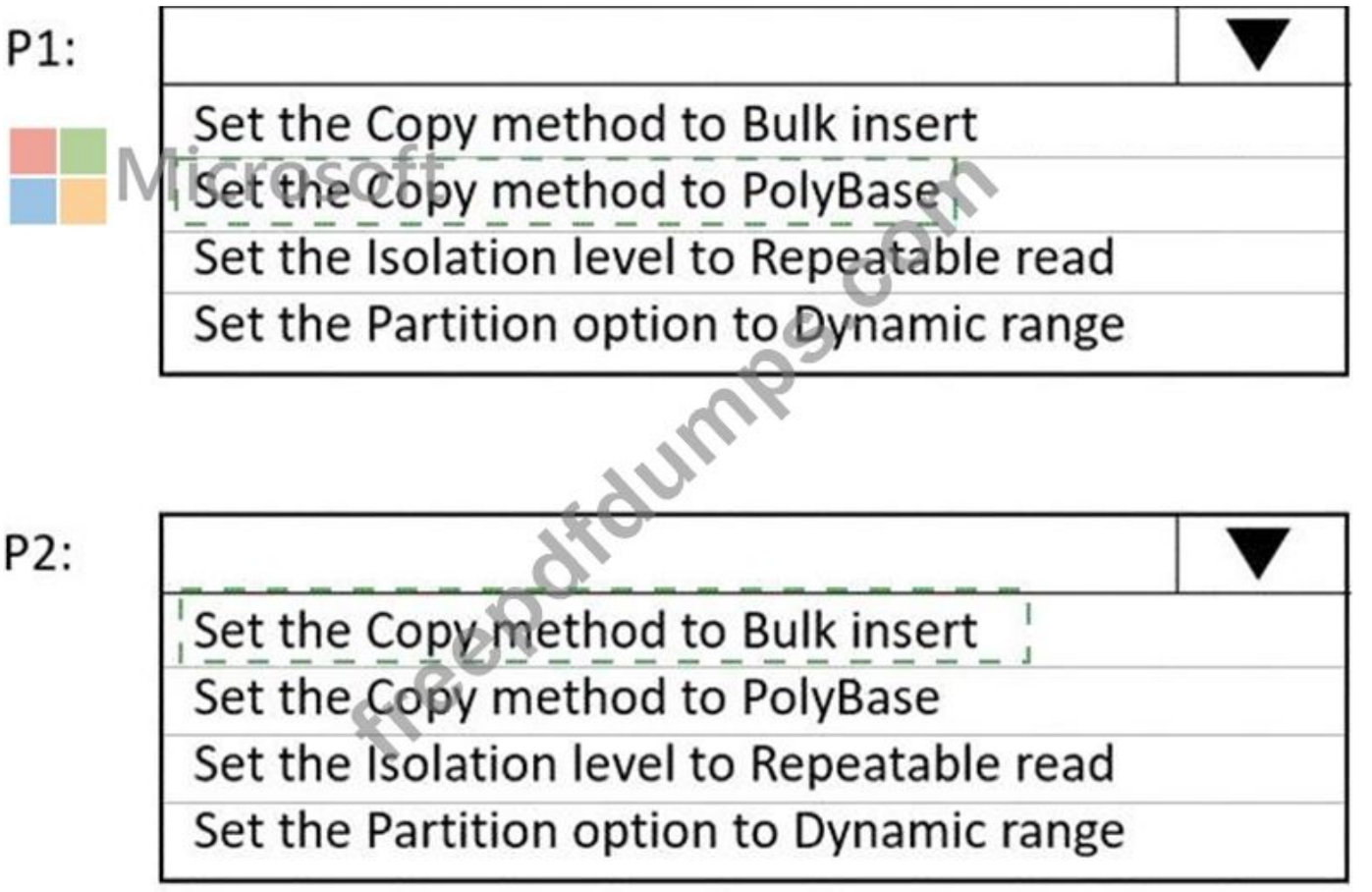
**C.** Create an Azure runbook that starts the cluster every 90 days.
**D.** Pin the cluster.
**Answer: D (LEAVE A REPLY)**
Explanation
To keep an interactive cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.
References:
https://docs.azuredatabricks.net/clusters/clusters-manage.html#automatic-termination

**NEW QUESTION: 65**
You have an Azure Data Factory instance named ADF1 and two Azure Synapse Analytics workspaces named WS1 and WS2.
ADF1 contains the following pipelines:
P1: Uses a copy activity to copy data from a nonpartitioned table in a dedicated SQL pool of WS1 to an Azure Data Lake Storage Gen2 account P2: Uses a copy activity to copy data from text-delimited files in an Azure Data Lake Storage Gen2 account to a nonpartitioned table in a dedicated SQL pool of WS2 You need to configure P1 and P2 to maximize parallelism and performance.
Which dataset settings should you configure for the copy activity if each pipeline? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

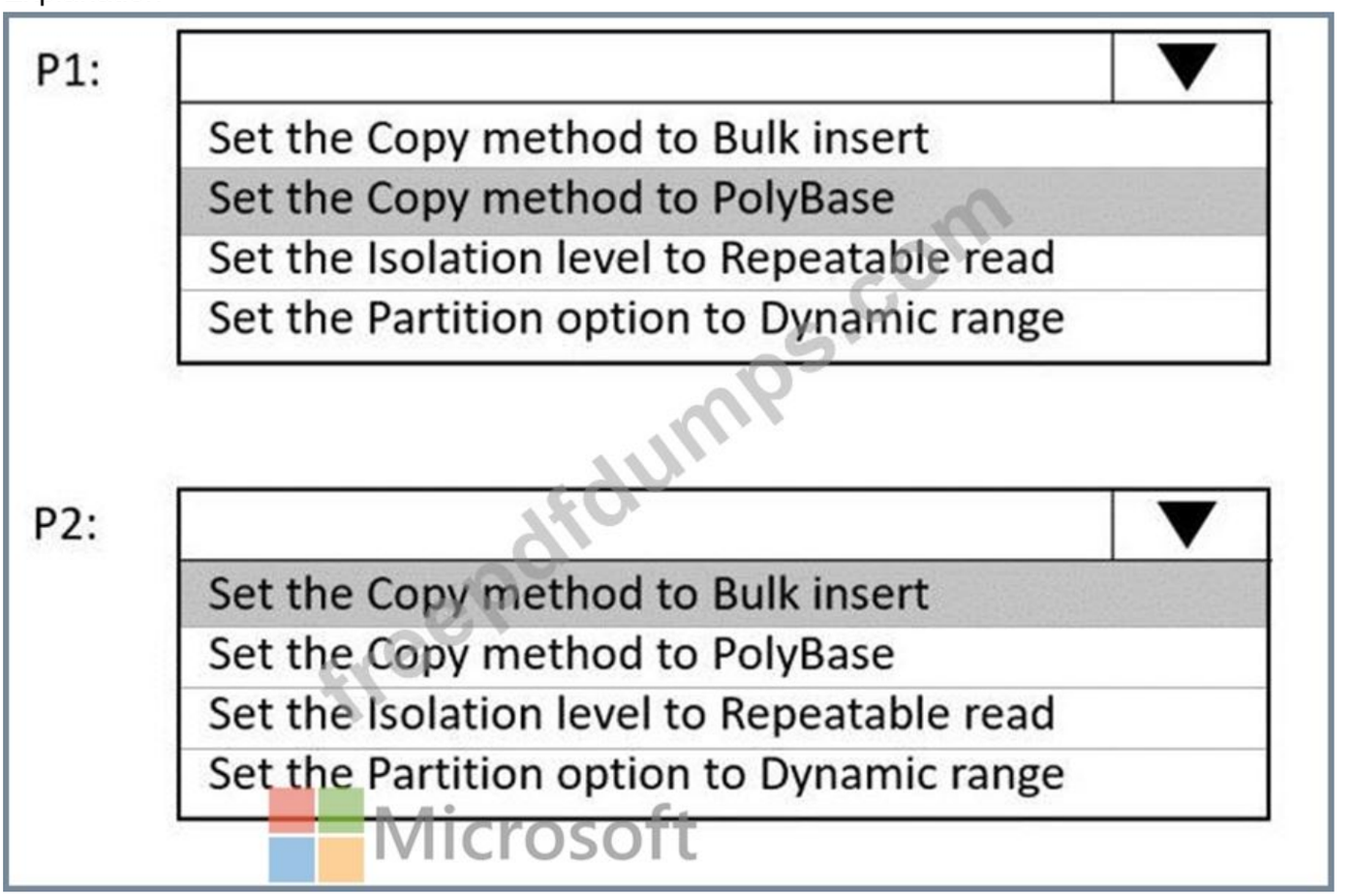


**Answer:**



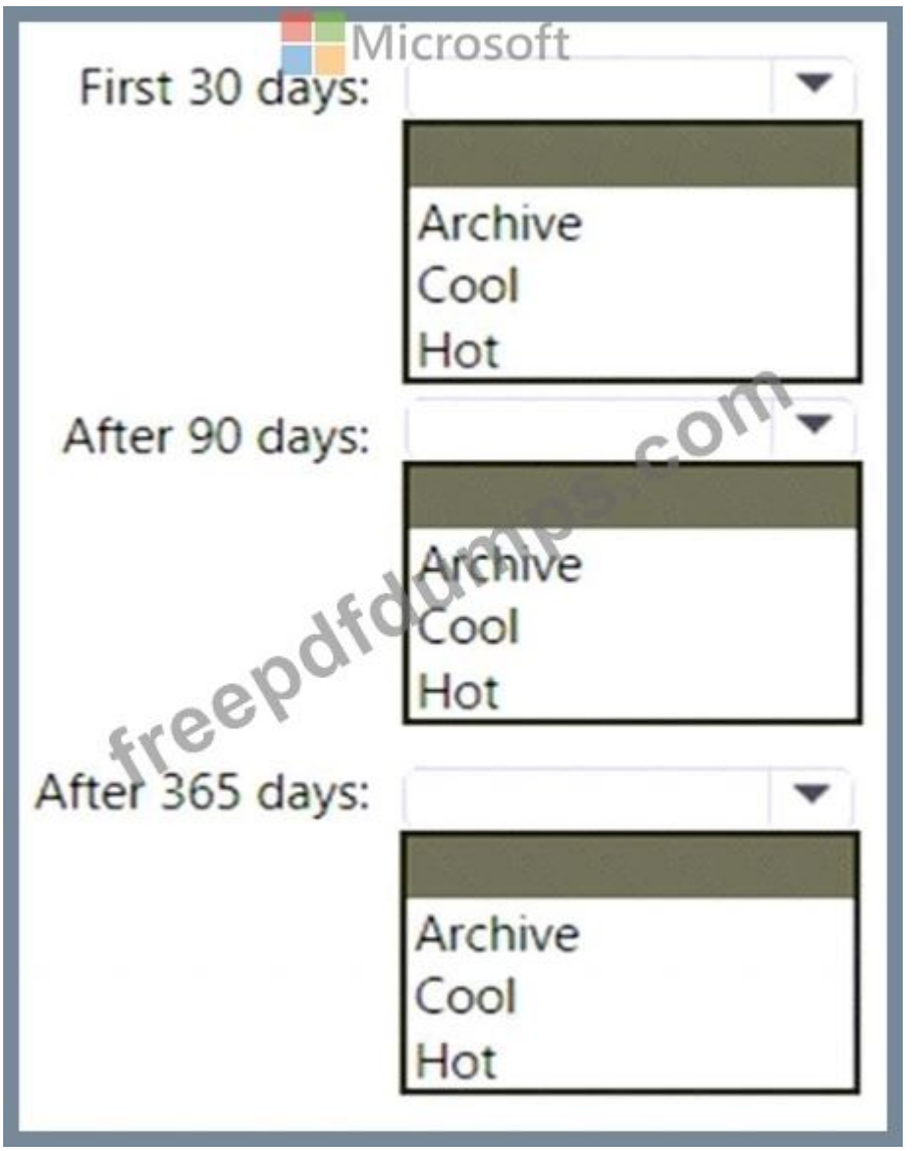

Explanation



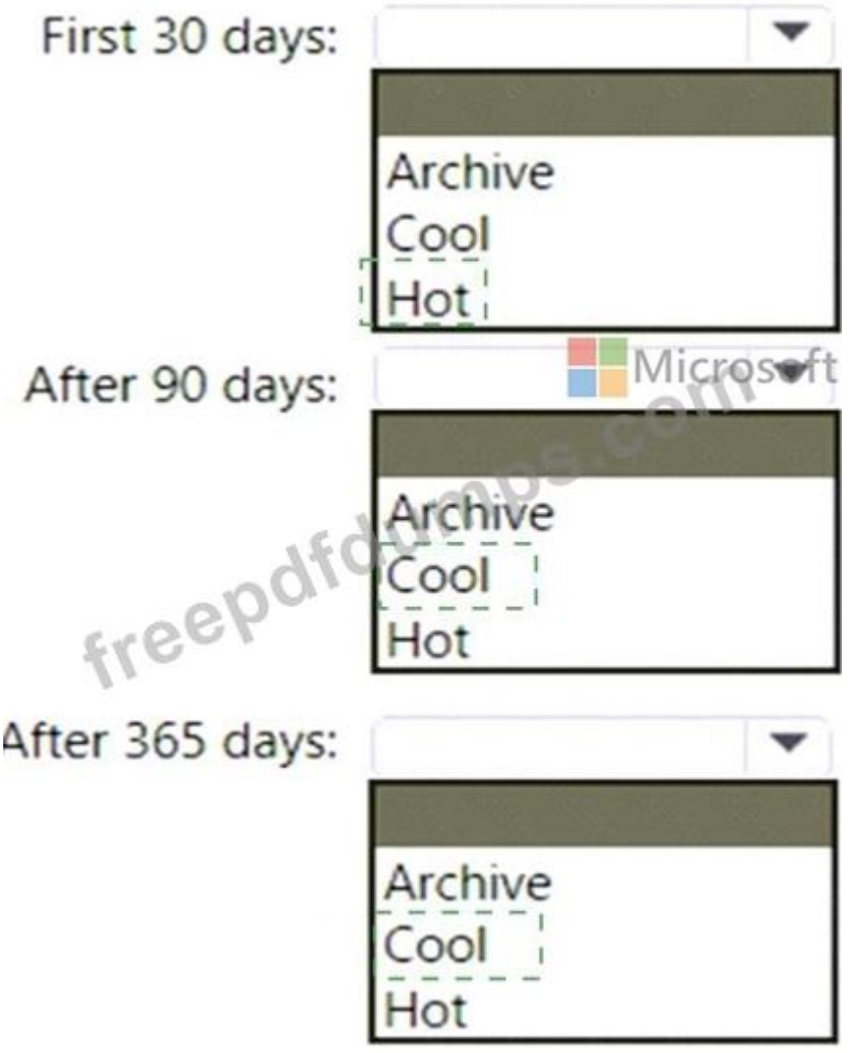
Box 1: Set the Copy method to PolyBase
While SQL pool supports many loading methods including non-Polybase options such as BCP and SQL BulkCopy API, the fastest and most scalable way to load data is through PolyBase. PolyBase is a technology that accesses external data stored in Azure Blob storage or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

**NEW QUESTION: 66**
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 vaul 1.
**B.** Enable Azure role-based access control on vault 1.
**C.** Remove the linked service from Df1.
**D.** Create a self-hosted integration runtime.
**Answer: C (LEAVE A REPLY)**
Explanation
Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources.
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

**NEW QUESTION: 67**
You are designing an application that will use an Azure Data Lake Storage Gen 2 account to store petabytes of license plate photos from toll booths. The account will use zone-redundant storage (ZRS).
You identify the following usage patterns:
* The data will be accessed several times a day during the first 30 days after the data is created. The data must meet an availability SU of 99.9%.
* After 90 days, the data will be accessed infrequently but must be available within 30 seconds.
* After 365 days, the data will be accessed infrequently but must be available within five minutes.


First 30 days:
Archive
Cool
Hot
After 90 days:
Archive
Cool
Hot
After 365 days:
Archive
Cool
Hot

**Answer:**




Explanation

Box 1: Hot

The data will be accessed several times a day during the first 30 days after the data is created. The data must meet an availability SLA of 99.9%.

Box 2: Cool

After 90 days, the data will be accessed infrequently but must be available within 30 seconds. Data in the Cool tier should be stored for a minimum of 30 days.

When your data is stored in an online access tier (either Hot or Cool), users can access it immediately. The Hot tier is the best choice for data that is in active use, while the Cool tier is ideal for data that is accessed less frequently, but that still must be available for reading and writing.

Box 3: Cool

After 365 days, the data will be accessed infrequently but must be available within five minutes.

Reference: https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview https://docs.microsoft.com/en-us/azure/storage/blobs/archive-rehydrate-overview

**NEW QUESTION: 68**

You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1.
You need to determine the size of the transaction log file for each distribution of DW1.
What should you do?
**A.** On DW1, execute a query against the sys.database_files dynamic management view.
**B.** From Azure Monitor in the Azure portal, execute a query against the logs of DW1.
**C.** Execute a query against the logs of DW1 by using the Get-AzOperationalInsightsSearchResult PowerShell cmdlet.
**D.** On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.
**Answer: A (LEAVE A REPLY)**
Explanation
For information about the current log file size, its maximum size, and the autogrow option for the file, you can also use the size, max_size, and growth columns for that log file in sys.database_files.
Reference:
https://docs.microsoft.com/en-us/sql/relational-databases/logs/manage-the-size-of-the-transaction-log-file

**NEW QUESTION: 69**
You haw an Azure data factory named ADF1.
You currently publish all pipeline authoring changes directly to ADF1.
You need to implement version control for the changes made to pipeline artifacts. The solution must ensure that you can apply version control to the resources currently defined m the UX Authoring canvas for ADF1.
Which two actions should you perform? Each correct answer presents part of the solution NOTE: Each correct selection is worth one point.
**A.** Create an Azure Data Factory trigger
**B.** From the UX Authoring canvas, select Set up code repository
**C.** Create a GitHub action
**D.** From the UX Authoring canvas, run Publish All.
**E.** Create a Git repository
**F.** From the UX Authoring canvas, select Publish
**Answer: D,E (LEAVE A REPLY)**
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/source-control

**NEW QUESTION: 70**
You plan to create a table in an Azure Synapse Analytics dedicated SQL pool.
Data in the table will be retained for five years. Once a year, data that is older than five years will be deleted.

You need to ensure that the data is distributed evenly across partitions. The solution must minimize the amount of time required to delete old data.

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

NOTE: Each correct selection is worth one point.

**Values**

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

**Answer Area**

```
CREATE TABLE [dbo].[FactSales]
(
    [ProductKey]        int       NOT NULL
,   [OrderDateKey]      int       NOT NULL
,   [CustomerKey]       int       NOT NULL
,   [SalesOrderNumber] nvarchar ( 20 )   NOT NULL
,   [OrderQuantity]          smallint     NOT NULL
,   [UnitPrice]              money        NOT NULL
)
WITH
(   CLUSTERED       COLUMNSTORE       INDEX
,   DISTRIBUTION =  [ Value ]  ([ProductKey])
,   PARTITION   (  [ Value ] ] RANGE RIGHT FOR VALUES
                (20170101,20180101,20190101,20200101,20210101)
                )
)
```




**Answer:**

**Values**

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

**Answer Area**

```
CREATE TABLE [dbo].[FactSales]
(
    [ProductKey]        int       NOT NULL
,   [OrderDateKey]      int       NOT NULL
,   [CustomerKey]       int       NOT NULL
,   [SalesOrderNumber] nvarchar ( 20 )   NOT NULL
,   [OrderQuantity]          smallint     NOT NULL
,   [UnitPrice]              money        NOT NULL
)
WITH
(   CLUSTERED       COLUMNSTORE       INDEX
,   DISTRIBUTION =  HASH  ([ProductKey])
,   PARTITION   (  [ OrderDateKey ] RANGE RIGHT FOR VALUES
                (20170101,20180101,20190101,20200101,20210101)
                )
)
```



Explanation

Box 1: HASH
Box 2: OrderDateKey
In most cases, table partitions are created on a date column.
A way to eliminate rollbacks is to use Metadata Only operations like partition switching for data management.
For example, rather than execute a DELETE statement to delete all rows in a table where the order_date was in October of 2001, you could partition your data early. Then you can switch out the partition with data for an empty partition from another table.
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

**NEW QUESTION: 71**
You have an Azure Data Factory pipeline named Pipeline1!. Pipelinel contains a copy activity that sends data to an Azure Data Lake Storage Gen2 account. Pipeline 1 is executed by a schedule trigger.
You change the copy activity sink to a new storage account and merge the changes into the collaboration branch.
After Pipelinel executes, you discover that data is NOT copied to the new storage account.
You need to ensure that the data is copied to the new storage account.
What should you do?
**A.** Publish from the collaboration branch.
**B.** Configure the change feed of the new storage account.
**C.** Create a pull request.
**D.** Modify the schedule trigger.
**Answer: A (LEAVE A REPLY)**
Explanation
CI/CD lifecycle
A development data factory is created and configured with Azure Repos Git. All developers should have permission to author Data Factory resources like pipelines and datasets.
A developer creates a feature branch to make a change. They debug their pipeline runs with their most recent changes After a developer is satisfied with their changes, they create a pull request from their feature branch to the main or collaboration branch to get their changes reviewed by peers.
After a pull request is approved and changes are merged in the main branch, the changes get published to the development factory.
Reference: https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-delivery

**NEW QUESTION: 72**
You are implementing a batch dataset in the Parquet format.

Data tiles will be produced by using Azure Data Factory and stored in Azure Data Lake Storage Gen2. The files will be consumed by an Azure Synapse Analytics serverless SQL pool.
You need to minimize storage costs for the solution.
What should you do?
**A.** Store all the data as strings in the Parquet tiles.
**B.** Use OPENROWEST to query the Parquet files.
**C.** Create an external table mat contains a subset of columns from the Parquet files.
**D.** Use Snappy compression for the files.
**Answer: C (LEAVE A REPLY)**
Explanation
An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless SQL pool.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

**NEW QUESTION: 73**
You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1.
You need to use an Azure Synapse pipeline to copy data from the selected tables in DB1 to the files in storage1. The solution must meet the following requirements:
* The Copy activity in the pipeline must be parameterized to use the data in File1.txt to identify the source and destination of the copy.
* Copy activities must occur in parallel as often as possible.
Which two pipeline activities should you include in the pipeline? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.
**A.** If Condition
**B.** ForEach
**C.** Lookup
**D.** Get Metadata
**Answer: (SHOW ANSWER)**
Lookup: This is a control activity that retrieves a dataset from any of the supported data sources and makes it available for use by subsequent activities in the pipeline. You can use a Lookup activity to read File1.txt from storage1 and store its content as an array variable ForEach: This is a control activity that iterates over a collection and executes specified activities in a loop. You can use a ForEach activity to loop over the array variable from the Lookup activity and pass each table name as a parameter to a Copy activity that copies data from DB1 to storage11.

**NEW QUESTION: 74**

You have an Azure Databricks workspace and an Azure Data Lake Storage Gen2 account named storage!
New files are uploaded daily to storage1.
* Incrementally process new files as they are upkorage1 as a structured streaming source. The solution must meet the following requirements:
* Minimize implementation and maintenance effort.
* Minimize the cost of processing millions of files.
* Support schema inference and schema drift.
Which should you include in the recommendation?
**A.** Apache Spark FileStreamSource
**B.** COPY INTO
**C.** Azure Data Factory
**D.** Auto Loader
**Answer: (SHOW ANSWER)**

**NEW QUESTION: 75**
You are designing a sales transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will contains approximately 60 million rows per month and will be partitioned by month. The table will use a clustered column store index and round-robin distribution.
Approximately how many rows will there be for each combination of distribution and partition?
**A.** 1 million
**B.** 5 million
**C.** 20 million
**D.** 60 million
**Answer: (SHOW ANSWER)**
Explanation
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition

**NEW QUESTION: 76**
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 |

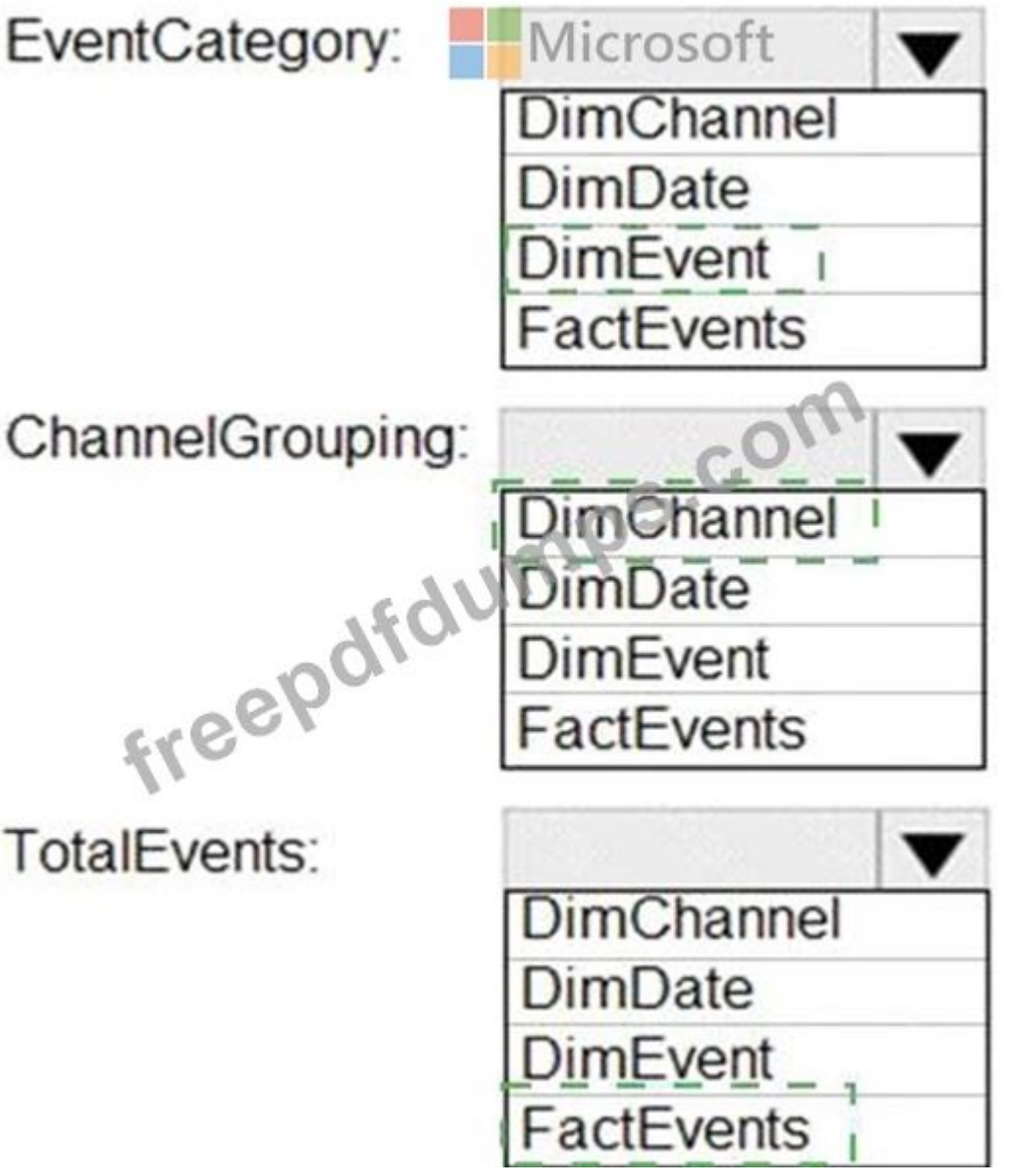
You need to design a star schema to support analytical queries of the data. The star schema will contain four tables including a date dimension.

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.




**Answer:**






Explanation

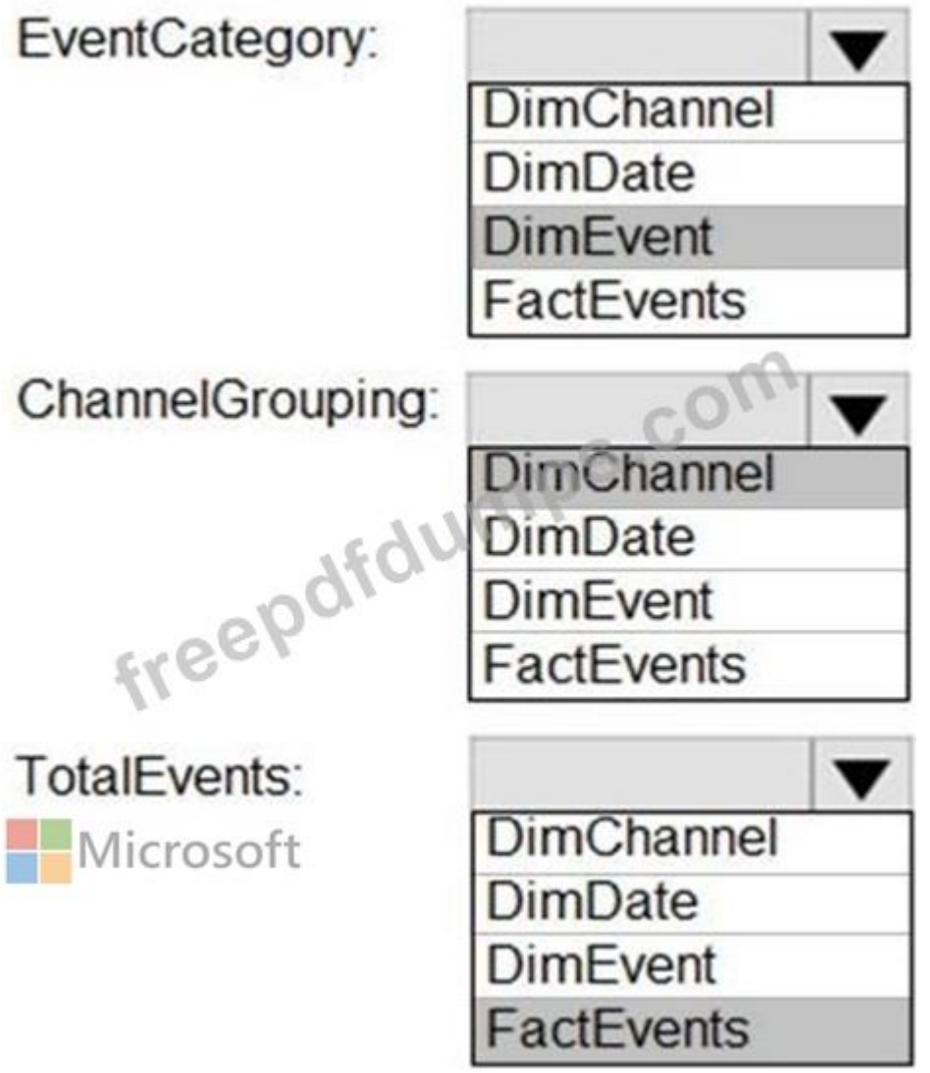



Table Description automatically generated

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



**NEW QUESTION: 77**

You are designing a data mart for the human resources (MR) department at your company. The data mart will contain information and employee transactions. From a source system you have a flat extract that has the following fields:

* EmployeeID
* FirstName
* LastName
* Recipient
* GrossArnount
* TransactionID
* GovernmentID
* NetAmountPaid
* TransactionDate

You need to design a start schema data model in an Azure Synapse analytics dedicated SQL pool for the data mart.

Which two tables should you create? Each Correct answer present part of the solution.

**A.** a dimension table for employee

**B.** a fabric for Employee

**C.** a dimension table far EmployeeTransaction

**D.** a dimension table for Transaction

**E.** a fact table for Transaction

**Answer: A,E (LEAVE A REPLY)**

Reference:

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

**NEW QUESTION: 78**

You are implementing an Azure Stream Analytics solution to process event data from devices. The devices output events when there is a fault and emit a repeat of the event every five seconds until the fault is resolved. The devices output a heartbeat event every five seconds after a previous event if there are no faults present.

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

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



You need to calculate the uptime between the faults.

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

NOTE: Each correct selection is worth one point.

```
SELECT

DeviceID,

MIN(EventTime) as StartTime,

MAX(EventTime) as EndTime,

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

FROM input TIMESTAMP BY EventTime
```

| |
|---|
| WHERE EventType='HeartBeat' |
| WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType |
| WHERE IsFirst(second,5) = 1 |

```
GROUP BY

DeviceID
```

| |
|---|
| ,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID) |
| ,TumblingWindow(second,5) |
| HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5 |

**Answer:**

```
SELECT

DeviceID,

MIN(EventTime) as StartTime,

MAX(EventTime) as EndTime,

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

FROM input TIMESTAMP BY EventTime
```

[ ▼ ]
- WHERE EventType='HeartBeat'
- WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType
- WHERE IsFirst(second,5) = 1

```
GROUP BY

DeviceID
```

[ ▼ ]
- ,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID)
- ,TumblingWindow(second,5)
- HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5



Explanation

Graphical user interface, text, application Description automatically generated

```
SELECT

DeviceID,

MIN(EventTime) as StartTime,

MAX(EventTime) as EndTime,

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

FROM input TIMESTAMP BY EventTime

[ WHERE EventType='HeartBeat' | WHERE LAG(EventType, 1) OVER (LIMIT DURATION(second,5)) <> EventType | WHERE IsFirst(second,5) = 1 ]

GROUP BY

DeviceID

[ ,SessionWindow(second, 5, 50000) OVER (PARTITION BY DeviceID) | ,TumblingWindow(second,5) | HAVING DATEDIFF(second, MIN(EventTime), MAX(EventTime)) > 5 ]
```



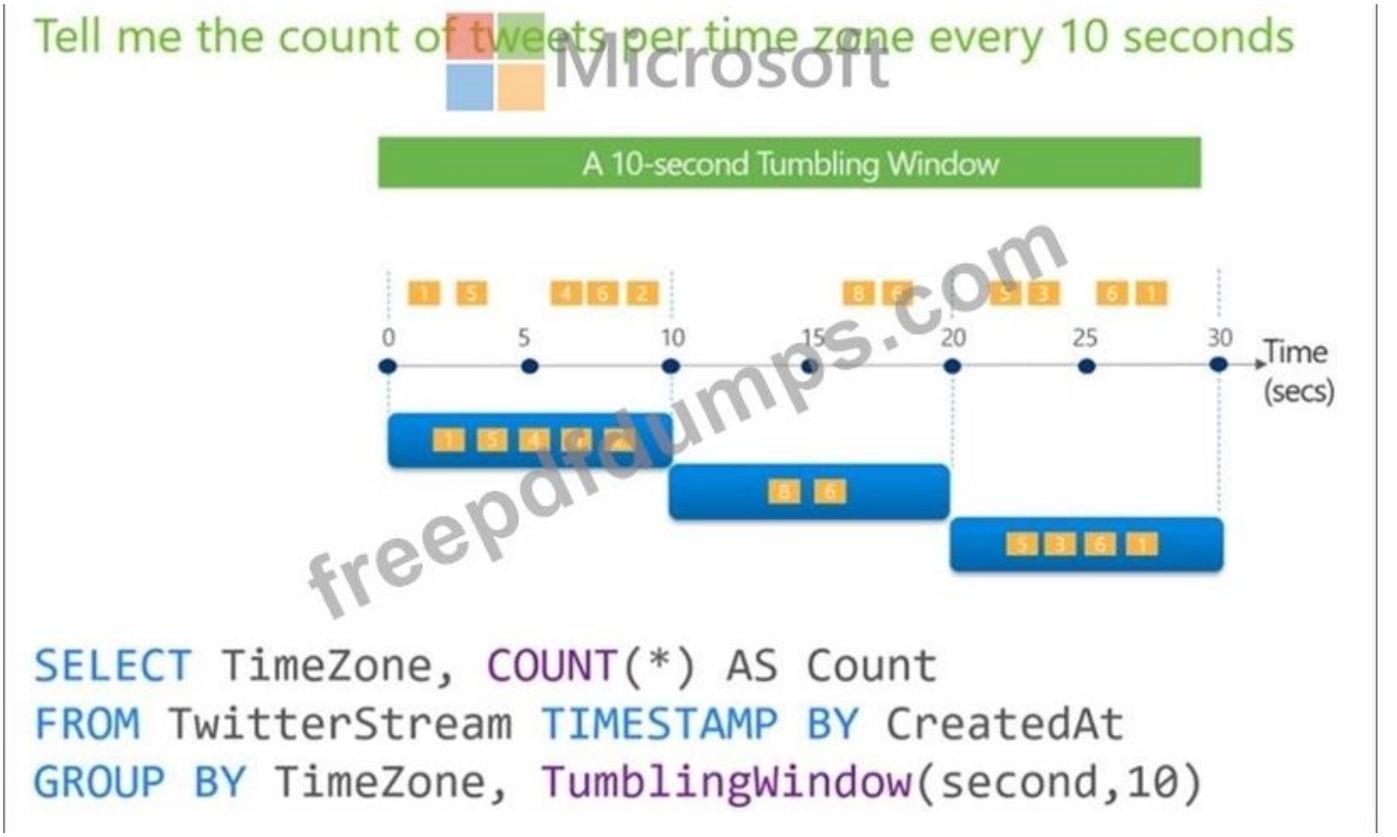
Box 1: WHERE EventType='HeartBeat'

Box 2: ,TumblingWindow(Second, 5)

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

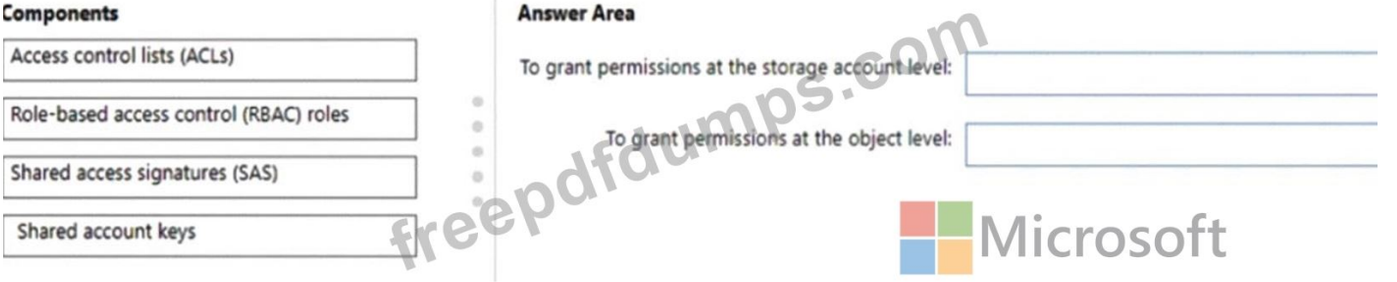
The following diagram illustrates a stream with a series of events and how they are mapped into 10-second tumbling windows.

Timeline Description automatically generated

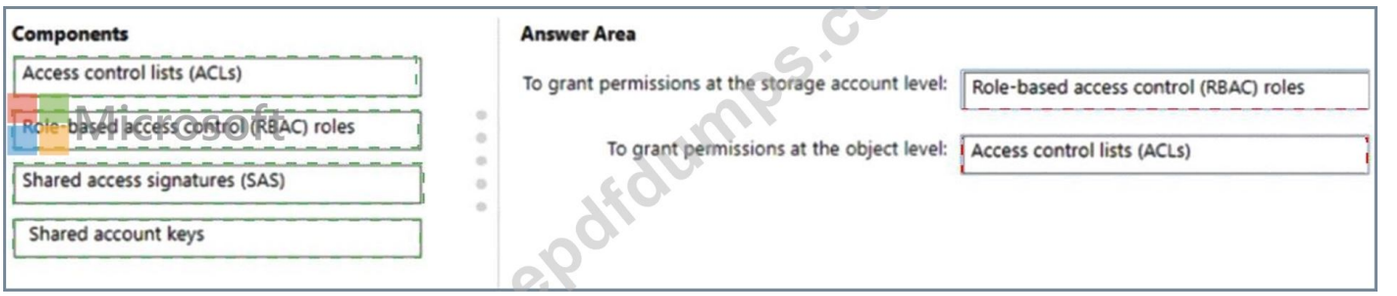



Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/session-window-azure-stream-analytics
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

**NEW QUESTION: 79**
You have an Azure Data Lake Storage Gen 2 account named storage1.
You need to recommend a solution for accessing the content in storage1. The solution must meet the following requirements:
List and read permissions must be granted at the storage account level.
Additional permissions can be applied to individual objects in storage1.
Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.
What should you use? To answer, drag the appropriate components to the correct requirements. Each component may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.




**Answer:**




Explanation

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

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

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

Role-based access control (Azure RBAC)

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

Box 2: Access control lists (ACLs)

Additional permissions can be applied to individual objects in storage1.

Access control lists (ACLs)

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

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

**NEW QUESTION: 80**

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named storage1.

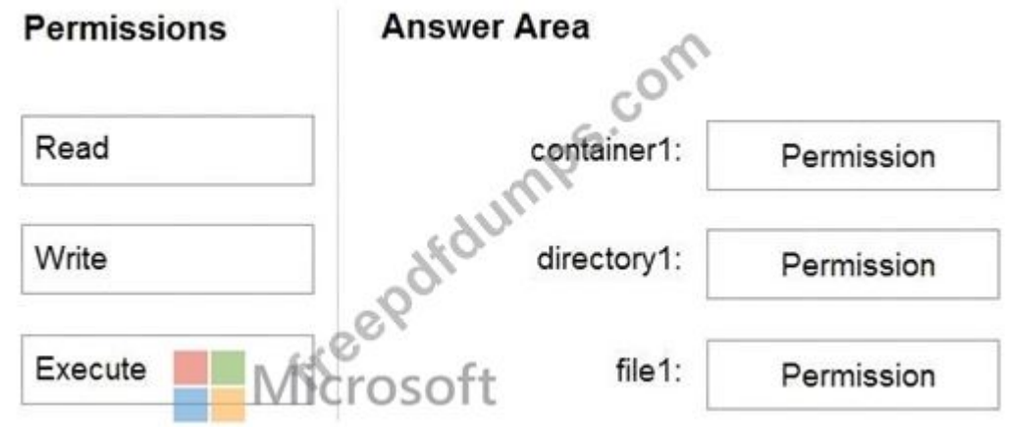
Storage1 contains a container named container1. Container1 contains a directory named directory1. Directory1 contains a file named file1.

You have an Azure Active Directory (Azure AD) user named User1 that is assigned the Storage Blob Data Reader role for storage1.

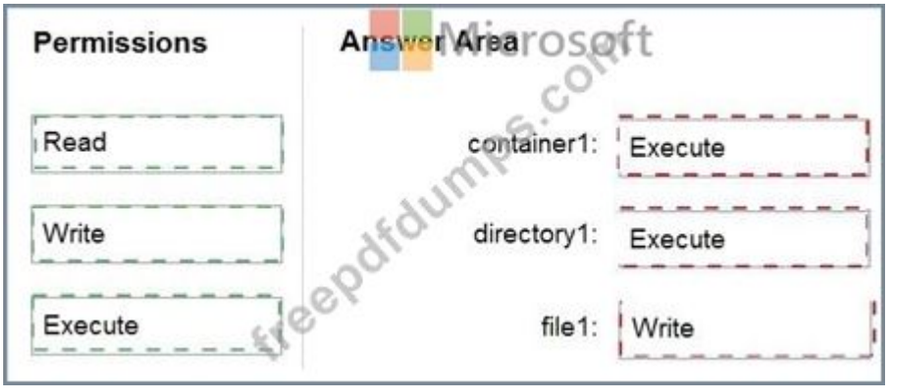
You need to ensure that User1 can append data to file1. The solution must use the principle of least privilege.

Which permissions should you grant? To answer, drag the appropriate permissions to the correct resources.

Each permission may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.




**Answer:**




Explanation

Box 1: Execute

If you are granting permissions by using only ACLs (no Azure RBAC), then to grant a security principal read or write access to a file, you'll need to give the security principal Execute permissions to the root folder of the container, and to each folder in the hierarchy of folders that lead to the file.

Box 2: Execute

On Directory: Execute (X): Required to traverse the child items of a directory Box 3: Write On file: Write (W): Can write or append to a file.

Reference:

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

**NEW QUESTION: 81**

You have a SQL pool in Azure Synapse that contains a table named dbo.Customers. The table contains a column name Email.

You need to prevent nonadministrative users from seeing the full email addresses in the Email column. The users must see values in a format of aXXX@XXXX.com instead.

What should you do?

**A.** From Microsoft SQL Server Management Studio, set an email mask on the Email column.

**B.** From the Azure portal, set a mask on the Email column.

**C.** From Microsoft SQL Server Management studio, grant the SELECT permission to the users for all the columns in the dbo.Customers table except Email.

**D.** From the Azure portal, set a sensitivity classification of Confidential for the Email column.

**Answer: D (LEAVE A REPLY)**

Explanation

From Microsoft SQL Server Management Studio, set an email mask on the Email column. This is because

"This feature cannot be set using portal for Azure Synapse (use PowerShell or REST API) or SQL Managed Instance." So use Create table statement with Masking e.g. CREATE TABLE Membership (MemberID int IDENTITY PRIMARY KEY, FirstName varchar(100) MASKED WITH (FUNCTION =

'partial(1,"XXXXXXX",0)') NULL, . .

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

upvoted 24 times

**NEW QUESTION: 82**

You have the following Azure Data Factory pipelines

* ingest Data from System 1
* Ingest Data from System2
* Populate Dimensions
* Populate facts

ingest Data from System1 and Ingest Data from System1 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest Data from System* Populate Facts must execute after the Populate Dimensions pipeline. All the pipelines must execute every eight hours.

What should you do to schedule the pipelines for execution?

**A.** Add an event trigger to all four pipelines.

**B.** Create a parent pipeline that contains the four pipelines and use an event trigger.

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

**D.** Add a schedule trigger to all four pipelines.

**Answer: C (LEAVE A REPLY)**

Explanation

Schedule trigger: A trigger that invokes a pipeline on a wall-clock schedule.

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers

**NEW QUESTION: 83**

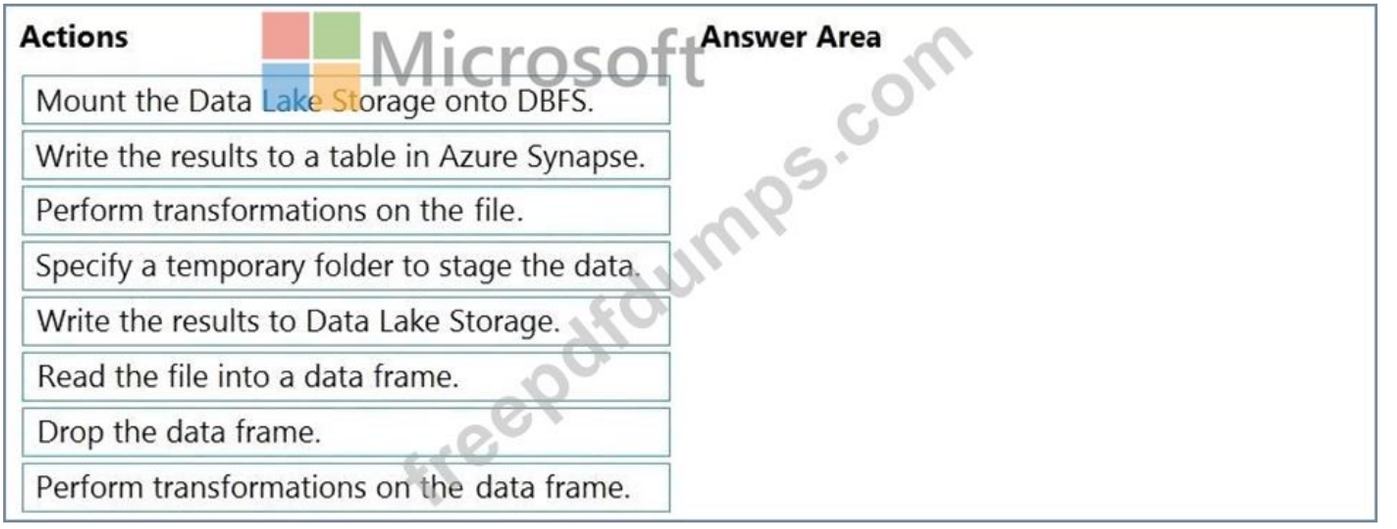
You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers. The file contains two attributes named FirstName and LastName.

You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks.

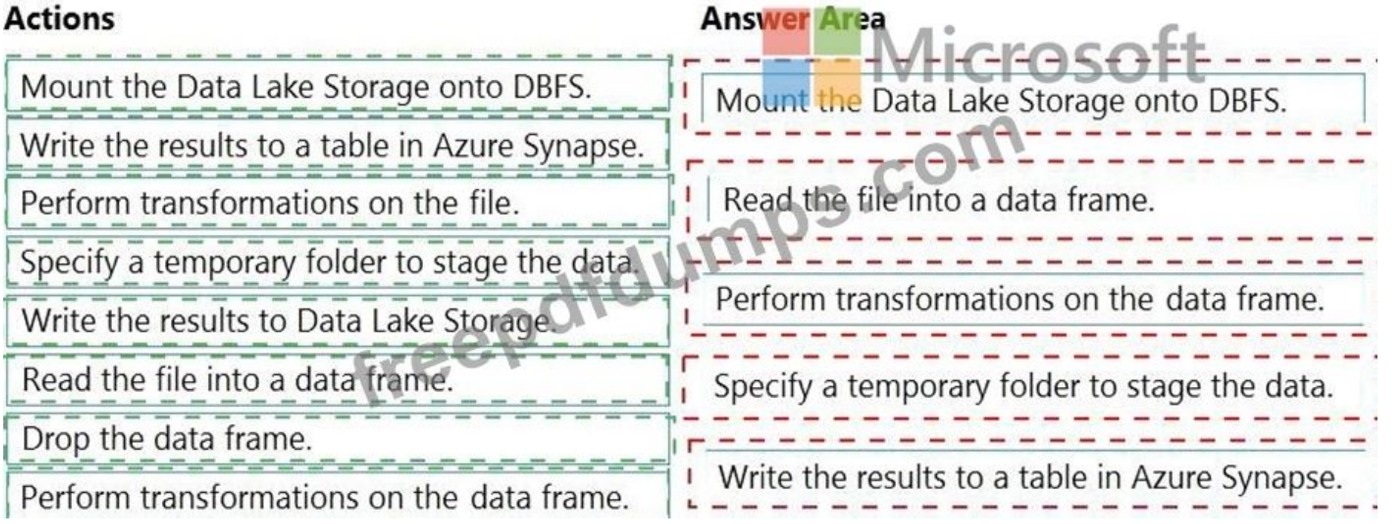
A new column must be created that concatenates the FirstName and LastName values.

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.



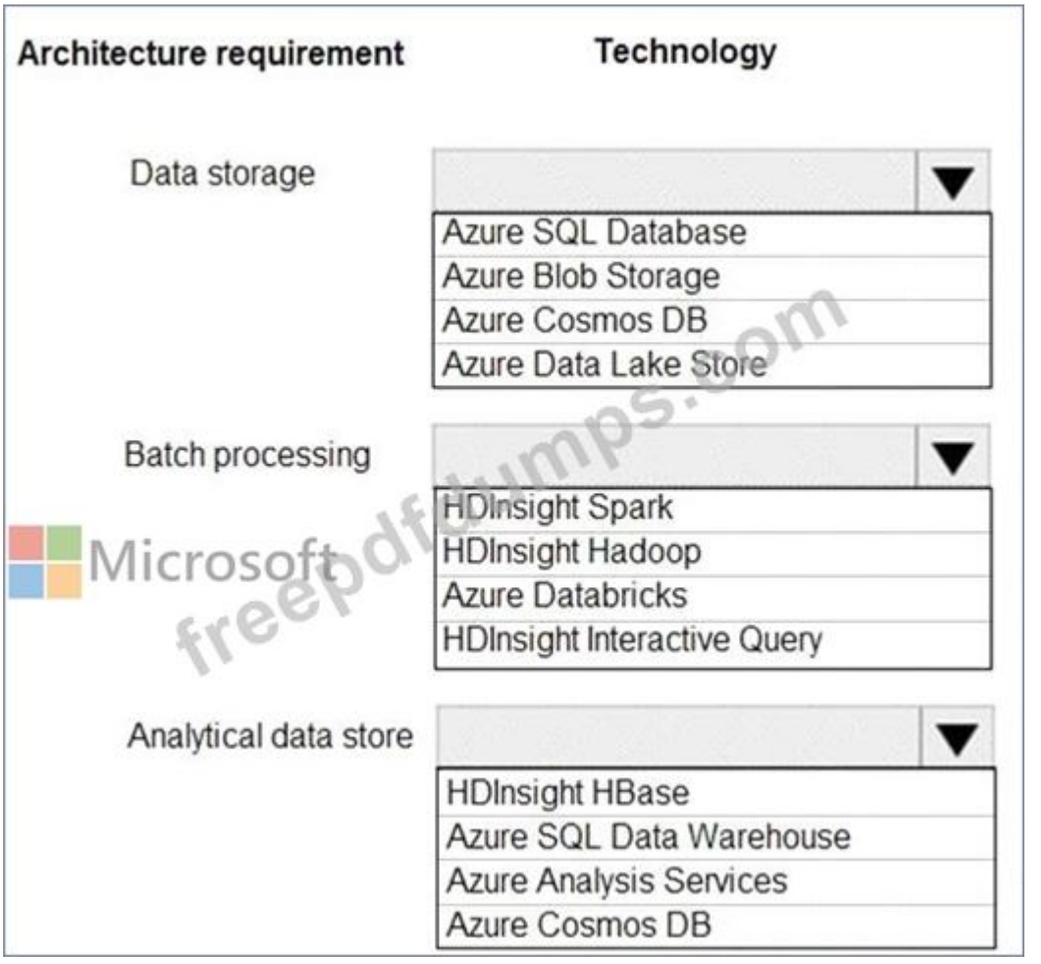

**Answer:**




Explanation
1) mount onto DBFS
2) read into data frame
3) transform data frame
4) specify temporary folder
5) write the results to table in in Azure Synapse
https://docs.databricks.com/data/data-sources/azure/azure-datalake-gen2.html
https://docs.microsoft.com/en-us/azure/databricks/scenarios/databricks-extract-load-sql-data-warehouse

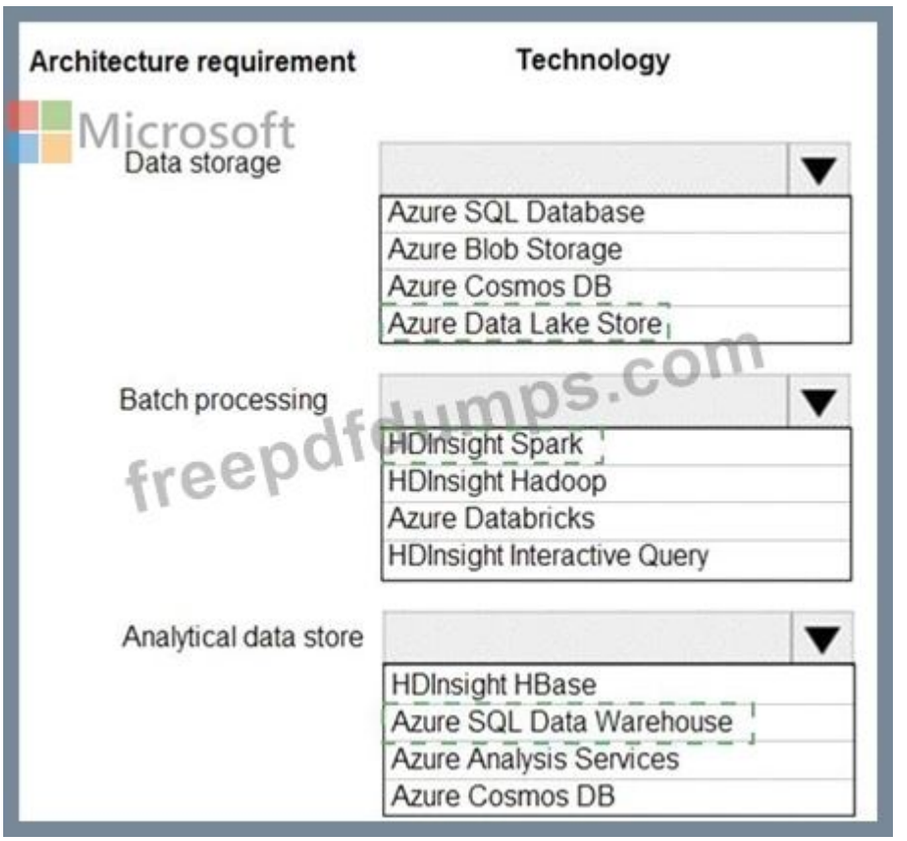
**NEW QUESTION: 84**
You are developing a solution using a Lambda architecture on Microsoft Azure.

The data at test layer must meet the following requirements:
Data storage:
*Serve as a repository (or high volumes of large files in various formats.
*Implement optimized storage for big data analytics workloads.
*Ensure that data can be organized using a hierarchical structure.
Batch processing:
*Use a managed solution for in-memory computation processing.
*Natively support Scala, Python, and R programming languages.
*Provide the ability to resize and terminate the cluster automatically.
Analytical data store:
*Support parallel processing.
*Use columnar storage.
*Support SQL-based languages.
You need to identify the correct technologies to build the Lambda architecture.
Which technologies should you use? To answer, select the appropriate options in the answer area NOTE: Each correct selection is worth one point.



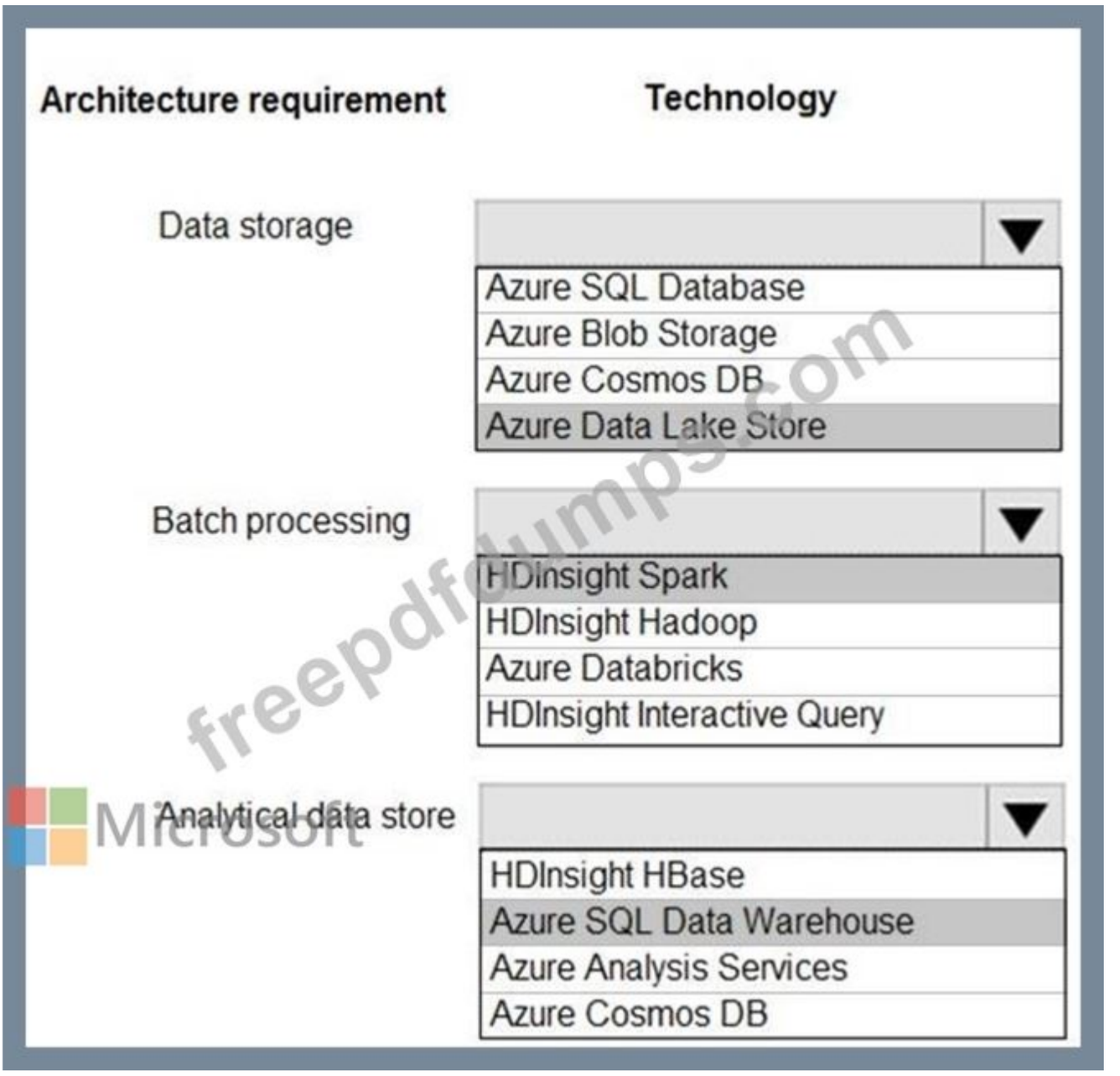


**Answer:**





Explanation






Data storage: Azure Data Lake Store

A key mechanism that allows Azure Data Lake Storage Gen2 to provide file system performance at object storage scale and prices is the addition of a hierarchical namespace. This allows the collection of objects/files within an account to be organized into a hierarchy of directories and nested subdirectories in the same way that the file system on your computer is organized. With the hierarchical namespace enabled, a storage account becomes capable of providing the scalability and cost-effectiveness of object storage, with file system semantics that are familiar to analytics engines and frameworks.

Batch processing: HD Insight Spark

Aparch Spark is an open-source, parallel-processing framework that supports in-memory processing to boost the performance of big-data analysis applications.

HDInsight is a managed Hadoop service. Use it deploy and manage Hadoop clusters in Azure. For batch processing, you can use Spark, Hive, Hive LLAP, MapReduce.

Languages: R, Python, Java, Scala, SQL

Analytic data store: SQL Data Warehouse
SQL Data Warehouse is a cloud-based Enterprise Data Warehouse (EDW) that uses Massively Parallel Processing (MPP).
SQL Data Warehouse stores data into relational tables with columnar storage.
References:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-namespace
https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/batch-processing
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-overview-what-is

**NEW QUESTION: 85**
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 SOL.
* 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 Standard cluster for the data engineers, and a High Concurrency cluster for the jobs.
Does this meet the goal?
**A.** Yes
**B.** No
**Answer: B (LEAVE A REPLY)**
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.

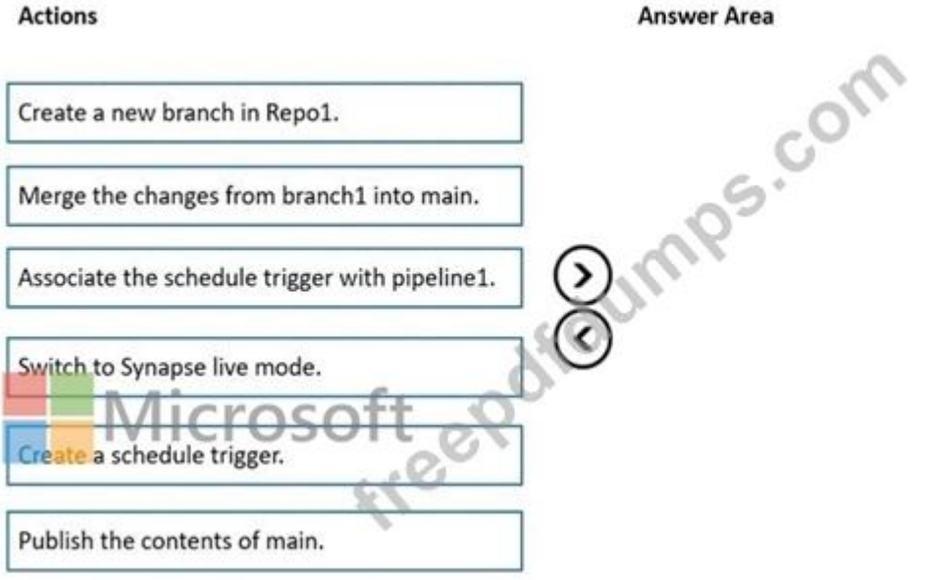
A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.
Reference:
https://docs.azuredatabricks.net/clusters/configure.html

**NEW QUESTION: 86**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values.
75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an Azure SQL data warehouse.
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
**Answer: (SHOW ANSWER)**
Explanation
Instead modify the files to ensure that each row is less than 1 MB.
References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

**NEW QUESTION: 87**
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
**Answer: B (LEAVE A REPLY)**
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.

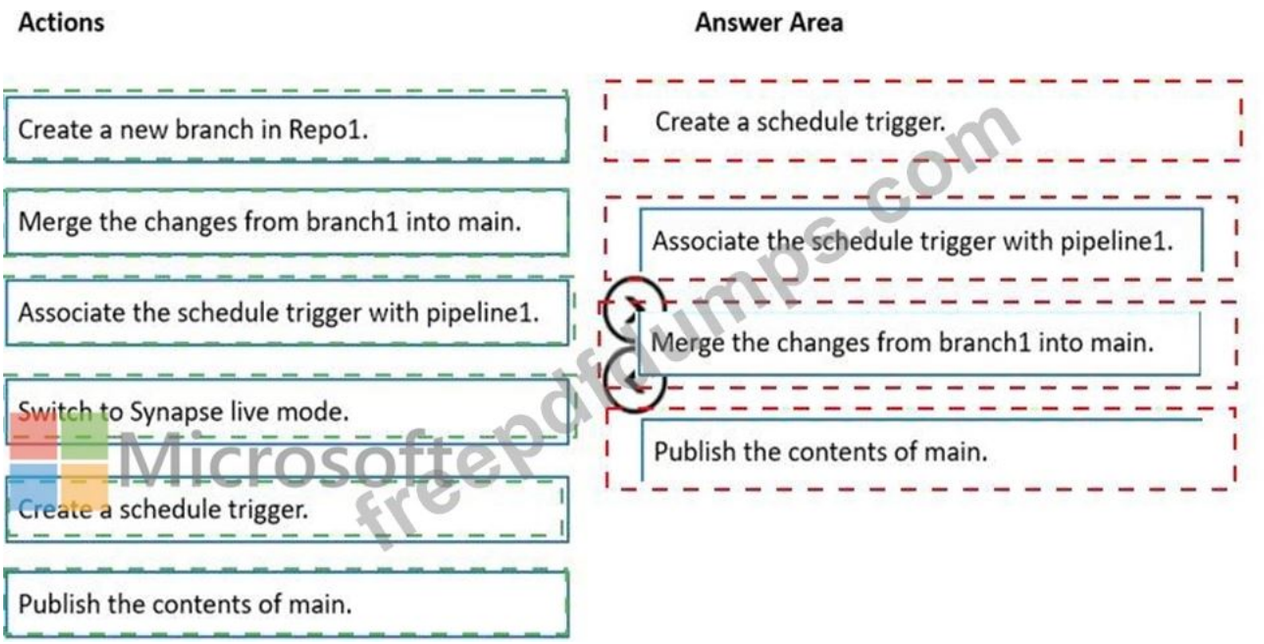
The suggested best practice is to launch a new cluster for each run of critical jobs. This helps avoid any issues (failures, missing SLA, and so on) due to an existing workload (noisy neighbor) on a shared cluster.
Reference:
https://docs.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-scheduled-bat

**NEW QUESTION: 88**
You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1.
Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1. Branch1 contains an Azure Synapse pipeline named pipeline1.
In workspace1, you complete testing of pipeline1.
You need to schedule pipeline1 to run daily at 6 AM.
Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

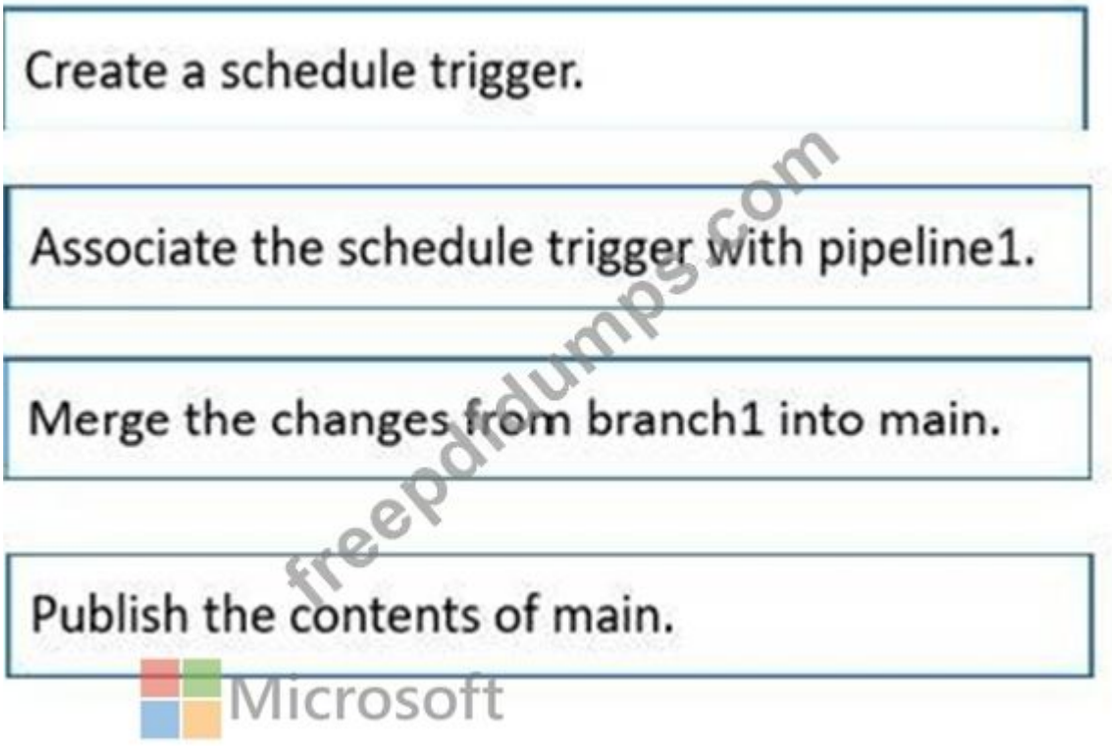





**Answer:**




Explanation




Timeline Description automatically generated

**NEW QUESTION: 89**

You are developing an application that uses Azure Data Lake Storage Gen 2.

You need to recommend a solution to grant permissions to a specific application for a limited time period.

What should you include in the recommendation?

**A.** Azure Active Directory (Azure AD) identities

**B.** shared access signatures (SAS)

**C.** account keys

**D.** role assignments

**Answer: (SHOW ANSWER)**

Explanation

A shared access signature (SAS) provides secure delegated access to resources in your storage account. With a SAS, you have granular control over how a client can access your data. For example:

What resources the client may access.

What permissions they have to those resources.

How long the SAS is valid.

Reference:

https://docs.microsoft.com/en-us/azure/storage/common/storage-sas-overview

**NEW QUESTION: 90**

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.

**Answer: (SHOW ANSWER)**

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

**NEW QUESTION: 91**

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Contacts. Contacts contains a column named Phone.

You need to ensure that users in a specific role only see the last four digits of a phone number when querying the Phone column.

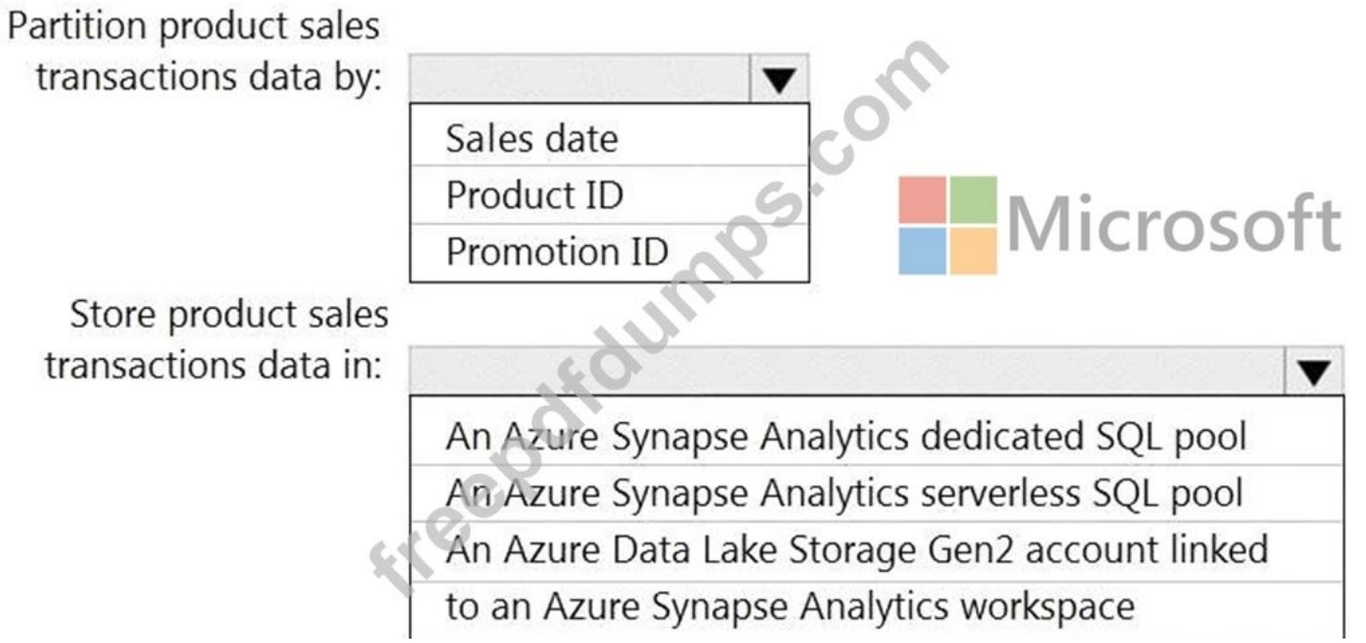
What should you include in the solution?

**A.** a default value

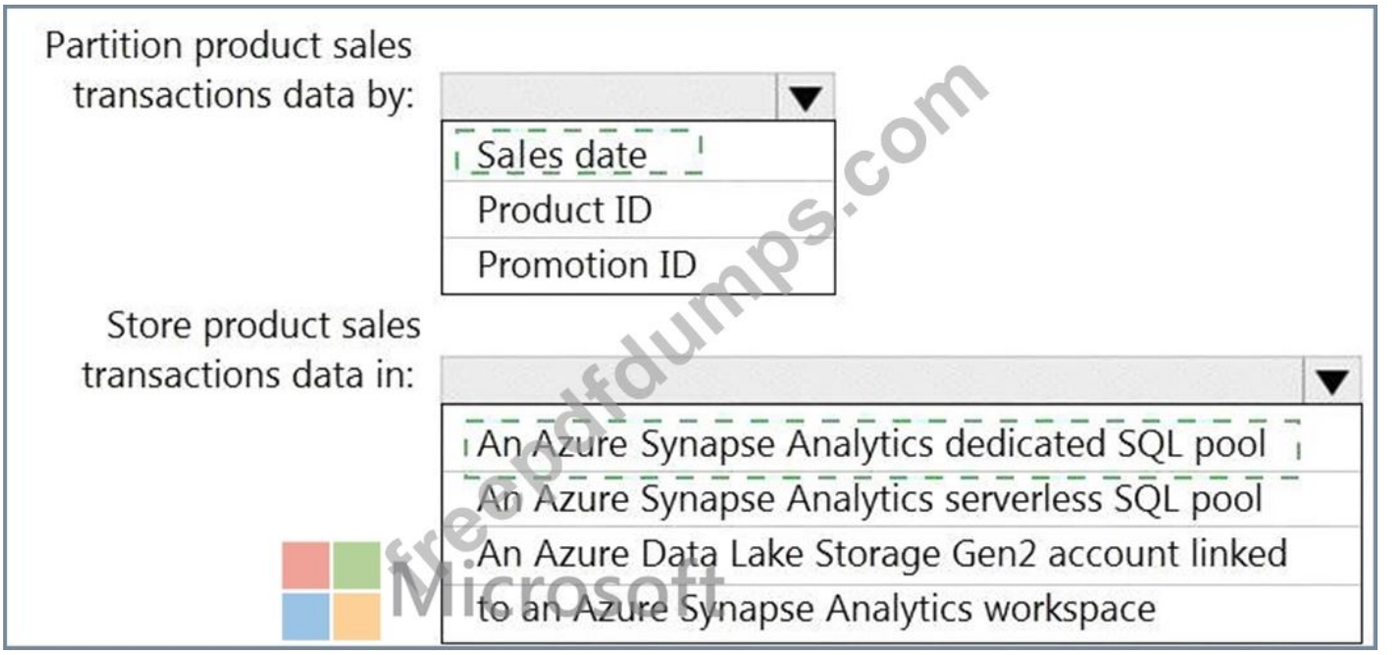
**B.** dynamic data masking
**C.** row-level security (RLS)
**D.** column encryption
**E.** table partitions
**Answer: B (LEAVE A REPLY)**
Explanation
Dynamic data masking helps prevent unauthorized access to sensitive data by enabling customers to designate how much of the sensitive data to reveal with minimal impact on the application layer. It's a policy-based security feature that hides the sensitive data in the result set of a query over designated database fields, while the data in the database is not changed.
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview



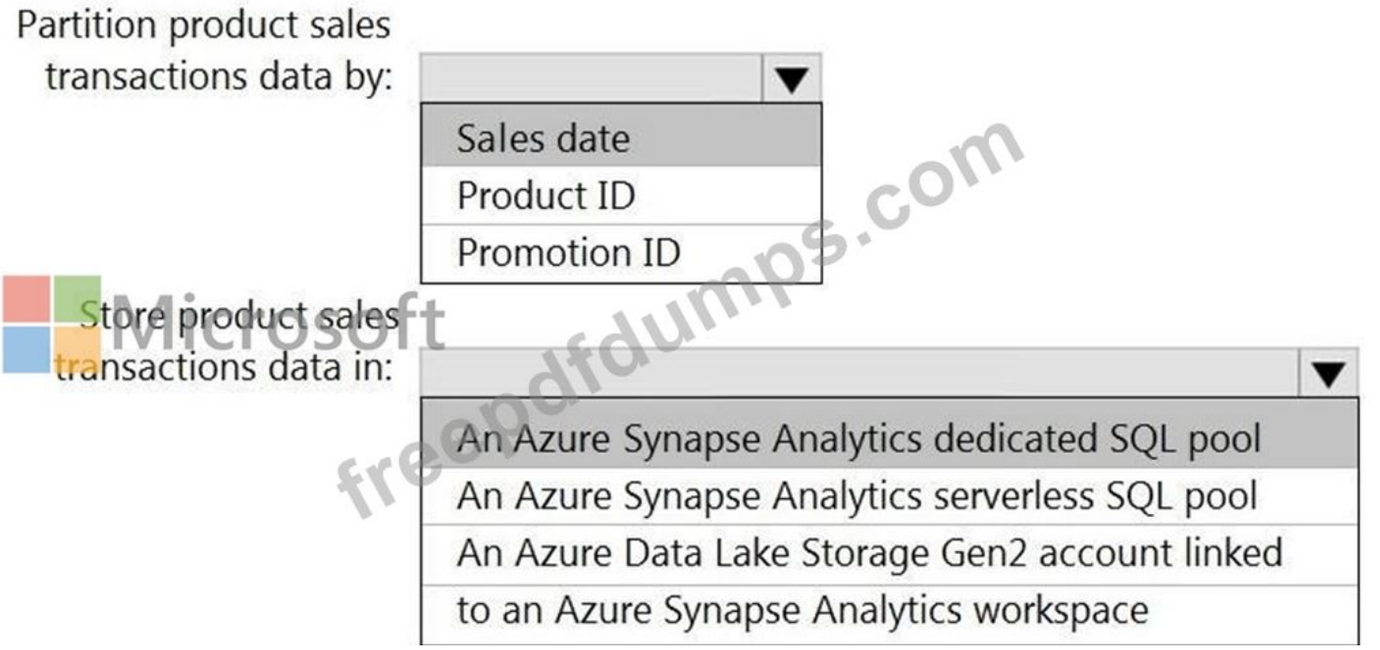
**NEW QUESTION: 92**
You need to design the partitions for the product sales transactions. The solution must meet the sales transaction dataset requirements.
What should you include in the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.




**Answer:**




Explanation





Box 1: Sales date

Scenario: Contoso requirements for data integration include:

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

Box 2: An Azure Synapse Analytics Dedicated SQL pool

Scenario: Contoso requirements for data integration include:

Ensure that data storage costs and performance are predictable.

The size of a dedicated SQL pool (formerly SQL DW) is determined by Data Warehousing Units (DWU).

Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage. This format significantly reduces the data storage costs, and improves query performance.

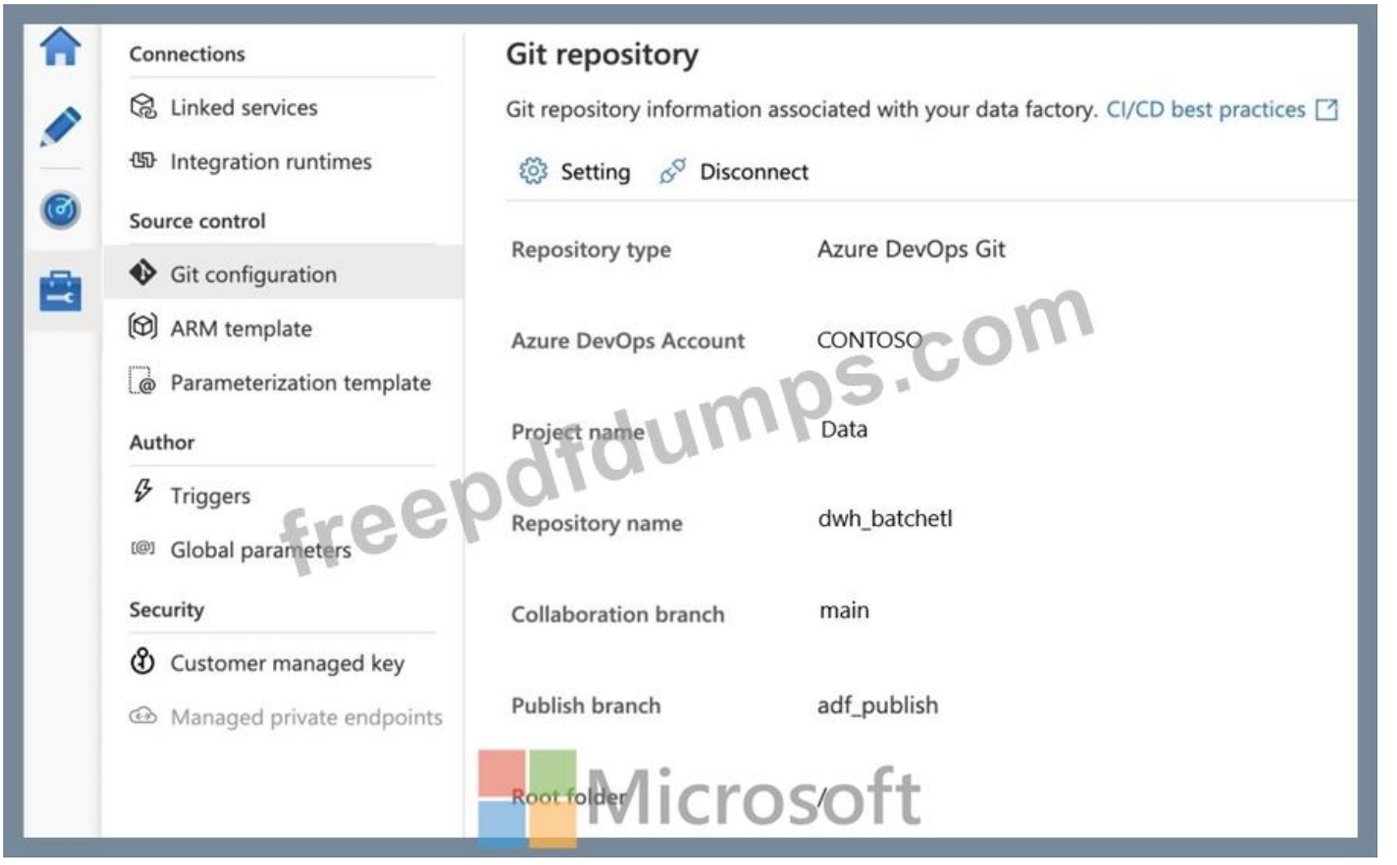
Synapse analytics dedicated sql pool

Reference:

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

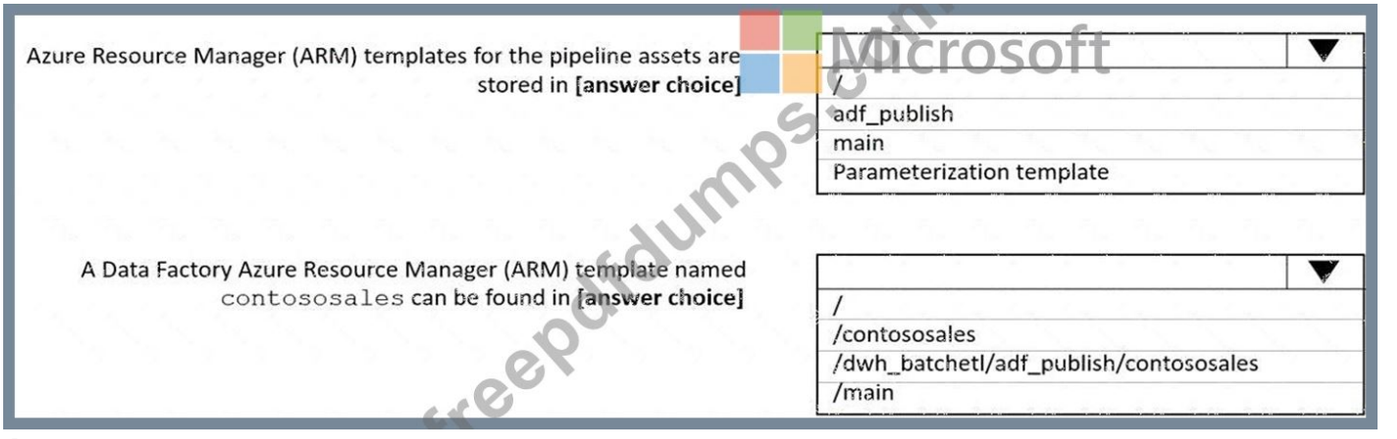
**NEW QUESTION: 93**

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

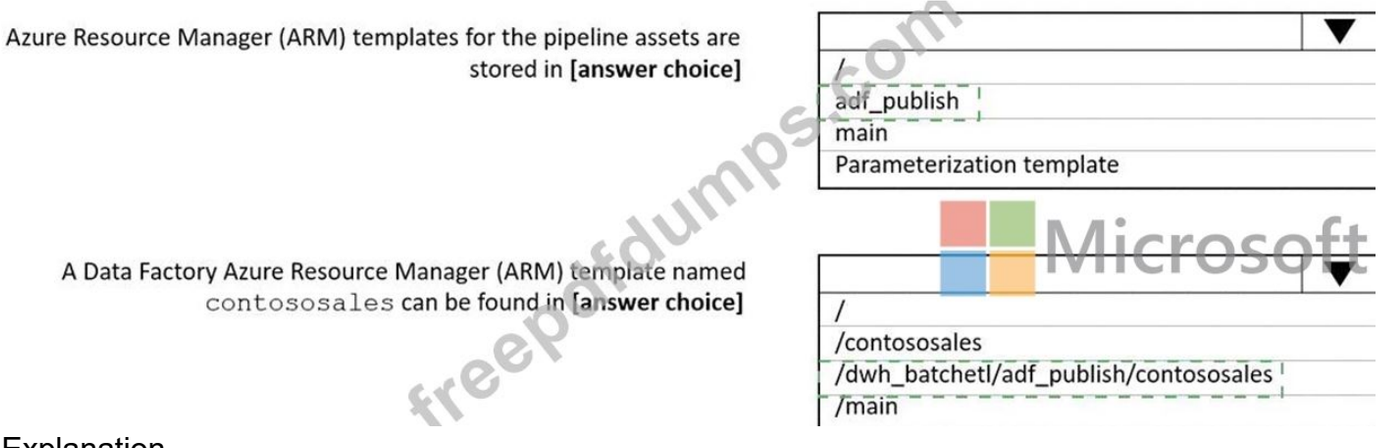



Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.

NOTE: Each correct selection is worth one point.

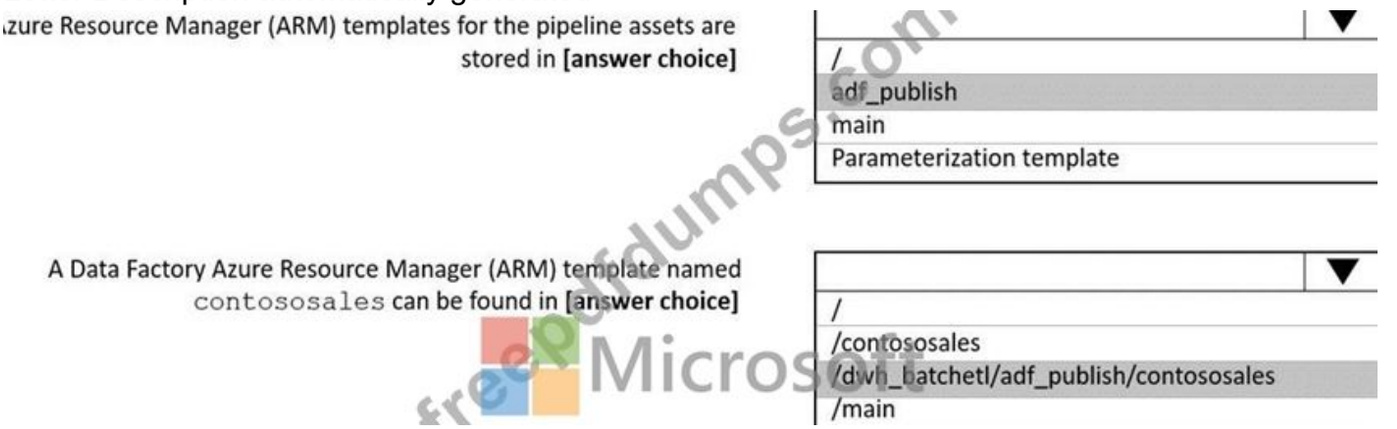



**Answer:**




Explanation

Letter Description automatically generated




Box 1: adf_publish

The Publish branch is the branch in your repository where publishing related ARM templates are stored and updated. By default, it's adf_publish.

Box 2: / dwh_batchetl/adf_publish/contososales

Note: RepositoryName (here dwh_batchetl): Your Azure Repos code repository name. Azure Repos projects contain Git repositories to manage your source code as your project grows. You can create a new repository or use an existing repository that's already in your project.

Reference:

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

**NEW QUESTION: 94**

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

**A.** Azure Data Factory

**B.** Azure Stream Analytics

**C.** Azure Databricks

**D.** Azure Event Hubs

**Answer: (SHOW ANSWER)**

Explanation

https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/stream-processing

**NEW QUESTION: 95**

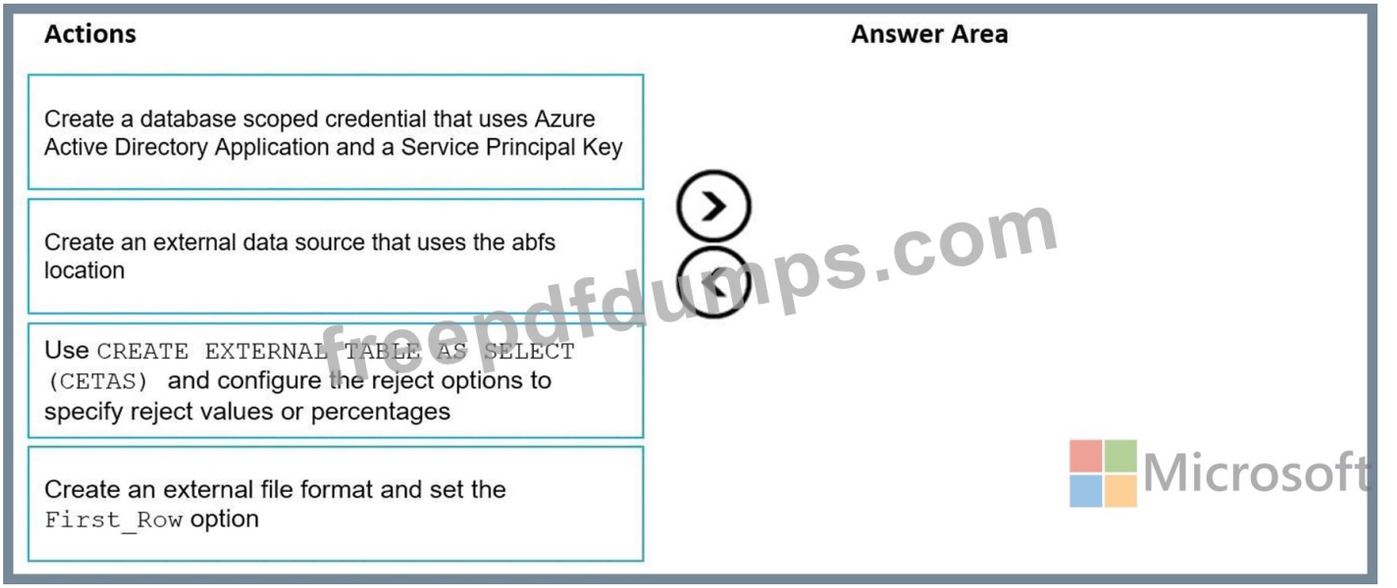
You have data stored in thousands of CSV files in Azure Data Lake Storage Gen2. Each file has a header row followed by a properly formatted carriage return (/r) and line feed (/n).

You are implementing a pattern that batch loads the files daily into an enterprise data warehouse in Azure Synapse Analytics by using PolyBase.

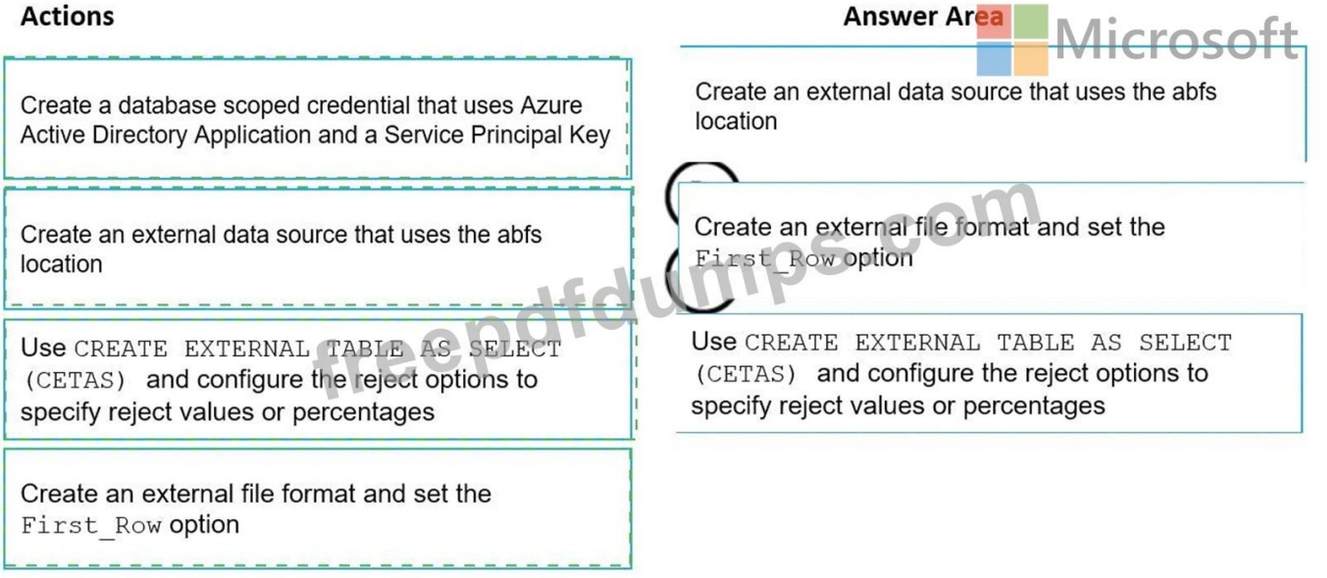
You need to skip the header row when you import the files into the data warehouse. Before building the loading pattern, you need to prepare the required database objects in Azure Synapse Analytics.

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

NOTE: Each correct selection is worth one point

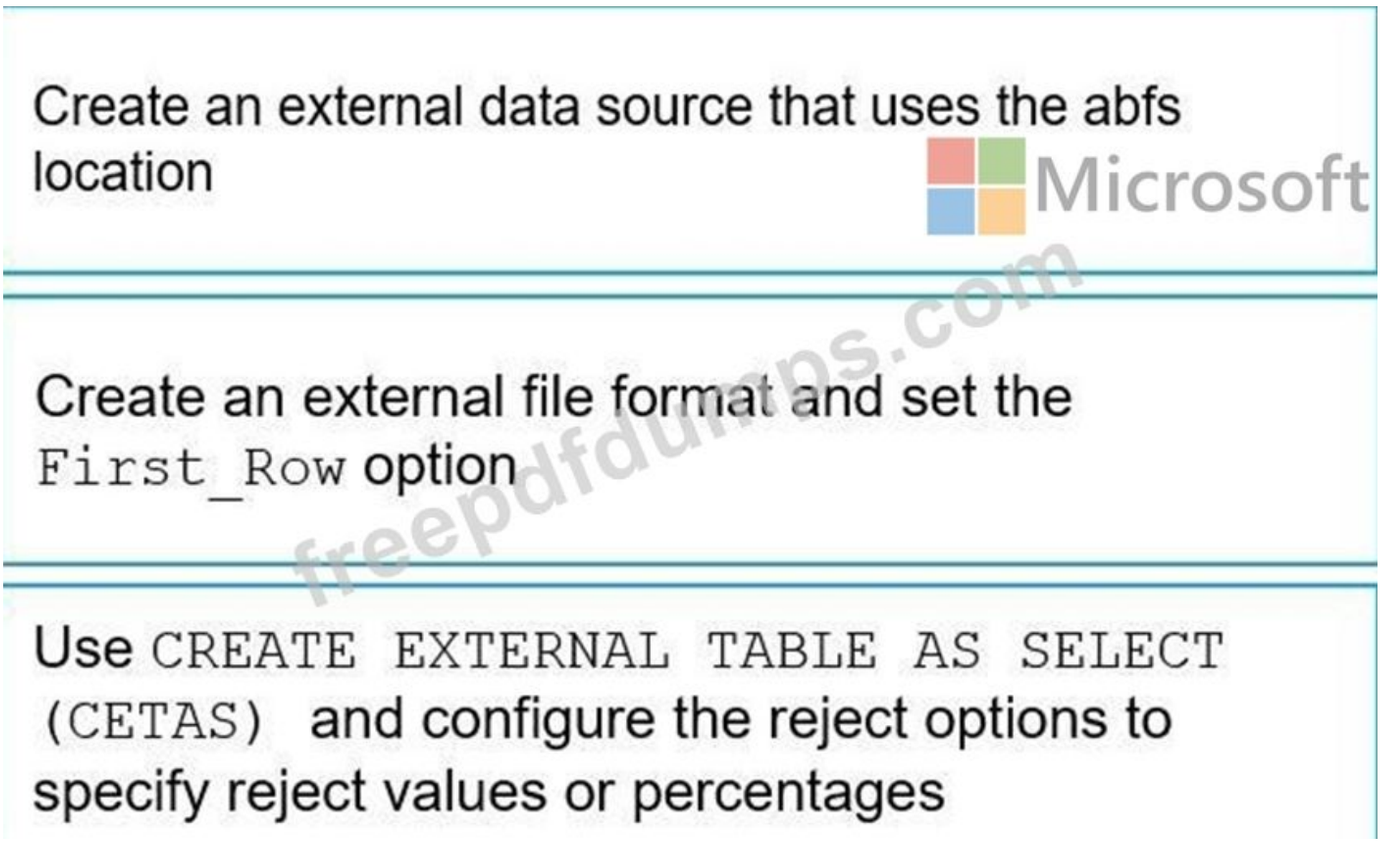



**Answer:**




Explanation

A picture containing timeline Description automatically generated



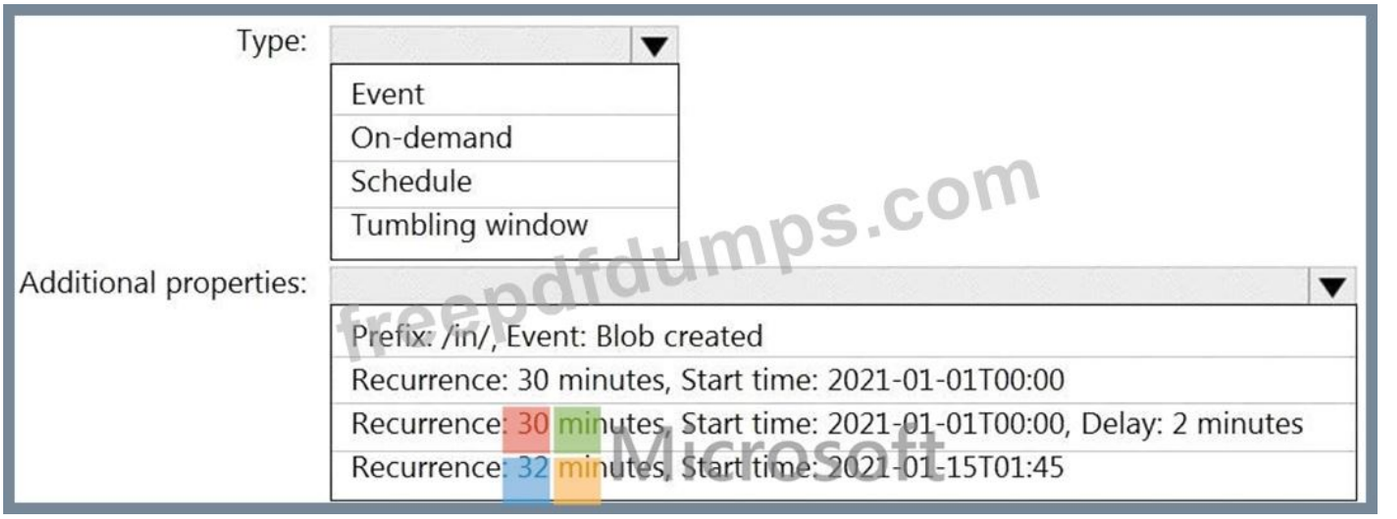


Step 1: Create an external data source that uses the abfs location
Create External Data Source to reference Azure Data Lake Store Gen 1 or 2 Step 2: Create an external file format and set the First_Row option.
Create External File Format.
Step 3: Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages To use PolyBase, you must create external tables to reference your external data.
Use reject options.
Note: REJECT options don't apply at the time this CREATE EXTERNAL TABLE AS SELECT statement is run. Instead, they're specified here so that the database can use them at a later time when it imports data from the external table. Later, when the CREATE TABLE AS SELECT statement selects data from the external table, the database will use the reject options to determine the number or percentage of rows that can fail to import before it stops the import.
Reference:
https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-as-select-transact-sql

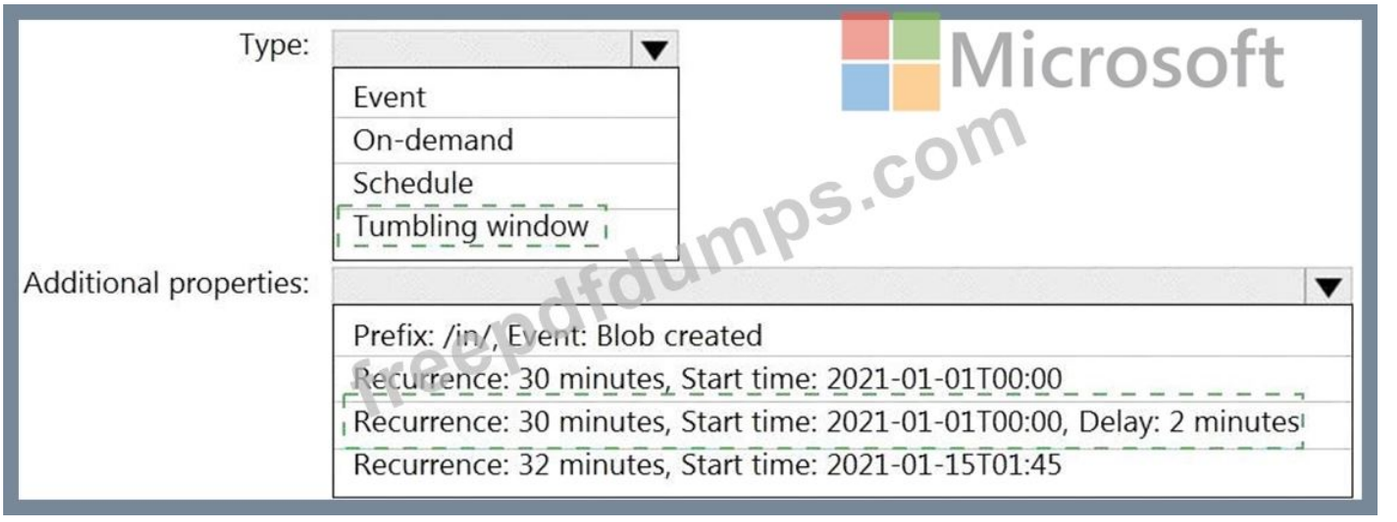
**NEW QUESTION: 96**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.
You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.
You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.
You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.
Solution: You use an Azure Synapse Analytics serverless SQL pool to create an external table that has an additional DateTime column.
Does this meet the goal?
**A.** Yes
**B.** No
**Answer: B (LEAVE A REPLY)**
Explanation
Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column

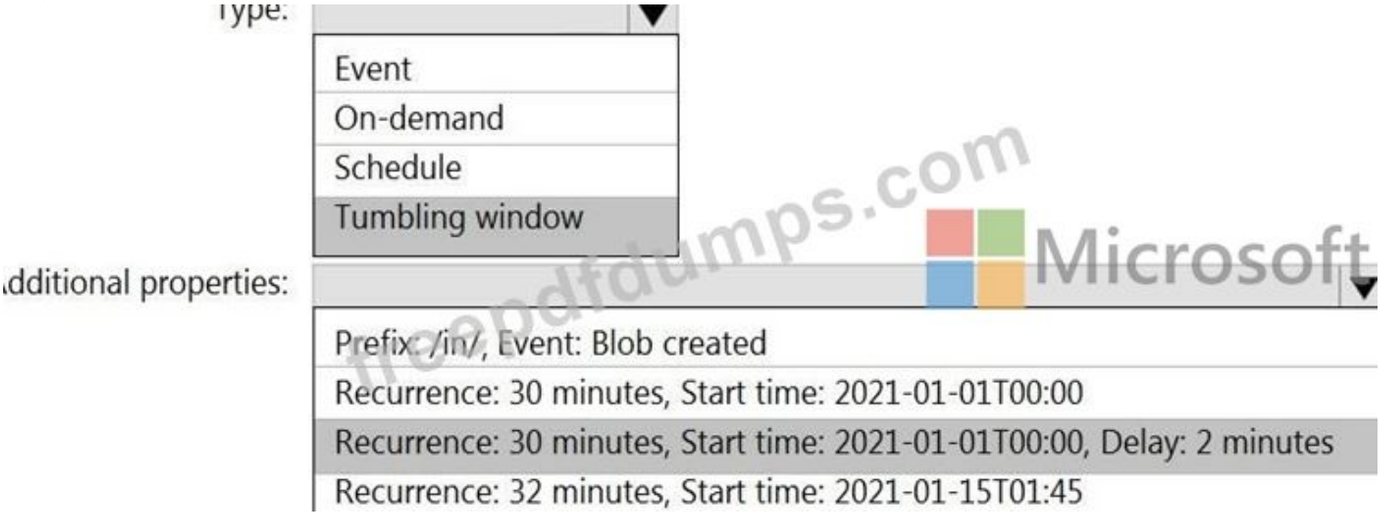
**NEW QUESTION: 97**
You build an Azure Data Factory pipeline to move data from an Azure Data Lake Storage Gen2 container to a database in an Azure Synapse Analytics dedicated SQL pool.
Data in the container is stored in the following folder structure.
/in/{YYYY}/{MM}/{DD}/{HH}/{mm}
The earliest folder 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.




**Answer:**




Explanation




Box 1: Tumbling window

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

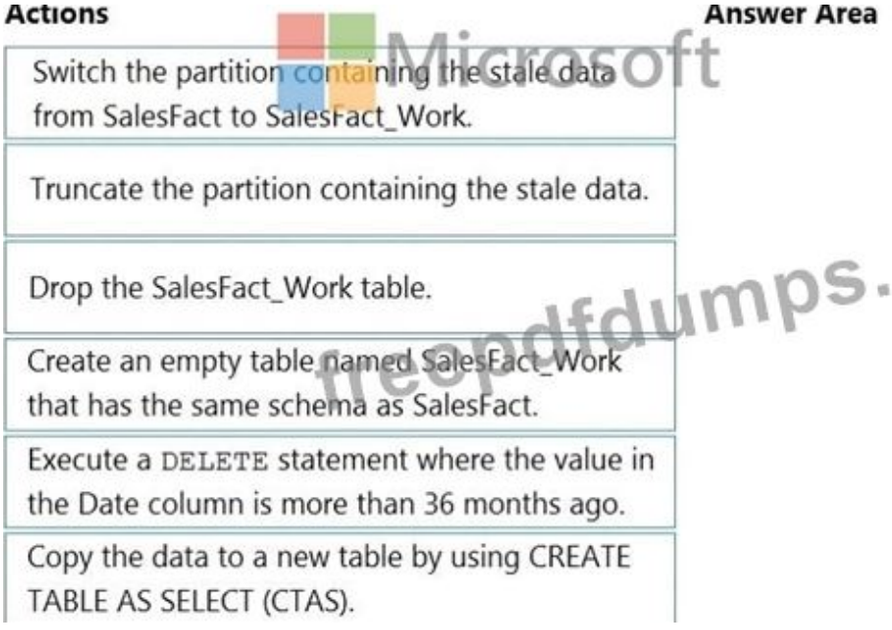
Box 2:

Recurrence: 30 minutes, not 32 minutes

Delay: 2 minutes.

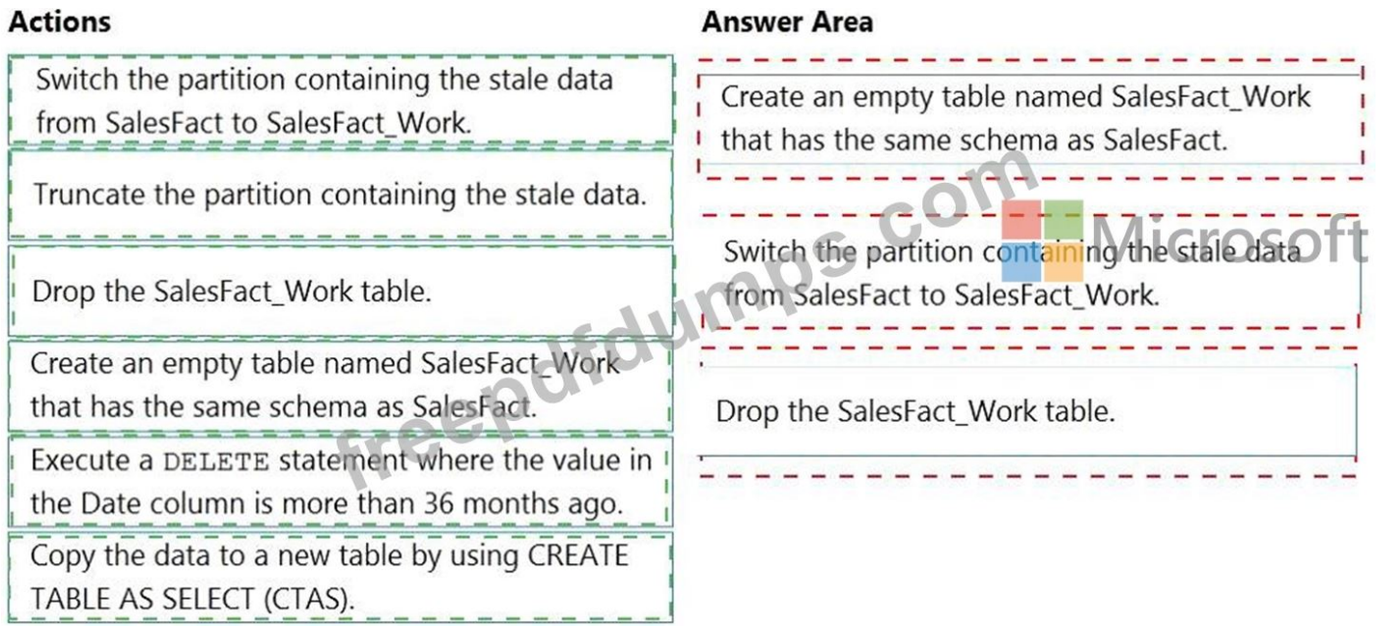
The amount of time to delay the start of data processing for the window. The pipeline run is started after the expected execution time plus the amount of delay. The delay defines how long the trigger waits past the due time before triggering a new run. The delay doesn't alter the window startTime.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger

**NEW QUESTION: 98**
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.

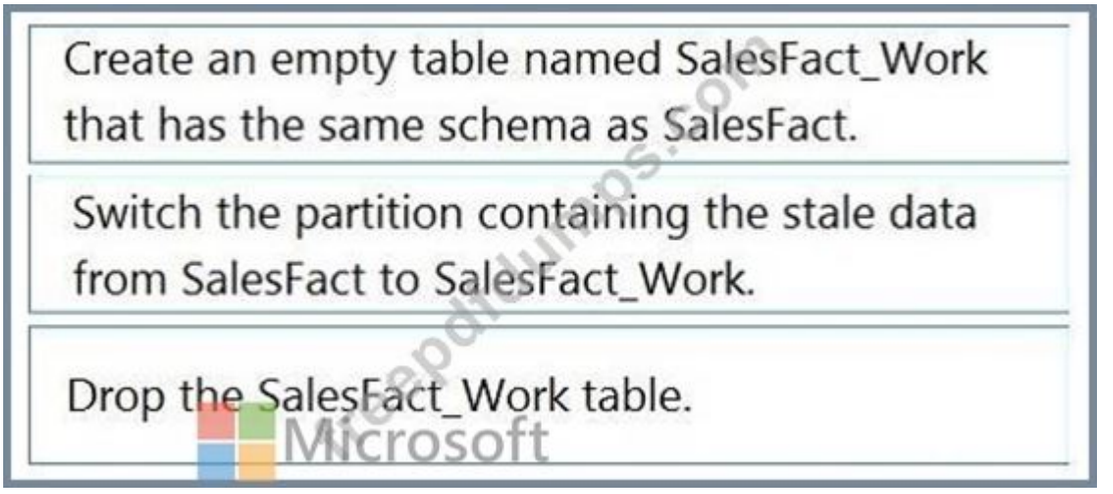




**Answer:**




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

**NEW QUESTION: 99**
You have an Azure Data Lake Storage Gen2 account named account1 that stores logs as shown in the following table.

| Type | Designated retention period |
|---|---|
| Application | 360 days |
| Infrastructure | 60 days |

You do not expect that the logs will be accessed during the retention periods.

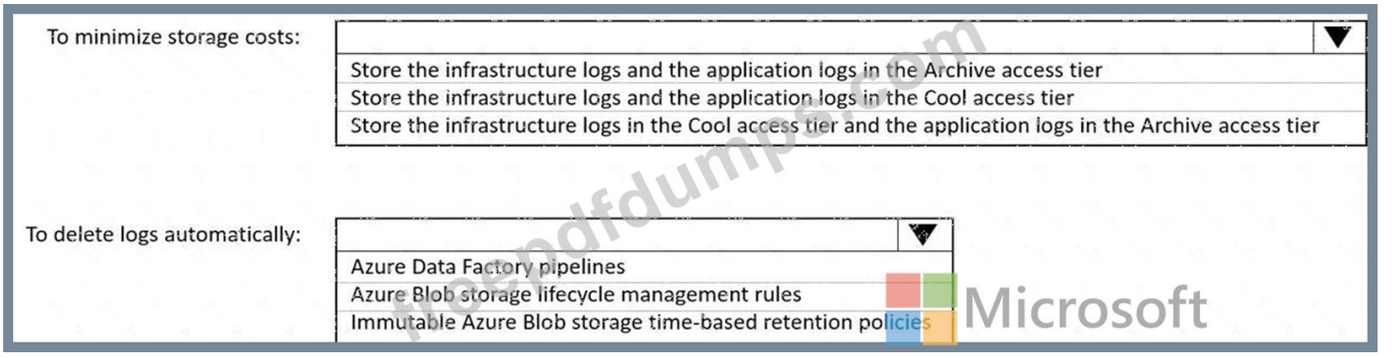
You need to recommend a solution for account1 that meets the following requirements:

Automatically deletes the logs at the end of each retention period

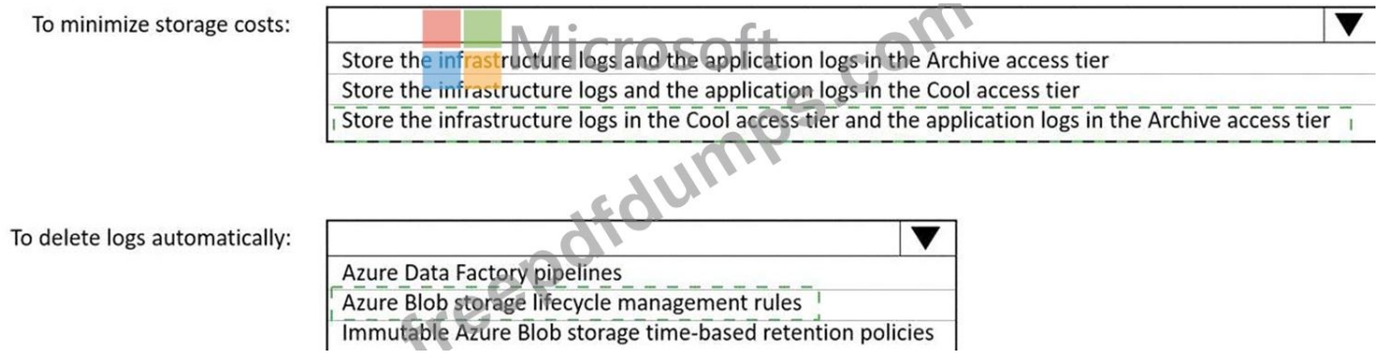
Minimizes storage costs

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

NOTE: Each correct selection is worth one point.

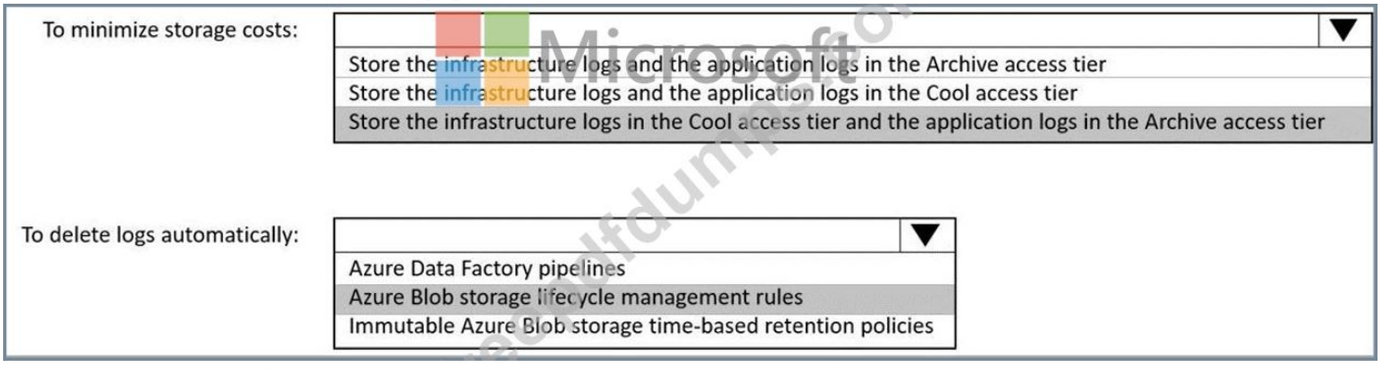



**Answer:**




Explanation

Table Description automatically generated




Box 1: Store the infrastructure logs in the Cool access tier and the application logs in the Archive access tier For infrastructure logs: Cool tier - An online tier optimized for storing data that is infrequently accessed or modified. Data in the cool tier should be stored for a minimum of 30 days. The cool tier has lower storage costs and higher access costs compared to the hot tier.

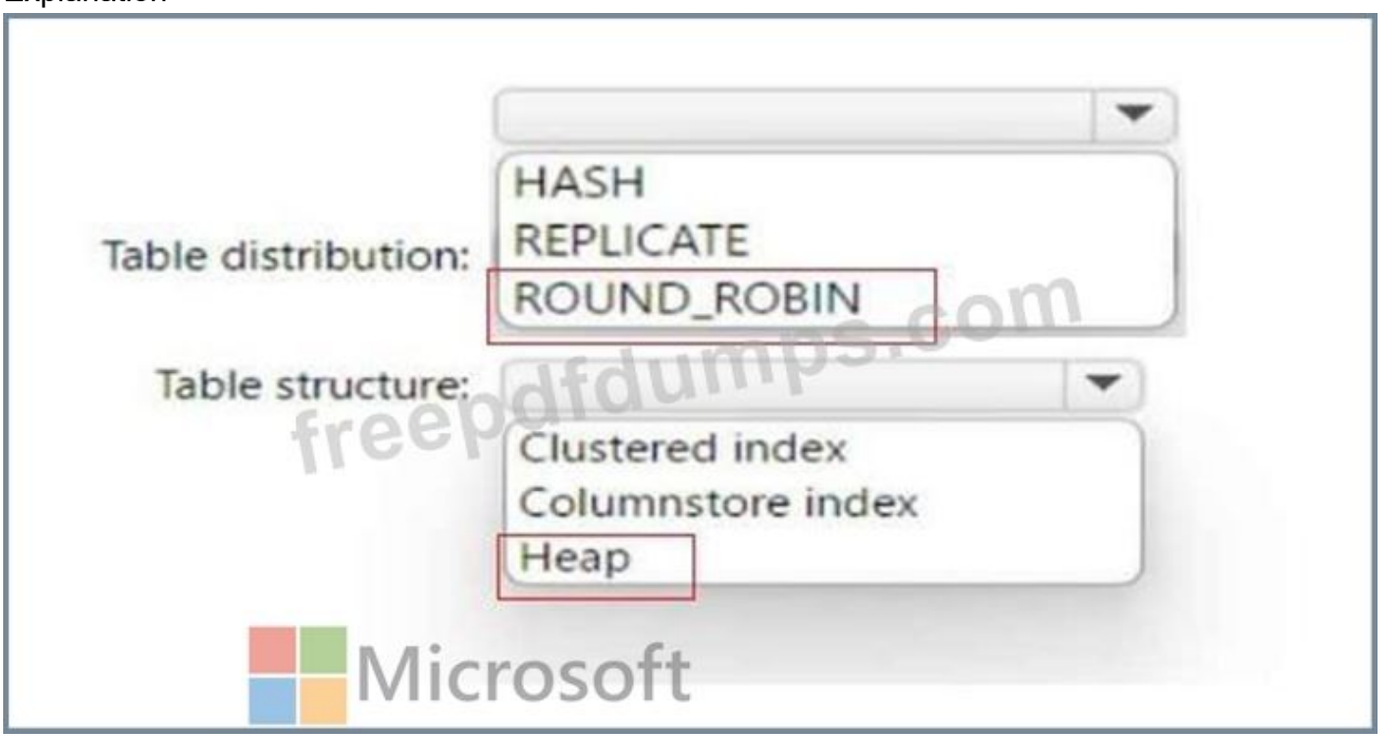
For application logs: Archive tier - An offline tier optimized for storing data that is rarely accessed, and that has flexible latency requirements, on the order of hours. Data in the archive tier should be stored for a minimum of 180 days.
Box 2: Azure Blob storage lifecycle management rules
Blob storage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

**NEW QUESTION: 100**
You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements:
* Can return an employee record from a given point in time.
* Maintains the latest employee information.
* Minimizes query complexity.
How should you model the employee data?
**A.** as a temporal table
**B.** as a SQL graph table
**C.** as a degenerate dimension table
**D.** as a Type 2 slowly changing dimension (SCD) table
**Answer: (SHOW ANSWER)**
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-p

**NEW QUESTION: 101**
You are incrementally loading data into fact tables in an Azure Synapse Analytics dedicated SQL pool.
Each batch of incoming data is staged before being loaded into the fact tables. | You need to ensure that the incoming data is staged as quickly as possible. | How should you configure the staging tables? To answer, select the appropriate options in the answer area.

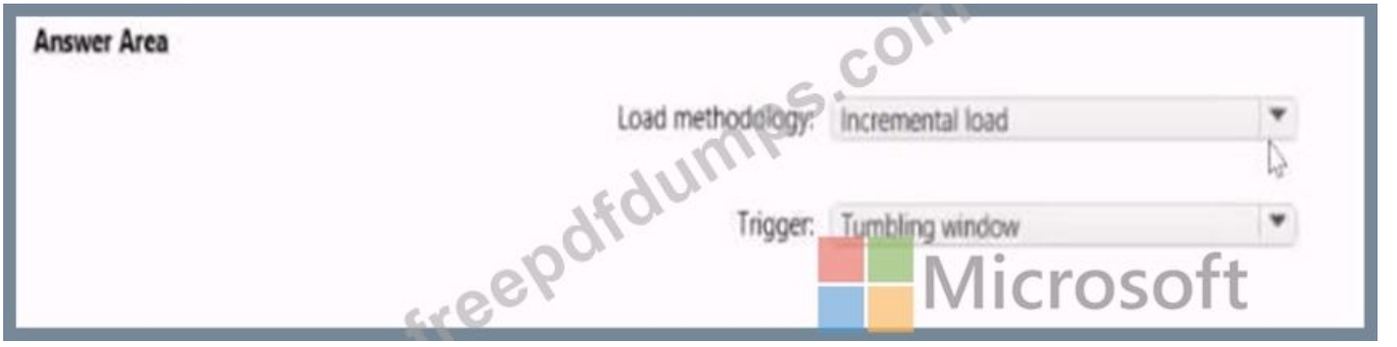

**Answer:**


Explanation



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

**NEW QUESTION: 102**
You have an Azure Storage account that generates 200.000 new files daily. The file names have a format of (YYY)/(MM)/(DD)/|HH])/(CustornerID).csv.
You need to design an Azure Data Factory solution that will toad new data from the storage account to an Azure Data lake once hourly. The solution must minimize load times and costs.
How should you configure the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
**Answer:**
See the answer below in explanation.
Explanation
answer as below




**NEW QUESTION: 103**
You are creating a new notebook in Azure Databricks that will support R as the primary language but will also support Scale and SOL Which switch should you use to switch between languages?
**A.** @<Language>
**B.** %<Language>
**C.** \\(<Language>)
**D.** \\(<Language>)
**Answer: (SHOW ANSWER)**
Explanation
To change the language in Databricks' cells to either Scala, SQL, Python or R, prefix the cell with '%', followed by the language.
%python //or r, scala, sql
Reference:
https://www.theta.co.nz/news-blogs/tech-blog/enhancing-digital-twins-part-3-predictive-maintenance-with-azure

**NEW QUESTION: 104**
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
)
```



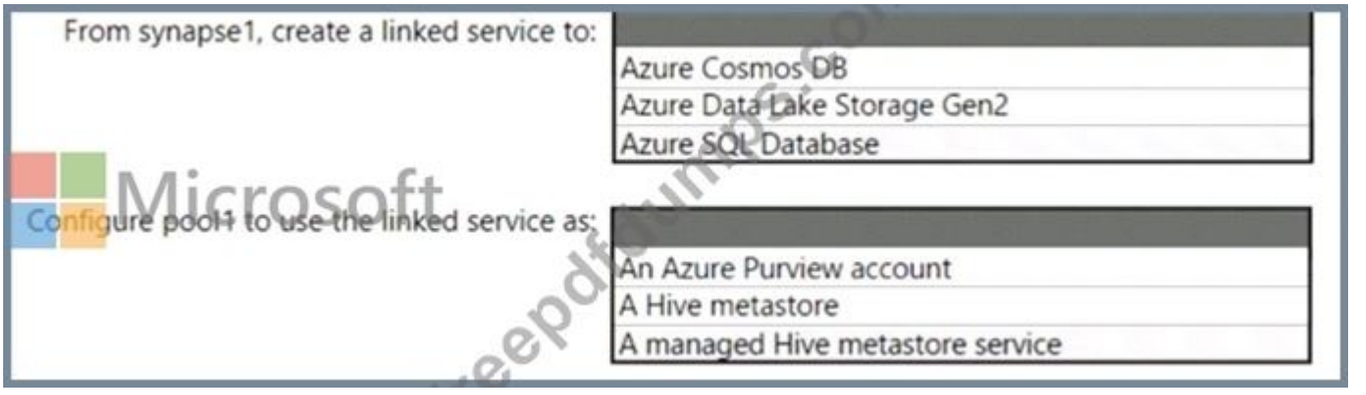
Which two columns should you add to the table? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
**A.** [EffectiveScarcDate] [datetime] NOT NULL,
**B.** [CurrentProduccCacegory] [nvarchar] (100) NOT NULL,
**C.** [EffectiveEndDace] [dacecime] NULL,
**D.** [ProductCategory] [nvarchar] (100) NOT NULL,
**E.** [OriginalProduccCacegory] [nvarchar] (100) NOT NULL,
**Answer: B,E (LEAVE A REPLY)**
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.
Graphical user interface, application, email Description automatically generated

| 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 |

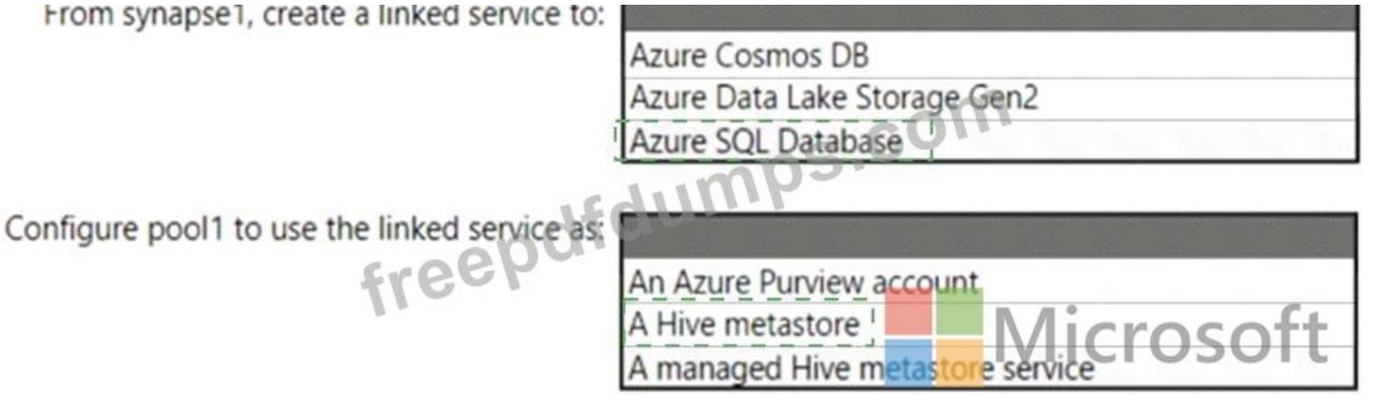
| CustomerID | FirstName | LastName | CurrentEmail | OriginalEmail | CompanyName | InsertedDate | ModifiedDate |
|---|---|---|---|---|---|---|---|
| 2 | Keith | Harris | keith0@aw.com | keith0@aw.com | Progressive Sports | 2021-03-20 | 2021-03-20 |
| 3 | Donna | Carreras | dc3@aw.com | donna0@aw.com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

**NEW QUESTION: 105**
You have an Azure subscription that contains an Azure Databricks workspace named databricks1 and an Azure Synapse Analytics workspace named synapse1. The synapse1 workspace contains an Apache Spark pool named pool1.
You need to share an Apache Hive catalog of pool1 with databricks1.
What should you do? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.




**Answer:**




Explanation
Box 1: Azure SQL Database
Use external Hive Metastore for Synapse Spark Pool
Azure Synapse Analytics allows Apache Spark pools in the same workspace to share a managed HMS (Hive Metastore) compatible metastore as their catalog.
Set up linked service to Hive Metastore

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

**NEW QUESTION: 106**
You are designing 2 solution that will use tables in Delta Lake on Azure Databricks.
You need to minimize how long it takes to perform the following:
*Queries against non-partitioned tables
* Joins on non-partitioned columns
Which two options should you include in the solution? Each correct answer presents part of the solution.
(Choose Correct Answer and Give Explanation and References to Support the answers based from Data Engineering on Microsoft Azure)
**A.** Z-Ordering
**B.** Apache Spark caching
**C.** dynamic file pruning (DFP)
**D.** the clone command
**Answer: (SHOW ANSWER)**
Explanation
According to the information I found on the web, two options that you should include in the solution to minimize how long it takes to perform queries and joins on non-partitioned tables are:
Z-Ordering: This is a technique to colocate related information in the same set of files. This co-locality is automatically used by Delta Lake in data-skipping algorithms. This behavior dramatically reduces the amount of data that Delta Lake on Azure Databricks needs to read123.
Apache Spark caching: This is a feature that allows you to cache data in memory or on disk for faster access. Caching can improve the performance of repeated queries and joins on the same data. You can cache Delta tables using the CACHE TABLE or CACHE LAZY commands.

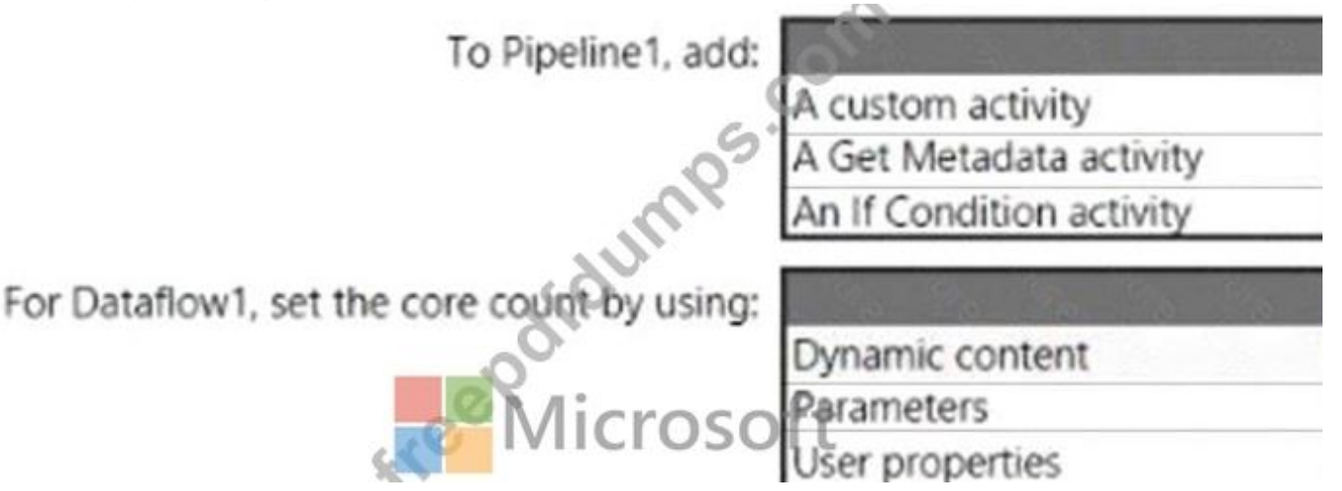
To minimize the time it takes to perform queries against non-partitioned tables and joins on non-partitioned columns in Delta Lake on Azure Databricks, the following options should be included in the solution:
A: Z-Ordering: Z-Ordering improves query performance by co-locating data that share the same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



**NEW QUESTION: 107**
You are monitoring an Azure Stream Analytics job by using metrics in Azure.
You discover that during the last 12 hours, the average watermark delay is consistently greater than the configured late arrival tolerance.
What is a possible cause of this behavior?
**A.** Events whose application timestamp is earlier than their arrival time by more than five minutes arrive as inputs.
**B.** There are errors in the input data.
**C.** The late arrival policy causes events to be dropped.
**D.** The job lacks the resources to process the volume of incoming data.
**Answer: (SHOW ANSWER)**
Explanation
Watermark Delay indicates the delay of the streaming data processing job.

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

**NEW QUESTION: 108**
You have an Azure Synapse Analytics pipeline named Pipeline1 that contains a data flow activity named Dataflow1.
Pipeline1 retrieves files from an Azure Data Lake Storage Gen 2 account named storage1.
Dataflow1 uses the AutoResolveIntegrationRuntime integration runtime configured with a core count of 128.
You need to optimize the number of cores used by Dataflow1 to accommodate the size of the files in storage1.
What should you configure? To answer, select the appropriate options in the answer area.




**Answer:**

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

**NEW QUESTION: 109**
What should you recommend using to secure sensitive customer contact information?
**A.** data labels
**B.** column-level security
**C.** row-level security
**D.** Transparent Data Encryption (TDE)
**Answer: B (LEAVE A REPLY)**
Explanation
Scenario: All cloud data must be encrypted at rest and in transit.
Always Encrypted is a feature designed to protect sensitive data stored in specific database columns from access (for example, credit card numbers, national identification numbers, or data on a need to know basis).
This includes database administrators or other privileged users who are authorized to access the database to perform management tasks, but have no business need to access the particular data in the encrypted columns.
The data is always encrypted, which means the encrypted data is decrypted only for processing by client applications with access to the encryption key.
References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-security-overview

**NEW QUESTION: 110**
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
**Answer: (SHOW ANSWER)**
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

**NEW QUESTION: 111**
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.
**Answer: C (LEAVE A REPLY)**
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-man

**NEW QUESTION: 112**
You are designing a folder structure for the files m an Azure Data Lake Storage Gen2 account. The account has one container that contains three years of data.
You need to recommend a folder structure that meets the following requirements:
* Supports partition elimination for queries by Azure Synapse Analytics serverless SQL pooh
* Supports fast data retrieval for data from the current month

* Simplifies data security management by department
Which folder structure should you recommend?
**A.** \YYY\MM\DD\Department\DataSource\DataFile_YYYMMMDD.parquet
**B.** \Depdftment\DataSource\YYY\MM\DataFile_YYYYMMDD.parquet
**C.** \DD\MM\YYYY\Department\DataSource\DataFile_DDMMYY.parquet
**D.** \DataSource\Department\YYYYMM\DataFile_YYYYMMDD.parquet
**Answer: (SHOW ANSWER)**
Explanation
Department top level in the hierarchy to simplify security management.
Month (MM) at the leaf/bottom level to support fast data retrieval for data from the current month.

**NEW QUESTION: 113**
You have two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.

| Table | Column |
|---|---|
| Flight | ArrivalAirportID<br>ArrivalDateTime |
| Weather | AirportID<br>ReportDateTime |

You need to recommend a solution that maximizes query performance.
What should you include in the recommendation?
**A.** In the tables use a hash distribution of ArrivalDateTime and ReportDateTime.
**B.** In the tables use a hash distribution of ArrivalAirportID and AirportID.
**C.** In each table, create an identity column.
**D.** In each table, create a column as a composite of the other two columns in the table.
**Answer: (SHOW ANSWER)**
Explanation
Hash-distribution improves query performance on large fact tables.

**NEW QUESTION: 114**
You are designing an Azure Data Lake Storage solution that will transform raw JSON files for use in an analytical workload.
You need to recommend a format for the transformed files. The solution must meet the following requirements:
Contain information about the data types of each column in the files.
Support querying a subset of columns in the files.
Support read-heavy analytical workloads.
Minimize the file size.
What should you recommend?
**A.** JSON

**B.** CSV
**C.** Apache Avro
**D.** Apache Parquet
**Answer: (SHOW ANSWER)**
Explanation
Parquet, an open-source file format for Hadoop, stores nested data structures in a flat columnar format.
Compared to a traditional approach where data is stored in a row-oriented approach, Parquet file format is more efficient in terms of storage and performance.
It is especially good for queries that read particular columns from a "wide" (with many columns) table since only needed columns are read, and IO is minimized.
Reference: https://www.clairvoyant.ai/blog/big-data-file-formats

**NEW QUESTION: 115**
You have an Azure Data Lake Storage account that contains a staging zone.
You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data Flow, and then inserts the data info the data warehouse.
Does this meet the goal?
**A.** Yes
**B.** No
**Answer: (SHOW ANSWER)**
Explanation
If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity, not a mapping flow,5 with your own data processing logic and use the activity in the pipeline. You can create a custom activity to run R scripts on your HDInsight cluster with R installed.
Reference:
https://docs.microsoft.com/en-US/azure/data-factory/transform-data

**NEW QUESTION: 116**
You have the following Azure Stream Analytics query.

```
WITH

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

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

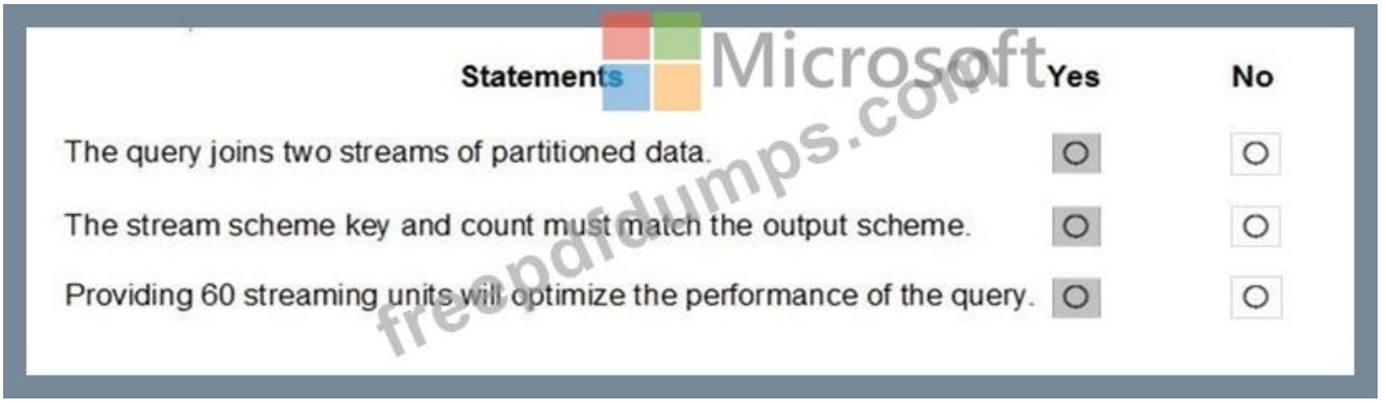
For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.

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

**Answer:**

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

Explanation





Box 1: Yes

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

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

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

step2 AS (SELECT * FROM [input2] PARTITION BY DeviceID INTO 10)

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

Box 2: Yes

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

Box 3: Yes

10 partitions x six SUs = 60 SUs is fine.

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

Reference:

https://azure.microsoft.com/en-in/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/

**NEW QUESTION: 117**

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

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

What should you recommend?

**A.** a managed identity

**B.** anonymous public read access
**C.** a shared key
**Answer: A (LEAVE A REPLY)**
Explanation
Managed Identity authentication is required when your storage account is attached to a VNet.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-exa

**NEW QUESTION: 118**
You have an Azure Data Lake Storage Gen2 account named adls2 that is protected by a virtual network.
You are designing a SQL pool in Azure Synapse that will use adls2 as a source.
What should you use to authenticate to adls2?
**A.** a shared access signature (SAS)
**B.** a managed identity
**C.** a shared key
**D.** an Azure Active Directory (Azure AD) user
**Answer: B (LEAVE A REPLY)**
Explanation
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. You can use the Managed Identity capability to authenticate to any service that support Azure AD authentication.
Managed Identity authentication is required when your storage account is attached to a VNet.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-exa

**NEW QUESTION: 119**
You have an Azure Stream Analytics query. The query returns a result set that contains 10,000 distinct values for a column named clusterID.
You monitor the Stream Analytics job and discover high latency.
You need to reduce the latency.
Which two actions should you perform? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.
**A.** Add a pass-through query.
**B.** Add a temporal analytic function.
**C.** Scale out the query by using PARTITION BY.
**D.** Convert the query to a reference query.
**E.** Increase the number of streaming units.
**Answer: (SHOW ANSWER)**

Explanation
C: Scaling a Stream Analytics job takes advantage of partitions in the input or output. Partitioning lets you divide data into subsets based on a partition key. A process that consumes the data (such as a Streaming Analytics job) can consume and write different partitions in parallel, which increases throughput.
E: 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. This capacity lets you focus on the query logic and abstracts the need to manage the hardware to run your Stream Analytics job in a timely manner.
References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption

**NEW QUESTION: 120**
A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU).
You need to optimize performance for the Azure Stream Analytics job.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
**A.** Implement event ordering.
**B.** Implement Azure Stream Analytics user-defined functions (UDF).
**C.** Implement query parallelization by partitioning the data output.
**D.** Scale the SU count for the job up.
**E.** Scale the SU count for the job down.
**F.** Implement query parallelization by partitioning the data input.
**Answer: D,F (LEAVE A REPLY)**
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization

**NEW QUESTION: 121**
You manage an enterprise data warehouse in Azure Synapse Analytics.
Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries.
You need to monitor resource utilization to determine the source of the performance issues.
Which metric should you monitor?
**A.** Data IO percentage
**B.** Local tempdb percentage
**C.** Cache used percentage
**D.** DWU percentage

**Answer: (SHOW ANSWER)**

Explanation

Monitor and troubleshoot slow query performance by determining whether your workload is optimally leveraging the adaptive cache for dedicated SQL pools.

Reference:

https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monito

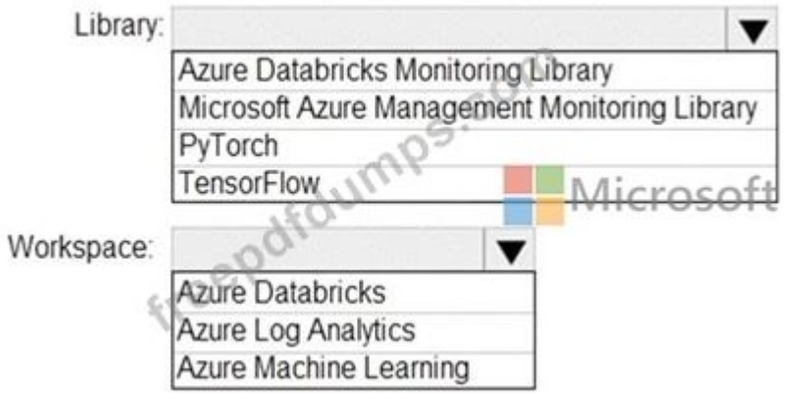


**NEW QUESTION: 122**

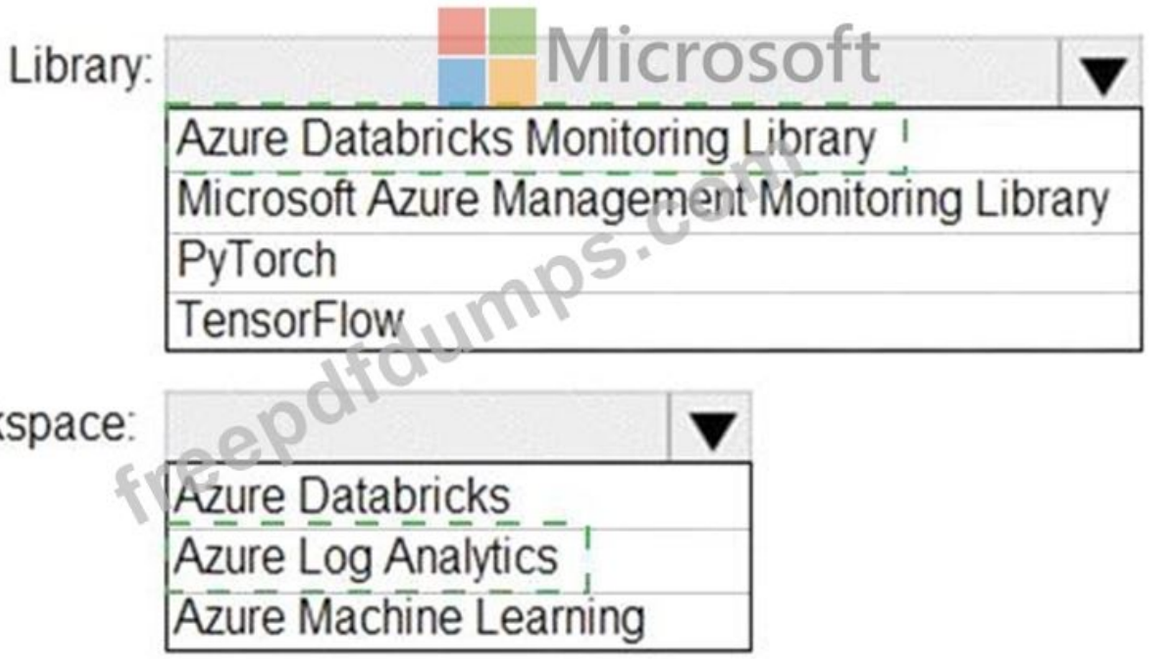
You need to collect application metrics, streaming query events, and application log messages for an Azure Databrick cluster.

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

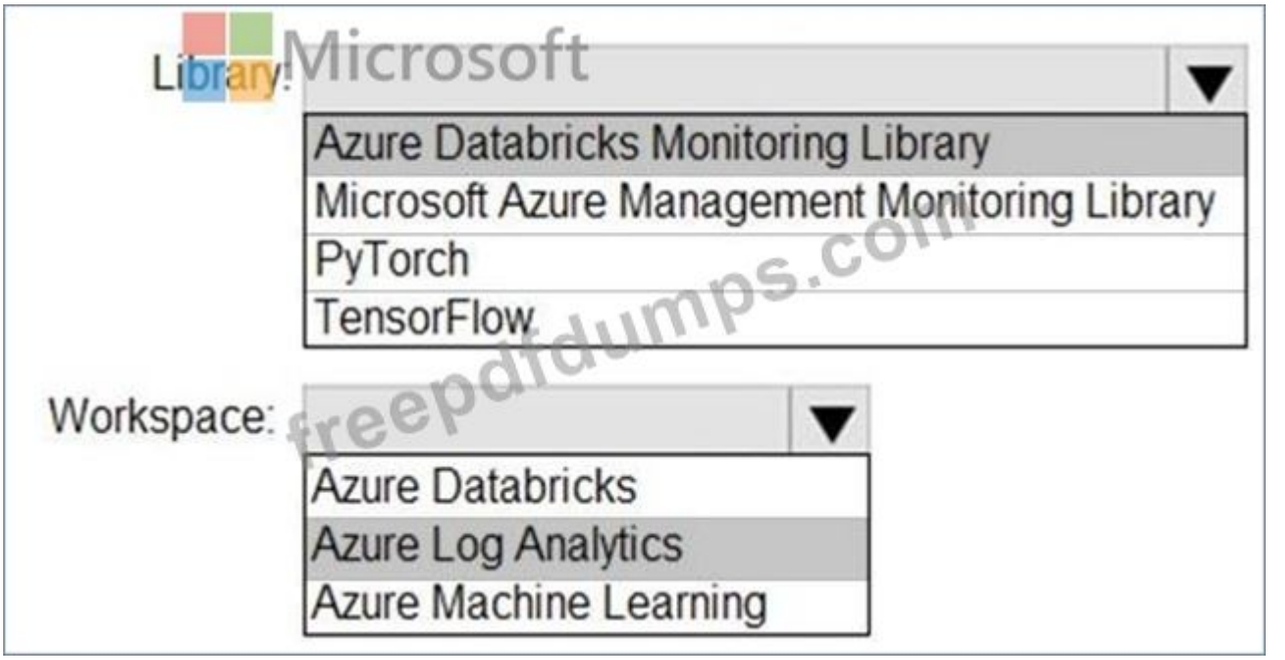
NOTE: Each correct selection is worth one point.




**Answer:**




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.

References:

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

**NEW QUESTION: 123**

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.


**Answer:**


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.

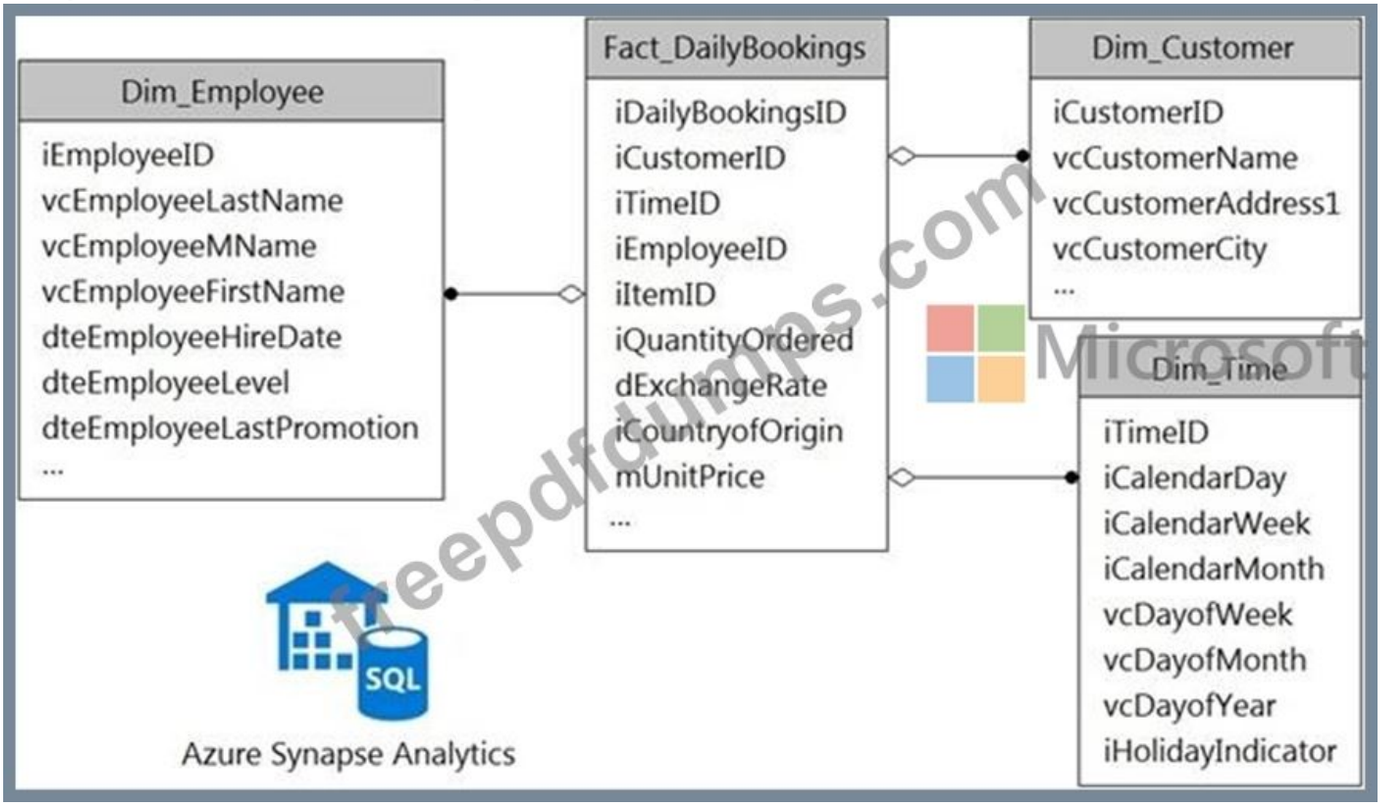
Syntax: DATEDIFF ( datepart , startdate, enddate )

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.

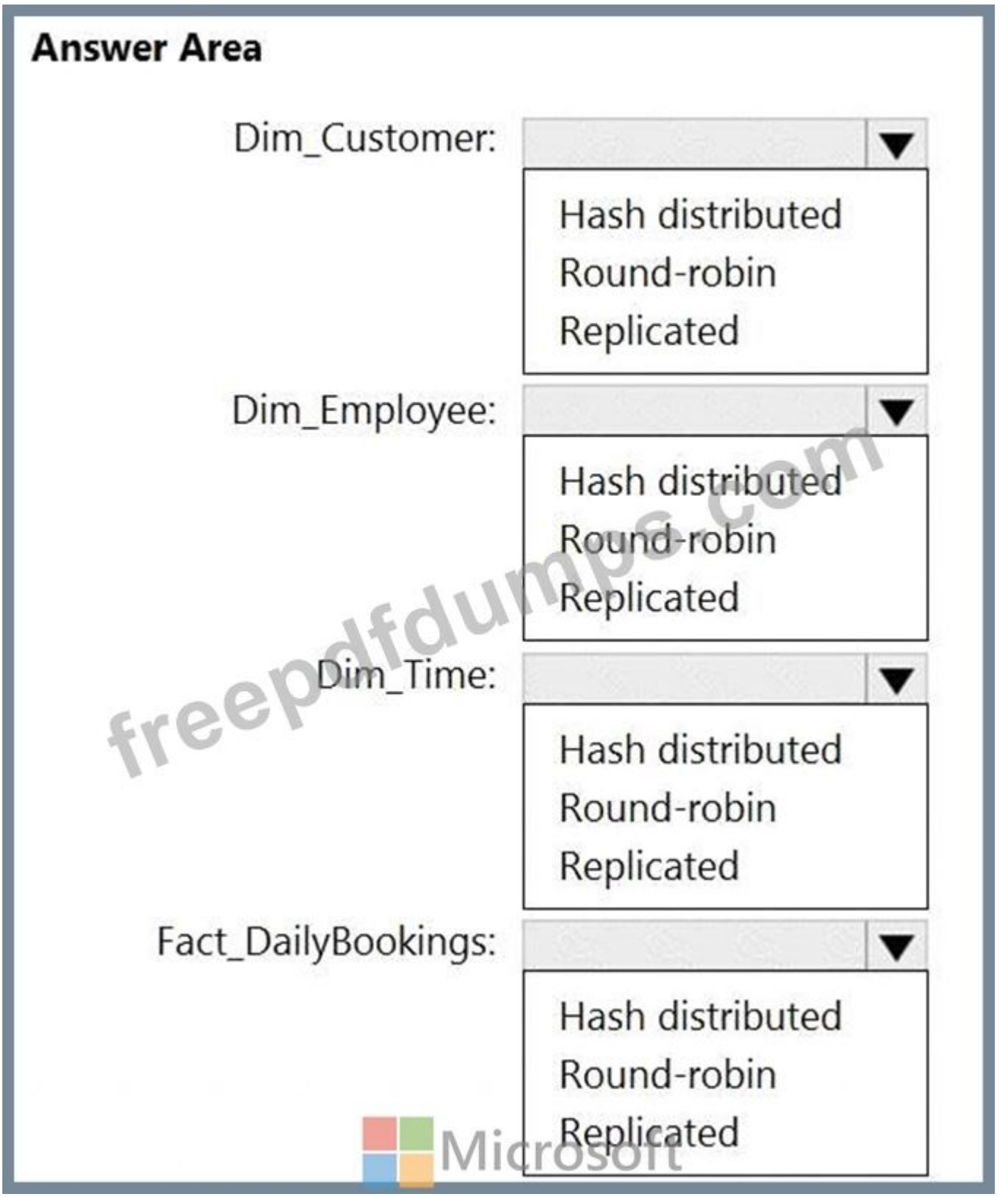
Example:
SELECT
[user],
feature,
DATEDIFF(
second,
LAST(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour,
1) WHEN Event = 'start'),
Time) as duration
FROM input TIMESTAMP BY Time
WHERE
Event = 'end'
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns

**NEW QUESTION: 124**
You have a data model that you plan to implement in a data warehouse in Azure Synapse Analytics as shown in the following exhibit.



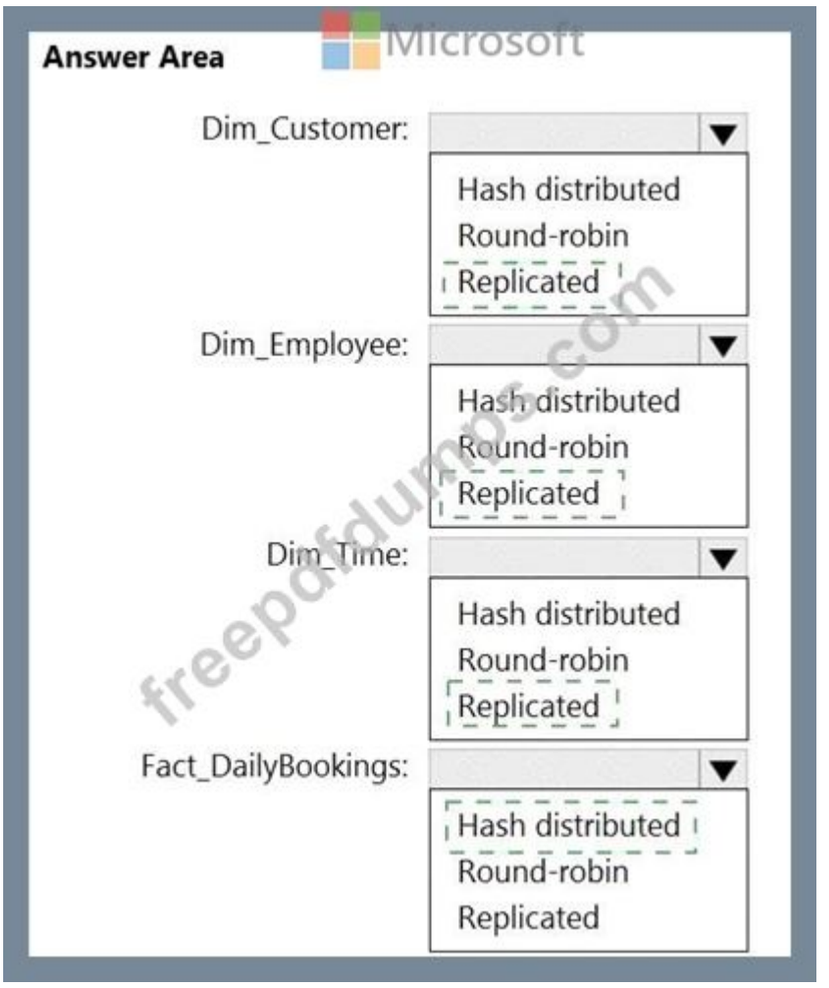

All the dimension tables will be less than 2 GB after compression, and the fact table will be approximately 6 TB.
Which type of table should you use for each table? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

**Answer Area**

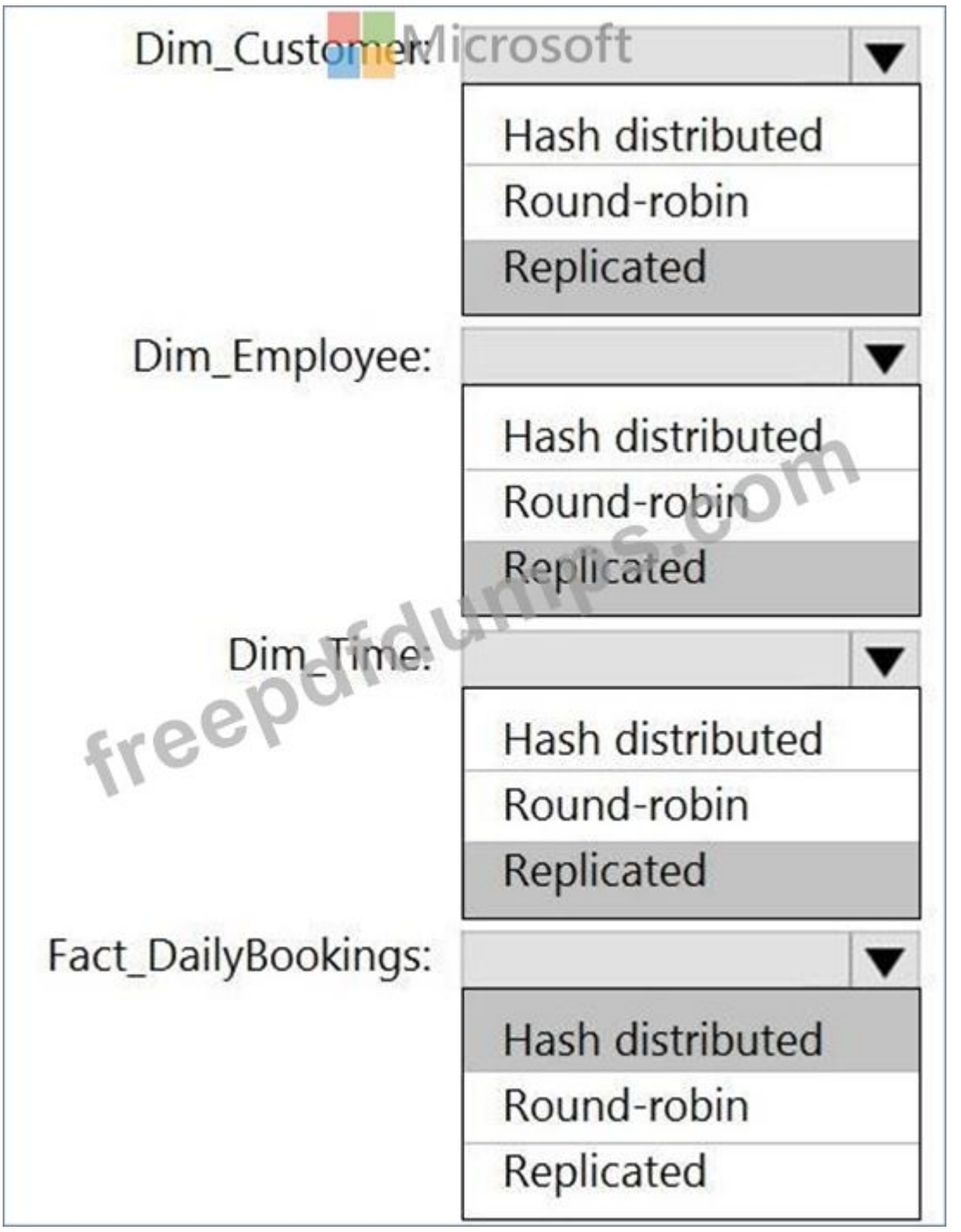





**Answer:**



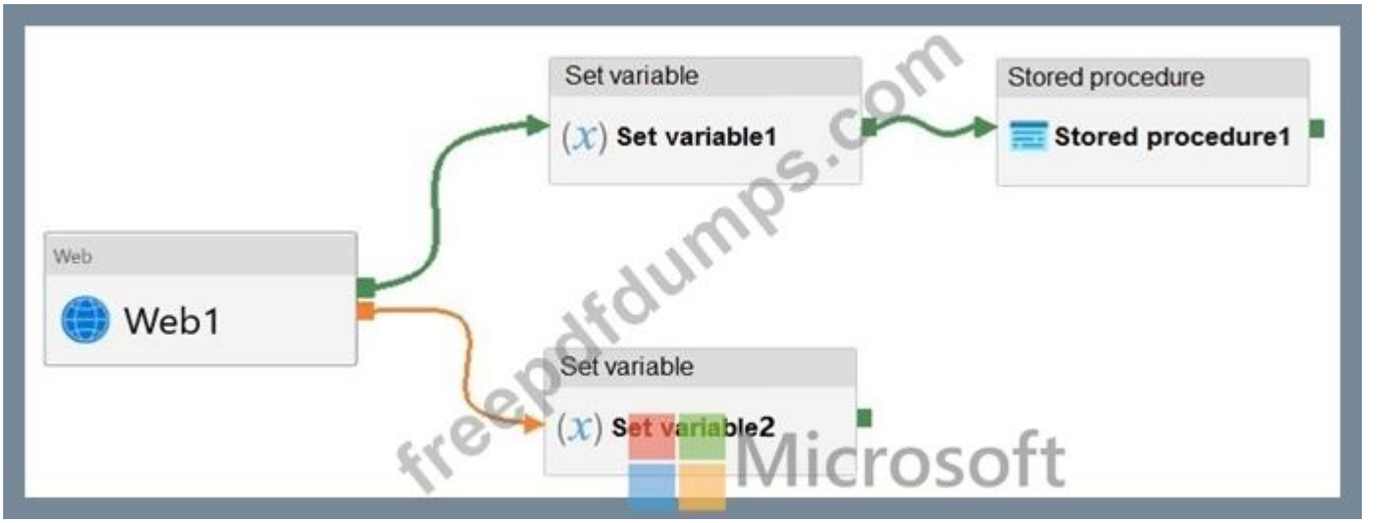

Explanation






**NEW QUESTION: 125**

You have an Azure Data Factory pipeline that has the activities shown in the following exhibit.




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.

Stored procedure1 will execute Web1 and Set variable1 [answer choice]

- complete
- fail
- succeed

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]

- Canceled
- Failed
- Succeeded

**Answer:**

Stored procedure1 will execute Web1 and Set variable1 [answer choice]: succeed

If Web1 fails and Set variable2 succeeds, the pipeline status will be [answer choice]: Failed

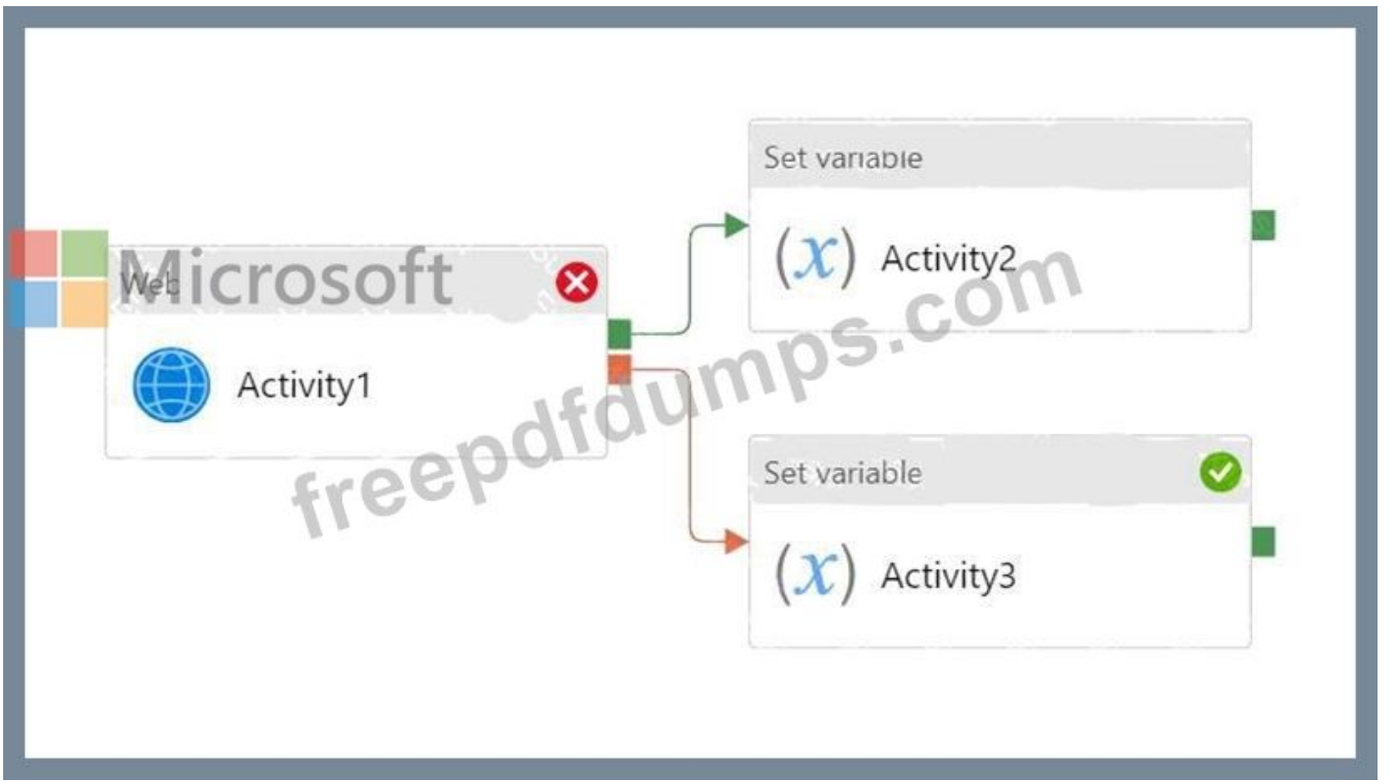
Explanation

Box 1: succeed

Box 2: failed

Example:

Now let's say we have a pipeline with 3 activities, where Activity1 has a success path to Activity2 and a failure path to Activity3. If Activity1 fails and Activity3 succeeds, the pipeline will fail. The presence of the success path alongside the failure path changes the outcome reported by the pipeline, even though the activity executions from the pipeline are the same as the previous scenario.




Activity1 fails, Activity2 is skipped, and Activity3 succeeds. The pipeline reports failure.

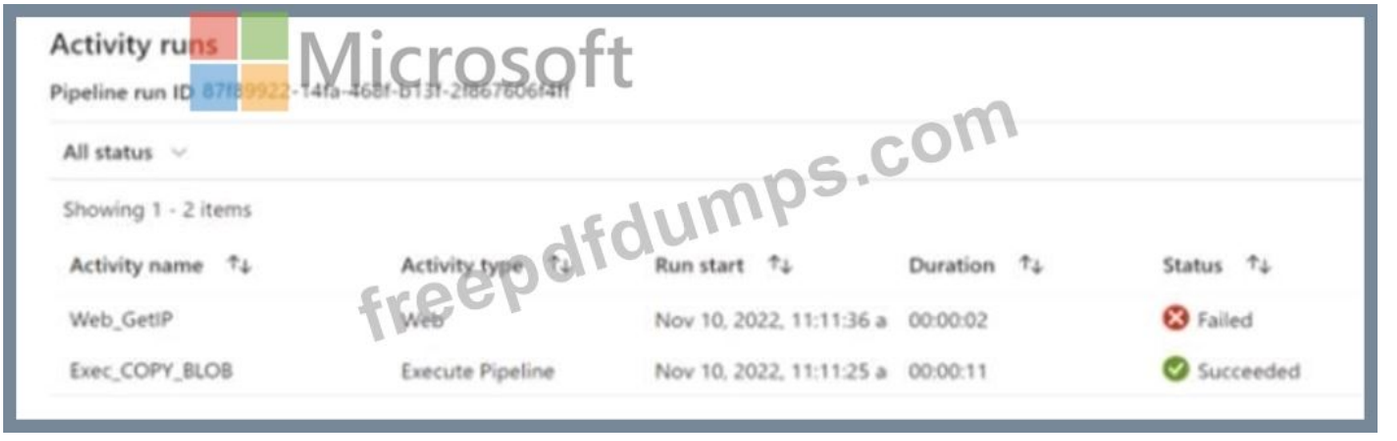
Reference:

https://datasavvy.me/2021/02/18/azure-data-factory-activity-failures-and-pipeline-outcomes/

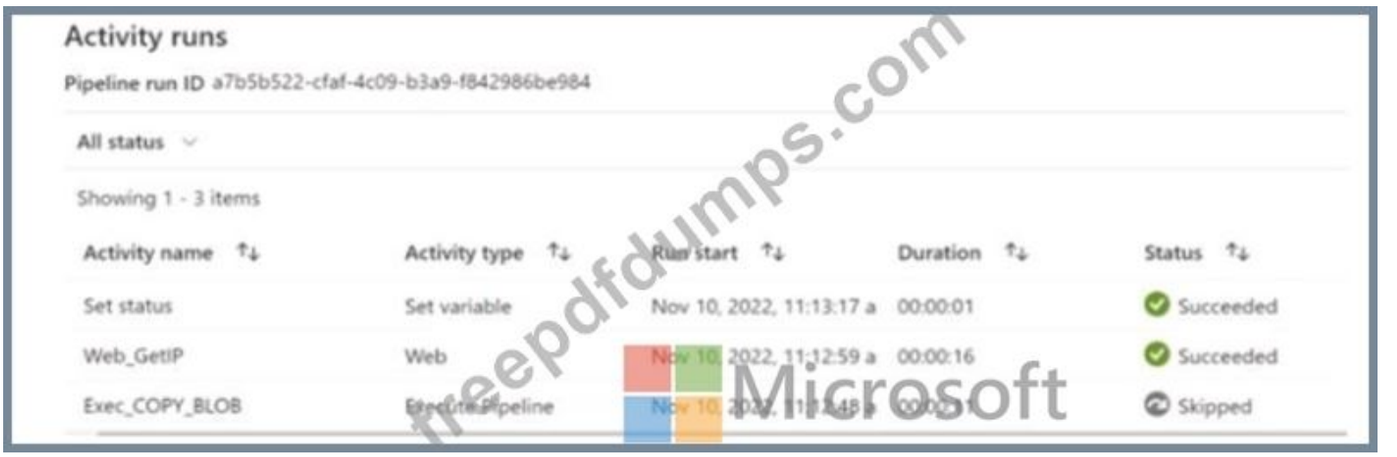
**NEW QUESTION: 126**

You have an Azure Data Factory pipeline shown the following exhibit.


The execution log for the first pipeline run is shown in the following exhibit.

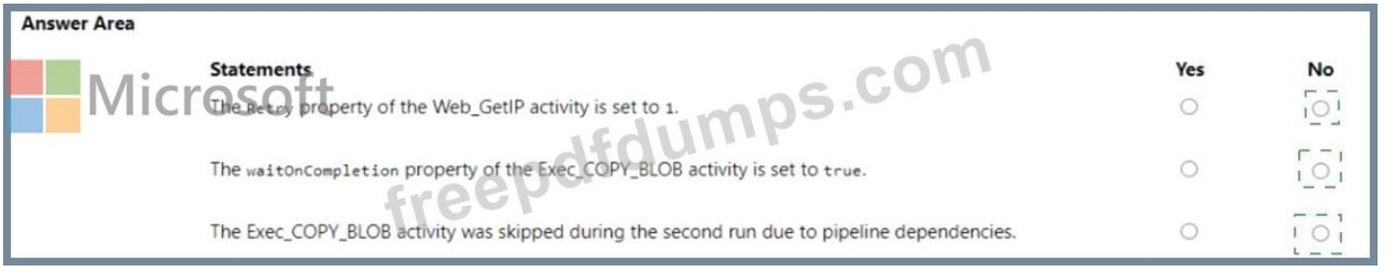



The execution log for the second pipeline run is shown in the following exhibit.

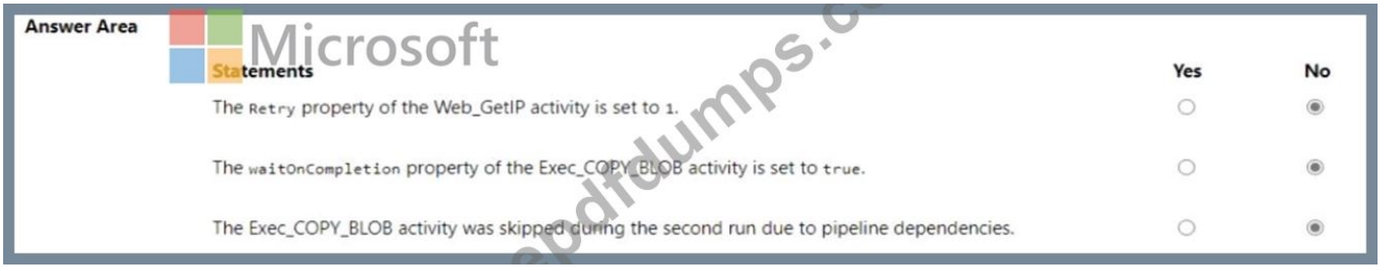



For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.


**Answer:**




Explanation

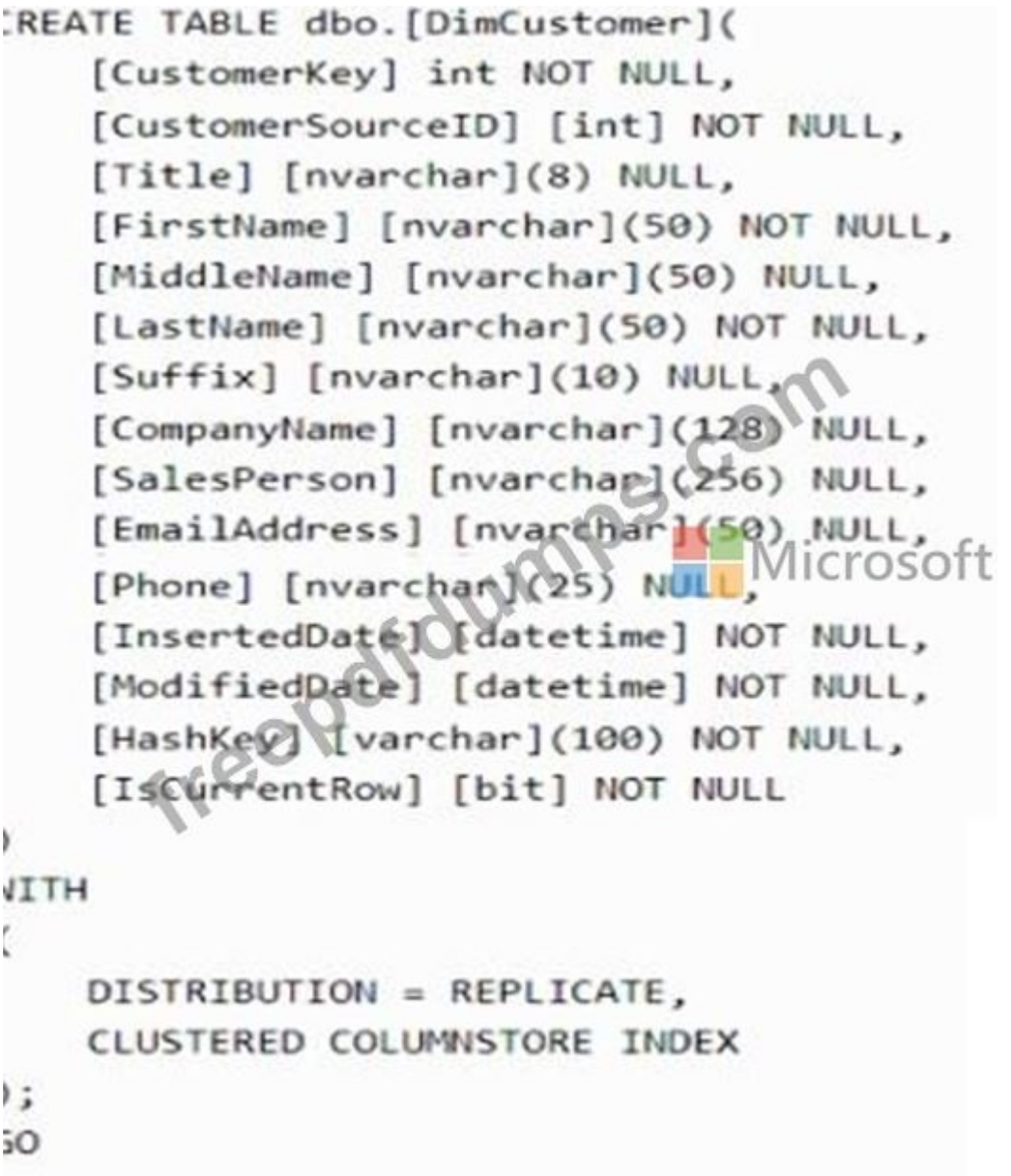



**NEW QUESTION: 127**
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.
You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.
You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.
Solution: You use a dedicated SQL pool to create an external table that has an additional DateTime column.
Does this meet the goal?
**A.** Yes
**B.** No
**Answer: B (LEAVE A REPLY)**
Explanation
Instead use the derived column transformation to generate new columns in your data flow or to modify existing fields.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column

**NEW QUESTION: 128**
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.** storage event
**Answer: D (LEAVE A REPLY)**
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

**NEW QUESTION: 129**
You have an Azure Synapse Analytics dedicated SQL pool named pool1.
You plan to implement a star schema in pool1 and create a new table named DimCustomer by using the following code.




You need to ensure that DimCustomer has the necessary columns to support a Type 2 slowly changing dimension (SCD). Which two columns should you add? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

**A.** [PreviousModifiedDate] [datetime] NOT NULL

**B.** [EffectiveEndDate] [datetime] NOT NULL

**C.** [EffectiveStartDate] [datetime] NOT NULL

**D.** [RowID] [bigint] NOT NULL

**E.** [HistoricalSalesPerson] [nvarchar] (256) NOT NULL

**Answer: (SHOW ANSWER)**

**NEW QUESTION: 130**

You need to design a solution that will process streaming data from an Azure Event Hub and output the data to Azure Data Lake Storage. The solution must ensure that analysts can interactively query the streaming data.

What should you use?

**A.** event triggers in Azure Data Factory

**B.** Azure Stream Analytics and Azure Synapse notebooks

**C.** Structured Streaming in Azure Databricks

**D.** Azure Queue storage and read-access geo-redundant storage (RA-GRS)

**Answer: (SHOW ANSWER)**

Explanation

Apache Spark Structured Streaming is a fast, scalable, and fault-tolerant stream processing API. You can use it to perform analytics on your streaming data in near real-time.

With Structured Streaming, you can use SQL queries to process streaming data in the same way that you would process static data.

Azure Event Hubs is a scalable real-time data ingestion service that processes millions of data in a matter of seconds. It can receive large amounts of data from multiple sources and stream the prepared data to Azure Data Lake or Azure Blob storage.

Azure Event Hubs can be integrated with Spark Structured Streaming to perform the processing of messages in near real-time. You can query and analyze the processed data as it comes by using a Structured Streaming query and Spark SQL.

Reference:

https://k21academy.com/microsoft-azure/data-engineer/structured-streaming-with-azure-event-hubs/

**NEW QUESTION: 131**

You have two Azure SQL databases named DB1 and DB2.

DB1 contains a table named Table 1. Table1 contains a timestamp column named LastModifiedOn.

LastModifiedOn contains the timestamp of the most recent update for each individual row.

DB2 contains a table named Watermark. Watermark contains a single timestamp column named WatermarkValue.

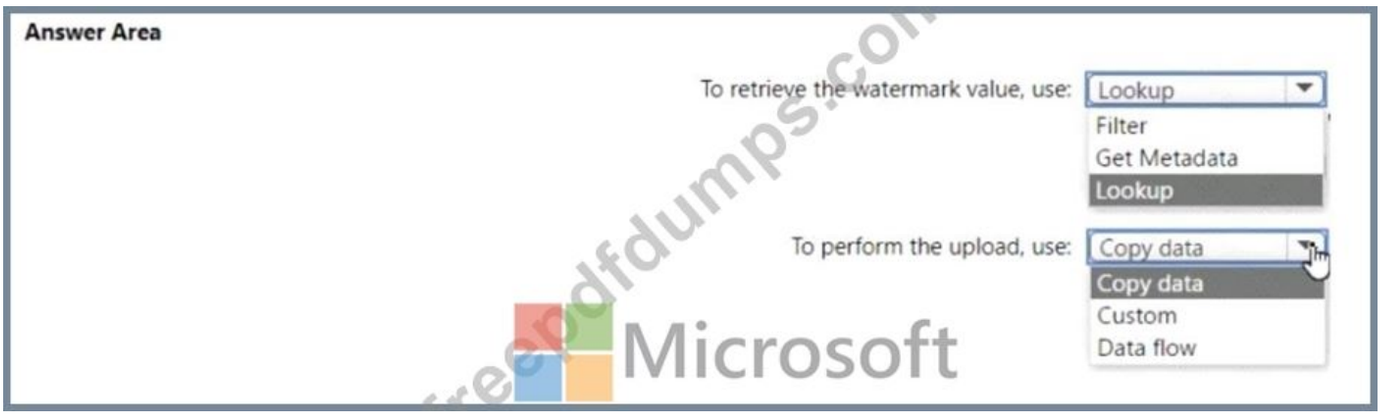
You plan to create an Azure Data Factory pipeline that will incrementally upload into Azure Blob Storage all the rows in Table1 for which the LastModifiedOn column contains a timestamp newer than the most recent value of the WatermarkValue column in Watermark.

You need to identify which activities to include in the pipeline. The solution must meet the following requirements:

* Minimize the effort to author the pipeline.

* Ensure that the number of data integration units allocated to the upload operation can be controlled.

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




**Answer:**


Explanation


**NEW QUESTION: 132**

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1.

Pool1 receives new data once every 24 hours.

You have the following function.

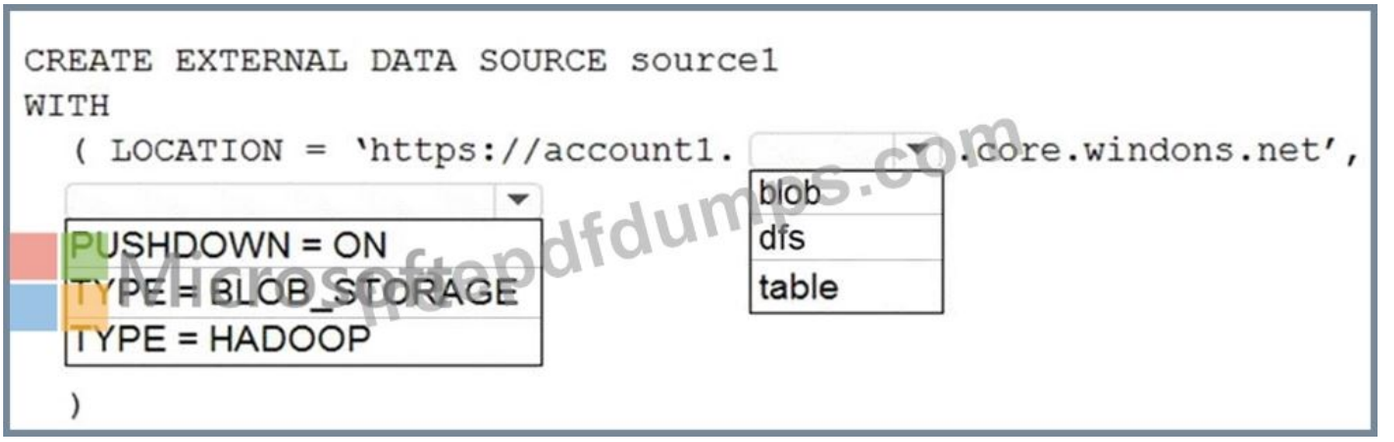
```
create function dbo.udfFtoC(F decimal)
return decimal
as
begin
return (F - 32) * 5.0 / 9
end
```

You have the following query.

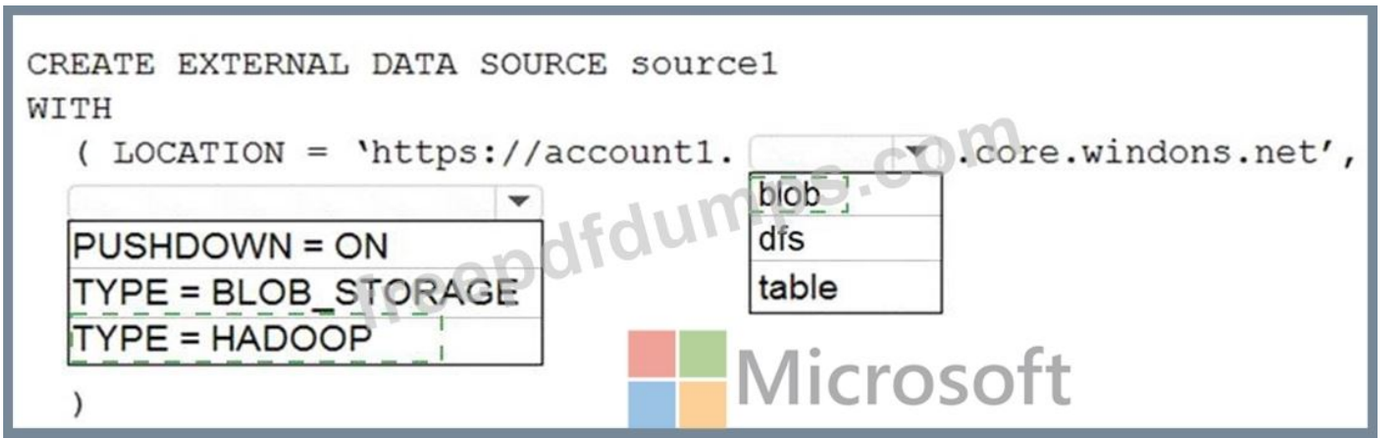
```
select avg_date, sensorid, avg_f, dbo.udfFtoC(avg_temperature) as avg_c from SensorTemps
where avg_date = @parameter
```

The query is executed once every 15 minutes and the @parameter value is set to the current date.
You need to minimize the time it takes for the query to return results.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
**A.** Create an index on the avg_f column.
**B.** Convert the avg_c column into a calculated column.
**C.** Create an index on the sensorid column.
**D.** Enable result set caching.
**E.** Change the table distribution to replicate.
**Answer: B,D (LEAVE A REPLY)**
Explanation
https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-cach

**NEW QUESTION: 133**
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named Account1.
You plan to access the files in Account1 by using an external table.
You need to create a data source in Pool1 that you can reference when you create the external table.
How should you complete the Transact-SQL statement? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

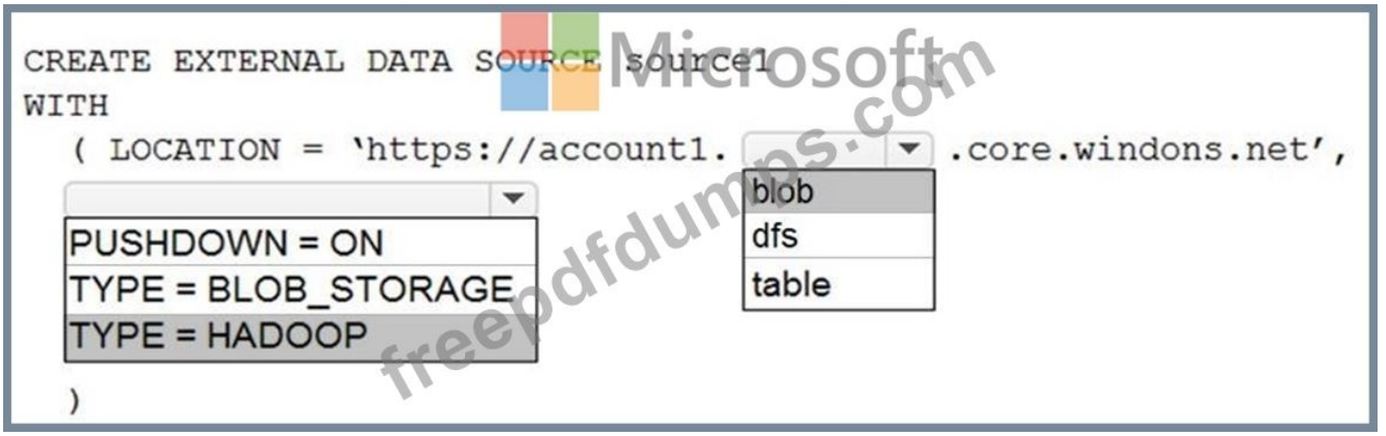



**Answer:**




Explanation

Graphical user interface, diagram Description automatically generated




Box 1: blob

The following example creates an external data source for Azure Data Lake Gen2 CREATE EXTERNAL DATA SOURCE YellowTaxi WITH ( LOCATION = 'https://azureopendatastorage.blob.core.windows.net/nyctlc/yellow/', TYPE = HADOOP) Box 2: HADOOP Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables