Microsoft.DP-203.v2023-10-05.q127

Exam Code:	DP-203
Exam Name:	Data Engineering on Microsoft Azure
Certification Provider:	Microsoft
Free Question Number:	127
Version:	v2023-10-05
# of views:	511
# of Questions views:	10859
https://www.freecram.net/torrent/Microsoft.DP-203.v2023-10-05.q127.html	

NEW QUESTION: 1
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: 2
You have an Azure Data Factory pipeline that is triggered hourly.
The pipeline has had 100% success for the past seven days.

The pipeline execution fails, and two retries that occur 15 minutes apart also fail. The third failure returns the following error.

```
ErrorCode=UserErrorFileNotFound,'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,Message=ADLS Gen2 operation failed for:
Operation returned an invalid status code 'NotFound'. Account: 'contosoproduksouth'. FileSystem: wwi. Path:
'BIKES/CARBON/year=2021/month=01/day=10/hour=06'. ErrorCode: 'PathNotFound'. Message: 'The specified path does not exist.'. RequestId: '6d269b78-
901f-001b-4924-e7a7bc000000'. TimeStamp: 'Sun, 10 Jan 2021 07:45:05
```

What is a possible cause of the error?

A. The pipeline was triggered too early.

B. From 06:00 to 07:00 on January 10, 2021, the file format of data in wwi/BIKES/CARBON was incorrect.

C. From 06:00 to 07:00 on January 10, 2021, there was no data in wwi/BIKES/CARBON.

D. The parameter used to generate year=2021/month=01/day=10/hour=06 was incorrect.

Answer: (SHOW ANSWER)

NEW QUESTION: 3

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.

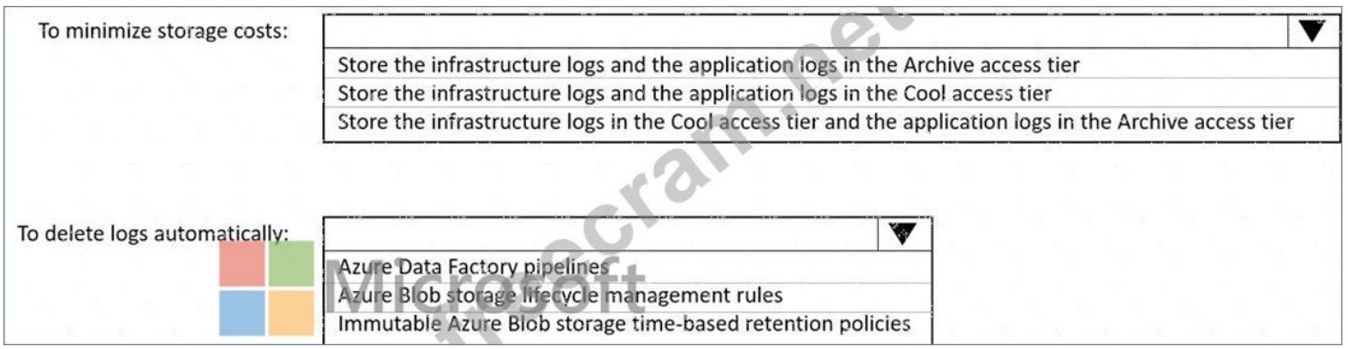
You need to recommend a solution for account1 that meets the following requirements:

Automatically deletes the logs at the end of each retention period

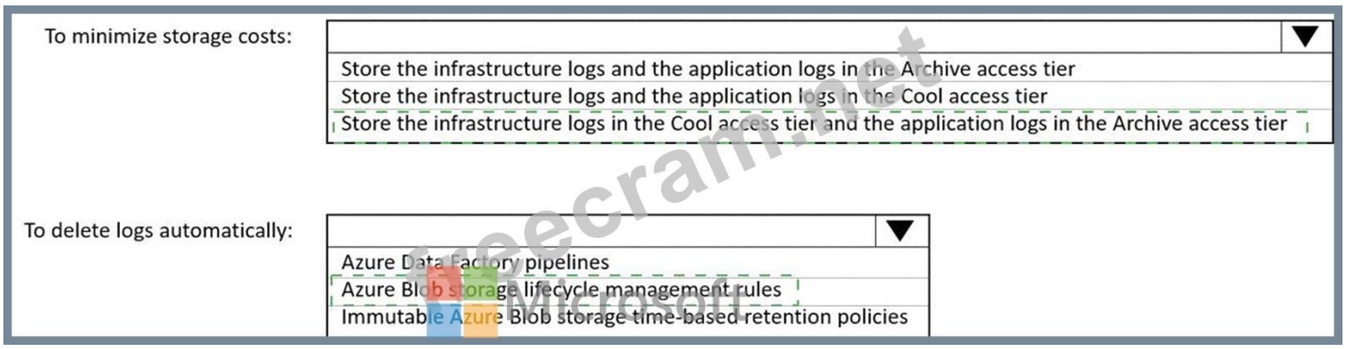
Minimizes storage costs

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.

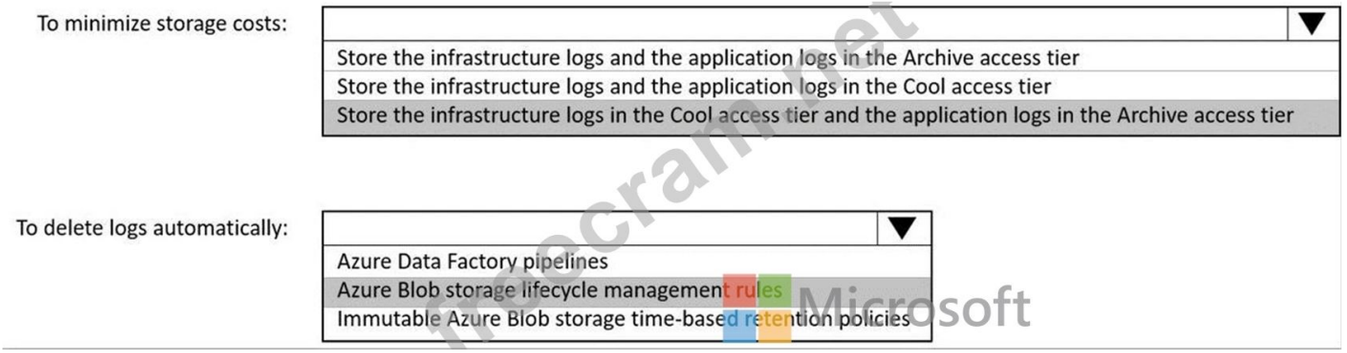


Answer:



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: 4

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

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

Report1: Reads three columns from a file that contains 50 columns.

Report2: Queries a single record based on a timestamp.

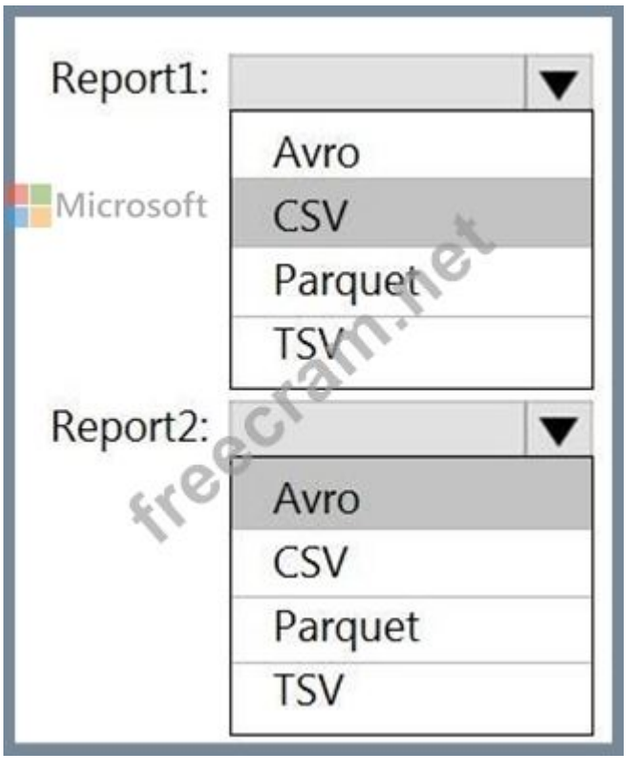
You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.

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

NOTE: Each correct selection is worth one point.

Answer:

Explanation



Report1: CSV

CSV: The destination writes records as delimited data.

Report2: AVRO

AVRO supports timestamps.

Not Parquet, TSV: Not options for Azure Data Lake Storage Gen2.

Reference:

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

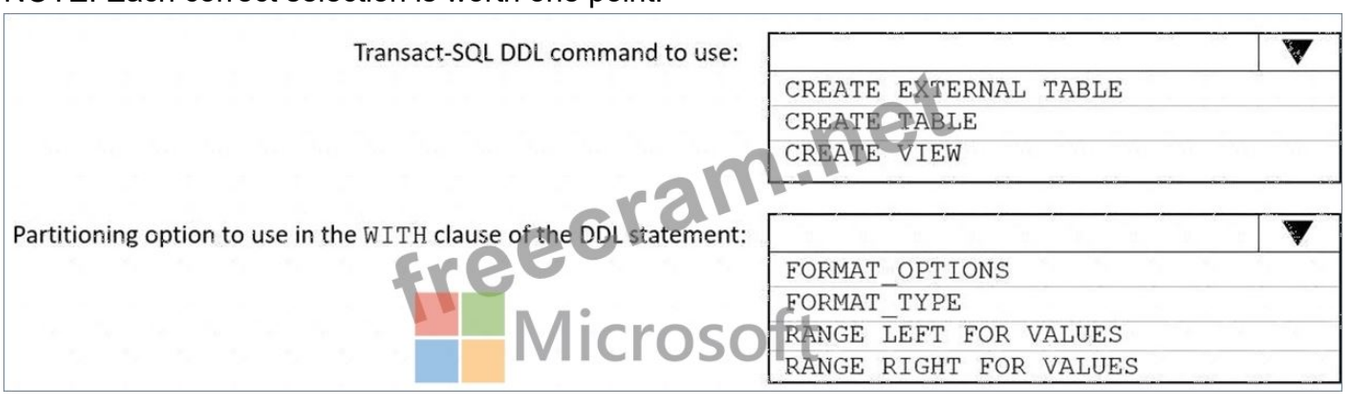
NEW QUESTION: 5

You need to implement an Azure Synapse Analytics database object for storing the sales transactions data. The solution must meet the sales transaction dataset requirements.

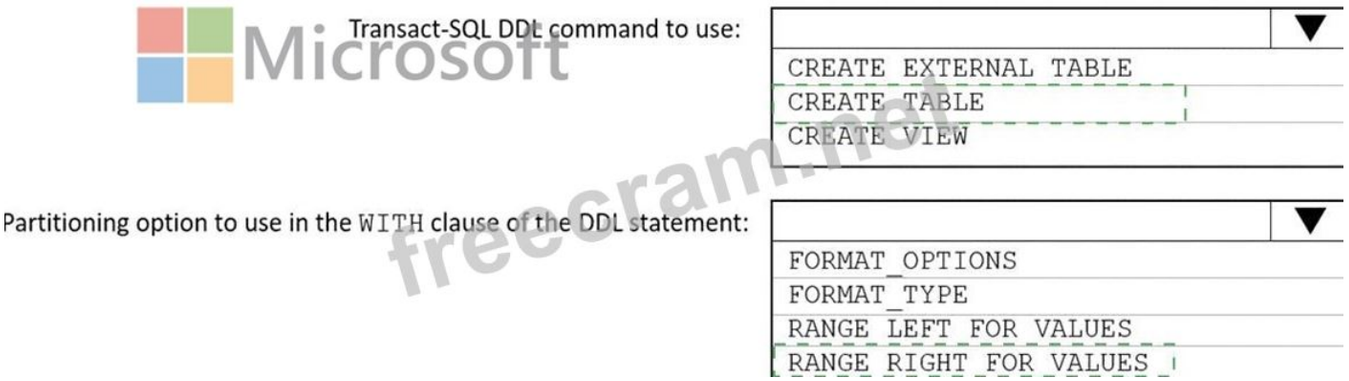
What solution must meet the sales transaction dataset requirements.

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

NOTE: Each correct selection is worth one point.

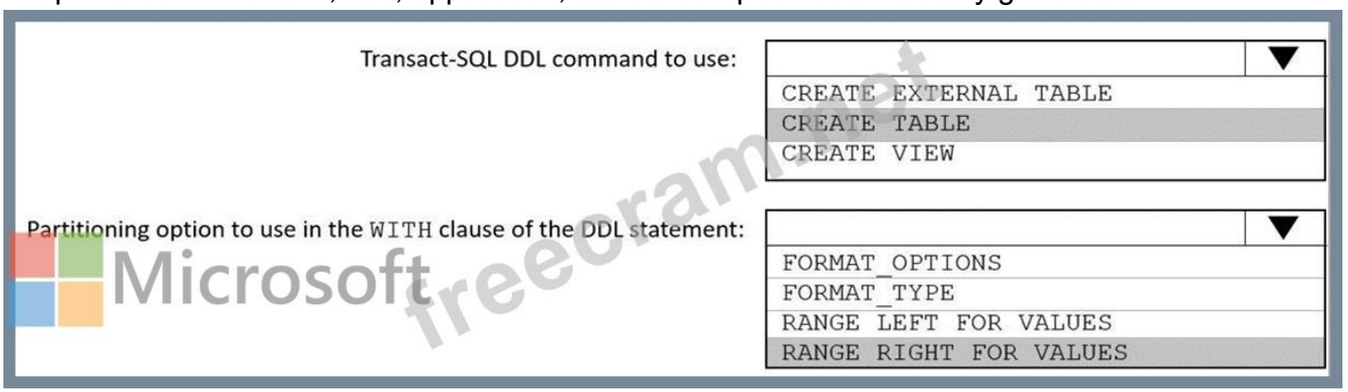


Answer:



Explanation

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



Box 1: Create table

Scenario: Load the sales transaction dataset to Azure Synapse Analytics Box 2: RANGE RIGHT FOR VALUES Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month. Boundary values must belong to the partition on the right.

RANGE RIGHT: Specifies the boundary value belongs to the partition on the right (higher values).

FOR VALUES (boundary_value [,...n]): Specifies the boundary values for the partition.

Scenario: Load the sales transaction dataset to Azure Synapse Analytics.

Contoso identifies the following requirements for the sales transaction dataset:

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

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

Implement a surrogate key to account for changes to the retail store addresses.

Ensure that data storage costs and performance are predictable.

Minimize how long it takes to remove old records.

Reference:

https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse

NEW QUESTION: 6

You have the following Azure Data Factory pipelines

* ingest Data from System 1

* Ingest Data from System2

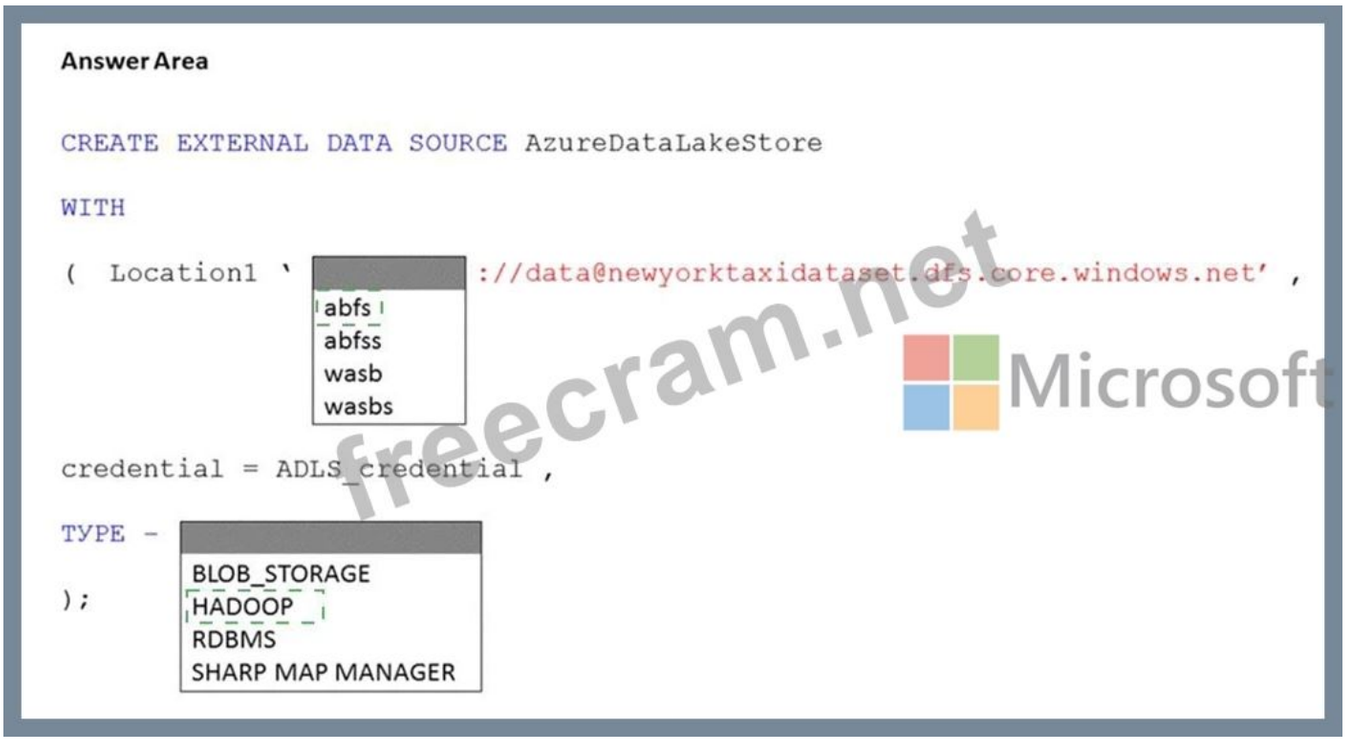
* Populate Dimensions
* Populate facts
ingest Data from System1 and Ingest Data from System1 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest Data from System* Populate Facts must execute after the Populate Dimensions pipeline. All the pipelines must execute 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: (SHOW ANSWER)
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: 7
You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers.
You need to create an external data source in Pool1 that will be used to read .orc files in storage1.
How should you complete the code? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area

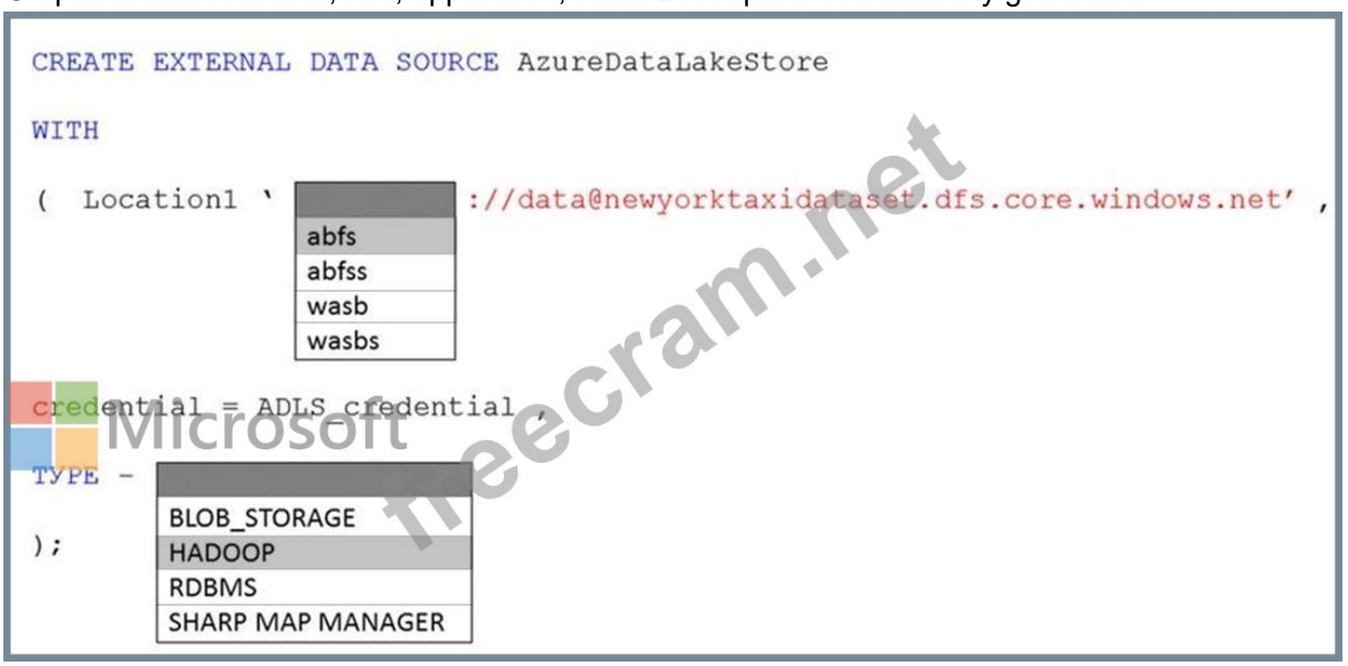
```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore

WITH

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

credential = ADLS_credential ,

TYPE -  [                  ]

);
                BLOB_STORAGE
                HADOOP
                RDBMS
                SHARP MAP MANAGER
```


Answer:



Explanation

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



Reference:

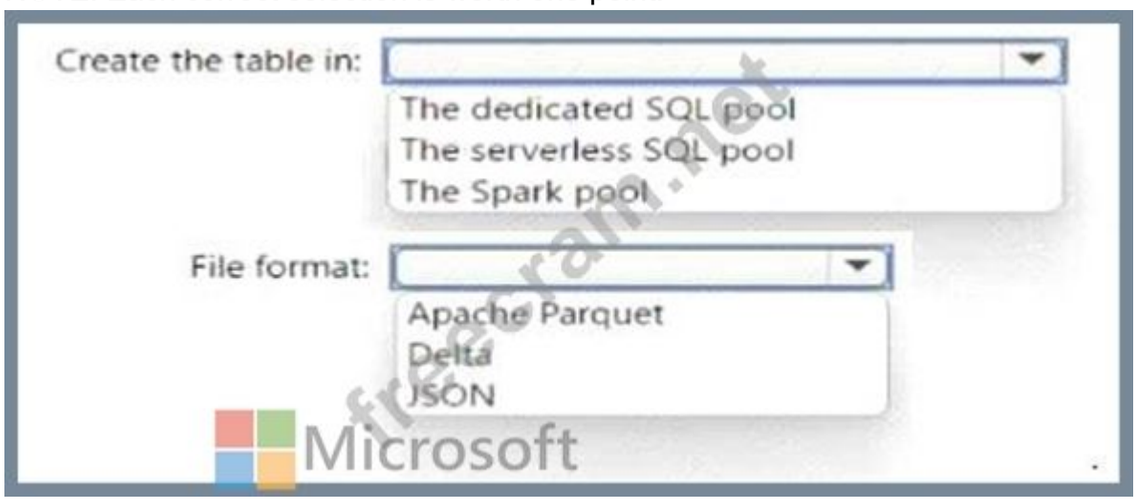
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw

NEW QUESTION: 8

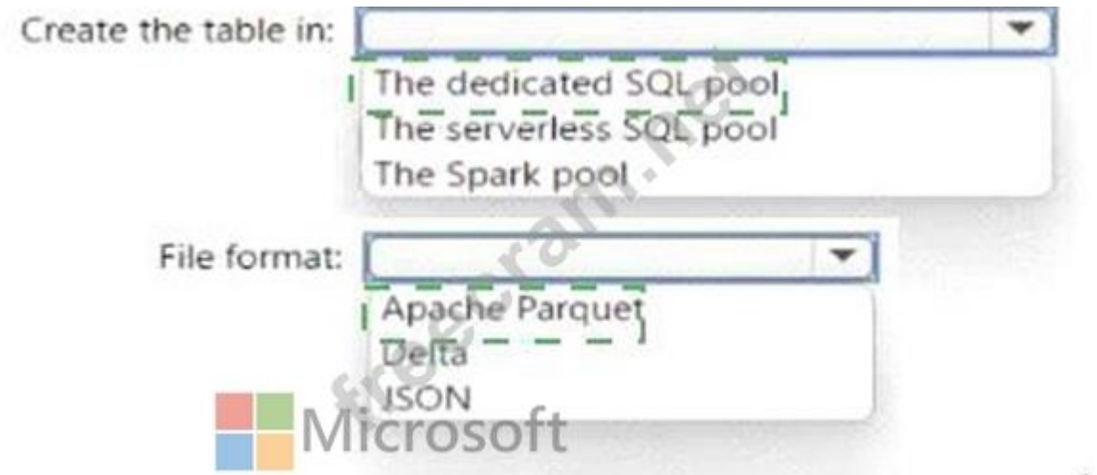
You have an Azure Synapse Analytics serverless SQL pool, an Azure Synapse Analytics dedicated SQL pool, an Apache Spark pool, and an Azure Data Lake Storage Gen2 account.

You need to create a table in a lake database. The table must be available to both the serverless SQL pool and the Spark pool.

Where should you create the table, and Which file format should you use for data in the table? TO answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

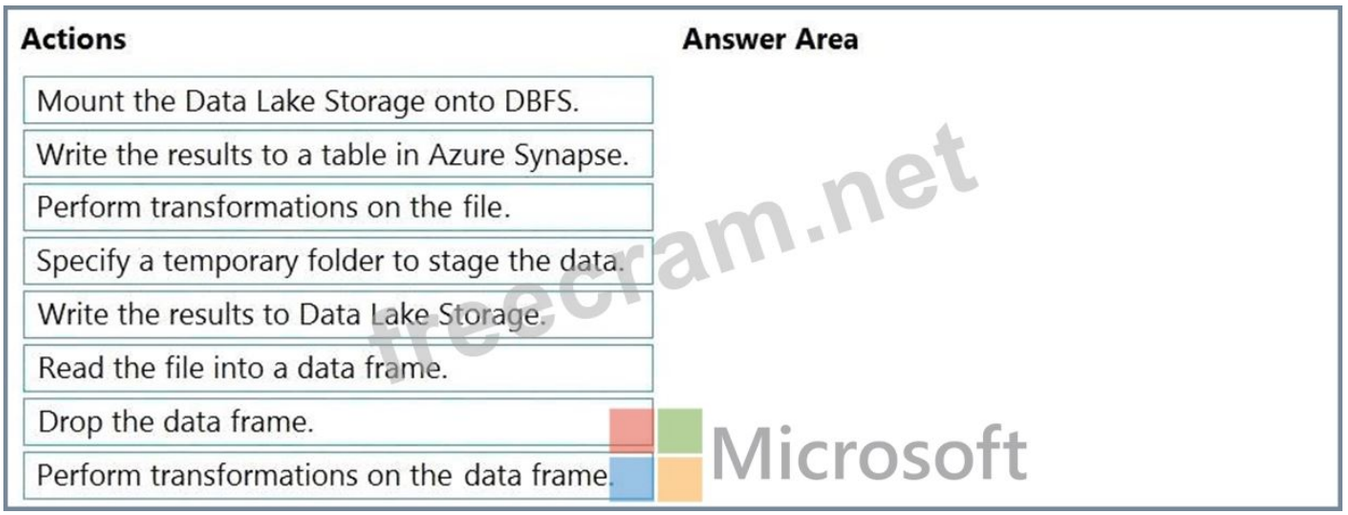


Answer:

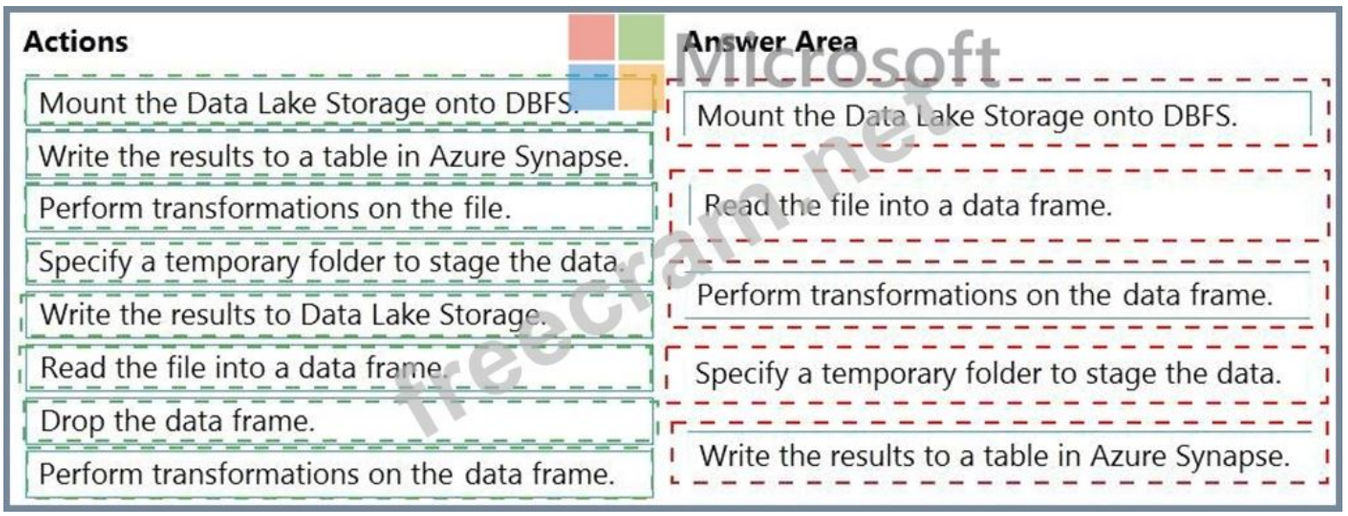


Explanation
The dedicated SQL pool
Apache Parquet

NEW QUESTION: 9
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: 10

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: 11

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: 12

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

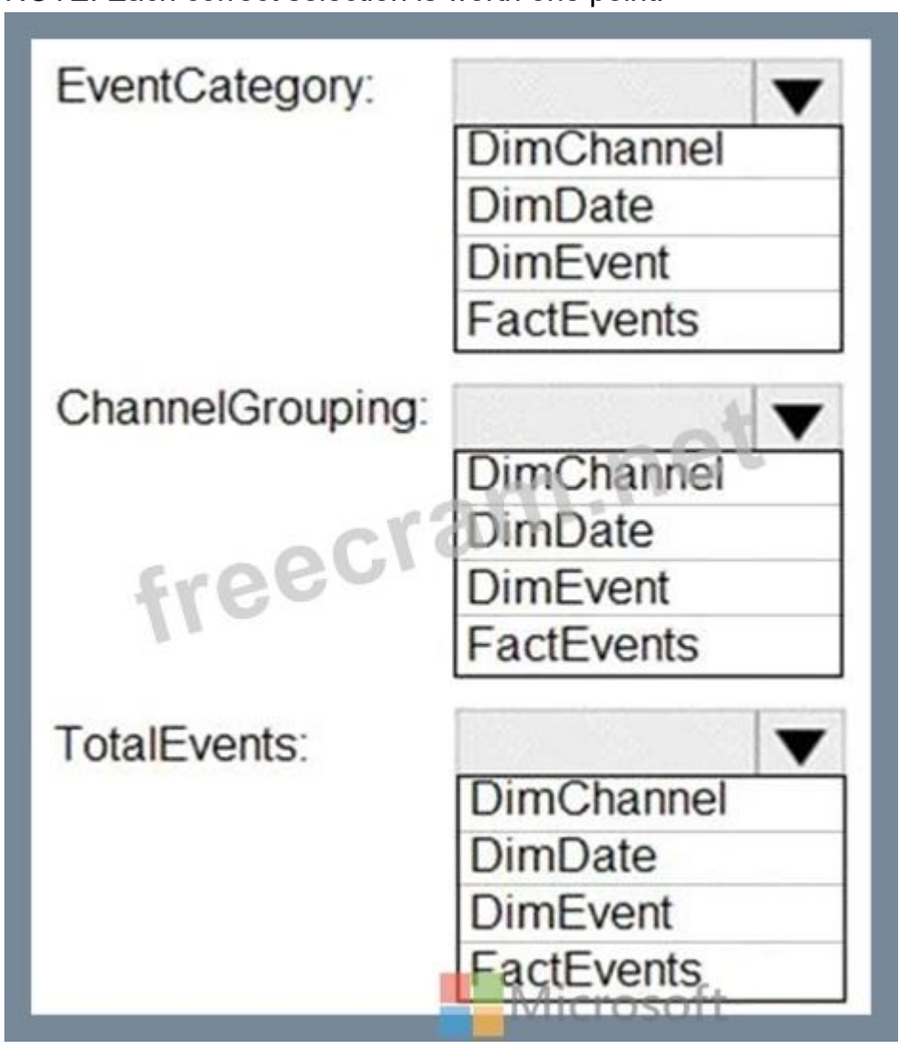
The data contains the following columns.

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

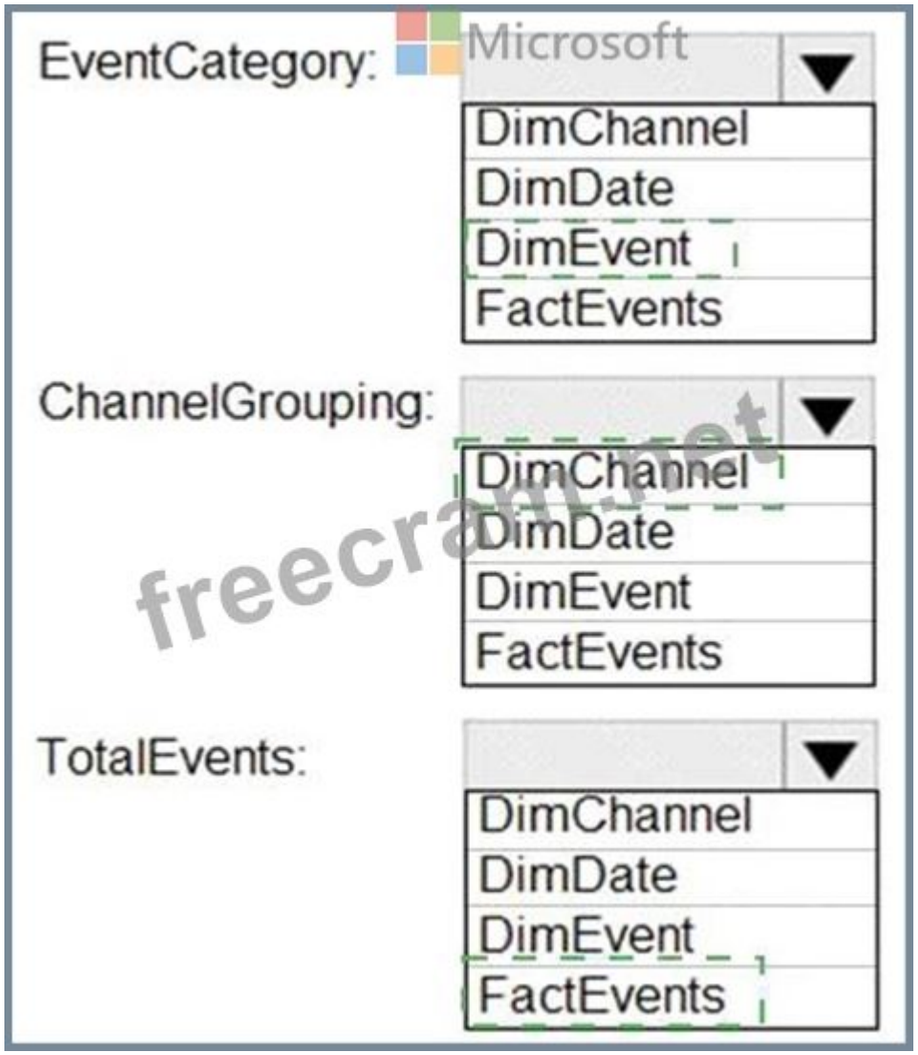
You need to design a star schema to support analytical queries of the data. The star schema will contain four tables including a date dimension.

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

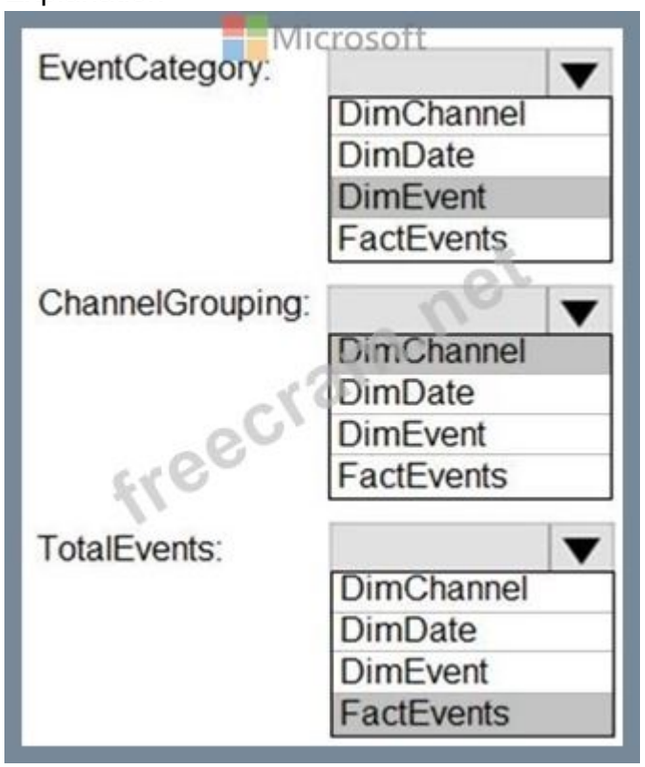


Table Description automatically generated
Box 1: DimEvent
Box 2: DimChannel
Box 3: FactEvents
Fact tables store observations or events, and can be sales orders, stock balances, exchange rates, temperatures, etc Reference:
https://docs.microsoft.com/en-us/power-bi/guidance/star-schema

NEW QUESTION: 13
You have an Azure Stream Analytics job that receives clickstream data from an Azure event hub.
You need to define a query in the Stream Analytics job. The query must meet the following requirements:
* Count the number of clicks within each 10-second window based on the country of a visitor.
* Ensure that each click is NOT counted more than once.
How should you define the Query?
A. SELECT Country, Avg(*) AS Average
FROM ClickStream TIMESTAMP BY CreatedAt
GROUP BY Country, SlidingWindow(second, 10)
B. SELECT Country, Count(*) AS Count
FROM ClickStream TIMESTAMP BY CreatedAt
GROUP BY Country, TumblingWindow(second, 10)
C. SELECT Country, Avg(*) AS Average
FROM ClickStream TIMESTAMP BY CreatedAt
GROUP BY Country, HoppingWindow(second, 10, 2)
D. SELECT Country, Count(*) AS Count
FROM ClickStream TIMESTAMP BY CreatedAt
GROUP BY Country, SessionWindow(second, 5, 10)
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.
Example:
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

NEW QUESTION: 14
You have an Apache Spark DataFrame named temperatures. A sample of the data is shown in the following table.

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

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

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

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

NOTE: Each correct selection is worth one point.

Values

CAST

COLLATE

CONVERT

FLATTEN

PIVOT

UNPIVOT

Answer Area

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

Answer:

Values **Answer Area**

CAST
COLLATE
CONVERT
FLATTEN
PIVOT
UNPIVOT

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


Explanation

Text Description automatically generated

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


Box 1: PIVOT

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

Reference:

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

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

NEW QUESTION: 15

You have an Azure Databricks workspace that contains a Delta Lake dimension table named Tablet.
Table1 is a Type 2 slowly changing dimension (SCD) table. You need to apply updates from a source table to Table1.
Which Apache Spark SQL operation should you use?
A. CREATE
B. UPDATE
C. MERGE
D. ALTER
Answer: (SHOW ANSWER)
Explanation
The Delta provides the ability to infer the schema for data input which further reduces the effort required in managing the schema changes. The Slowly Changing Data(SCD) Type 2 records all the changes made to each key in the dimensional table. These operations require updating the existing rows to mark the previous values of the keys as old and then inserting new rows as the latest values. Also, Given a source table with the updates and the target table with dimensional data, SCD Type 2 can be expressed with the merge.
Example:

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

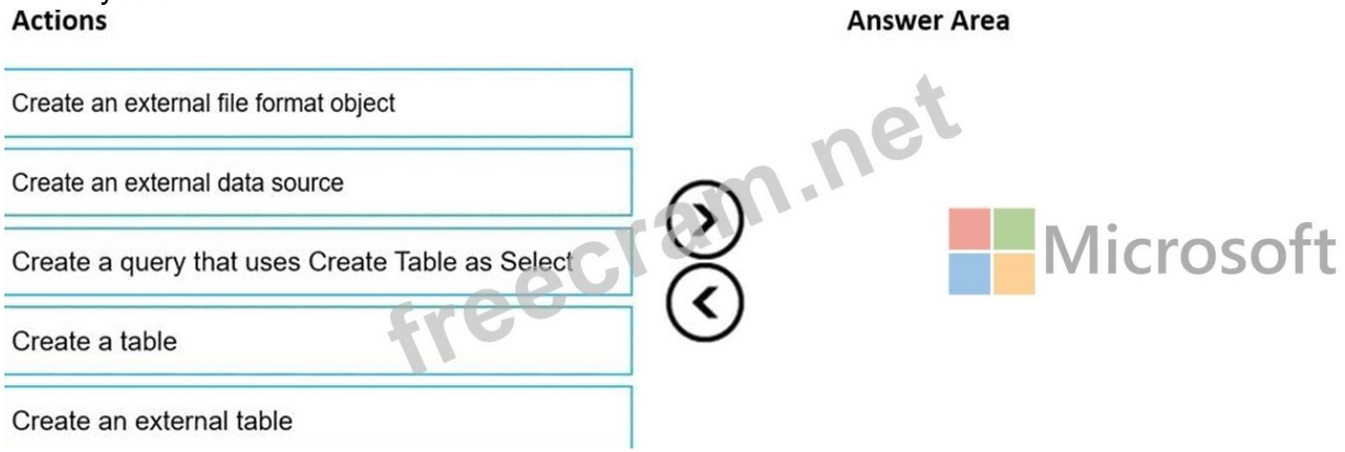
Reference:
https://www.projectpro.io/recipes/what-is-slowly-changing-data-scd-type-2-operation-delta-table-databricks

NEW QUESTION: 16

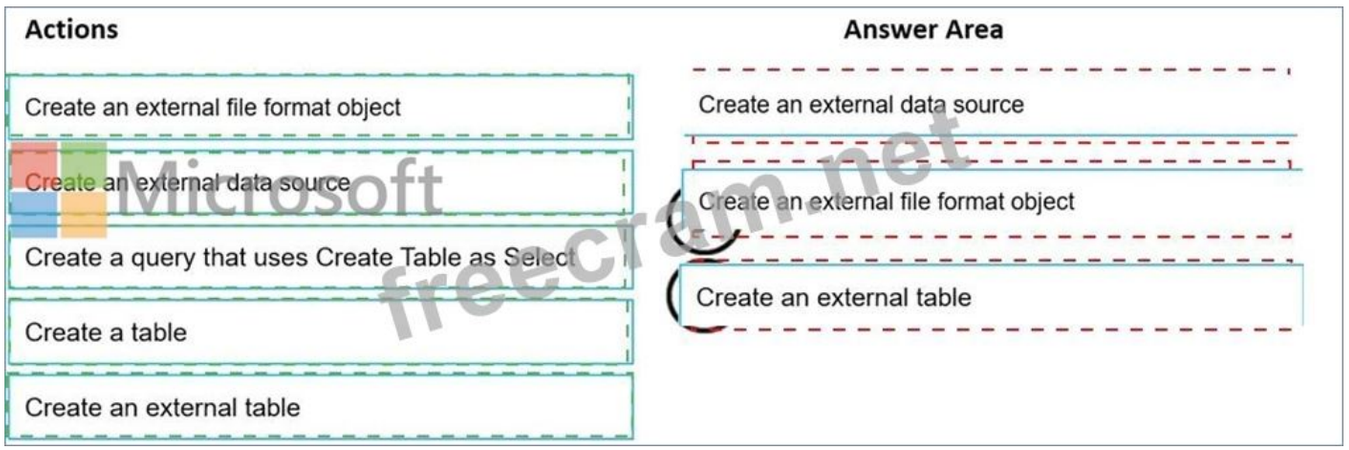
You need to build a solution to ensure that users can query specific files in an Azure Data Lake Storage Gen2 account from an Azure Synapse Analytics serverless SQL pool.

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

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

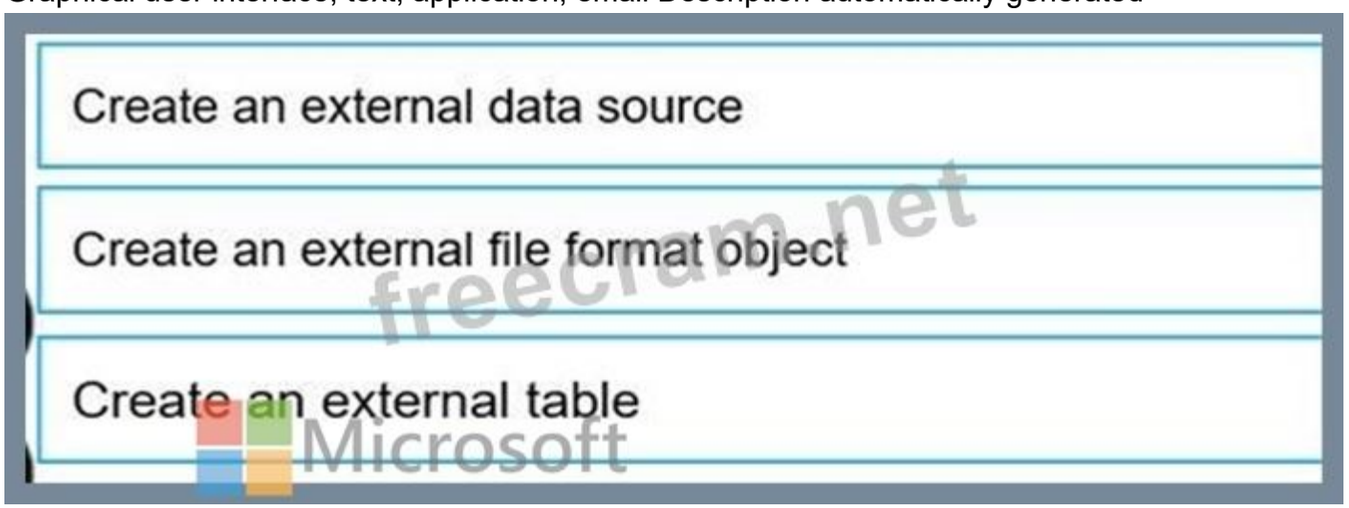


Answer:



Explanation

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



Step 1: Create an external data source

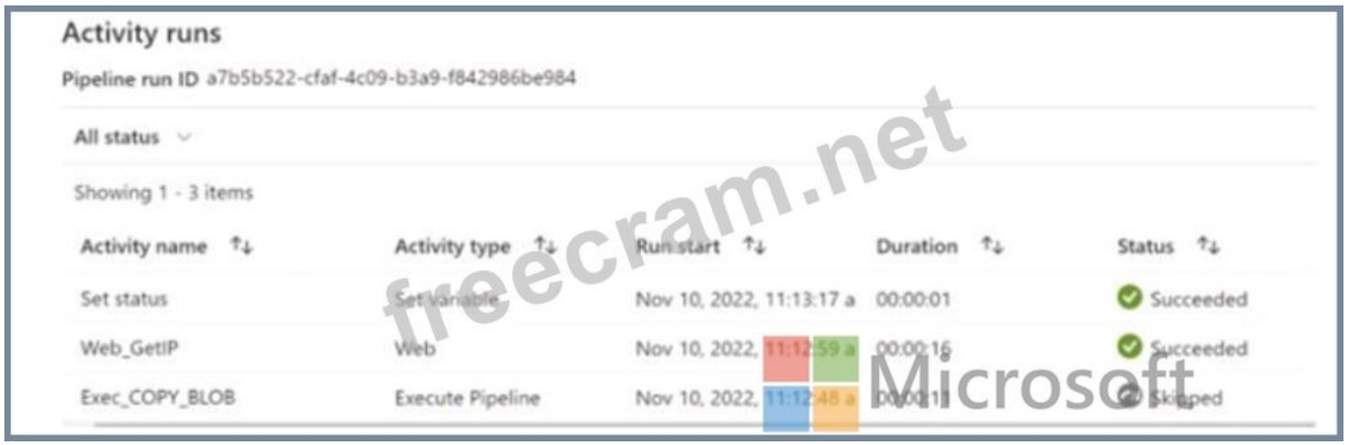
You can create external tables in Synapse SQL pools via the following steps:

CREATE EXTERNAL DATA SOURCE to reference an external Azure storage and specify the credential that should be used to access the storage.
CREATE EXTERNAL FILE FORMAT to describe format of CSV or Parquet files.
CREATE EXTERNAL TABLE on top of the files placed on the data source with the same file format.
Step 2: Create an external file format object
Creating an external file format is a prerequisite for creating an external table.
Step 3: Create an external table
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

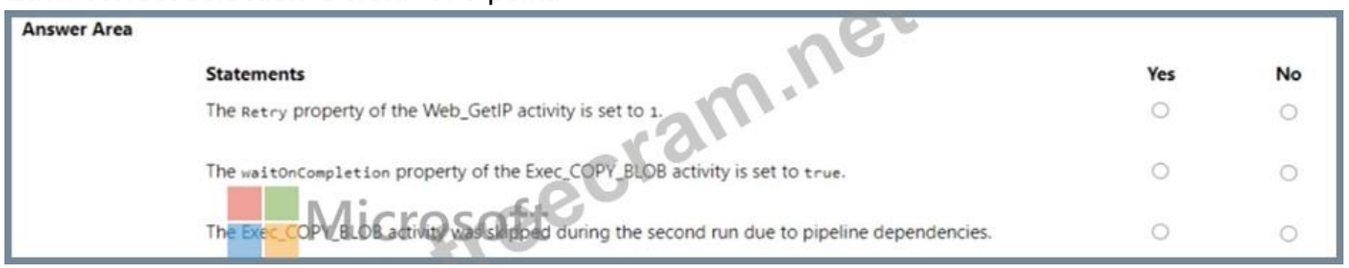
NEW QUESTION: 17
You have an Azure Data Factory pipeline shown the following exhibit.

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.

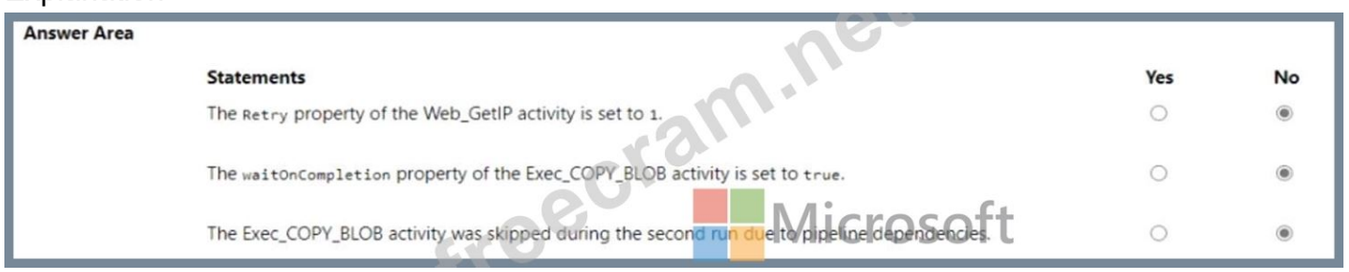


For each of the following statements, select Yes if the statement is true. Otherwise, select No. NOTE: Each correct selection is worth one point.



Answer:

Explanation



NEW QUESTION: 18

You have an activity in an Azure Data Factory pipeline. The activity calls a stored procedure in a data warehouse in Azure Synapse Analytics and runs daily.

You need to verify the duration of the activity when it ran last.

What should you use?

A. activity runs in Azure Monitor

B. Activity log in Azure Synapse Analytics

C. the sys.dm_pdw_wait_stats data management view in Azure Synapse Analytics

D. an Azure Resource Manager template

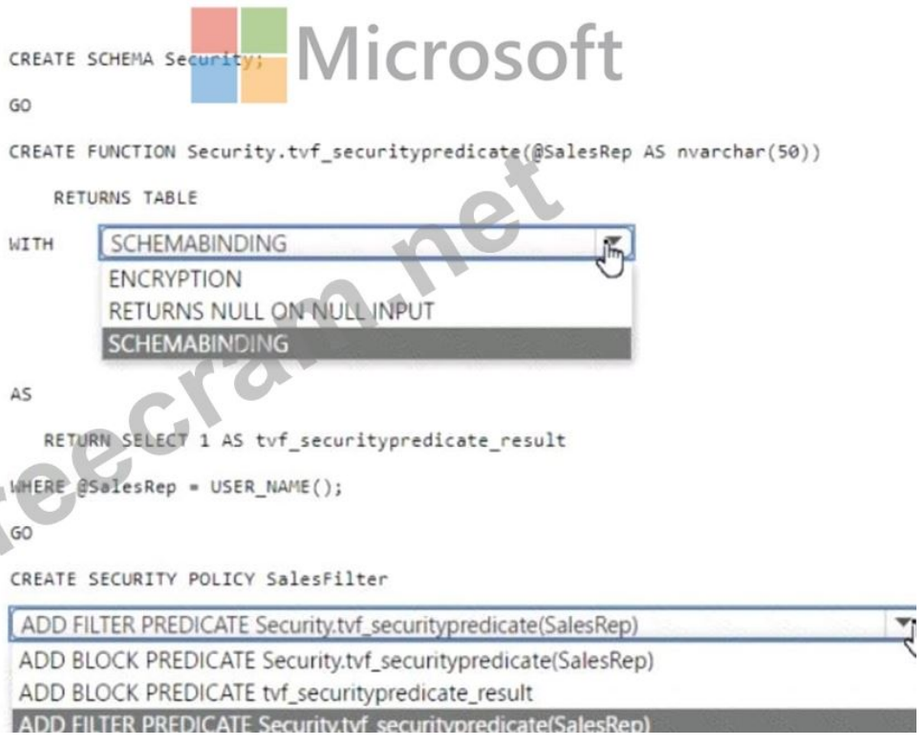
Answer: (SHOW ANSWER)

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

NEW QUESTION: 19
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: (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.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger

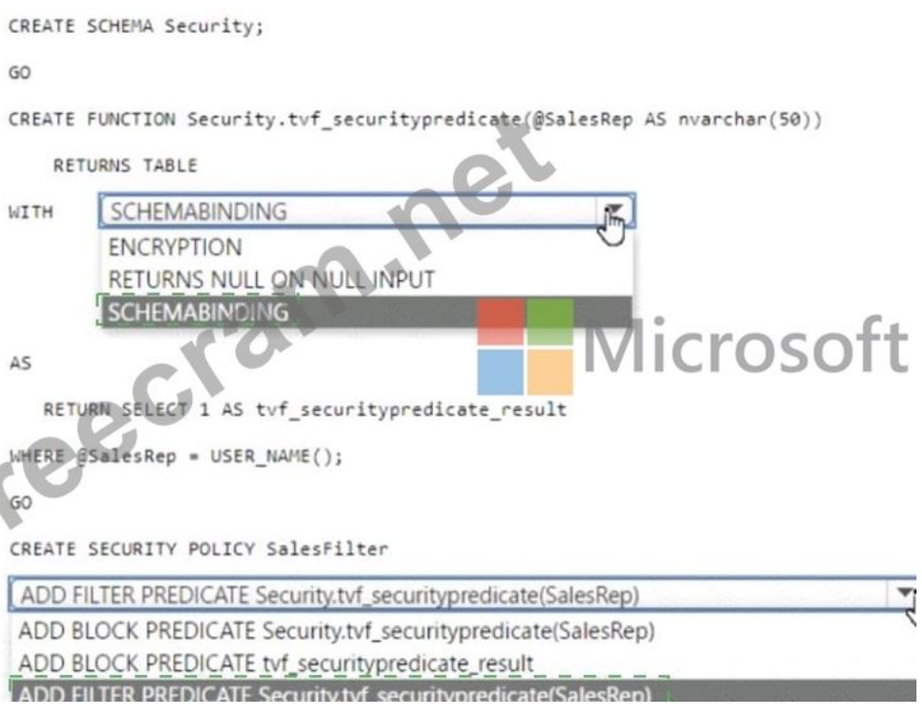
NEW QUESTION: 20
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Sales.Orders. Sales.Orders contains a column named SalesRep.
You plan to implement row-level security (RLS) for Sales.Orders.
You need to create the security policy that will be used to implement RLS. 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 Area



Answer:

Answer Area



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: 21

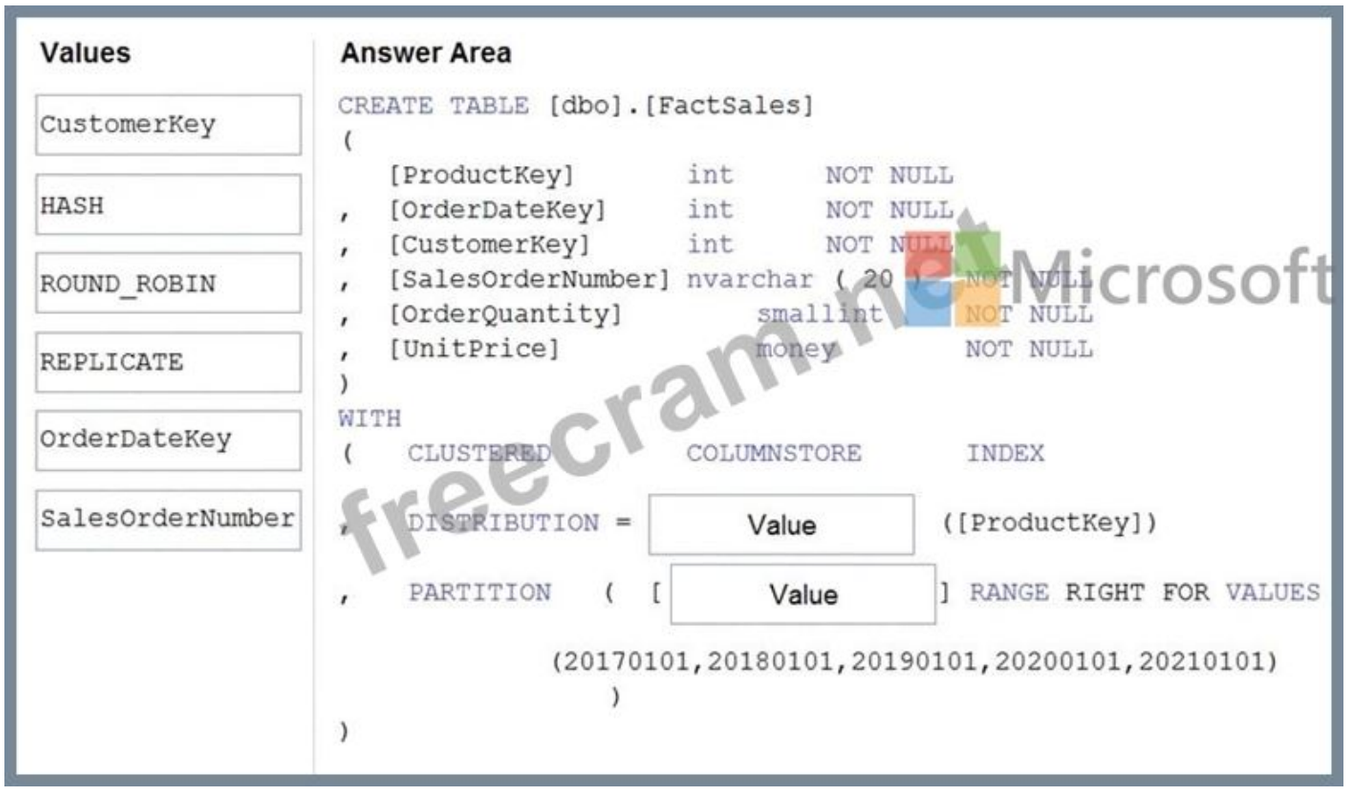
You plan to create a table in an Azure Synapse Analytics dedicated SQL pool.

Data in the table will be retained for five years. Once a year, data that is older than five years will be deleted.

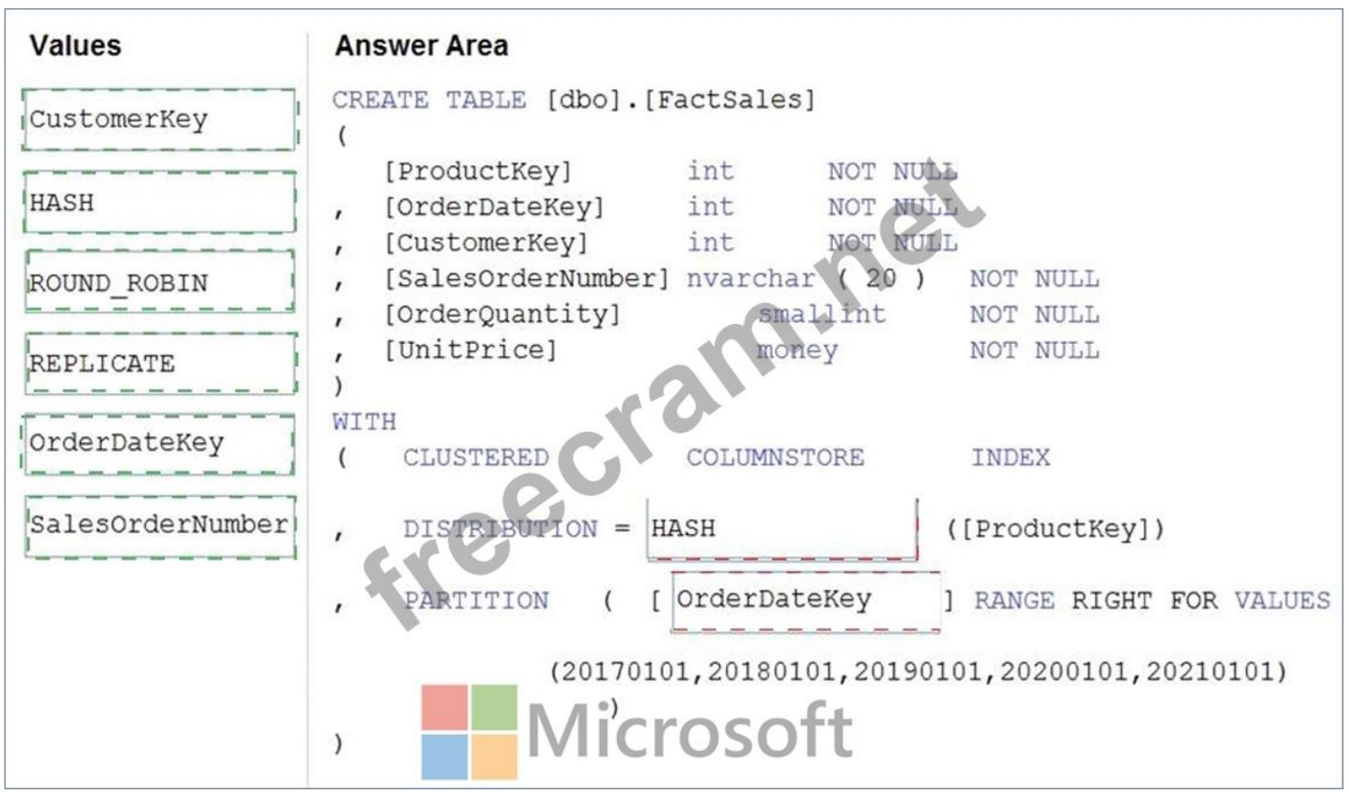
You need to ensure that the data is distributed evenly across partitions. The solution must minimize the amount of time required to delete old data.

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.



Answer:



Explanation

Box 1: HASH

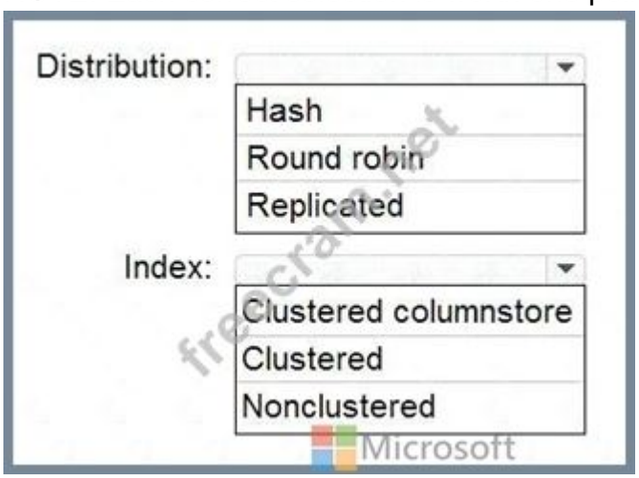
Box 2: OrderDateKey

In most cases, table partitions are created on a date column.

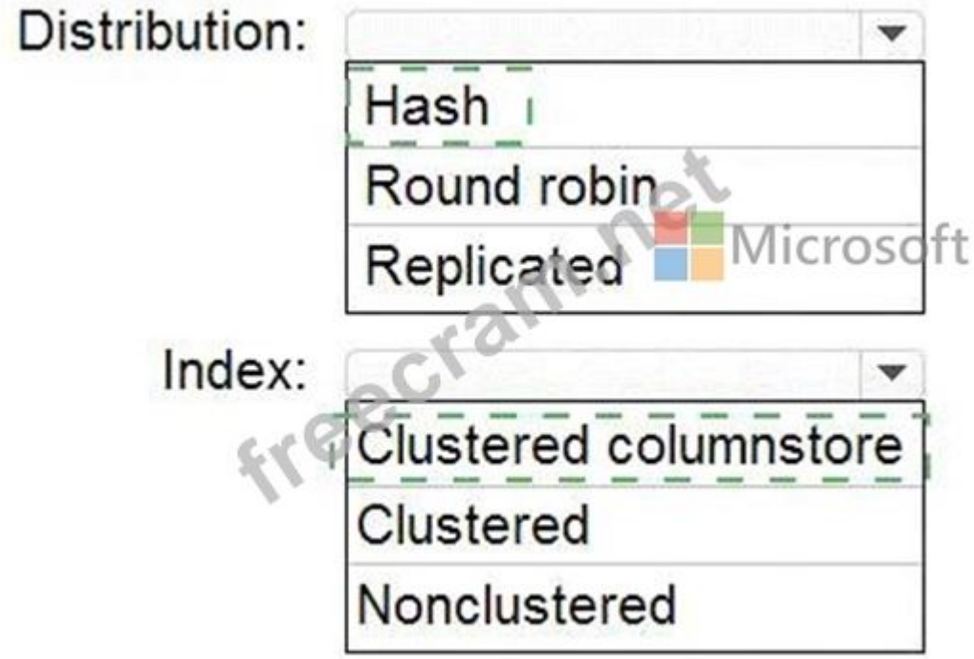
A way to eliminate rollbacks is to use Metadata Only operations like partition switching for data management.

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: 22
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:



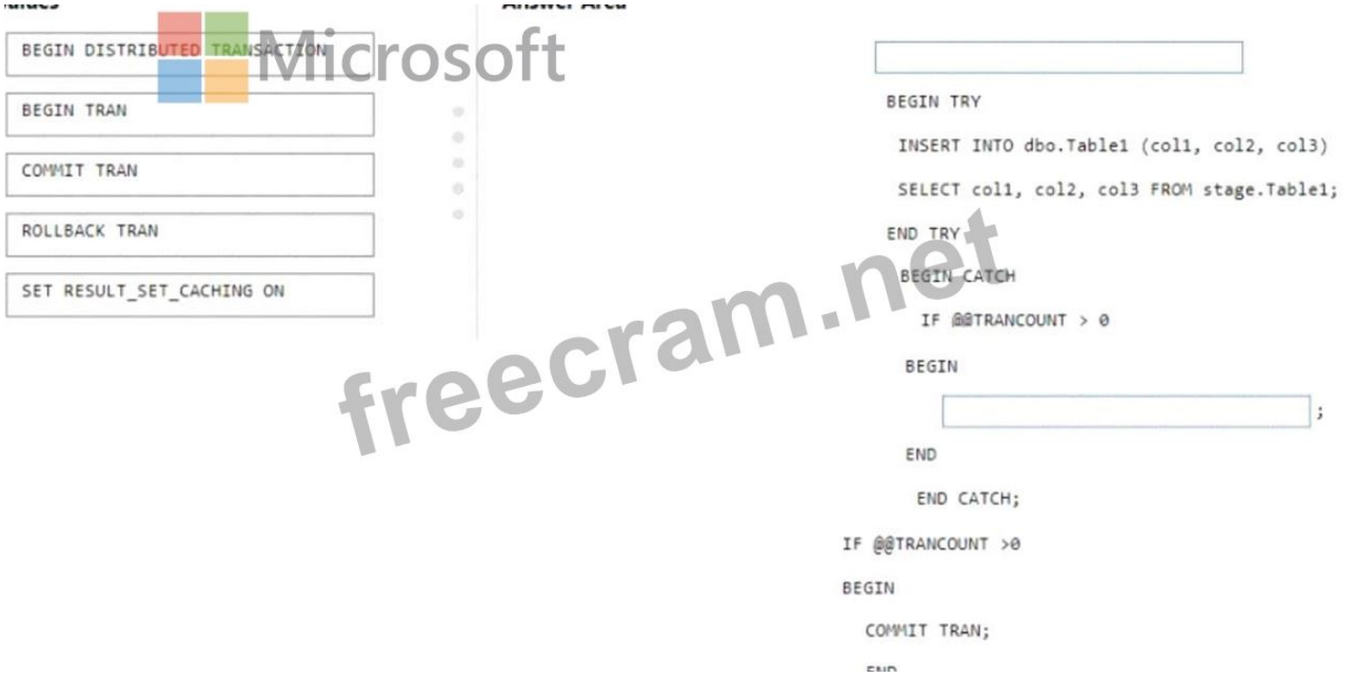
Explanation
Box 1: Hash
Consider using a hash-distributed table when:
The table size on disk is more than 2 GB.
The table has frequent insert, update, and delete operations.
Box 2: Clustered columnstore
Clustered columnstore tables offer both the highest level of data compression and the best overall 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: 23
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: (SHOW ANSWER)

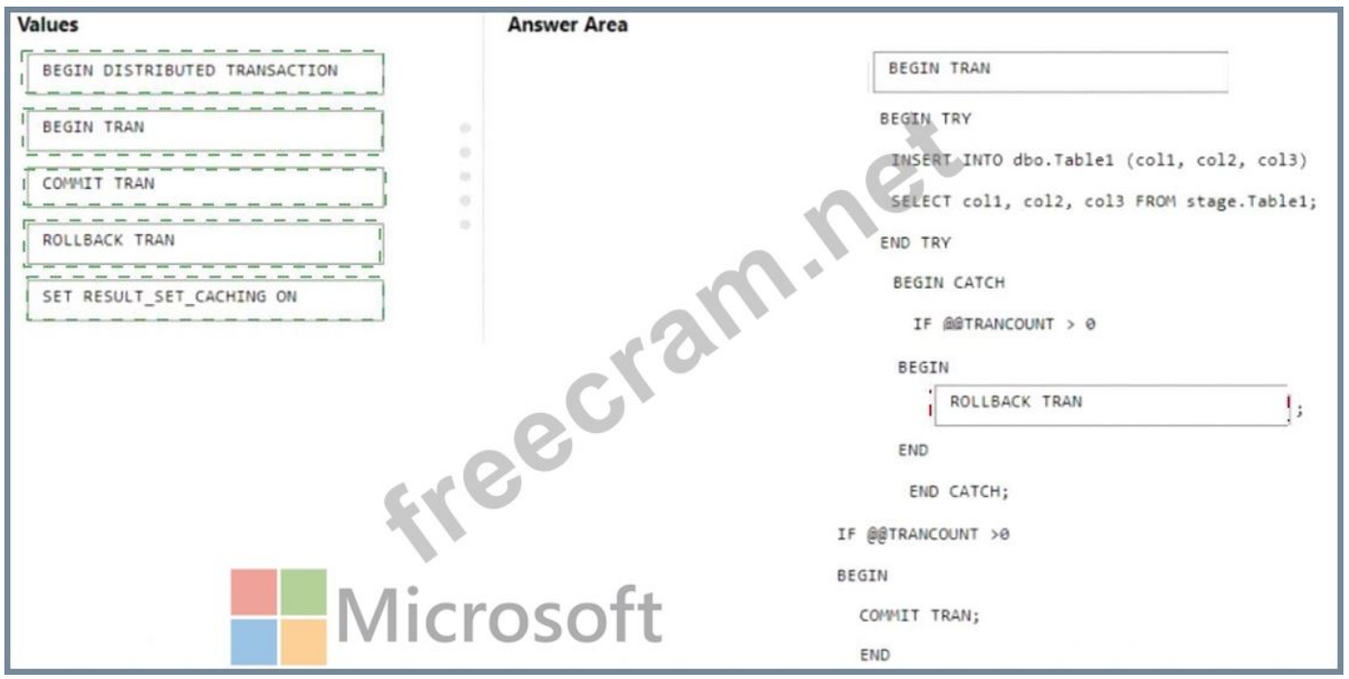
Explanation
An external table points to data located in Hadoop, Azure Storage blob, or Azure Data Lake Storage. External tables are used to read data from files or write data to files in Azure Storage. With Synapse SQL, you can use external tables to read external data using dedicated SQL pool or serverless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

NEW QUESTION: 24
You have an Azure Data Factory pipeline that performs an incremental load of source data to an Azure Data Lake Storage Gen2 account.
Data to be loaded is identified by a column named LastUpdatedDate in the source table.
You plan to execute the pipeline every four hours.
You need to ensure that the pipeline execution meets the following requirements:
* Automatically retries the execution when the pipeline run fails due to concurrency or throttling limits.
* Supports backfilling existing data in the table.
Which type of trigger should you use?
A. event
B. on-demand
C. schedule
D. tumbling window
Answer: D (LEAVE A REPLY)
Explanation
In case of pipeline failures, tumbling window trigger can retry the execution of the referenced pipeline automatically, using the same input parameters, without the user intervention. This can be specified using the property "retryPolicy" in the trigger definition.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-tumbling-window-trigger

NEW QUESTION: 25
You are batch loading a table in an Azure Synapse Analytics dedicated SQL pool.
You need to load data from a staging table to the target table. The solution must ensure that if an error occurs while loading the data to the target table, all the inserts in that batch are undone.
How should you complete the Transact-SQL code? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE Each correct selection is worth one point.

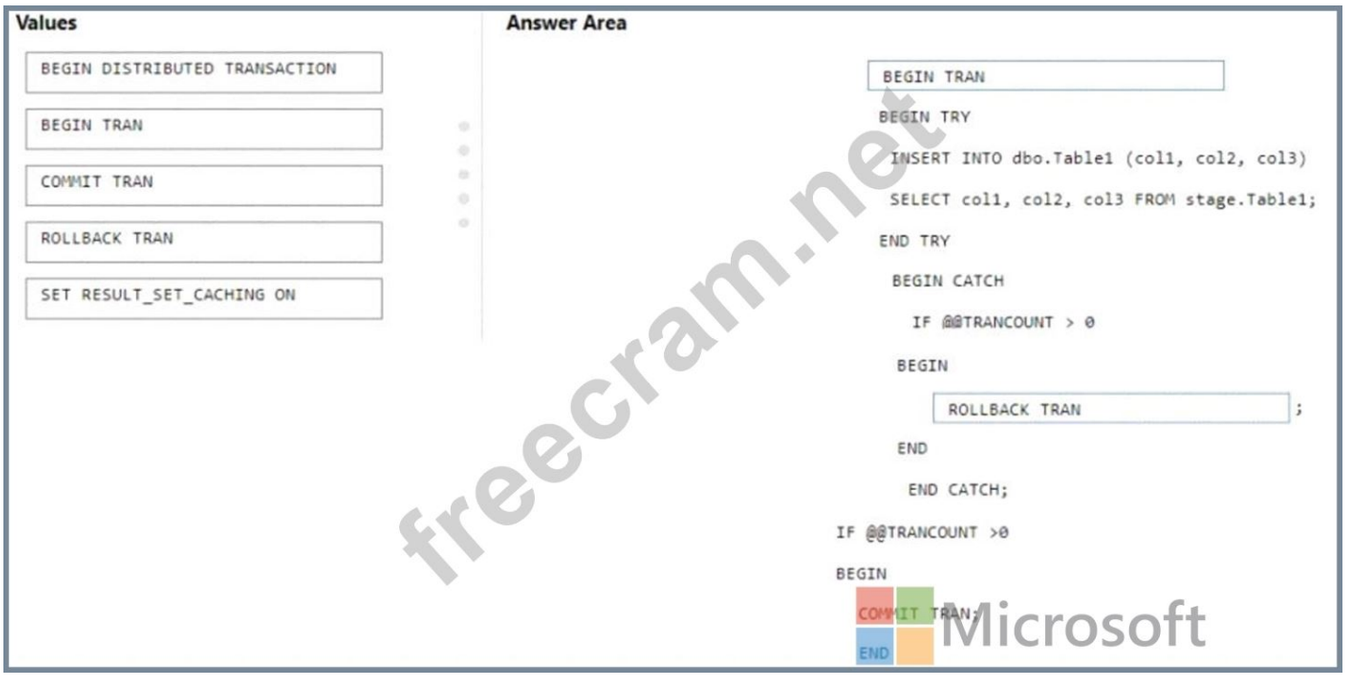


Answer:



Explanation

D:\mudassar\Untitled.jpg



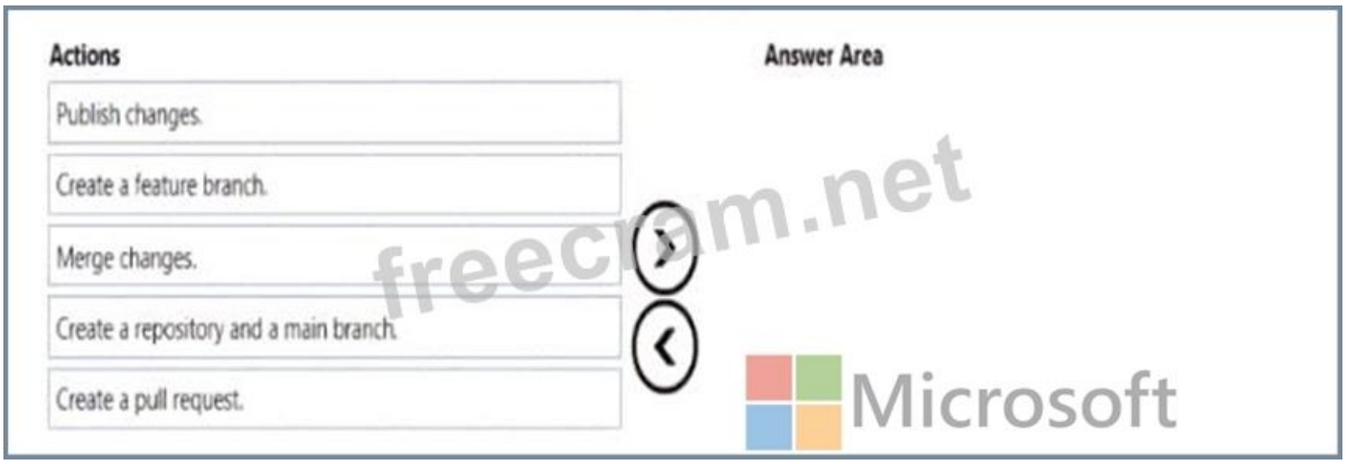
NEW QUESTION: 26
You are developing an application that uses Azure Data Lake Storage Gen 2.
You need to recommend a solution to grant permissions to a specific application for a limited time period.
What should you include in the recommendation?
A. Azure Active Directory (Azure AD) identities
B. shared 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: 27
You have an Azure Synapse Analytics dedicated SQL pool that contains a large fact table. The table contains
50 columns and 5 billion rows and is a heap.
Most queries against the table aggregate values from approximately 100 million rows and return only two columns.
You discover that the queries against the fact table are very slow.

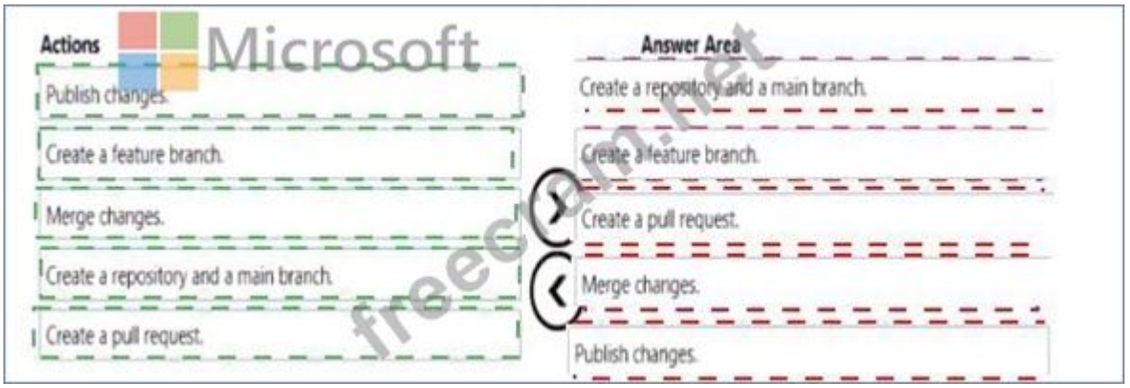
Which type of index should you add to provide the fastest query times?
A. nonclustered columnstore
B. clustered columnstore
C. nonclustered
D. clustered
Answer: (SHOW ANSWER)
Explanation
Clustered columnstore indexes are one of the most efficient ways you can store your data in dedicated SQL pool.
Columnstore tables won't benefit a query unless the table has more than 60 million rows.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

NEW QUESTION: 28
You are designing an Azure Synapse Analytics dedicated SQL pool.
You need to ensure that you can audit access to Personally Identifiable information (PII).
What should you include in the solution?
A. dynamic data masking
B. row-level security (RLS)
C. sensitivity classifications
D. column-level security
Answer: (SHOW ANSWER)
Explanation
Data Discovery & Classification is built into Azure SQL Database, Azure SQL Managed Instance, and Azure Synapse Analytics. It provides basic capabilities for discovering, classifying, labeling, and reporting the sensitive data in your databases.
Your most sensitive data might include business, financial, healthcare, or personal information. Discovering and classifying this data can play a pivotal role in your organization's information-protection approach. It can serve as infrastructure for:
* Helping to meet standards for data privacy and requirements for regulatory compliance.
* Various security scenarios, such as monitoring (auditing) access to sensitive data.
* Controlling access to and hardening the security of databases that contain highly sensitive data.
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview

NEW QUESTION: 29
You need to implement versioned changes to the integration pipelines. The solution must meet the data integration requirements.
In which order should you perform the actions? To answer, move all actions from the list of actions to the answer area and arrange them in the correct order.

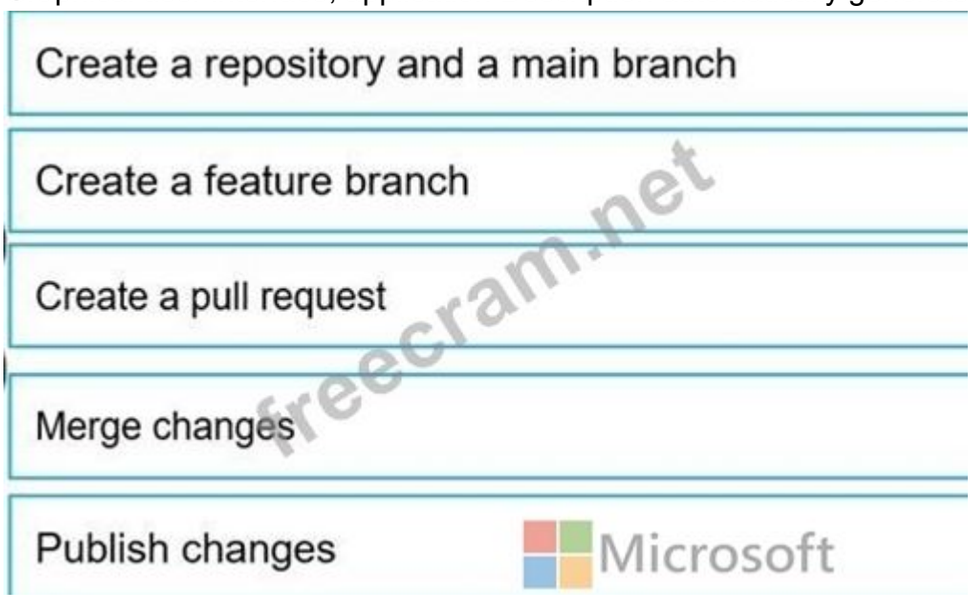


Answer:



Explanation

Graphical user interface, application Description automatically generated



Scenario: Identify a process to ensure that changes to the ingestion and transformation activities can be version-controlled and developed independently by multiple data engineers.

Step 1: Create a repository and a main branch

You need a Git repository in Azure Pipelines, TFS, or GitHub with your app.

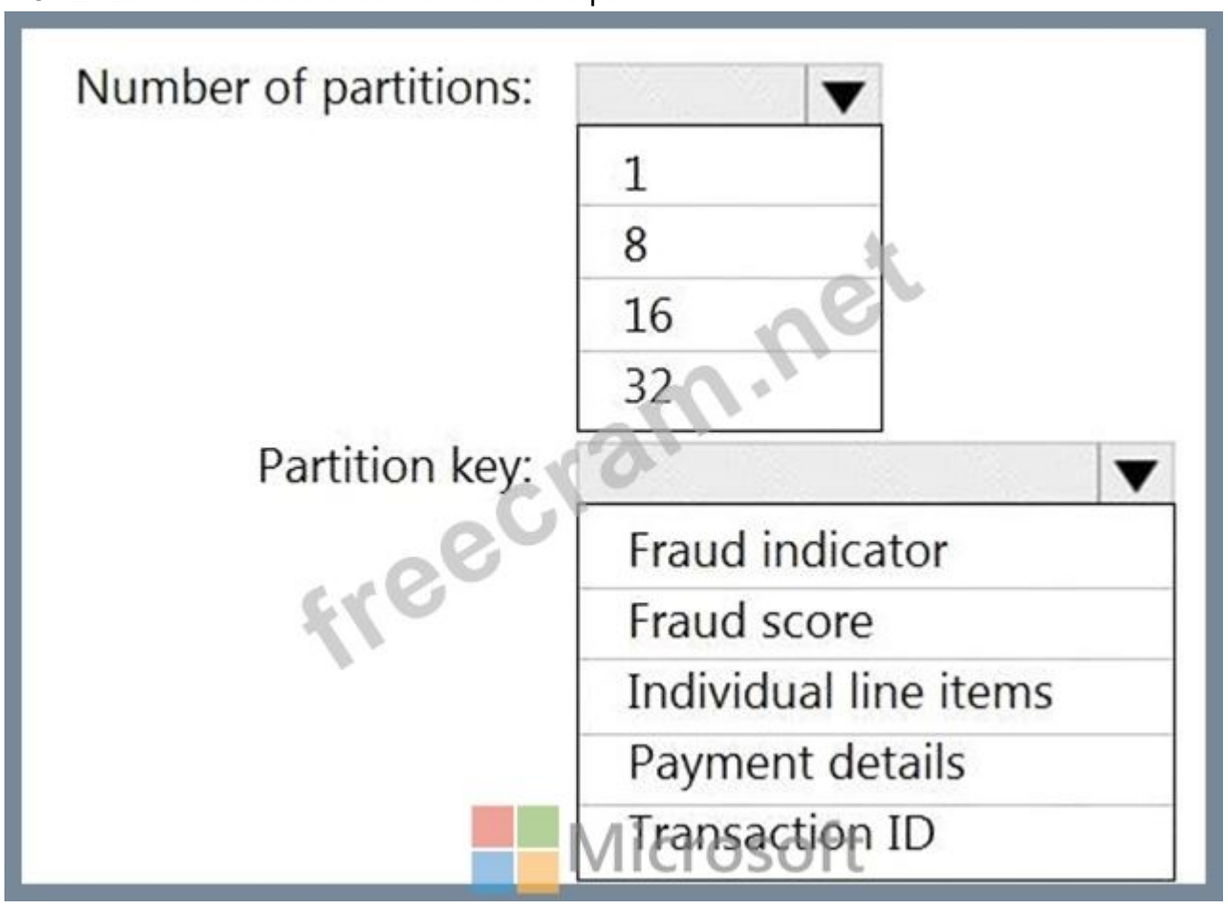
Step 2: Create a feature branch

Step 3: Create a pull request

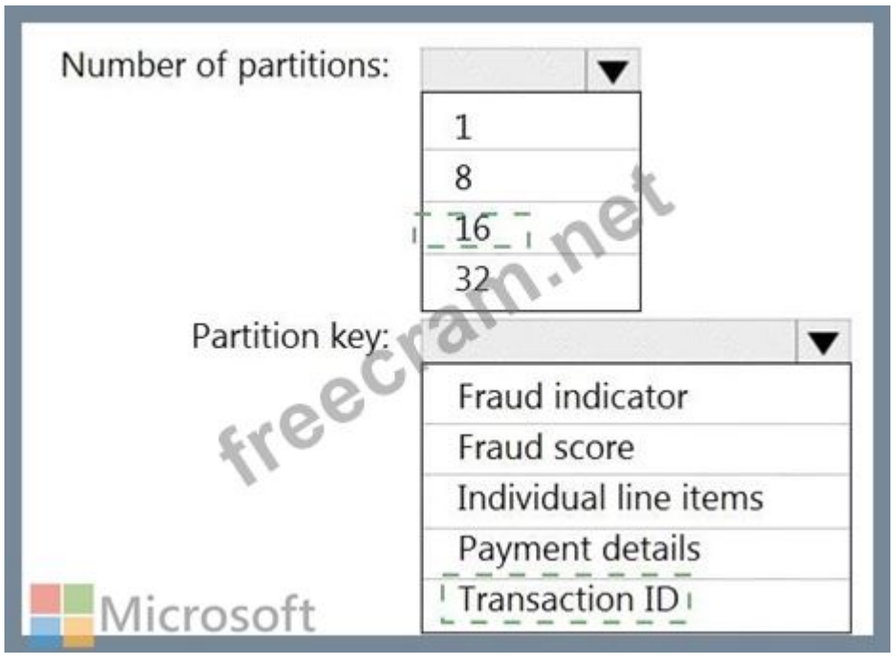
Step 4: Merge changes

Merge feature branches into the main branch using pull requests.
Step 5: Publish changes
Reference:
https://docs.microsoft.com/en-us/azure/devops/pipelines/repos/pipeline-options-for-git

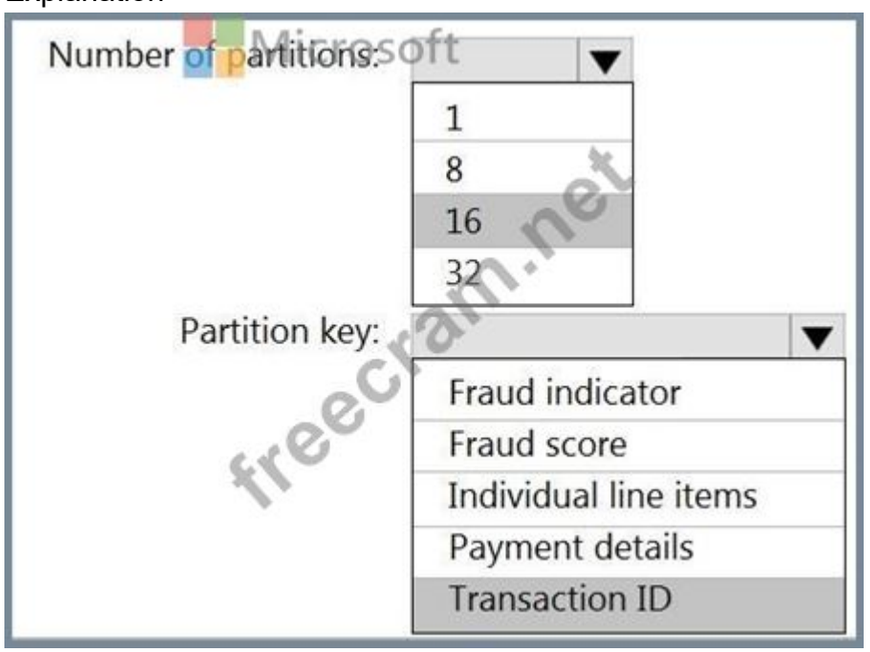
NEW QUESTION: 30
You have an Azure event hub named retailhub that has 16 partitions. Transactions are posted to retailhub. Each transaction includes the transaction ID, the individual line items, and the payment details. The transaction ID is used as the partition key.
You are designing an Azure Stream 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

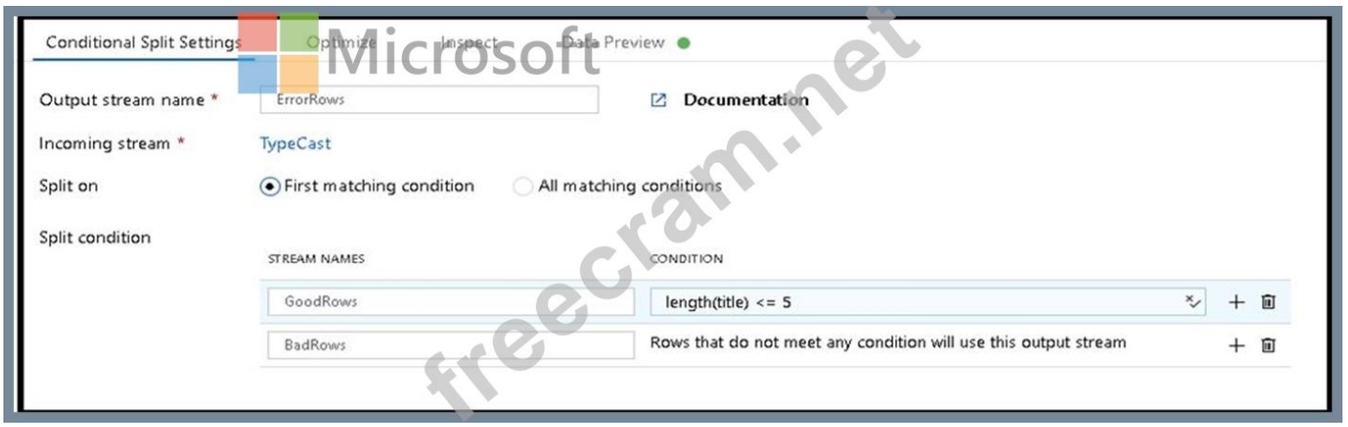
Reference:

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

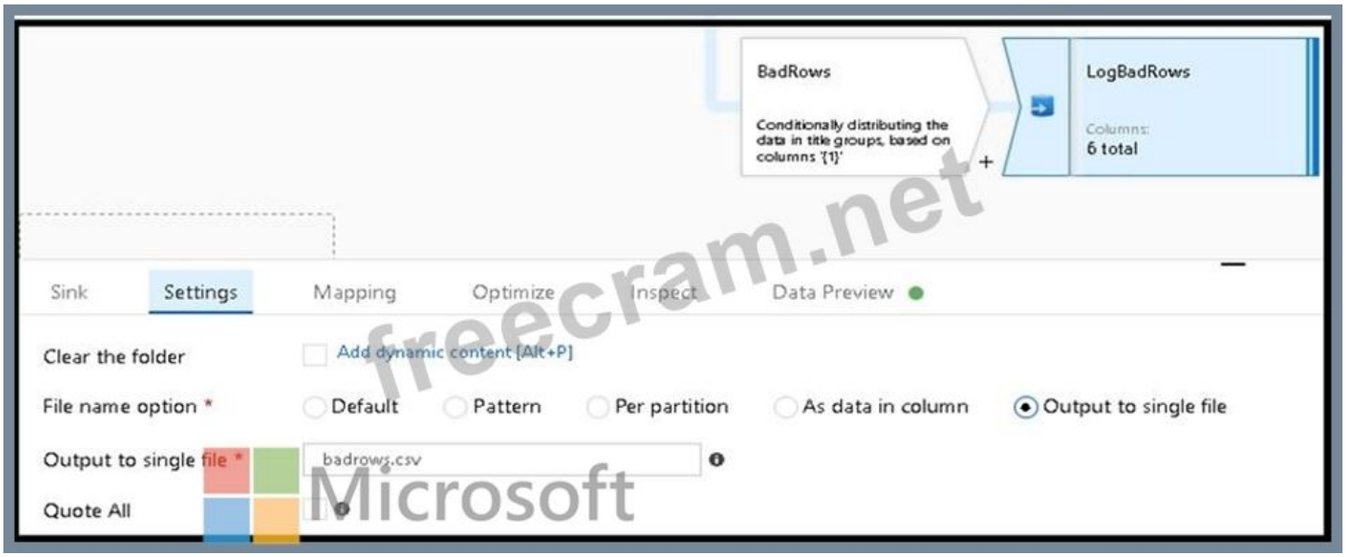
NEW QUESTION: 31

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

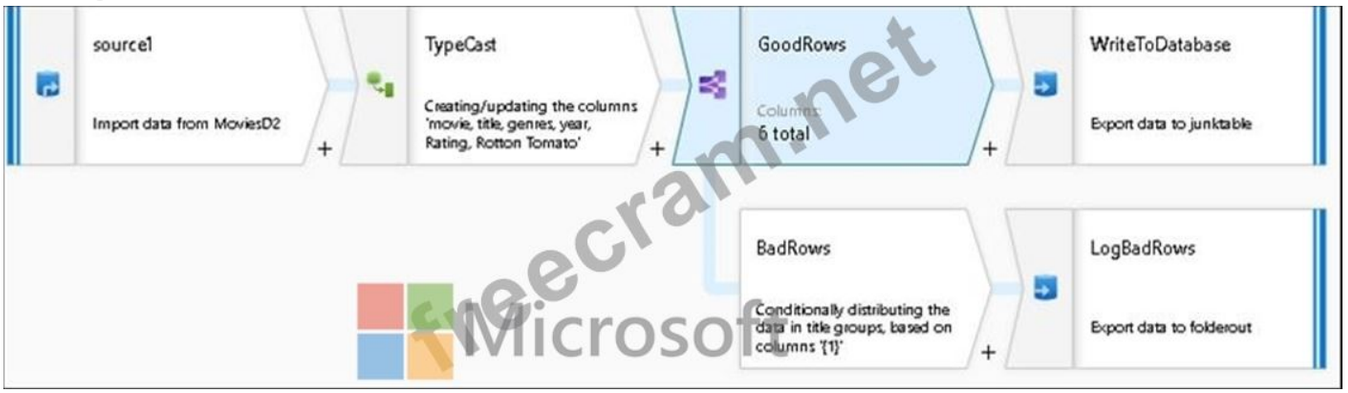
The data flow already contains the following:
A source transformation.
A Derived Column transformation to set the appropriate types of data.
A sink transformation to land the data in the pool.
You need to ensure that the data flow meets the following requirements:
All valid rows must be written to the destination table.
Truncation errors in the comment column must be avoided proactively.
Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
A. To the data flow, add a sink transformation to write the rows to a file in blob storage.
B. To the data flow, add a Conditional Split transformation to separate the rows that will cause truncation errors.
C. To the data flow, add a filter transformation to filter out rows that will cause truncation errors.
D. Add a select transformation to select only the rows that will cause truncation errors.
Answer: (SHOW ANSWER)
Explanation
B: Example:
1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.



2. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.
A:
3. Now we need to log the rows that failed. Add a sink transformation to the BadRows stream for logging. Here, we'll "auto-map" all of the fields so that we have logging of the complete transaction record. This is a text-delimited CSV file output to a single file in Blob Storage. We'll call the log file "badrows.csv".



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



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

NEW QUESTION: 32
You plan to create a dimension table in Azure Synapse Analytics that will be less than 1 GB.
You need to create the table to meet the following requirements:
* Provide the fastest Query time.
* Minimize data movement during queries.
Which type of table should you use?
A. hash distributed

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

NEW QUESTION: 33
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. `\YYYY\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: (SHOW ANSWER)
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: 34
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: 35
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

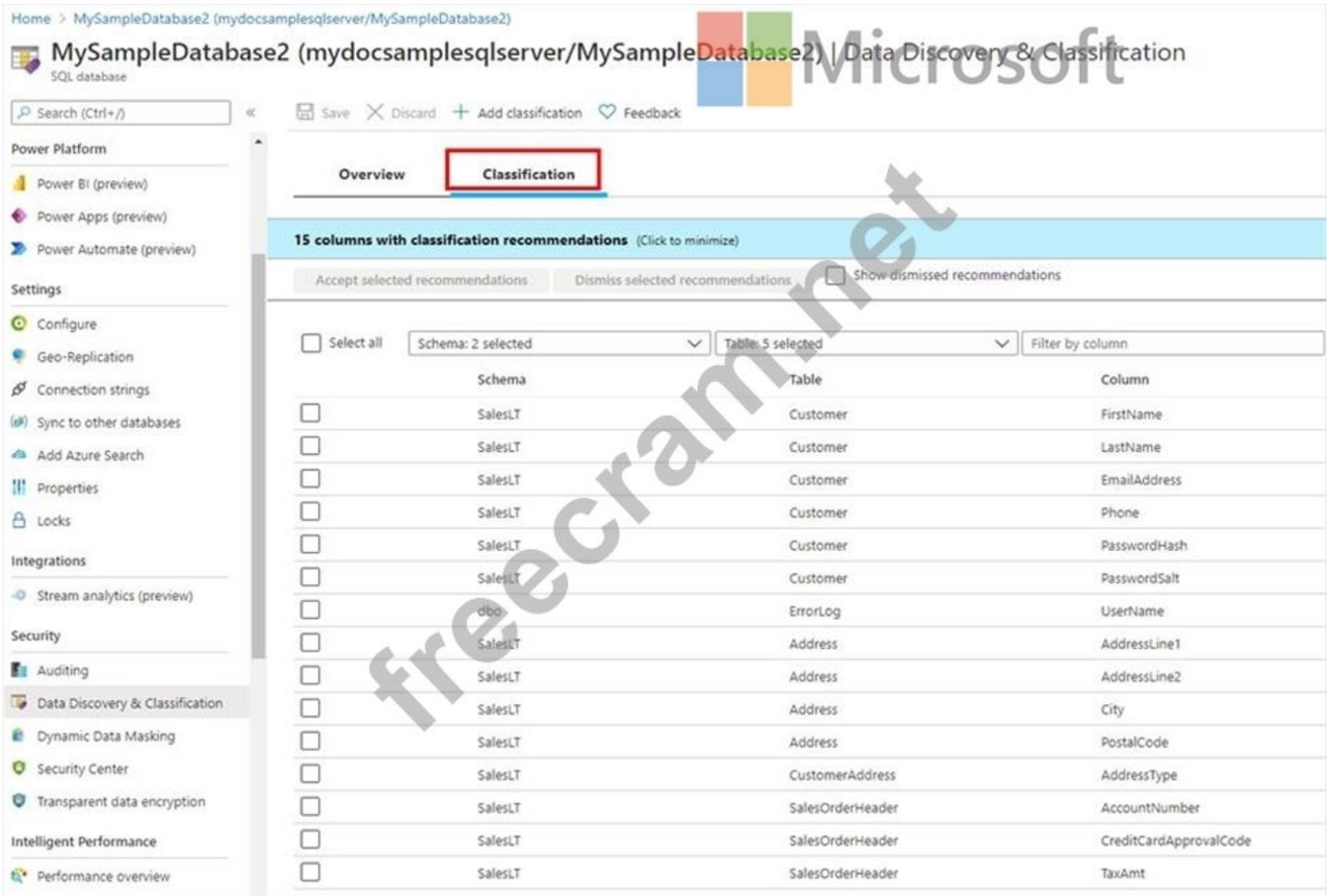
C. audit logs sent to a Log Analytics workspace

D. dynamic data masking for columns that contain confidential information

Answer: A,C (LEAVE A REPLY)

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: 36

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: 37

You are planning a streaming data solution that will use Azure Databricks. The solution will stream sales transaction data from an online store. The solution has the following specifications:

* The output data will contain items purchased, quantity, line total sales amount, and line total tax amount.
* Line total sales amount and line total tax amount will be aggregated in Databricks.
* Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.

You need to recommend an output mode for the dataset that will be processed by using Structured Streaming.

The solution must minimize duplicate data.

What should you recommend?

A. Append

B. Update

C. Complete

Answer: (SHOW ANSWER)

Explanation

By default, streams run in append mode, which adds new records to the table.

https://docs.databricks.com/delta/delta-streaming.html

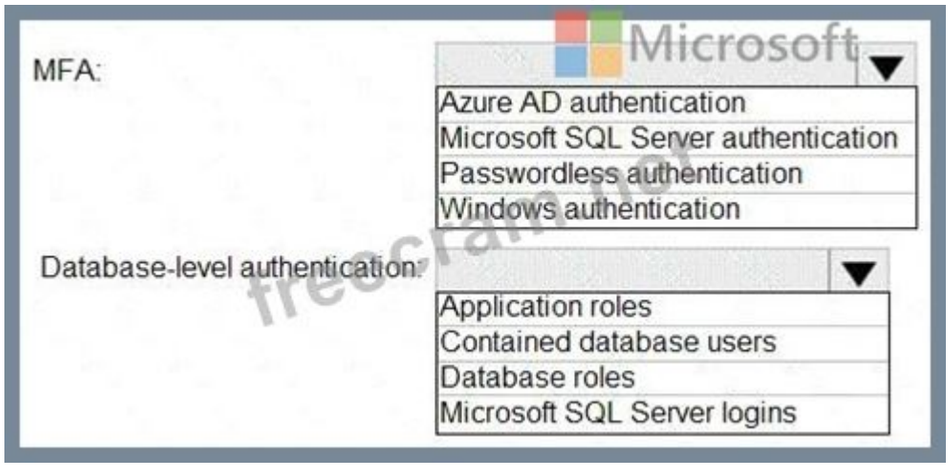
NEW QUESTION: 38

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

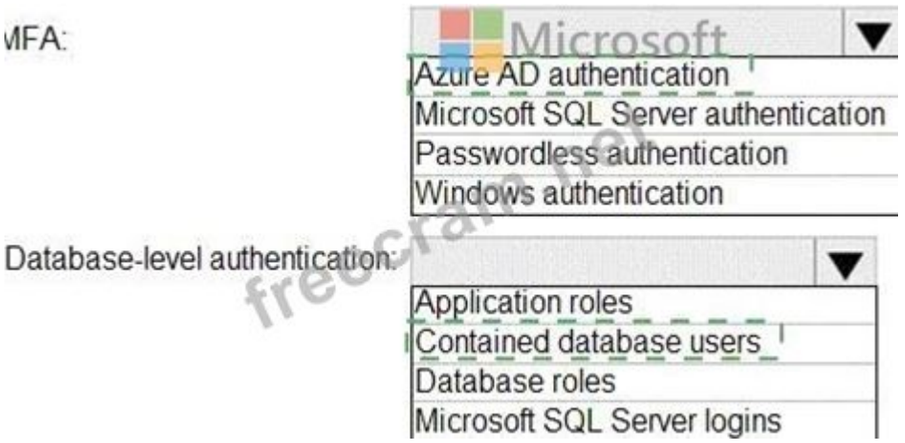
You need to recommend an authentication solution for Pool1. The solution must support multi-factor authentication (MFA) and database-level authentication.

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

NOTE: Each correct selection is worth one point.

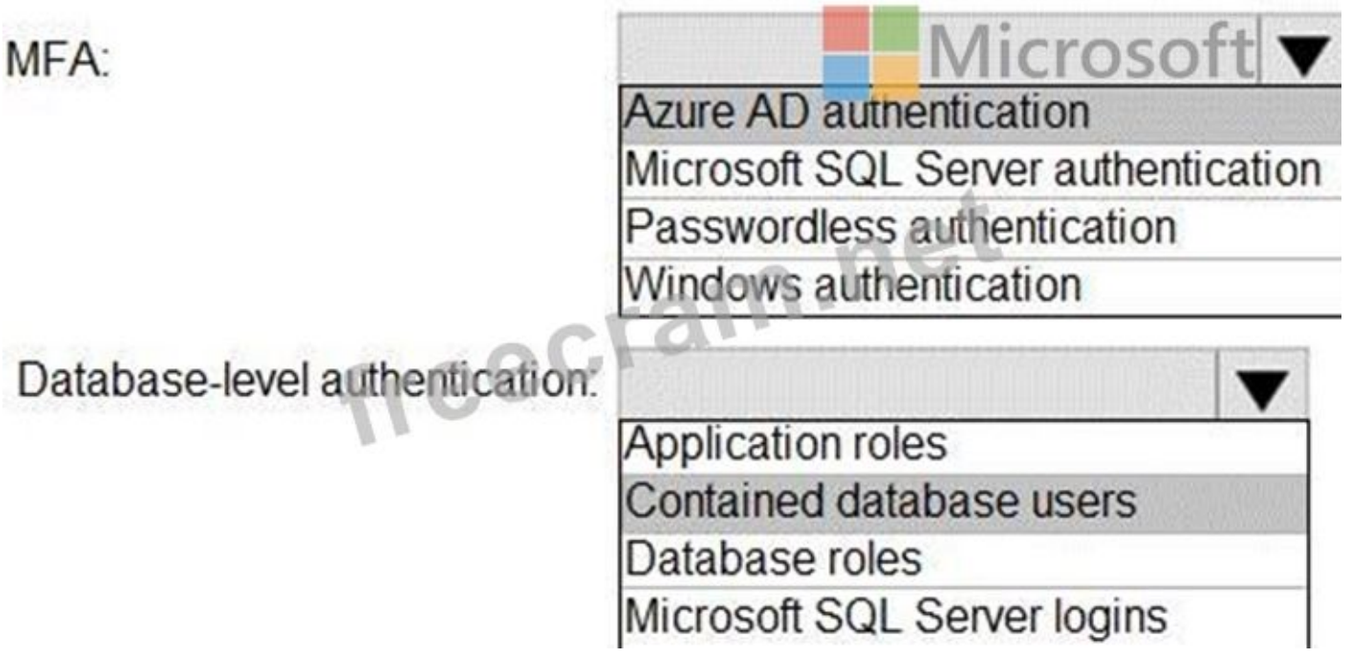


Answer:



Explanation

Graphical user interface, text, application, chat or text message Description automatically generated



Box 1: Azure AD authentication

Azure Active Directory authentication supports Multi-Factor authentication through Active Directory Universal Authentication.

Box 2: Contained database users

Azure Active Directory Uses contained database users to authenticate identities at the database level.

Reference:

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

NEW QUESTION: 39

You have an Azure Databricks resource.

You need to log actions that relate to changes in compute for the Databricks resource.

Which Databricks services should you log?

A. clusters

B. workspace

C. DBFS

D. SSH

E lobs

Answer: (SHOW ANSWER)

Explanation

Cloud Provider Infrastructure Logs.Databricks logging allows security and admin teams to demonstrate conformance to data governance standards within or from a Databricks workspace. Customers, especially in the regulated industries, also need records on activities like:- User access control to cloud data storage- Cloud Identity and Access Management roles- User access to cloud network and compute Azure Databricks offers three distinct workloads on several VM Instances tailored for your data analytics workflow-the Jobs Compute and Jobs Light Compute workloads make it easy for data engineers to build

and execute jobs, and the All-Purpose Compute workload makes it easy for data scientists to explore, visualize, manipulate, and share data and insights interactively.

NEW QUESTION: 40

You have an Azure data factory.
You need to examine the pipeline failures from the last 60 days.
What should you use?

A. the Activity log blade for the Data Factory resource
B. the Monitor & Manage app in Data Factory
C. the Resource health blade for the Data Factory resource
D. Azure Monitor
Answer: (SHOW ANSWER)
Explanation
Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor

NEW QUESTION: 41
You have an Azure Synapse Analystics 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: (SHOW ANSWER)
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: 42
You are designing database for an Azure Synapse Analytics dedicated SQL pool to support workloads for detecting ecommerce transaction fraud.

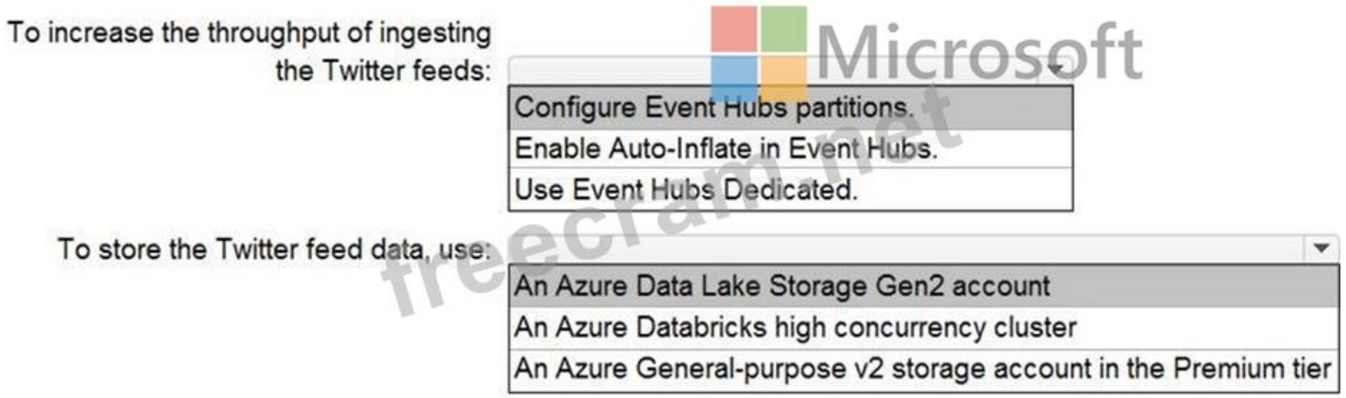
Data will be combined from multiple ecommerce sites and can include sensitive financial information such as credit card numbers.
You need to recommend a solution that meets the following requirements:
* Users must be able to identify potentially fraudulent transactions.
* Users must be able to use credit cards as a potential feature in models.
* Users must NOT be able to access the actual credit card numbers.
What should you include in the recommendation?
A. Transparent Data Encryption (TDE)
B. row-level security (RLS)
C. column-level encryption
D. Azure Active Directory (Azure AD) pass-through authentication
Answer: (SHOW ANSWER)
Explanation
Use Always Encrypted to secure the required columns. You can configure Always Encrypted for individual database columns containing your sensitive data. Always Encrypted is a feature designed to protect sensitive data, such as credit card numbers or national identification numbers (for example, U.S. social security numbers), stored in Azure SQL Database or SQL Server databases.
Reference:
https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/always-encrypted-database-engine

NEW QUESTION: 43
You need to design a data ingestion and storage solution for the Twitter feeds. The solution must meet the customer sentiment analytics requirements.
What should you include in the solution? To answer, select the appropriate options in the answer area
NOTE: Each correct selection b worth one point.

Answer:

Explanation

Graphical user interface, text Description automatically generated



Box 1: Configure Evegent Hubs partitions

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

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

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

Box 2: An Azure Data Lake Storage Gen2 account

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

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

Reference:

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

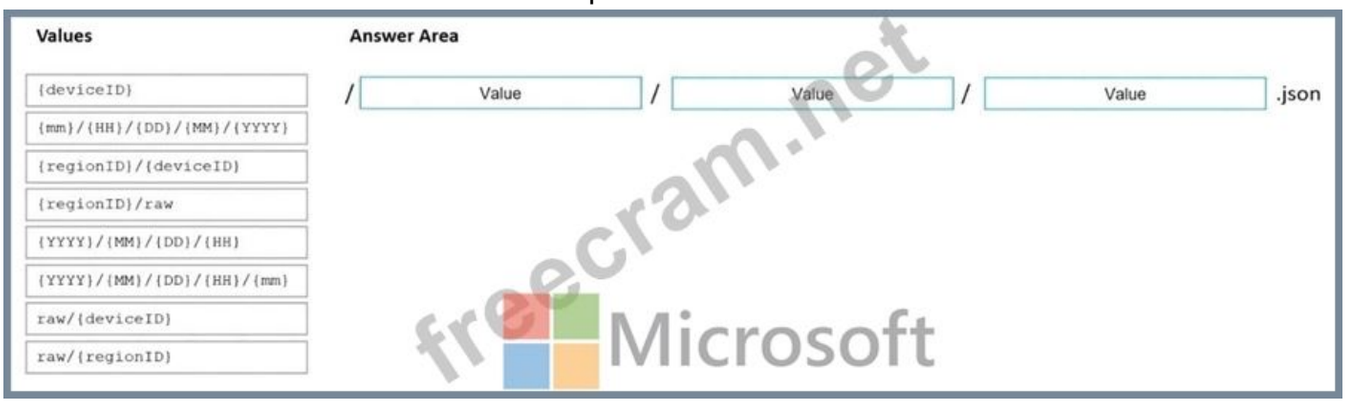
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control

NEW QUESTION: 44

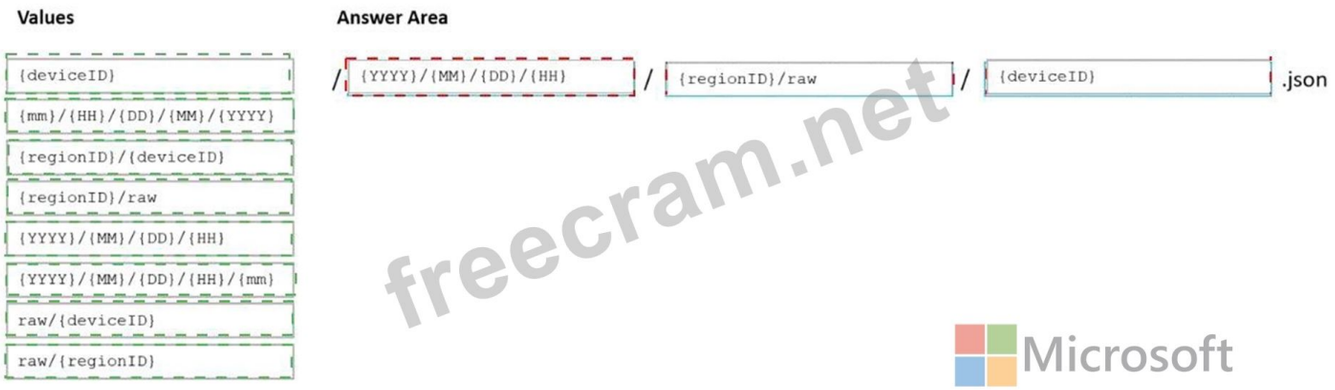
You are designing an Azure Data Lake Storage Gen2 structure for telemetry data from 25 million devices distributed across seven key geographical regions. Each minute, the devices will send a JSON payload of metrics to Azure Event Hubs.

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: 45

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

Ingest:

Access multiple data sources.

Provide the ability to orchestrate workflow.

Provide the capability to run SQL Server Integration Services packages.

Store:

Optimize storage for big data workloads.

Provide encryption of data at rest.

Operate with no size limits.

Prepare and Train:

Provide a fully-managed and interactive workspace for exploration and visualization.

Provide the ability to program in R, SQL, Python, Scala, and Java.

Provide seamless user authentication with Azure Active Directory.

Model & Serve:

Implement native columnar storage.

Support for the SQL language

Provide support for structured streaming.

You need to build the data integration pipeline.

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

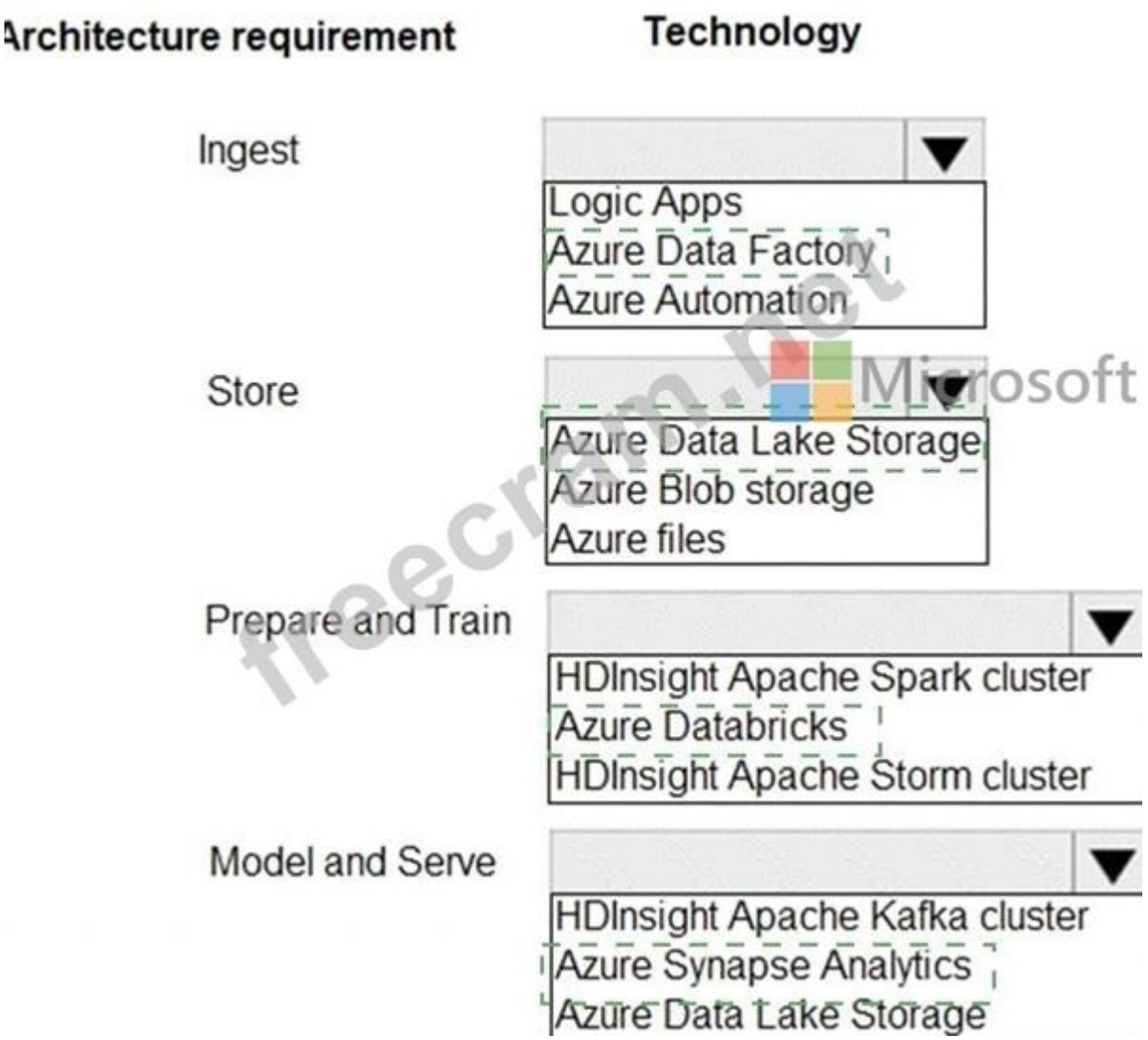
NOTE: Each correct selection is worth one point.

Answer Area

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

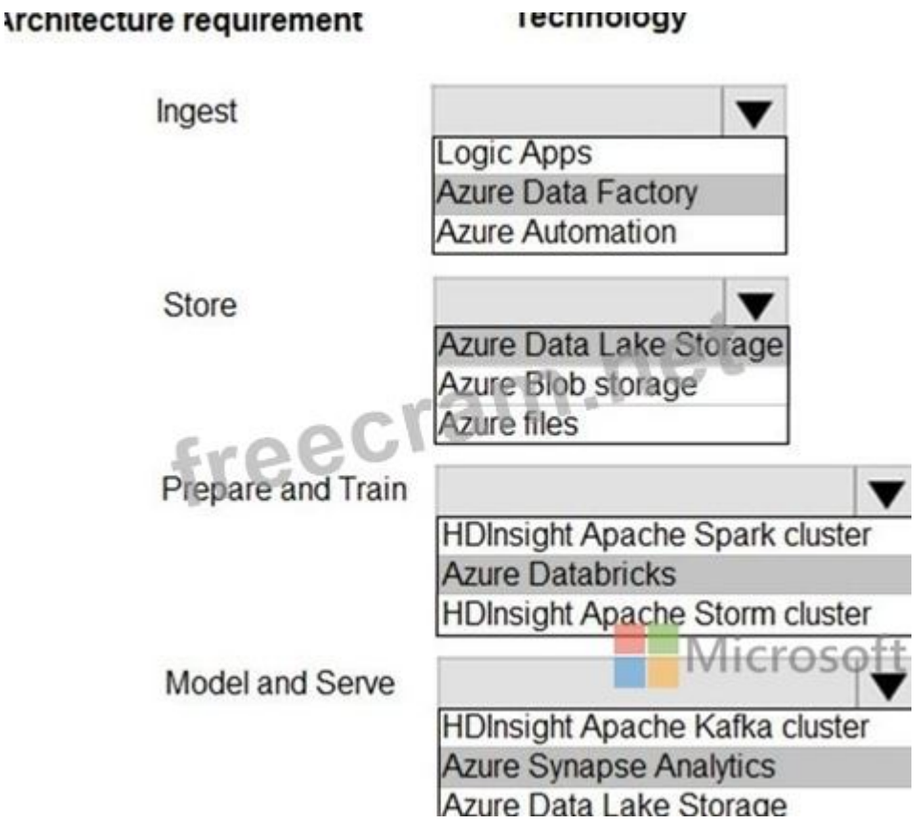
Answer:

Answer Area



Explanation

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



NEW QUESTION: 46
You are designing a dimension table in an Azure Synapse Analytics dedicated SQL pool.
You need to create a surrogate key for the table. The solution must provide the fastest query performance.
What should you use for the surrogate key?
A. a GUID column
B. a sequence object
C. an IDENTITY column
Answer: (SHOW ANSWER)
Explanation
Use IDENTITY to create surrogate keys using dedicated SQL pool in AzureSynapse Analytics.
Note: A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identity

NEW QUESTION: 47

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: (SHOW ANSWER)

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: 48

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 ArrivalDateTime
Weather	AirportID ReportDateTime

You need to recommend a solution that maximum query performance.

What should you include in the recommendation?

A. In the tables, use a hash distribution of ArriveDateTime and ReportDateTime.

B. In the tables, use a hash distribution of ArriveAirPortID 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)

NEW QUESTION: 49

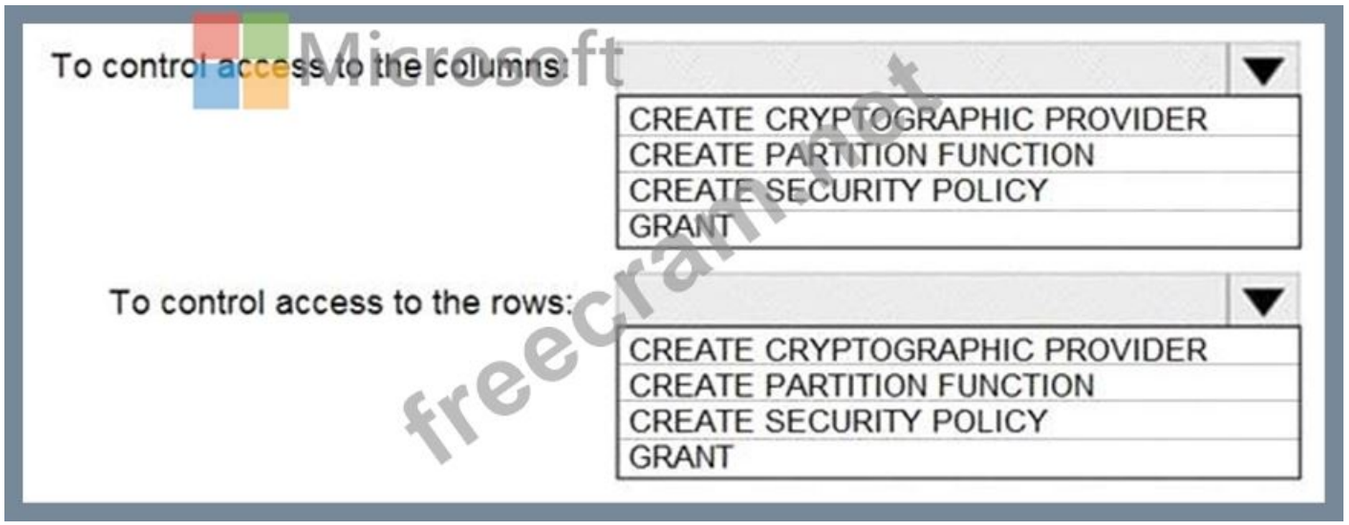
You are designing a slowly changing dimension (SCD) for supplier data in an Azure Synapse Analytics dedicated SQL pool.
You plan to keep a record of changes to the available fields.
The supplier data contains the following columns.

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

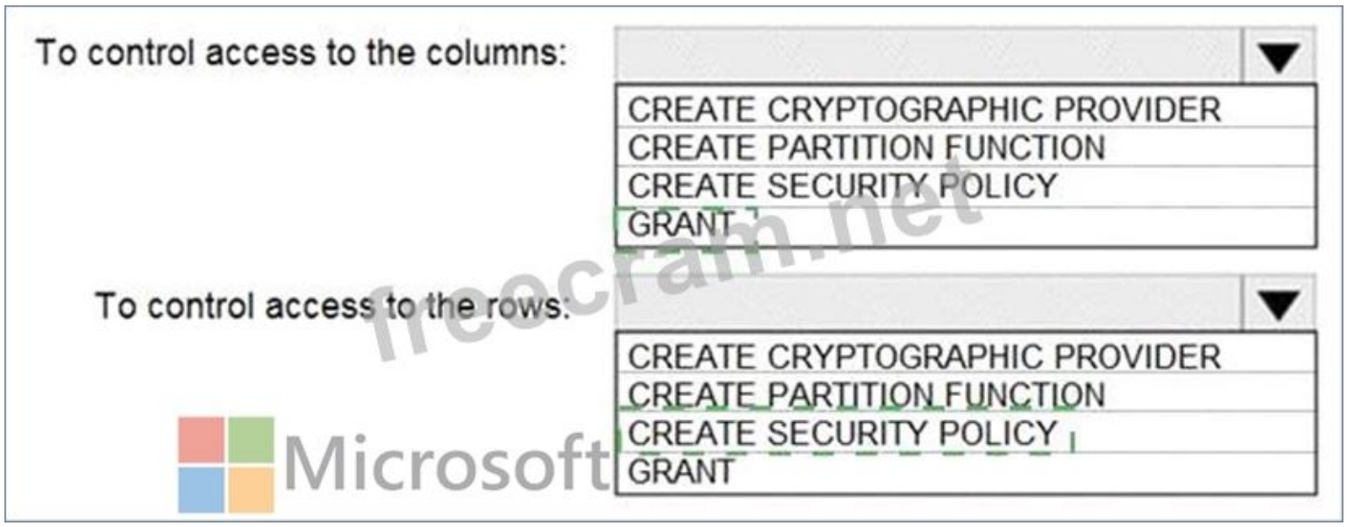
Which three additional columns should you add to the data to create a Type 2 SCD? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
A. surrogate primary key
B. foreign key
C. effective start date
D. effective end date
E. last modified date
F. business key
Answer: (SHOW ANSWER)
Reference:
https://docs.microsoft.com/en-us/sql/integration-services/data-flow/transformations/slowly-changing-dimension-

NEW QUESTION: 50
You have an Azure subscription that contains the following resources:
An Azure Active Directory (Azure AD) tenant that contains a security group named Group1 An Azure Synapse Analytics SQL pool named Pool1 You need to control the access of Group1 to specific columns and rows in a table in Pool1.

Which Transact-SQL commands should you use? To answer, select the appropriate options in the answer area.

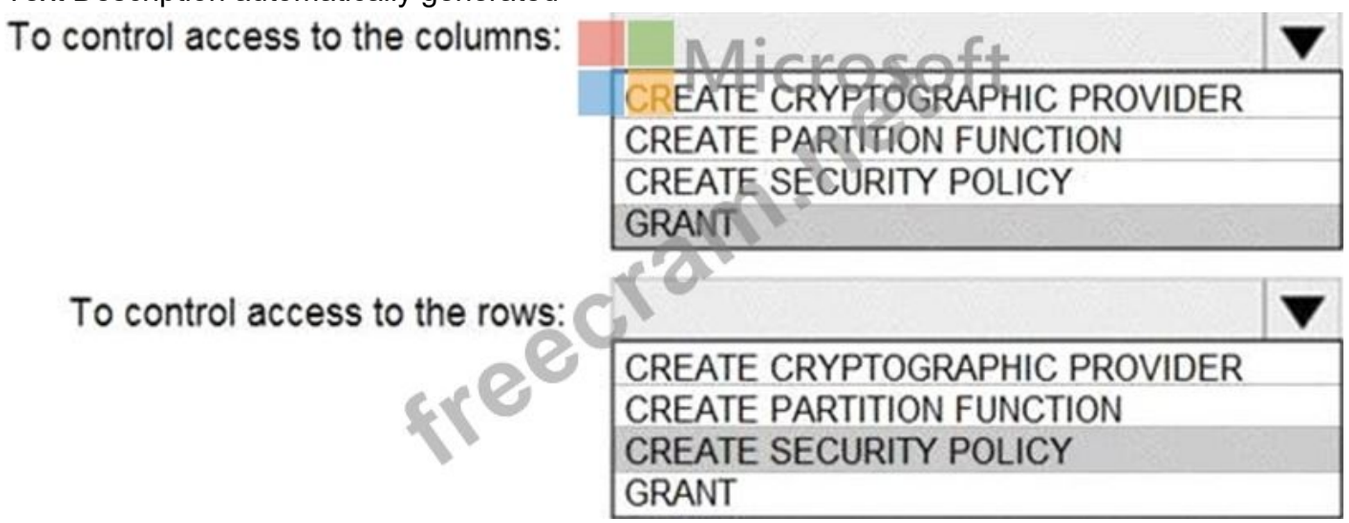


Answer:



Explanation

Text Description automatically generated

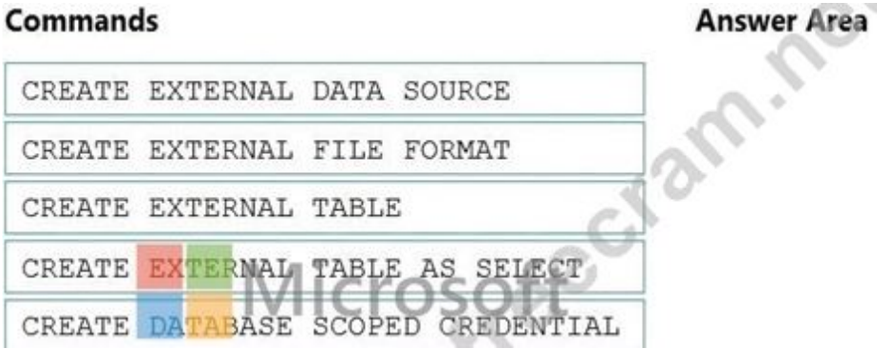


Box 1: GRANT

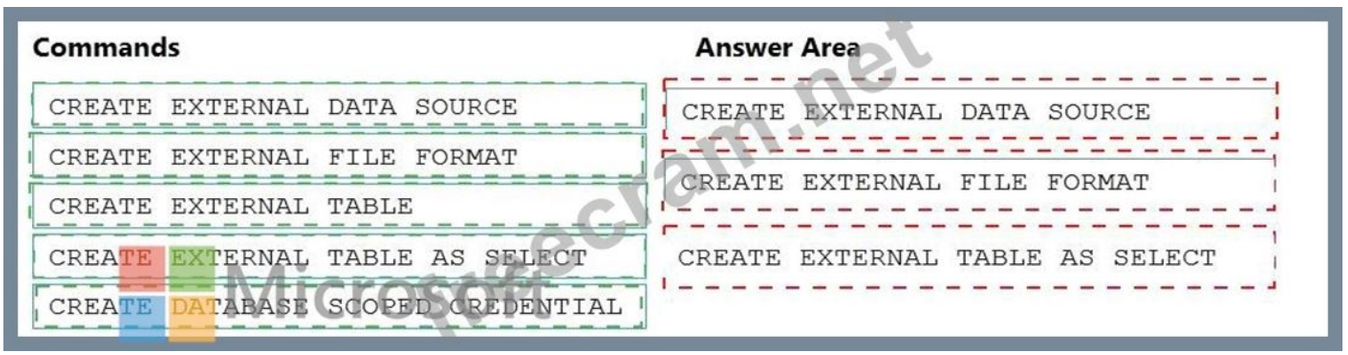
You can implement column-level security with the GRANT T-SQL statement.

Box 2: CREATE SECURITY POLICY
Implement Row Level Security by using the CREATE SECURITY POLICY Transact-SQL statement
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/column-level-security

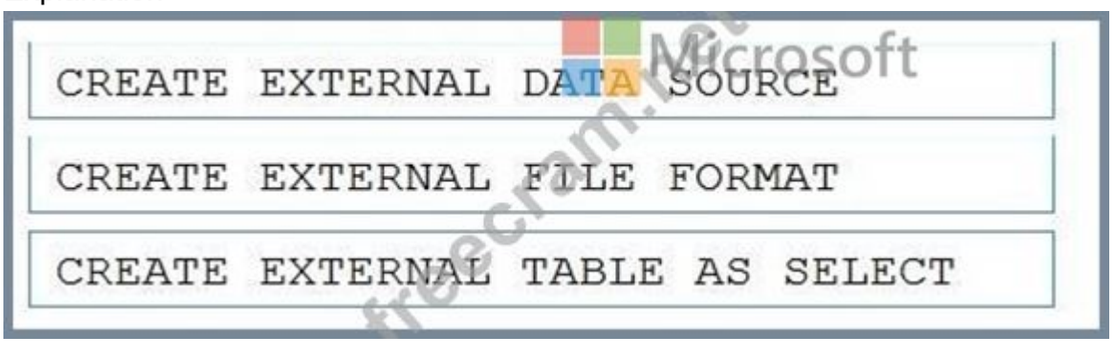
NEW QUESTION: 51
You need to ensure that the Twitter feed data can be analyzed in the dedicated SQL pool. The solution must meet the customer sentiment analytics requirements.
Which three Transaction-SQL DDL commands should you run in sequence? To answer, move the appropriate commands from the list of commands to the answer area and arrange them in the correct order.
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.



Answer:



Explanation



Scenario: Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds. Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.
Box 1: CREATE EXTERNAL DATA SOURCE

External data sources are used to connect to storage accounts.
Box 2: CREATE EXTERNAL FILE FORMAT
CREATE EXTERNAL FILE FORMAT creates an external file format object that defines external data stored in Azure Blob Storage or Azure Data Lake Storage. Creating an external file format is a prerequisite for creating an external table.
Box 3: CREATE EXTERNAL TABLE AS SELECT
When used in conjunction with the CREATE TABLE AS SELECT statement, selecting from an external table imports data into a table within the SQL pool. In addition to the COPY statement, external tables are useful for loading data.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

NEW QUESTION: 52
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

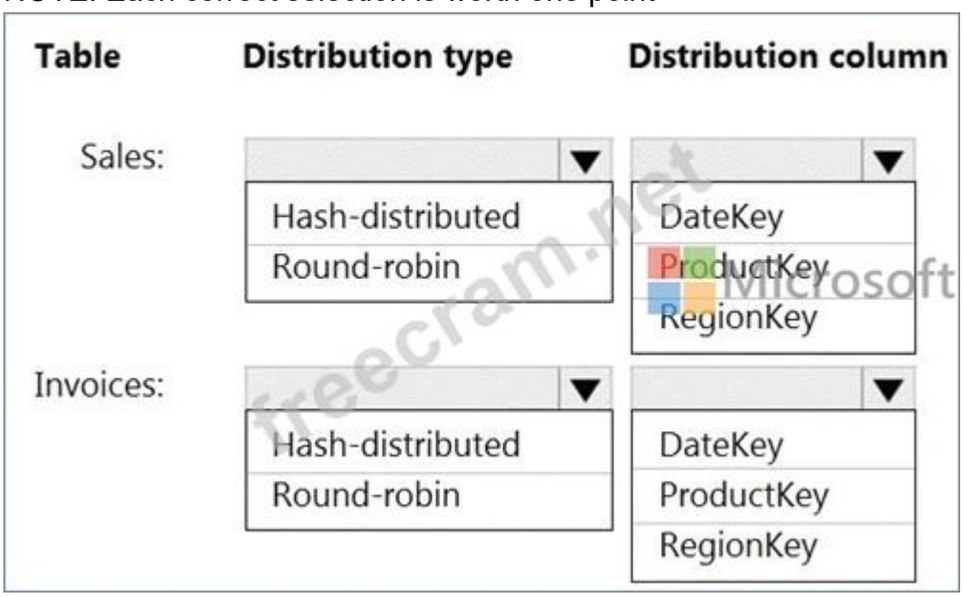
Box 1: LAG
The LAG analytic operator allows one to look up a "previous" event in an event stream, within certain constraints. It is very useful for computing the rate of growth of a variable, detecting when a variable crosses a threshold, or when a condition starts or stops being true.
Box 2: LIMIT DURATION

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

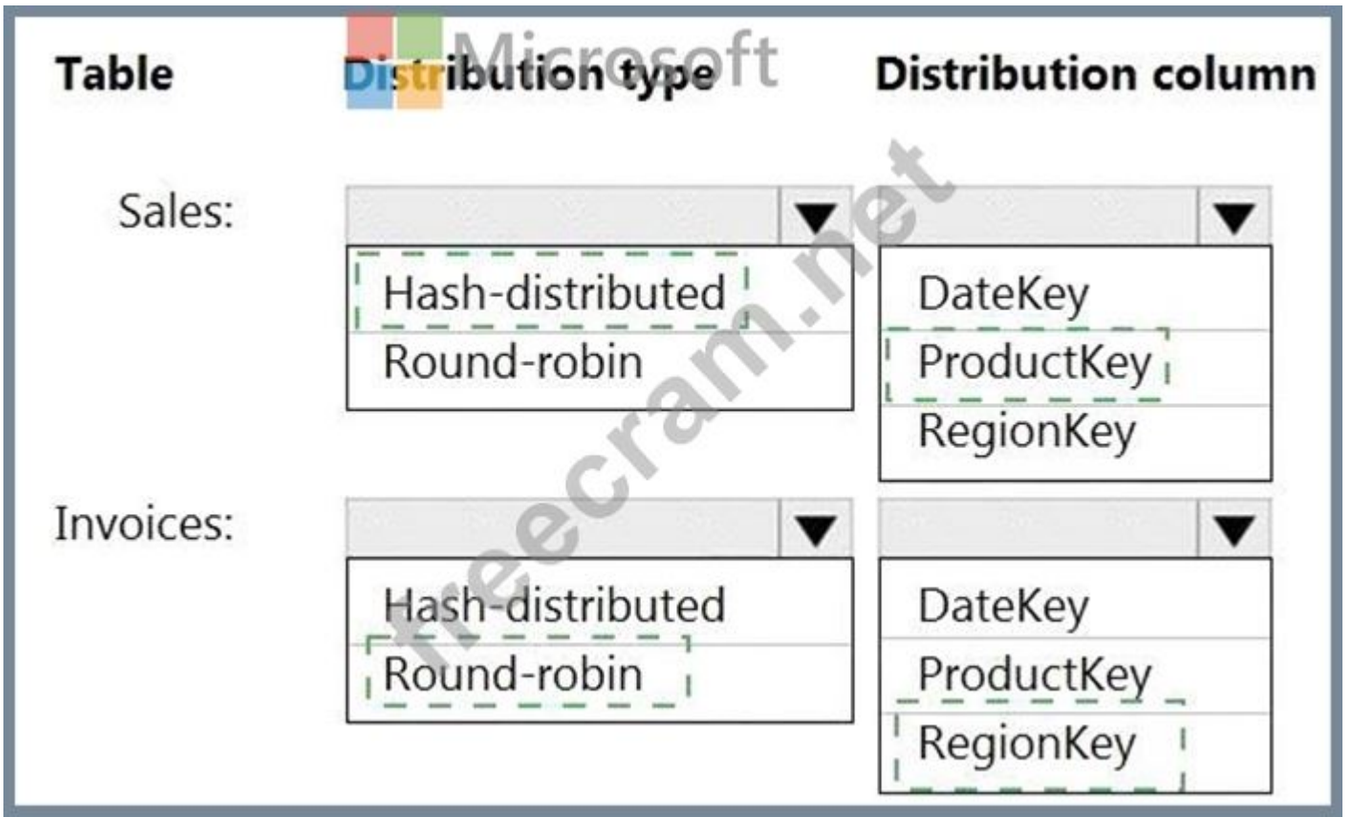
NEW QUESTION: 53
You have an on-premises data warehouse that includes the following fact tables. Both tables have the following columns: DateKey, ProductKey, RegionKey. There are 120 unique product keys and 65 unique region keys.

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

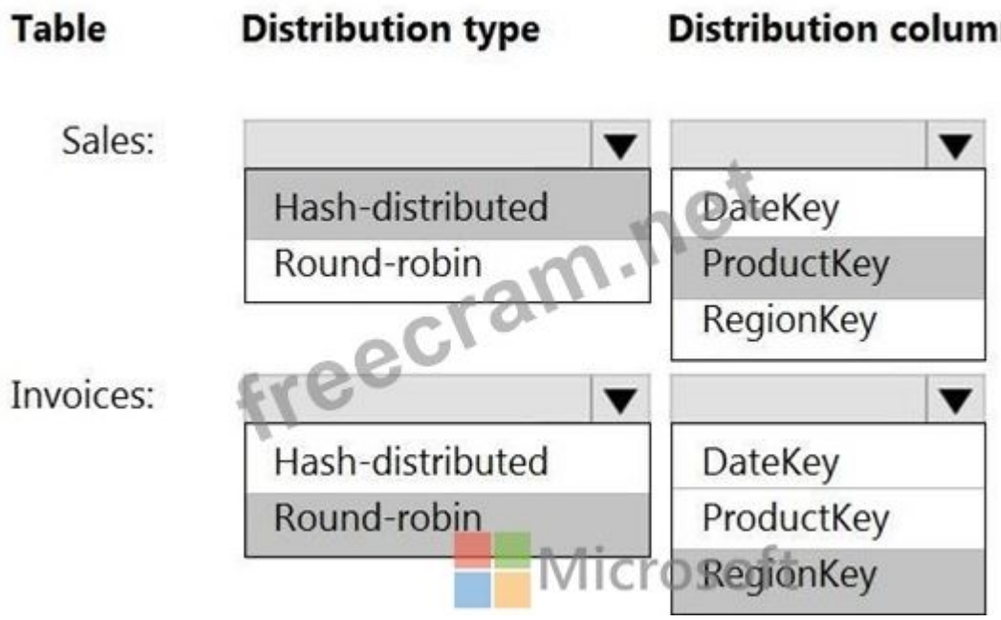
Queries that use the data warehouse take a long time to complete.
You plan to migrate the solution to use Azure Synapse Analytics. You need to ensure that the Azure-based solution optimizes query performance and minimizes processing skew.
What should you recommend? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point



Answer:



Explanation



Box 1: Hash-distributed

Box 2: ProductKey

ProductKey is used extensively in joins.

Hash-distributed tables improve query performance on large fact tables.

Box 3: Round-robin

Box 4: RegionKey

Round-robin tables are useful for improving loading speed.

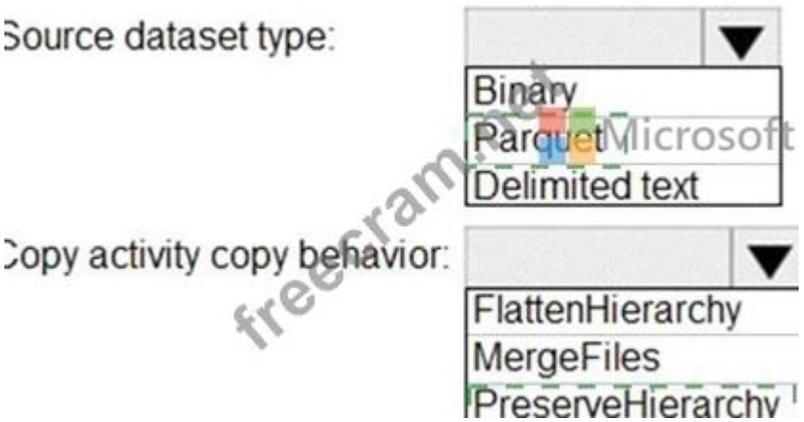
Consider using the round-robin distribution for your table in the following scenarios:

When getting started as a simple starting point since it is the default If there is no obvious joining key If there is not good candidate column for hash distributing the table If the table does not share a common join key with other tables If the join is less significant than other joins in the query When the table is a temporary staging table Note: A distributed table appears as a single table, but the rows are actually stored across 60 distributions. The rows are distributed with a hash or round-robin algorithm.
Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute

NEW QUESTION: 54
You plan to ingest streaming social media data by using Azure Stream Analytics. The data will be stored in files in Azure Data Lake Storage, and then consumed by using Azure Datiabricks and PolyBase in Azure Synapse Analytics.
You need to recommend a Stream Analytics data output format to ensure that the queries from Databricks and PolyBase against the files encounter the fewest possible errors. The solution must ensure that the tiles can be queried quickly and that the data type information is retained.
What should you recommend?
A. Parquet
B. Avro
C. CSV
D. JSON
Answer: (SHOW ANSWER)
Explanation
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-define-outputs

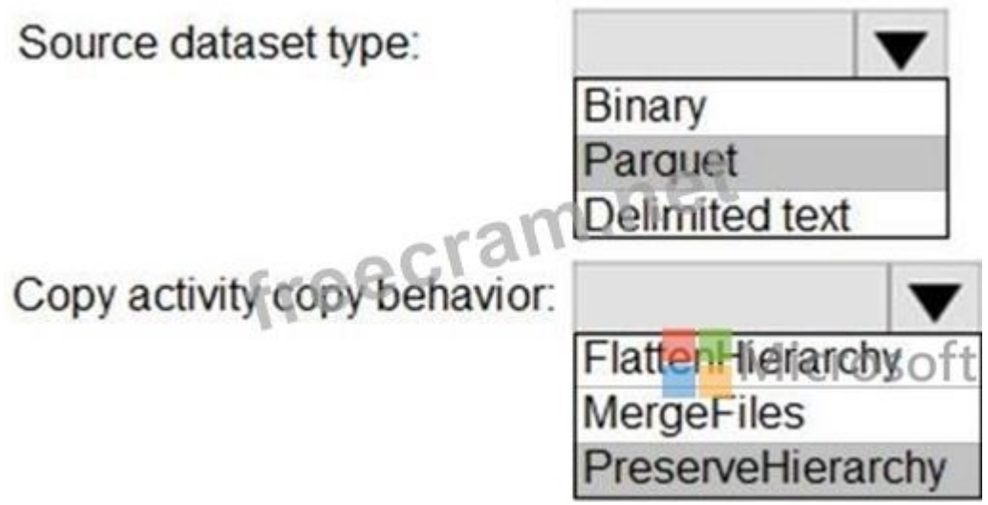
NEW QUESTION: 55
You have two Azure Storage accounts named Storage1 and Storage2. Each account holds one container and has the hierarchical namespace enabled. The system has files that contain data stored in the Apache Parquet format.
You need to copy folders and files from Storage1 to Storage2 by using a Data Factory copy activity. The solution must meet the following requirements:
No transformations must be performed.
The original folder structure must be retained.
Minimize time required to perform the copy activity.
How should you configure the copy activity? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer:



Explanation

Graphical user interface, text, application, chat or text message Description automatically generated



Box 1: Parquet

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

Box 2: PreserveHierarchy
PreserveHierarchy (default): Preserves the file hierarchy in the target folder. The relative path of the source file to the source folder is identical to the relative path of the target file to the target folder.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/format-parquet
https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage

NEW QUESTION: 56
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 High Concurrency cluster for the data engineers, and a High Concurrency cluster for the jobs.
Does this meet the goal?
A. Yes
B. No
Answer: (SHOW ANSWER)
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: 57
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: 58
You have an Azure Synapse Analytics dedicated SQL pool named pool1.
You need to perform a monthly audit of SQL statements that affect sensitive data. The solution must minimize administrative effort.
What should you include in the solution?
A. Microsoft Defender for SQL
B. dynamic data masking
C. workload management
D. sensitivity labels
Answer: (SHOW ANSWER)

NEW QUESTION: 59
You need to design an analytical storage solution for the transactional data. The solution must meet the sales transaction dataset requirements.

What should you include in the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer:

Explanation
Graphical user interface, text, application, table Description automatically generated

Box 1: Round-robin
Round-robin tables are useful for improving loading speed.
Scenario: Partition data that contains sales transaction records. Partitions must be designed to provide efficient loads by month.
Box 2: Hash
Hash-distributed tables improve query performance on large fact tables.
Reference:

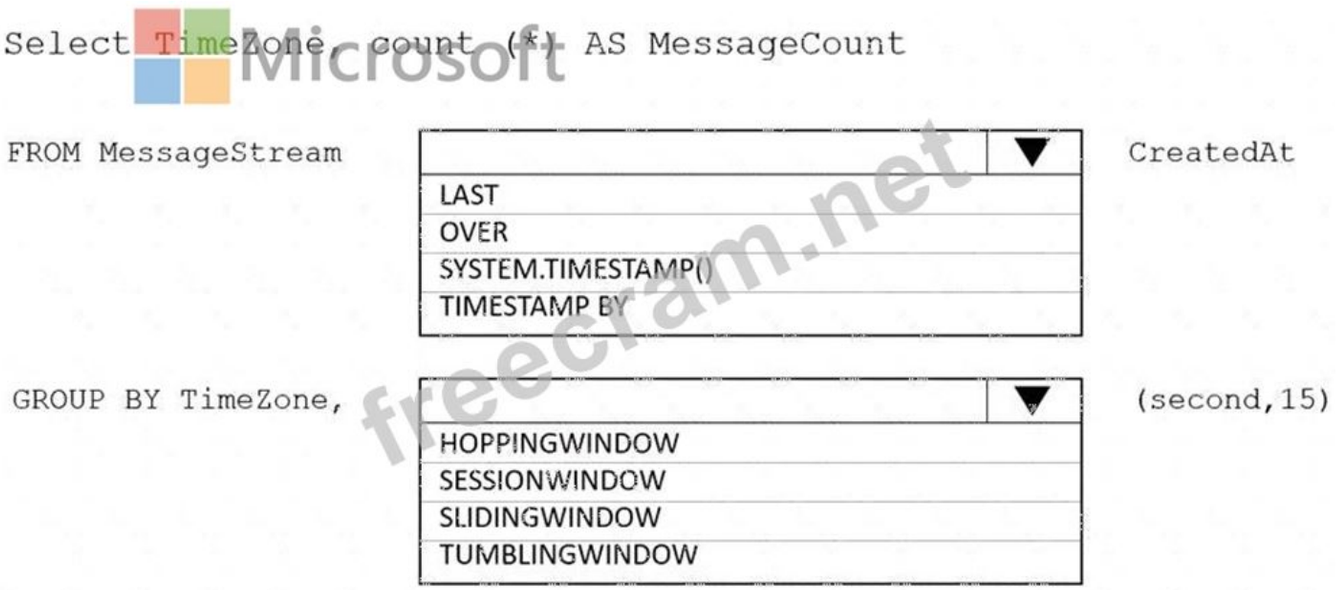
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

NEW QUESTION: 60

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: (SHOW ANSWER)
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: 61
You are designing an Azure Stream Analytics solution that receives instant messaging data from an Azure Event Hub.
You need to ensure that the output from the Stream Analytics job counts the number of messages per time zone every 15 seconds.
How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

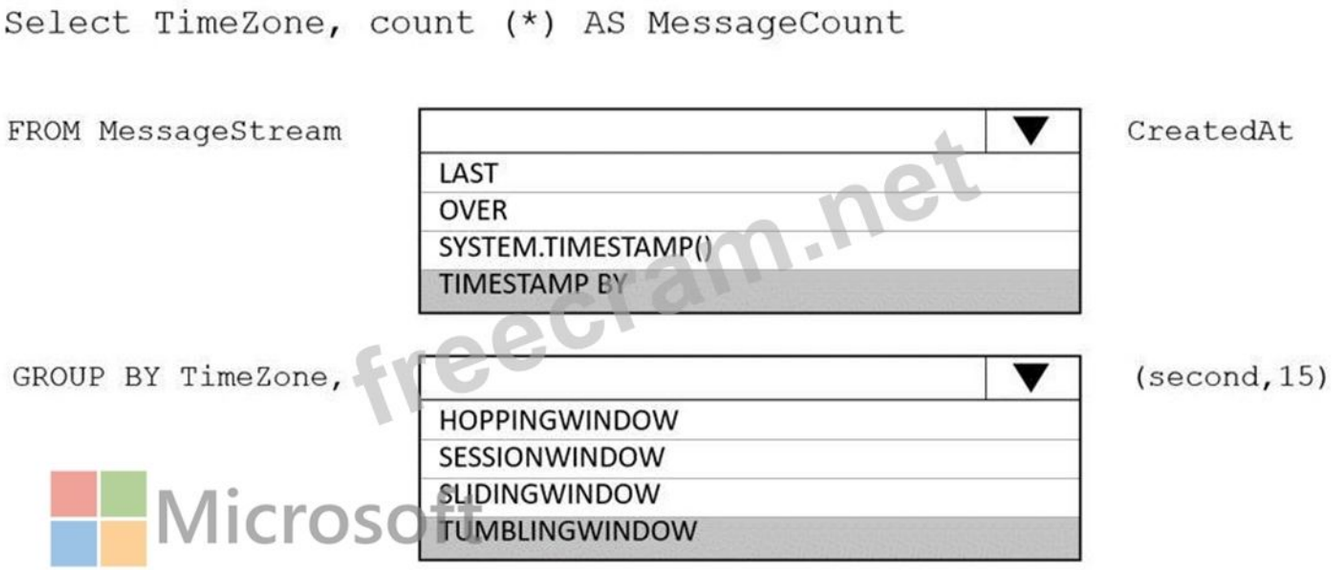


Answer:

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

FROM MessageStream [ ▼ ]                    CreatedAt
                   LAST
                   OVER
                   SYSTEM.TIMESTAMP()
                   TIMESTAMP BY

GROUP BY TimeZone, [ ▼ ]                    (second,15)
                   HOPPINGWINDOW
                   SESSIONWINDOW
                   SLIDINGWINDOW
                   TUMBLINGWINDOW
```

Explanation

Table Description automatically generated

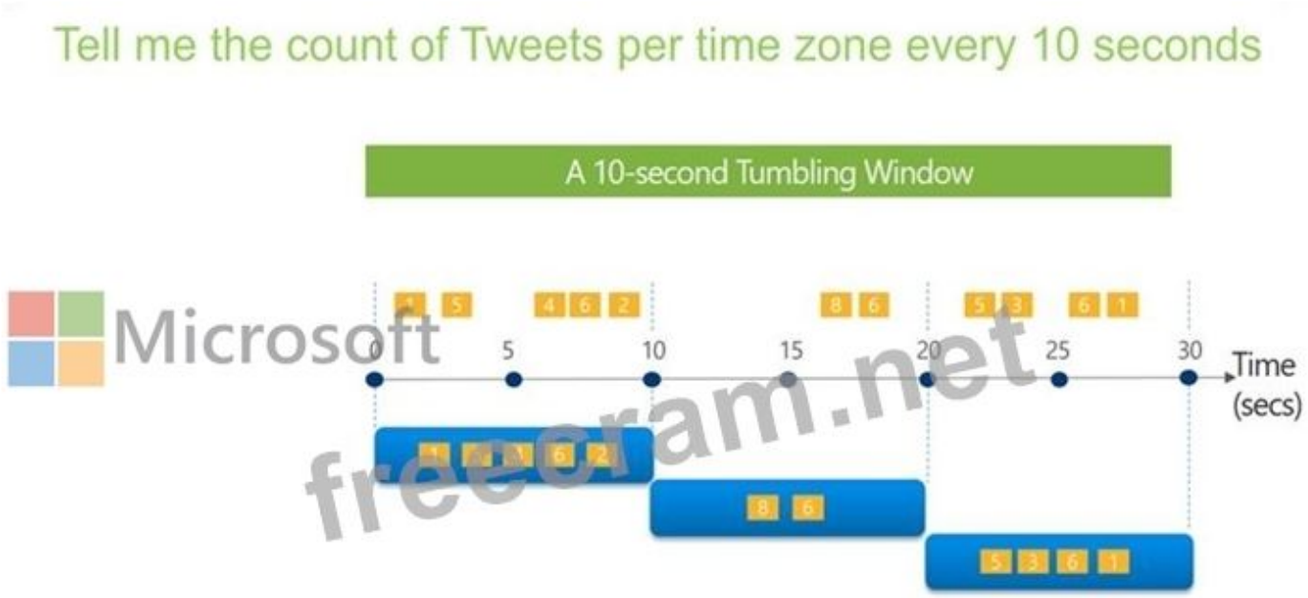


Box 1: timestamp by

Box 2: TUMBLINGWINDOW

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

Timeline Description automatically generated



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

NEW QUESTION: 62
You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1.
You need to verify whether the size of the transaction log file for each distribution of DW1 is smaller than 160 GB.
What should you do?
A. On the master database, execute a query against the sys.dm_pdw_nodes_os_performance_counters dynamic management view.
B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.
C. On DW1, execute a query against the sys.database_files dynamic management view.
D. Execute a query against the logs of DW1 by using the

Get-AzOperationalInsightSearchResult PowerShell cmdlet.

Answer: (SHOW ANSWER)

Explanation

The following query returns the transaction log size on each distribution. If one of the log files is reaching 160 GB, you should consider scaling up your instance or limiting your transaction size.

```
-- Transaction log size
SELECT
instance_name as distribution_db,
cntr_value*1.0/1048576 as log_file_size_used_GB,
pdw_node_id
FROM sys.dm_pdw_nodes_os_performance_counters
WHERE
instance_name like 'Distribution_%'
AND counter_name = 'Log File(s) Used Size (KB)'
```

References:

https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-manage-monitor

NEW QUESTION: 63

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: 64
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: (SHOW ANSWER)
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: 65
You are designing 2 solution that will use tables in Delta Lake on Azure Databricks.
You need to minimize how long it takes to perform the following:
*Queries against non-partitioned tables
* Joins on non-partitioned columns
Which two options should you include in the solution? Each correct answer presents part of the solution.
(Choose Correct Answer and Give Explanation and References to Support the answers based from Data Engineering on Microsoft Azure)
A. Z-Ordering
B. Apache Spark caching
C. dynamic file pruning (DFP)
D. the clone command
Answer: A,B (LEAVE A REPLY)
Explanation

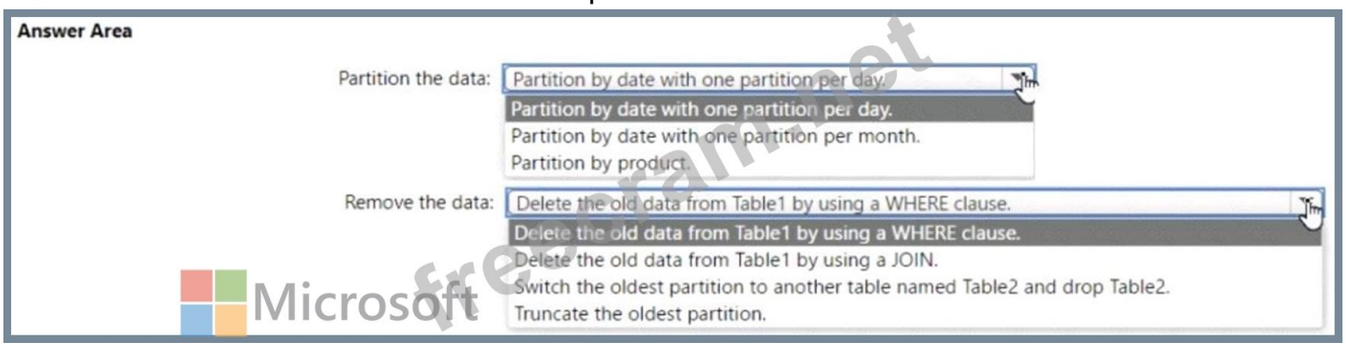
According to the information I found on the web, two options that you should include in the solution to minimize how long it takes to perform queries and joins on non-partitioned tables are:
Z-Ordering: This is a technique to colocate related information in the same set of files. This co-locality is 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: 66
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: (SHOW ANSWER)
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

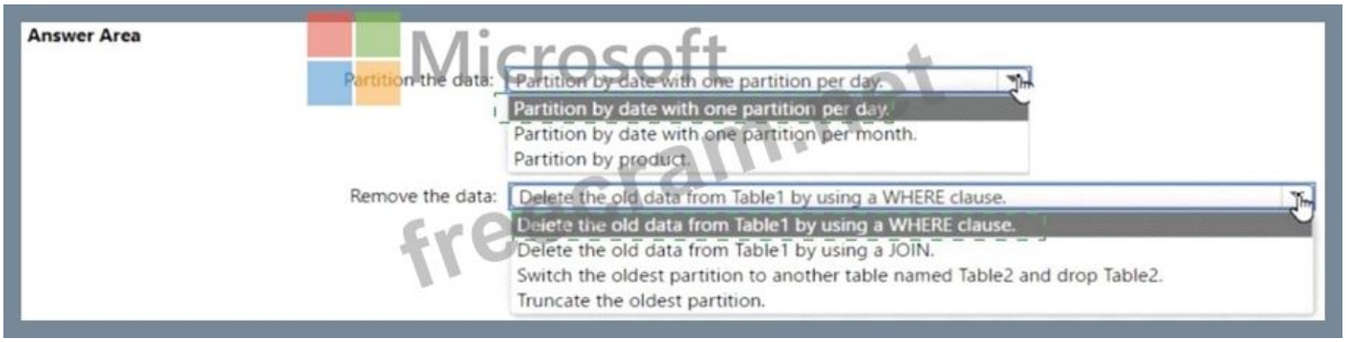
NEW QUESTION: 67

You are designing a highly available Azure Data Lake Storage solution that will induce geo-zone-redundant storage (GZRS).
You need to monitor for replication delays that can affect the recovery point objective (RPO).
What should you include m the monitoring solution?
A. Last Sync Time
B. Average Success Latency
C. Error errors
D. availability
Answer: (SHOW ANSWER)
Explanation
Because geo-replication is asynchronous, it is possible that data written to the primary region has not yet been written to the secondary region at the time an outage occurs. The Last Sync Time property indicates the last time that data from the primary region was written successfully to the secondary region. All writes made to the primary region before the last sync time are available to be read from the secondary location. Writes made to the primary region after the last sync time property may or may not be available for reads yet.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/last-sync-time-get

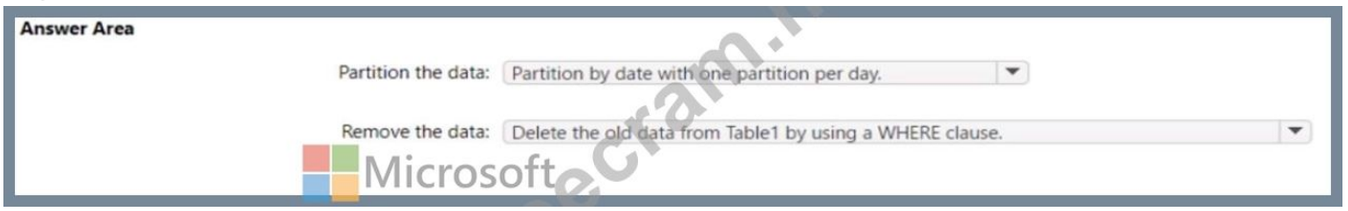
NEW QUESTION: 68
You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Tablet. Table1 contains sales data. Sixty-five million rows of data are added to Table1 monthly. At the end of each month, you need to remove data that is older than 36 months. The solution must minimize how long it takes to remove the data.
How should you partition Table1, and how should you remove the old data? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



Explanation



NEW QUESTION: 69

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: (SHOW ANSWER)

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: 70

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: (SHOW ANSWER)

Reference:

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

NEW QUESTION: 71

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.

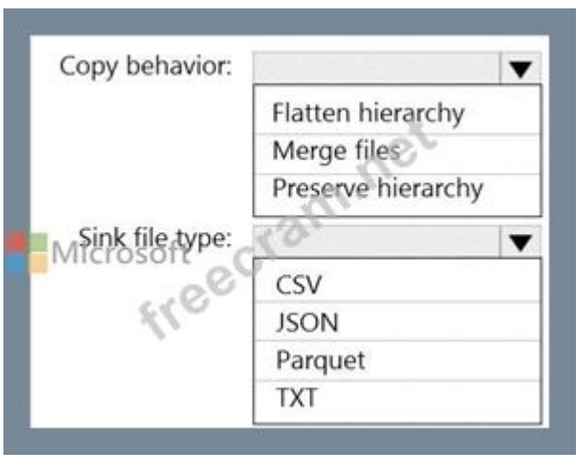
You need to move the files to a different folder and transform the data to meet the following requirements:

Provide the fastest possible query times.

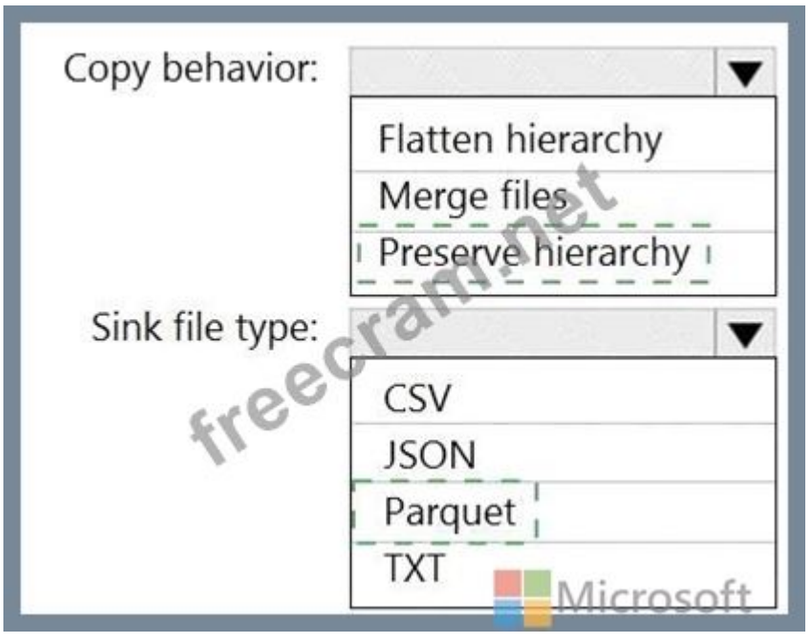
Automatically infer the schema from the underlying files.

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

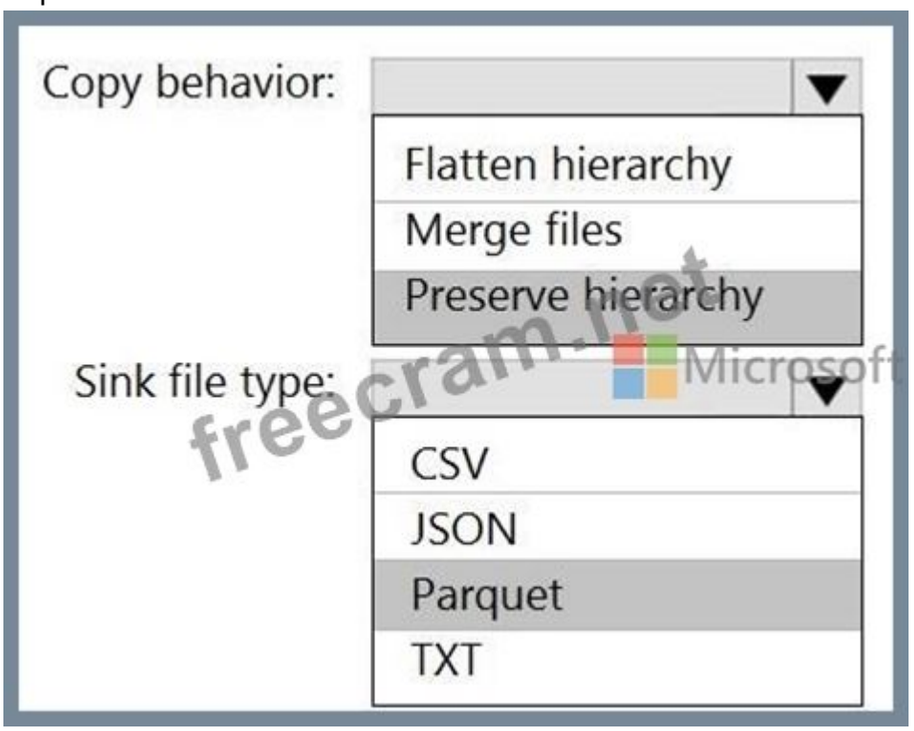
NOTE: Each correct selection is worth one point.



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.

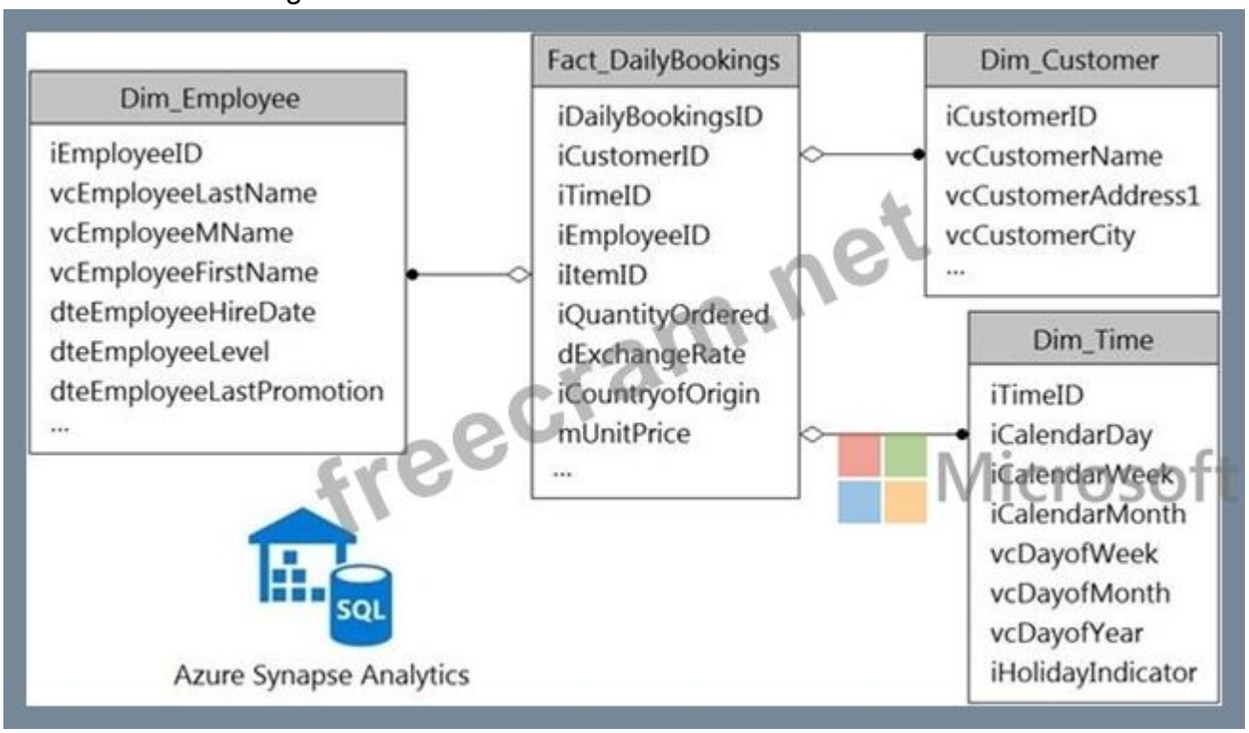
Reference:

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: 72

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

Dim_Customer:

- Hash distributed
- Round-robin
- Replicated

Dim_Employee:

- Hash distributed
- Round-robin
- Replicated

Dim_Time:

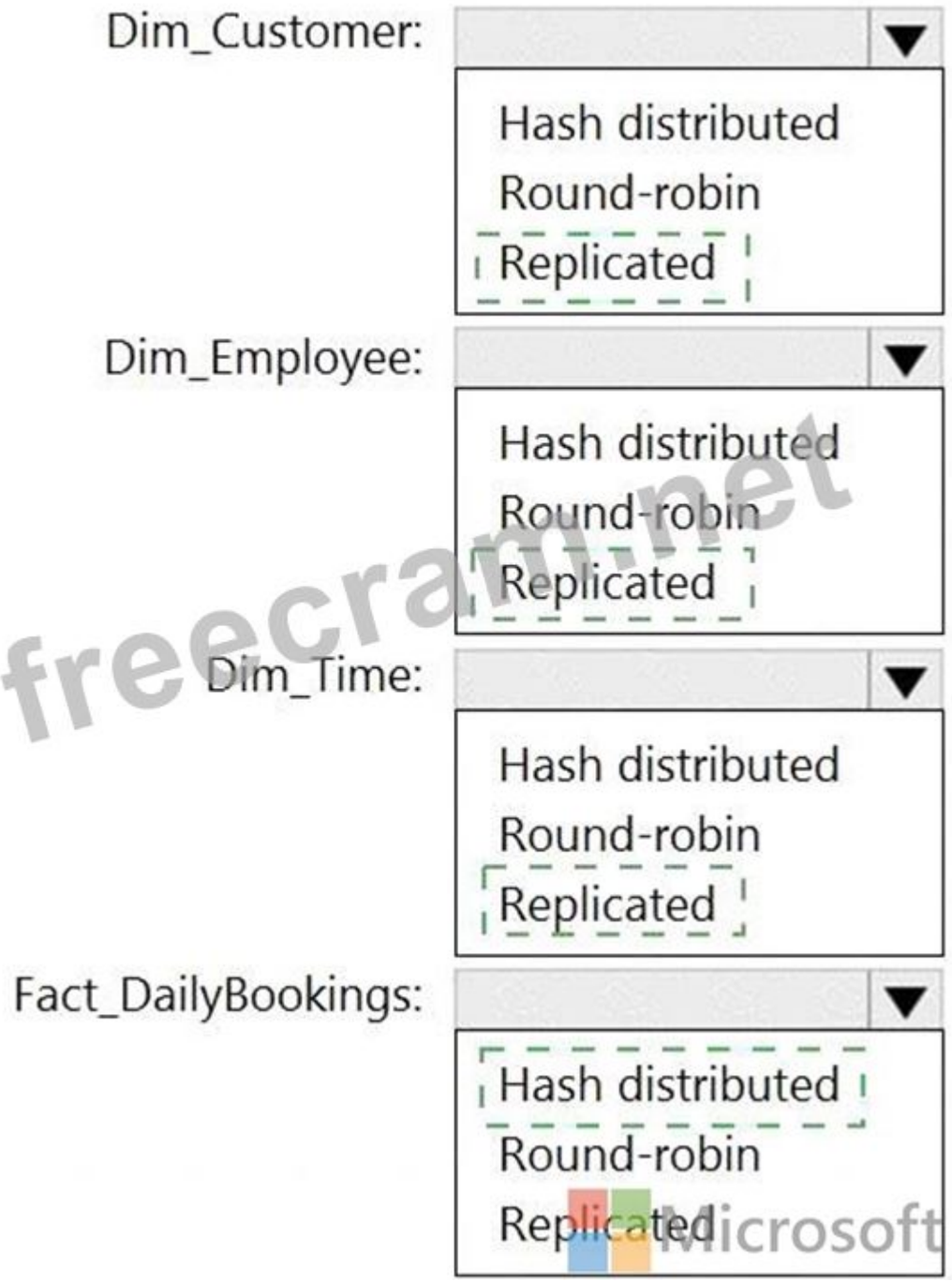
- Hash distributed
- Round-robin
- Replicated

Fact_DailyBookings:

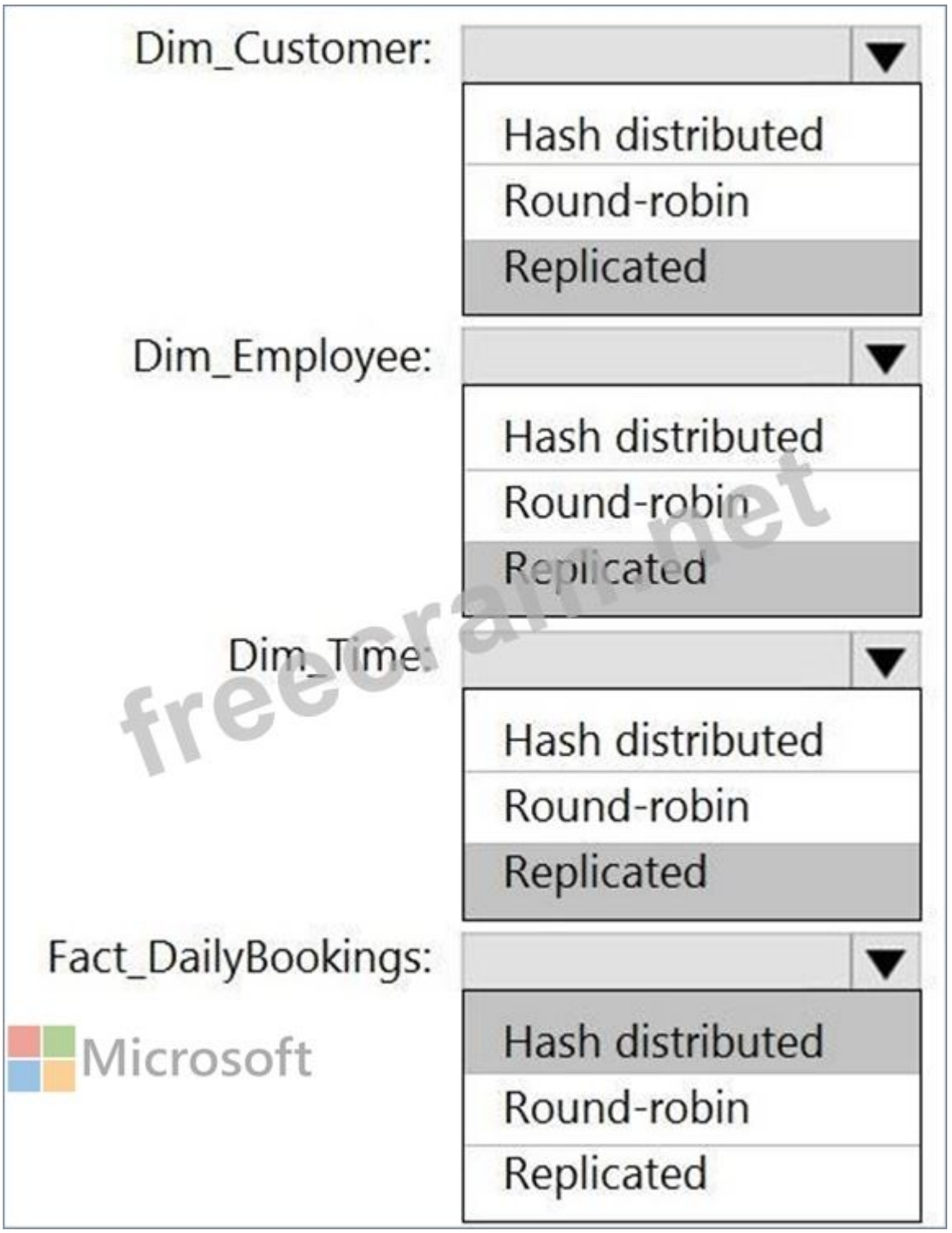
- Hash distributed
- Round-robin
- Replicated

Answer:

Answer Area



Explanation



NEW QUESTION: 73

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.

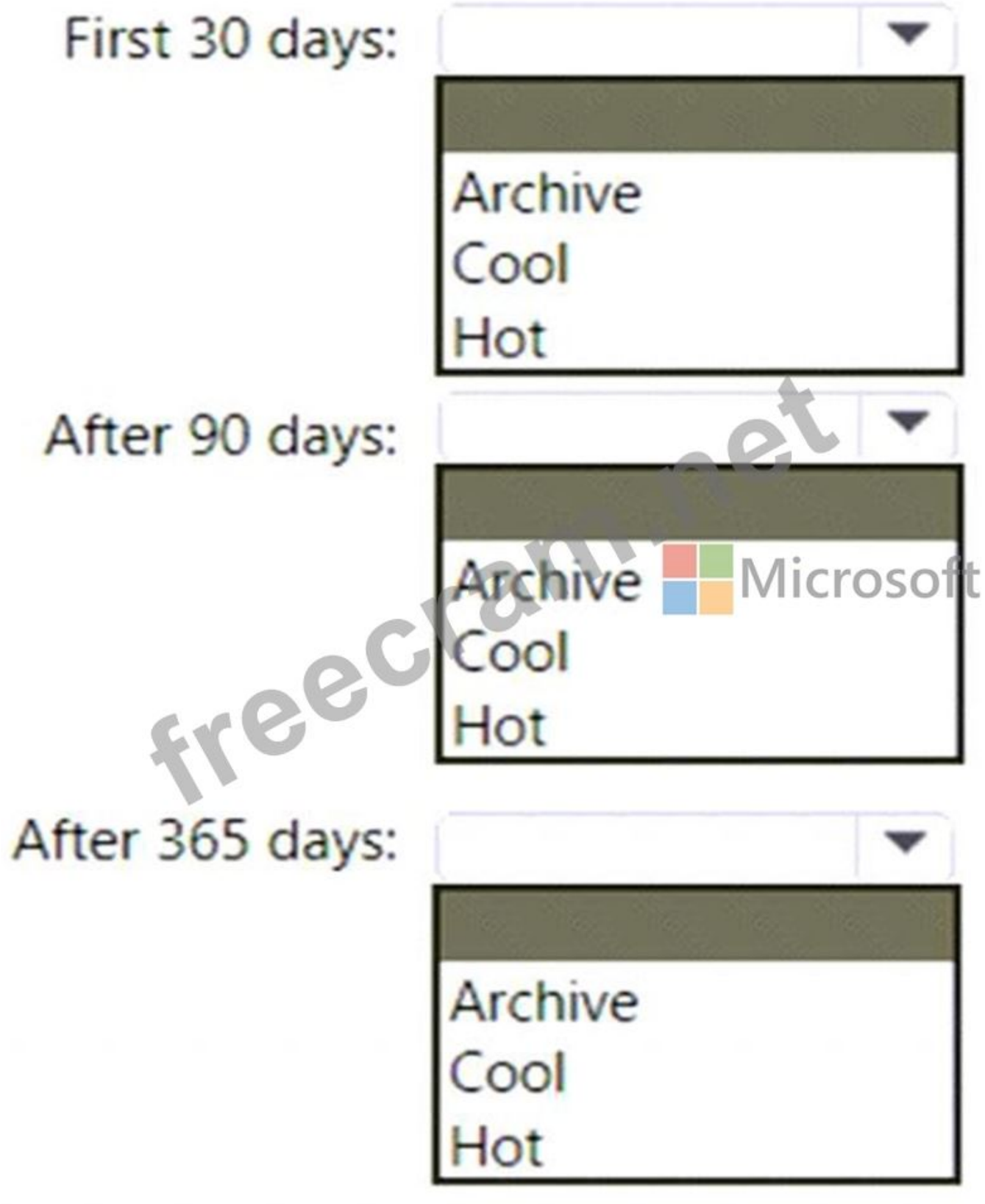
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

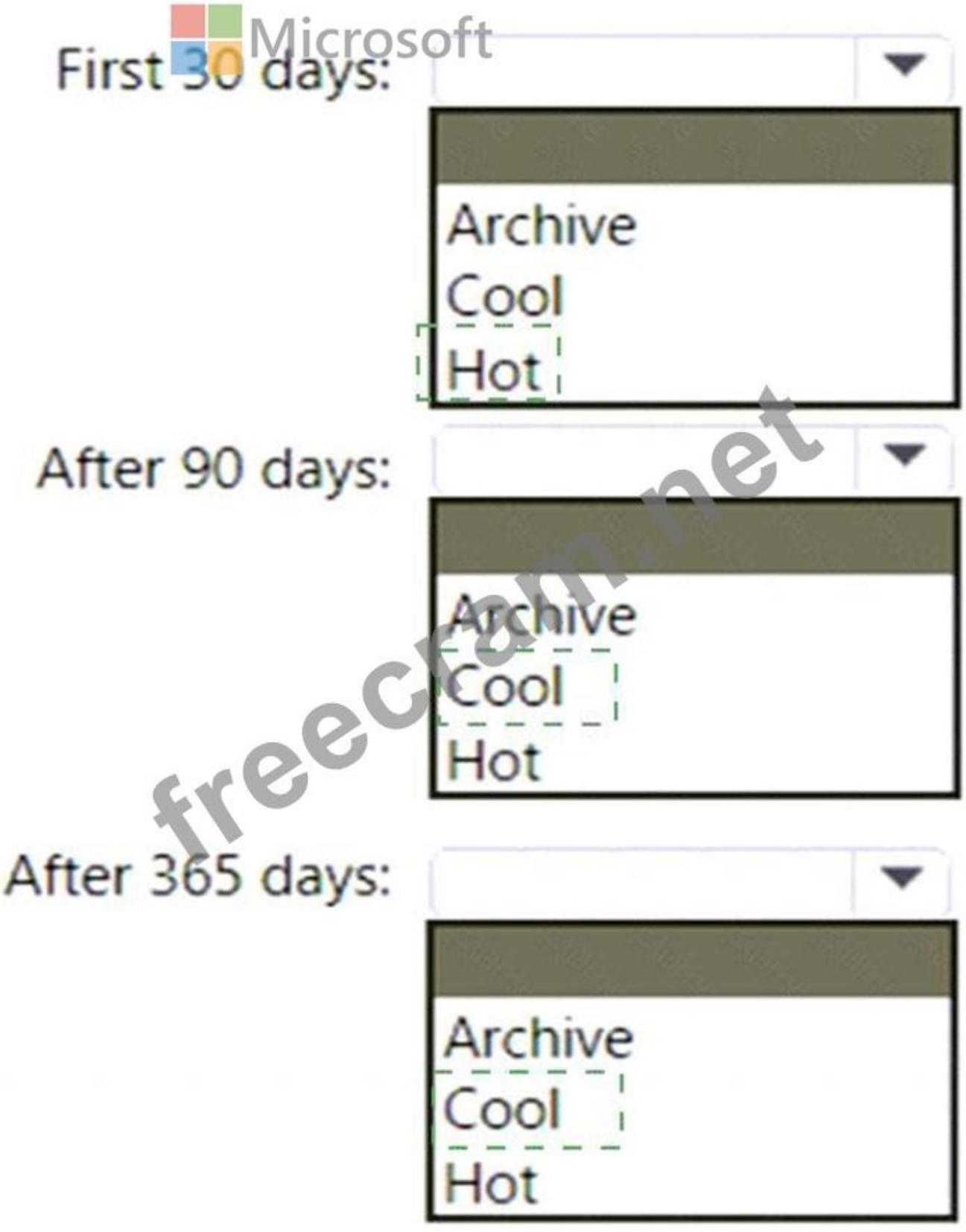
* A workload for data engineers who will use Python and SQL.

* 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: (SHOW ANSWER)
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: 74
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.



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: 75
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: 76
You are designing a streaming data solution that will ingest variable volumes of data.
You need to ensure that you can change the partition count after creation.
Which service should you use to ingest the data?
A. Azure Event Hubs Dedicated
B. Azure Stream Analytics
C. Azure Data Factory

D. Azure Synapse Analytics

Answer: (SHOW ANSWER)

Explanation

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

Reference:

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

NEW QUESTION: 77

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.

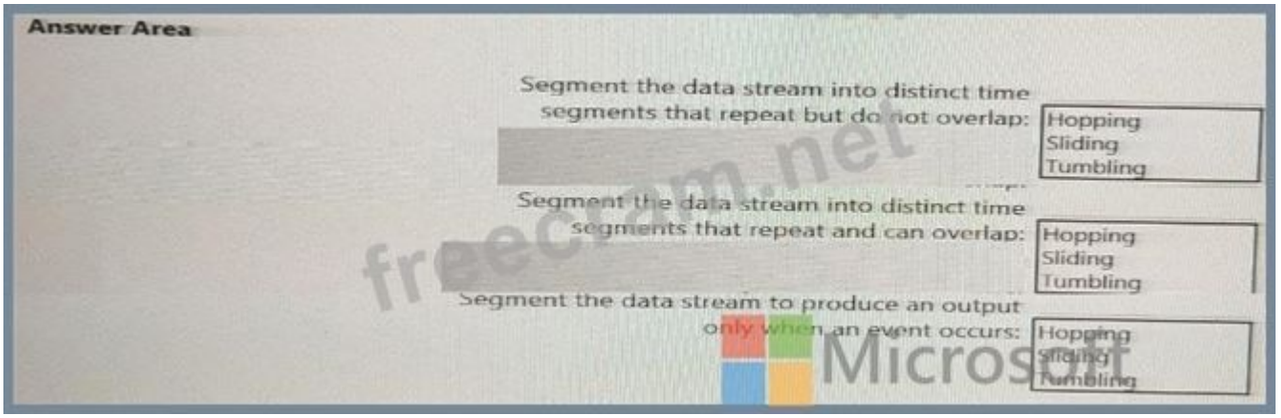
* Output sinks are not provisioned with enough capacity, so they are throttled. The possible solutions vary widely based on the flavor of output service being used.

Reference:

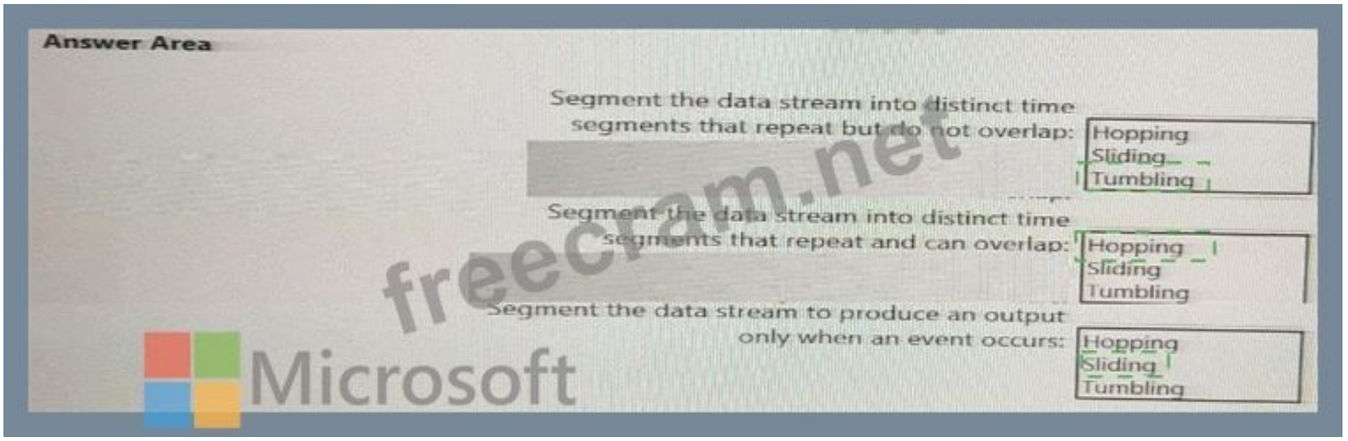
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling

NEW QUESTION: 78

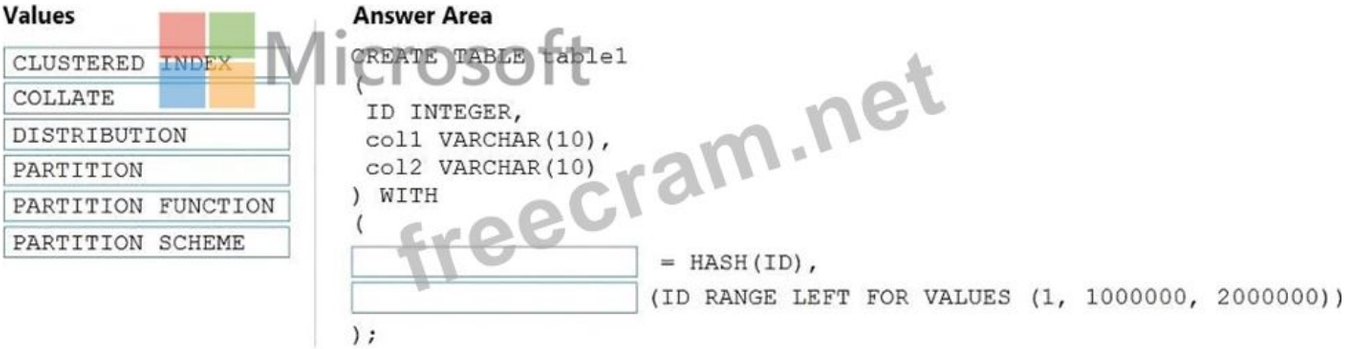
You are implementing Azure Stream Analytics windowing functions.
Which windowing function should you use for each requirement? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



NEW QUESTION: 79
You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.
How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.



Answer:

Explanation

Box 1: DISTRIBUTION

Table distribution options include DISTRIBUTION = HASH (distribution_column_name), assigns each row to one distribution by hashing the value stored in distribution_column_name.

Box 2: PARTITION

Table partition options. Syntax:

PARTITION (partition_column_name RANGE [LEFT | RIGHT] FOR VALUES ([boundary_value [,...n]])) Reference:

https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse?

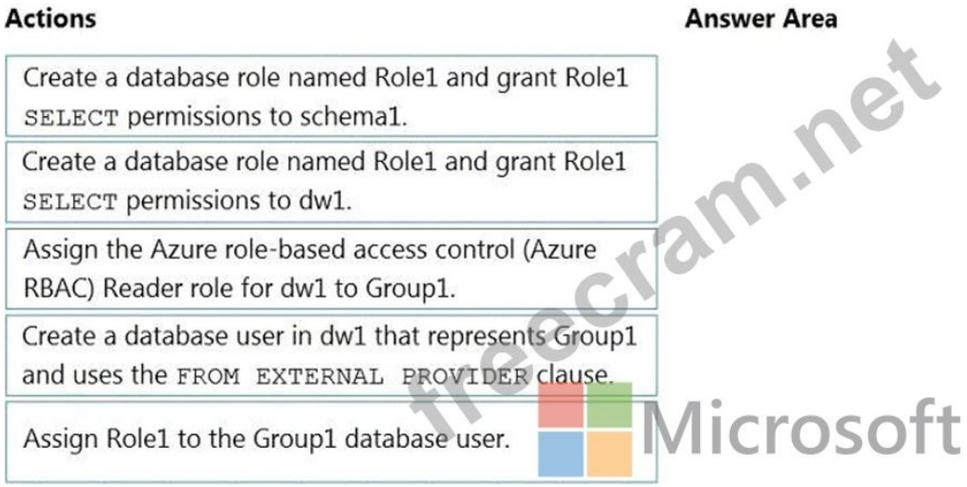
NEW QUESTION: 80

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

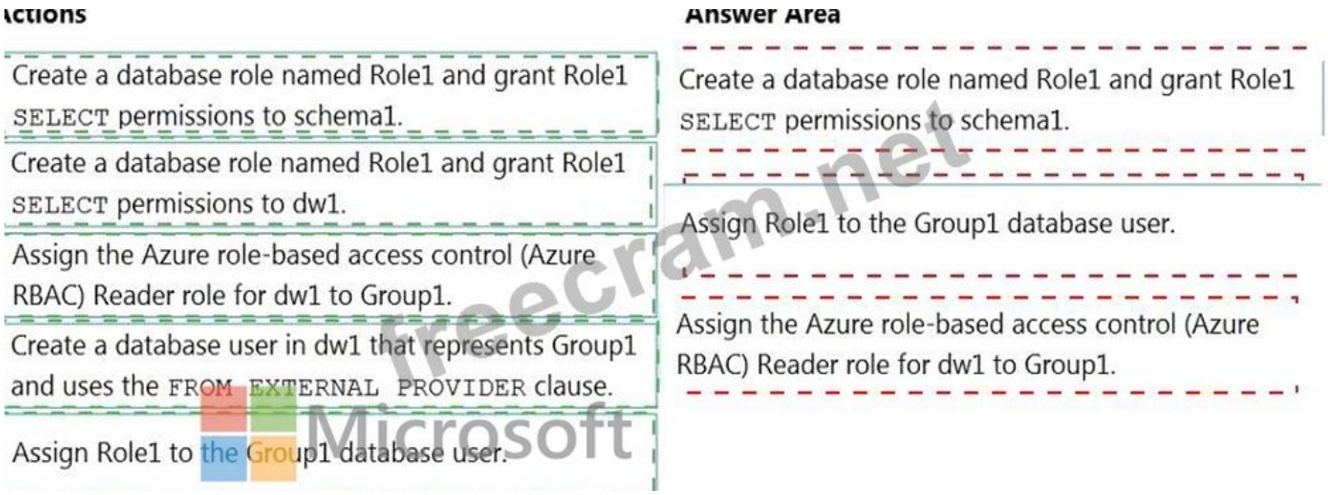
You need to grant Group1 read-only permissions to all the tables and views in schema1. The solution must use the principle of least privilege.

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

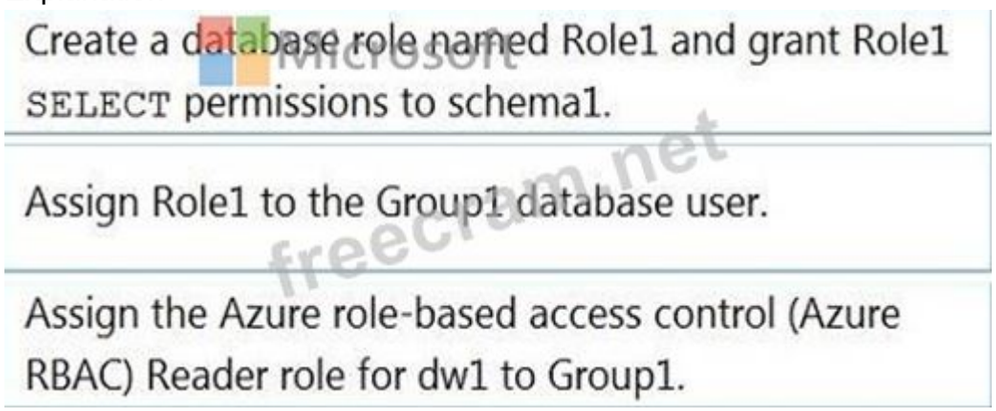
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.



Answer:



Explanation



Step 1: Create a database role named Role1 and grant Role1 SELECT permissions to schema You need to grant Group1 read-only permissions to all the tables and views in schema1.

Place one or more database users into a database role and then assign permissions to the database role.

Step 2: Assign Rol1 to the Group database user

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

Reference:

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

NEW QUESTION: 81

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

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

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

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

How should you complete the PySpark code? To answer, drag the appropriate values to he 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

Graphical user interface, text, application Description automatically generated

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

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

NEW QUESTION: 82
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. All files that have file names that beginning with "tripdata_2020".
B. All CSV files that have file names that contain "tripdata_2020".
C. Only CSV that have file names that beginning with "tripdata_2020".
D. Only CSV files in the tripdata_2020 subfolder.
Answer: (SHOW ANSWER)

NEW QUESTION: 83
You have an Azure subscription that contains an Azure Data Lake Storage account named myaccount1. The myaccount1 account contains two containers named container1 and contained. The subscription is linked to an Azure Active Directory (Azure AD) tenant that contains a security group named Group1.
You need to grant Group1 read access to contamer1. The solution must use the principle of least privilege.
Which role should you assign to Group1?
A. Storage Blob Data Reader for myaccount1
B. Storage Table Data Reader for myaccount1
C. Storage Blob Data Reader for container1

D. Storage Table Data Reader for container1

Answer: C (LEAVE A REPLY)

NEW QUESTION: 84

You are designing an Azure Synapse Analytics dedicated SQL pool.

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

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

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

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

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

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

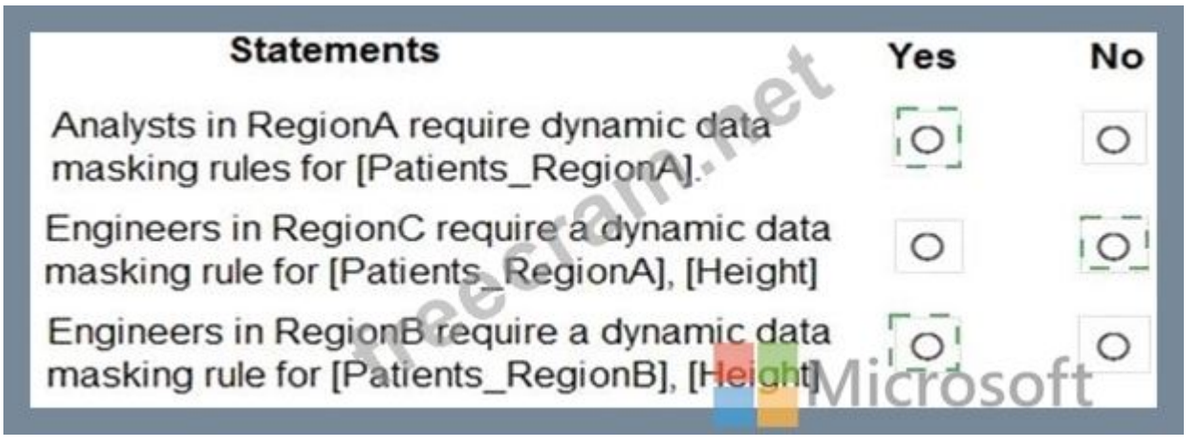
You are designing dynamic data masking to maintain compliance.

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

NOTE: Each correct selection is worth one point.

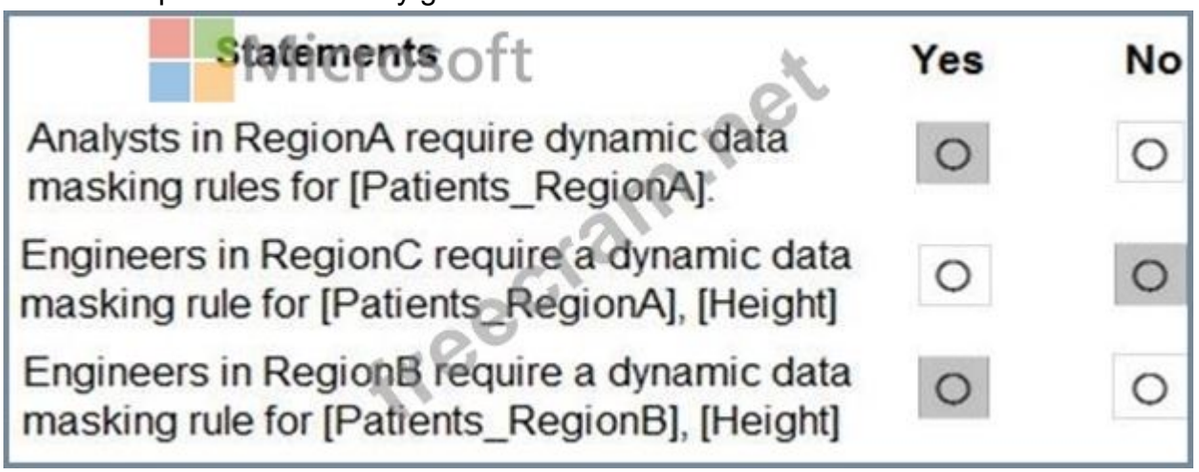
Statements	Yes	No
Analysts in RegionA require dynamic data masking rules for [Patients_RegionA].	○	○
Engineers in RegionC require a dynamic data masking rule for [Patients_RegionA], [Height]	○	○
Engineers in RegionB require a dynamic data masking rule for [Patients_RegionB], [Height]	○	○

Answer:



Explanation

Text Description automatically generated



Reference:

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

NEW QUESTION: 85

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.

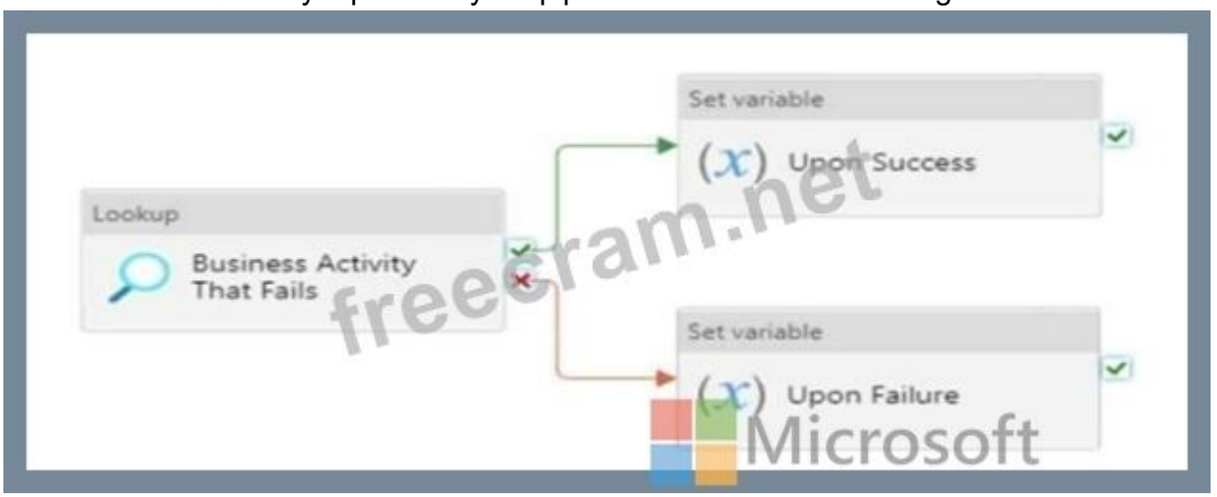
Answer: (SHOW ANSWER)

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: 86
You have the Azure Synapse Analytics pipeline shown in the following exhibit.



You need to add a set variable activity to the pipeline to ensure that after the pipeline's completion, the status of the pipeline is always successful.
What should you configure for the set variable activity?
A. a success dependency on the Business Activity That Fails activity
B. a failure dependency on the Upon Failure activity
C. a skipped dependency on the Upon Success activity
D. a skipped dependency on the Upon Failure activity
Answer: (SHOW ANSWER)
Explanation
A failure dependency means that the activity will run only if the previous activity fails. In this case, setting a failure dependency on the Upon Failure activity will ensure that the set variable activity will run after the pipeline fails and set the status of the pipeline to successful.

NEW QUESTION: 87
You are building a data flow in Azure Data Factory that upserts data into a table in an Azure Synapse Analytics dedicated SQL pool.
You need to add a transformation to the data flow. The transformation must specify logic indicating when a row from the input data must be upserted into the sink.
Which type of transformation should you add to the data flow?
A. join
B. select

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

NEW QUESTION: 88
You have an Azure Stream Analytics job.
You need to ensure that the job has enough streaming units provisioned.
You configure monitoring of the SU % Utilization metric.
Which two additional metrics should you monitor? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
A. Backlogged Input Events
B. Watermark Delay
C. Function Events
D. Out of order Events
E. Late Input Events
Answer: (SHOW ANSWER)
Explanation
To react to increased workloads and increase streaming units, consider setting an alert of 80% on the SU Utilization metric. Also, you can use watermark delay and backlogged events metrics to see if there is an impact.
Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job, by increasing the SUs.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-monitoring

NEW QUESTION: 89
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: (SHOW ANSWER)
Explanation

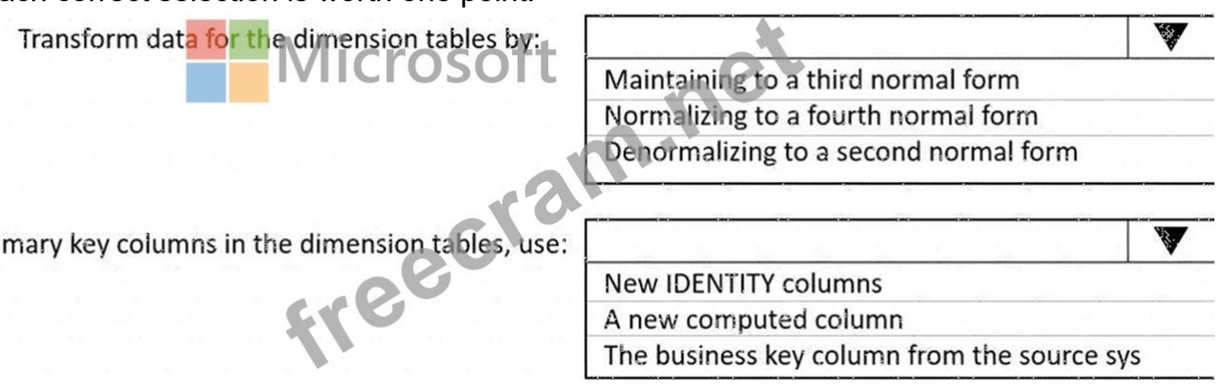
Use Synapse Studio to monitor your Apache Spark applications. To monitor running Apache Spark application Open Monitor, then select Apache Spark applications. To view the details about the Apache Spark applications that are running, select the submitting Apache Spark application and view the details. 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: 90

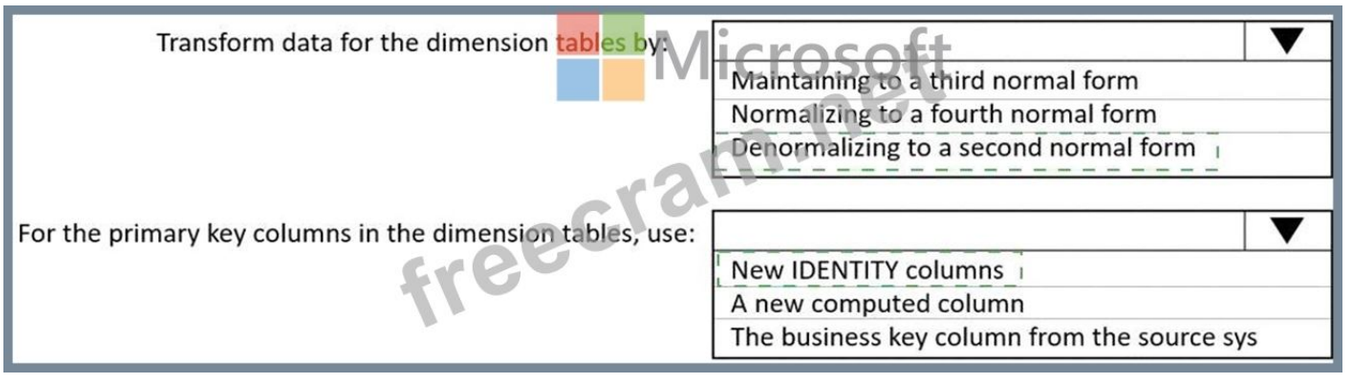
You plan to implement an Azure Data Lake Gen2 storage account.
You need to ensure that the data lake will remain available if a data center fails in the primary Azure region.
The solution must minimize costs.
Which type of replication should you use for the storage account?

A. geo-redundant storage (GRS)
B. zone-redundant storage (ZRS)
C. locally-redundant storage (LRS)
D. geo-zone-redundant storage (GZRS)
Answer: (SHOW ANSWER)
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: 91
You have a Microsoft SQL Server database that uses a third normal form schema.
You plan to migrate the data in the database to a star schema in an Azure Synapse Analytics dedicated SQI pool.
You need to design the dimension tables. The solution must optimize read operations.
What should you include in the solution? to answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

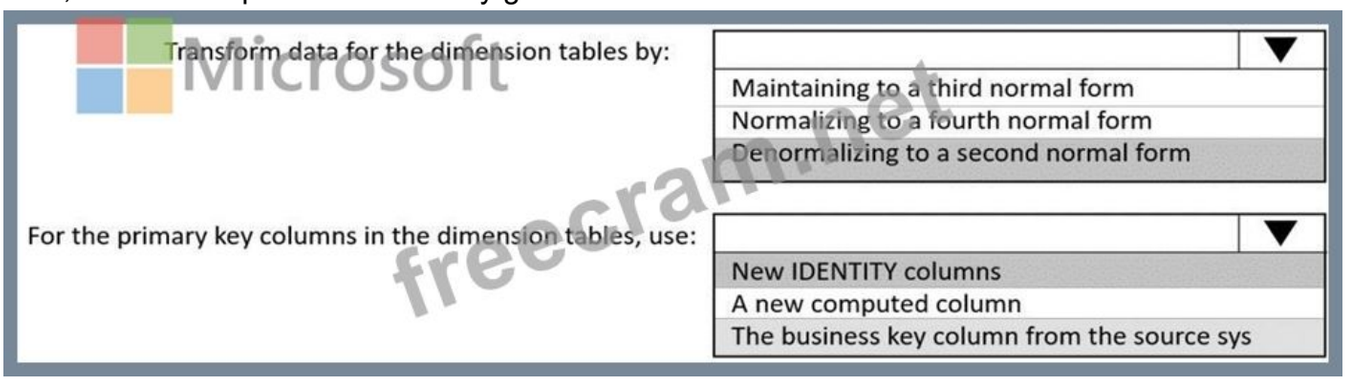


Answer:



Explanation

Text, table Description automatically generated



Box 1: Denormalize to a second normal form

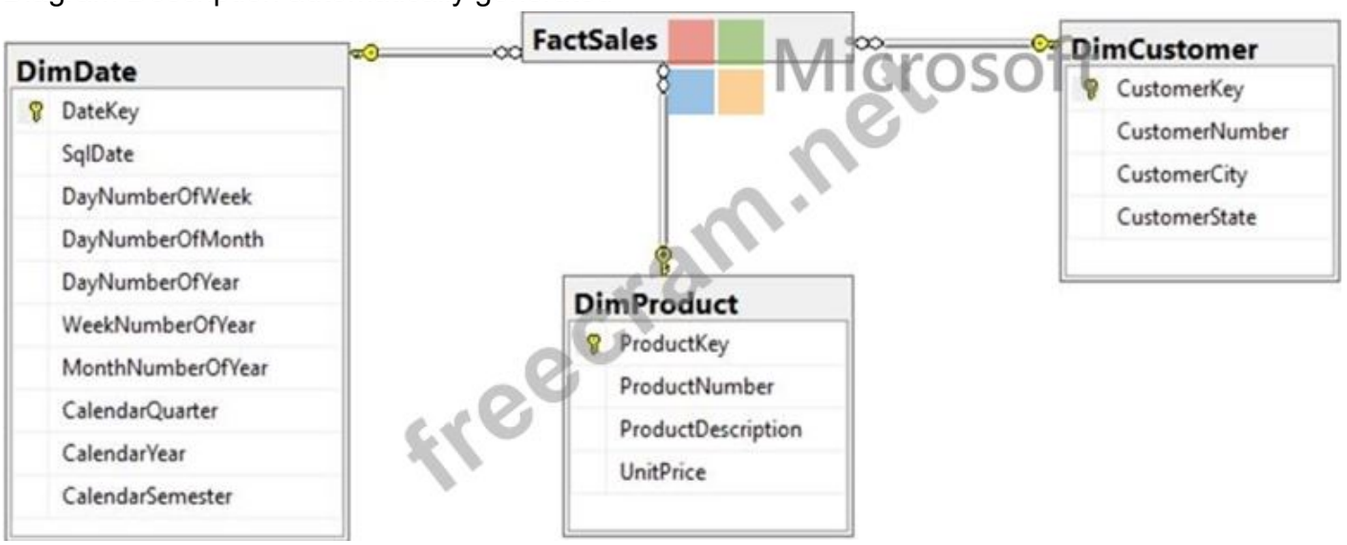
Denormalization is the process of transforming higher normal forms to lower normal forms via storing the join of higher normal form relations as a base relation. Denormalization increases the performance in data retrieval at cost of bringing update anomalies to a database.

Box 2: New identity columns

The collapsing relations strategy can be used in this step to collapse classification entities into component entities to obtain at dimension tables with single-part keys that connect directly to the fact table. The single-part key is a surrogate key generated to ensure it remains unique over time.

Example:

Diagram Description automatically generated



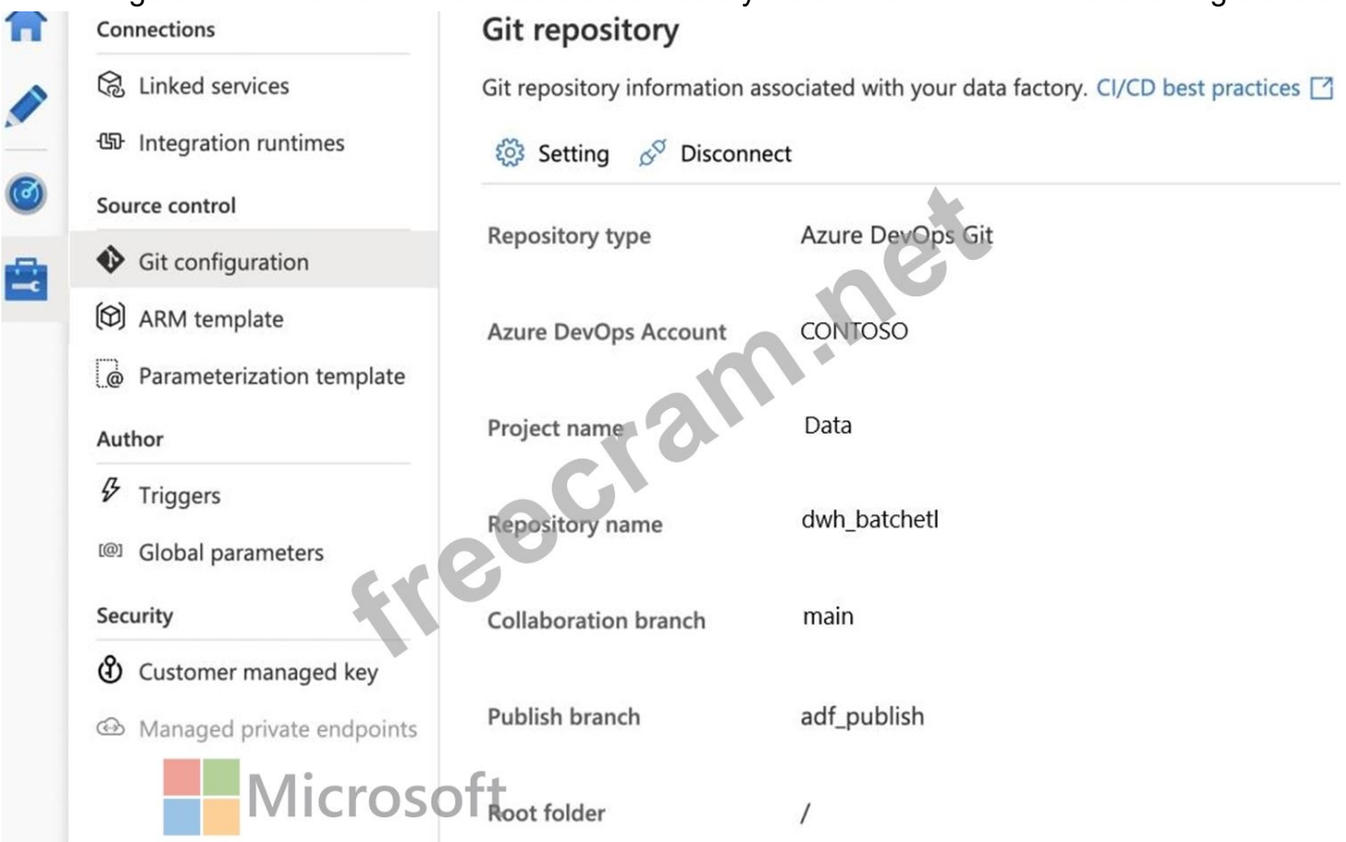
Note: A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.
Reference:
https://www.mssqltips.com/sqlservertip/5614/explore-the-role-of-normal-forms-in-dimensional-modeling/
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-identity

NEW QUESTION: 92
You develop data engineering solutions for a company.
A project requires the deployment of data to Azure Data Lake Storage.
You need to implement role-based access control (RBAC) so that project members can manage the Azure Data Lake Storage resources.
Which three actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
A. Assign Azure AD security groups to Azure Data Lake Storage.
B. Configure end-user authentication for the Azure Data Lake Storage account.
C. Configure service-to-service authentication for the Azure Data Lake Storage account.
D. Create security groups in Azure Active Directory (Azure AD) and add project members.
E. Configure access control lists (ACL) for the Azure Data Lake Storage account.
Answer: (SHOW ANSWER)
Explanation
References:
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-secure-data

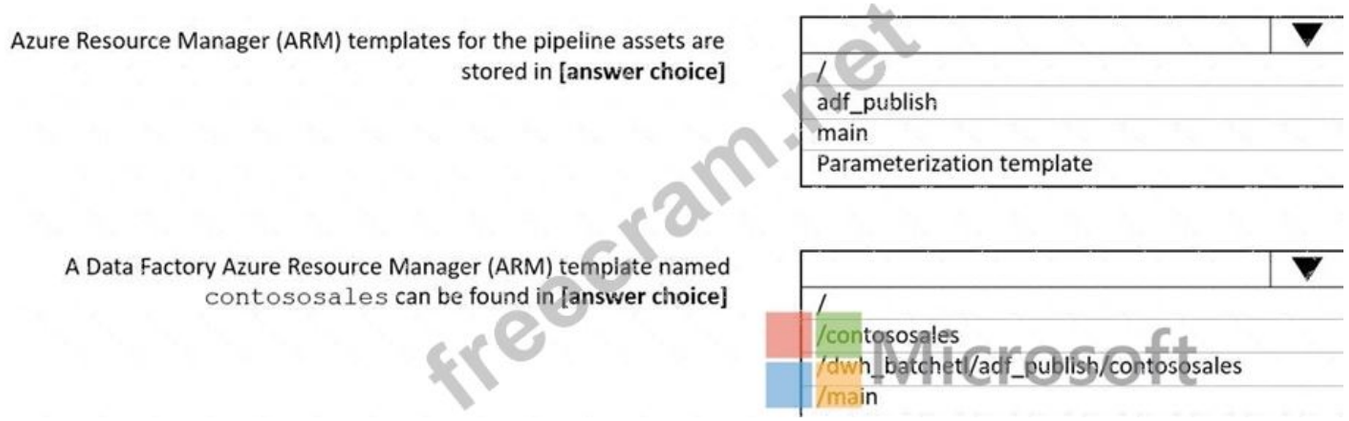
NEW QUESTION: 93
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
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: (SHOW ANSWER)
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

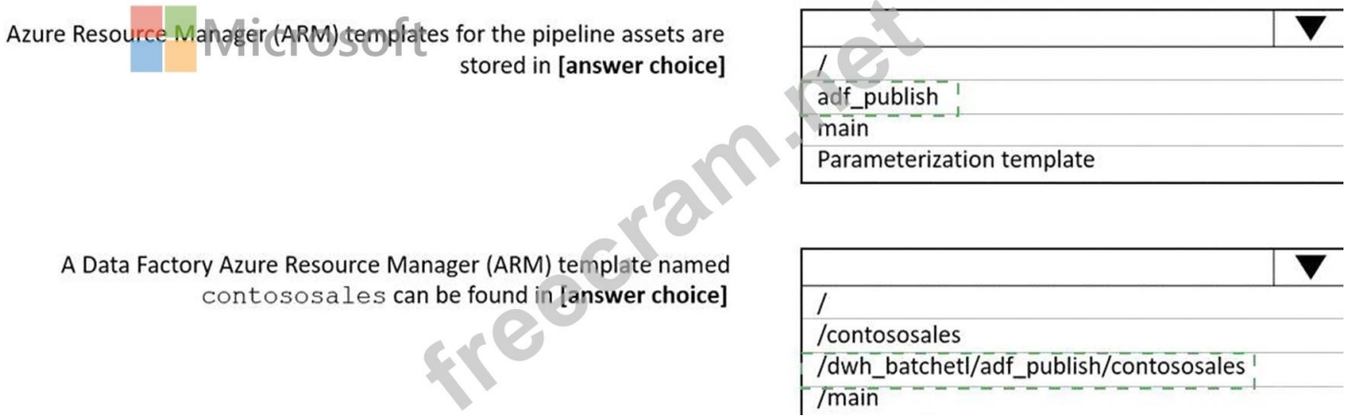
NEW QUESTION: 94
You configure version control for an Azure Data Factory instance as shown in the following exhibit.



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.

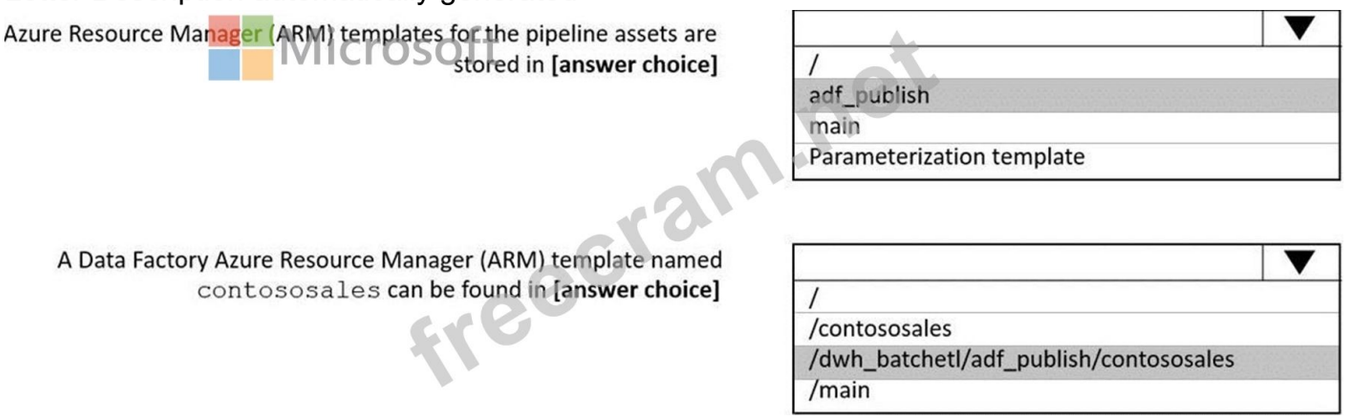


Answer:



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.

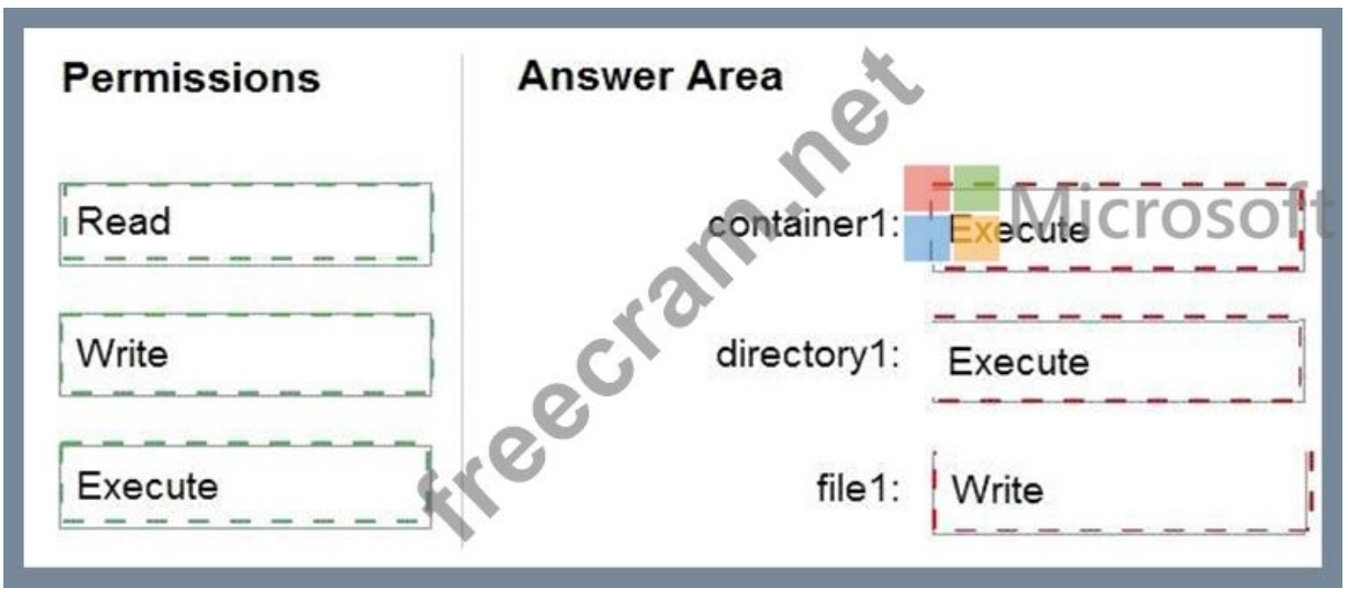
Reference:

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

NEW QUESTION: 95

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.

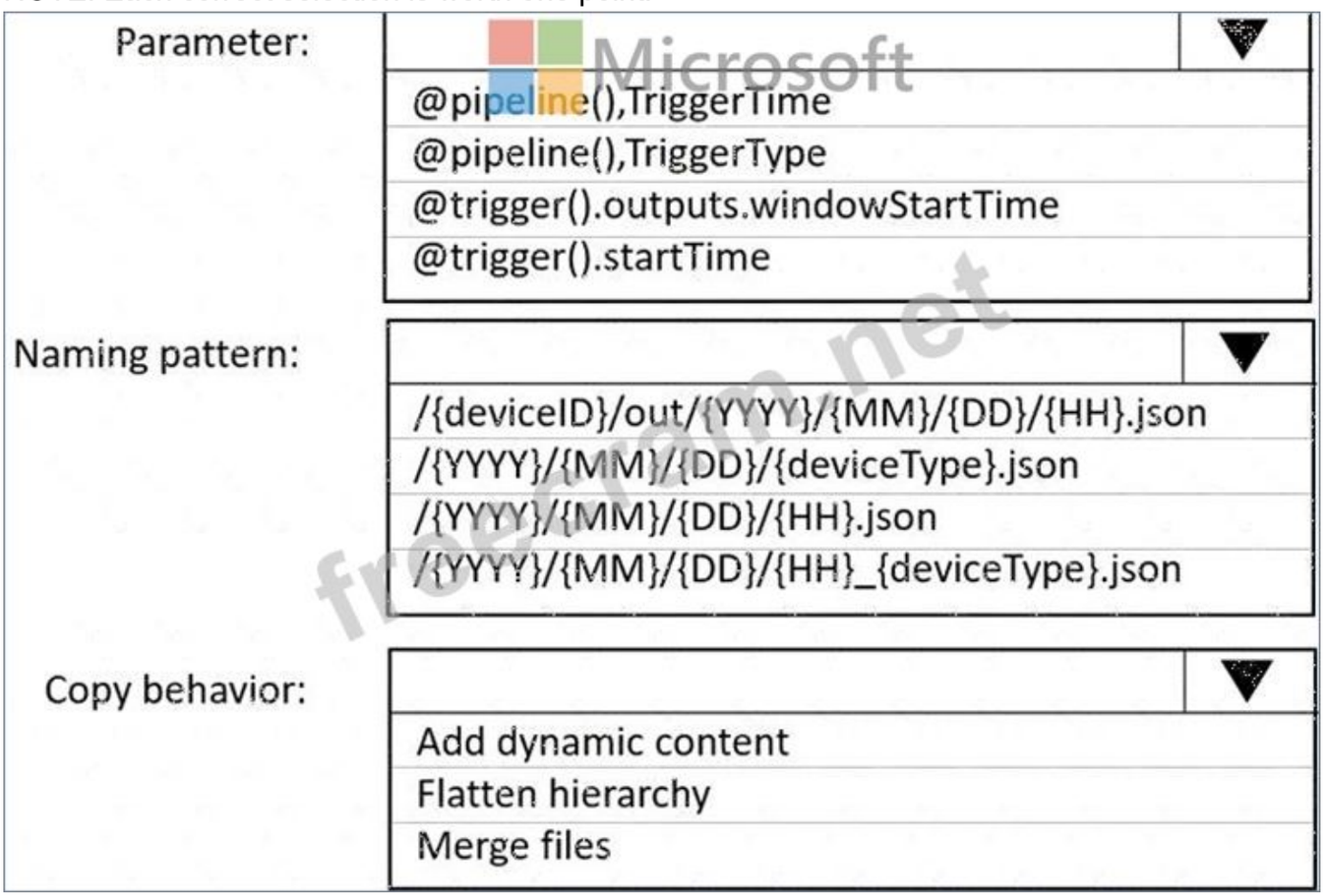
Answer:



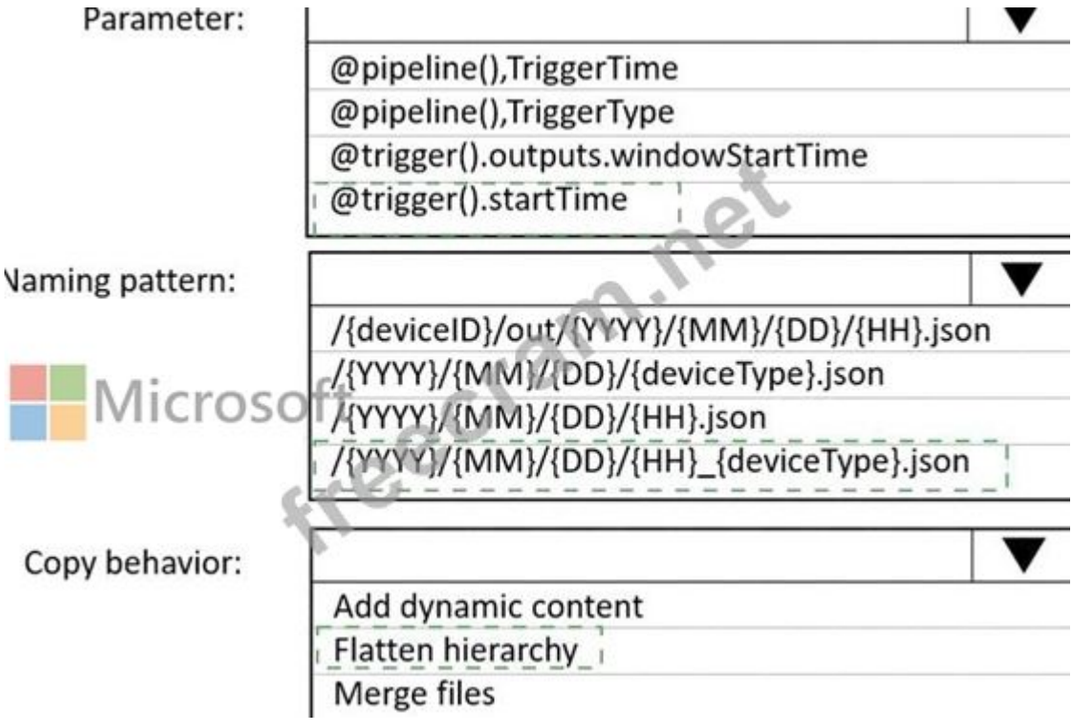
Explanation
Box 1: Execute
If you are granting permissions by using only ACLs (no Azure RBAC), then to grant a security principal read or write access to a file, you'll need to give the security principal Execute permissions to the root folder of the container, and to each folder in the hierarchy of 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: 96
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: 97

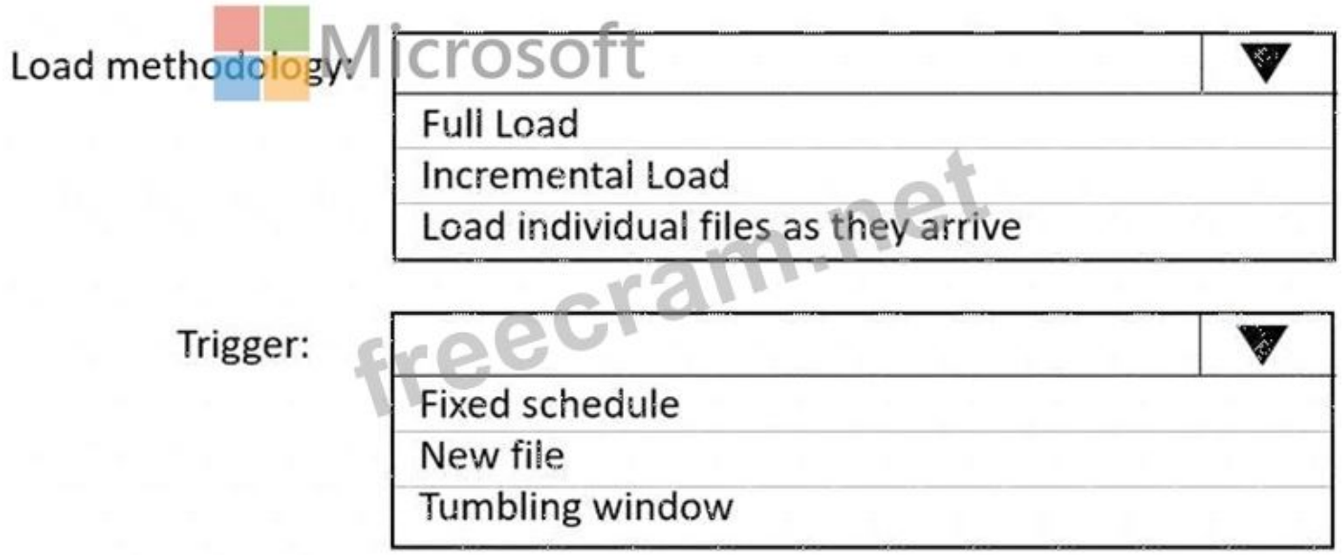
You have an Azure Storage account that generates 200,000 new files daily. The file names have a format of

{YYYY}/{MM}/{DD}/{HH}/{CustomerID}.csv.

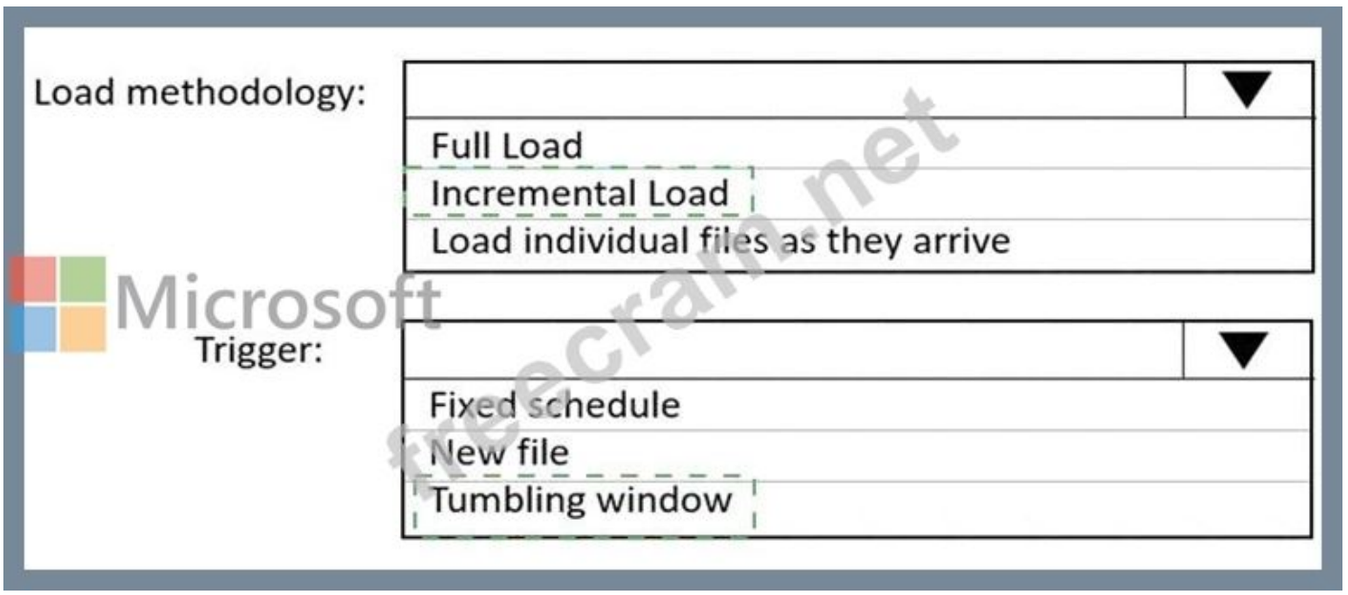
You need to design an Azure Data Factory solution that will load new data from the storage account to an Azure Data Lake once hourly. The solution must minimize load times and costs.

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

NOTE: Each correct selection is worth one point.

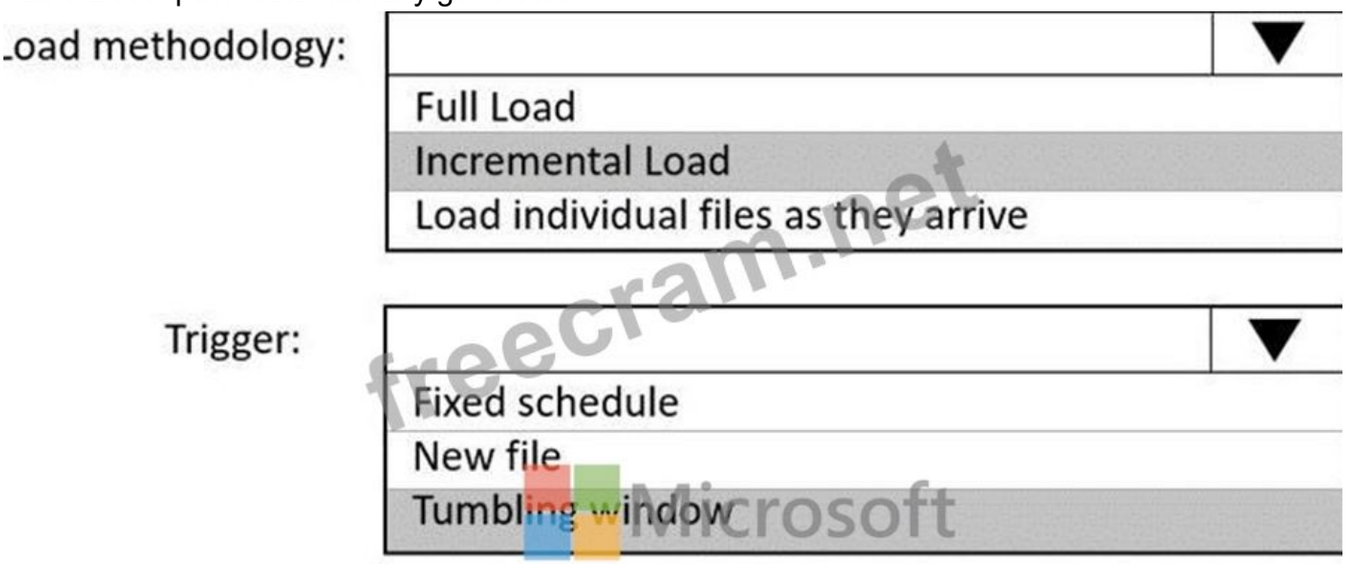


Answer:



Explanation

Table Description automatically generated

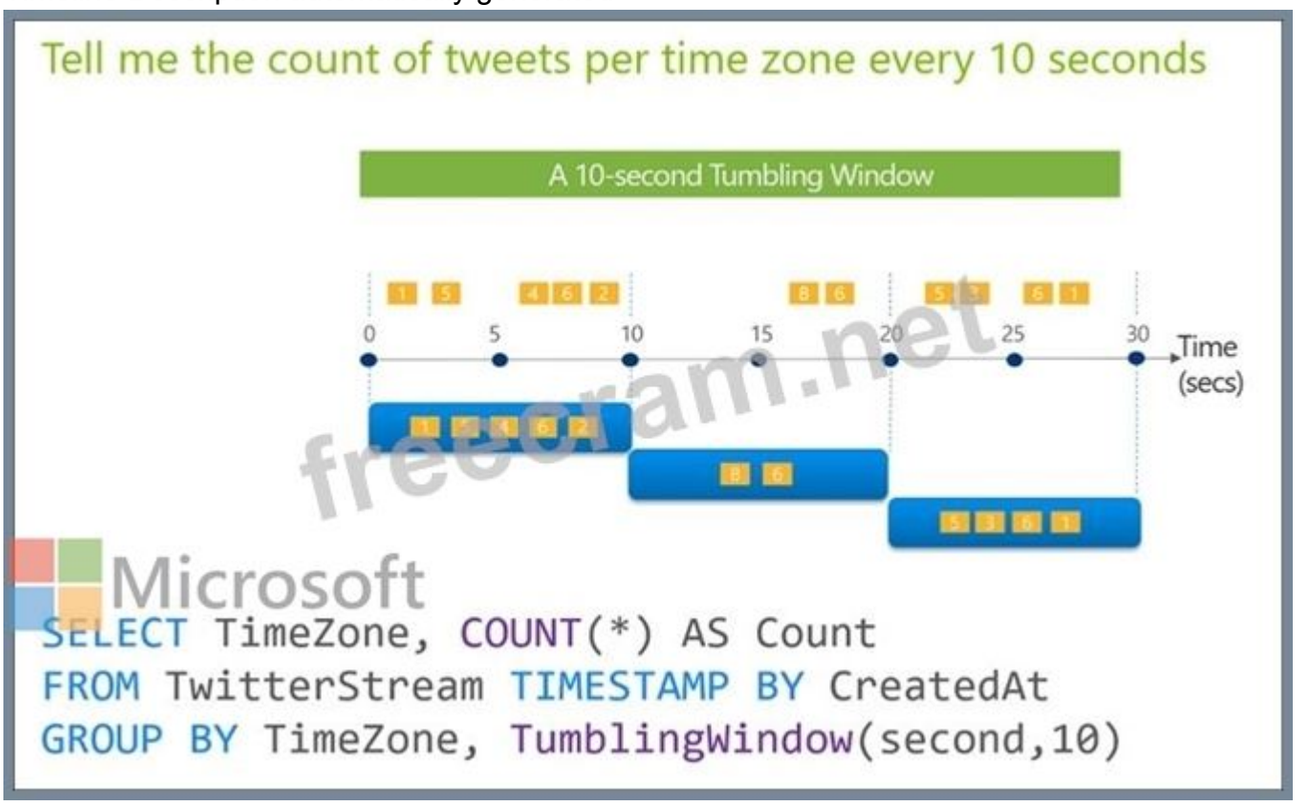


Box 1: Incremental load

Box 2: Tumbling window

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

Timeline Description automatically generated



Reference:

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

NEW QUESTION: 98

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 • Contains columns named Year, Month, Quarter, and IsWeekend
Marketing.WebSessions	1,500,500,000	Hourly inserts and updates	Fact table that contains counts of and updates sessions and page views, including foreign key values for date, channel, device, and medium
Staging.WebSessions	300,000	Hourly truncation and inserts	Staging table for web session data, truncation and including descriptive fields for inserts channel, device, and medium

You need to design the table storage for pool1. The solution must meet the following requirements:
Maximize the performance of data loading operations to Staging.WebSessions.
Minimize query times for reporting queries against the dimensional model.
Which type of table distribution should you use for each table? To answer, drag the appropriate table distribution types to the correct tables. Each table distribution type may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.

Table distribution types

- Hash
- Replicated
- Round-robin

Answer Area

- Common.Data: _____
- Marketing.Web.Sessions: _____
- Staging. Web.Sessions: _____

Answer:

Explanation

Box 1: Replicated

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

Box 2: Hash

Hash-distribution improves query performance on large fact tables.

Box 3: Round-robin

Round-robin distribution is useful for improving loading speed.

Reference:

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

Topic 2, Contoso Case StudyTransactional Date

Contoso has three years of customer, transactional, operation, sourcing, and supplier data comprised of 10 billion records stored across multiple on-premises Microsoft SQL Server servers. The SQL server instances contain data from various operational systems. The data is loaded into the instances by using SQL server integration Services (SSIS) packages.

You estimate that combining all product sales transactions into a company-wide sales transactions dataset will result in a single table that contains 5 billion rows, with one row per transaction.

Most queries targeting the sales transactions data will be used to identify which products were sold in retail stores and which products were sold online during different time period. Sales transaction data that is older than three years will be removed monthly.

You plan to create a retail store table that will contain the address of each retail store. The table will be approximately 2 MB. Queries for retail store sales will include the retail store addresses.

You plan to create a promotional table that will contain a promotion ID. The promotion ID will be associated to a specific product. The product will be identified by a product ID. The table will be approximately 5 GB.

Streaming Twitter Data

The ecommerce department at Contoso develops and Azure logic app that captures trending Twitter feeds referencing the company's products and pushes the products to Azure Event Hubs.

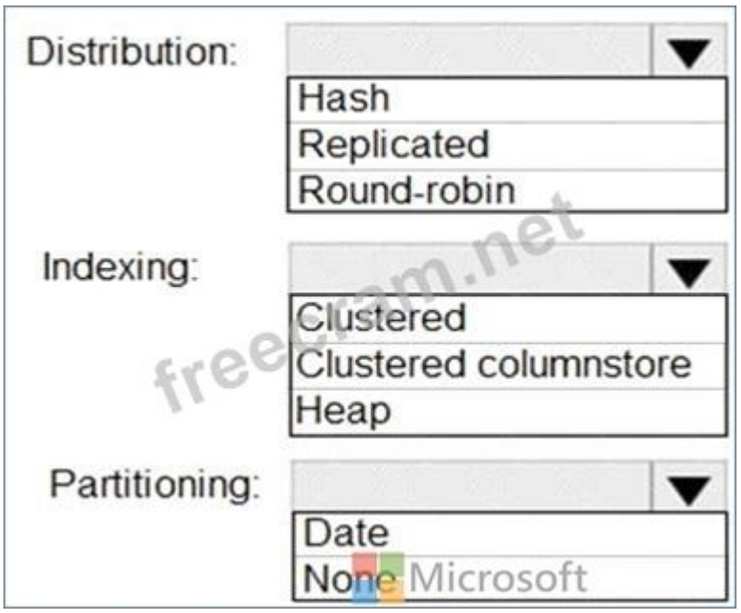
Planned Changes

Contoso plans to implement the following changes:

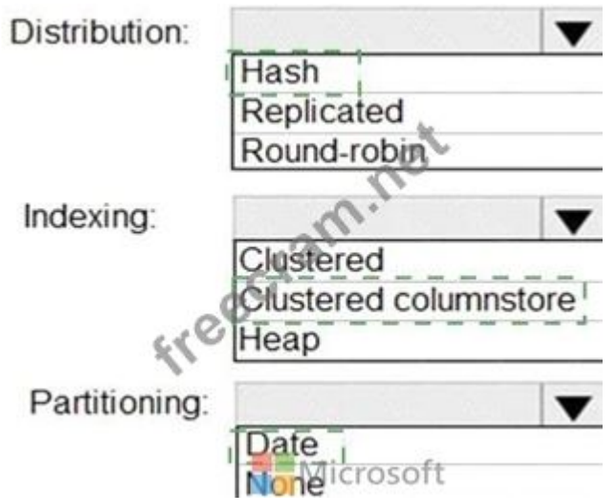
* Load the sales transaction dataset to Azure Synapse Analytics.

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

NEW QUESTION: 99
You have a SQL pool in Azure Synapse.
You plan to load data from Azure Blob storage to a staging table. Approximately 1 million rows of data will be loaded daily. The table will be truncated before each daily load.
You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.
How should you configure the table? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

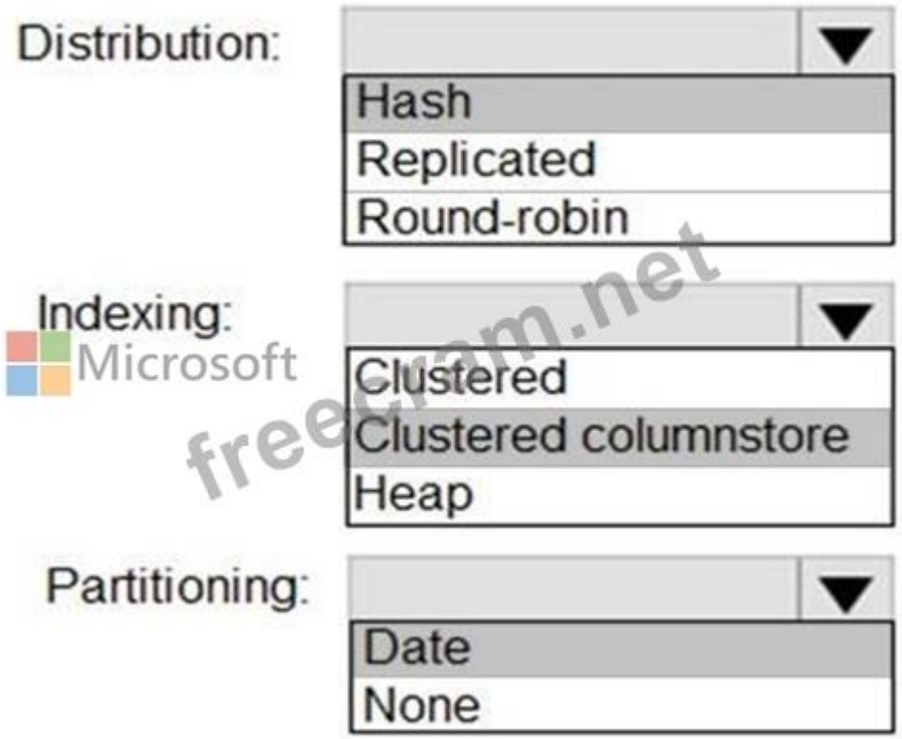


Answer:



Explanation

Graphical user interface, application, table Description automatically generated



Box 1: Hash
Hash-distributed tables improve query performance on large fact tables. They can have very large numbers of rows and still achieve high performance.
Box 2: Clustered columnstore
When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed.
Box 3: Date
Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.
Partition switching can be used to quickly remove or replace a section of a table.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribu

NEW QUESTION: 100
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. `[EffectiveStartDate] [datetime] NOT NULL`

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

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

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

E. `[EffectiveEndDate] [datetime] NOT NULL`

F. `[IsCurrentRow] [bit] NOT NULL`

Answer: (SHOW ANSWER)

NEW QUESTION: 101

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: (SHOW ANSWER)

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

NEW QUESTION: 102
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: (SHOW ANSWER)
Explanation
Scenario: All cloud data must be encrypted at rest and in transit.
Always Encryption 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: 103
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: (SHOW ANSWER)
Explanation
https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/performance-tuning-result-set-cach

NEW QUESTION: 104
A company plans to use Apache Spark analytics to analyze intrusion detection data.
You need to recommend a solution to analyze network and system activity data for malicious activities and policy violations. The solution must minimize administrative efforts.
What should you recommend?
A. Azure Data Lake Storage
B. Azure Databricks
C. Azure HDInsight
D. Azure Data Factory
Answer: (SHOW ANSWER)
Explanation
Three common analytics use cases with Microsoft Azure Databricks
Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.
Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.
Note: Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.
Reference:
https://azure.microsoft.com/es-es/blog/three-critical-analytics-use-cases-with-microsoft-azure-databricks/

NEW QUESTION: 105
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.

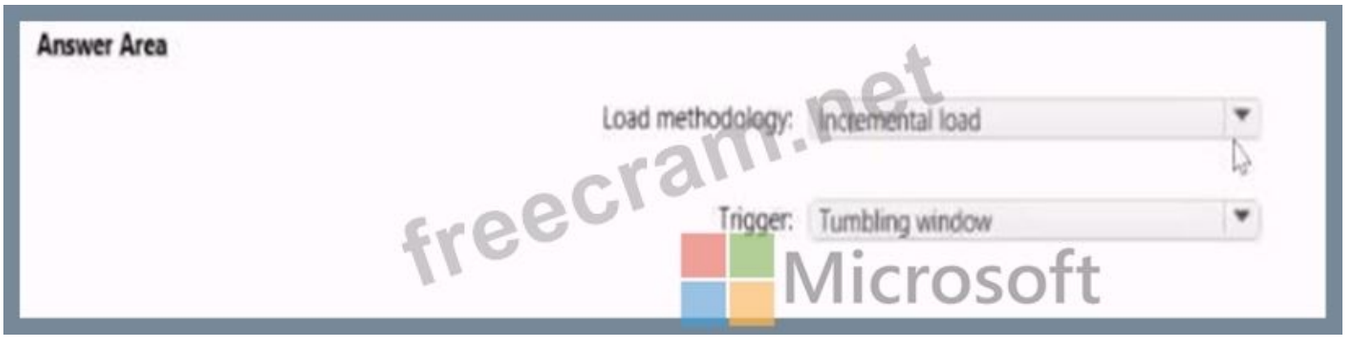
NOTE: Each correct selection is worth one point.

Answer:

See the answer below in explanation.

Explanation

answer as below



NEW QUESTION: 106

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.

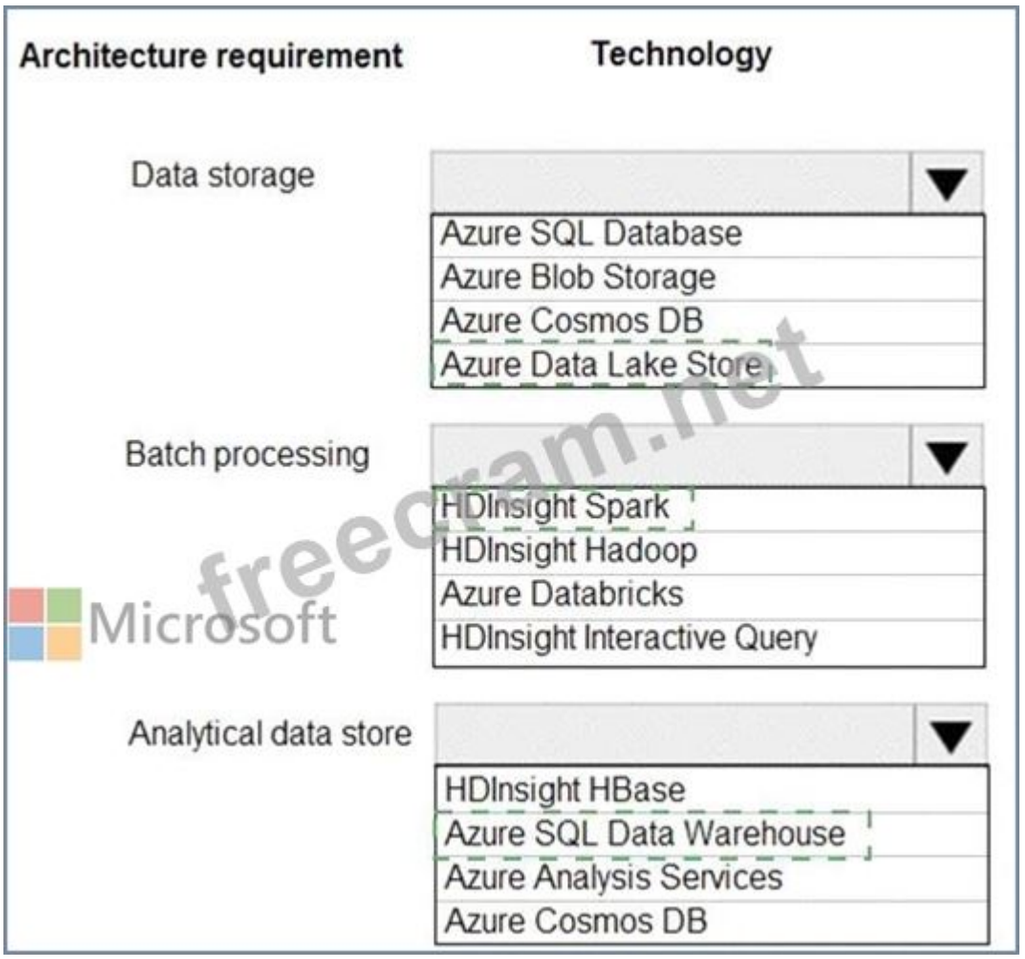
*Support SQL-based languages.

You need to identify the correct technologies to build the Lambda architecture.

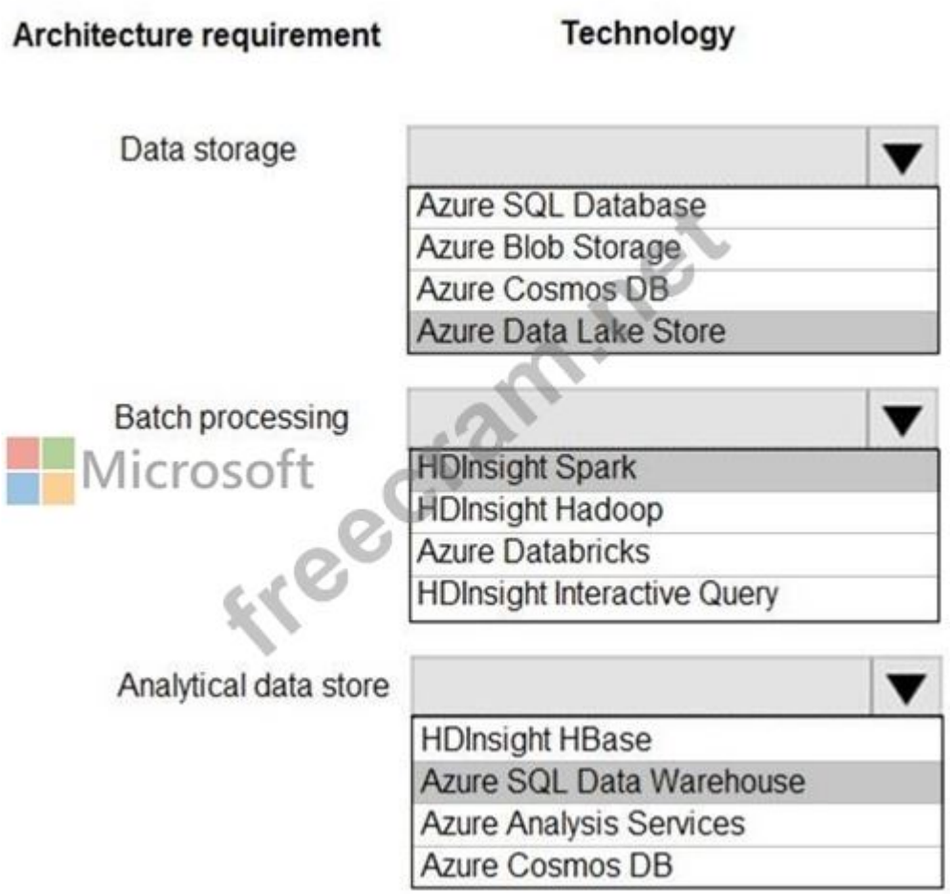
Which technologies should you use? To answer, select the appropriate options in the answer area NOTE: Each correct selection is worth one point.

Architecture requirement	Technology
Data storage	Azure SQL Database Azure Blob Storage Azure Cosmos DB Azure Data Lake Store
Batch processing	HDInsight Spark HDInsight Hadoop Azure Databricks HDInsight Interactive Query
Analytical data store	HDInsight HBase Azure SQL Data Warehouse Azure Analysis Services Azure Cosmos DB

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:

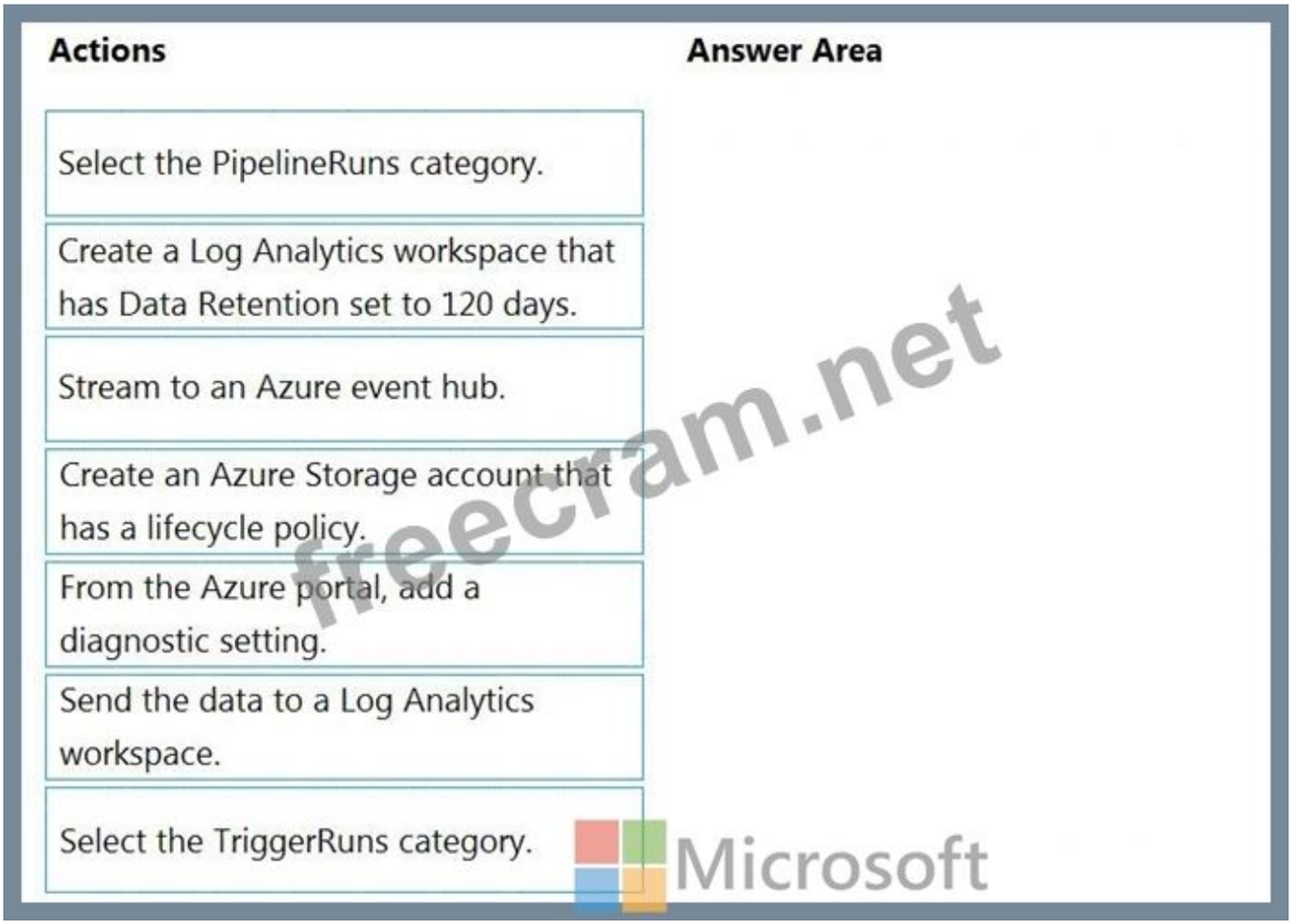
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-namespace

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

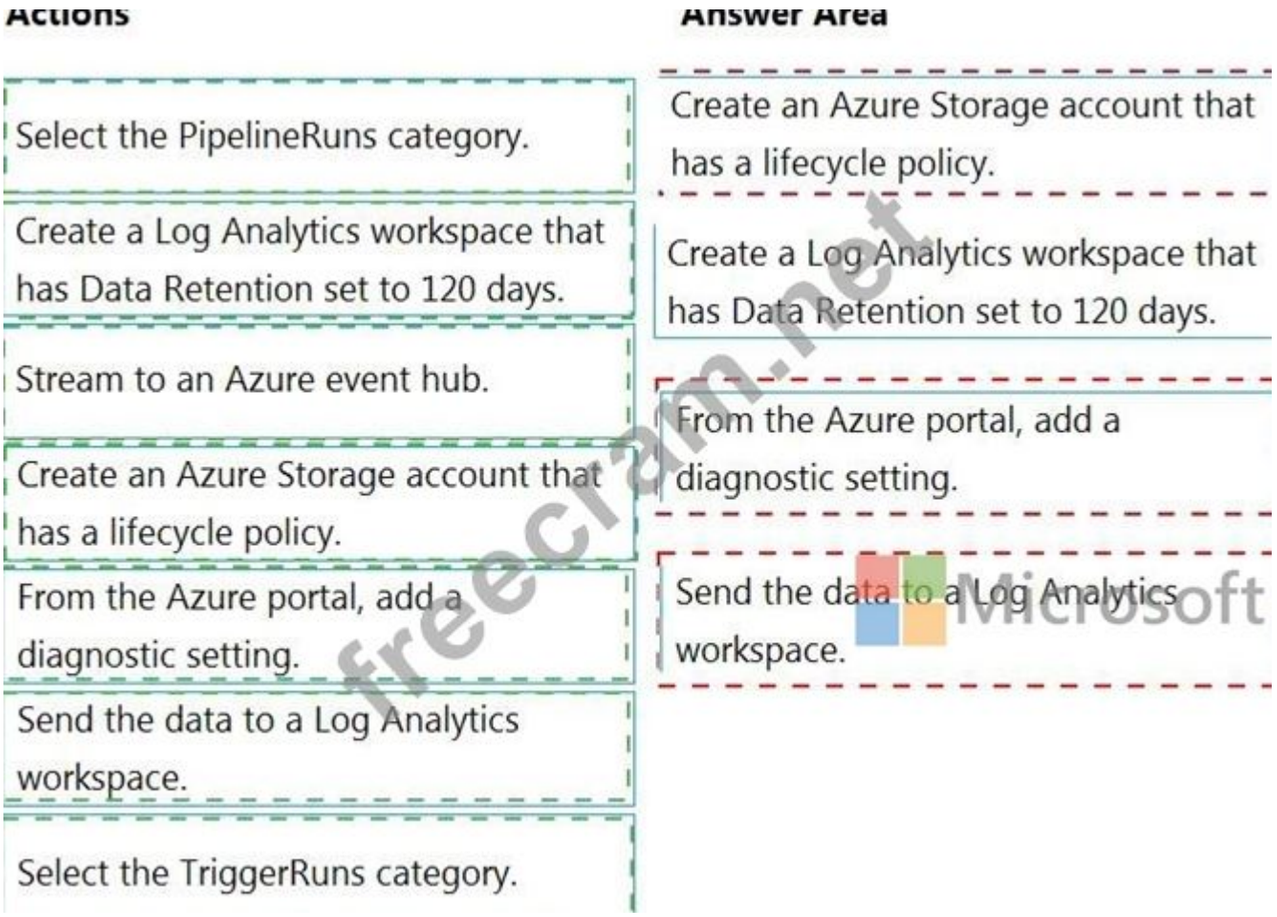
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-overview-what-is

NEW QUESTION: 107
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

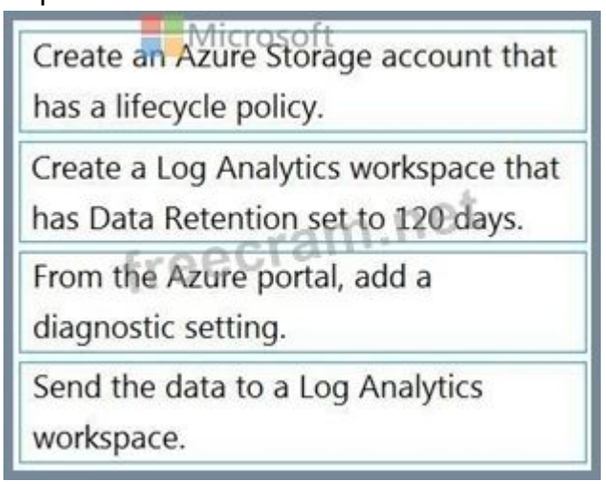
NEW QUESTION: 108
You have an Azure data factory.
You need to ensure that pipeline-run data is retained for 120 days. The solution must ensure that you can query the data by using the Kusto query language.
Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.


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

Answer:



Explanation



Step 1: Create an Azure Storage account that has a lifecycle policy

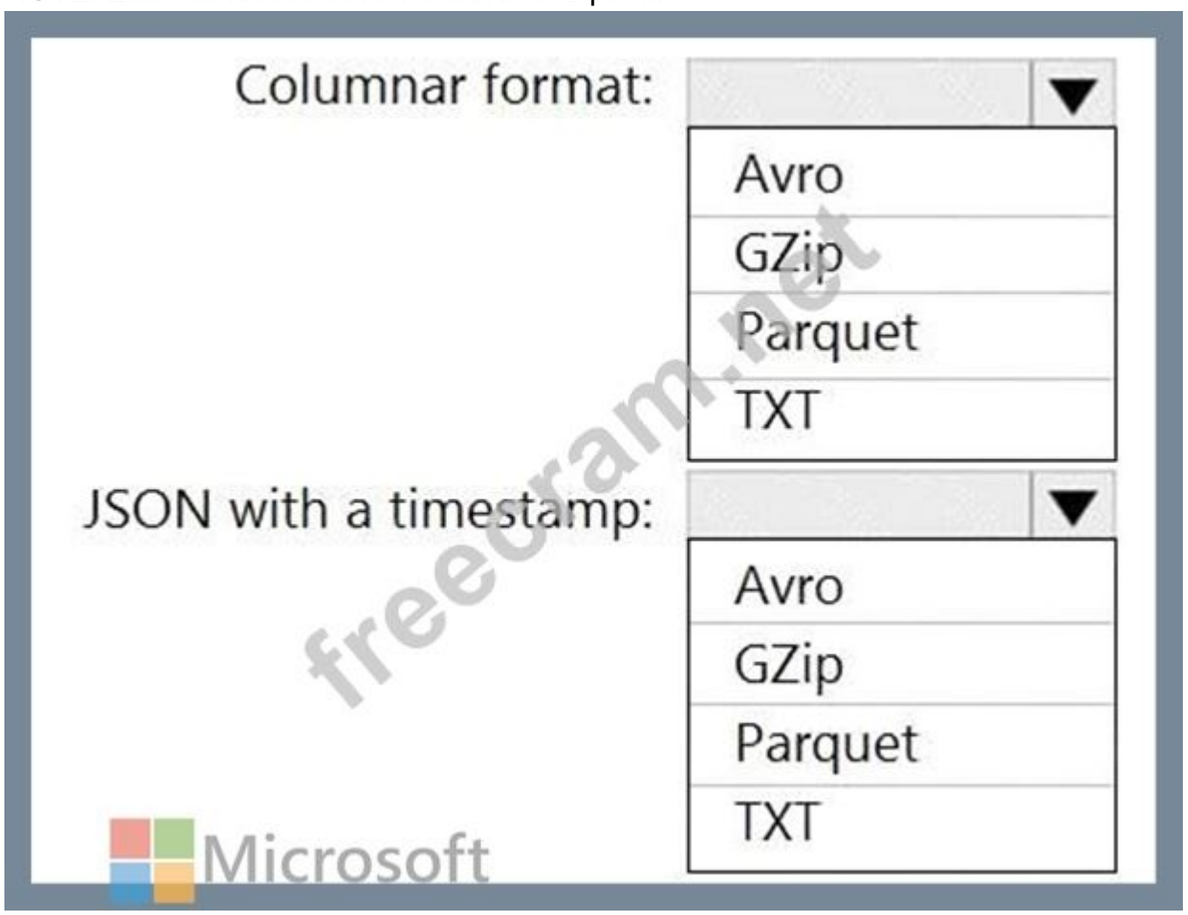
To automate common data management tasks, Microsoft created a solution based on Azure Data Factory. The service, Data Lifecycle Management, makes frequently accessed data available and archives or purges other data according to retention policies. Teams across the company use the service to reduce storage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

Step 2: Create a Log Analytics workspace that has Data Retention set to 120 days.

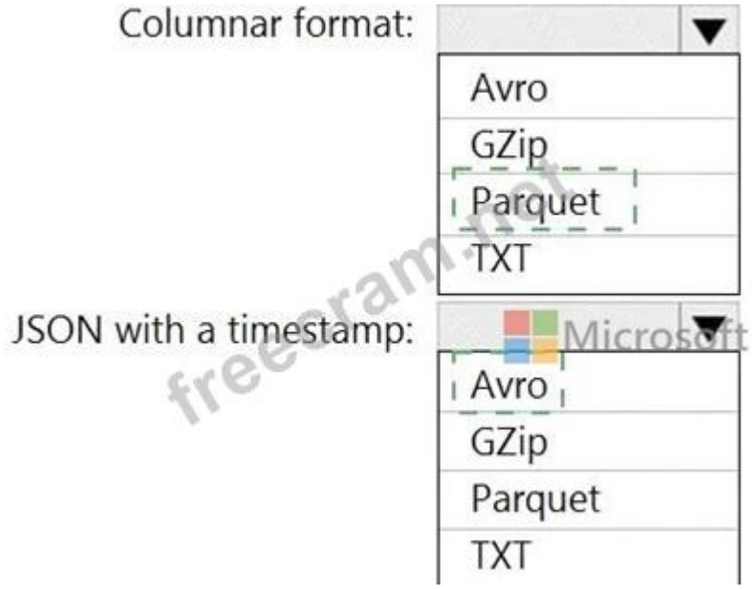
Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time. With Monitor, you can route diagnostic logs for analysis to multiple different targets, such as a Storage Account: Save your diagnostic logs to a storage account for auditing or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

Step 3: From Azure Portal, add a diagnostic setting.
Step 4: Send the data to a log Analytics workspace,
Event Hub: A pipeline that transfers events from services to Azure Data Explorer.
Keeping Azure Data Factory metrics and pipeline-run data.
Configure diagnostic settings and workspace.
Create or add diagnostic settings for your data factory.
In the portal, go to Monitor. Select Settings > Diagnostic settings.
Select the data factory for which you want to set a diagnostic setting.
If no settings exist on the selected data factory, you're prompted to create a setting. Select Turn on diagnostics.
Give your setting a name, select Send to Log Analytics, and then select a workspace from Log Analytics Workspace.
Select Save.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor

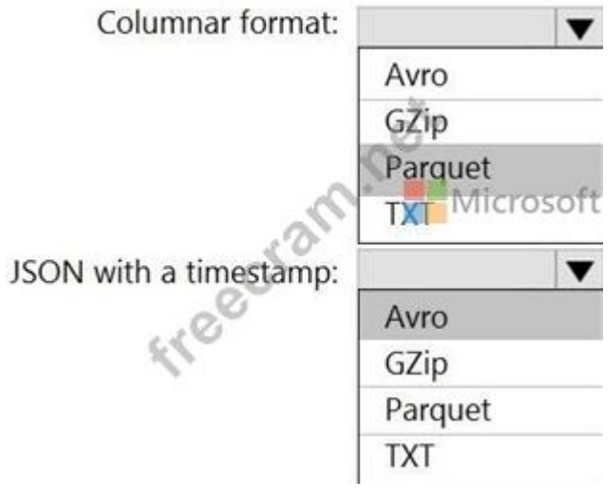
NEW QUESTION: 109
You need to output files from Azure Data Factory.
Which file format should you use for each type of output? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



Explanation



Box 1: Parquet

Parquet stores data in columns, while Avro stores data in a row-based format. By their very nature, column-oriented data stores are optimized for read-heavy analytical workloads, while row-based databases are best for write-heavy transactional workloads.

Box 2: Avro

An Avro schema is created using JSON format.

AVRO supports timestamps.

Note: Azure Data Factory supports the following file formats (not GZip or TXT).

Avro format

Binary format

Delimited text format

Excel format

JSON format

ORC format

Parquet format

XML format
Reference:
https://www.datanami.com/2018/05/16/big-data-file-formats-demystified

NEW QUESTION: 110

You are designing a security model for an Azure Synapse Analytics dedicated SQL pool that will support multiple companies. You need to ensure that users from each company can view only the data of their respective company. Which two objects should you include in the solution? Each correct answer presents part of the solution NOTE: Each correct selection it worth one point.

A. a custom role-based access control (RBAC) role.
B. asymmetric keys
C. a predicate function
D. a column encryption key
E. a security policy
Answer: (SHOW ANSWER)
Reference:
https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-access-control-overview

NEW QUESTION: 111
You have an Azure Data Lake Storage Gen2 account that contains two folders named Folder and Folder2.
You use Azure Data Factory to copy multiple files from Folder1 to Folder2.

```
Operation on target Copy_sks failed: Failure happened on 'Sink' side.
ErrorCode=DelimitedTextMoreColumnsThanDefined,
'Type=Microsoft.DataTransfer.Common.Shared.HybridDeliveryException,
Message=Error found when processing 'Csv/Tsv Format Text' source
'0_2020_11_09_11_43_32.avro' with row number 53: found more columns
than expected column count 27., Source=Microsoft.DataTransfer.Common,'
```


You receive the following error.
What should you do to resolve the error.
A. Add an explicit mapping.
B. Enable fault tolerance to skip incompatible rows.
C. Lower the degree of copy parallelism
D. Change the Copy activity setting to Binary Copy
Answer: A (LEAVE A REPLY)
Reference:
https://knowledge.informatica.com/s/article/Microsoft-Azure-Data-Lake-Store-Gen2-target-file-names-not-gener

NEW QUESTION: 112
You are monitoring an Azure Stream Analytics job.

The Backlogged Input Events count has been 20 for the last hour.
You need to reduce the Backlogged Input Events count.
What should you do?
A. Drop late arriving events from the job.
B. Add an Azure Storage account to the job.
C. Increase the streaming units for the job.
D. Stop the job.
Answer: (SHOW ANSWER)
Explanation
General symptoms of the job hitting system resource limits include:
* If the backlog event metric keeps increasing, it's an indicator that the system resource is constrained (either because of output sink throttling, or high CPU).
Note: Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job: adjust Streaming Units.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-scale-jobs
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-monitoring

NEW QUESTION: 113
You are designing a highly available Azure Data Lake Storage solution that will include geo-zone-redundant storage (GZRS).
You need to monitor for replication delays that can affect the recovery point objective (RPO).
What should you include in the monitoring solution?
A. availability
B. Average Success E2E Latency
C. 5xx: Server Error errors
D. Last Sync Time
Answer: D (LEAVE A REPLY)
Explanation
Because geo-replication is asynchronous, it is possible that data written to the primary region has not yet been written to the secondary region at the time an outage occurs. The Last Sync Time property indicates the last time that data from the primary region was written successfully to the secondary region. All writes made to the primary region before the last sync time are available to be read from the secondary location. Writes made to the primary region after the last sync time property may or may not be available for reads yet.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/last-sync-time-get

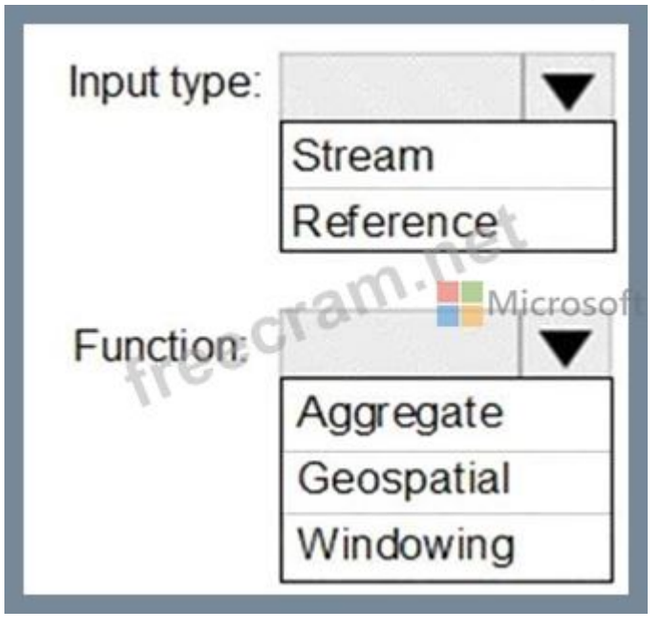
NEW QUESTION: 114

You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account.
The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/.
You need to design a daily Azure Data Factory data load to minimize the data transfer between the two accounts.
Which two configurations should you include in the design? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
A. Delete the files in the destination before loading new data.
B. Filter by the last modified date of the source files.
C. Delete the source files after they are copied.
D. Specify a file naming pattern for the destination.
Answer: (SHOW ANSWER)
Explanation
Copy data from one place to another. The requirements are : 1- need to minimize transfert and 2- need to adapte data to the destination folder structure. Filter on LastModifiedDate will copy everything that have changed since the latest load while minimizing the data transfert. Specifying the file naming pattern allows to copy data at the right place to the destination Data Lake.

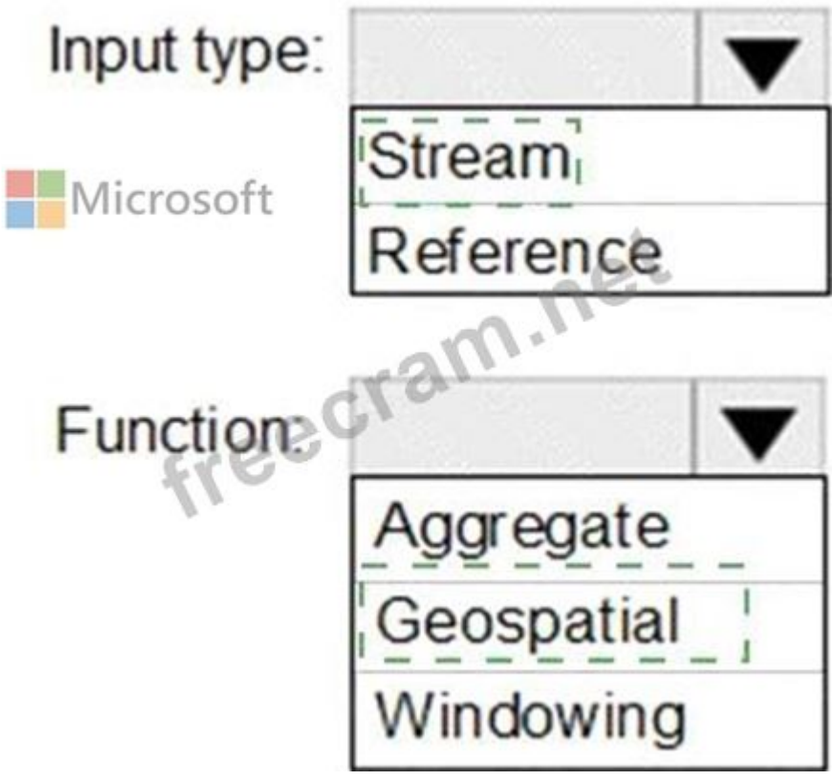
NEW QUESTION: 115
You are designing an Azure Synapse Analytics workspace.
You need to recommend a solution to provide double encryption of all the data at rest.
Which two components should you include in the recommendation? Each coned answer presents part of the solution NOTE: Each correct selection is worth one point.
A. an X509 certificate
B. an RSA key
C. an Azure key vault that has purge protection enabled
D. an Azure virtual network that has a network security group (NSG)
E. an Azure Policy initiative
Answer: (SHOW ANSWER)
Explanation
Synapse workspaces encryption uses existing keys or new keys generated in Azure Key Vault. A single key is used to encrypt all the data in a workspace. Synapse workspaces support RSA 2048 and 3072 byte-sized keys, and RSA-HSM keys.
The Key Vault itself needs to have purge protection enabled.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption

NEW QUESTION: 116
You plan to create a real-time monitoring app that alerts users when a device travels more than 200 meters away from a designated location.

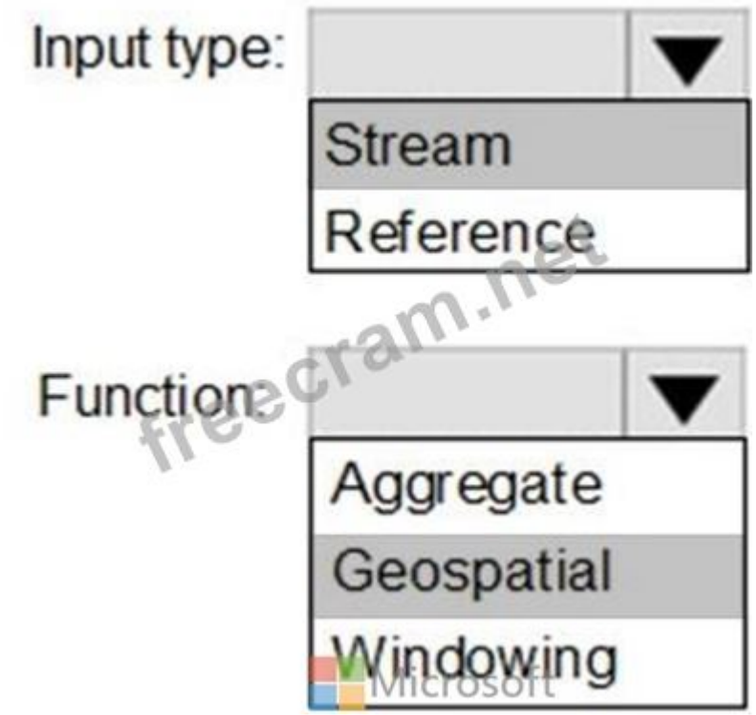
You need to design an Azure Stream Analytics job to process the data for the planned app. The solution must minimize the amount of code developed and the number of technologies used.
What should you include in the Stream Analytics job? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



Explanation



Diagram, table Description automatically generated
Input type: Stream
You can process real-time IoT data streams with Azure Stream Analytics.
Function: Geospatial
With built-in geospatial functions, you can use Azure Stream Analytics to build applications for scenarios such as fleet management, ride sharing, connected cars, and asset tracking.
Note: In a real-world scenario, you could have hundreds of these sensors generating events as a stream. Ideally, a gateway device would run code to push these events to Azure Event Hubs or Azure IoT Hubs.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-get-started-with-azure-stream-analytics
https://docs.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios

NEW QUESTION: 117
You need to design a data retention solution for the Twitter feed data records. The solution must meet the customer sentiment analytics requirements.
Which Azure Storage functionality should you include in the solution?
A. change feed
B. soft delete
C. time-based retention
D. lifecycle management
Answer: (SHOW ANSWER)
Explanation
Scenario: Purge Twitter feed data records that are older than two years.

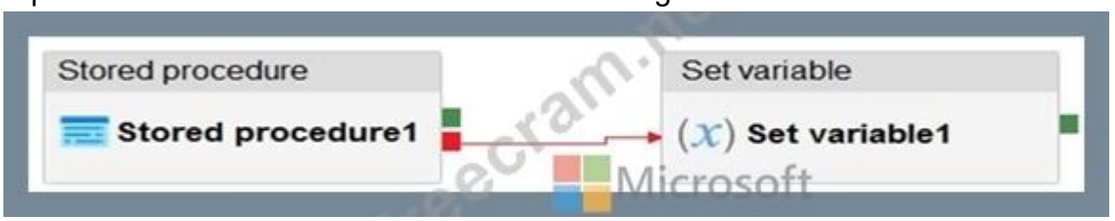
Data sets have unique lifecycles. Early in the lifecycle, people access some data often. But the need for access often drops drastically as the data ages. Some data remains idle in the cloud and is rarely accessed once stored.
Some data sets expire days or months after creation, while other data sets are actively read and modified throughout their lifetimes. Azure Storage lifecycle management offers a rule-based policy that you can use to transition blob data to the appropriate access tiers or to expire data at the end of the data lifecycle.
Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/lifecycle-management-overview

NEW QUESTION: 118
You have two Azure Data Factory instances named ADFdev and ADFprod. ADFdev connects to an Azure DevOps Git repository.
You publish changes from the main branch of the Git repository to ADFdev.
You need to deploy the artifacts from ADFdev to ADFprod.
What should you do first?
A. From ADFdev, modify the Git configuration.
B. From ADFdev, create a linked service.
C. From Azure DevOps, create a release pipeline.
D. From Azure DevOps, update the main branch.
Answer: C (LEAVE A REPLY)
Explanation
In Azure Data Factory, continuous integration and delivery (CI/CD) means moving Data Factory pipelines from one environment (development, test, production) to another.
Note:
The following is a guide for setting up an Azure Pipelines release that automates the deployment of a data factory to multiple environments.
In Azure DevOps, open the project that's configured with your data factory.
On the left side of the page, select Pipelines, and then select Releases.
Select New pipeline, or, if you have existing pipelines, select New and then New release pipeline.
In the Stage name box, enter the name of your environment.
Select Add artifact, and then select the git repository configured with your development data factory.
Select the publish branch of the repository for the Default branch. By default, this publish branch is adf_publish.
Select the Empty job template.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment
Topic 1, Litware, inc.
Case study
This is a case study. Case studies are not timed separately. You can use as much exam time as you would like to complete each case. However, there may be additional case studies and sections on this

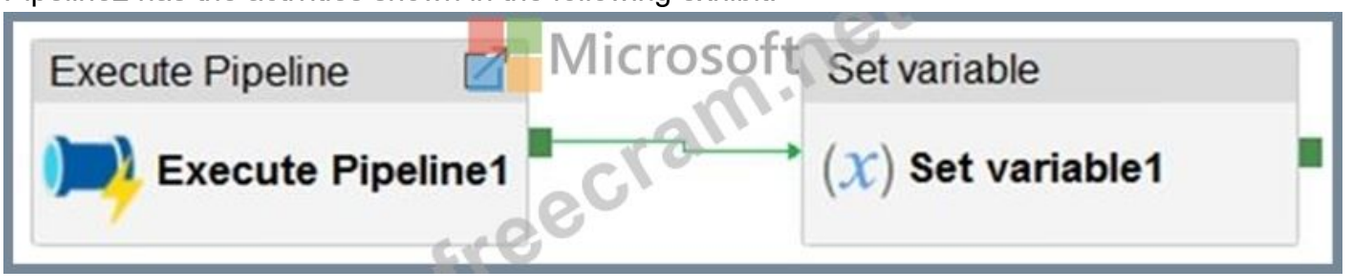
exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.
To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other questions in this case study.
At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.
To start the case study
To display the first question in this case study, click the button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an All Information tab, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the button to return to the question.
Overview
Litware, Inc. owns and operates 300 convenience stores across the US. The company sells a variety of packaged foods and drinks, as well as a variety of prepared foods, such as sandwiches and pizzas.
Litware has a loyalty club whereby members can get daily discounts on specific items by providing their membership number at checkout.
Litware employs business analysts who prefer to analyze data by using Microsoft Power BI, and data scientists who prefer analyzing data in Azure Databricks notebooks.
Requirements
Business Goals
Litware wants to create a new analytics environment in Azure to meet the following requirements:
See inventory levels across the stores. Data must be updated as close to real time as possible.
Execute ad hoc analytical queries on historical data to identify whether the loyalty club discounts increase sales of the discounted products.
Every four hours, notify store employees about how many prepared food items to produce based on historical demand from the sales data.
Technical Requirements
Litware identifies the following technical requirements:
Minimize the number of different Azure services needed to achieve the business goals.
Use platform as a service (PaaS) offerings whenever possible and avoid having to provision virtual machines that must be managed by Litware.
Ensure that the analytical data store is accessible only to the company's on-premises network and Azure services.
Use Azure Active Directory (Azure AD) authentication whenever possible.
Use the principle of least privilege when designing security.
Stage Inventory data in Azure Data Lake Storage Gen2 before loading the data into the analytical data store. Litware wants to remove transient data from Data Lake Storage once the data is no longer in use.

Files that have a modified date that is older than 14 days must be removed.
Limit the business analysts' access to customer contact information, such as phone numbers, because this type of data is not analytically relevant.
Ensure that you can quickly restore a copy of the analytical data store within one hour in the event of corruption or accidental deletion.
Planned Environment
Litware plans to implement the following environment:
The application development team will create an Azure event hub to receive real-time sales data, including store number, date, time, product ID, customer loyalty number, price, and discount amount, from the point of sale (POS) system and output the data to data storage in Azure.
Customer data, including name, contact information, and loyalty number, comes from Salesforce, a SaaS application, and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
Product data, including product ID, name, and category, comes from Salesforce and can be imported into Azure once every eight hours. Row modified dates are not trusted in the source table.
Daily inventory data comes from a Microsoft SQL server located on a private network.
Litware currently has 5 TB of historical sales data and 100 GB of customer data. The company expects approximately 100 GB of new data per month for the next year.
Litware will build a custom application named FoodPrep to provide store employees with the calculation results of how many prepared food items to produce every four hours.
Litware does not plan to implement Azure ExpressRoute or a VPN between the on-premises network and Azure.

NEW QUESTION: 119
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: (SHOW ANSWER)

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:

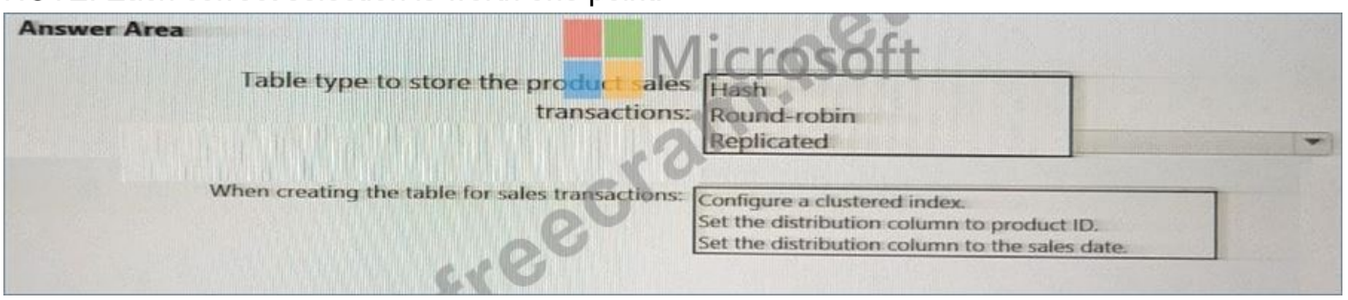
https://datasavvy.me/category/azure-data-factory/

NEW QUESTION: 120

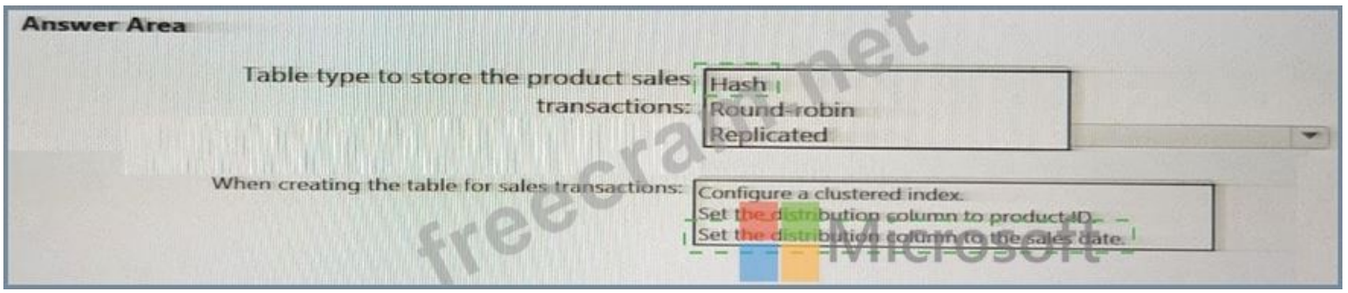
You need to design a data storage structure for the product sales transactions. The solution must meet the sales transaction dataset requirements.

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

NOTE: Each correct selection is worth one point.

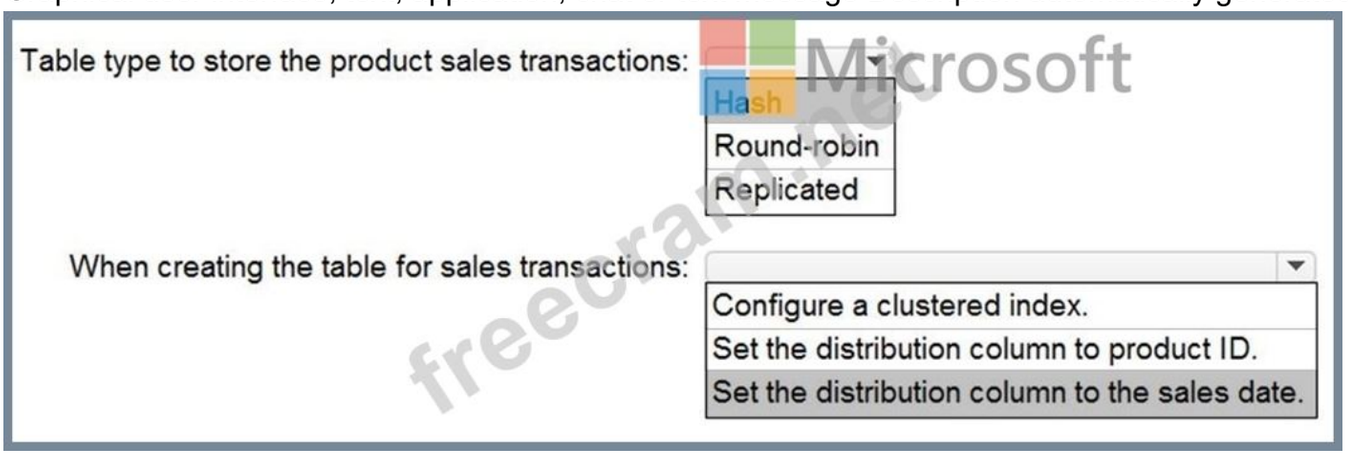


Answer:



Explanation

Graphical user interface, text, application, chat or text message Description automatically generated



Box 1: Hash

Scenario:

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

A hash distributed table can deliver the highest query performance for joins and aggregations on large tables.

Box 2: Set the distribution column to the sales date.

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

Reference:

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

NEW QUESTION: 121

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

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You are designing an Azure Stream Analytics solution that will analyze Twitter data.

You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.

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

Does this meet the goal?

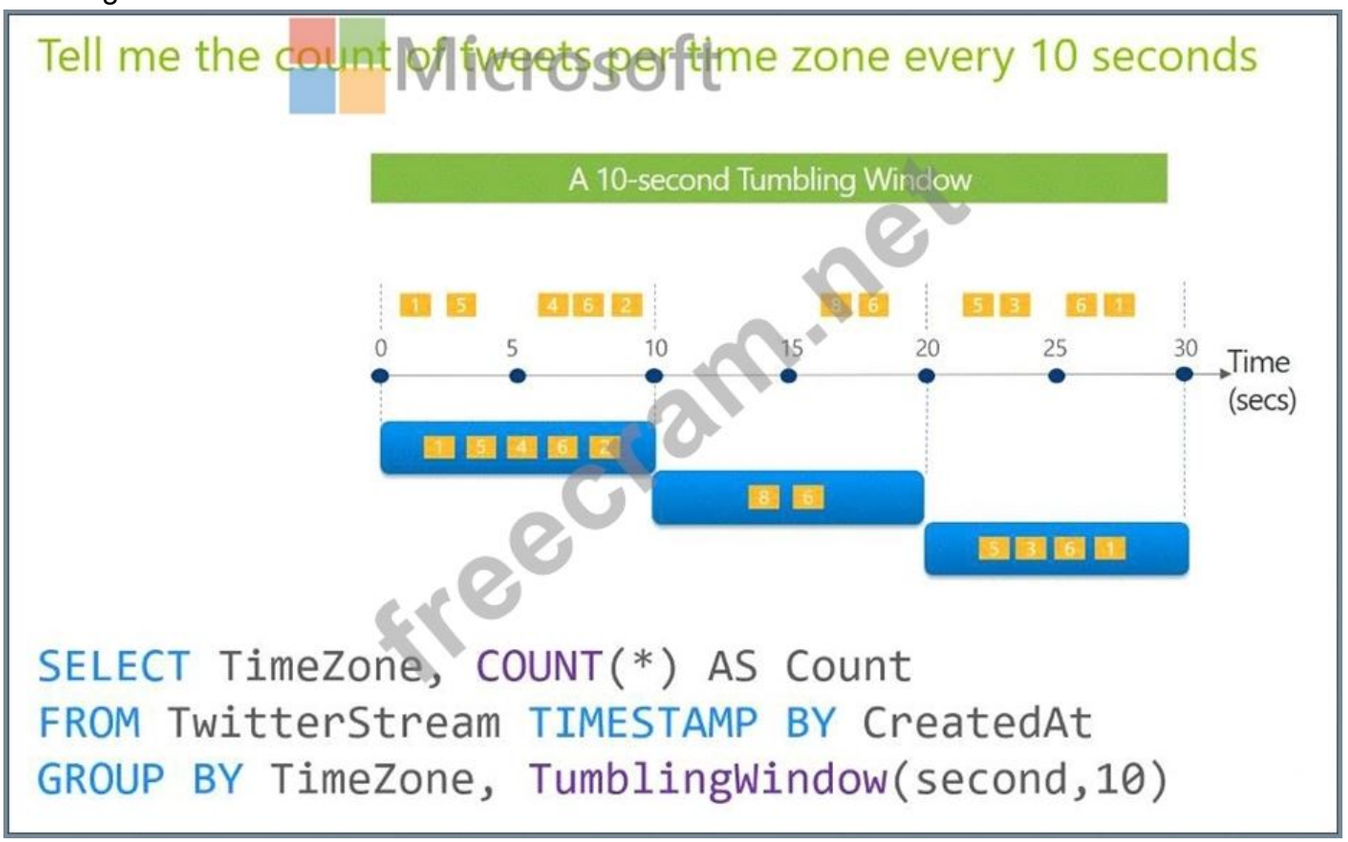
A. Yes

B. No

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: 122

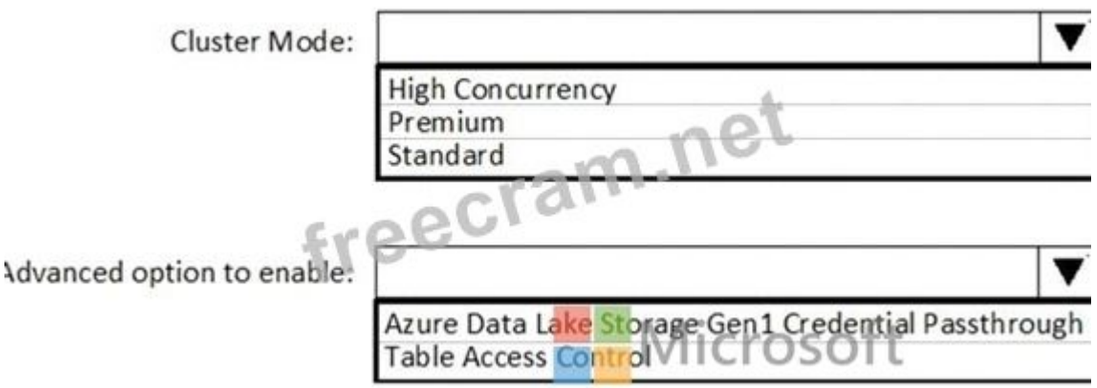
You need to design a data retention solution for the Twitter teed data records. The solution must meet the customer sentiment analytics requirements.

Which Azure Storage functionality should you include in the solution?

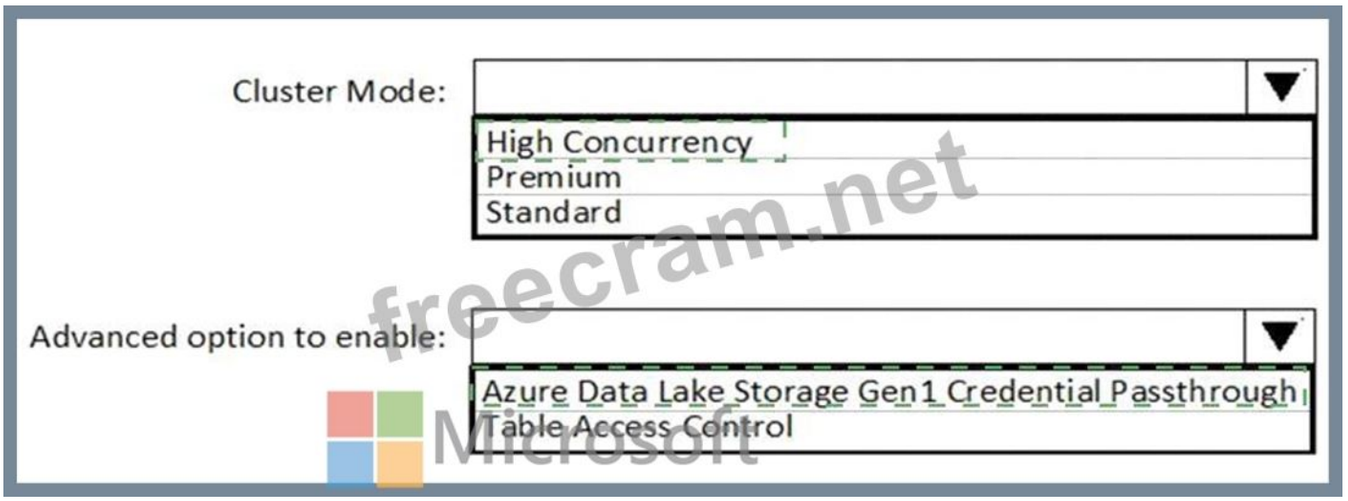
A. change feed
B. soft delete
C. time-based retention
D. lifecycle management
Answer: (SHOW ANSWER)

NEW QUESTION: 123
You have a C# application that process data from an Azure IoT hub and performs complex transformations.
You need to replace the application with a real-time solution. The solution must reuse as much code as possible from the existing application.
A. Azure Databricks
B. Azure Event Grid
C. Azure Stream Analytics
D. Azure Data Factory
Answer: (SHOW ANSWER)
Explanation
Azure Stream Analytics on IoT Edge empowers developers to deploy near-real-time analytical intelligence closer to IoT devices so that they can unlock the full value of device-generated data. UDF are available in C# for IoT Edge jobs Azure Stream Analytics on IoT Edge runs within the Azure IoT Edge framework. Once the job is created in Stream Analytics, you can deploy and manage it using IoT Hub.
References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-edge

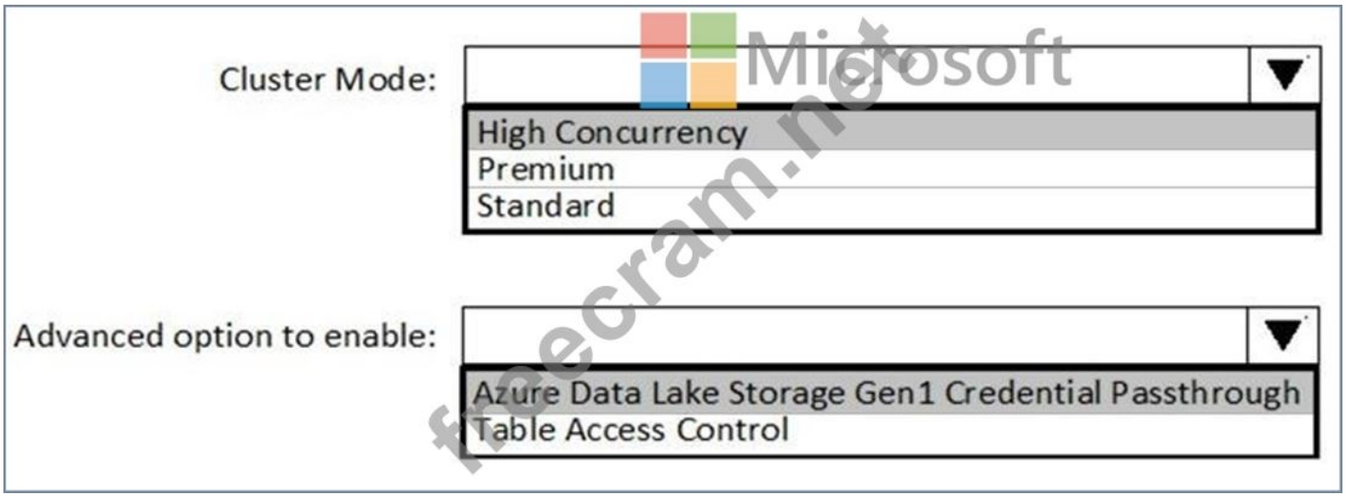
NEW QUESTION: 124
You need to implement an Azure Databricks cluster that automatically connects to Azure Data Lake Storage Gen2 by using Azure Active Directory (Azure AD) integration.
How should you configure the new cluster? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



Explanation



Box 1: High Concurrency

Enable Azure Data Lake Storage credential passthrough for a high-concurrency cluster.

Incorrect:

Support for Azure Data Lake Storage credential passthrough on standard clusters is in Public Preview. Standard clusters with credential passthrough are supported on Databricks Runtime 5.5 and above and are limited to a single user.

Box 2: Azure Data Lake Storage Gen1 Credential Passthrough

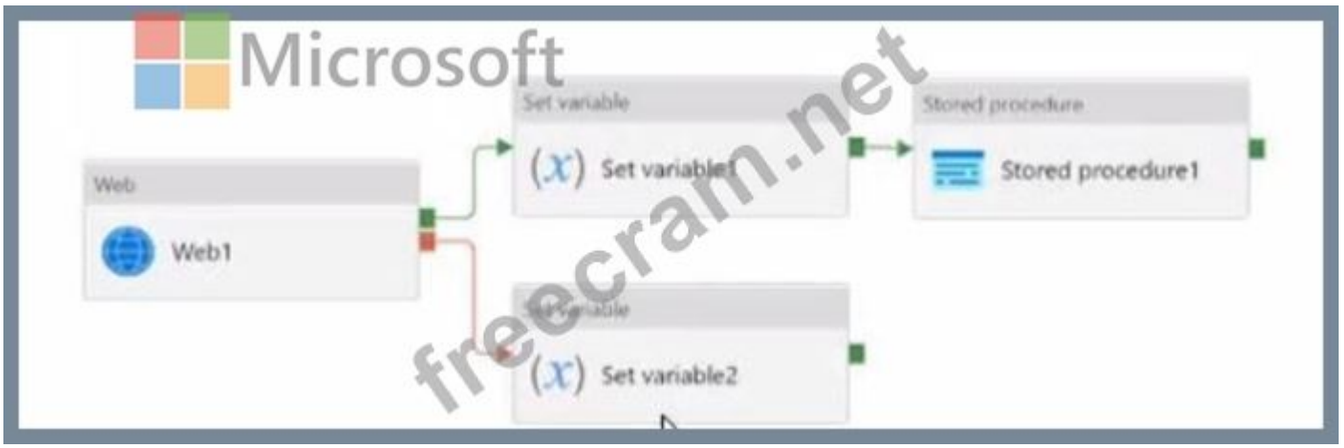
You can authenticate automatically to Azure Data Lake Storage Gen1 and Azure Data Lake Storage Gen2 from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that you use to log into Azure Databricks. When you enable your cluster for Azure Data Lake Storage credential passthrough, commands that you run on that cluster can read and write data in Azure Data Lake Storage without requiring you to configure service principal credentials for access to storage.

References:

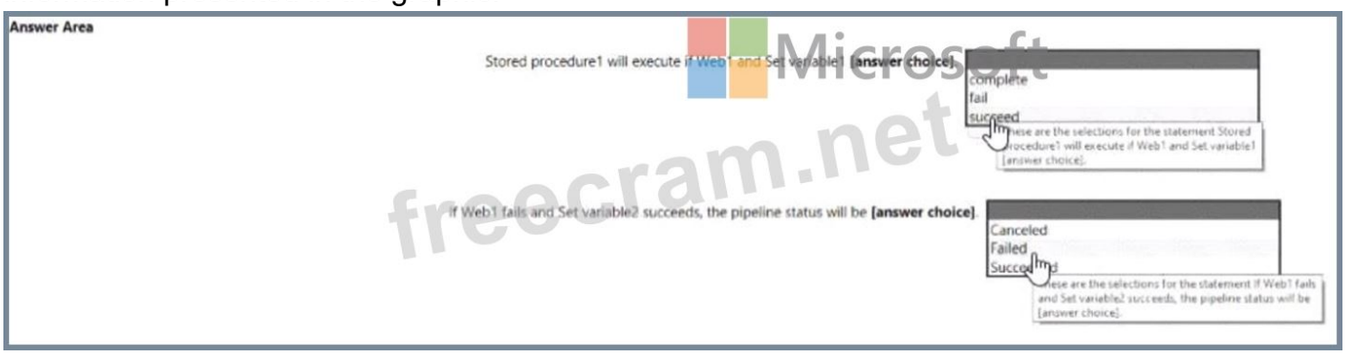
https://docs.azuredatabricks.net/spark/latest/data-sources/azure/adls-passthrough.html

NEW QUESTION: 125

You have an Azure Data Factory pipeline that has the activity 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.



Answer:

Explanation

NEW QUESTION: 126

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>)

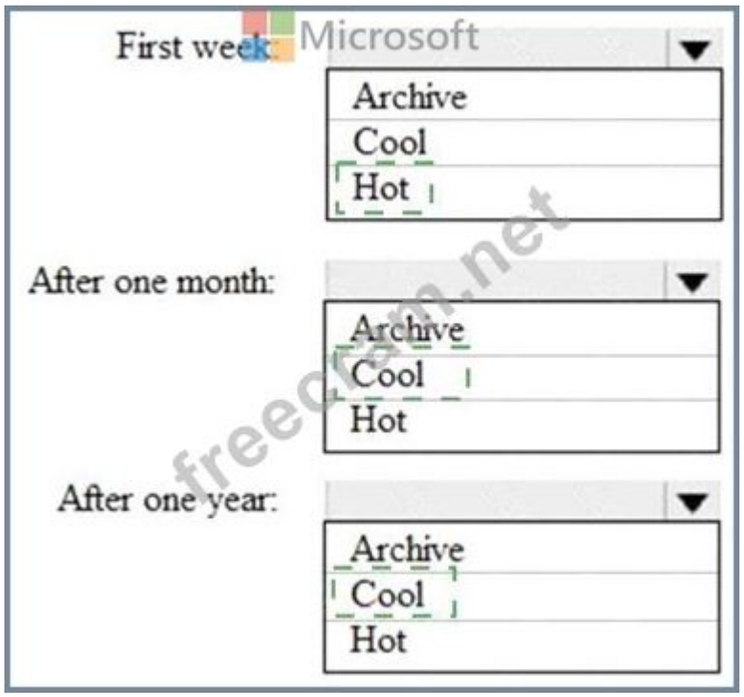
D. \\(<Language>)

Answer: (SHOW ANSWER)

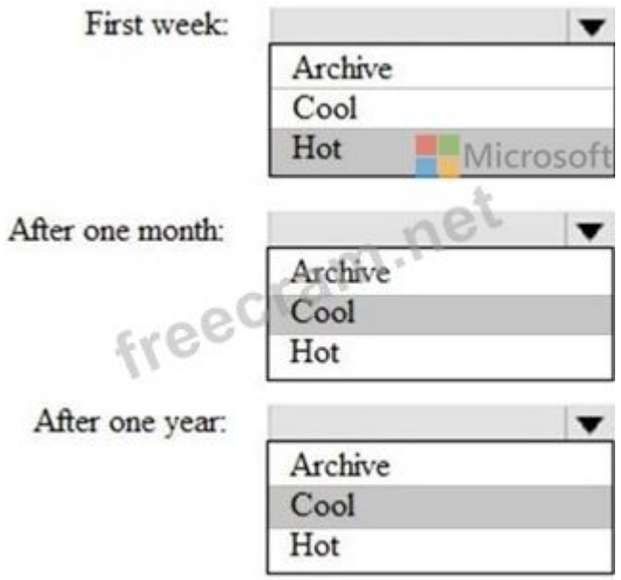
Explanation
To change the language in Databricks' cells to either Scala, SQL, Python or R, prefix the cell with '%', followed by the language.
%python //or r, scala, sql
Reference:
https://www.theta.co.nz/news-blogs/tech-blog/enhancing-digital-twins-part-3-predictive-maintenance-with-azure

NEW QUESTION: 127
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