Actual exam question from Microsoft's DP-203

Question #: 1

Topic #: 1

[All DP-203 Questions]

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

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

You need to alter the table to meet the following requirements:

- ✑ Ensure that users can identify the current manager of employees.
- ✑ Support creating an employee reporting hierarchy for your entire company.
- ✑ Provide fast lookup of the managers' attributes such as name and job title.

Which column should you add to the table?

A. [ManagerEmployeeID] [smallint] NULL

B. [ManagerEmployeeKey] [smallint] NULL

C. [ManagerEmployeeKey] [int] NULL

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 2

Topic #: 1

[All DP-203 Questions]

You have an Azure Synapse workspace named MyWorkspace that contains an Apache Spark database named mytestdb.

You run the following command in an Azure Synapse Analytics Spark pool in MyWorkspace.

CREATE TABLE mytestdb.myParquetTable(

EmployeeID int,

EmployeeName string,

EmployeeStartDate date)

USING Parquet -

You then use Spark to insert a row into mytestdb.myParquetTable. The row contains the following data.

EmployeeName	EmployeeID	EmployeeStartDate
Alice	24	2020-01-25

One minute later, you execute the following query from a serverless SQL pool in MyWorkspace.

SELECT EmployeeID -

FROM mytestdb.dbo.myParquetTable

WHERE EmployeeName = 'Alice';

What will be returned by the query?

A. 24

B. an error

C. a null value

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 3

Topic #: 1

[All DP-203 Questions]

DRAG DROP -

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

- ✑ Is partitioned by month
- ✑ Contains one billion rows
- ✑ Has clustered columnstore index

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

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

Select and Place:

Actions

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

Answer Area

Show Suggested Answer

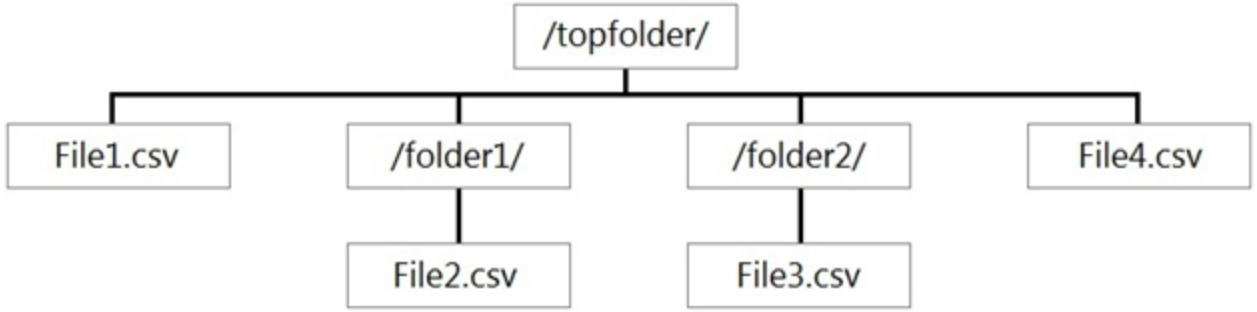
Actual exam question from Microsoft's DP-203

Question #: 4

Topic #: 1

[All DP-203 Questions]

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



You create an external table named ExtTable that has LOCATION='/topfolder/'.

When you query ExtTable by using an Azure Synapse Analytics serverless SQL pool, which files are returned?

A. File2.csv and File3.csv only

B. File1.csv and File4.csv only

C. File1.csv, File2.csv, File3.csv, and File4.csv

D. File1.csv only

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 5

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

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

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

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 6

Topic #: 1

[All DP-203 Questions]

You are designing the folder structure for an Azure Data Lake Storage Gen2 container.

Users will query data by using a variety of services including Azure Databricks and Azure Synapse Analytics serverless SQL pools. The data will be secured by subject area. Most queries will include data from the current year or current month.

Which folder structure should you recommend to support fast queries and simplified folder security?

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

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

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

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 7

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

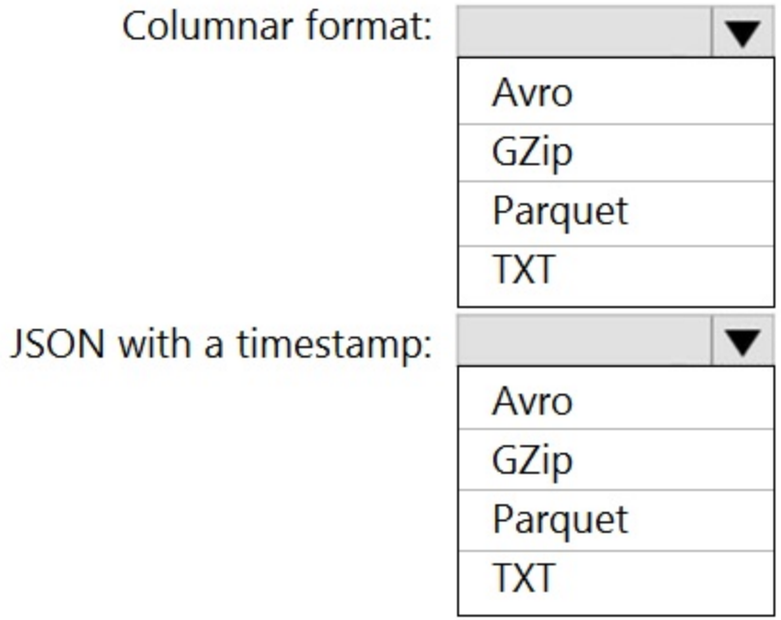
You need to output files from Azure Data Factory.

Which file format should you use for each type of output? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area



Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 8

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

You use Azure Data Factory to prepare data to be queried by Azure Synapse Analytics serverless SQL pools.

Files are initially ingested into an Azure Data Lake Storage Gen2 account as 10 small JSON files. Each file contains the same data attributes and data from a subsidiary of your company.

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

✑ Provide the fastest possible query times.

✑ Automatically infer the schema from the underlying files.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Copy behavior:

- Flatten hierarchy
- Merge files
- Preserve hierarchy

Sink file type:

- CSV
- JSON
- Parquet
- TXT

Show Suggested Answer

Actual exam question from Microsoft's DP-203

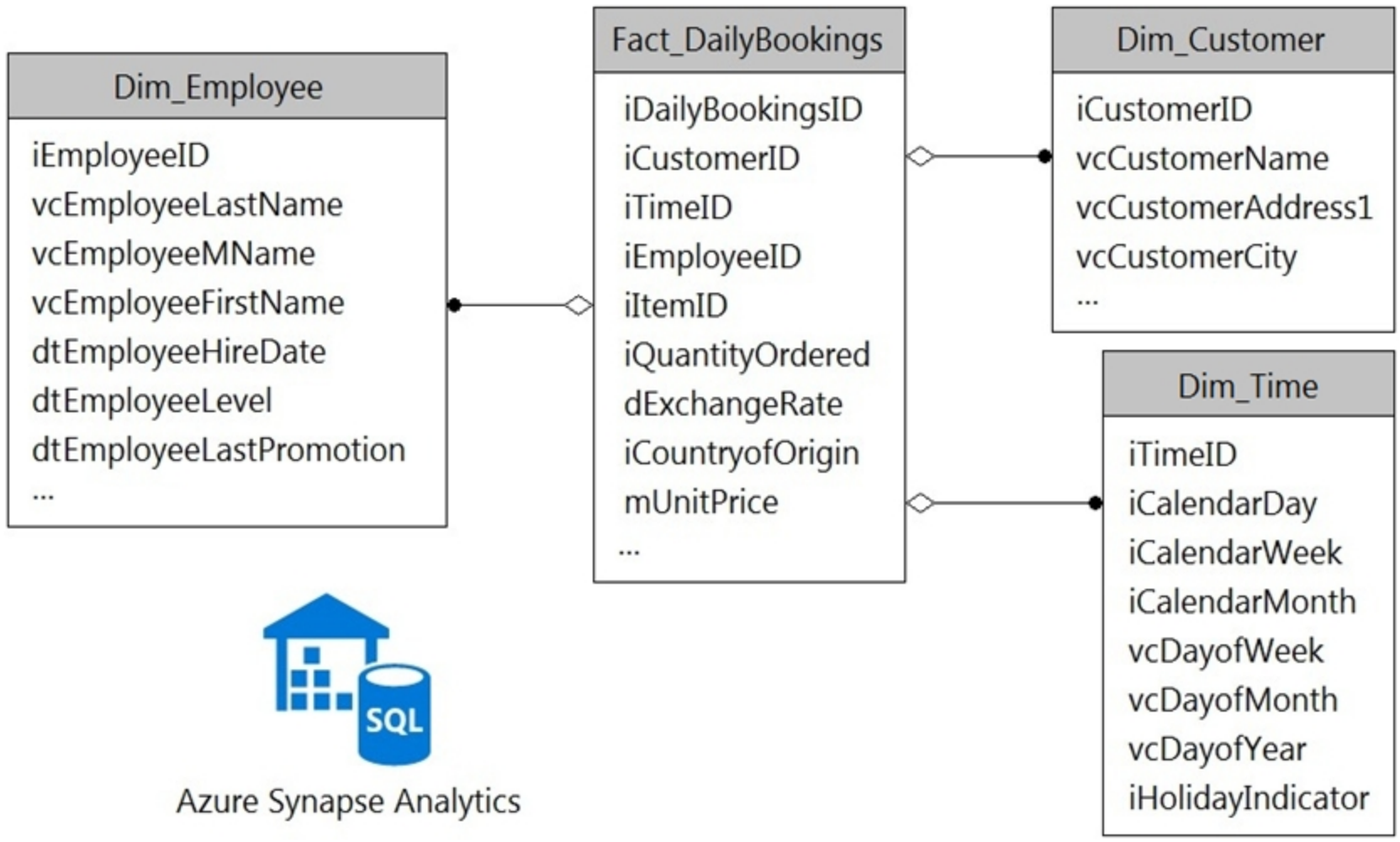
Question #: 9

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

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



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

Which type of table should you use for each table? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Dim_Customer: ▼
- Hash distributed
- Round-robin
- Replicated

Dim_Employee: ▼
- Hash distributed
- Round-robin
- Replicated

Dim_Time: ▼
- Hash distributed
- Round-robin
- Replicated

Fact_DailyBookings: ▼
- Hash distributed
- Round-robin
- Replicated

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 10

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

You have an Azure Data Lake Storage Gen2 container.

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

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

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

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

NOTE: Each correct selection is worth one point

Hot Area:

Answer Area

Five-year-old data:

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

Seven-year-old data:

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 11

Topic #: 1

[All DP-203 Questions]

DRAG DROP -

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

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

NOTE: Each correct selection is worth one point.

Select and Place:

Values

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

Answer Area

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 12

Topic #: 1

[All DP-203 Questions]

You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements:

- ✑ Can return an employee record from a given point in time.
- ✑ Maintains the latest employee information.
- ✑ Minimizes query complexity.

How should you model the employee data?

A. as a temporal table

B. as a SQL graph table

C. as a degenerate dimension table

D. as a Type 2 slowly changing dimension (SCD) table

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 13

Topic #: 1

[All DP-203 Questions]

You have an enterprise-wide Azure Data Lake Storage Gen2 account. The data lake is accessible only through an Azure virtual network named VNET1.

You are building a SQL pool in Azure Synapse that will use data from the data lake.

Your company has a sales team. All the members of the sales team are in an Azure Active Directory group named Sales. POSIX controls are used to assign the Sales group access to the files in the data lake.

You plan to load data to the SQL pool every hour.

You need to ensure that the SQL pool can load the sales data from the data lake.

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

NOTE: Each area selection is worth one point.

A. Add the managed identity to the Sales group.

B. Use the managed identity as the credentials for the data load process.

C. Create a shared access signature (SAS).

D. Add your Azure Active Directory (Azure AD) account to the Sales group.

E. Use the shared access signature (SAS) as the credentials for the data load process.

F. Create a managed identity.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 14

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

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

Name	Role
User1	Server admin
User2	db_datereader

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

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

Results Messages

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

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

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

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 15

Topic #: 1

[All DP-203 Questions]

You have an enterprise data warehouse in Azure Synapse Analytics.

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

The external table has three columns.

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

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

A.

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

B.

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

C.

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

D.

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 16

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

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

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

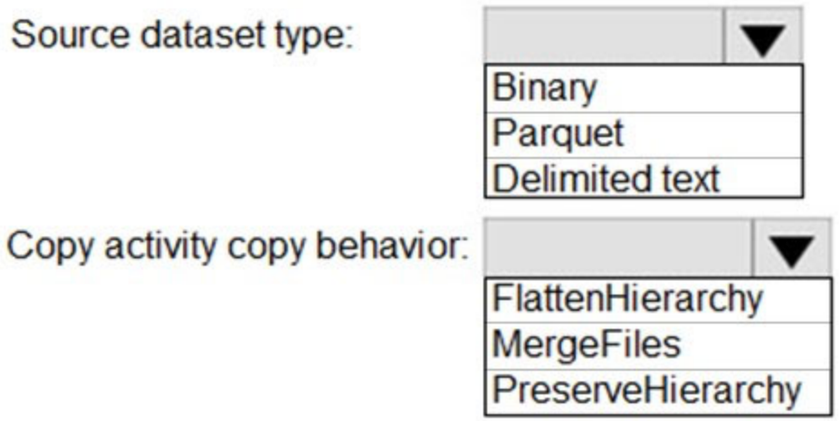
- ✑ No transformations must be performed.
- ✑ The original folder structure must be retained.
- ✑ Minimize time required to perform the copy activity.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area



Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 17

Topic #: 1

[All DP-203 Questions]

You have an Azure Data Lake Storage Gen2 container that contains 100 TB of data.

You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region. The solution must minimize costs.

Which type of data redundancy should you use?

A. geo-redundant storage (GRS)

B. read-access geo-redundant storage (RA-GRS)

C. zone-redundant storage (ZRS)

D. locally-redundant storage (LRS)

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 18

Topic #: 1

[All DP-203 Questions]

You plan to implement an Azure Data Lake Gen 2 storage account.

You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs.

Which type of replication should you use for the storage account?

A. geo-redundant storage (GRS)

B. geo-zone-redundant storage (GZRS)

C. locally-redundant storage (LRS)

D. zone-redundant storage (ZRS)

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 19

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

You have a SQL pool in Azure Synapse.

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

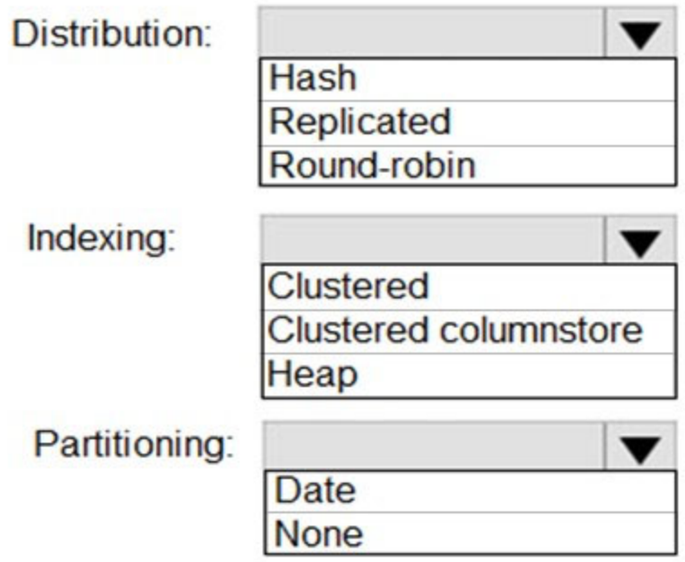
You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area



Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 20

Topic #: 1

[All DP-203 Questions]

You are designing a fact table named FactPurchase in an Azure Synapse Analytics dedicated SQL pool. The table contains purchases from suppliers for a retail store. FactPurchase will contain the following columns.

Name	Data type	Nullable
PurchaseKey	Bigint	No
DateKey	Int	No
SupplierKey	Int	No
StockItemKey	Int	No
PurchaseOrderID	Int	Yes
OrderedQuantity	Int	No
OrderedOuters	Int	No
ReceivedOuters	Int	No
Package	Nvarchar(50)	No
IsOrderFinalized	Bit	No
LineageKey	Int	No

FactPurchase will have 1 million rows of data added daily and will contain three years of data.

Transact-SQL queries similar to the following query will be executed daily.

```
SELECT -
SupplierKey, StockItemKey, IsOrderFinalized, COUNT(*)

FROM FactPurchase -

WHERE DateKey >= 20210101 -

AND DateKey <= 20210131 -
GROUP By SupplierKey, StockItemKey, IsOrderFinalized
```

Which table distribution will minimize query times?

A. replicated

B. hash-distributed on PurchaseKey

C. round-robin

D. hash-distributed on IsOrderFinalized

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 21

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

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

The data contains the following columns.

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

EventCategory: DimChannel / DimDate / DimEvent / FactEvents

ChannelGrouping: DimChannel / DimDate / DimEvent / FactEvents

TotalEvents: DimChannel / DimDate / DimEvent / FactEvents

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 22
Topic #: 1
[All DP-203 Questions]

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly.
Solution: You convert the files to compressed delimited text files.
Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 23

Topic #: 1

[All DP-203 Questions]

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

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

You have an Azure Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You copy the files to a table that has a columnstore index.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 24

Topic #: 1

[All DP-203 Questions]

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

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

You have an Azure Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You modify the files to ensure that each row is more than 1 MB.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 25
Topic #: 1
[All DP-203 Questions]

You build a data warehouse in an Azure Synapse Analytics dedicated SQL pool.
Analysts write a complex SELECT query that contains multiple JOIN and CASE statements to transform data for use in inventory reports. The inventory reports will use the data and additional WHERE parameters depending on the report. The reports will be produced once daily.
You need to implement a solution to make the dataset available for the reports. The solution must minimize query times.
What should you implement?

A. an ordered clustered columnstore index

B. a materialized view

C. result set caching

D. a replicated table

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 26

Topic #: 1

[All DP-203 Questions]

You have an Azure Synapse Analytics workspace named WS1 that contains an Apache Spark pool named Pool1.

You plan to create a database named DB1 in Pool1.

You need to ensure that when tables are created in DB1, the tables are available automatically as external tables to the built-in serverless SQL pool.

Which format should you use for the tables in DB1?

A. CSV

B. ORC

C. JSON

D. Parquet

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 27
Topic #: 1
[All DP-203 Questions]

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

A. Azure Event Hubs

B. Azure Data Factory

C. Azure Stream Analytics

D. Azure Databricks

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 28
Topic #: 1
[All DP-203 Questions]

You plan to implement an Azure Data Lake Storage Gen2 container that will contain CSV files. The size of the files will vary based on the number of events that occur per hour.
File sizes range from 4 KB to 5 GB.
You need to ensure that the files stored in the container are optimized for batch processing.
What should you do?

A. Convert the files to JSON

B. Convert the files to Avro

C. Compress the files

D. Merge the files

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 29

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

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

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

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 30

Topic #: 1

[All DP-203 Questions]

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

- TransactionType: 40 million rows per transaction type
- CustomerSegment: 4 million per customer segment
- TransactionMonth: 65 million rows per month

AccountType: 500 million per account type

▪

You have the following query requirements:

- Analysts will most commonly analyze transactions for a given month.
- Transactions analysis will typically summarize transactions by transaction type, customer segment, and/or account type

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

On which column should you recommend partitioning the table?

A. CustomerSegment

B. AccountType

C. TransactionType

D. TransactionMonth

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 31

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

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

Type	Designated retention period
Application	360 days
Infrastructure	60 days

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

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

- Automatically deletes the logs at the end of each retention period
- Minimizes storage costs

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

To minimize storage costs:

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

To delete logs automatically:

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 32
Topic #: 1
[All DP-203 Questions]

You plan to ingest streaming social media data by using Azure Stream Analytics. The data will be stored in files in Azure Data Lake Storage, and then consumed by using Azure Databricks and PolyBase in Azure Synapse Analytics.
You need to recommend a Stream Analytics data output format to ensure that the queries from Databricks and PolyBase against the files encounter the fewest possible errors. The solution must ensure that the files can be queried quickly and that the data type information is retained.
What should you recommend?

A. JSON

B. Parquet

C. CSV

D. Avro

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 33
Topic #: 1
[All DP-203 Questions]

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

A. Insert the data from stg.Sales into dbo.Sales.

B. Switch the first partition from dbo.Sales to stg.Sales.

C. Switch the first partition from stg.Sales to dbo.Sales.

D. Update dbo.Sales from stg.Sales.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 34

Topic #: 1

[All DP-203 Questions]

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 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-test description of the supplier company
SupplierCategory	Category of goods provided by the supplier company

Which three additional columns should you add to the data to create a Type 2 SCD? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. surrogate primary key

B. effective start date

C. business key

D. last modified date

E. effective end date

F. foreign key

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 35

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

You have a Microsoft SQL Server database that uses a third normal form schema.

You plan to migrate the data in the database to a star schema in an Azure Synapse Analytics dedicated SQL pool.

You need to design the dimension tables. The solution must optimize read operations.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Transform data for the dimension tables by:

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

For the primary key columns in the dimension tables, use:

- New IDENTITY columns
- A new computed column
- The business key column from the source sys

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 36
Topic #: 1
[All DP-203 Questions]

HOTSPOT -

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

✑ ProductID
✑ ItemPrice
✑ LineTotal
✑ Quantity
✑ StoreID
✑ Minute
✑ Month
✑ Hour

Year -
·
✑ Day

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

df.write

.bucketBy
.partitionBy
.range
.sortBy

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

.mode("append")

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 37

Topic #: 1

[All DP-203 Questions]

You are designing a partition strategy for a fact table in an Azure Synapse Analytics dedicated SQL pool. The table has the following specifications:

✑ Contain sales data for 20,000 products.

Use hash distribution on a column named ProductID.

▪

✑ Contain 2.4 billion records for the years 2019 and 2020.

Which number of partition ranges provides optimal compression and performance for the clustered columnstore index?

A. 40

B. 240

C. 400

D. 2,400

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 38

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

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

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

```
CREATE TABLE [DBO].[DimProduct](
        [ProductKey] [int] IDENTITY(1,1) NOT NULL,
        [ProductSourceID] [int] NOT NULL,
        [ProductName] [nvarchar](100) NOT NULL,
        [ProductNumber] [nvarchar](25) NOT NULL,
        [Color] [nvarchar](15) NULL,
        [Size] [nvarchar](5) NULL,
        [Weight] [decimal](8, 2) NULL,
        [ProductCategory] [nvarchar](100) NULL,
        [SellStartDate] [date] NOT NULL,
        [SellEndDate] [date] NULL,
        [RowInsertedDateTime] [datetime] NOT NULL,
        [RowUpdatedDateTime] [datetime] NOT NULL,
        [ETLAuditID] [int] NOT NULL
)
```

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

- Type 0
- Type 1
- Type 2

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

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 39

Topic #: 1

[All DP-203 Questions]

You are designing a fact table named FactPurchase in an Azure Synapse Analytics dedicated SQL pool. The table contains purchases from suppliers for a retail store.

FactPurchase will contain the following columns.

Name	Data type	Nullable
PurchaseKey	Bigint	No
DateKey	Int	No
SupplierKey	Int	No
StockItemKey	Int	No
PurchaseOrderID	Int	Yes
OrderedQuantity	Int	No
OrderedOuters	Int	No
ReceivedOuters	Int	No
Package	Nvarchar(50)	No
IsOrderFinalized	Bit	No
LineageKey	Int	No

FactPurchase will have 1 million rows of data added daily and will contain three years of data.

Transact-SQL queries similar to the following query will be executed daily.

```
SELECT -
SupplierKey, StockItemKey, COUNT(*)

FROM FactPurchase -

WHERE DateKey >= 20210101 -

AND DateKey <= 20210131 -
GROUP By SupplierKey, StockItemKey
```

Which table distribution will minimize query times?

A. replicated

B. hash-distributed on PurchaseKey

C. round-robin

D. hash-distributed on DateKey

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 40
Topic #: 1
[All DP-203 Questions]

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

A. Use Snappy compression for the files.

B. Use OPENROWSET to query the Parquet files.

C. Create an external table that contains a subset of columns from the Parquet files.

D. Store all data as string in the Parquet files.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 41

Topic #: 1

[All DP-203 Questions]

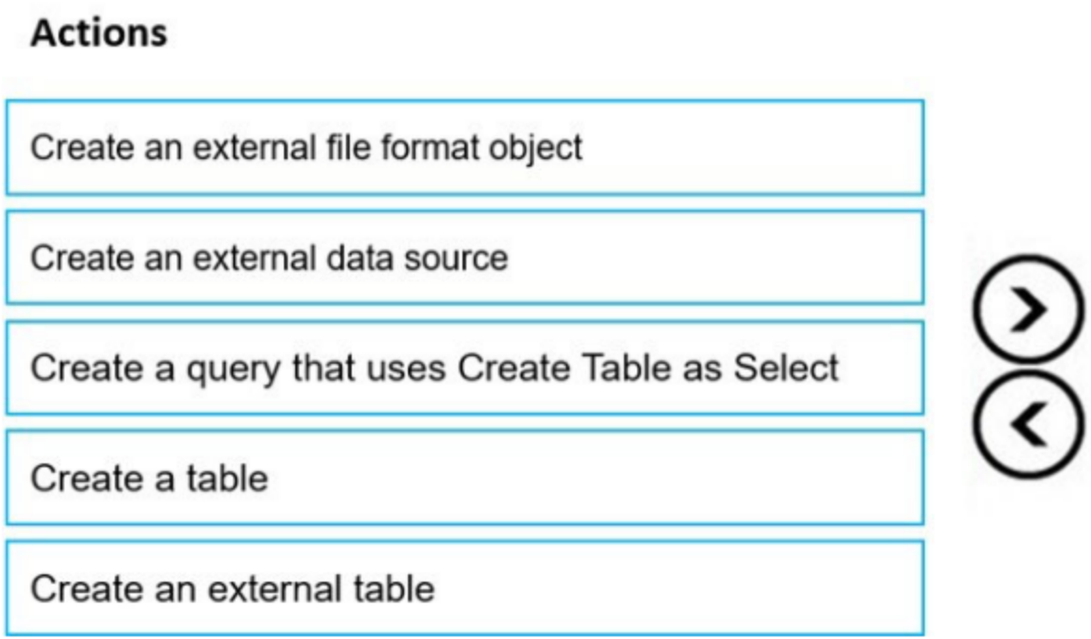
DRAG DROP -

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

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

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

Select and Place:



Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 42
Topic #: 1
[All DP-203 Questions]

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

FirstName -
▪
✑ LastName
✑ Recipient
✑ GrossAmount
✑ TransactionID
✑ GovernmentID
✑ NetAmountPaid
✑ TransactionDate
You need to design a star schema data model in an Azure Synapse Analytics dedicated SQL pool for the data mart.
Which two tables should you create? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. a dimension table for Transaction

B. a dimension table for EmployeeTransaction

C. a dimension table for Employee

D. a fact table for Employee

E. a fact table for Transaction

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 43

Topic #: 1

[All DP-203 Questions]

You are designing a dimension table for a data warehouse. The table will track the value of the dimension attributes over time and preserve the history of the data by adding new rows as the data changes.

Which type of slowly changing dimension (SCD) should you use?

A. Type 0

B. Type 1

C. Type 2

D. Type 3

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 44
Topic #: 1
[All DP-203 Questions]

DRAG DROP -
You have data stored in thousands of CSV files in Azure Data Lake Storage Gen2. Each file has a header row followed by a properly formatted carriage return (/ r) and line feed (/n).
You are implementing a pattern that batch loads the files daily into a dedicated SQL pool in Azure Synapse Analytics by using PolyBase.
You need to skip the header row when you import the files into the data warehouse. Before building the loading pattern, you need to prepare the required database objects in Azure Synapse Analytics.
Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
NOTE: Each correct selection is worth one point
Select and Place:

Actions

- Create a database scoped credential that uses Azure Active Directory Application and a Service Principal Key
- Create an external data source that uses the abfs location
- Use `CREATE EXTERNAL TABLE AS SELECT (CETAS)` and configure the reject options to specify reject values or percentages
- Create an external file format and set the `First_Row` option

Answer Area

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 45
Topic #: 1
[All DP-203 Questions]

HOTSPOT -
You are building an Azure Synapse Analytics dedicated SQL pool that will contain a fact table for transactions from the first half of the year 2020.
You need to ensure that the table meets the following requirements:
✑ Minimizes the processing time to delete data that is older than 10 years
✑ Minimizes the I/O for queries that use year-to-date values
How should you complete the Transact-SQL statement? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Hot Area:

Answer Area

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 46

Topic #: 1

[All DP-203 Questions]

You are performing exploratory analysis of the bus fare data in an Azure Data Lake Storage Gen2 account by using an Azure Synapse Analytics serverless SQL pool.

You execute the Transact-SQL query shown in the following exhibit.

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

What do the query results include?

A. Only CSV files in the tripdata_2020 subfolder.

B. All files that have file names that beginning with "tripdata_2020".

C. All CSV files that have file names that contain "tripdata_2020".

D. Only CSV that have file names that beginning with "tripdata_2020".

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 47

Topic #: 1

[All DP-203 Questions]

DRAG DROP -

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

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

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

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

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

NOTE: Each correct selection is worth one point.

Select and Place:

Values

alias

array_union

createDataFrame

explode

select

translate

Answer Area

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 48

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

You are designing an application that will store petabytes of medical imaging data.

When the data is first created, the data will be accessed frequently during the first week. After one month, the data must be accessible within 30 seconds, but files will be accessed infrequently. After one year, the data will be accessed infrequently but must be accessible within five minutes.

You need to select a storage strategy for the data. The solution must minimize costs.

Which storage tier should you use for each time frame? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

First week: [Archive / Cool / Hot]

After one month: [Archive / Cool / Hot]

After one year: [Archive / Cool / Hot]

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 49

Topic #: 1

[All DP-203 Questions]

You have an Azure Synapse Analytics Apache Spark pool named Pool1.

You plan to load JSON files from an Azure Data Lake Storage Gen2 container into the tables in Pool1. The structure and data types vary by file.

You need to load the files into the tables. The solution must maintain the source data types.

What should you do?

A. Use a Conditional Split transformation in an Azure Synapse data flow.

B. Use a Get Metadata activity in Azure Data Factory.

C. Load the data by using the OPENROWSET Transact-SQL command in an Azure Synapse Analytics serverless SQL pool.

D. Load the data by using PySpark.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 50

Topic #: 1

[All DP-203 Questions]

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier. Workspace1 contains an all-purpose cluster named cluster1.

You need to reduce the time it takes for cluster1 to start and scale up. The solution must minimize costs.

What should you do first?

A. Configure a global init script for workspace1.

B. Create a cluster policy in workspace1.

C. Upgrade workspace1 to the Premium pricing tier.

D. Create a pool in workspace1.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 51

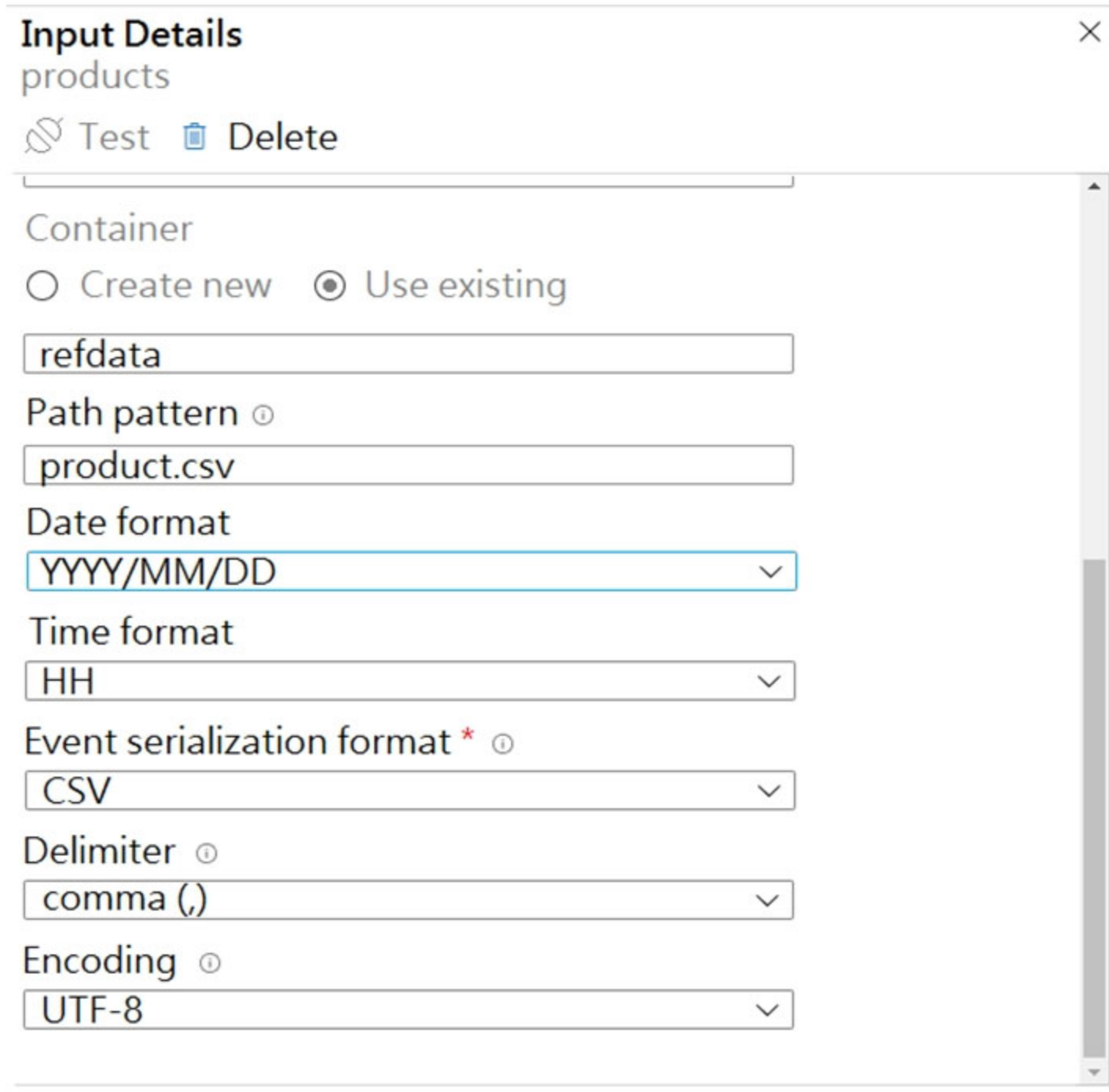
Topic #: 1

[All DP-203 Questions]

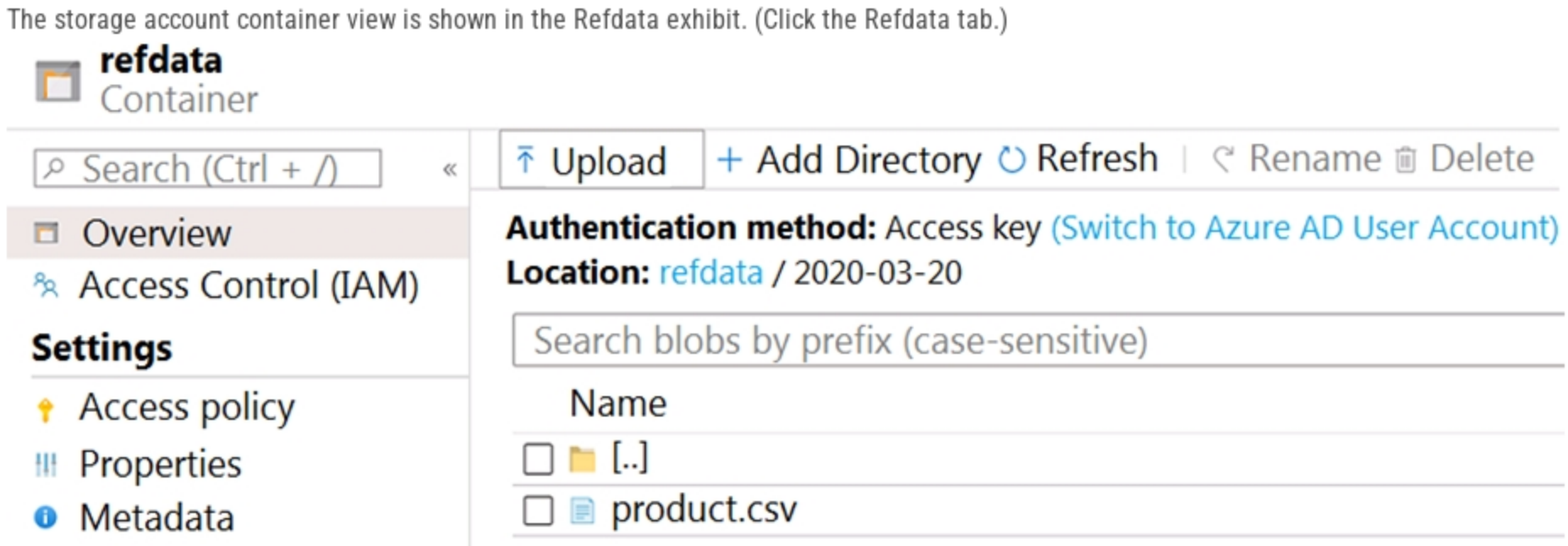
HOTSPOT -

You are building an Azure Stream Analytics job that queries reference data from a product catalog file. The file is updated daily.

The reference data input details for the file are shown in the Input exhibit. (Click the Input tab.)



The storage account container view is shown in the Refdata exhibit. (Click the Refdata tab.)

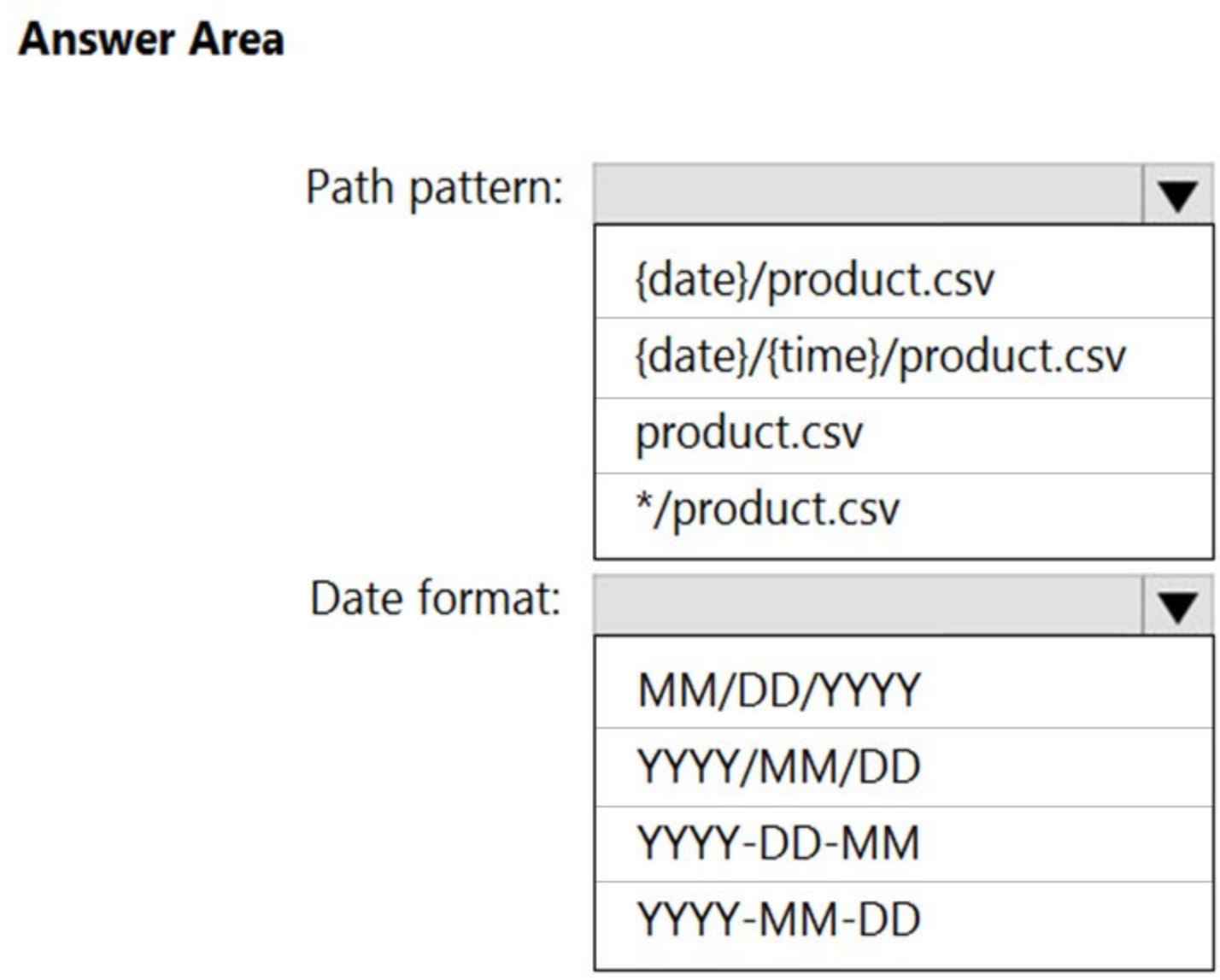


You need to configure the Stream Analytics job to pick up the new reference data.

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

NOTE: Each correct selection is worth one point.

Hot Area:



Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 52

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

You have the following Azure Stream Analytics query.

```
WITH

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 53

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

You are building a database in an Azure Synapse Analytics serverless SQL pool.

You have data stored in Parquet files in an Azure Data Lake Storege Gen2 container.

Records are structured as shown in the following sample.

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

The records contain two applicants at most.

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 54

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named Account1.

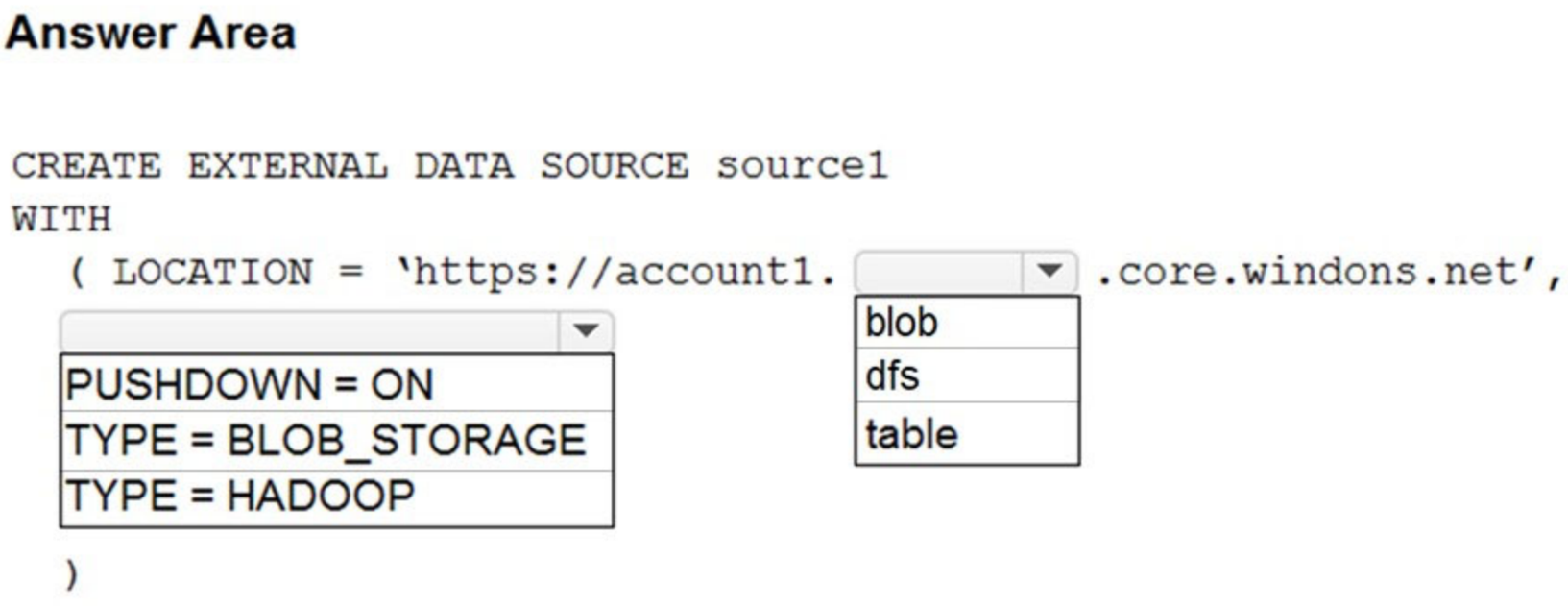
You plan to access the files in Account1 by using an external table.

You need to create a data source in Pool1 that you can reference when you create the external table.

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

NOTE: Each correct selection is worth one point.

Hot Area:



Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 55

Topic #: 1

[All DP-203 Questions]

You have an Azure subscription that contains an Azure Blob Storage account named storage1 and an Azure Synapse Analytics dedicated SQL pool named Pool1.

You need to store data in storage1. The data will be read by Pool1. The solution must meet the following requirements:

Enable Pool1 to skip columns and rows that are unnecessary in a query.

- ✑ Automatically create column statistics.
- ✑ Minimize the size of files.

Which type of file should you use?

A. JSON

B. Parquet

C. Avro

D. CSV

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 56

Topic #: 1

[All DP-203 Questions]

DRAG DROP -

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

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

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

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

NOTE: Each correct selection is worth one point.

Select and Place:

Values

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

Answer Area

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

,   DISTRIBUTION =  [ Value ]  ([ProductKey])

,   PARTITION   (  [ Value ] ] RANGE RIGHT FOR VALUES

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 57

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

You have an Azure Data Lake Storage Gen2 service.

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

- ✑ Data that is older than five years is accessed infrequently but must be available within one second when requested.
- ✑ Data that is older than seven years is NOT accessed.
- ✑ Costs must be minimized while maintaining the required availability.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Data over five years old:

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

Data over seven years old:

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 58

Topic #: 1

[All DP-203 Questions]

HOTSPOT -

You plan to create an Azure Data Lake Storage Gen2 account.

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

- ✑ Provides the highest degree of data resiliency
- ✑ Ensures that content remains available for writes if a primary data center fails

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Replication mechanism:	Change feed Zone-redundant storage (ZRS) Read-access geo-redundant storage (RA-GRS) Read-access geo-zone-redundant storage (RA-GRS)
Failover process:	Failover initiated by Microsoft Failover manually initiated by the customer Failover automatically initiated by an Azure Automation job

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 59

Topic #: 1

[All DP-203 Questions]

You need to implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool.

You have a table that was created by using the following Transact-SQL statement.

```
CREATE TABLE [DB0].[DimProduct](
[ProductKey] [int] IDENTITY(1,1) NOT NULL,
[ProductSourceID] [int] NOT NULL,
[ProductNane] [nvarchar](100) NOT NULL,
[Color] [nvarchar] (15) NULL,
[SellStartDate] [date] NOT NULL,
[SellEnd0ate] [date] NULL,
[RowInsertedDateTime] [datetime] NOT NULL,
[RowipdatedDateTine] [datetime] NOT NULL,
[ETLAuditID] [int] NOT NULL
)
```

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

NOTE: Each correct selection is worth one point.

A.

```
[EffectiveEndDate] [datetime] NULL,
```

B.

```
[CurrentProductCategory] [nvarchar] (100) NOT NULL,
```

C.

```
[ProductCategory] [nvarchar](100) NOT NULL,
```

D.

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

E.

```
[OriginalProductCategory] [nvarchar] (100) NOT NULL,
```

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 60

Topic #: 1

[All DP-203 Questions]

DRAG DROP -

You have an Azure subscription.

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

Pool1 will contain the following tables.

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

Select and Place:

Table distribution types

Hash

Replicated

Round-robin

Answer Area

Common.Data: [ ]

Marketing.Web.Sessions: [ ]

Staging. Web.Sessions: [ ]

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 61
Topic #: 1
[All DP-203 Questions]

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

Answer Area

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

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

```
, DISTRIBUTION =
);
```

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 62

Topic #: 1

[All DP-203 Questions]

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. Table1 contains the following:

- One billion rows
- A clustered columnstore index
- A hash-distributed column named Product Key
- A column named Sales Date that is of the date data type and cannot be null

Thirty million rows will be added to Table1 each month.

You need to partition Table1 based on the Sales Date column. The solution must optimize query performance and data loading.

How often should you create a partition?

A. once per month

B. once per year

C. once per day

D. once per week

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 63

Topic #: 1

[All DP-203 Questions]

You have an Azure Databricks workspace that contains a Delta Lake dimension table named Table1.

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. ALTER

D. MERGE

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 64

Topic #: 1

[All DP-203 Questions]

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 65
Topic #: 1
[All DP-203 Questions]

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You have an Azure Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.
You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly.
Solution: You modify the files to ensure that each row is less than 1 MB.
Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 66

Topic #: 1

[All DP-203 Questions]

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. replicated

B. hash distributed

C. heap

D. round-robin

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 67

Topic #: 1

[All DP-203 Questions]

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 68

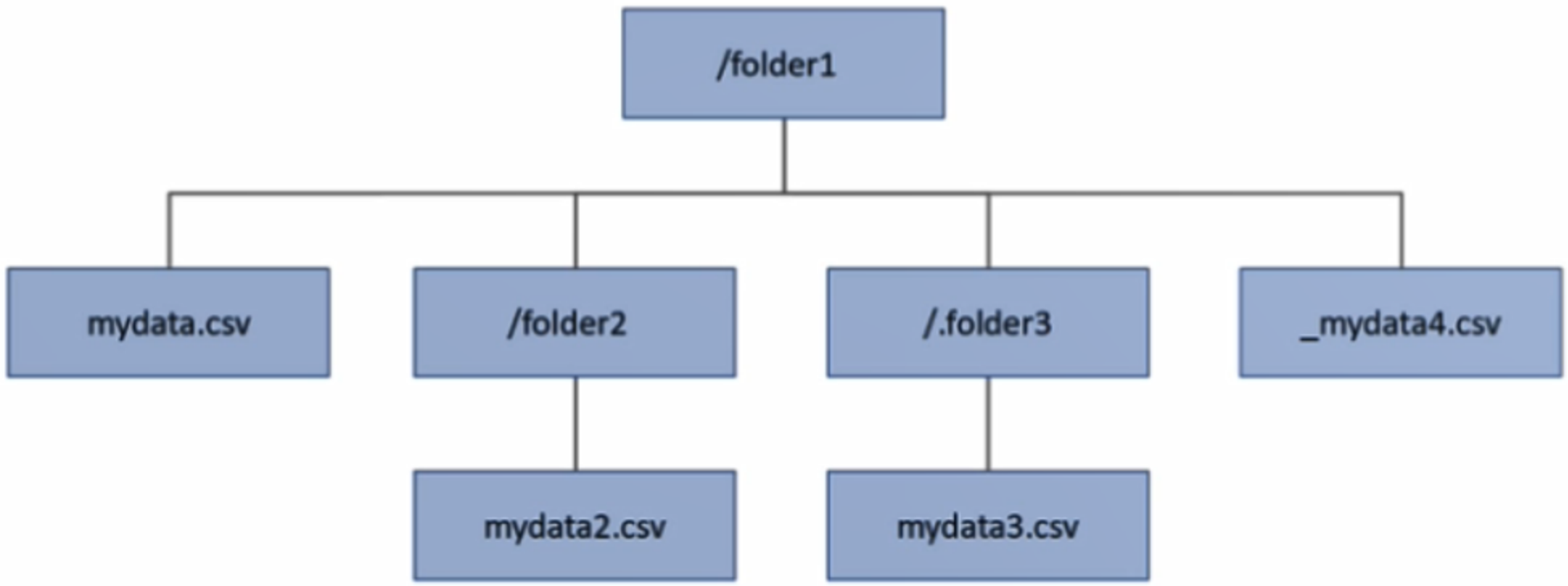
Topic #: 1

[All DP-203 Questions]

HOTSPOT

-

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



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

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

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

NOTE: Each correct selection is worth one point.

Answer Area

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 69

Topic #: 1

[All DP-203 Questions]

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a fact table named Table1 that will store sales data from the last three years. The solution must be optimized for the following query operations:

- Show order counts by week.
- Calculate sales totals by region.
- Calculate sales totals by product.
- Find all the orders from a given month.

Which data should you use to partition Table1?

A. product

B. month

C. week

D. region

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 70
Topic #: 1
[All DP-203 Questions]

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 pools.
- 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. \DataSource\SubjectArea\YYYY\WW\FileData_YYYY_MM_DD.parquet

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

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

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

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 71

Topic #: 1

[All DP-203 Questions]

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a table named table1.

You load 5 TB of data into table1.

You need to ensure that columnstore compression is maximized for table1.

Which statement should you execute?

A. DBCC INDEXDEFRAG (pool1, table1)

B. DBCC DBREINDEX (table1)

C. ALTER INDEX ALL on table1 REORGANIZE

D. ALTER INDEX ALL on table1 REBUILD

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 72

Topic #: 1

[All DP-203 Questions]

You have an Azure Synapse Analytics dedicated SQL pool named pool1.

You plan to implement a star schema in pool and create a new table named DimCustomer by using the following code.

```
CREATE TABLE dbo.[DimCustomer](
    [CustomerKey] int NOT NULL,
    [CustomerSourceID] [int] NOT NULL,
    [Title] [nvarchar](8) NULL,
    [FirstName] [nvarchar](50) NOT NULL,
    [MiddleName] [nvarchar](50) NULL,
    [LastName] [nvarchar](50) NOT NULL,
    [Suffix] [nvarchar](10) NULL,
    [CompanyName] [nvarchar](128) NULL,
    [SalesPerson] [nvarchar](256) NULL,
    [EmailAddress] [nvarchar](50) NULL,
    [Phone] [nvarchar](25) NULL,
    [InsertedDate] [datetime] NOT NULL,
    [ModifiedDate] [datetime] NOT NULL,
    [HashKey] [varchar](100) NOT NULL,
    [IsCurrentRow] [bit] NOT NULL
)
WITH
(
    DISTRIBUTION = REPLICATE,
    CLUSTERED COLUMNSTORE INDEX
);
GO
```

You need to ensure that DimCustomer has the necessary columns to support a Type 2 slowly changing dimension (SCD).

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

NOTE: Each correct selection is worth one point.

A. [HistoricalSalesPerson] [nvarchar] (256) NOT NULL

B. [EffectiveEndDate] [datetime] NOT NULL

C. [PreviousModifiedDate] [datetime] NOT NULL

D. [RowID] [bigint] NOT NULL

E. [EffectiveStartDate] [datetime] NOT NULL

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 73
Topic #: 1
[All DP-203 Questions]

HOTSPOT
-

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool.
You plan to deploy a solution that will analyze sales data and include the following:

- A table named Country that will contain 195 rows
- A table named Sales that will contain 100 million rows
- A query to identify total sales by country and customer from the past 30 days

You need to create the tables. The solution must maximize query performance.

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

NOTE: Each correct selection is worth one point.

Answer Area

```
CREATE TABLE [dbo].[Sales]
(
      [OrderDate]         date          NOT NULL
,     [CustomerId] int NOT NULL
,     [CountryId] int NOT NULL
,     [Total] money NOT NULL
)
WITH
(
      DISTRIBUTION =
      CLUSTERED COLUMNSTORE INDEX
)
CREATE TABLE [dbo].[Country]
(
      [CountryId] int NOT NULL
,     [CountryCode] varchar(10) NOT NULL
)
WITH
(
      DISTRIBUTION =
      CLUSTERED COLUMNSTORE INDEX
)
```

DISTRIBUTION = HASH([CustomerId]) / HASH([OrderDate]) / REPLICATE / ROUND_ROBIN

DISTRIBUTION = HASH([CountryCode]) / HASH([CountryId]) / REPLICATE / ROUND_ROBIN

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 74
Topic #: 1
[All DP-203 Questions]

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named account1 and an Azure Synapse Analytics workspace named workspace1.

You need to create an external table in a serverless SQL pool in workspace1. The external table will reference CSV files stored in account1. The solution must maximize performance.

How should you configure the external table?

A. Use a native external table and authenticate by using a shared access signature (SAS).

B. Use a native external table and authenticate by using a storage account key.

C. Use an Apache Hadoop external table and authenticate by using a shared access signature (SAS).

D. Use an Apache Hadoop external table and authenticate by using a service principal in Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 75

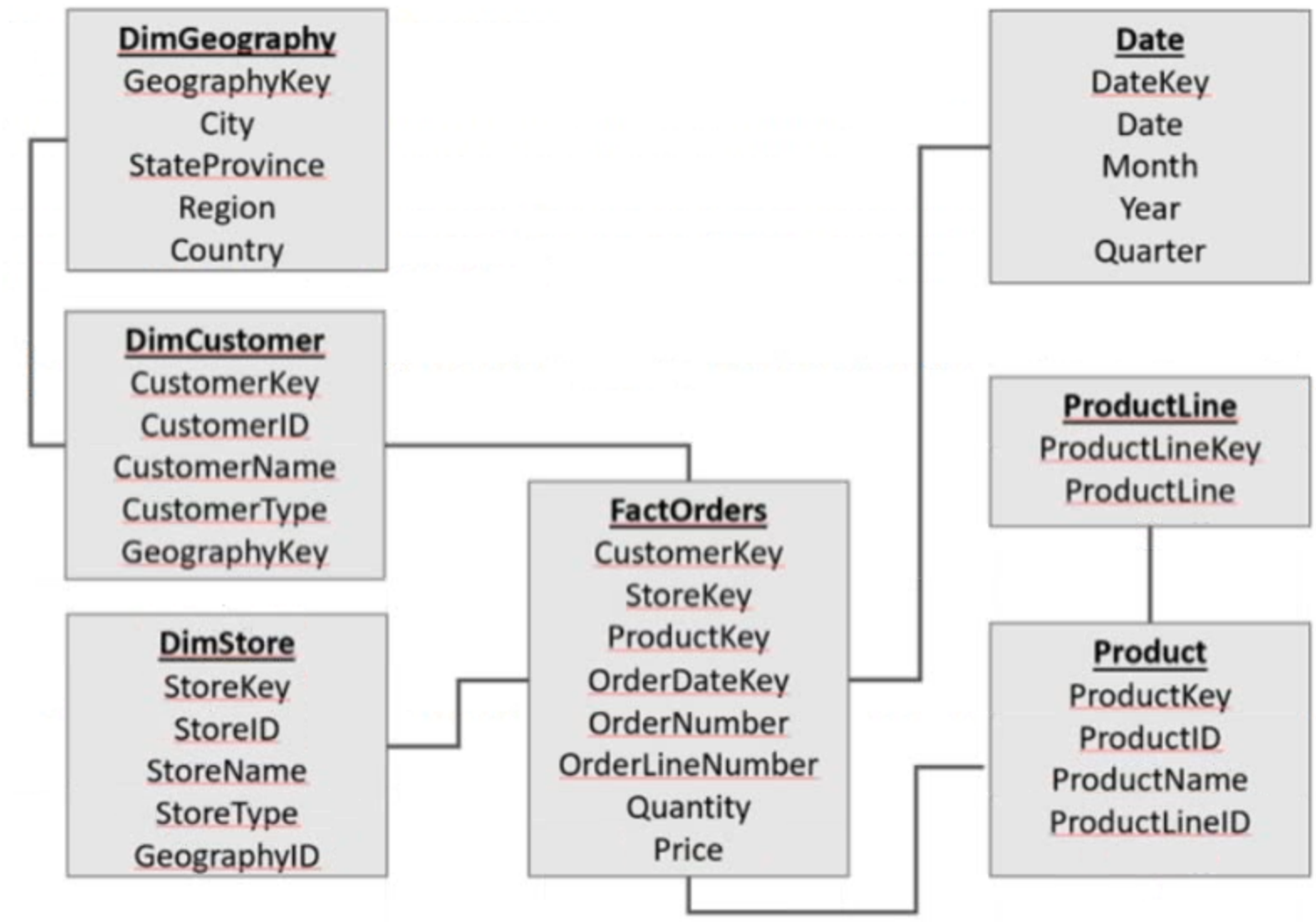
Topic #: 1

[All DP-203 Questions]

HOTSPOT

-

You have an Azure Synapse Analytics serverless SQL pool that contains a database named db1. The data model for db1 is 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 exhibit.

NOTE: Each correct selection is worth one point.

Answer Area

To convert the data model to a star schema, **[answer choice]**.

- join DimGeography and DimCustomer
- join DimGeography and FactOrders
- union DimGeography and DimCustomer
- union DimGeography and FactOrders

Once the data model is converted into a star schema, there will be **[answer choice]** tables.

- 4
- 5
- 6
- 7

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 76

Topic #: 1

[All DP-203 Questions]

You have an Azure Databricks workspace and an Azure Data Lake Storage Gen2 account named storage1.

New files are uploaded daily to storage1.

You need to recommend a solution that configures storage1 as a structured streaming source. The solution must meet the following requirements:

- Incrementally process new files as they are uploaded to storage1.
- Minimize implementation and maintenance effort.
- Minimize the cost of processing millions of files.
- Support schema inference and schema drift.

Which should you include in the recommendation?

A. COPY INTO

B. Azure Data Factory

C. Auto Loader

D. Apache Spark FileStreamSource

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 77

Topic #: 1

[All DP-203 Questions]

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

Name	Type	Description
storage1	Azure Blob storage account	Contains publicly accessible TSV files that do **NOT** have a header row
WS1	Azure Synapse Analytics workspace	Contains a serverless SQL pool

You need to read the TSV files by using ad-hoc queries and the OPENROWSET function. The solution must assign a name and override the inferred data type of each column.

What should you include in the OPENROWSET function?

A. the WITH clause

B. the ROWSET_OPTIONS bulk option

C. the DATAFILETYPE bulk option

D. the DATA_SOURCE parameter

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 78

Topic #: 1

[All DP-203 Questions]

You have an Azure Synapse Analytics dedicated SQL pool.

You plan to create a fact table named Table1 that will contain a clustered columnstore index.

You need to optimize data compression and query performance for Table1.

What is the minimum number of rows that Table1 should contain before you create partitions?

A. 100,000

B. 600,000

C. 1 million

D. 60 million

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 79

Topic #: 1

[All DP-203 Questions]

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named DimSalesPerson. DimSalesPerson contains the following columns:

- RepSourceID
- SalesRepID
- FirstName
- LastName
- StartDate
- EndDate
- Region

You are developing an Azure Synapse Analytics pipeline that includes a mapping data flow named Dataflow1. Dataflow1 will read sales team data from an external source and use a Type 2 slowly changing dimension (SCD) when loading the data into DimSalesPerson.

You need to update the last name of a salesperson in DimSalesPerson.

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

NOTE: Each correct selection is worth one point.

A. Update three columns of an existing row.

B. Update two columns of an existing row.

C. Insert an extra row.

D. Update one column of an existing row.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 80

Topic #: 1

[All DP-203 Questions]

HOTSPOT

-

You plan to use an Azure Data Lake Storage Gen2 account to implement a Data Lake development environment that meets the following requirements:

- Read and write access to data must be maintained if an availability zone becomes unavailable.
- Data that was last modified more than two years ago must be deleted automatically.
- Costs must be minimized.

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

NOTE: Each correct selection is worth one point.

Answer Area

For storage redundancy:

- Geo-zone-redundant storage (GZRS)
- Locally-redundant storage (LRS)
- Zone-redundant storage (ZRS)

For data deletion:

- A lifecycle management policy
- Soft delete
- Versioning

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 81
Topic #: 1
[All DP-203 Questions]

HOTSPOT
-

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 and rarely accessed.
- 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.

What should you include in the storage policy for each department? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

HR:
- Archive storage after one day and delete storage after 2,555 days.
- Archive storage after 2,555 days.
- Cool storage after one day.
- Cool storage after 180 days.
- Cool storage after 180 days and delete storage after 2,555 days.
- Delete after one day.
- Delete after 180 days.

Operations:
- Archive storage after one day and delete storage after 2,555 days.
- Archive storage after 2,555 days.
- Cool storage after one day.
- Cool storage after 180 days.
- Cool storage after 180 days and delete storage after 2,555 days.
- Delete after one day.
- Delete after 180 days.

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 82
Topic #: 1
[All DP-203 Questions]

HOTSPOT
-

You are developing an Azure Synapse Analytics pipeline that will include a mapping data flow named Dataflow1. Dataflow1 will read customer data from an external source and use a Type 1 slowly changing dimension (SCD) when loading the data into a table named DimCustomer in an Azure Synapse Analytics dedicated SQL pool.

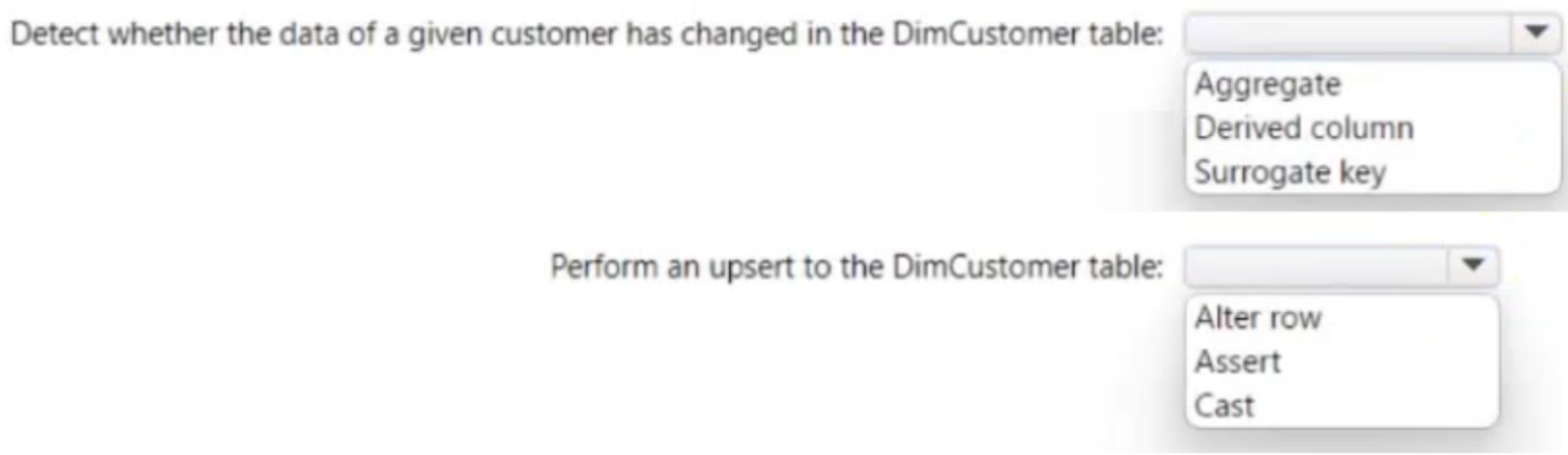
You need to ensure that Dataflow1 can perform the following tasks:

- Detect whether the data of a given customer has changed in the DimCustomer table.
- Perform an upsert to the DimCustomer table.

Which type of transformation should you use for each task? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area



Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 83
Topic #: 1
[All DP-203 Questions]

DRAG DROP
-

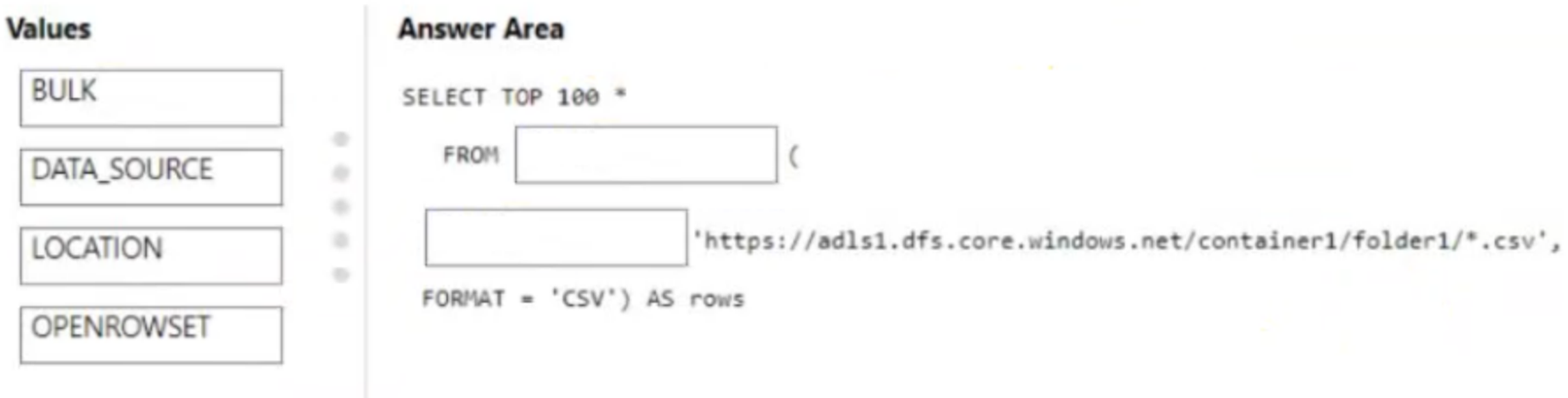
You have an Azure Synapse Analytics serverless SQL pool.

You have an Azure Data Lake Storage account named adls1 that contains a public container named container1. The container1 container contains a folder named folder1.

You need to query the top 100 rows of all the CSV files in folder1.

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



Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 84

Topic #: 1

[All DP-203 Questions]

You have an Azure Synapse Analytics workspace named WS1 that contains an Apache Spark pool named Pool1.

You plan to create a database named DB1 in Pool1.

You need to ensure that when tables are created in DB1, the tables are available automatically as external tables to the built-in serverless SQL pool.

Which format should you use for the tables in DB1?

A. Parquet

B. ORC

C. JSON

D. HIVE

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 85

Topic #: 1

[All DP-203 Questions]

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

You plan to implement query acceleration for storage1.

Which two file types support query acceleration? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. JSON

B. Apache Parquet

C. XML

D. CSV

E. Avro

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 86

Topic #: 1

[All DP-203 Questions]

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

Name	Type	Description
storage1	Azure Blob storage account	Contains publicly accessible JSON files
WS1	Azure Synapse Analytics workspace	Contains a serverless SQL pool

You need to read the files in storage1 by using ad-hoc queries and the OPENROWSET function. The solution must ensure that each rowset contains a single JSON record.

To what should you set the FORMAT option of the OPENROWSET function?

A. JSON

B. DELTA

C. PARQUET

D. CSV

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 87

Topic #: 1

[All DP-203 Questions]

HOTSPOT

-

You have an Azure subscription that contains the Azure Synapse Analytics workspaces shown in the following table.

Name	Primary storage account
workspace1	datalake1
workspace2	datalake2
workspace3	datalake1

Each workspace must read and write data to datalake1.

Each workspace contains an unused Apache Spark pool.

You plan to configure each Spark pool to share catalog objects that reference datalake1.

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 Area

Statements	Yes	No
The shared catalog objects can be stored in Azure Database for MySQL.	○	○
For the Apache Hive Metastore of each workspace, you must configure a linked service that uses user-password authentication.	○	○
The users of workspace1 must be assigned the Storage Blob Contributor role for datalake1.	○	○

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 88
Topic #: 1
[All DP-203 Questions]

DRAG DROP
-

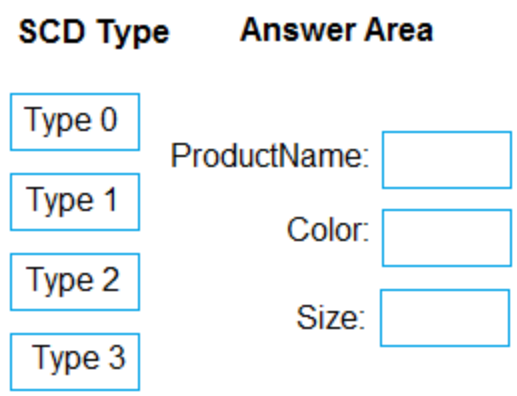
You have a data warehouse.

You need to implement a slowly changing dimension (SCD) named Product that will include three columns named ProductName, ProductColor, and ProductSize. The solution must meet the following requirements:

- Prevent changes to the values stored in ProductName.
- Retain only the current and the last values in ProductSize.
- Retain all the current and previous values in ProductColor.

Which type of SCD should you implement for each column? To answer, drag the appropriate types to the correct columns. Each 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.



Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 89
Topic #: 1
[All DP-203 Questions]

HOTSPOT
-

You are incrementally loading data into fact tables in an Azure Synapse Analytics dedicated SQL pool.

Each batch of incoming data is staged before being loaded into the fact tables.

You need to ensure that the incoming data is staged as quickly as possible.

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

NOTE: Each correct selection is worth one point.

Answer Area

Table distribution:
- HASH
- REPLICATE
- ROUND_ROBIN

Table structure:
- Clustered index
- Columnstore index
- Heap

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 90
Topic #: 1
[All DP-203 Questions]

You have an Azure subscription that contains an Azure Synapse Analytics workspace named ws1 and an Azure Cosmos DB database account named Cosmos1. Cosmos1 contains a container named container1 and ws1 contains a serverless SQL pool.

You need to ensure that you can query the data in container1 by using the serverless SQL pool.

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

NOTE: Each correct selection is worth one point.

A. Enable Azure Synapse Link for Cosmos1.

B. Disable the analytical store for container1.

C. In ws1, create a linked service that references Cosmos1.

D. Enable the analytical store for container1.

E. Disable indexing for container1.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 91

Topic #: 1

[All DP-203 Questions]

HOTSPOT

-

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

Name	Type	Description
Workspace1	Azure Synapse workspace	Contains the Built-in serverless SQL pool
Pool1	Azure Synapse Analytics dedicated SQL pool	Deployed to Workspace1
storage1	Storage account	Hierarchical namespace enabled

The storage1 account contains a container named container1. The container1 container contains the following files.

```
Webdata <root folder>
    Monthly <folder>
           _monthly.csv
           Monthly.csv
    .testdata.csv
    testdata.csv
```

In Pool1, you run the following script.

```
CREATE EXTERNAL DATA SOURCE Ds1
WITH
    ( LOCATION = 'abfss://container1@storage1.dfs.core.windows.net' ,
    CREDENTIAL = credential1,
    TYPE = HADOOP
  ) ;
```

In the Built-in serverless SQL pool, you run the following script.

```
CREATE EXTERNAL DATA SOURCE Ds2
WITH (
    LOCATION = 'https://storage1.blob.core.windows.net/container1/Webdata/',
    CREDENTIAL = credential2
);
```

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 Area

Statements	Yes	No
An external table that uses Ds1 can read the _monthly.csv file.	○	○
An external table that uses Ds1 can read the Monthly.csv file.	○	○
An external table that uses Ds2 can read the .testdata.csv file.	○	○

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 92

Topic #: 1

[All DP-203 Questions]

DRAG DROP

-

You have an Azure subscription that contains an Azure Data Lake Storage Gen2 account named account1 and a user named User1.

In account1, you create a container named container1. In container1, you create a folder named folder1.

You need to ensure that User1 can list and read all the files in folder1. The solution must use the principle of least privilege.

How should you configure the permissions for each folder? To answer, drag the appropriate permissions to the correct folders. 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.

NOTE: Each correct selection is worth one point.

Permissions

Execute	None
Read	Read and Execute
Read and Write	Write

Answer Area

container1/:

container1/folder1:

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 93

Topic #: 1

[All DP-203 Questions]

You have an Azure Data Factory pipeline named pipeline1.

You need to execute pipeline1 at 2 AM every day. The solution must ensure that if the trigger for pipeline1 stops, the next pipeline execution will occur at 2 AM, following a restart of the trigger.

Which type of trigger should you create?

A. schedule

B. tumbling

C. storage event

D. custom event

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 94
Topic #: 1
[All DP-203 Questions]

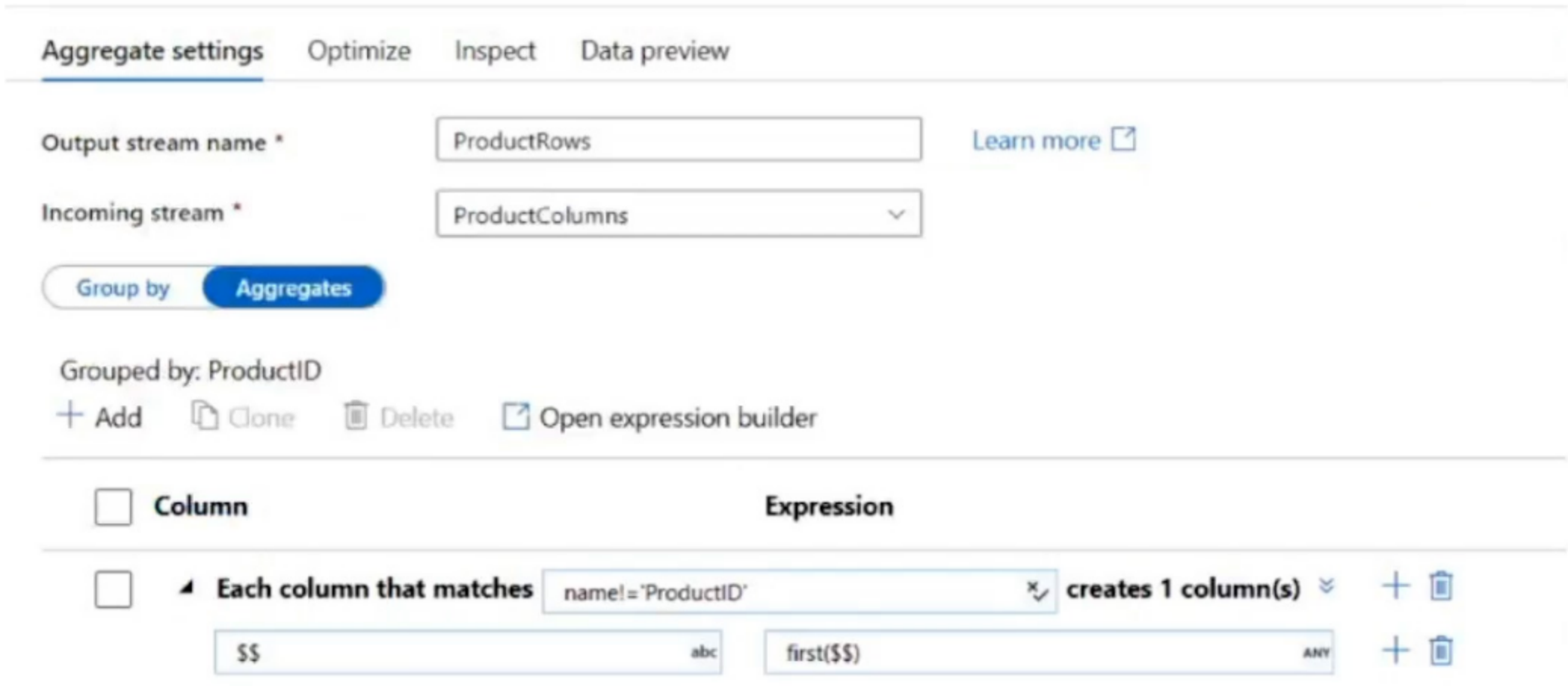
HOTSPOT
-

You have an Azure data factory named adf1 that contains a pipeline named ExecProduct. ExecProduct contains a data flow named Product.

The Product data flow contains the following transformations:

1. WeeklyData: A source that points to a CSV file in an Azure Data Lake Storage Gen2 account with 20 columns
2. ProductColumns: A select transformation that selects from WeeklyData six columns named ProductID, ProductDescr, ProductSubCategory, ProductCategory, ProductStatus, and ProductLastUpdated
3. ProductRows: An aggregate transformation
4. ProductList: A sink that outputs data to an Azure Synapse Analytics dedicated SQL pool

The Aggregate settings for ProductRows are configured as 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 Area

Statements	Yes	No
There will be six columns in the output of ProductRows.	○	○
There will always be one output row for each unique value of ProductDescr.	○	○
There will always be one output row for each unique value of ProductID.	○	○

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 95

Topic #: 1

[All DP-203 Questions]

You manage an enterprise data warehouse in Azure Synapse Analytics.

Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries.

You need to monitor resource utilization to determine the source of the performance issues.

Which metric should you monitor?

A. DWU limit

B. Cache hit percentage

C. Local tempdb percentage

D. Data IO percentage

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 96

Topic #: 1

[All DP-203 Questions]

HOTSPOT

-

You have an Azure Synapse Analytics serverless SQL pool.

You have an Apache Parquet file that contains 10 columns.

You need to query data from the file. The solution must return only two columns.

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 Area

```
SELECT * FROM
OPENROWSET( [ BULK | DELTA | OPENQUERY | SINGLE_BLOB ] N'https://myaccount.dfs.core.windows.net/mycontainer/mysubfolder/data.parquet', FORMAT = 'PARQUET')
WITH [ (Col1 int, Col2 varchar(20)) | FILEPATH(2) | PARSER_VERSION = '2.0' | SINGLE_BLOB ] as rows
```

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 97
Topic #: 1
[All DP-203 Questions]

You have an Azure Synapse Analytics workspace that contains an Apache Spark pool named SparkPool1. SparkPool1 contains a Delta Lake table named SparkTable1.

You need to recommend a solution that supports Transact-SQL queries against the data referenced by SparkTable1. The solution must ensure that the queries can use partition elimination.

What should you include in the recommendation?

A. a partitioned table in a dedicated SQL pool

B. a partitioned view in a dedicated SQL pool

C. a partitioned index in a dedicated SQL pool

D. a partitioned view in a serverless SQL pool

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 98

Topic #: 1

[All DP-203 Questions]

You are designing a sales transactions table in an Azure Synapse Analytics dedicated SQL pool. The table will contain 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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 99

Topic #: 1

[All DP-203 Questions]

You have an Azure Synapse Analytics workspace.

You plan to deploy a lake database by using a database template in Azure Synapse.

Which two elements are included in the template? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. relationships

B. data formats

C. linked services

D. table permissions

E. table definitions

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 100

Topic #: 1

[All DP-203 Questions]

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 (SCD) 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
);
```

Which three columns should you add? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. [EffectiveStartDate] [datetime] NOT NULL

B. [EffectiveEndDate] [datetime] NOT NULL

C. [OriginalProductDescription] NVARCHAR(2000) NOT NULL

D. [IsCurrentRow] [bit] NOT NULL

E. [OriginalColor] NVARCHAR(50) NOT NULL

F. [OriginalProductName] NVARCHAR(100) NULL

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 101
Topic #: 1
[All DP-203 Questions]

HOTSPOT
-

You have an Azure Synapse Analytics dedicated SQL pool named Pool1. Pool1 contains a fact table named Table1. 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 Area

Partition the data:
- Partition by date with one partition per day.
- Partition by date with one partition per month.
- Partition by product.

Remove the data:
- Delete the old data from Table1 by using a WHERE clause.
- Delete the old data from Table1 by using a JOIN.
- Switch the oldest partition to another table named Table2 and drop Table2.
- Truncate the oldest partition.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 102

Topic #: 1

[All DP-203 Questions]

You have an Azure subscription that contains an Azure Synapse Analytics serverless SQL pool.

You execute the following query.

```
CREATE EXTERNAL TABLE Orders
WITH
(
     LOCATION = 'orders/',
     DATA_SOURCE = sales,
     FILE_FORMAT = SalesOrders
)
AS
SELECT OrderID, CustomerName, OrderTotal
FROM OPENROWSET
(
     BULK 'sales_orders/*.csv',
     DATA_SOURCE = 'sales',
     FORMAT = 'CSV',
     PARSER_VERSION = '2.0',
     HEADER_ROW = TRUE
) AS source_data
WHERE OrderType = 'Customer Order';
```

Where will the rows returned by the query be stored?

A. in a file in a data lake

B. in a relational database

C. in a global temporary table

D. in a session temporary table

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 103

Topic #: 1

[All DP-203 Questions]

You are deploying a lake database by using an Azure Synapse database template.

You need to add additional tables to the database. The solution must use the same grouping method as the template tables.

Which grouping method should you use?

A. partition style

B. business area

C. size

D. facts and dimensions

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 104

Topic #: 1

[All DP-203 Questions]

You have an Azure data factory connected to a Git repository that contains the following branches:

- main: Collaboration branch
- abc: Feature branch
- xyz: Feature branch

You save changes to a pipeline in the xyz branch.

You need to publish the changes to the live service.

What should you do first?

A. Publish the data factory.

B. Create a pull request to merge the changes into the main branch.

C. Create a pull request to merge the changes into the abc branch.

D. Push the code to a remote origin.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 105

Topic #: 1

[All DP-203 Questions]

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

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

You have an Azure subscription that contains an Azure data factory named ADF1.

From Azure Data Factory Studio, you build a complex data pipeline in ADF1.

You discover that the Save button is unavailable, and there are validation errors that prevent the pipeline from being published.

You need to ensure that you can save the logic of the pipeline.

Solution: You enable Git integration for ADF1.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 106
Topic #: 1
[All DP-203 Questions]

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

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

You have an Azure subscription that contains an Azure data factory named ADF1.

From Azure Data Factory Studio, you build a complex data pipeline in ADF1.

You discover that the Save button is unavailable, and there are validation errors that prevent the pipeline from being published.

You need to ensure that you can save the logic of the pipeline.

Solution: You view the JSON code representation of the resource and copy the JSON to a file.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 107
Topic #: 1
[All DP-203 Questions]

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

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

You have an Azure subscription that contains an Azure data factory named ADF1.

From Azure Data Factory Studio, you build a complex data pipeline in ADF1.

You discover that the Save button is unavailable, and there are validation errors that prevent the pipeline from being published.

You need to ensure that you can save the logic of the pipeline.

Solution: You export ADF1 as an Azure Resource Manager (ARM) template.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 108

Topic #: 1

[All DP-203 Questions]

HOTSPOT

-

You have an Azure Databricks workspace.

You read data from a CSV file by using a notebook, and then load the data to a DataFrame.

You need to add rows from the DataFrame to an existing Delta table by using Python code.

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

```
new_rows_df.write. [ ▼ ] [ ▼ ] .save(delta_table_path)
```

format.("csv").	mode("append")
format.("delta").	mode("error")
format.("json").	mode("ignore")
format.("parquet").	mode("overwrite")

Show Suggested Answer

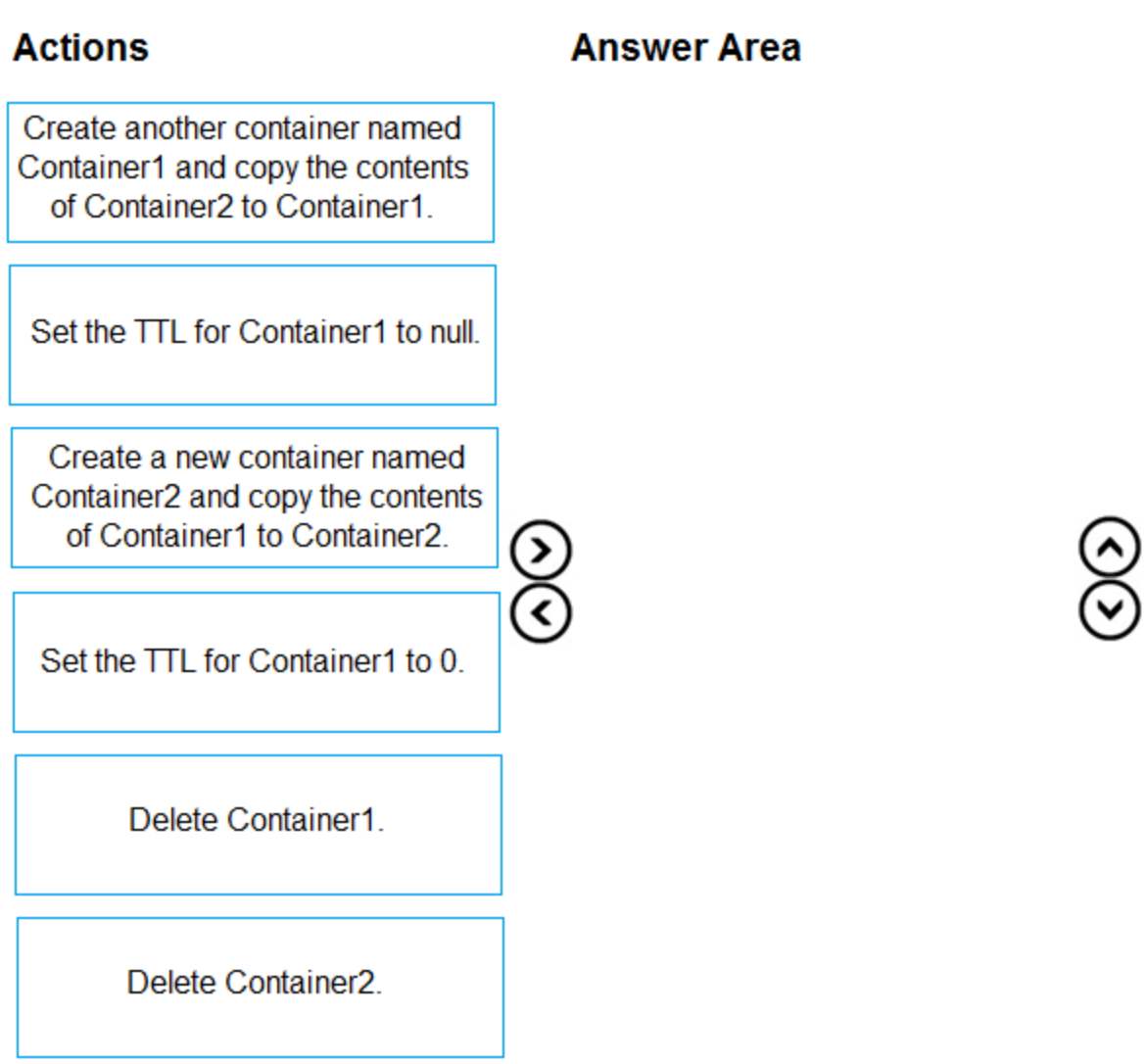
Actual exam question from Microsoft's DP-203
Question #: 109
Topic #: 1
[All DP-203 Questions]

DRAG DROP
-

You have an Azure subscription that contains an Azure Cosmos DB for NoSQL account named account1. The account1 account contains a container named Container1 that has the following configurations:
• Analytical store: On
• TTL: 3600

You need to remove analytical store support from Container1. The solution must meet the following requirements:
• Minimize the impact on the apps that reference Container1.
• Minimize storage usage.

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.



Show Suggested Answer

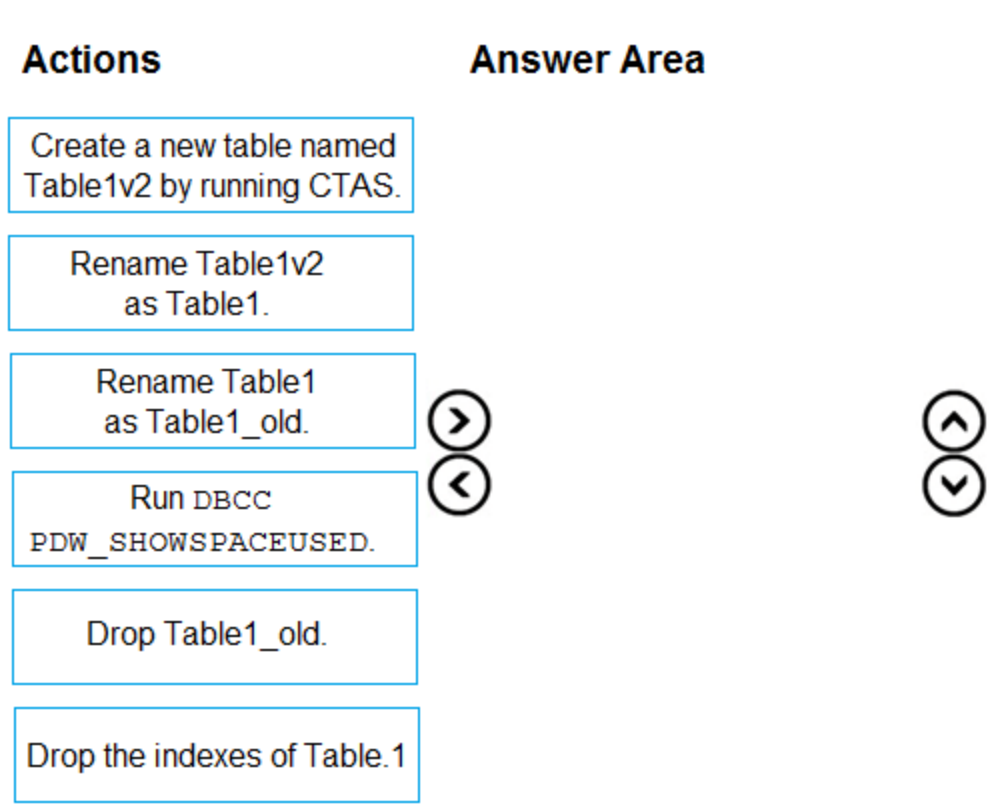
Actual exam question from Microsoft's DP-203
Question #: 110
Topic #: 1
[All DP-203 Questions]

DRAG DROP
-

You have an Azure Synapse Analytics dedicated SQL pool named SQL1 that contains a hash-distributed fact table named Table1.

You need to recreate Table1 and add a new distribution column. The solution must maximize the availability of data.

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.



Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 111
Topic #: 1
[All DP-203 Questions]

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

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

You have an Azure subscription that contains an Azure data factory named ADF1.

From Azure Data Factory Studio, you build a complex data pipeline in ADF1.

You discover that the Save button is unavailable, and there are validation errors that prevent the pipeline from being published.

You need to ensure that you can save the logic of the pipeline.

Solution: You disable all the triggers for ADF1.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 112

Topic #: 1

[All DP-203 Questions]

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

Name	Type	Description
workspace1	Azure Synapse Analytics workspace	*None*
sql1	Azure SQL Database server	*None*
SQLDb1	Azure SQL database	Hosted on sql1

You need to implement Azure Synapse Link for Azure SQL Database.

Which two actions should you perform on sql1? Each correct answer presents a part of the solution.

NOTE: Each correct selection is worth one point.

A. Update the firewall rules to allow Azure services to access sql1.

B. Enable the system-assigned managed identity.

C. From the Access control (IAM) settings, assign the Contributor role to the system-assigned managed identity of workspace1.

D. Disable Transparent Data Encryption (TDE).

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 113

Topic #: 1

[All DP-203 Questions]

You have an Azure subscription that contains an Azure Cosmos DB database. Azure Synapse Link is implemented on the database.

You configure a full fidelity schema for the analytical store.

You perform the following actions:

- Insert {"customerID": 12, "customer": "Tailspin Toys"} as the first document in the container.
- Insert {"customerID": "14", "customer": "Contoso"} as the second document in the container.

How many columns will the analytical store contain?

A. 1

B. 2

C. 3

D. 4

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 114

Topic #: 1

[All DP-203 Questions]

HOTSPOT

-

You have an Azure Data Lake Storage account that contains CSV files. The CSV files contain sales order data and are partitioned by using the following format.

/data/salesorders/year=xxxx/month=y

You need to retrieve only the sales orders from January 2023 and February 2023.

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 Area

```
SELECT product_id, product, list_price
FROM OPENROWSET
(
  BULK 'https://salesdatalake.blob.core.windows.net/data/salesorders/**',
  FORMAT = 'CSV',
 PARSER_VERSION = '2.0'
)
WITH
(
 product_id int,
  product varchar(35) COLLATE Latin1_General_100_BIN2_UTF8,
  list_price decimal(18, 2)
) AS so
WHERE [ so.filepath(0) = 2023 | so.filepath(1) = '2023' | so.year = 2023 ]
AND [ (so.month = 1 OR so.month = 2) | so.filepath(1) IN (1,2) | so.filepath(2) IN ('1', '2') ]
```

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 1

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

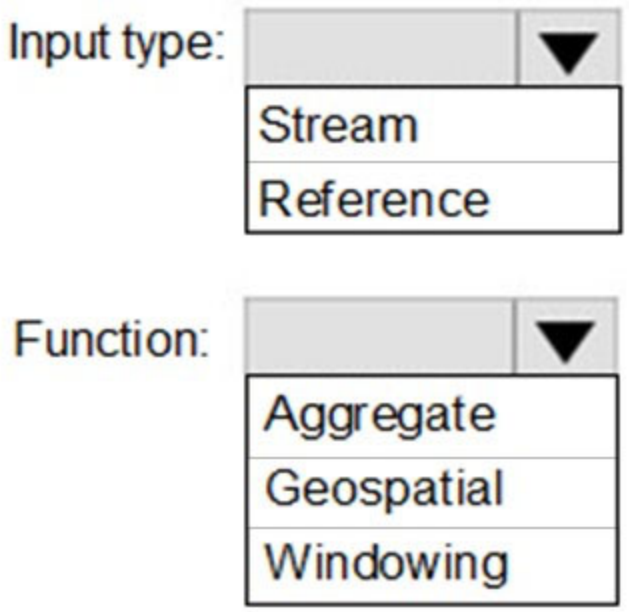
You need to design an Azure Stream Analytics job to process the data for the planned app. The solution must minimize the amount of code developed and the number of technologies used.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area



Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 2
Topic #: 2
[All DP-203 Questions]

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.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 3

Topic #: 2

[All DP-203 Questions]

You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container.

Which resource provider should you enable?

A. Microsoft.Sql

B. Microsoft.Automation

C. Microsoft.EventGrid

D. Microsoft.EventHub

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 4

Topic #: 2

[All DP-203 Questions]

You plan to perform batch processing in Azure Databricks once daily.

Which type of Databricks cluster should you use?

A. High Concurrency

B. automated

C. interactive

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 5

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
WITH LastInWindow AS
(
        SELECT
                [ ▼ ] (Time) AS LastEventTime
        FROM
                Input TIMESTAMP BY Time
        GROUP BY
                [ ▼ ] (minute, 10)
)
SELECT
        Input.License_plate,
        Input.Make,
        Input.Time
FROM
        Input TIMESTAMP BY Time
        INNER JOIN LastInWindow
        ON [ ▼ ] (minute, Input, LastInWindow) BETWEEN 0 AND 10
AND Input.Time = LastInWindow.LastEventTime
```

- COUNT
- MAX
- MIN
- TOPONE

- HoppingWindow
- SessionWindow
- SlidingWindow
- TumblingWindow

- DATEADD
- DATEDIFF
- DATENAME
- DATEPART

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 6

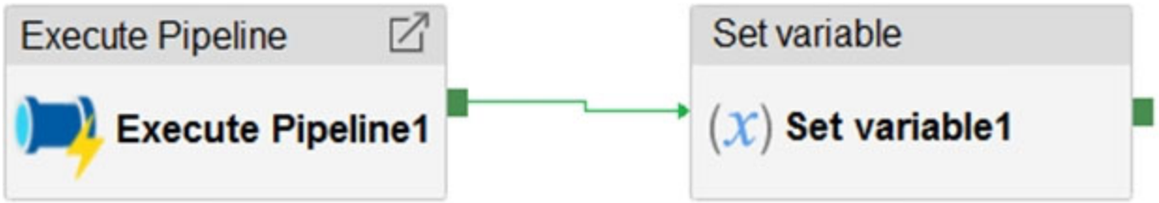
Topic #: 2

[All DP-203 Questions]

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.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 7

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

Ingest:

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

Store:

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

Prepare and Train:

- ✑ Provide a fully-managed and interactive workspace for exploration and visualization.
- ✑ Provide the ability to program in R, SQL, Python, Scala, and Java.

Provide seamless user authentication with Azure Active Directory.

▪

Model & Serve:

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

You need to build the data integration pipeline.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 8

Topic #: 2

[All DP-203 Questions]

DRAG DROP -

You have the following table named Employees.

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

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

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

NOTE: Each correct selection is worth one point.

Select and Place:

Values

CASE

ELSE

OVER

PARTITION BY

ROW_NUMBER

Answer Area

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 9

Topic #: 2

[All DP-203 Questions]

DRAG DROP -

You have an Azure Synapse Analytics workspace named WS1.

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

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

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

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

NOTE: Each correct selection is worth one point.

Select and Place:

Values

opendatasource

openjson

openquery

openrowset

Answer Area

```
select*

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

) as q
cross apply [          ](contextcustomdimensions)

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 10

Topic #: 2

[All DP-203 Questions]

DRAG DROP -

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

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

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

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

How should you complete the query? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all.

You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.

Select and Place:

Values

CAST

COLLATE

CONVERT

FLATTEN

PIVOT

UNPIVOT

Answer Area

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 11
Topic #: 2
[All DP-203 Questions]

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

A. a resource tag

B. a correlation ID

C. a run group ID

D. an annotation

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 12

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 13

Topic #: 2

[All DP-203 Questions]

You are designing a statistical analysis solution that will use custom proprietary Python functions on near real-time data from Azure Event Hubs.

You need to recommend which Azure service to use to perform the statistical analysis. The solution must minimize latency.

What should you recommend?

A. Azure Synapse Analytics

B. Azure Databricks

C. Azure Stream Analytics

D. Azure SQL Database

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 14

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 15

Topic #: 2

[All DP-203 Questions]

You need to implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool.

You have a table that was created by using the following Transact-SQL statement.

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

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

NOTE: Each correct selection is worth one point.

A. [EffectiveStartDate] [datetime] NOT NULL,

B. [CurrentProductCategory] [nvarchar] (100) NOT NULL,

C. [EffectiveEndDate] [datetime] NULL,

D. [ProductCategory] [nvarchar] (100) NOT NULL,

E. [OriginalProductCategory] [nvarchar] (100) NOT NULL,

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 16

Topic #: 2

[All DP-203 Questions]

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

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

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

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

Solution: You use a hopping window that uses a hop size of 10 seconds and a window size of 10 seconds.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 17

Topic #: 2

[All DP-203 Questions]

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 18

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 19
Topic #: 2
[All DP-203 Questions]

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.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 20

Topic #: 2

[All DP-203 Questions]

DRAG DROP -

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

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

NOTE: Each correct selection is worth one point.

Select and Place:

Values

- `all, ecommerce, retail, wholesale`
- `dept=='ecommerce', dept=='retail', dept=='wholesale'`
- `dept=='ecommerce', dept=='wholesale', dept=='retail'`
- `disjoint: false`
- `disjoint: true`
- `ecommerce, retail, wholesale, all`

Answer Area

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 21

Topic #: 2

[All DP-203 Questions]

DRAG DROP -

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

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

You create the following components:

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

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

Select and Place:

Actions	Answer Area
Mount the Data Lake Storage onto DBFS.	
Write the results to a table in Azure Synapse.	
Perform transformations on the file.	
Specify a temporary folder to stage the data.	
Write the results to Data Lake Storage.	
Read the file into a data frame.	
Drop the data frame.	
Perform transformations on the data frame.	

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 22

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

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

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

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

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Type: ▼

- Event
- On-demand
- Schedule
- Tumbling window

Additional properties: ▼

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 23

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

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

- ✑ Minimize latency from an Azure Event hub to the dashboard.
- ✑ Minimize the required storage.
- ✑ Minimize development effort.

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

NOTE: Each correct selection is worth one point

Hot Area:

Answer Area

Azure Stream Analytics input type:

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

Azure Stream Analytics output type:

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

Aggregation query location:

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 24

Topic #: 2

[All DP-203 Questions]

DRAG DROP -

You have an Azure Stream Analytics job that is a Stream Analytics project solution in Microsoft Visual Studio. The job accepts data generated by IoT devices in the JSON format.

You need to modify the job to accept data generated by the IoT devices in the Protobuf format.

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

Select and Place:

Actions	Answer Area
Change the Event Serialization Format to Protobuf in the input.json file of the job and reference the DLL.	
Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution.	
Add .NET deserializer code for Protobuf to the custom deserializer project.	
Add .NET deserializer code for Protobuf to the Stream Analytics project.	
Add an Azure Stream Analytics Application project to the solution.	

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 25

Topic #: 2

[All DP-203 Questions]

You have an Azure Storage account and a data warehouse in Azure Synapse Analytics in the UK South region.

You need to copy blob data from the storage account to the data warehouse by using Azure Data Factory. The solution must meet the following requirements:

✑ Ensure that the data remains in the UK South region at all times.

✑ Minimize administrative effort.

Which type of integration runtime should you use?

A. Azure integration runtime

B. Azure-SSIS integration runtime

C. Self-hosted integration runtime

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 26

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

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

You need to implement Azure Stream Analytics to calculate the average fare per mile by driver.

How should you configure the Stream Analytics input for each source? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

HubA: [▼]
- Stream
- Reference

HubB: [▼]
- Stream
- Reference

Database1: [▼]
- Stream
- Reference

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 27

Topic #: 2

[All DP-203 Questions]

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)

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 28

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

You need to calculate the difference in the number of readings per sensor per hour.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 29

Topic #: 2

[All DP-203 Questions]

You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container.

Which type of trigger should you use?

A. on-demand

B. tumbling window

C. schedule

D. event

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 30

Topic #: 2

[All DP-203 Questions]

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.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 31

Topic #: 2

[All DP-203 Questions]

You are developing a solution that will stream to Azure Stream Analytics. The solution will have both streaming data and reference data.

Which input type should you use for the reference data?

A. Azure Cosmos DB

B. Azure Blob storage

C. Azure IoT Hub

D. Azure Event Hubs

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 32

Topic #: 2

[All DP-203 Questions]

You are designing an Azure Stream Analytics job to process incoming events from sensors in retail environments.

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

Which type of window should you use?

A. snapshot

B. tumbling

C. hopping

D. sliding

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 33

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

You are designing a monitoring solution for a fleet of 500 vehicles. Each vehicle has a GPS tracking device that sends data to an Azure event hub once per minute.

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

You need to ensure that when a GPS position is outside the expected area, a message is added to another event hub for processing within 30 seconds. The solution must minimize cost.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

Window:
- Hopping
- No window
- Session
- Tumbling

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 34

Topic #: 2

[All DP-203 Questions]

You are designing an Azure Databricks table. The table will ingest an average of 20 million streaming events per day.

You need to persist the events in the table for use in incremental load pipeline jobs in Azure Databricks. The solution must minimize storage costs and incremental load times.

What should you include in the solution?

A. Partition by DateTime fields.

B. Sink to Azure Queue storage.

C. Include a watermark column.

D. Use a JSON format for physical data storage.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 35

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

You have a self-hosted integration runtime in Azure Data Factory.

The current status of the integration runtime has the following configurations:

- ✑ Status: Running
- ✑ Type: Self-Hosted
- ✑ Version: 4.4.7292.1
- ✑ Running / Registered Node(s): 1/1
- ✑ High Availability Enabled: False
- ✑ Linked Count: 0
- ✑ Queue Length: 0
- ✑ Average Queue Duration. 0.00s

The integration runtime has the following node details:

- ✑ Name: X-M
- ✑ Status: Running
- ✑ Version: 4.4.7292.1
- ✑ Available Memory: 7697MB
- ✑ CPU Utilization: 6%
- ✑ Network (In/Out): 1.21KBps/0.83KBps
- ✑ Concurrent Jobs (Running/Limit): 2/14
- ✑ Role: Dispatcher/Worker
- ✑ Credential Status: In Sync

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

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

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

- raised
- lowered
- left as is

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 36

Topic #: 2

[All DP-203 Questions]

You have an Azure Databricks workspace named workspace1 in the Standard pricing tier.

You need to configure workspace1 to support autoscaling all-purpose clusters. The solution must meet the following requirements:

- ✑ Automatically scale down workers when the cluster is underutilized for three minutes.
- ✑ Minimize the time it takes to scale to the maximum number of workers.
- ✑ Minimize costs.

What should you do first?

A. Enable container services for workspace1.

B. Upgrade workspace1 to the Premium pricing tier.

C. Set Cluster Mode to High Concurrency.

D. Create a cluster policy in workspace1.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 37

Topic #: 2

[All DP-203 Questions]

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 38

Topic #: 2

[All DP-203 Questions]

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

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

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

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

Solution: You use a session window that uses a timeout size of 10 seconds.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 39

Topic #: 2

[All DP-203 Questions]

You use Azure Stream Analytics to receive data from Azure Event Hubs and to output the data to an Azure Blob Storage account.

You need to output the count of records received from the last five minutes every minute.

Which windowing function should you use?

A. Session

B. Tumbling

C. Sliding

D. Hopping

Show Suggested Answer

Actual exam question from Microsoft's DP-203

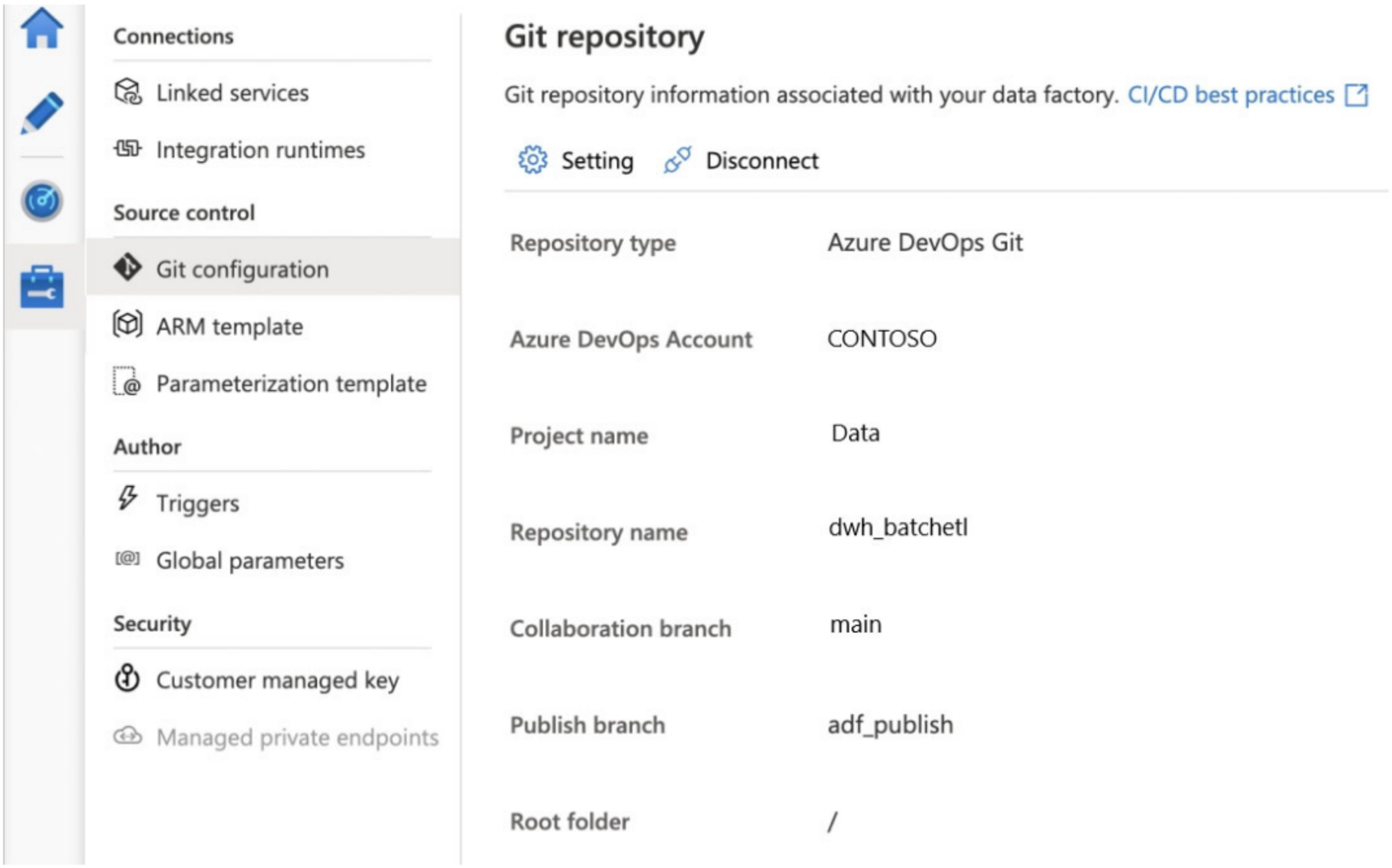
Question #: 40

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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.

Hot Area:

Answer Area

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

- /
- adf_publish
- main
- Parameterization template

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

- /
- /contososales
- /dwh_batchetl/adf_publish/contososales
- /main

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 41

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

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

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 42

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

ADF1 contains the following pipelines:

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

P1:

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

P2:

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 43

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Load methodology:

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

Trigger:

- Fixed schedule
- New file
- Tumbling window

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 44

Topic #: 2

[All DP-203 Questions]

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

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

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

- A workload for data engineers who will use Python and SQL.
- A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a Standard cluster for the data engineers, and a High Concurrency cluster for the jobs.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 45

Topic #: 2

[All DP-203 Questions]

You have the following Azure Data Factory pipelines:

- ✑ Ingest Data from System1
- ✑ Ingest Data from System2
- ✑ Populate Dimensions
- ✑ Populate Facts

Ingest Data from System1 and Ingest Data from System2 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest Data from System2. Populate Facts must execute after Populate Dimensions pipeline. All the pipelines must execute every eight hours.

What should you do to schedule the pipelines for execution?

A. Add an event trigger to all four pipelines.

B. Add a schedule trigger to all four pipelines.

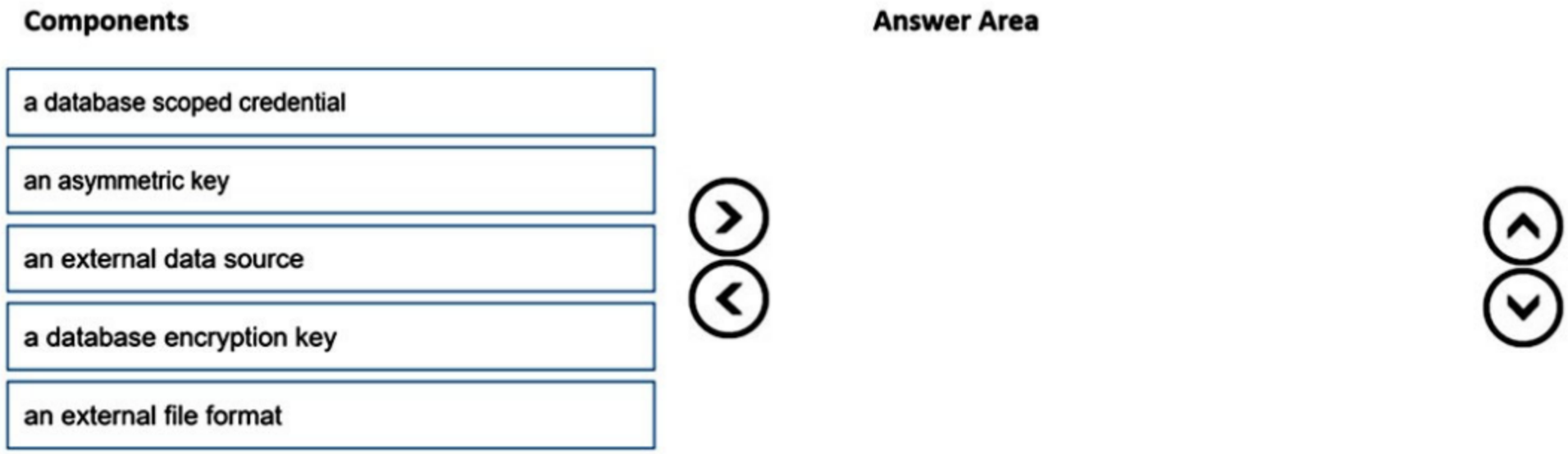
C. Create a patient pipeline that contains the four pipelines and use a schedule trigger.

D. Create a patient pipeline that contains the four pipelines and use an event trigger.

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 46
Topic #: 2
[All DP-203 Questions]

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



Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 47

Topic #: 2

[All DP-203 Questions]

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.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 48

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

You are building an Azure Stream Analytics job to retrieve game data.

You need to ensure that the job returns the highest scoring record for each five-minute time interval of each game.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

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

FROM input TIMESTAMP BY CreatedAt

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 49
Topic #: 2
[All DP-203 Questions]

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

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 50
Topic #: 2
[All DP-203 Questions]

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

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 51

Topic #: 2

[All DP-203 Questions]

You are designing an Azure Databricks cluster that runs user-defined local processes.

You need to recommend a cluster configuration that meets the following requirements:

- ✑ Minimize query latency.
- ✑ Maximize the number of users that can run queries on the cluster at the same time.
- ✑ Reduce overall costs without compromising other requirements.

Which cluster type should you recommend?

A. Standard with Auto Termination

B. High Concurrency with Autoscaling

C. High Concurrency with Auto Termination

D. Standard with Autoscaling

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 52

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

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

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

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

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

Parameter:

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

Naming pattern:

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

Copy behavior:

- Add dynamic content
- Flatten hierarchy
- Merge files

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 53

Topic #: 2

[All DP-203 Questions]

DRAG DROP -

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

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

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

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

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

NOTE: Each correct selection is worth one point.

Select and Place:

Values	Answer Area
{deviceID}	/ Value / Value / Value .json
{mm}/{HH}/{DD}/{MM}/{YYYY}	
{regionID}/{deviceID}	
{regionID}/raw	
{YYYY}/{MM}/{DD}/{HH}	
{YYYY}/{MM}/{DD}/{HH}/{mm}	
raw/{deviceID}	
raw/{regionID}	

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 54

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

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

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

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

You need to calculate the uptime between the faults.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
SELECT

DeviceID,

MIN(EventTime) as StartTime,

MAX(EventTime) as EndTime,

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

FROM input TIMESTAMP BY EventTime
```

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

```
GROUP BY

DeviceID
```

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 55

Topic #: 2

[All DP-203 Questions]

You are creating a new notebook in Azure Databricks that will support R as the primary language but will also support Scala and SQL.

Which switch should you use to switch between languages?

A. %<language>

B. @<Language >

C. \\[<language >]

D. \\(<language >)

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 56

Topic #: 2

[All DP-203 Questions]

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 57

Topic #: 2

[All DP-203 Questions]

You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account.

The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/.

You need to design a daily Azure Data Factory data load to minimize the data transfer between the two accounts.

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

NOTE: Each correct selection is worth one point

A. Specify a file naming pattern for the destination.

B. Delete the files in the destination before loading the data.

C. Filter by the last modified date of the source files.

D. Delete the source files after they are copied.

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 58
Topic #: 2
[All DP-203 Questions]

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

A. update

B. complete

C. append

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 59

Topic #: 2

[All DP-203 Questions]

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

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

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.

You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: In an Azure Synapse Analytics pipeline, you use a data flow that contains a Derived Column transformation.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 60

Topic #: 2

[All DP-203 Questions]

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

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

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.

You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: You use a dedicated SQL pool to create an external table that has an additional DateTime column.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 61

Topic #: 2

[All DP-203 Questions]

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

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

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.

You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.

You plan to insert data from the files in container1 into Table1 and transform the data. Each row of data in the files will produce one row in the serving layer of Table1.

You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.

Solution: You use an Azure Synapse Analytics serverless SQL pool to create an external table that has an additional DateTime column.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 62
Topic #: 2
[All DP-203 Questions]

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

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 63
Topic #: 2
[All DP-203 Questions]

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

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 64

Topic #: 2

[All DP-203 Questions]

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

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

You have an Azure Data Lake Storage account that contains a staging zone.

You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.

Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that executes mapping data flow, and then inserts the data into the data warehouse.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 65

Topic #: 2

[All DP-203 Questions]

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

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

You have an Azure Data Lake Storage account that contains a staging zone.

You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.

Solution: You schedule an Azure Databricks job that executes an R notebook, and then inserts the data into the data warehouse.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 66

Topic #: 2

[All DP-203 Questions]

You plan to create an Azure Data Factory pipeline that will include a mapping data flow.

You have JSON data containing objects that have nested arrays.

You need to transform the JSON-formatted data into a tabular dataset. The dataset must have one row for each item in the arrays.

Which transformation method should you use in the mapping data flow?

A. new branch

B. unpivot

C. alter row

D. flatten

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 67

Topic #: 2

[All DP-203 Questions]

You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account.

You need to output the count of tweets during the last five minutes every five minutes. Each tweet must only be counted once.

Which windowing function should you use?

A. a five-minute Sliding window

B. a five-minute Session window

C. a five-minute Hopping window that has a one-minute hop

D. a five-minute Tumbling window

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 68
Topic #: 2
[All DP-203 Questions]

You are planning a streaming data solution that will use Azure Databricks. The solution will stream sales transaction data from an online store. The solution has the following specifications:
The output data will contain items purchased, quantity, line total sales amount, and line total tax amount.

- ✑ Line total sales amount and line total tax amount will be aggregated in Databricks.
- ✑ Sales transactions will never be updated. Instead, new rows will be added to adjust a sale.

You need to recommend an output mode for the dataset that will be processed by using Structured Streaming. The solution must minimize duplicate data.
What should you recommend?

A. Update

B. Complete

C. Append

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 69

Topic #: 2

[All DP-203 Questions]

You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1.

You need to determine the size of the transaction log file for each distribution of DW1.

What should you do?

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

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

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

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 70

Topic #: 2

[All DP-203 Questions]

You are designing an anomaly detection solution for streaming data from an Azure IoT hub. The solution must meet the following requirements:

- ✑ Send the output to Azure Synapse.
- ✑ Identify spikes and dips in time series data.
- ✑ Minimize development and configuration effort.

Which should you include in the solution?

A. Azure Databricks

B. Azure Stream Analytics

C. Azure SQL Database

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 71

Topic #: 2

[All DP-203 Questions]

A company uses Azure Stream Analytics to monitor devices.

The company plans to double the number of devices that are monitored.

You need to monitor a Stream Analytics job to ensure that there are enough processing resources to handle the additional load.

Which metric should you monitor?

A. Early Input Events

B. Late Input Events

C. Watermark delay

D. Input Deserialization Errors

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 72

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

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

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

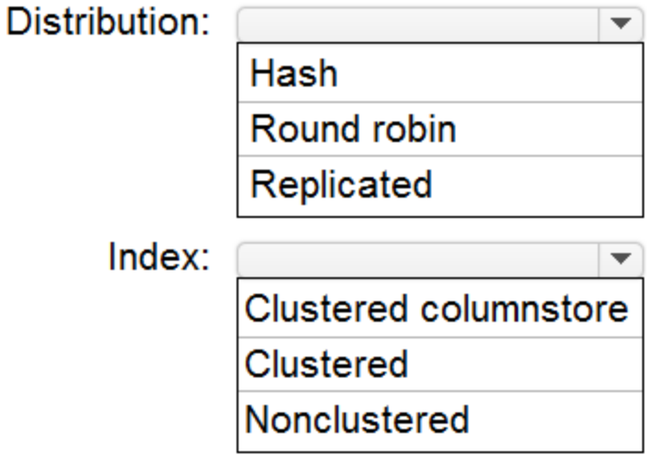
You need to recommend which distribution type and index type to use for the table. The solution must provide the fastest query performance.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area



Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 73

Topic #: 2

[All DP-203 Questions]

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 74

Topic #: 2

[All DP-203 Questions]

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 75

Topic #: 2

[All DP-203 Questions]

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

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

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

D. The pipeline was triggered too early.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 76

Topic #: 2

[All DP-203 Questions]

You have an Azure Synapse Analytics job that uses Scala.

You need to view the status of the job.

What should you do?

A. From Synapse Studio, select the workspace. From Monitor, select SQL requests.

B. From Azure Monitor, run a Kusto query against the AzureDiagnostics table.

C. From Synapse Studio, select the workspace. From Monitor, select Apache Sparks applications.

D. From Azure Monitor, run a Kusto query against the SparkLoggingEvent_CL table.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 77

Topic #: 2

[All DP-203 Questions]

DRAG DROP -

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

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

You create the following components:

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

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

Select and Place:

Actions	Answer Area
Mount the Data Lake Storage onto DBFS.	
Write the results to a table in Azure Synapse.	
Specify a temporary folder to stage the data.	
Read the file into a data frame.	
Perform transformations on the data frame.	

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 78
Topic #: 2
[All DP-203 Questions]

You have an Azure data factory named ADF1.
You currently publish all pipeline authoring changes directly to ADF1.
You need to implement version control for the changes made to pipeline artifacts. The solution must ensure that you can apply version control to the resources currently defined in the UX Authoring canvas for ADF1.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. From the UX Authoring canvas, select Set up code repository.

B. Create a Git repository.

C. Create a GitHub action.

D. Create an Azure Data Factory trigger.

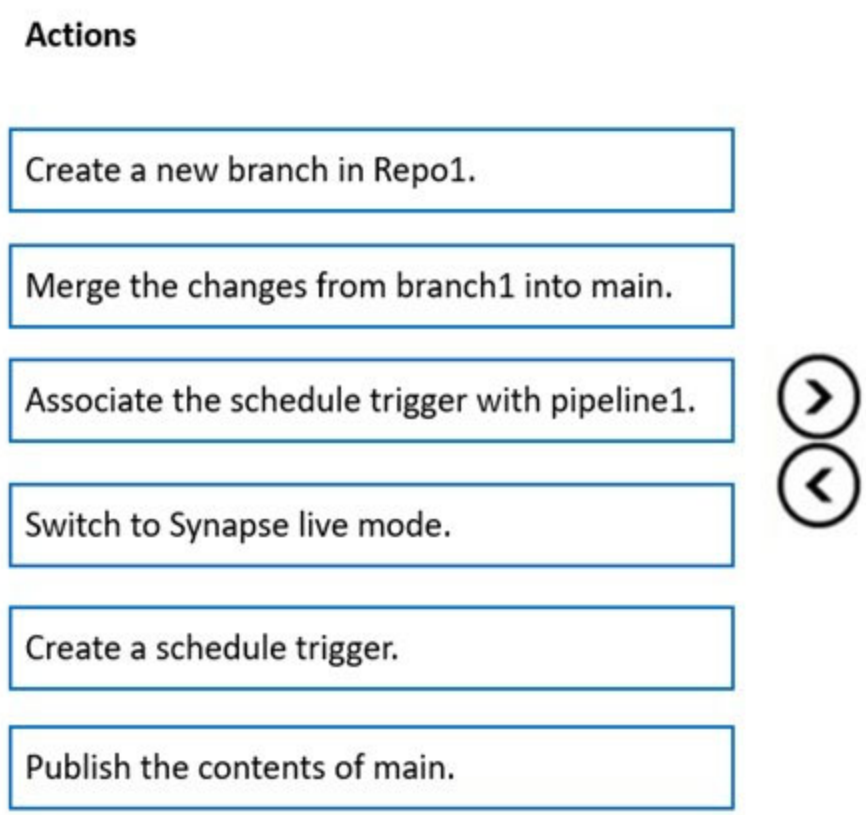
E. From the UX Authoring canvas, select Publish.

F. From the UX Authoring canvas, run Publish All.

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 79
Topic #: 2
[All DP-203 Questions]

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



Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 80

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers.

You need to create an external data source in Pool1 that will be used to read .orc files in storage1.

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

NOTE: Each correct selection is worth one point.

Hot Area:

Answer Area

```
CREATE EXTERNAL DATA SOURCE AzureDataLakeStore

WITH

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

credential = ADLS_credential ,

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

);
```

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 81

Topic #: 2

[All DP-203 Questions]

You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named SQLPool1.

SQLPool1 is currently paused.

You need to restore the current state of SQLPool1 to a new SQL pool.

What should you do first?

A. Create a workspace.

B. Create a user-defined restore point.

C. Resume SQLPool1.

D. Create a new SQL pool.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 82

Topic #: 2

[All DP-203 Questions]

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 correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. an X.509 certificate

B. an RSA key

C. an Azure virtual network that has a network security group (NSG)

D. an Azure Policy initiative

E. an Azure key vault that has purge protection enabled

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 83

Topic #: 2

[All DP-203 Questions]

You have an Azure Synapse Analytics serverless SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named storage1. The AllowBlobPublicAccess property is disabled for storage1.

You need to create an external data source that can be used by Azure Active Directory (Azure AD) users to access storage from Pool1.

What should you create first?

A. an external resource pool

B. an external library

C. database scoped credentials

D. a remote service binding

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 84

Topic #: 2

[All DP-203 Questions]

You have an Azure Data Factory pipeline named Pipeline1. Pipeline1 contains a copy activity that sends data to an Azure Data Lake Storage Gen2 account.

Pipeline1 is executed by a schedule trigger.

You change the copy activity sink to a new storage account and merge the changes into the collaboration branch.

After Pipeline1 executes, you discover that data is NOT copied to the new storage account.

You need to ensure that the data is copied to the new storage account.

What should you do?

A. Publish from the collaboration branch.

B. Create a pull request.

C. Modify the schedule trigger.

D. Configure the change feed of the new storage account.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 85

Topic #: 2

[All DP-203 Questions]

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

You need to ensure that pipeline1 will execute only if the previous execution completes successfully.

How should you configure the self-dependency for Trigger1?

A. offset: "-00:01:00" size: "00:01:00"

B. offset: "01:00:00" size: "-01:00:00"

C. offset: "01:00:00" size: "01:00:00"

D. offset: "-01:00:00" size: "01:00:00"

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 86

Topic #: 2

[All DP-203 Questions]

HOTSPOT -

You have an Azure Synapse Analytics pipeline named Pipeline1 that contains a data flow activity named Dataflow1.

Pipeline1 retrieves files from an Azure Data Lake Storage Gen 2 account named storage1.

Dataflow1 uses the AutoResolveIntegrationRuntime integration runtime configured with a core count of 128.

You need to optimize the number of cores used by Dataflow1 to accommodate the size of the files in storage1.

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

Hot Area:

Answer Area

To Pipeline1, add:

A custom activity
A Get Metadata activity
An If Condition activity

For Dataflow1, set the core count by using:

Dynamic content
Parameters
User properties

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 87

Topic #: 2

[All DP-203 Questions]

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

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

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

- A workload for data engineers who will use Python and SQL.
- A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 88

Topic #: 2

[All DP-203 Questions]

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

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

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

- A workload for data engineers who will use Python and SQL.
- A workload for jobs that will run notebooks that use Python, Scala, and SQL.
- A workload that data scientists will use to perform ad hoc analysis in Scala and R.

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databricks clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a High Concurrency cluster for the jobs.

Does this meet the goal?

A. Yes

B. No

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 89

Topic #: 2

[All DP-203 Questions]

You are designing a folder structure for the files in an Azure Data Lake Storage Gen2 account. The account has one container that contains three years of data.

You need to recommend a folder structure that meets the following requirements:

- ✑ Supports partition elimination for queries by Azure Synapse Analytics serverless SQL pools
- ✑ Supports fast data retrieval for data from the current month
- ✑ Simplifies data security management by department

Which folder structure should you recommend?

A. \Department\DataSource\YYYY\MM\DataFile_YYYYMMDD.parquet

B. \DataSource\Department\YYYYMM\DataFile_YYYYMMDD.parquet

C. \DD\MM\YYYY\Department\DataSource\DataFile_DDMMYY.parquet

D. \YYYY\MM\DD\Department\DataSource\DataFile_YYYYMMDD.parquet

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 90

Topic #: 2

[All DP-203 Questions]

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.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 91

Topic #: 2

[All DP-203 Questions]

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. Azure Stream Analytics and Azure Synapse notebooks

B. Structured Streaming in Azure Databricks

C. event triggers in Azure Data Factory

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 92

Topic #: 2

[All DP-203 Questions]

You are creating an Apache Spark job in Azure Databricks that will ingest JSON-formatted data.

You need to convert a nested JSON string into a DataFrame that will contain multiple rows.

Which Spark SQL function should you use?

A. explode

B. filter

C. coalesce

D. extract

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 93

Topic #: 2

[All DP-203 Questions]

DRAG DROP

-

You have an Azure subscription that contains an Azure Databricks workspace. The workspace contains a notebook named Notebook1.

In Notebook1, you create an Apache Spark DataFrame named df_sales that contains the following columns:

- Customer
- SalesPerson
- Region
- Amount

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

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

NOTE: Each correct selection is worth one point.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 94

Topic #: 2

[All DP-203 Questions]

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 95
Topic #: 2
[All DP-203 Questions]

DRAG DROP
-

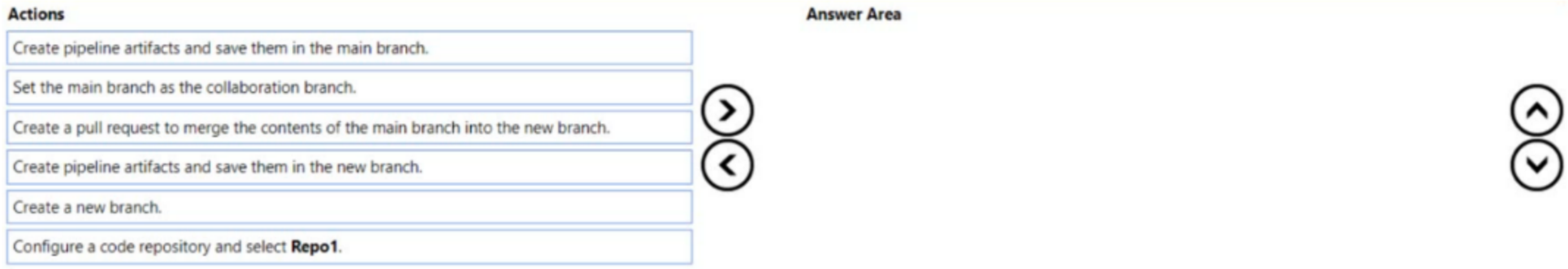
You have a project in Azure DevOps that contains a repository named Repo1. Repo1 contains a branch named main.

You create a new Azure Synapse workspace named Workspace1.

You need to create data processing pipelines in Workspace1. The solution must meet the following requirements:

- Pipeline artifacts must be stored in Repo1
- Source control must be provided for pipeline artifacts.
- All development must be performed in a feature branch.

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



Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 96
Topic #: 2
[All DP-203 Questions]

You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1.

You need to use an Azure Synapse pipeline to copy data from the selected tables in DB1 to the files in storage1. The solution must meet the following requirements:

- The Copy activity in the pipeline must be parameterized to use the data in File1.txt to identify the source and destination of the copy.
- Copy activities must occur in parallel as often as possible.

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

NOTE: Each correct selection is worth one point.

A. Get Metadata

B. Lookup

C. ForEach

D. If Condition

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 97

Topic #: 2

[All DP-203 Questions]

You have an Azure data factory that connects to a Microsoft Purview account. The data factory is registered in Microsoft Purview.

You update a Data Factory pipeline.

You need to ensure that the updated lineage is available in Microsoft Purview.

What should you do first?

A. Disconnect the Microsoft Purview account from the data factory.

B. Execute the pipeline.

C. Execute an Azure DevOps build pipeline.

D. Locate the related asset in the Microsoft Purview portal.

Show Suggested Answer

Actual exam question from Microsoft's DP-203

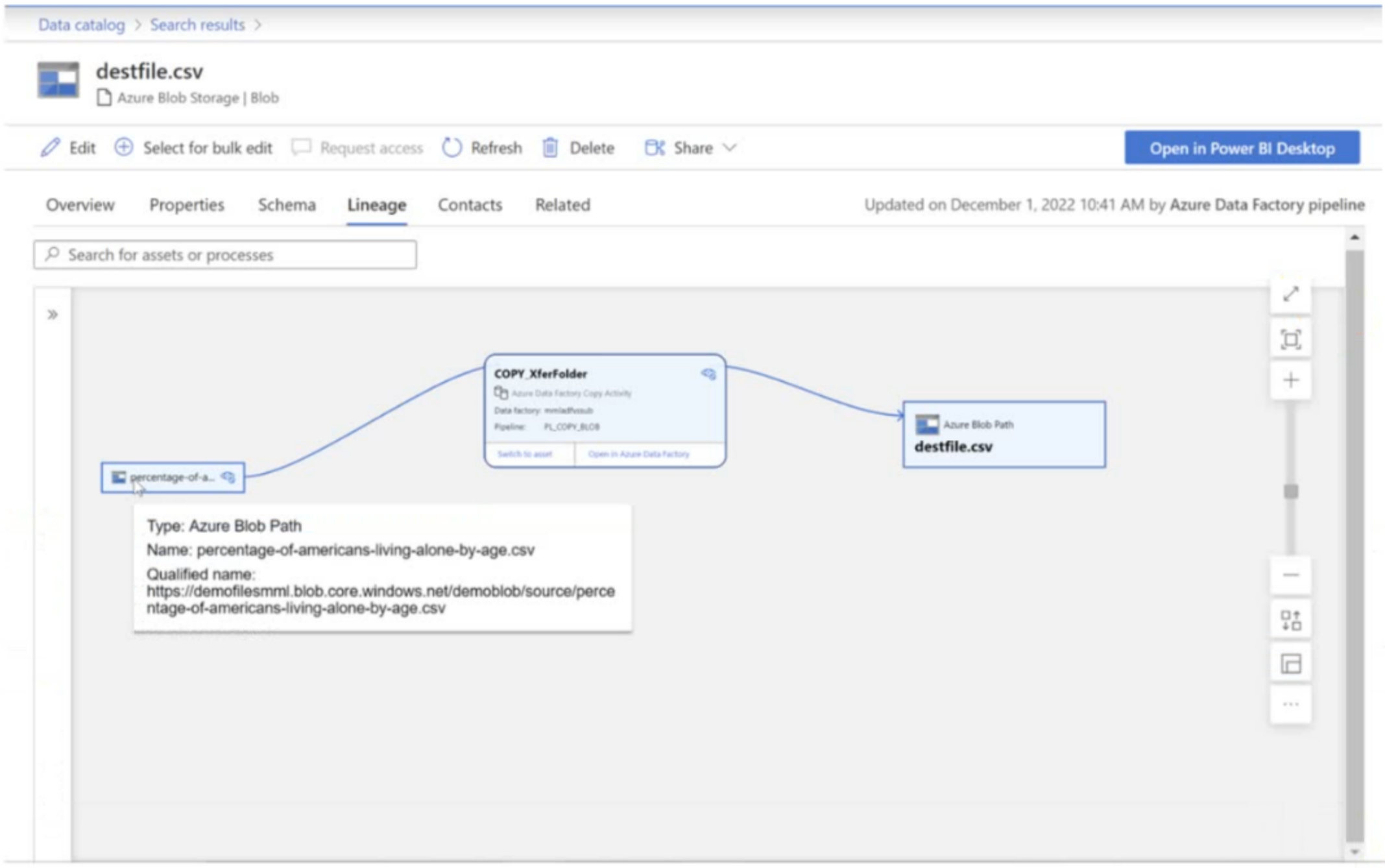
Question #: 98

Topic #: 2

[All DP-203 Questions]

You have a Microsoft Purview account.

The Lineage view of a CSV file is shown in the following exhibit.



How is the data for the lineage populated?

A. manually

B. by scanning data stores

C. by executing a Data Factory pipeline

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 99

Topic #: 2

[All DP-203 Questions]

You have an Azure subscription that contains a Microsoft Purview account named MP1, an Azure data factory named DF1, and a storage account named storage1. MP1 is configured to scan storage1. DF1 is connected to MP1 and contains a dataset named DS1. DS1 references a file in storage1.

In DF1, you plan to create a pipeline that will process data from DS1.

You need to review the schema and lineage information in MP1 for the data referenced by DS1.

Which two features can you use to locate the information? Each correct answer presents a complete solution.

NOTE: Each correct answer is worth one point.

A. the search bar in the Microsoft Purview governance portal

B. the Storage browser of storage1 in the Azure portal

C. the search bar in the Azure portal

D. the search bar in Azure Data Factory Studio

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 100
Topic #: 2
[All DP-203 Questions]

HOTSPOT
-

You have an Azure Blob storage account that contains a folder. The folder contains 120,000 files. Each file contains 62 columns.

Each day, 1,500 new files are added to the folder.

You plan to incrementally load five data columns from each new file into an Azure Synapse Analytics workspace.

You need to minimize how long it takes to perform the incremental loads.

What should you use to store the files and in which format? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Storage:
- Multiple blob storage accounts
- Multiple containers in the blob storage account
- Timeslice partitioning in the folders

Format:
- Apache Parquet
- CSV
- JSON

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 101
Topic #: 2
[All DP-203 Questions]

DRAG DROP
-

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.

Values

- BEGIN DISTRIBUTED TRANSACTION
- BEGIN TRAN
- COMMIT TRAN
- ROLLBACK TRAN
- SET RESULT_SET_CACHING ON

Answer Area

```
[                                    ]
BEGIN TRY
 INSERT INTO dbo.Table1 (col1, col2, col3)
 SELECT col1, col2, col3 FROM stage.Table1;
END TRY
 BEGIN CATCH
   IF @@TRANCOUNT > 0
     BEGIN
         [                                    ];
     END
      END CATCH;
IF @@TRANCOUNT >0
BEGIN
  COMMIT TRAN;
END
```

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 102
Topic #: 2
[All DP-203 Questions]

HOTSPOT
-

You have two Azure SQL databases named DB1 and DB2.

DB1 contains a table named Table1. Table1 contains a timestamp column named LastModifiedOn. LastModifiedOn contains the timestamp of the most recent update for each individual row.

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

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

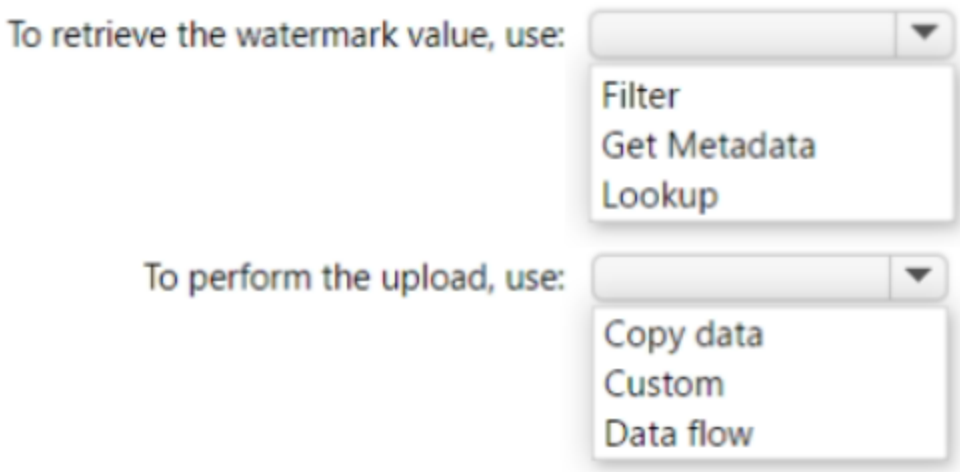
You need to identify which activities to include in the pipeline. The solution must meet the following requirements:

- Minimize the effort to author the pipeline.
- Ensure that the number of data integration units allocated to the upload operation can be controlled.

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

NOTE: Each correct answer is worth one point.

Answer Area



Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 103

Topic #: 2

[All DP-203 Questions]

HOTSPOT

-

You have an Azure Synapse serverless SQL pool.

You need to read JSON documents from a file by using the OPENROWSET function.

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 Area

```
SELECT *
FROM OPENROWSET
(
  BULK
'https://sourcedatalake.blob.core.windows.net/public/docs.json',
  FORMAT =
  FIELDTERMINATOR = '0x0b',
  FIELDQUOTE =
  ROWTERMINATOR = '0x0b'
)
WITH (jsondoc nvarchar(max) AS JsonDocuments
```

FORMAT = 'CSV' / 'DELTA' / 'JSON' / 'PARQUET'

FIELDQUOTE = '0x09' / '0x0a' / '0x0b' / '0x0c'

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 104

Topic #: 2

[All DP-203 Questions]

You use Azure Data Factory to create data pipelines.

You are evaluating whether to integrate Data Factory and GitHub for source and version control.

What are two advantages of the integration? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. additional triggers

B. lower pipeline execution times

C. the ability to save without publishing

D. the ability to save pipelines that have validation issues

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 105

Topic #: 2

[All DP-203 Questions]

DRAG DROP

-

You have an Azure Synapse Analytics workspace named Workspace1.

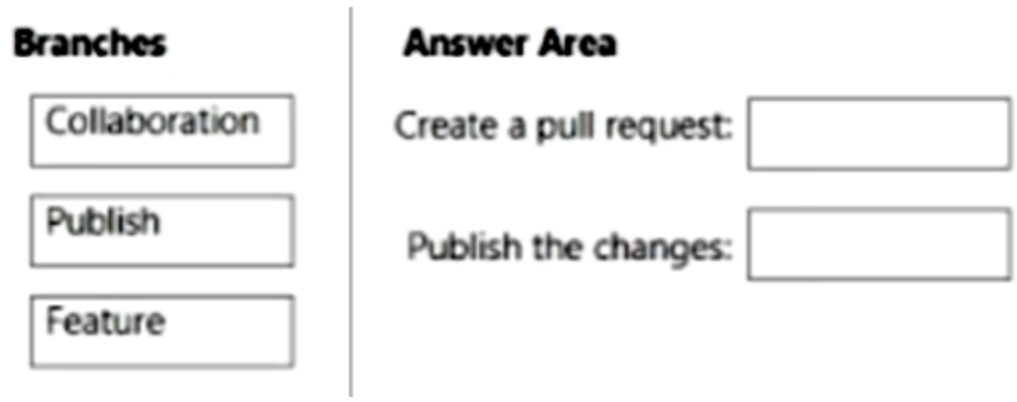
You perform the following changes:

- Implement source control for Workspace1.
- Create a branch named Feature based on the collaboration branch.
- Switch to the Feature branch.
- Modify Workspace1.

You need to publish the changes to Azure Synapse.

From which branch should you perform each change? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point



Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 106
Topic #: 2
[All DP-203 Questions]

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

You plan to create an Azure Data Factory pipeline that will use scheduled intervals to replicate newly created or modified blobs from account1 to account2.

You need to recommend a solution to implement the pipeline. The solution must meet the following requirements:
• Ensure that the pipeline only copies blobs that were created or modified since the most recent replication event.
• Minimize the effort to create the pipeline.

What should you recommend?

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

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

C. Create a pipeline that contains a flowlet.

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

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 107

Topic #: 2

[All DP-203 Questions]

You have an Azure Data Factory pipeline named pipeline1 that contains a data flow activity named activity1.

You need to run pipeline1.

Which runtime will be used to run activity1?

A. Azure Integration runtime

B. Self-hosted integration runtime

C. SSIS integration runtime

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 108

Topic #: 2

[All DP-203 Questions]

HOTSPOT

-

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 contains a dedicated SQL pool named SQLPool1 and an Apache Spark pool named sparkpool1. Sparkpool1 contains a DataFrame named pyspark_df.

You need to write the contents of pyspark_df to a table in SQLPool1 by using a PySpark notebook.

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

NOTE: Each correct selection is worth one point.

Answer Area

```
pyspark_df.createOrReplaceTempView("pysparkdftemptable")

[ ▼ ]
%%local
%%spark
%%sql

val scala_df = spark.sqlContext.sql ("select * from pysparkdftemptable")
scala_df.write. [ ▼ ] ("sqlpool1.dbo.PySparkTable", Constants.INTERNAL)
jdbc
saveAsTable
synapsesql
```

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 109
Topic #: 2
[All DP-203 Questions]

You have an Azure data factory named ADF1 and an Azure Synapse Analytics workspace that contains a pipeline named SynPipeLine1. SynPipeLine1 includes a Notebook activity.

You create a pipeline in ADF1 named ADFPipeline1.

You need to invoke SynPipeLine1 from ADFPipeline1.

Which type of activity should you use?

A. Web

B. Spark

C. Custom

D. Notebook

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 110

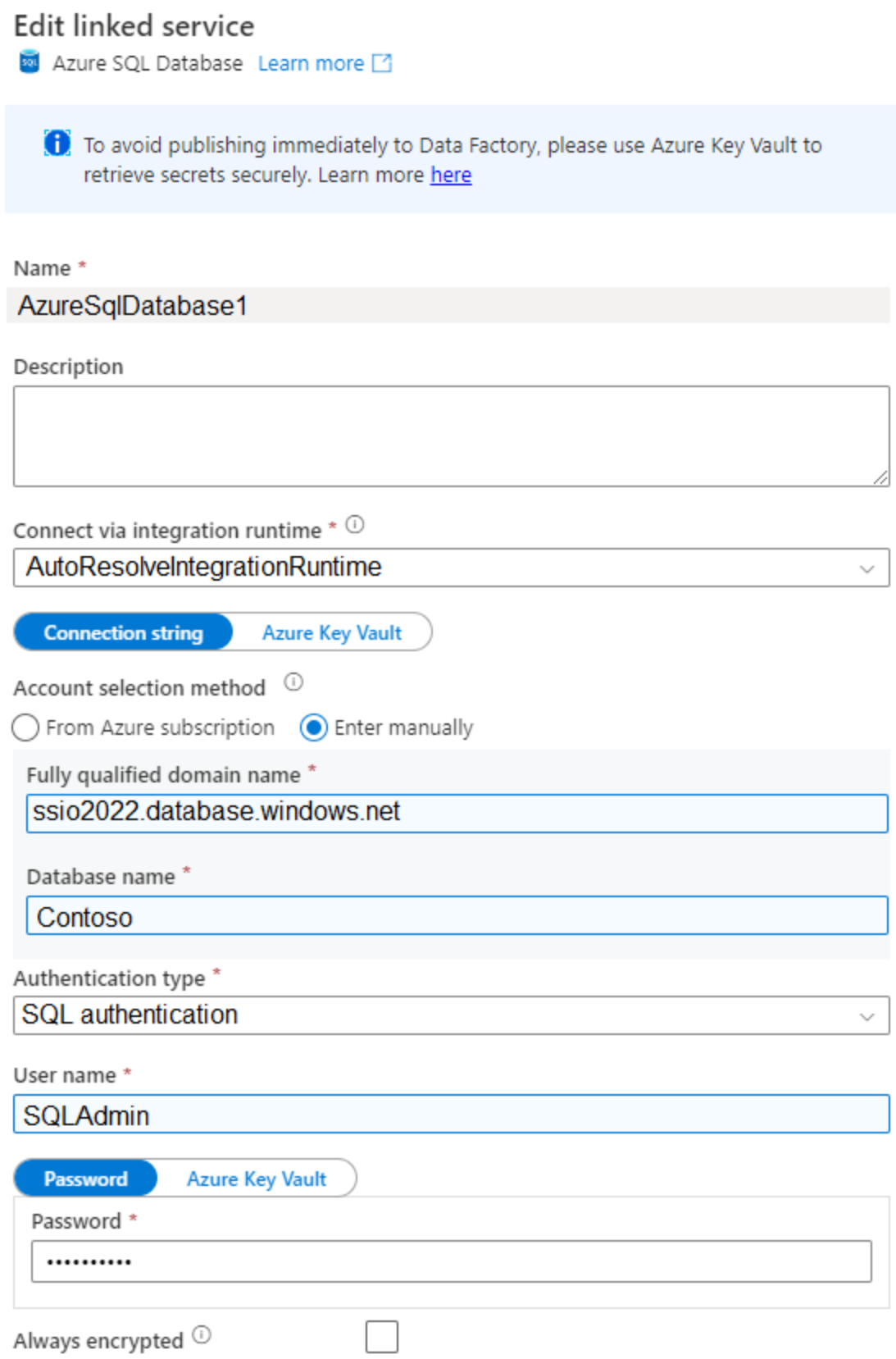
Topic #: 2

[All DP-203 Questions]

HOTSPOT

-

You have an Azure data factory that contains the linked service 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 answer is worth one point.

Answer Area

When working in a feature branch, changes to the linked service will be published to the live service

- upon publishing the changes
- upon saving the changes
- when the changes are merged into the collaboration branch

A Copy activity that uses the linked service as the source will perform the Copy activity

- in the region of the data factory
- in the region of the selected external compute
- in the region of the source database

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 111

Topic #: 2

[All DP-203 Questions]

HOTSPOT

-

In Azure Data Factory, you have a schedule trigger that is scheduled in Pacific Time.

Pacific Time observes daylight saving time.

The trigger has the following JSON file.

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

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

NOTE: Each correct selection is worth one point.

Answer Area

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

- one time
- two times
- zero times

The trigger **[answer choice]** daylight saving time.

- is unaffected by
- will automatically adjust for
- will require an adjustment for

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 112
Topic #: 2
[All DP-203 Questions]

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a pipeline that will execute a stored procedure in the dedicated SQL pool and use the returned result set as the input for a downstream activity. The solution must minimize development effort.

Which type of activity should you use in the pipeline?

A. U-SQL

B. Stored Procedure

C. Script

D. Notebook

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 113

Topic #: 2

[All DP-203 Questions]

You have an Azure SQL database named DB1 and an Azure Data Factory data pipeline named pipeline1.

From Data Factory, you configure a linked service to DB1.

In DB1, you create a stored procedure named SP1. SP1 returns a single row of data that has four columns.

You need to add an activity to pipeline1 to execute SP1. The solution must ensure that the values in the columns are stored as pipeline variables.

Which two types of activities can you use to execute SP1? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. Script

B. Copy

C. Lookup

D. Stored Procedure

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 114
Topic #: 2
[All DP-203 Questions]

You have an Azure data factory named ADF1.

You currently publish all pipeline authoring changes directly to ADF1.

You need to implement version control for the changes made to pipeline artifacts. The solution must ensure that you can apply version control to the resources currently defined in the Azure Data Factory Studio for ADF1.

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

NOTE: Each correct selection is worth one point.

A. From the Azure Data Factory Studio, run Publish All.

B. Create an Azure Data Factory trigger.

C. Create a Git repository.

D. Create a GitHub action.

E. From the Azure Data Factory Studio, select Set up code repository.

F. From the Azure Data Factory Studio, select Publish.

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 115
Topic #: 2
[All DP-203 Questions]

You have an Azure data factory named ADF1 that contains a pipeline named Pipeline1.

Pipeline1 must execute every 30 minutes with a 15-minute offset.

You need to create a trigger for Pipeline1. The trigger must meet the following requirements:

- Backfill data from the beginning of the day to the current time.
- If Pipeline1 fails, ensure that the pipeline can re-execute within the same 30-minute period.
- Ensure that only one concurrent pipeline execution can occur.
- Minimize development and configuration effort.

Which type of trigger should you create?

A. schedule

B. event-based

C. manual

D. tumbling window

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 116

Topic #: 2

[All DP-203 Questions]

You have an Azure Data Lake Storage Gen2 account named account1 and an Azure event hub named Hub1. Data is written to account1 by using Event Hubs Capture.

You plan to query account by using an Apache Spark pool in Azure Synapse Analytics.

You need to create a notebook and ingest the data from account1. The solution must meet the following requirements:

- Retrieve multiple rows of records in their entirety.
- Minimize query execution time.
- Minimize data processing.

Which data format should you use?

A. Parquet -

O. Avro

C. ORC

D. JSON

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 117

Topic #: 2

[All DP-203 Questions]

You have an Azure Blob Storage account named blob1 and an Azure Data Factory pipeline named pipeline1.

You need to ensure that pipeline1 runs when a file is deleted from a container in blob1. The solution must minimize development effort.

Which type of trigger should you use?

A. schedule

B. storage event

C. tumbling window

D. custom event

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 118

Topic #: 2

[All DP-203 Questions]

HOTSPOT

-

You have Azure Data Factory configured with Azure Repos Git integration. The collaboration branch and the publish branch are set to the default values.

You have a pipeline named pipeline1.

You build a new version of pipeline1 in a branch named feature1.

From the Data Factory Studio, you select Publish.

The source code of which branch will be built, and which branch will contain the output of the Azure Resource Manager (ARM) template? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Answer Area

Source code: [ ]
- adf_publish
- feature1
- main

ARM template output: [ ]
- adf_publish
- feature1
- main

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 119
Topic #: 2
[All DP-203 Questions]

DRAG DROP
-

You have an Azure subscription that contains an Azure data factory.

You are editing an Azure Data Factory activity JSON.

The script needs to copy a file from Azure Blob Storage to multiple destinations. The solution must ensure that the source and destination files have consistent folder paths.

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

NOTE: Each correct selection is worth one point.

Values

- FlattenHierarchy
- ForEach
- MergeFiles
- PreserveHierarchy
- Switch
- Until

Answer Area

```
{
    "name": "Pipeline1",
    "properties": {
       "activities": [
         {
         "name": "Activity1",
         "type": [          ],
         "typeProperties": {
                        "isSequential": "true",
                        "items": {
                              "value": "@pipeline
        ().parameters.mySinkDatasetFolderPath",
                              "type": "Expression"},
                        "activities" [
                              {
                                    "name": "MyCopyActivity",
                                "type": "Copy",
                                "typeProperties": {
                                   "source": {
                                        "type": "BlobSource",
                                        "recursive": "false" },
                                   "sink": {
                                        "type": "BlobSink",
                                        "CopyBehavior": [          ]
                                …
                                }
```

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 120

Topic #: 2

[All DP-203 Questions]

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. alter row

C. surrogate key

D. select

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 121

Topic #: 2

[All DP-203 Questions]

You have an on-premises database named db1 and a set-hosted integration runtime.

You have an Azure subscription that contains an Azure Data Lake Storage account named dl1.

You need to develop four data pipeline projects that will use Microsoft Power Query to copy data from db1 to dl1. The solution must meet the following requirements:

- All pipelines must use the self-hosted integration runtime.
- Each project must be stored in a separate Git repository.
- Development effort must be minimized.

What should you use?

A. Azure Synapse Analytics

B. Azure Logic Apps.

C. Azure Data Factory

D. Microsoft Power BI

Show Suggested Answer

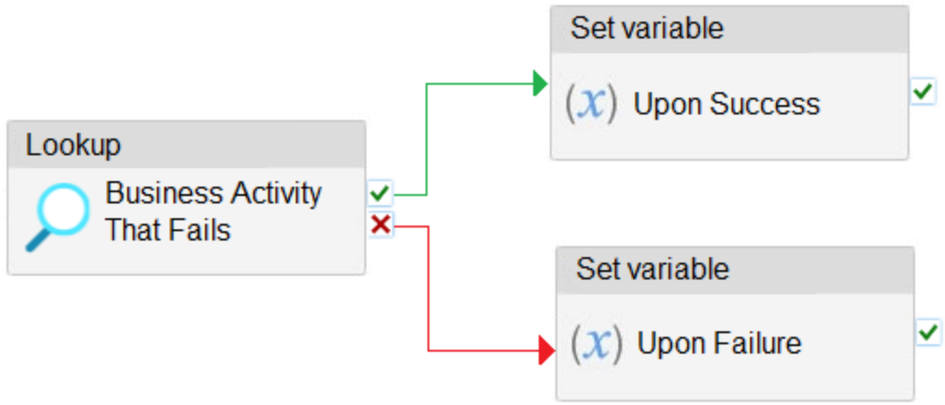
Actual exam question from Microsoft's DP-203

Question #: 122

Topic #: 2

[All DP-203 Questions]

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 skipped dependency on the Upon Failure activity

B. a skipped dependency on the Upon Success activity

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

D. a failure dependency on the Upon Failure activity

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 123

Topic #: 2

[All DP-203 Questions]

You have an on-premises Linux server that contains a database named DB1.

You have an Azure subscription that contains an Azure data factory named ADF1 and an Azure Data Lake Storage account named ADLS1.

You need to create a pipeline in ADF1 that will copy data from DB1 to ADLS1.

Which type of integration runtime should you use to read the data from DB1?

A. self-hosted integration runtime

B. Azure integration runtime

C. Azure-SQL Server Integration Services (SSIS)

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 124

Topic #: 2

[All DP-203 Questions]

DRAG DROP

-

You have an Azure Data Lake Storage account named account1.

You use an Azure Synapse Analytics serverless SQL pool to access sales data stored in account1.

You need to create a bar chart that displays sales by product. The solution must minimize development effort.

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.

Actions

- Add a `SELECT` statement that will return the sales by product data.
- Switch to the Chart view.
- Modify the Chart settings.
- Create a SQL script by using Synapse Studio.
- Execute the script.

Answer Area

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 125
Topic #: 2
[All DP-203 Questions]

DRAG DROP
-

You have an Azure Synapse Analytics dedicated SQL pool.

You need to create a copy of the data warehouse and make the copy available for 28 days. The solution must minimize costs.

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.

Actions

- Create a new user-defined restore point.
- Restore the latest automatic restore point to a new data warehouse.
- Pause the restored data warehouse.
- Restore the copy from the latest automatic restore point to the current data warehouse.
- Restore the copy from the new user-defined restore point to a new data warehouse.

Answer Area

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 126

Topic #: 2

[All DP-203 Questions]

HOTSPOT

-

You have an Azure Synapse Analytics workspace that contains an Apache Spark pool named Pool1.

You need to read data from a CSV file and write the data to a Delta table by using Pool1.

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

NOTE: Each correct selection is worth one point.

Answer Area

```
from delta.tables import *
from pyspark.sql.functions import *
df = spark.read.load
('abfss://container@mydatalake.dfs.core.windows.net/stage/
products.csv', format = 'csv', header = True)
delta_table_path = "/delta/products-delta"
df. [ ▼ ] .save(delta_table_path)
```

- cache()
- inputFiles()
- write.format("delta")
- write.parquet

```
deltaTable = [ ▼ ] (spark, delta_table_path)
```

- deltaTable.alias
- deltaTable.convertToDelta
- deltaTable.forPath
- deltaTable.update

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 127
Topic #: 2
[All DP-203 Questions]

HOTSPOT
-

You have an Azure Data Lake Storage account that contains one CSV file per hour for January 1, 2020, through January 31, 2023. The files are partitioned by using the following folder structure.

```
csv/system1/{year}/{month}/{filename}.csv
```

You need to query the files by using an Azure Synapse Analytics serverless SQL pool. The solution must return the row count of each file created during the last three months of 2022.

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

Answer Area

```
SELECT
     r.filepath() AS filepath
     ,COUNT_BIG(*) AS [rows]
FROM OPENROWSET(
     BULK [ ▼ ]
          'csv/system1/2022',
          'csv/system1/2022/',
          'csv/system1/2022/*/*.csv',
DATA_SOURCE = 'MyDataLake',
     FORMAT = 'CSV',
     PARSER_VERSION = '2.0',
     FIRSTROW = 2 )
WITH (vendor_id INT) AS [r]
WHERE
  [ ▼ ] IN ('10', '11', '12')
     r.filepath()
     r.filepath(1)
     r.filepath(2)
GROUP BY
```

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 128

Topic #: 2

[All DP-203 Questions]

HOTSPOT

-

You have an Azure Synapse Analytics dedicated SQL pool named Pool1 that contains an external table named Sales. Sales contains sales data. Each row in Sales contain data on a single sale, including the name of the salesperson.

You need to implement row-level security (RLS). The solution must ensure that the salespeople can access only their respective sales.

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

NOTE: Each correct selection is worth one point.

Answer Area

Create:

- A materialized view in Pool1
- A security policy for Sales
- Database scoped credentials in Pool1

Restrict row access by using:

- A masking rule
- A table-valued function
- The CONTAINS predicate

Show Suggested Answer

Actual exam question from Microsoft's DP-203
Question #: 129
Topic #: 2
[All DP-203 Questions]

You have an Azure Data Factory pipeline named P1.

You need to schedule P1 to run at 10:15 AM, 12:15 PM, 2:15 PM, and 4:15 PM every day.

Which frequency and interval should you configure for the scheduled trigger?

A. Frequency: Month -
Interval: 1

B. Frequency: Day -
Interval: 1

C. Frequency: Minute -
Interval: 60

D. Frequency: Hour -
Interval: 2

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 130

Topic #: 2

[All DP-203 Questions]

You are creating an Azure Data Factory pipeline.

You need to add an activity to the pipeline. The activity must execute a Transact-SQL stored procedure that has the following characteristics:

- Returns the number of sales invoices for a current date
- Does NOT require input parameters

Which type on activity should you use?

A. Stored Procedure

B. Get Metadata

C. Append Variable

D. Lookup

Show Suggested Answer

Actual exam question from Microsoft's DP-203

Question #: 131

Topic #: 2

[All DP-203 Questions]

HOTSPOT

-

You have an Azure Synapse Analytics workspace that contains three pipelines and three triggers named Trigger1, Trigger2, and Trigger3.

Trigger3 has the following definition.

```
...
{
  "name": "Trigger3",
  "properties": {
    "annotations": [],
    "runtimeState": "Stopped",
    "pipeline": {
      "pipelineReference": {
         "referenceName": "Pipeline 3",
         "type": "PipelineReference"
       }
    },
  },
  "type": "TumblingWindowTrigger",
  "typeProperties": {
    "frequency": "Hour",
    "interval": 1,
    "startTime": "2023-11-12T11:00:00Z",
    "delay": "00:00:00",
    "maxConcurrency": 1,
    "retryPolicy": {
      "intervalInSeconds": 30
    },
    "dependsOn": [
      {
        "type": "TumblingWindowTriggerDependencyReference",
        "size": "0.03:00:00",
        "offset": "-0.02:00:00",
        "referenceTrigger": {
          "referenceName": "Trigger1",
          "type": "TriggerReference"
        }
      },
      {
        "type": "TumblingWindowTriggerDependencyReference",
        "size": "0.03:00:00",
        "offset": "-0.02:00:00",
        "referenceTrigger": {
          "referenceName": "Trigger2",
          "type": "TriggerReference"
        }
      },
      {
        "type": "SelfDependencyTumblingWindowTriggerReference",
        "offset": "-0.03:00:00"
      }
    ]
  }
}
...
```

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 Area

Statements	Yes	No
Pipeline3 will execute when Trigger3 fires.	○	○
Up to three instances of Trigger3 can fire simultaneously.	○	○
Trigger3 will fire three hours after Trigger1 has fired three times, and Trigger2 has fired three times.	○	○

Show Suggested Answer