QUESTIONS & ANSWERS

Kill your exam at first Attempt

KILL EXAMS

DP-203 Dumps
DP-203 Braindumps
DP-203 Real Questions
DP-203 Practice Test
DP-203 dumps free

Microsoft

DP-203

Data Engineering on Microsoft Azure

Question: 74

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

Answer: D

Explanation:

There's an important reason to put the date at the end of the directory structure. If you want to lock down certain regions or subject matters to users/groups, then you can easily do so with the POSIX permissions. Otherwise, if there was a need to restrict a certain security group to viewing just the UK data or certain planes, with the date structure in front a separate permission would be required for numerous directories under every hour directory. Additionally, having the date structure in front would exponentially increase the number of directories as time went on.

Note: In IoT workloads, there can be a great deal of data being landed in the data store that spans across numerous products, devices, organizations, and customers. It's important to pre-plan the directory layout for organization, security, and efficient processing of the data for down-stream consumers. A general template to consider might be the following layout:

{Region}/{SubjectMatter(s)}/{yyyy}/{mm}/{dd}/{hh}/

Question: 75

HOTSPOT

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

You need to improve the performance of queries against FactOnlineSales by using table partitions.

The solution must meet the following requirements:

Create four partitions based on the order date.

Ensure that each partition contains all the orders places during a given calendar year.

How should you complete the T-SQL command? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

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

[RIGHT / LEFT]

([▼])

[20090101,20121231 / 20100101,20110101,20120101 / 20090101,20100101,20110101,20120101]

Answer:

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

Explanation:

Text

Description automatically generated

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

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

But if you use range RIGHT and 20100101,20110101,20120101

Partition will be, datecol<20100101, datecol>=20100101 and datecol<20110101, datecol>=20110101 and datecol<20120101, datecol>=20120101

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

Question: 76

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

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

You plan to insert data from the files into Table1 and azure Data Lake Storage Gen2 container named container1.

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

row in the serving layer of Table1.

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

Solution: In an Azure Synapse Analytics pipeline, you use a data flow that contains a Derived Column transformation.
A. Yes
B. No

Answer: A

Explanation:

Use the derived column transformation to generate new columns in your data flow or to modify existing fields.

Reference: https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column

Question: 77

You have an Azure Data Lake Storage Gen2 account named adls2 that is protected by a virtual network.

You are designing a SQL pool in Azure Synapse that will use adls2 as a source.

What should you use to authenticate to adls2?
A. a shared access signature (SAS)
B. a managed identity
C. a shared key
D. an Azure Active Directory (Azure AD) user

Answer: B

Explanation:

Managed identity for Azure resources is a feature of Azure Active Directory. The feature provides Azure services with an automatically managed identity in Azure AD. You can use the Managed Identity capability to authenticate to any service that support Azure AD authentication.

Managed Identity authentication is required when your storage account is attached to a VNet.

Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-examples

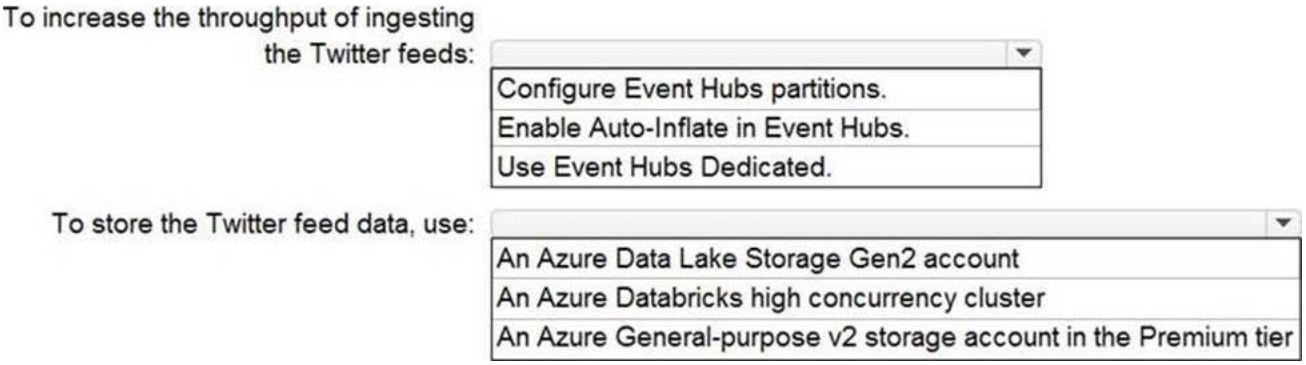
Question: 78

HOTSPOT

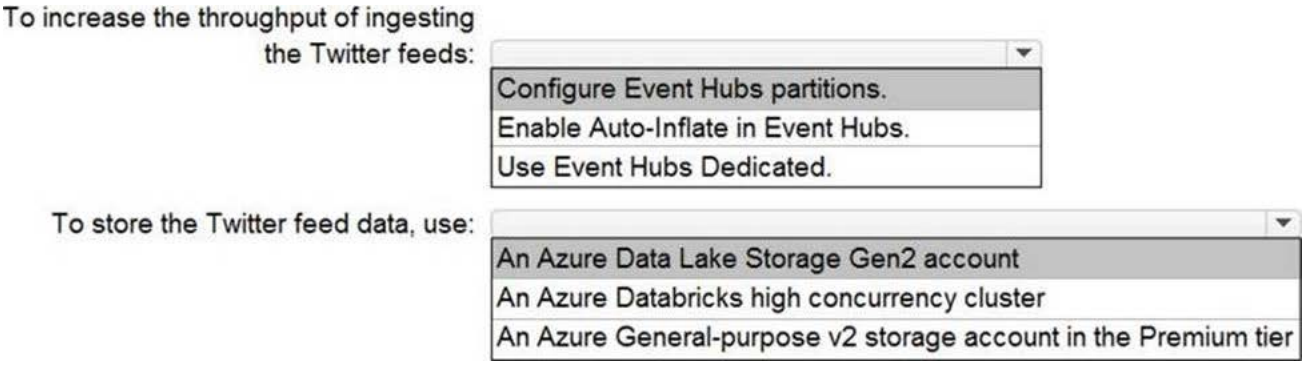
You need to design a data ingestion and storage solution for the Twitter feeds. The solution must meet the customer sentiment analytics requirements.

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

correct selection b worth one point.



Answer:



Explanation:

Graphical user interface, text

Description automatically generated

Box 1: Configure Evegent Hubs partitions

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

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

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

Box 2: An Azure Data Lake Storage Gen2 account

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

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

Question: 79

You plan to ingest streaming social media data by using Azure Stream Analytics. The data will be stored in files in Azure Data Lake Storage, and then consumed by using Azure Datiabricks and PolyBase in Azure Synapse Analytics.

You need to recommend a Stream Analytics data output format to ensure that the queries from Databricks and PolyBase against the files encounter the fewest possible errors. The solution must ensure that the tiles can be queried quickly and that the data type information is retained.

What should you recommend?
A. Parquet
B. Avro
C. CSV
D. JSON

Answer: A

Explanation:

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

Question: 80

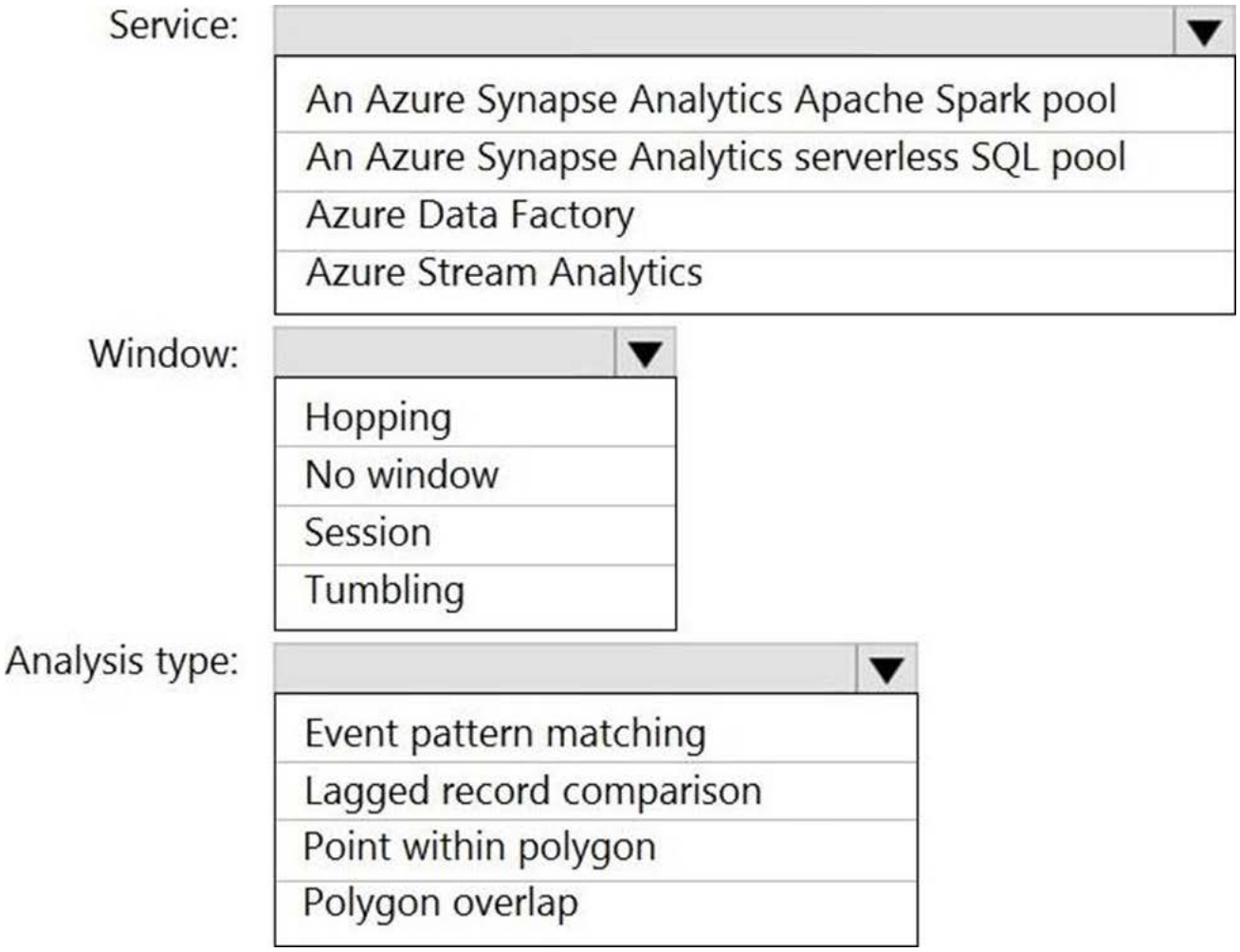
HOTSPOT

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.

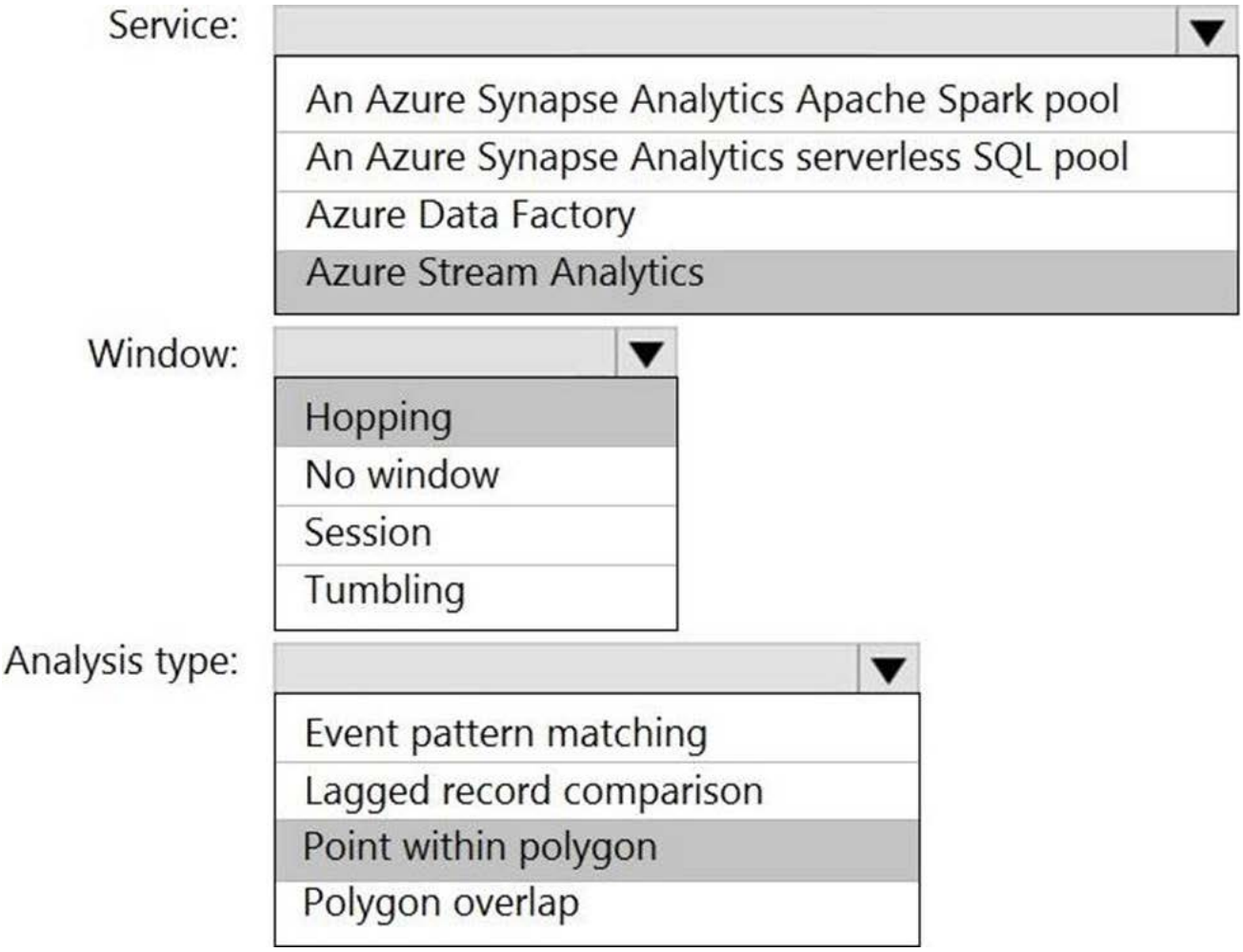
You have a CSV file in an Azure Data Lake Storage Gen2 container. The file maintains the expected geographical area in which each vehicle should be.

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.


Service:
An Azure Synapse Analytics Apache Spark pool
An Azure Synapse Analytics serverless SQL pool
Azure Data Factory
Azure Stream Analytics
Window:
Hopping
No window
Session
Tumbling
Analysis type:
Event pattern matching
Lagged record comparison
Point within polygon
Polygon overlap

Answer:



Explanation:

Box 1: Azure Stream Analytics

Box 2: Hopping

Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.

Box 3: Point within polygon

Question: 81

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

Which resource provider should you enable?
A. Microsoft.Sql
B. Microsoft-Automation
C. Microsoft.EventGrid
D. Microsoft.EventHub

Answer: C

Explanation:

Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure Blob Storage account. Data Factory natively integrates with Azure Event Grid, which lets you trigger pipelines on such events.

Reference:

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

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

Question: 82

You are designing an anomaly detection solution for streaming data from an Azure IoT hub.

The solution must meet the following requirements:

Send the output to Azure Synapse.

Identify spikes and dips in time series data.

Minimize development and configuration effort.

Which should you include in the solution?
A. Azure Databricks
B. Azure Stream Analytics
C. Azure SQL Database

Answer: B

Explanation:

You can identify anomalies by routing data via IoT Hub to a built-in ML model in Azure Stream Analytics.

Reference: https://docs.microsoft.com/en-us/learn/modules/data-anomaly-detection-using-azure-iot-hub/

Question: 83

HOTSPOT

You have the following Azure Stream Analytics query.

```
WITH

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

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

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

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

Answer:

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

Explanation:

Box 1: No

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

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

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

SELECT * INTO [output] FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceID

Note: The new extension of Azure Stream Analytics SQL includes a keyword INTO that allows you to specify the number of partitions for a stream when performing reshuffling using a PARTITION BY statement.

Box 2: Yes

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

Box 3: Yes

Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job.

In general, the best practice is to start with 6 SUs for queries that don't use PARTITION BY.

Here there are 10 partitions, so 6×10 = 60 SUs is good.

Note: Remember, Streaming Unit (SU) count, which is the unit of scale for Azure Stream

Analytics, must be adjusted so the number of physical resources available to the job can fit the partitioned flow. In general, six SUs is a good number to assign to each partition. In case there are insufficient resources assigned to the job, the system will only apply the repartition if it benefits the job.

Question: 84

DRAG DROP

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

You need to grant Group1 read-only permissions to all the tables and views in schema1.

The solution must use the principle of least privilege.

Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order. NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

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

Answer:

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

Explanation:

Step 1: Create a database role named Role1 and grant Role1 SELECT permissions to schema

You need to grant Group1 read-only permissions to all the tables and views in schema1. Place one or more database users into a database role and then assign permissions to the database role.

Step 2: Assign Rol1 to the Group database user

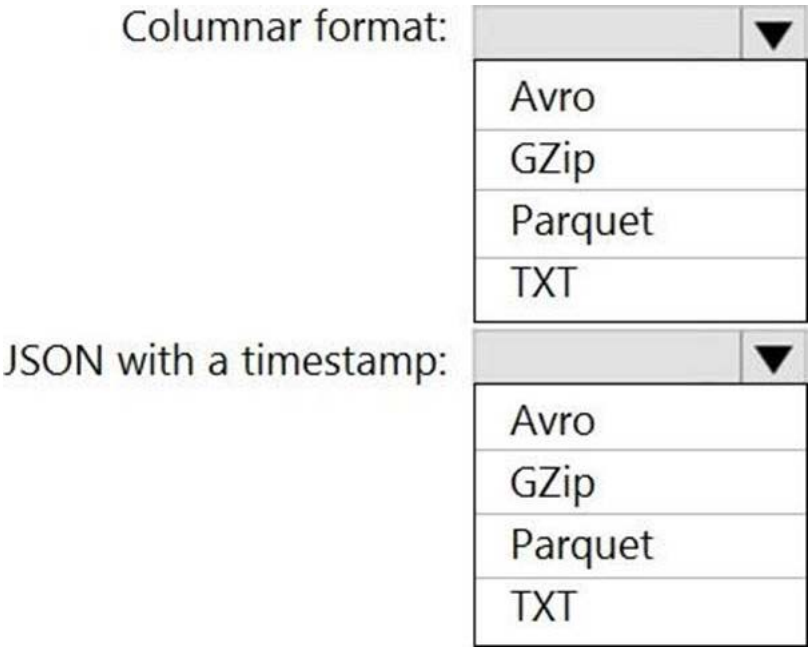
Step 3: Assign the Azure role-based access control (Azure RBAC) Reader role for dw1 to Group1

Question: 85

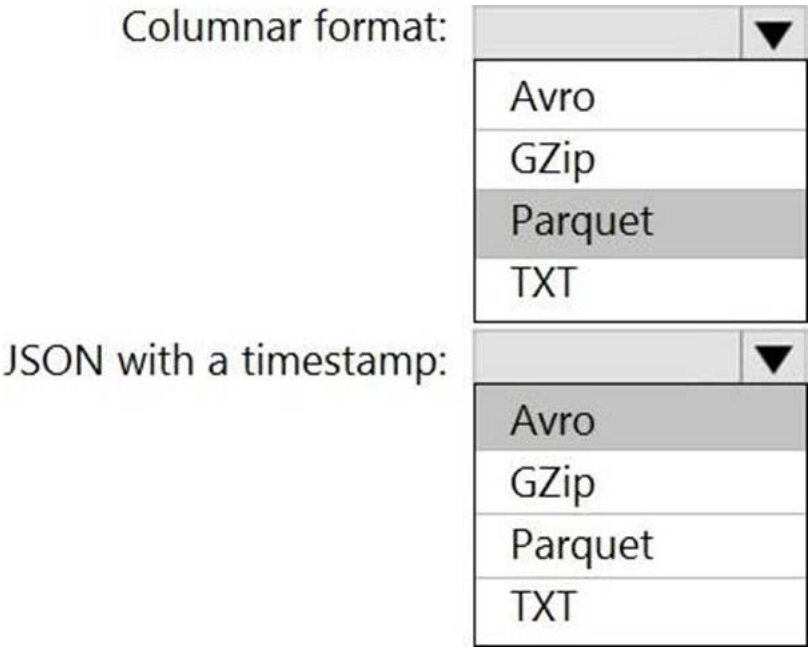
HOTSPOT

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.



Answer:



Explanation:

Box 1: Parquet

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

Box 2: Avro

An Avro schema is created using JSON format.

AVRO supports timestamps.

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

- Avro format
- Binary format
- Delimited text format
- Excel format
- JSON format
- ORC format
- Parquet format
- XML format