

# Microsoft

## Exam Questions DP-203

Data Engineering on Microsoft Azure



## NEW QUESTION 1

- (Exam Topic 1)

You need to design the partitions for the product sales transactions. The solution must mee the sales transaction dataset requirements.

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

### Answer Area

Partition product sales transactions data by:	<input type="checkbox"/> Sales date <input checked="" type="checkbox"/> Product ID <input type="checkbox"/> Promotion ID
Store product sales transactions data in:	<input checked="" type="checkbox"/> An Azure Synapse Analytics dedicated SQL pool <input type="checkbox"/> An Azure Synapse Analytics serverless SQL pool <input type="checkbox"/> An Azure Data Lake Storage Gen2 account linked to an Azure Synapse Analytics workspace

- A. Mastered
- B. Not Mastered

**Answer: A**

### Explanation:

Box 1: Sales date

Scenario: Contoso requirements for data integration include:

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

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

➤ Ensure that data storage costs and performance are predictable.

The size of a dedicated SQL pool (formerly SQL DW) is determined by Data Warehousing Units (DWU). Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage. This format

significantly reduces the data storage costs, and improves query performance.

Synapse analytics dedicated sql pool Reference:

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

## NEW QUESTION 2

- (Exam Topic 3)

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.

Copy behavior:	<input type="text"/> <input checked="" type="checkbox"/> Flatten hierarchy <input type="checkbox"/> Merge files <input type="checkbox"/> Preserve hierarchy
Sink file type:	<input type="text"/> <input type="checkbox"/> CSV <input checked="" type="checkbox"/> JSON <input type="checkbox"/> Parquet <input type="checkbox"/> TXT

- A. Mastered
- B. Not Mastered

**Answer: A**

### Explanation:

Box 1: Preserver herarchy

Compared to the flat namespace on Blob storage, the hierarchical namespace greatly improves the performance of directory management operations, which improves overall job performance.

Box 2: Parquet

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

Reference:

<https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-introduction> <https://docs.microsoft.com/en-us/azure/data-factory/format-parquet>

### NEW QUESTION 3

- (Exam Topic 3)

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

**Answer: D**

#### Explanation:

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

Reference:

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

### NEW QUESTION 4

- (Exam Topic 3)

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.

SELECT

[user],

feature,

second,

Time) as duration

FROM input TIMESTAMP BY Time

WHERE

Event = 'end'

(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour, 1) WHEN Event = 'start'),

- A. Mastered
- B. Not Mastered

**Answer: A**

#### Explanation:

Box 1: DATEDIFF

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

Syntax: DATEDIFF ( datepart , startdate, enddate ) Box 2: LAST

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

Example: SELECT

[user], feature, DATEDIFF(

second,

LAST(Time) OVER (PARTITION BY [user], feature LIMIT DURATION(hour,

1) WHEN Event = 'start'), Time) as duration

FROM input TIMESTAMP BY Time

WHERE

Event = 'end' Reference:

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

### NEW QUESTION 5

- (Exam Topic 3)

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

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

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

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

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

Which three actions should you perform? Each correct answer presents part of the solution. NOTE: Each area selection is worth one point.

- A. Add the managed identity to the Sales group.
- B. Use the managed identity as the credentials for the data load process.
- C. Create a shared access signature (SAS).
- D. Add your Azure Active Directory (Azure AD) account to the Sales group.
- E. Use the snared access signature (SAS) as the credentials for the data load process.
- F. Create a managed identity.

**Answer:** ADF

**Explanation:**

The managed identity grants permissions to the dedicated SQL pools in the workspace.

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

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

**NEW QUESTION 6**

- (Exam Topic 3)

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: You use a dedicated SQL pool to create an external table that has a additional DateTime column. Does this meet the goal?

- A. Yes
- B. No

**Answer:** A

**NEW QUESTION 7**

- (Exam Topic 3)

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

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Box 1: Azure Stream Analytics Box 2: Hopping

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

Box 3: Point within polygon Reference:  
<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

**NEW QUESTION 8**

- (Exam Topic 3)

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

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

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

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

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

**Values**

CAST

COLLATE

CONVERT

FLATTEN

PIVOT

UNPIVOT

**Answer Area**

```

SELECT * FROM (
  SELECT YEAR(Date) Year, MONTH(Date)
    FROM temperatures
  WHERE date BETWEEN DATE '2019-01-01' AND DATE
    '2021-08-31'
    Value (
      Value (Temp AS DECIMAL(4, 1)))
  AVG (
    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

```

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:



Values

CAST

COLLATE

CONVERT

FLATTEN

PIVOT

UNPIVOT

Answer Area

```
SELECT * FROM (
    SELECT YEAR(Date) Year, MONTH(Date)
    FROM Temperatures
    WHERE date BETWEEN DATE '2019-01-01' AND DATE
    '2021-08-31'
    CONVERT (
        COLLATE (Temp AS DECIMAL(4, 1)))
    AVG (
        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
```

**NEW QUESTION 9**

- (Exam Topic 3)

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

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

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

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

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

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

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

**NEW QUESTION 10**

- (Exam Topic 3)

You have a C# application that process data from an Azure IoT hub and performs complex transformations. You need to replace the application with a real-time

solution. The solution must reuse as much code as possible from the existing application.

- A. Azure Databricks
- B. Azure Event Grid
- C. Azure Stream Analytics
- D. Azure Data Factory

**Answer: C**

**Explanation:**

Azure Stream Analytics on IoT Edge empowers developers to deploy near-real-time analytical intelligence closer to IoT devices so that they can unlock the full value of device-generated data. UDF are available in C# for IoT Edge jobs  
 Azure Stream Analytics on IoT Edge runs within the Azure IoT Edge framework. Once the job is created in Stream Analytics, you can deploy and manage it using IoT Hub.

References:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-edge>

**NEW QUESTION 10**

- (Exam Topic 3)

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.

Does this meet the goal?

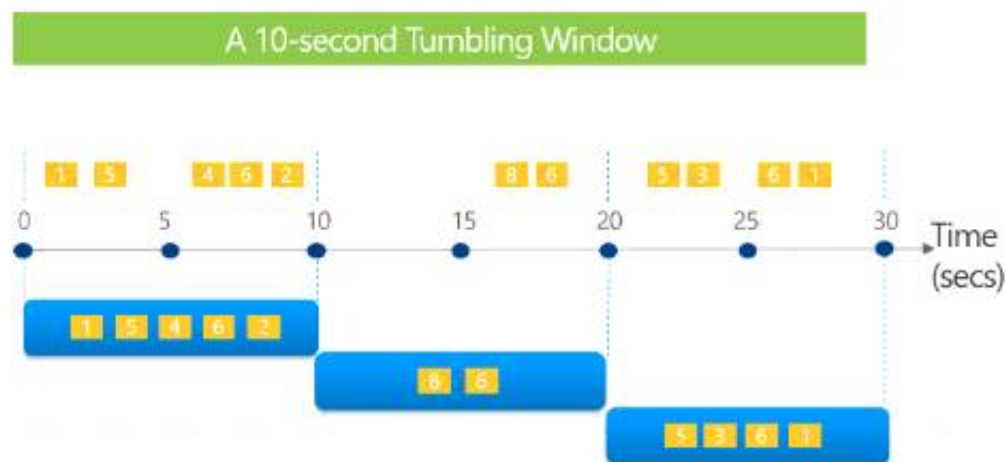
- A. Yes
- B. No

**Answer: A**

**Explanation:**

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

Tell me the count of tweets per time zone every 10 seconds



```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

Reference:

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

**NEW QUESTION 13**

- (Exam Topic 3)

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.

**Answer: B**

**Explanation:**

For clusters running Databricks Runtime 6.4 and above, optimized autoscaling is used by all-purpose clusters in the Premium plan

Optimized autoscaling:

Scales up from min to max in 2 steps.

Can scale down even if the cluster is not idle by looking at shuffle file state. Scales down based on a percentage of current nodes.

On job clusters, scales down if the cluster is underutilized over the last 40 seconds.

On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds.

The spark.databricks.aggressiveWindowDownS Spark configuration property specifies in seconds how often a cluster makes down-scaling decisions. Increasing the value causes a cluster to scale down more slowly. The maximum value is 600.

Note: Standard autoscaling

Starts with adding 8 nodes. Thereafter, scales up exponentially, but can take many steps to reach the max. You can customize the first step by setting the spark.databricks.autoscaling.standardFirstStepUp Spark configuration property.

Scales down only when the cluster is completely idle and it has been underutilized for the last 10 minutes. Scales down exponentially, starting with 1 node.

Reference:

#### NEW QUESTION 15

- (Exam Topic 3)

You have an Azure subscription that contains the following resources:

\* An Azure Active Directory (Azure AD) tenant that contains a security group named Group1.

\* An Azure Synapse Analytics SQL pool named Pool1.

You need to control the access of Group1 to specific columns and rows in a table in Pool1

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

#### Answer Area

To control access to the columns:

CREATE CRYPTOGRAPHIC PROVIDER  
 CREATE PARTITION FUNCTION  
 CREATE SECURITY POLICY  
 GRANT

To control access to the rows:

CREATE CRYPTOGRAPHIC PROVIDER  
 CREATE PARTITION FUNCTION  
 CREATE SECURITY POLICY  
 GRANT

- A. Mastered
- B. Not Mastered

**Answer: A**

**Explanation:**

#### Answer Area

To control access to the columns:

CREATE CRYPTOGRAPHIC PROVIDER  
 CREATE PARTITION FUNCTION  
 CREATE SECURITY POLICY  
 GRANT

To control access to the rows:

CREATE CRYPTOGRAPHIC PROVIDER  
 CREATE PARTITION FUNCTION  
 CREATE SECURITY POLICY  
 GRANT

#### NEW QUESTION 18

- (Exam Topic 3)

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

**Answer: B**

**Explanation:**



Stream Analytics supports Azure Blob storage and Azure SQL Database as the storage layer for Reference Data.  
Reference:  
<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data>

**NEW QUESTION 20**

- (Exam Topic 3)

You plan to monitor an Azure data factory by using the Monitor & Manage app.

You need to identify the status and duration of activities that reference a table in a source database.

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

Actions

From the Data Factory monitoring app, add the Source user property to the Activity Runs table.

From the Data Factory monitoring app, add the Source user property to the Pipeline Runs table.

From the Data Factory authoring UI, publish the pipelines.

From the Data Factory monitoring app, add a linked service to the Pipeline Runs table.

From the Data Factory authoring UI, generate a user property for Source on all activities.

From the Data Factory authoring UI, generate a user property for Source on all datasets.

Answer Area

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Step 1: From the Data Factory authoring UI, generate a user property for Source on all activities. Step 2: From the Data Factory monitoring app, add the Source user property to Activity Runs table.

You can promote any pipeline activity property as a user property so that it becomes an entity that you can monitor. For example, you can promote the Source and Destination properties of the copy activity in your pipeline as user properties. You can also select Auto Generate to generate the Source and Destination user properties for a copy activity.

Step 3: From the Data Factory authoring UI, publish the pipelines

Publish output data to data stores such as Azure SQL Data Warehouse for business intelligence (BI) applications to consume.

References:

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

**NEW QUESTION 22**

- (Exam Topic 3)

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

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

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

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

NOTE: Each correct selection is worth one point.

To track encryption key usage:

▼

Always Encrypted

TDE with customer-managed keys

TDE with platform-managed keys

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

▼

Create and configure Azure key vaults in two Azure regions.

Enable Advanced Data Security on Server1.

Implement the client apps by using a Microsoft .NET Framework data provider.

- A. Mastered  
 B. Not Mastered

**Answer:** A

**Explanation:**

Box 1: TDE with customer-managed keys

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

Box 2: Create and configure Azure key vaults in two Azure regions

The contents of your key vault are replicated within the region and to a secondary region at least 150 miles away, but within the same geography to maintain high durability of your keys and secrets.

Reference:

<https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption> <https://docs.microsoft.com/en-us/azure/key-vault/general/logging>

**NEW QUESTION 25**

- (Exam Topic 3)

You have an Azure data factory.

You need to examine the pipeline failures from the last 60 days. What should you use?

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

**Answer:** D

**Explanation:**

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

Reference:

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

**NEW QUESTION 26**

- (Exam Topic 3)

You implement an enterprise data warehouse in Azure Synapse Analytics. You have a large fact table that is 10 terabytes (TB) in size.

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

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

You need to distribute the large fact table across multiple nodes to optimize performance of the table. Which technology should you use?

- A. hash distributed table with clustered index

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

**Answer:** B

**Explanation:**

Hash-distributed tables improve query performance on large fact tables. Columnstore indexes can achieve up to 100x better performance on analytics and data warehousing workloads and up to 10x better data compression than traditional rowstore indexes.  
Reference:  
<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute> <https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-query-performance>

**NEW QUESTION 28**

- (Exam Topic 3)  
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.

**Actions**

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

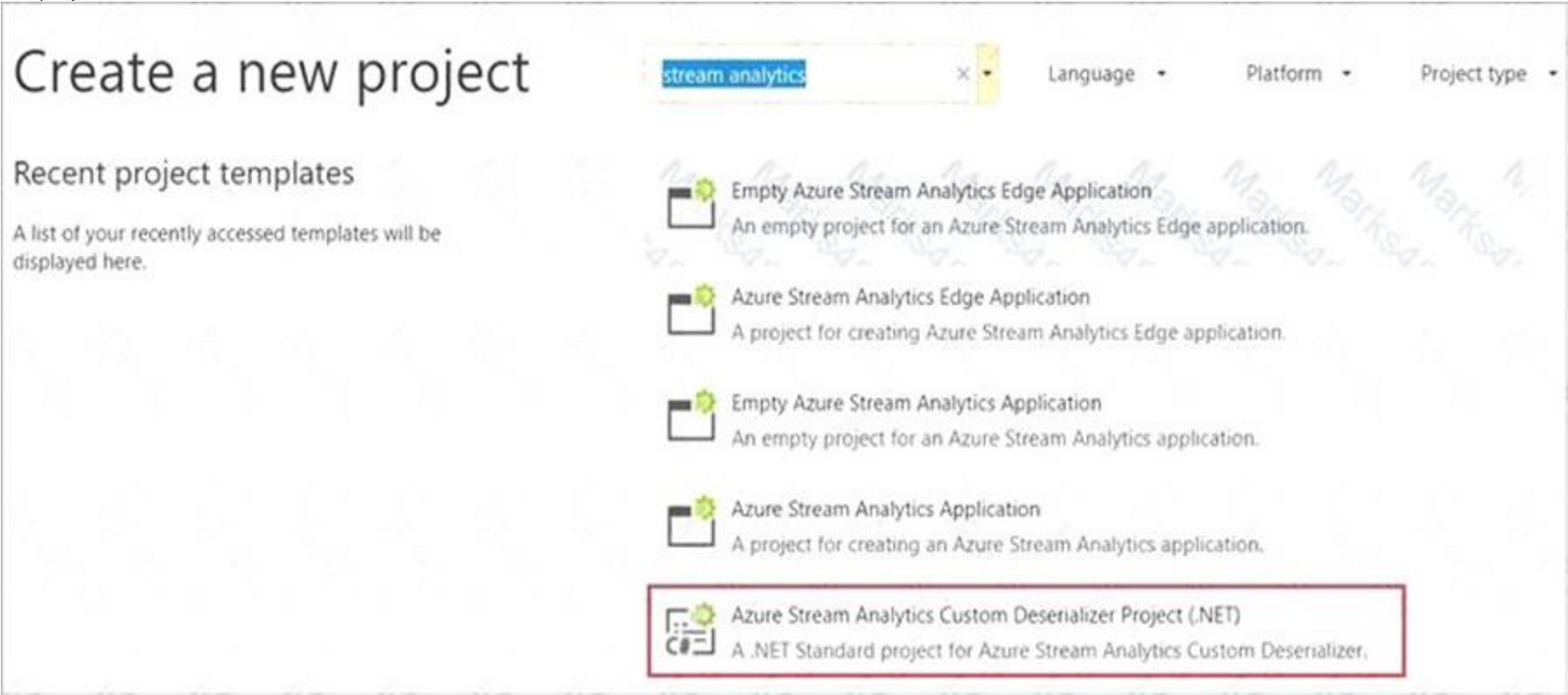
**Answer Area**

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Step 1: Add an Azure Stream Analytics Custom Deserializer Project (.NET) project to the solution. Create a custom deserializer  
\* 1. Open Visual Studio and select File > New > Project. Search for Stream Analytics and select Azure Stream Analytics Custom Deserializer Project (.NET). Give the project a name, like Protobuf Deserializer.



- \* 2. In Solution Explorer, right-click your Protobuf Deserializer project and select Manage NuGet Packages from the menu. Then install the Microsoft.Azure.StreamAnalytics and Google.Protobuf NuGet packages.
- \* 3. Add the MessageBodyProto class and the MessageBodyDeserializer class to your project.
- \* 4. Build the Protobuf Deserializer project.



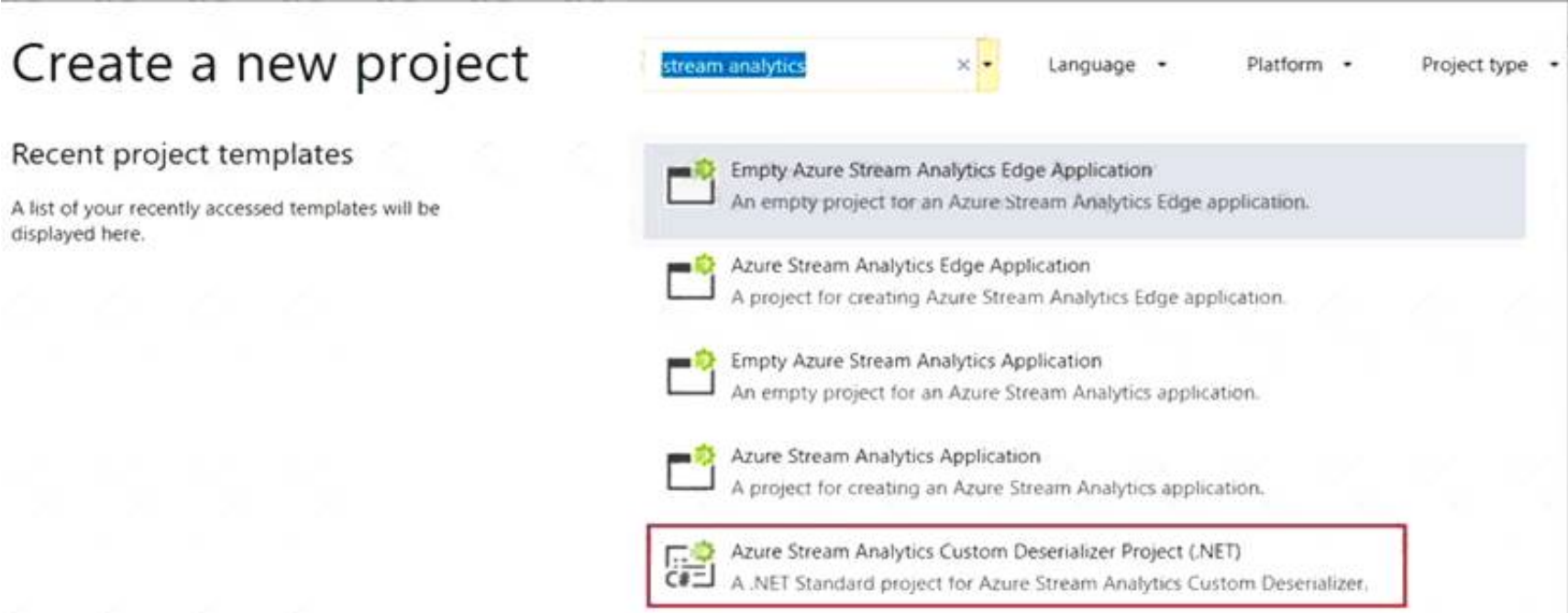
Step 2: Add .NET deserializer code for Protobuf to the custom deserializer project  
Azure Stream Analytics has built-in support for three data formats: JSON, CSV, and Avro. With custom .NET deserializers, you can read data from other formats such as Protocol Buffer, Bond and other user defined formats for both cloud and edge jobs.

Step 3: Add an Azure Stream Analytics Application project to the solution Add an Azure Stream Analytics project

> In Solution Explorer, right-click the Protobuf Deserializer solution and select Add > New Project. Under Azure Stream Analytics > Stream Analytics, choose Azure Stream Analytics Application. Name it ProtobufCloudDeserializer and select OK.

> Right-click References under the ProtobufCloudDeserializer Azure Stream Analytics project. Under Projects, add Protobuf Deserializer. It should be automatically populated for you.

Reference:  
<https://docs.microsoft.com/en-us/azure/stream-analytics/custom-deserializer>



**NEW QUESTION 31**  
- (Exam Topic 3)  
Which Azure Data Factory components should you recommend using together to import the daily inventory data from the SQL server to Azure Data Lake Storage?  
To answer, select the appropriate options in the answer area.  
NOTE: Each correct selection is worth one point.

**Answer Area:**

Integration runtime type:	<div><div>Azure integration runtime</div><div>Azure-SSIS integration runtime</div><div>Self-hosted integration runtime</div></div>
Trigger type:	<div><div>Event-based trigger</div><div>Schedule trigger</div><div>Tumbling window trigger</div></div>
Activity type:	<div><div>Copy activity</div><div>Lookup activity</div><div>Stored procedure activity</div></div>

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

**Answer Area:**

Integration runtime type:	<div><div>Azure integration runtime</div><div>Azure-SSIS integration runtime</div><div>Self-hosted integration runtime</div></div>
Trigger type:	<div><div>Event-based trigger</div><div>Schedule trigger</div><div>Tumbling window trigger</div></div>
Activity type:	<div><div>Copy activity</div><div>Lookup activity</div><div>Stored procedure activity</div></div>

**NEW QUESTION 33**



- (Exam Topic 3)

You use Azure Data Lake Storage Gen2.

You need to ensure that workloads can use filter predicates and column projections to filter data at the time the data is read from disk.

Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

- A. Reregister the Microsoft Data Lake Store resource provider.
- B. Reregister the Azure Storage resource provider.
- C. Create a storage policy that is scoped to a container.
- D. Register the query acceleration feature.
- E. Create a storage policy that is scoped to a container prefix filter.

**Answer:** BD

#### NEW QUESTION 36

- (Exam Topic 3)

You have an Azure Stream Analytics query. The query returns a result set that contains 10,000 distinct values for a column named clusterID.

You monitor the Stream Analytics job and discover high latency. You need to reduce the latency.

Which two actions should you perform? Each correct answer presents a complete solution. NOTE: Each correct selection is worth one point.

- A. Add a pass-through query.
- B. Add a temporal analytic function.
- C. Scale out the query by using PARTITION BY.
- D. Convert the query to a reference query.
- E. Increase the number of streaming units.

**Answer:** CE

#### Explanation:

C: Scaling a Stream Analytics job takes advantage of partitions in the input or output. Partitioning lets you divide data into subsets based on a partition key. A process that consumes the data (such as a Streaming Analytics job) can consume and write different partitions in parallel, which increases throughput.

E: Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job. This capacity lets you focus on the query logic and abstracts the need to manage the hardware to run your Stream Analytics job in a timely manner.

References:

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization> <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption>

#### NEW QUESTION 41

- (Exam Topic 3)

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.

- A. Mastered
- B. Not Mastered

**Answer:** A

#### 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 Role1 to the Group database user

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

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

#### NEW QUESTION 44

- (Exam Topic 3)

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.

Values	Answer Area
CLUSTERED INDEX	CREATE TABLE table1
COLLATE	(
DISTRIBUTION	ID INTEGER,
PARTITION	col1 VARCHAR(10),
PARTITION FUNCTION	col2 VARCHAR(10)
PARTITION SCHEME	) WITH
	(
	= HASH(ID),
	(ID RANGE LEFT FOR VALUES (1, 1000000, 2000000))
	);

- A. Mastered
- B. Not Mastered

**Answer:** A

#### Explanation:

Box 1: DISTRIBUTION

Table distribution options include DISTRIBUTION = HASH ( distribution\_column\_name ), assigns each row to one distribution by hashing the value stored in distribution\_column\_name. Box 2: PARTITION

Table partition options. Syntax:

PARTITION ( partition\_column\_name RANGE [ LEFT | RIGHT ] FOR VALUES ( [ boundary\_value [...n] ] ) )

Reference:

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

#### NEW QUESTION 45

- (Exam Topic 3)

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.

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

- A. Mastered
- B. Not Mastered

**Answer:** A

#### Explanation:

Box 1: Tumbling window

To be able to use the Delay parameter we select Tumbling window. Box 2:

Recurrence: 30 minutes, not 32 minutes

Delay: 2 minutes.

The amount of time to delay the start of data processing for the window. The pipeline run is started after the expected execution time plus the amount of delay. The delay defines how long the trigger waits past the due time before triggering a new run. The delay doesn't alter the window startTime.

Reference:

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

**NEW QUESTION 46**

- (Exam Topic 3)

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

**Answer:** A

**Explanation:**

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

Reference:

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime>

**NEW QUESTION 47**

- (Exam Topic 3)

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 AverageFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SlidingWindow(second, 10)
- B. SELECT Country, Count(\*) AS CountFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, TumblingWindow(second, 10)
- C. SELECT Country, Avg(\*) AS AverageFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, HoppingWindow(second, 10, 2)
- D. SELECT Country, Count(\*) AS CountFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SessionWindow(second, 5, 10)

**Answer:** B

**Explanation:**

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

Example: Reference:

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

**NEW QUESTION 48**

- (Exam Topic 3)

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. zone-redundant storage (ZRS)
- B. read-access geo-redundant storage (RA-GRS)
- C. locally-redundant storage (LRS)
- D. geo-redundant storage (GRS)

**Answer:** C

**NEW QUESTION 50**

- (Exam Topic 3)

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

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

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

NOTE: Each correct selection is worth one point.

```
SELECT sensorId,  
       growth = reading -  
           (reading) OVER (PARTITION BY sensorId  
                           LAG  
                           LAST  
                           LEAD  
                           (hour, 1))  
FROM input
```

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Box 1: LAG

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

Box 2: LIMIT DURATION

Example: Compute the rate of growth, per sensor: `SELECT sensorId,  
growth = reading  
LAG(reading) OVER (PARTITION BY sensorId LIMIT DURATION(hour, 1)) FROM input`

Reference:

<https://docs.microsoft.com/en-us/stream-analytics-query/lag-azure-stream-analytics>

**NEW QUESTION 51**

- (Exam Topic 3)

You configure monitoring for a Microsoft Azure SQL Data Warehouse implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Gen 2 using an external table.

Files with an invalid schema cause errors to occur. You need to monitor for an invalid schema error. For which error should you monitor?

- A. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge\_Connect: Error[com.microsoft.polybase.client.KerberosSecureLogin] occurred while accessing external files.'
- B. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge\_Connect: Error [No FileSystem for scheme: wasbs] occurred while accessing external file.'
- C. Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11": for linked server "(null)", Query aborted- the maximum reject threshold (orows) was reached while regarding from an external source: 1 rows rejected out of total 1 rows processed.
- D. EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge\_Connect: Error [Unable to instantiate LoginClass] occurredwhile accessing external files.'

**Answer:** C

**Explanation:**

Customer Scenario:

SQL Server 2016 or SQL DW connected to Azure blob storage. The CREATE EXTERNAL TABLE DDL points to a directory (and not a specific file) and the directory contains files with different schemas.

SSMS Error:

Select query on the external table gives the following error: Msg 7320, Level 16, State 110, Line 14

Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11" for linked server "(null)". Query aborted-- the maximum reject threshold (0 rows) was reached while reading from an external source: 1 rows rejected out of total 1 rows processed.

Possible Reason:

The reason this error happens is because each file has different schema. The PolyBase external table DDL when pointed to a directory recursively reads all the files in that directory. When a column or data type mismatch happens, this error could be seen in SSMS.

Possible Solution:

If the data for each table consists of one file, then use the filename in the LOCATION section prepended by the directory of the external files. If there are multiple files per table, put each set of files into different directories in Azure Blob Storage and then you can point LOCATION to the directory instead of a particular file. The latter suggestion is the best practices recommended by SQLCAT even if you have one file per table.

**NEW QUESTION 55**

- (Exam Topic 3)

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 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.	<input type="radio"/>	<input type="radio"/>
The stream scheme key and count must match the output scheme.	<input type="radio"/>	<input type="radio"/>
Providing 60 streaming units will optimize the performance of the query.	<input type="radio"/>	<input type="radio"/>

- A. Mastered
- B. Not Mastered

Answer: A

Explanation:

Box 1: Yes  
You can now use a new extension of Azure Stream Analytics SQL to specify the number of partitions of a stream when reshuffling the data. The outcome is a stream that has the same partition scheme. Please see below for an example: WITH step1 AS (SELECT \* FROM [input1] PARTITION BY DeviceID INTO 10), step2 AS (SELECT \* FROM [input2] PARTITION BY DeviceID INTO 10) SELECT \* INTO [output] FROM step1 PARTITION BY DeviceID UNION step2 PARTITION BY DeviceID Note: The new extension of Azure Stream Analytics SQL includes a keyword INTO that allows you to specify the number of partitions for a stream when performing reshuffling using a PARTITION BY statement.

Box 2: Yes  
When joining two streams of data explicitly repartitioned, these streams must have the same partition key and partition count.

Box 3: Yes  
10 partitions x six SUs = 60 SUs is fine.  
Note: Remember, Streaming Unit (SU) count, which is the unit of scale for Azure Stream Analytics, must be adjusted so the number of physical resources available to the job can fit the partitioned flow. In general, six SUs is a good number to assign to each partition. In case there are insufficient resources assigned to the job, the system will only apply the repartition if it benefits the job.

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

NEW QUESTION 58

- (Exam Topic 3)  
You are designing an Azure Stream Analytics solution that receives instant messaging data from an Azure event hub. You need to ensure that the output from the Stream Analytics job counts the number of messages per time zone every 15 seconds.  
How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.  
NOTE: Each correct selection is worth one point.

## Answer Area

Select TimeZone, count(\*) AS MessageCount  
 FROM  
 MessageStream  
 HOPPINGWINDOW  
 SESSIONWINDOW  
 SLIDINGWINDOW  
 TUMBLINGWINDOW  
 CreatedAt  
 (second,15)

- A. Mastered  
 B. Not Mastered

**Answer:** A

**Explanation:**

## Answer Area

Select TimeZone, count(\*) AS MessageCount  
 FROM  
 MessageStream  
 HOPPINGWINDOW  
 SESSIONWINDOW  
 SLIDINGWINDOW  
 TUMBLINGWINDOW  
 CreatedAt  
 (second,15)

### NEW QUESTION 63

- (Exam Topic 3)

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. Parquet  
 B. Avro  
 C. CSV  
 D. JSON

**Answer:** B

**Explanation:**

The Avro format is great for data and message preservation. Avro schema with its support for evolution is essential for making the data robust for streaming architectures like Kafka, and with the metadata that schema provides, you can reason on the data. Having a schema provides robustness in providing meta-data about the data stored in Avro records which are self- documenting the data. References: <http://cloudurable.com/blog/avro/index.html>

### NEW QUESTION 66

- (Exam Topic 3)

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

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

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

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

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

NOTE: Each correct selection is worth one point.

Number of partitions:

	▼
1	
8	
16	
32	

Partition key:

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

- A. Mastered
- B. Not Mastered

**Answer:** A

**Explanation:**

Box 1: 16

For Event Hubs you need to set the partition key explicitly.

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

Box 2: Transaction ID Reference:

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

**NEW QUESTION 69**

- (Exam Topic 3)

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

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

What should you do?

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

**Answer:** D

**Explanation:**

To keep an interactive cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.

References:

<https://docs.azuredatabricks.net/clusters/clusters-manage.html#automatic-termination>

**NEW QUESTION 70**

- (Exam Topic 3)

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

You need to design queries to maximize the benefits of partition elimination. What should you include in the Transact-SQL queries?

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

**Answer:** B

**NEW QUESTION 71**

- (Exam Topic 3)

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 Get Metadata activity that retrieves the DateTime of the files.

Does this meet the goal?

- A. Yes
- B. No

**Answer:** B

#### NEW QUESTION 72

- (Exam Topic 3)

You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Contacts. Contacts contains a column named Phone. You need to ensure that users in a specific role only see the last four digits of a phone number when querying the Phone column. What should you include in the solution?

- A. a default value
- B. dynamic data masking
- C. row-level security (RLS)
- D. column encryption
- E. table partitions

**Answer:** C

#### NEW QUESTION 73

- (Exam Topic 3)

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

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

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

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

Which column should you add to the table?

- A. [ManagerEmployeeID] [int] NULL
- B. [ManagerEmployeeID] [smallint] NULL
- C. [ManagerEmployeeKey] [int] NULL
- D. [ManagerName] [varchar](200) NULL

**Answer:** A

#### Explanation:

Use the same definition as the EmployeeID column. Reference:

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

#### NEW QUESTION 78

.....



## Thank You for Trying Our Product

### We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questions and Answers in PDF Format

### DP-203 Practice Exam Features:

- \* DP-203 Questions and Answers Updated Frequently
- \* DP-203 Practice Questions Verified by Expert Senior Certified Staff
- \* DP-203 Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- \* DP-203 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

**100% Actual & Verified — Instant Download, Please Click**  
**[Order The DP-203 Practice Test Here](#)**