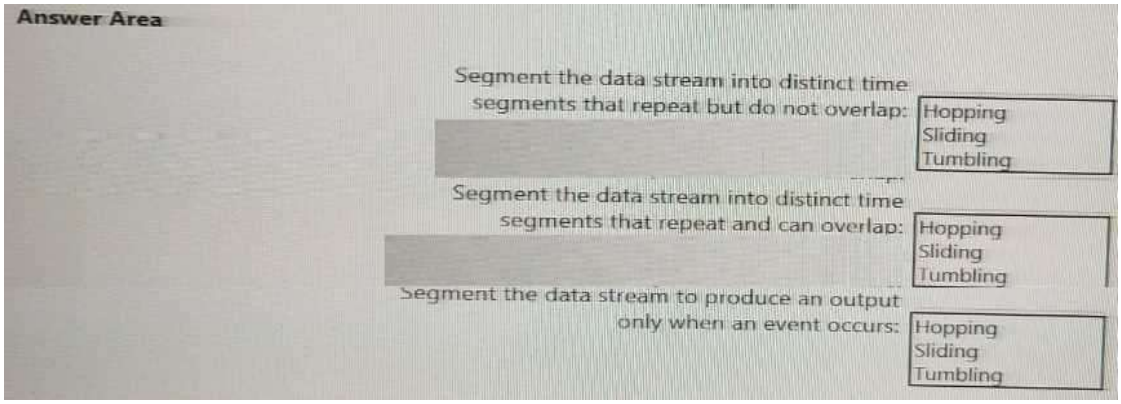
- **Vendor: Microsoft**
- **Exam Code: DP-203**
- **Exam Name: Data Engineering on Microsoft Azure**
- **New Updated Questions from Braindump2go**
- **(Updated in March/2021)**

Question: 34

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



Answer:

1. Tumbling
2. Hopping
3. Hopping

Question: 35

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

You need to output the count of tweets during the last five minutes every five minutes. Each tweet must only
be counted once.
Which windowing function should you use?

A. a five-minute Session window
B. a five-minute Sliding window
C. a five-minute Tumbling window
D. a five-minute Hopping window that has one-minute hop

Answer: C

Explanation:
Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.
References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

Question: 36

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 refere ce query.
E. Increase the number of streaming units.

Answer: C, E

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

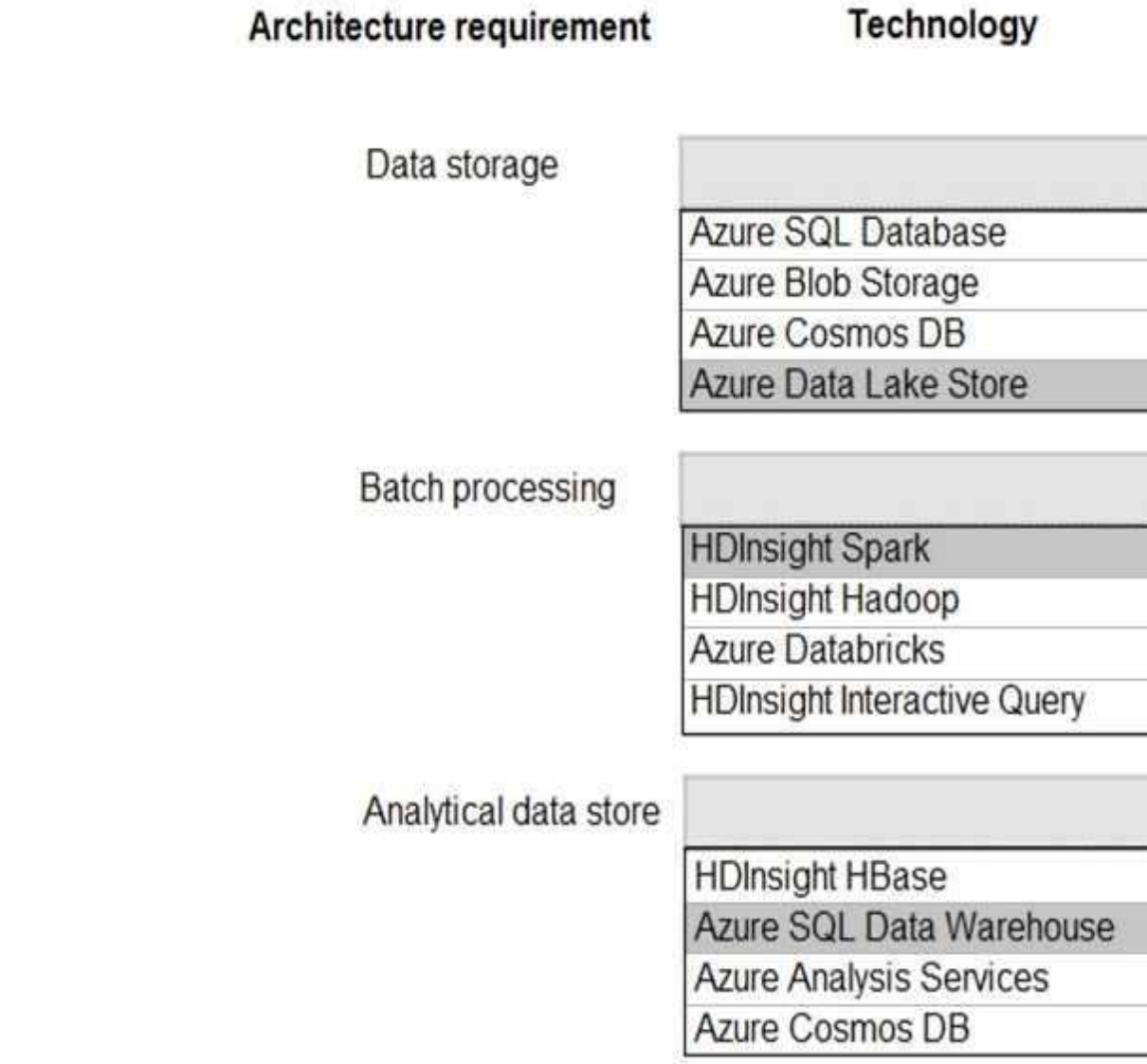
job. The higher the number of SUs, the more CPU and memory resources are allocated for your job.
This
capacity lets you focus on the query logic and abstracts the need to manage the hardware to run your Stream
Analytics job in a timely manner.
References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption

Question: 37

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

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

Answer:



Explanation:
Data storage: Azure Data Lake Store
A key mechanism that allows Azure Data Lake Storage Gen2 to provide file system performance at object storage scale and prices is the addition of a hierarchical namespace. This allows the collection of objects/files within an account to be organized into a hierarchy of directories and nested subdirectories in he same way that the file system on your computer is organized. With the hierarchical namespace enabled, a storage account becomes capable of providing the scalability and cost-effectiveness of object storage, with file system semantics that are familiar to analytics engines and frameworks.
Batch processing: HD Insight Spark
Aparch Spark is an open-source, parallel-processing framework that supports in-memory processing to boost the performance of big-data analysis applications.
HDInsight is a managed Hadoop service. Use it deploy and manage Hadoop clusters in Azure. For batch processing, you can use Spark, Hive, Hive LLAP, MapReduce.

Languages: R, Python, Java, Scala, SQL
Analytic data store: SQL Data Warehouse

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

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

Question: 38

You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account.
The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/.
You need to design a daily Azure Data Factory data load to minimize the data transfer between the two
accounts.
Which two configurations should you include in the design? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Delete the files in the destination before loading new data.
B. Filter by the last modified date of the source files.

C. Delete the source files after they are copied.
D. Specify a file naming pattern for the destination.

Answer: B,C

Reference:
https://docs.microsoft.com/en-us/azu e/data-factory/connector-azure-data-lake-storage

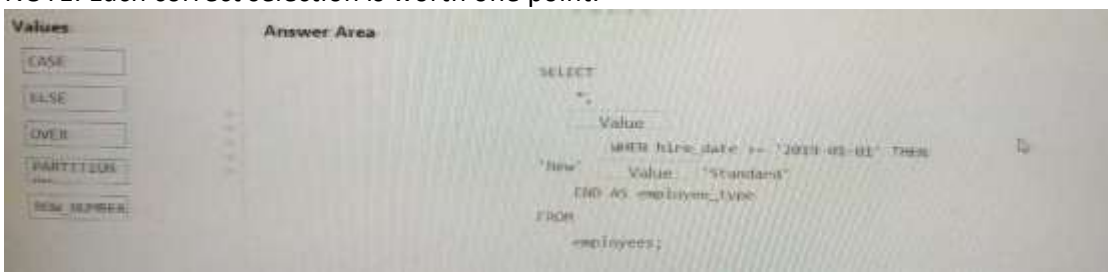
Question: 39

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.



Answer:

Question: 40

A company purchases IoT devices to monitor manufacturing machinery. The company uses an IoT appliance to communicate with the IoT devices.
The company must be able to monitor the devices in real-time.
You need to design the solution.
What should you recommend?

A. Azure Stream Analytics cloud job using Azure PowerShell
B. Azure Analysis Services using Azure Portal
C. Azure Data Factory instance using Azure Portal
D. Azure Analysis Services using Azure PowerShell

Answer: A

Explanation:
Stream Analytics is a cost-effective event p ocessing engine that helps uncover real-time insights from devices, sensors, infrastructure, applications and data quickly and easily.
Monitor and manage Stream Analytic resources with Azure PowerShell cmdlets and powershell scripting that execute basic Stream An lytics tasks.
Reference:
https://cloudblogs.microsoft. om/sqlserver/2014/10/29/microsoft-adds-iot-streaming-analytics-data-production-and-workflow-services-to-azure/

Question: 41

You are designing a statistical analysis solution that will use custom proprietary1 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 Stream Analytics
B. Azure SQL Database
C. Azure Databricks
D. Azure Synapse Analytics

Answer: A

Question: 42

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
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

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

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.

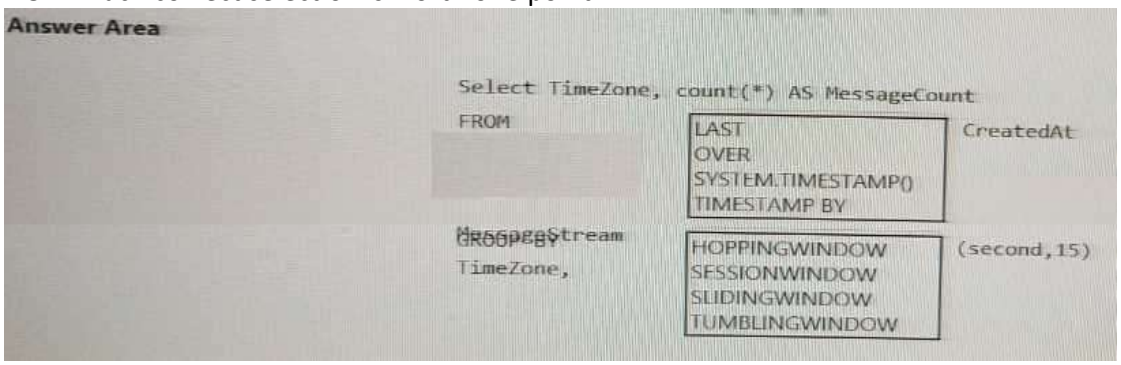
Note: Remember Streaming Unit (SU) count, which is the unit of scale for Azure Stream Analytics, must be adjus ed so the number of physical resources available to the job can fit the

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/

Question: 43

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.



Answer:

1. SYSTEM.TIMESTAMP()
2. HOPPINGWINDOW

Question: 44

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