DP-203 Q&As

Data Engineering on Microsoft Azure

Pass Microsoft DP-203 Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.certbus.com/dp-203.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Microsoft Official Exam Center

QUESTION 1

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:

1.

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

2.

A workload for jobs that will run notebooks that use Python, Scala, and SOL.

3.

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:

1.

The data engineers must share a cluster.

2.

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.

3.

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

Correct Answer: B

We need a High Concurrency cluster for the data engineers and the jobs.

Note:

Standard clusters are recommended for a single user. Standard can run workloads developed in any language:

Python, R, Scala, and SQL.

A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

Reference:

https://docs.azuredatabricks.net/clusters/configure.html

QUESTION 2

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. MERGE

D. ALTER

Correct Answer: C

The Delta provides the ability to infer the schema for data input which further reduces the effort required in managing the schema changes. The Slowly Changing Data(SCD) Type 2 records all the changes made to each key in the dimensional

table. These operations require updating the existing rows to mark the previous values of the keys as old and then inserting new rows as the latest values. Also, Given a source table with the updates and the target table with dimensional data,

SCD Type 2 can be expressed with the merge.

Example:

// Implementing SCD Type 2 operation using merge function customersTable as("customers")

merge(

stagedUpdates.as("staged_updates"),

"customers.customerId = mergeKey")

whenMatched("customers.current = true AND customers.address staged_updates.address") updateExpr(Map(

"current" -> "false",

"endDate" -> "staged_updates.effectiveDate"))

whenNotMatched()

insertExpr(Map(

"customerid" -> "staged_updates.customerId",

"address" -> "staged_updates.address",

"current" -> "true",

"effectiveDate" -> "staged_updates.effectiveDate", "endDate" -> "null")) execute()

}

Reference:

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

QUESTION 3

A company purchases IoT devices to monitor manufacturing machinery. The company uses an Azure IoT Hub 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 Analysis Services using Azure PowerShell

B. Azure Stream Analytics Edge application using Microsoft Visual Studio

C. Azure Analysis Services using Microsoft Visual Studio

D. Azure Data Factory instance using Azure Portal

Correct Answer: B

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.

You can use Stream Analytics tools for Visual Studio to author, debug, and create your Stream Analytics Edge jobs. After you create and test the job, you can go to the Azure portal to deploy it to your devices.

Incorrect:

Not A, not C: Azure Analysis Services is a fully managed platform as a service (PaaS) that provides enterprise-grade data models in the cloud. Use advanced mashup and modeling features to combine data from multiple data sources,

define

metrics, and secure your data in a single, trusted tabular semantic data model.

Reference:

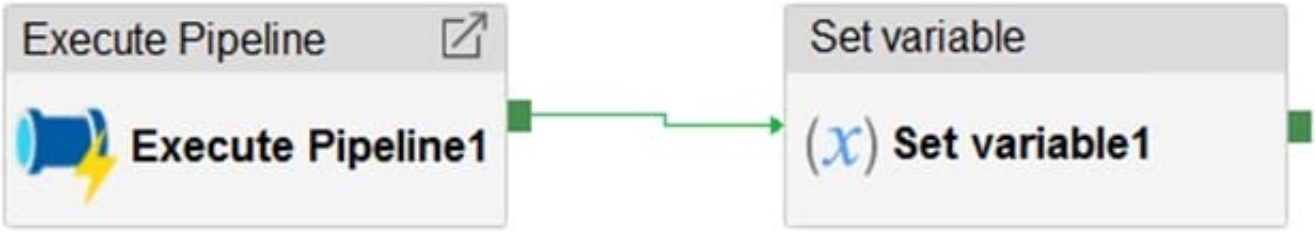
https://docs.microsoft.com/en-us/azure/iot-hub/monitor-iot-hub

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-tools-for-visual-studio-edge-jobs

QUESTION 4

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.

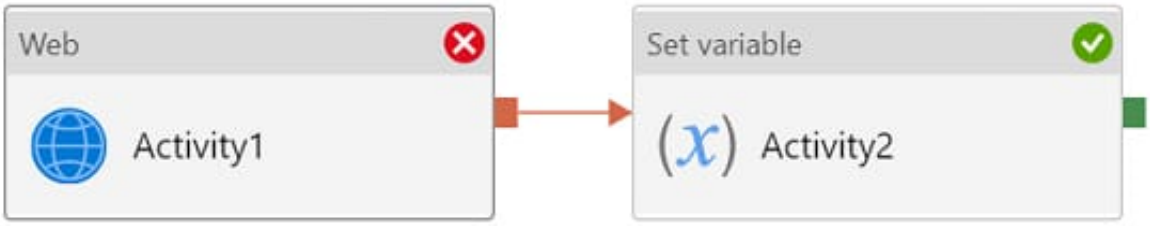
Correct Answer: A

Activities are linked together via dependencies. A dependency has a condition of one of the following: Succeeded, Failed,Skipped, or Completed.

Consider Pipeline1:

If we have a pipeline with two activities where Activity2 has a failure dependency on Activity1, the pipeline will not fail just because Activity1 failed. If Activity1 fails and Activity2 succeeds, the pipeline will succeed. This scenario is treated as a

try-catch block by Data Factory.



The failure dependency means this pipeline reports success.

Note:

If we have a pipeline containing Activity1 and Activity2, and Activity2 has a success dependency on Activity1, it will only execute if Activity1 is successful. In this scenario, if Activity1 fails, the pipeline will fail.

Reference:

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

QUESTION 5

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:

1.

New data is accessed frequently and must be available as quickly as possible.

2.

Data that is older than five years is accessed infrequently but must be available within one second when requested.

3.

Data that is older than seven years is NOT accessed. After seven years, the data must be persisted at the lowest cost possible.

4.

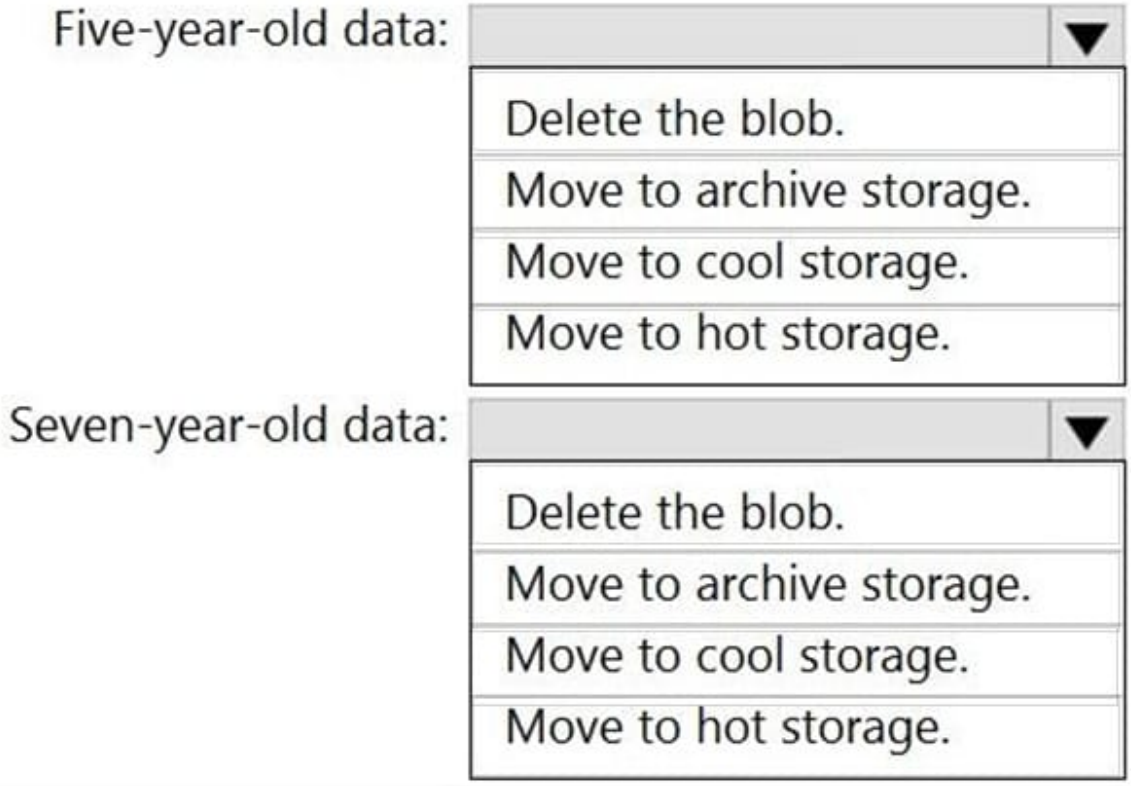
Costs must be minimized while maintaining the required availability.

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

NOTE: Each correct selection is worth one point

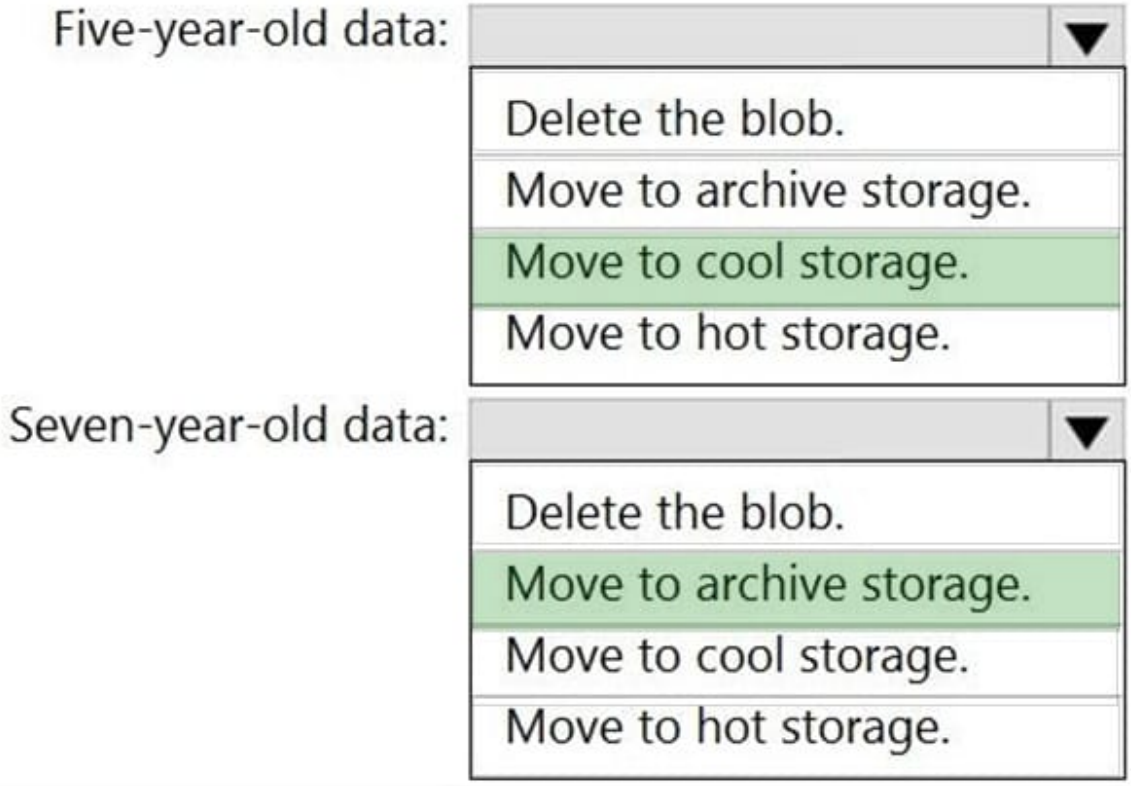
Hot Area:

Answer Area



Correct Answer:

Answer Area



Box 1: Move to cool storage

Box 2: Move to archive storage

Archive - Optimized for storing data that is rarely accessed and stored for at least 180 days with flexible latency requirements, on the order of hours.

The following table shows a comparison of premium performance block blob storage, and the hot, cool, and archive access tiers.

	Premium performance	Hot tier	Cool tier	Archive tier
Availability	99.9%	99.9%	99%	Offline
Availability (RA-GRS reads)	N/A	99.99%	99.9%	Offline
Usage charges	Higher storage costs, lower access, and transaction cost	Higher storage costs, lower access, and transaction costs	Lower storage costs, higher access, and transaction costs	Lowest storage costs, highest access, and transaction costs
Minimum storage duration	N/A	N/A	30 days[1]	180 days
Latency (Time to first byte)	Single-digit milliseconds	milliseconds	milliseconds	hours[2]

Reference: https://docs.microsoft.com/en-us/azure/storage/blobs/storage-blob-storage-tiers

QUESTION 6

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

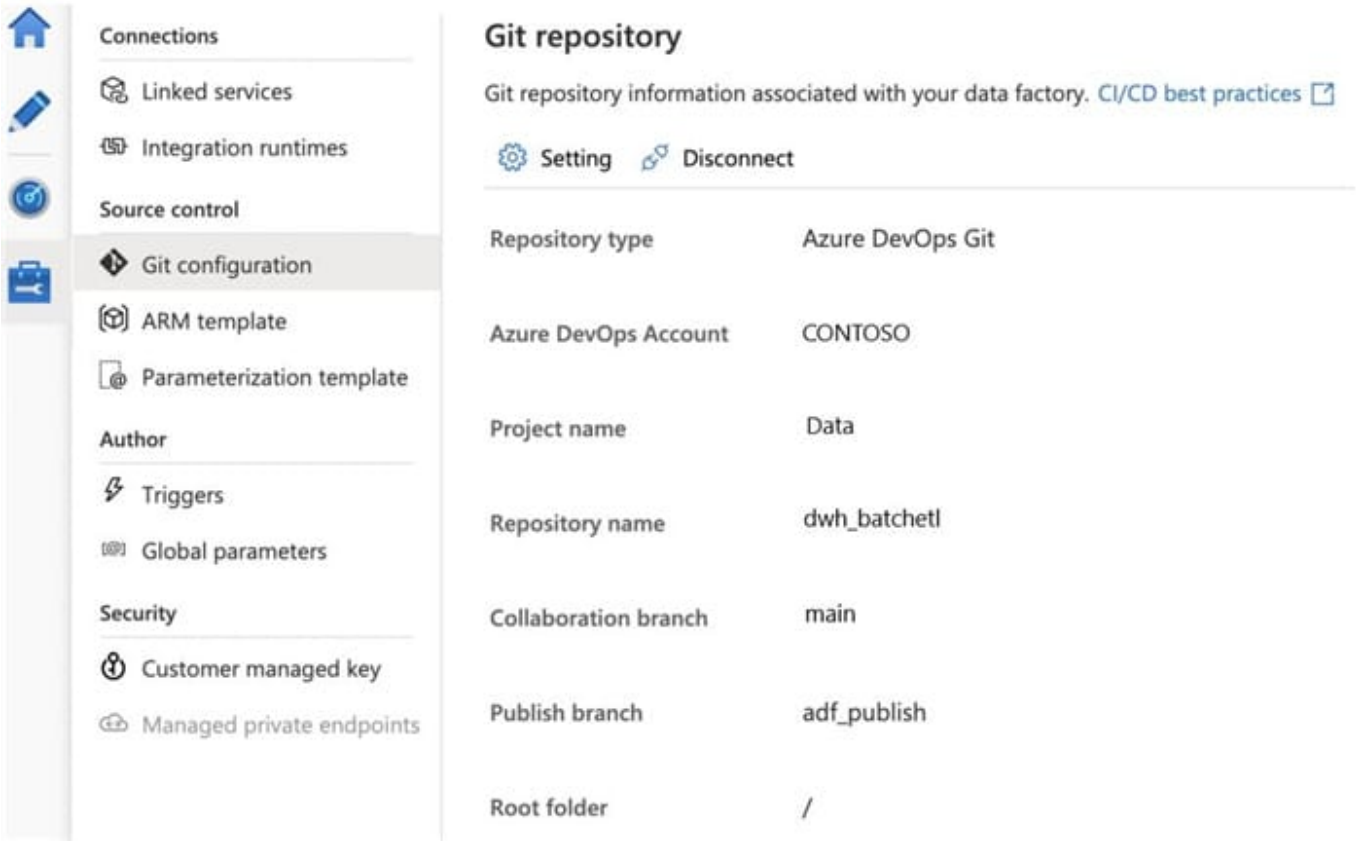
Correct Answer: A

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

QUESTION 7

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]**

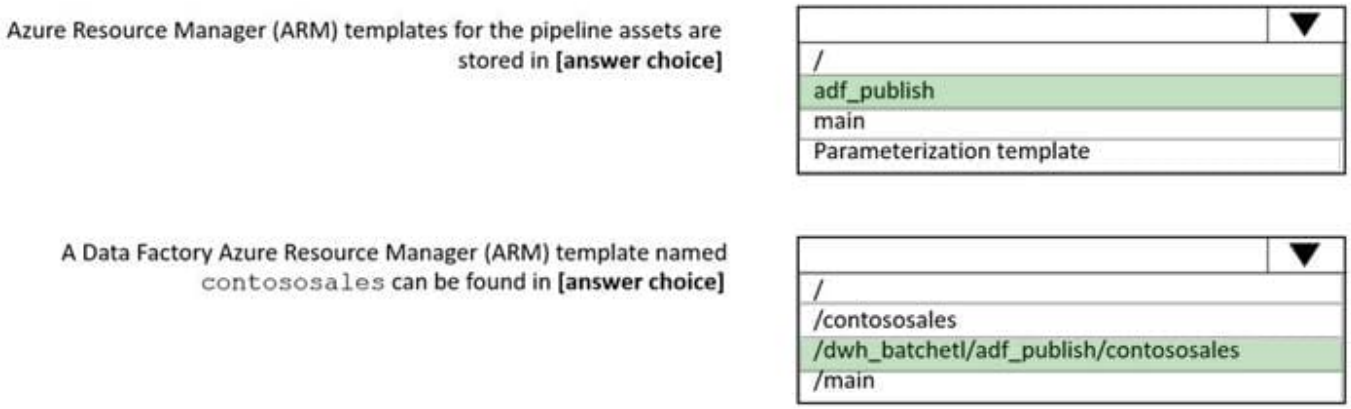
- /
- adf_publish
- main
- Parameterization template

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

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

Correct Answer:

Answer Area



Box 1: adf_publish

The Publish branch is the branch in your repository where publishing related ARM templates are stored and updated. By default, it\\'s adf_publish.

Box 2: / dwh_batchetl/adf_publish/contososales

Note: RepositoryName (here dwh_batchetl): Your Azure Repos code repository name. Azure Repos projects contain Git repositories to manage your source code as your project grows. You can create a new repository or use an existing

repository that\\'s already in your project.

Reference:

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

QUESTION 8

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 use?

A. Type 0

B. Type 1

C. Type 2

D. Type 3

Correct Answer: C

Type 2 - Creating a new additional record. In this methodology all history of dimension changes is kept in the database. You capture attribute change by adding a new row with a new surrogate key to the dimension table. Both the prior and

new rows contain as attributes the natural key(or other durable identifier). Also \\'effective date\\' and \\'current indicator\\' columns are used in this method. There could be only one record with current indicator set to \\'Y\\'. For

\\'effective date\\' columns,

i.e. start_date and end_date, the end_date

for current record usually is set to value 9999-12-31. Introducing changes to the dimensional model in type 2 could be very expensive database operation so it is not recommended to use it in dimensions where a new attribute could be added

in the future.

https://www.datawarehouse4u.info/SCD-Slowly-Changing-Dimensions.html

QUESTION 9

HOTSPOT

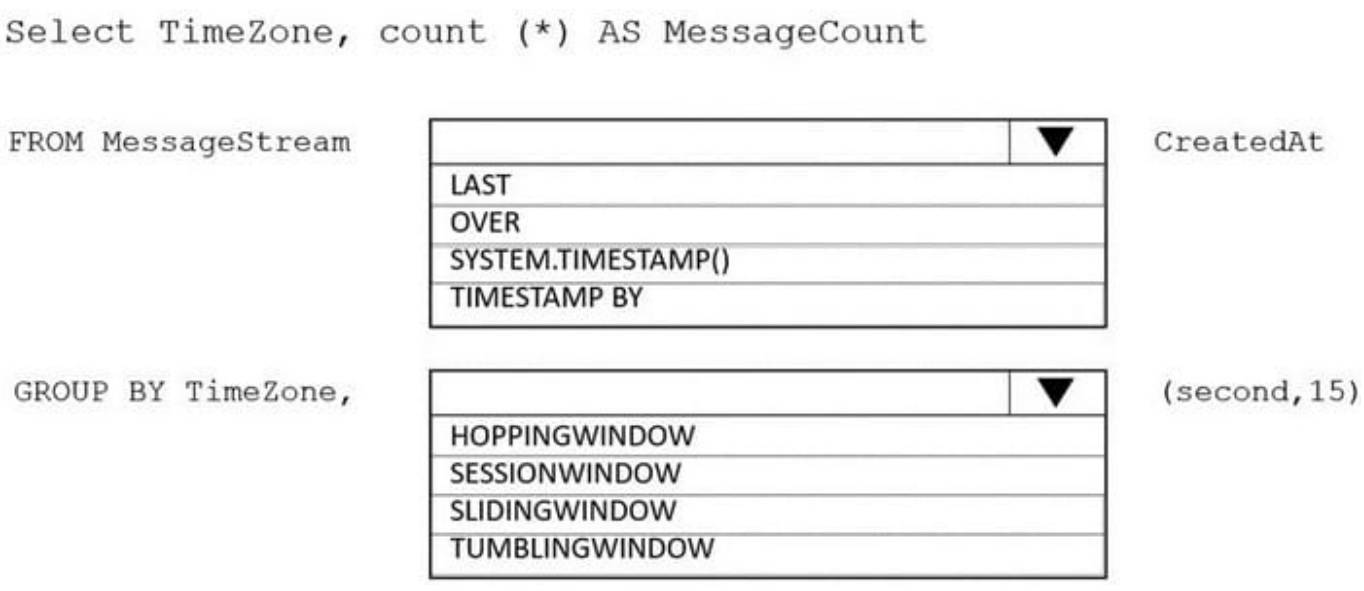
You are designing an Azure Stream Analytics solution that receives instant messaging data from an Azure event hub.

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:



Correct Answer:

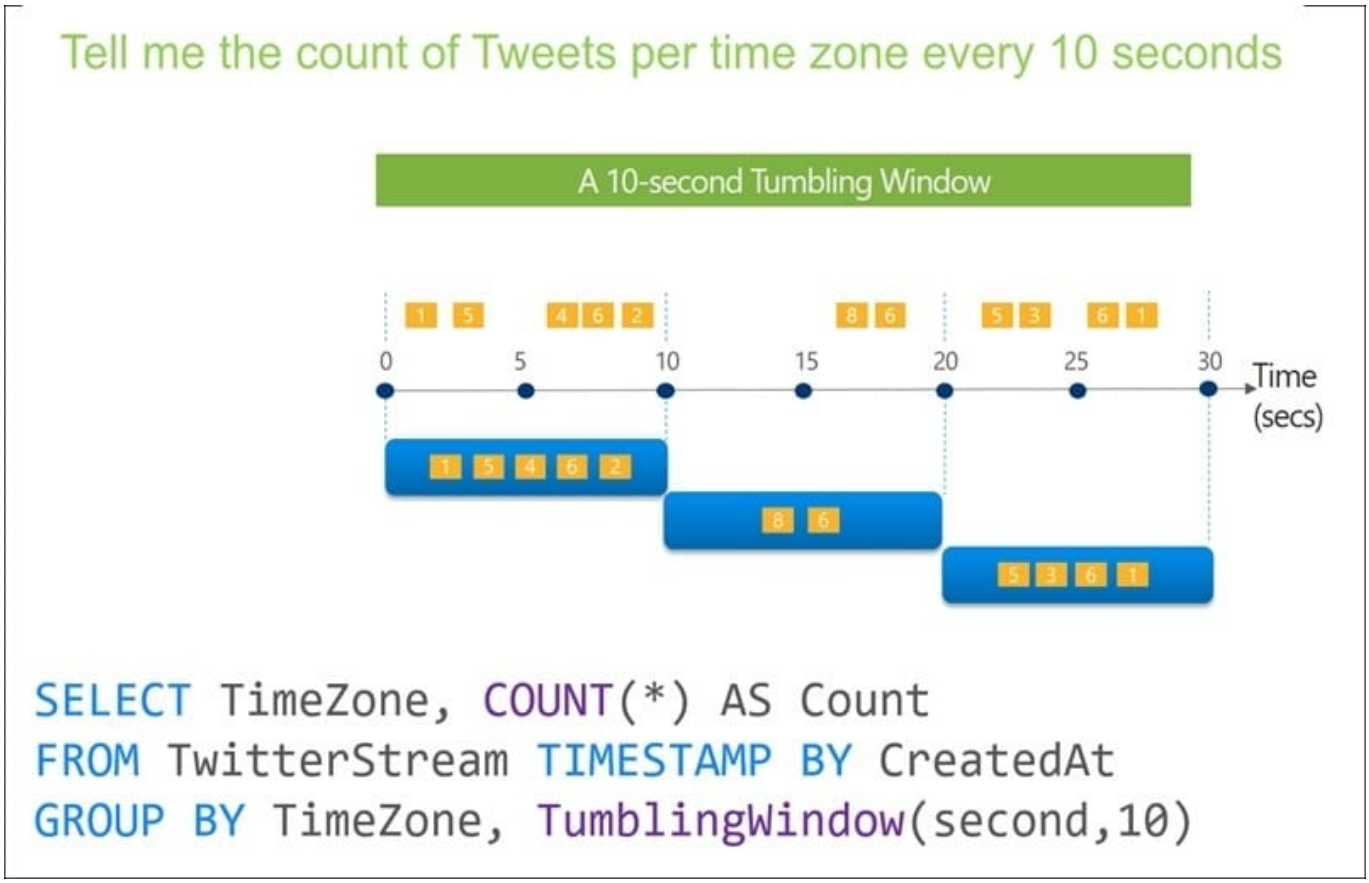
```
Select TimeZone, count (*) AS MessageCount

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

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

Box 1: timestamp by

Box 2: TUMBLINGWINDOW

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

and an event cannot belong to more than one tumbling window.



QUESTION 10

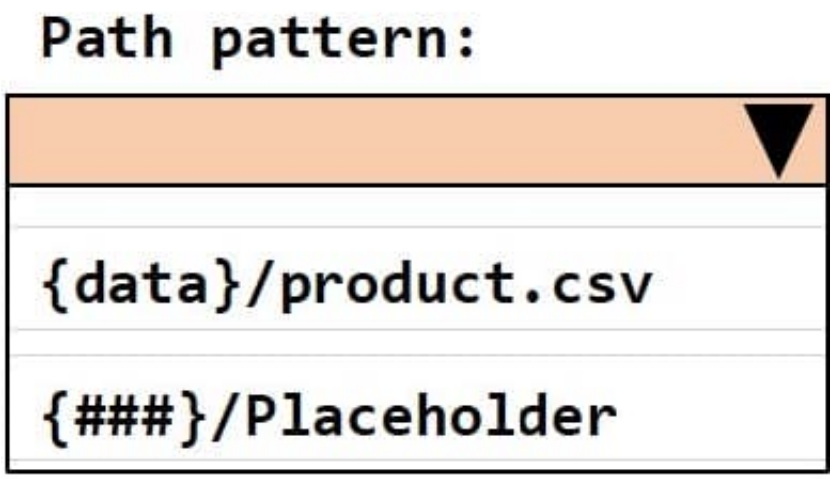
HOTSPOT

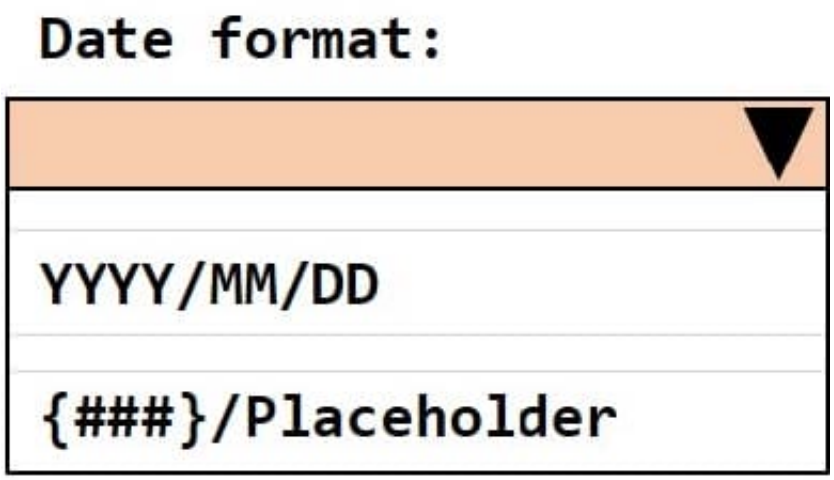
The storage account container view is shown in the Refdata exhibit. (Click the Refdata tab.) You need to configure the Stream Analytics job to pick up the new reference data. What should you configure? To answer, select the appropriate

options in the answer area NOTE:

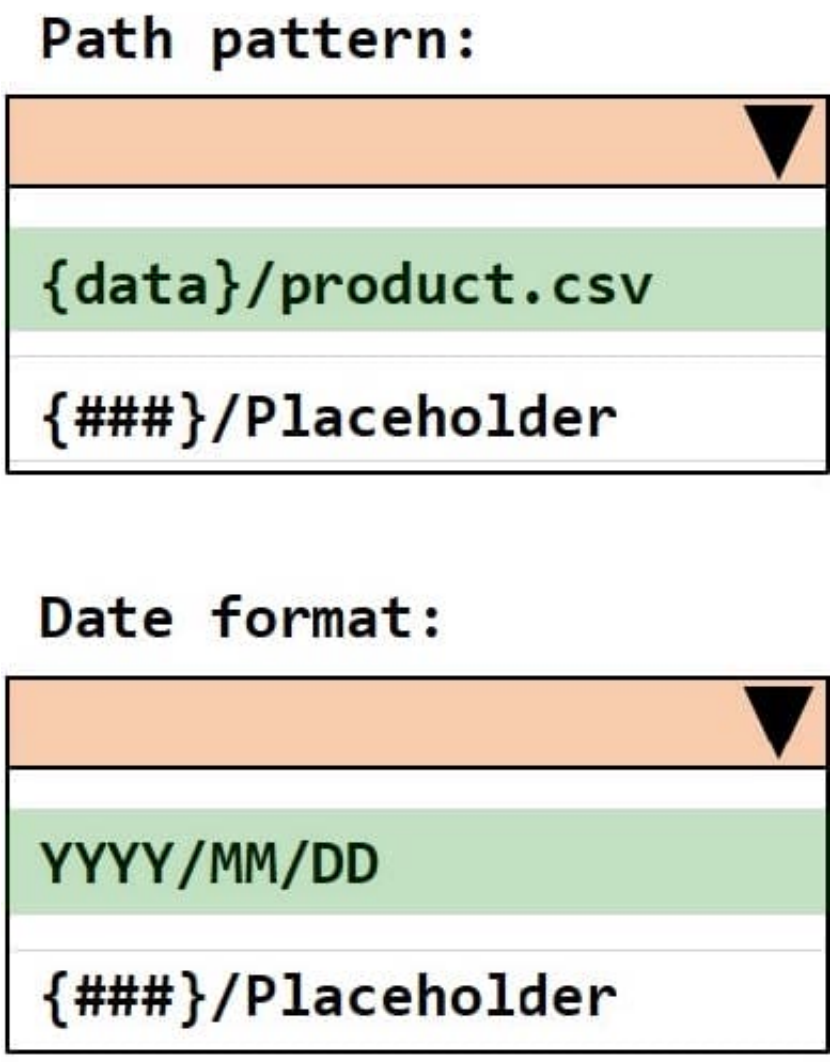
Each correct selection is worth one point.

Hot Area:





Correct Answer:



QUESTION 11

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. B. a five-minute Sliding window

C. a five-minute Tumbling window

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

Correct Answer: C

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 12

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:

Correct Answer:

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

QUESTION 13

You have an Azure Data Lake Storage Gen2 account named account1 that contains a container named container1.

You plan to create lifecycle management policy rules for container1.

You need to ensure that you can create rules that will move blobs between access tiers based on when each blob was accessed last.

What should you do first?

A. Configure object replication

B. Create an Azure application

C. Enable access time tracking

D. Enable the hierarchical namespace

Correct Answer: C

Generally available: Access time-based lifecycle management rules for Data Lake Storage Gen2

Some data in Azure Storage is written once and read many times. To effectively manage the lifecycle of such data and optimize your storage costs, it is important to know the last time of access for the data. When access time tracking is

enabled for a storage account, the last access time property on the file is updated when it is read. You can then define lifecycle management policies based on last access time:

Transition objects from hotter to cooler access tiers if the file has not been accessed for a specified duration.

Automatically transition objects from cooler to hotter access tiers when a file is accessed again.

Delete objects if they have not been accessed for an extended duration.

Access time tracking is only available for files in Data Lake Storage Gen2.

Reference:

https://azure.microsoft.com/en-us/updates/access-time-based-lifecycle-management-adls-gen2/

QUESTION 14

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)

Correct Answer: D

Geo-redundant storage (GRS) copies your data synchronously three times within a single physical location in the primary region using LRS. It then copies your data asynchronously to a single physical location in the secondary region.
Incorrect Answers:

B: Zone-redundant storage (ZRS) copies your data synchronously across three Azure availability zones in the primary region. For applications requiring high availability, Microsoft recommends using ZRS in the primary region, and also

replicating to a secondary region.

C: Locally redundant storage (LRS) copies your data synchronously three times within a single physical location in the primary region. LRS is the least expensive replication option, but is not recommended for applications requiring high

availability.

D: GZRS is more expensive compared to GRS.

Reference:

https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy

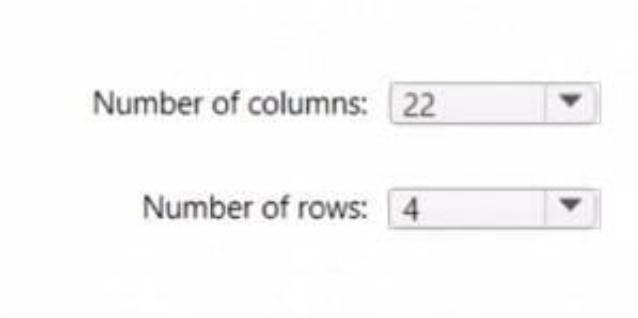
QUESTION 15

HOTSPOT

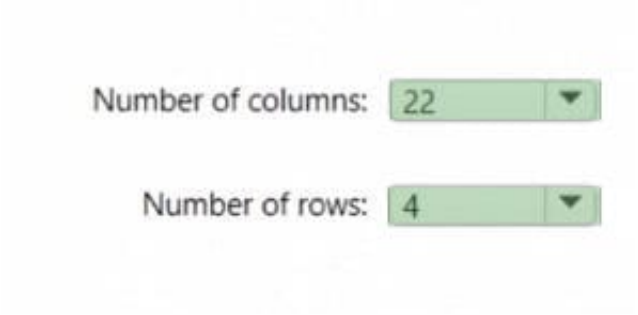
You have an Azure Data Factory pipeline that contains a data flow. The data flow contains the following expression.

```
source(output(
    License_plate as string,
    Make as string,
    Time as string
  ),
  allowSchemaDrift: true,
```

Hot Area:



Correct Answer:



Latest DP-203 Dumps DP-203 VCE Dumps DP-203 Braindumps