DUMPSBOSS.COM

# Data Engineering on Microsoft Azure

## Microsoft DP-203

## Version Demo

**Total Demo Questions: 20**

**Total Premium Questions: 439**

**Buy Premium PDF**

https://dumpsboss.com

support@dumpsboss.com

# Topic Break Down

| Topic | No. of Questions |
|---|---|
| Topic 2, New Update | 227 |
| Topic 3, Case Study 1 | 7 |
| Topic 4, Case Study 2 | 2 |
| Topic 5, Mixed Questions | 203 |
| Total | 439 |

## QUESTION NO: 1

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

You plan to implement a star schema in pool1 and create a new table named DimCustomer by using the following code.

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

You need to ensure that DimCustomer has the necessary columns to support a Type 2 slowly changing dimension (SCD). Which two columns should you add? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

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

**B.** [EffectiveEndDate] [datetime] NOT NULL

**C.** [PreviousModifiedDate] [datetime] NOT NULL

**D.** [RowID] [bigint] NOT NULL

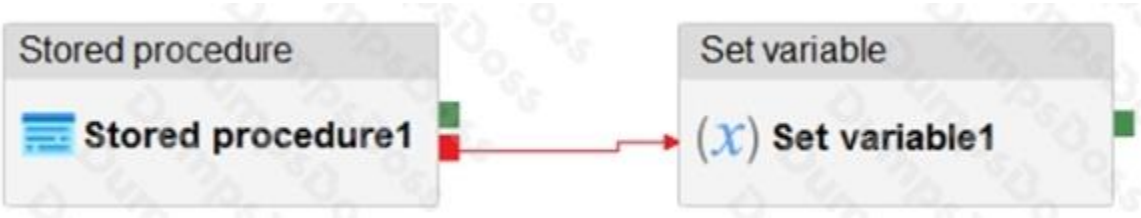
**E.** [EffectiveStartDate] [datetime] NOT NULL

**ANSWER: A B**

## QUESTION NO: 2

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.

**ANSWER: A**

**Explanation:**

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 NO: 3

You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date.

You need to ensure that the design provides the fastest query times of the records when querying for arbitrary date ranges and aggregating by fiscal calendar attributes.

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

NOTE: Each correct selection is worth one point.

**A.** Create a date dimension table that has a DateTime key.

**B.** Use built-in SQL functions to extract date attributes.

**C.** Create a date dimension table that has an integer key in the format of yyyymmdd.

**D.** In the fact table, use integer columns for the date fields.

**E.** Use DateTime columns for the date fields.

**ANSWER: B D**

## QUESTION NO: 4

You have an Azure SQL database named DB1 and an Azure Data Factory data pipeline named pipeline.

From Data Factory, you configure a linked service to DB1.

In DB1, you create a stored procedure named SP1. SP1 returns a single row of data that has four columns.

You need to add an activity to pipeline to execute SP1. The solution must ensure that the values in the columns are stored as pipeline variables.

Which two types of activities can you use to execute SP1? (Refer to Data Engineering on Microsoft Azure documents or guide for Answers/Explanation available at Microsoft.com)

**A.** Stored Procedure

**B.** Lookup

**C.** Script

**D.** Copy

**ANSWER: A B**

**Explanation:**

the two types of activities that you can use to execute SP1 are Stored Procedure and Lookup.

A Stored Procedure activity executes a stored procedure on an Azure SQL Database or Azure Synapse Analytics or SQL Server1. You can specify the stored procedure name and parameters in the activity settings1.

A Lookup activity retrieves a dataset from any data source that returns a single row of data with four columns2. You can use a query to execute a stored procedure as the source of the Lookup activity2. You can then store the values in the columns as pipeline variables by using expressions2.

https://learn.microsoft.com/en-us/azure/data-factory/transform-data-using-stored-procedure

**QUESTION NO: 5**

You have an Azure Data Factory pipeline that performs an incremental load of source data to an Azure Data Lake Storage Gen2 account.

Data to be loaded is identified by a column named LastUpdatedDate in the source table.

You plan to execute the pipeline every four hours.

You need to ensure that the pipeline execution meets the following requirements:

▪ Automatically retries the execution when the pipeline run fails due to concurrency or throttling limits. ▪ Supports backfilling existing data in the table.

Which type of trigger should you use?

**A.** event

**B.** on-demand

**C.** schedule

**D.** tumbling window

**ANSWER: D**

**Explanation:**

In case of pipeline failures, tumbling window trigger can retry the execution of the referenced pipeline automatically, using the same input parameters, without the user intervention. This can be specified using the property "retryPolicy" in the trigger definition.

Reference:

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

## QUESTION NO: 6 - (HOTSPOT)

HOTSPOT

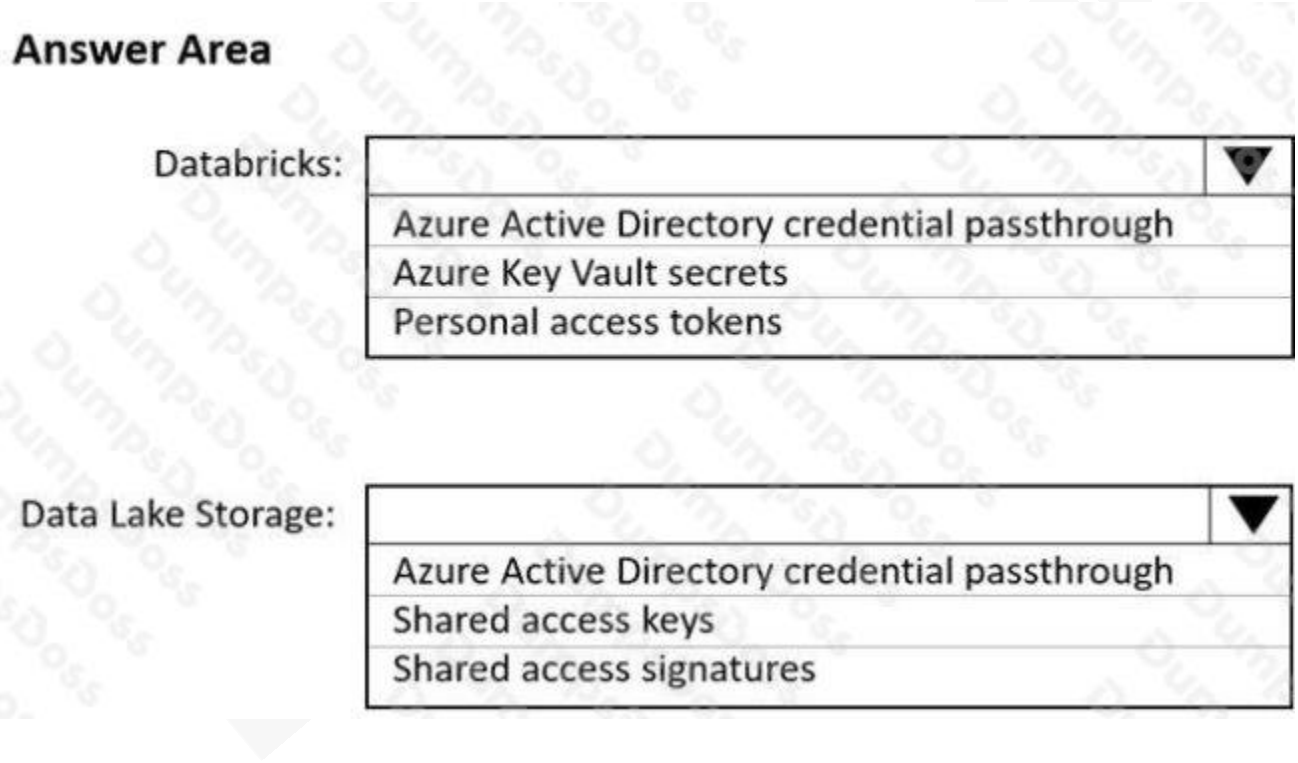
You use Azure Data Lake Storage Gen2 to store data that data scientists and data engineers will query by using Azure Databricks interactive notebooks. Users will have access only to the Data Lake Storage folders that relate to the projects on which they work.

You need to recommend which authentication methods to use for Databricks and Data Lake Storage to provide the users with the appropriate access. The solution must minimize administrative effort and development effort.

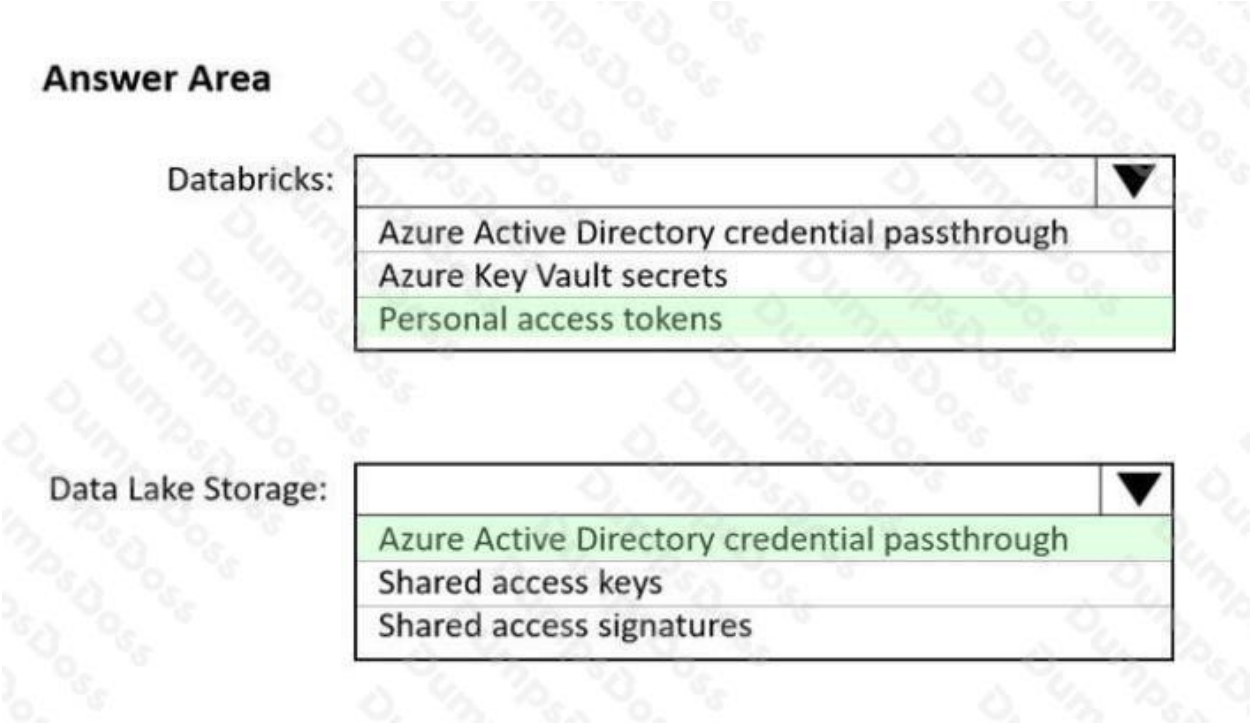
Which authentication method should you recommend for each Azure service? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

**Hot Area:**




**ANSWER:**




**Explanation:**

Box 1: Personal access tokens

You can use storage shared access signatures (SAS) to access an Azure Data Lake Storage Gen2 storage account directly. With SAS, you can restrict access to a storage account using temporary tokens with fine-grained access control.

You can add multiple storage accounts and configure respective SAS token providers in the same Spark session.

Box 2: Azure Active Directory credential passthrough

You can authenticate automatically to Azure Data Lake Storage Gen1 (ADLS Gen1) and Azure Data Lake Storage Gen2 (ADLS Gen2) from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that you use to log into Azure Databricks. When you enable your cluster for Azure Data Lake Storage credential passthrough, commands that you run on that cluster can read and write data in Azure Data Lake Storage without requiring you to configure service principal credentials for access to storage.

After configuring Azure Data Lake Storage credential passthrough and creating storage containers, you can access data directly in Azure Data Lake Storage Gen1 using an adl:// path and Azure Data Lake Storage Gen2 using an abfss:// path:

Reference:

https://docs.microsoft.com/en-us/azure/databricks/data/data-sources/azure/adls-gen2/azure-datalake-gen2-sas-access
https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough

## QUESTION NO: 7

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 NO: 8 - (DRAG DROP)

DRAG DROP

You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1. Branch1 contains an Azure Synapse pipeline named pipeline1.

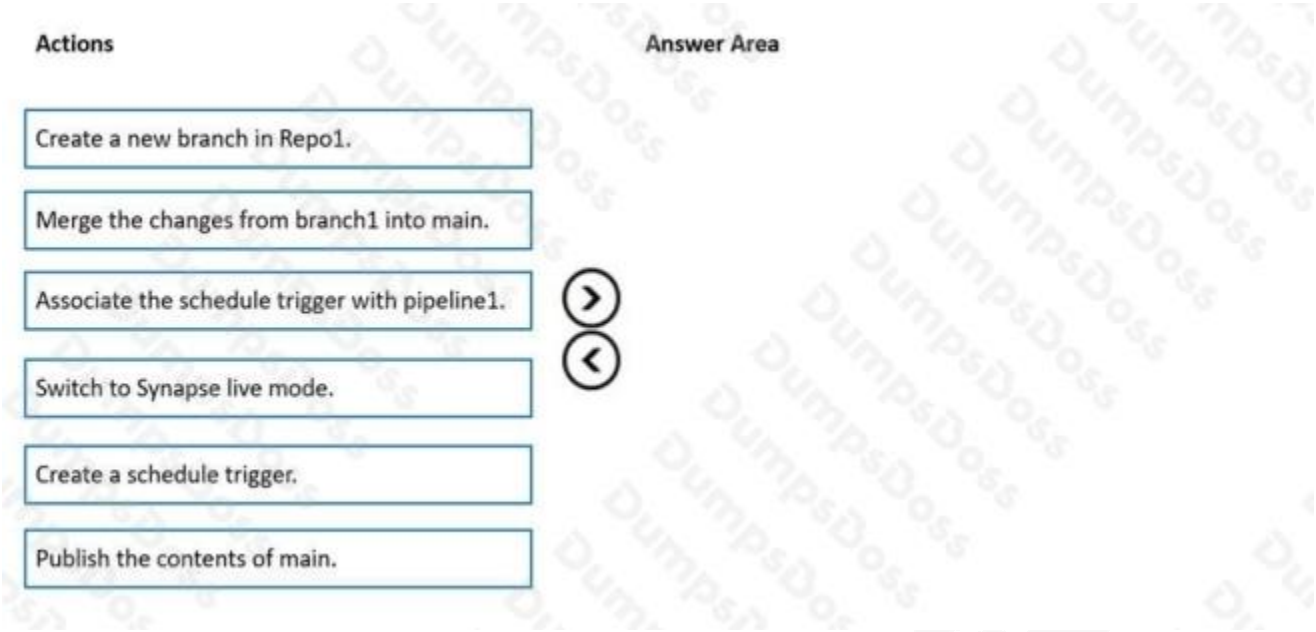
In workspace1, you complete testing of pipeline1.

You need to schedule pipeline1 to run daily at 6 AM.

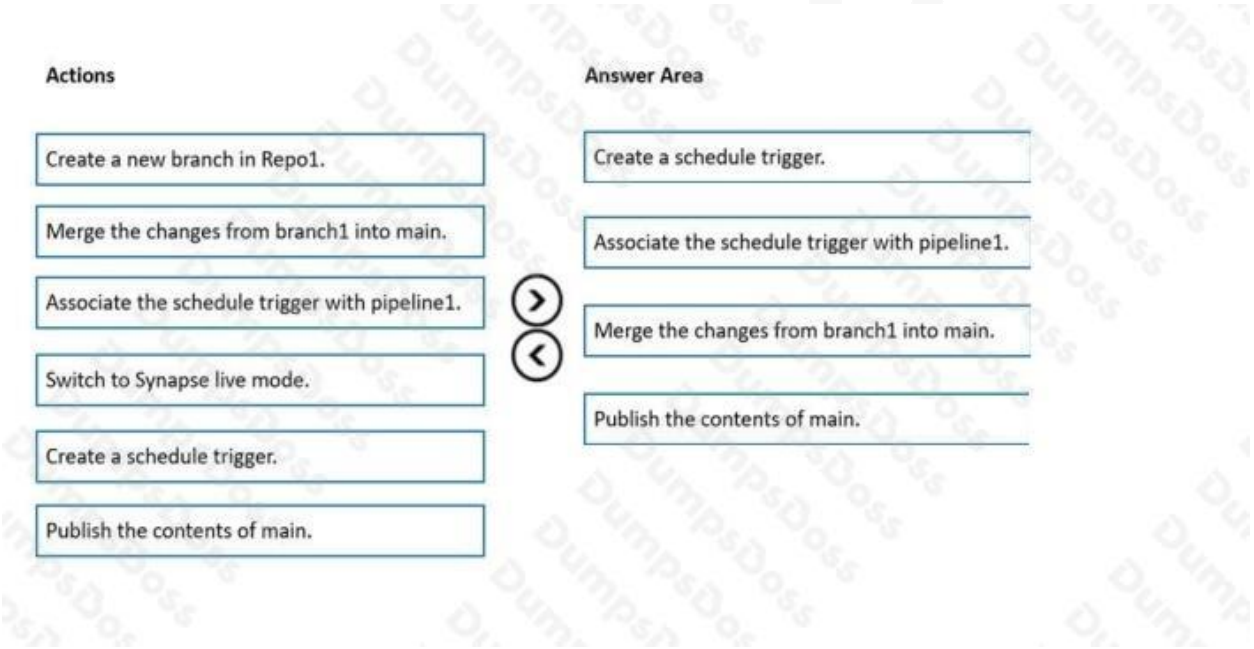
Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

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

**Select and Place:**




**ANSWER:**




**Explanation:**

## QUESTION NO: 9

You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1.

You need to use an Azure Synapse pipeline to copy data from the selected tables in DB1 to the files in storage1. The solution must meet the following requirements:

• The Copy activity in the pipeline must be parameterized to use the data in File1.txt to identify the source and destination of the copy.

• Copy activities must occur in parallel as often as possible.

Which two pipeline activities should you include in the pipeline? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

**A.** If Condition

**B.** ForEach

**C.** Lookup

**D.** Get Metadata

**ANSWER: B C**

## QUESTION NO: 10

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

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

Which type of window should you use?

**A.** snapshot

**B.** tumbling

**C.** hopping

**D.** sliding

**ANSWER: C**

**Explanation:**

Unlike tumbling windows, hopping windows model scheduled overlapping windows. A hopping window specification consist of three parameters: the timeunit, the windowsize (how long each window lasts) and the hopsize (by how much each window moves forward relative to the previous one).

Reference:

https://docs.microsoft.com/en-us/stream-analytics-query/hopping-window-azure-stream-analytics

## QUESTION NO: 11

You have an enterprise data warehouse in Azure Synapse Analytics.

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

The external table has three columns.

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

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

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

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

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

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

**A.** Option A

**B.** Option B

**C.** Option C

**D.** Option D

**ANSWER: C**

**Explanation:**

Incorrect Answers:

A, D: Only these Data Definition Language (DDL) statements are allowed on external tables:

- CREATE TABLE and DROP TABLE
- CREATE STATISTICS and DROP STATISTICS
- CREATE VIEW and DROP VIEW

Reference:

https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql

## QUESTION NO: 12

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

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

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

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

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

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

You need to create the Databricks clusters for the workloads.

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

Does this meet the goal?

**A.** Yes

**B.** No

**ANSWER: A**

**Explanation:**

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 NO: 13 - (SIMULATION)

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.

**ANSWER: Seetheanswerbelowinexplanation.**

**Explanation:**

Answer as below




## QUESTION NO: 14

You are designing a security model for an Azure Synapse Analytics dedicated SQL pool that will support multiple companies. You need to ensure that users from each company can view only the data of their respective company. Which two objects should you include in the solution? Each correct answer presents part of the solution

NOTE: Each correct selection it worth one point.

**A.** a custom role-based access control (RBAC) role.

**B.** asymmetric keys

**C.** a predicate function

**D.** a column encryption key

**E.** a security policy

**ANSWER: A E**

**Explanation:**

Reference:

https://docs.microsoft.com/en-us/sql/relational-databases/security/row-level-security

https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-access-control-overview

## QUESTION NO: 15

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

## QUESTION NO: 16 - (HOTSPOT)

HOTSPOT

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

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

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

NOTE: Each correct selection is worth one point.

**Hot Area:**

**Answer Area**




**ANSWER:**

**Answer Area**



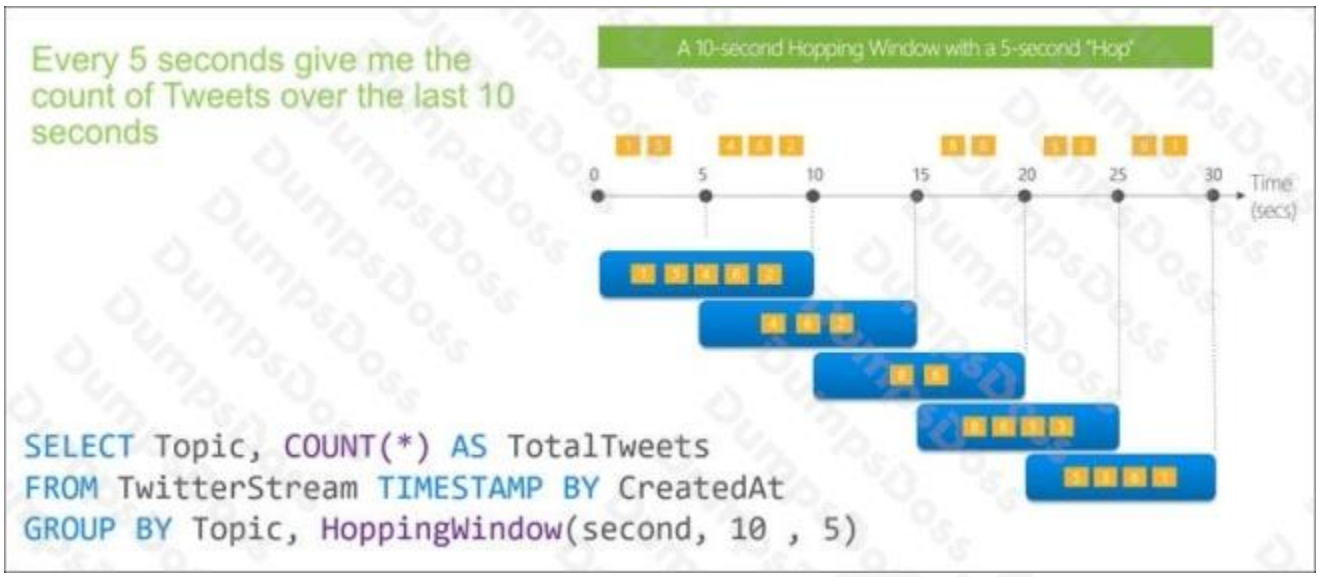

**Explanation:**

Box 1: TopOne OVER(PARTITION BY Game ORDER BY Score Desc)

TopOne returns the top-rank record, where rank defines the ranking position of the event in the window according to the specified ordering. Ordering/ranking is based on event columns and can be specified in ORDER BY clause.

Box 2: Hopping(minute,5)

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.




Reference: https://docs.microsoft.com/en-us/stream-analytics-query/topone-azure-stream-analytics
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

## QUESTION NO: 17

You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytics dedicated SQL pool. The CSV file contains columns named username, comment and date.

The data flow already contains the following:

• A source transformation

• A Derived Column transformation to set the appropriate types of data

• A sink transformation to land the data in the pool

You need to ensure that the data flow meets the following requirements;

• All valid rows must be written to the destination table.

• Truncation errors in the comment column must be avoided proactively.

• Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.

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

**A.** Add a select transformation that selects only the rows which will cause truncation errors.

**B.** Add a sink transformation that writes the rows to a file in blob storage.

**C.** Add a filter transformation that filters out rows which will cause truncation errors.

**D.** Add a Conditional Split transformation that separates the rows which will cause truncation errors.

**ANSWER: B D**

## QUESTION NO: 18

A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU).

You need to optimize performance for the Azure Stream Analytics job.

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

NOTE: Each correct selection is worth one point.

**A.** Implement event ordering.

**B.** Implement Azure Stream Analytics user-defined functions (UDF).

**C.** Implement query parallelization by partitioning the data output.

**D.** Scale the SU count for the job up.

**E.** Scale the SU count for the job down.

**F.** Implement query parallelization by partitioning the data input.

**ANSWER: D F**

**Explanation:**

D: Scale out the query by allowing the system to process each input partition separately.

F: A Stream Analytics job definition includes inputs, a query, and output. Inputs are where the job reads the data stream from.

Reference: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization

## QUESTION NO: 19

You have an Azure Databricks workspace and an Azure Data Lake Storage Gen2 account named storage!

New files are uploaded daily to storage1.

• Incrementally process new files as they are upkorage1 as a structured streaming source. The solution must meet the following requirements:

• Minimize implementation and maintenance effort.

• Minimize the cost of processing millions of files.

• Support schema inference and schema drift.

Which should you include in the recommendation?

**A.** Auto Loader

**B.** Apache Spark FileStreamSource

**C.** COPY INTO

**D.** Azure Data Factory

**ANSWER: D**

## QUESTION NO: 20 - (DRAG DROP)

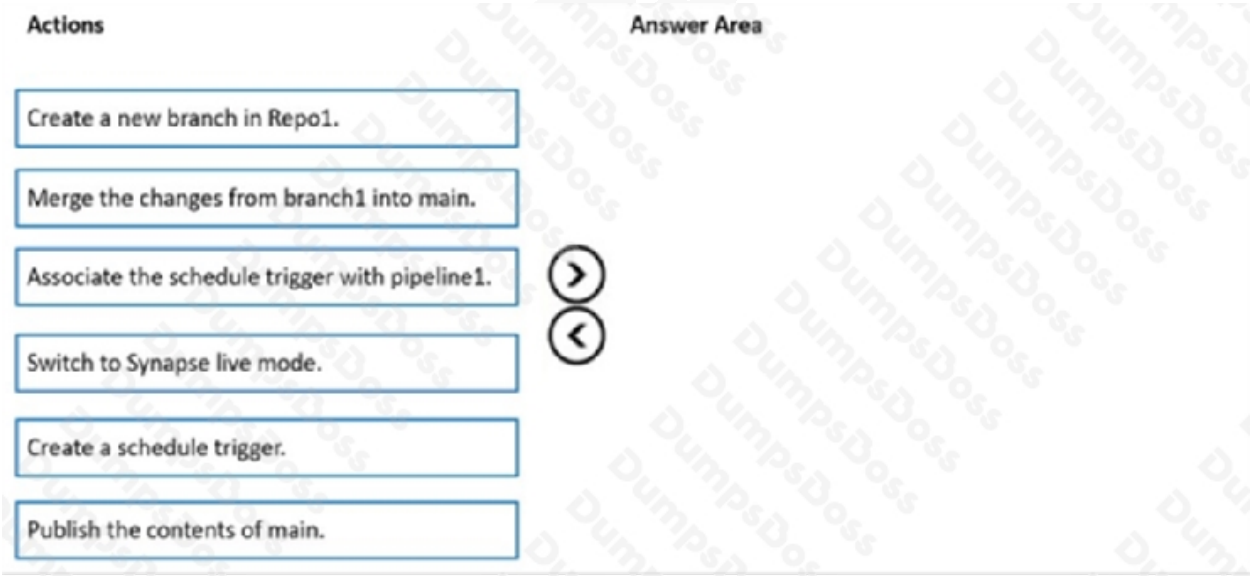
You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1. Branch1 contains an Azure Synapse pipeline named pipeline1.

In workspace1, you complete testing of pipeline1.

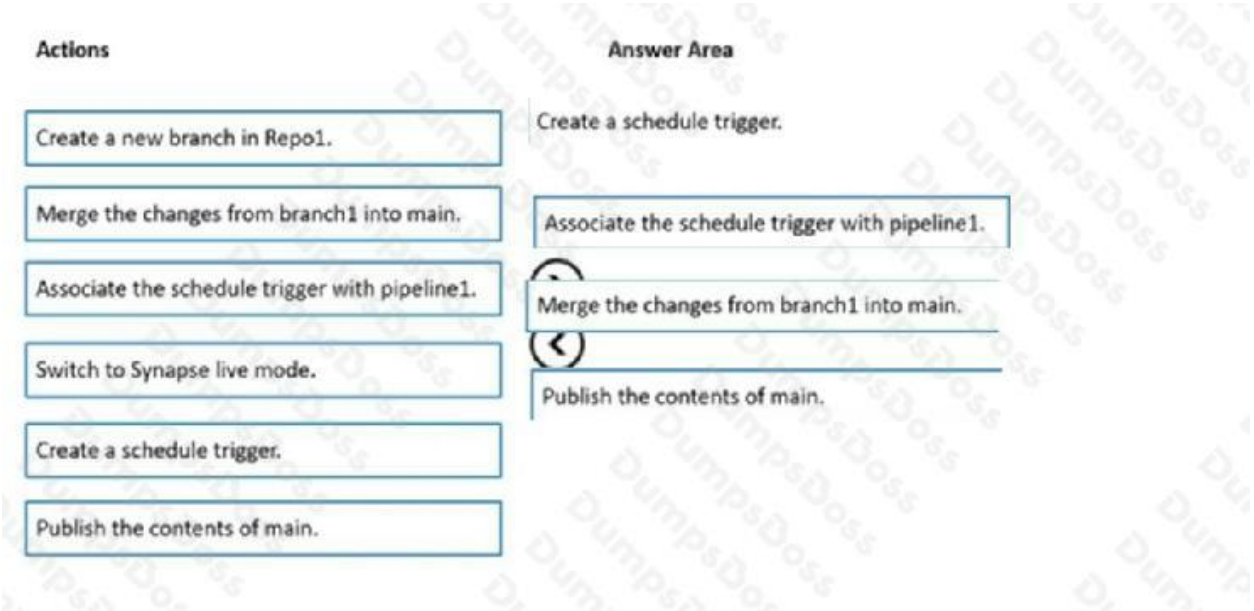
You need to schedule pipeline1 to run daily at 6 AM.

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

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




**ANSWER:**




**Explanation:**

