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.pass2lead.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

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:

1.

Ensure that the data remains in the UK South region at all times.

2.

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

Correct Answer: A

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

QUESTION 2

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 have an Azure Storage account that contains 100 GB of files. The files contain rows of text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics.

You need to prepare the files to ensure that the data copies quickly.

Solution: You copy the files to a table that has a columnstore index.

Does this meet the goal?

A. Yes

B. No

Correct Answer: B

Instead convert the files to compressed delimited text files.

Reference: https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

QUESTION 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

Correct Answer: B

Dynamic data masking helps prevent unauthorized access to sensitive data by enabling customers to designate how much of the sensitive data to reveal with minimal impact on the application layer. It\\'s a policy-based security feature that hides the sensitive data in the result set of a query over designated database fields, while the data in the database is not changed.

Reference: https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

QUESTION 4

You are designing a solution that will use tables in Delta Lake on Azure Databricks. You need to minimize how long it takes to perform the following:

1.

Queries against non-partitioned tables

2.

Joins on non-partitioned columns

Which two options should you include in the solution? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. the clone command

B. Z-Ordering

C. Apache Spark caching

D. dynamic file pruning (DFP)

Correct Answer: BD

Best practices: Delta Lake

B: Provide data location hints If you expect a column to be commonly used in query predicates and if that column has high cardinality (that is, a large number of distinct values), then use Z-ORDER BY. Delta Lake automatically lays out the data in the files based on the column values and uses the layout information to skip irrelevant data while querying.

BD: Dynamic file pruning, can significantly improve the performance of many queries on Delta Lake tables. Dynamic file pruning is especially efficient for non-partitioned tables, or for joins on non-partitioned columns. The performance impact

of dynamic file pruning is often correlated to the clustering of data so consider using Z-Ordering to maximize the benefit.

Incorrect:

Not C: Spark caching

Databricks does not recommend that you use Spark caching for the following reasons:

You lose any data skipping that can come from additional filters added on top of the cached DataFrame.

The data that gets cached might not be updated if the table is accessed using a different identifier (for example, you do spark.table(x).cache() but then write to the table using spark.write.save(/some/path).

Reference: https://learn.microsoft.com/en-us/azure/databricks/delta/best-practices#spark-caching
https://learn.microsoft.com/en-us/azure/databricks/optimizations/dynamic-file-pruning

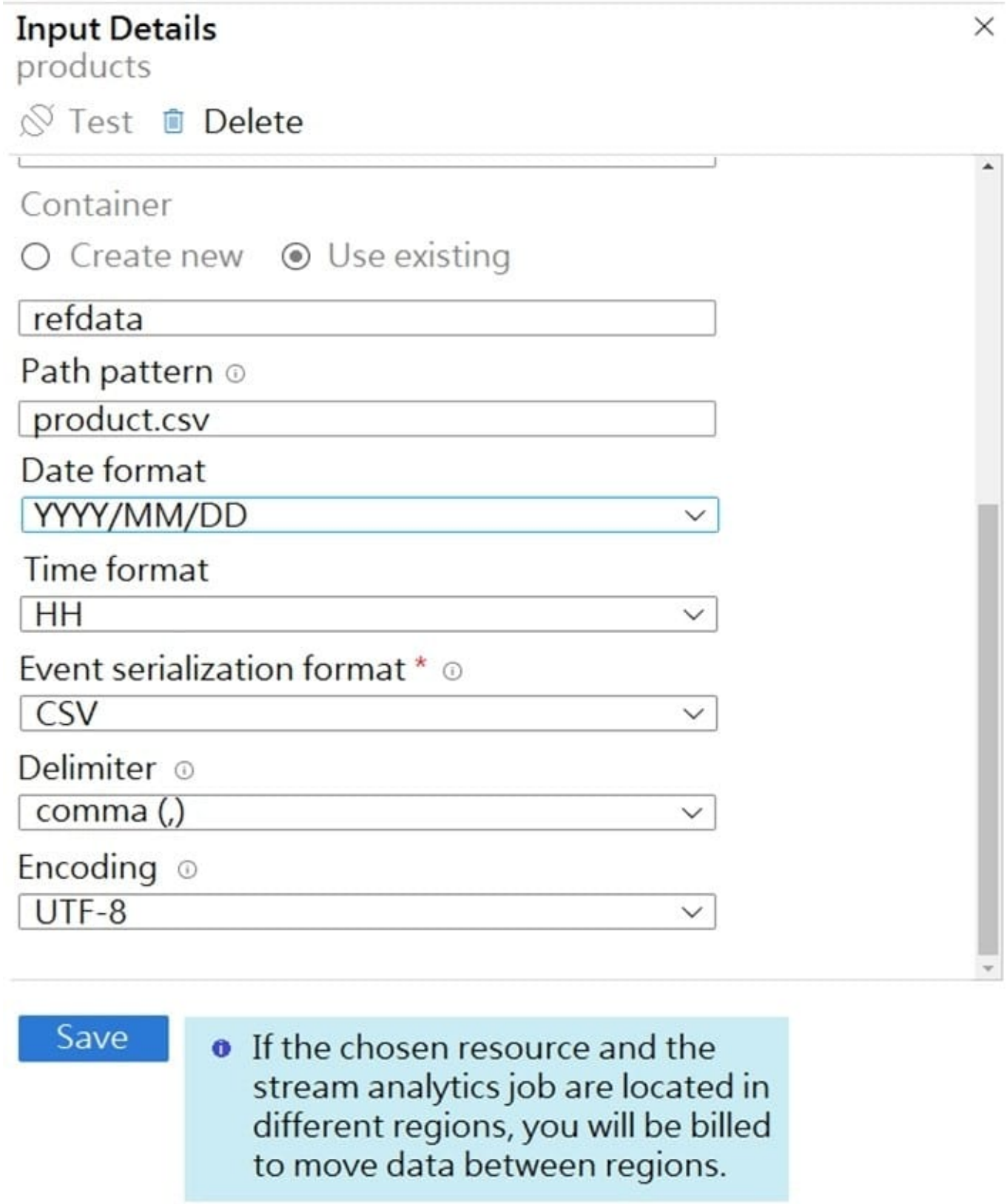
---

QUESTION 5

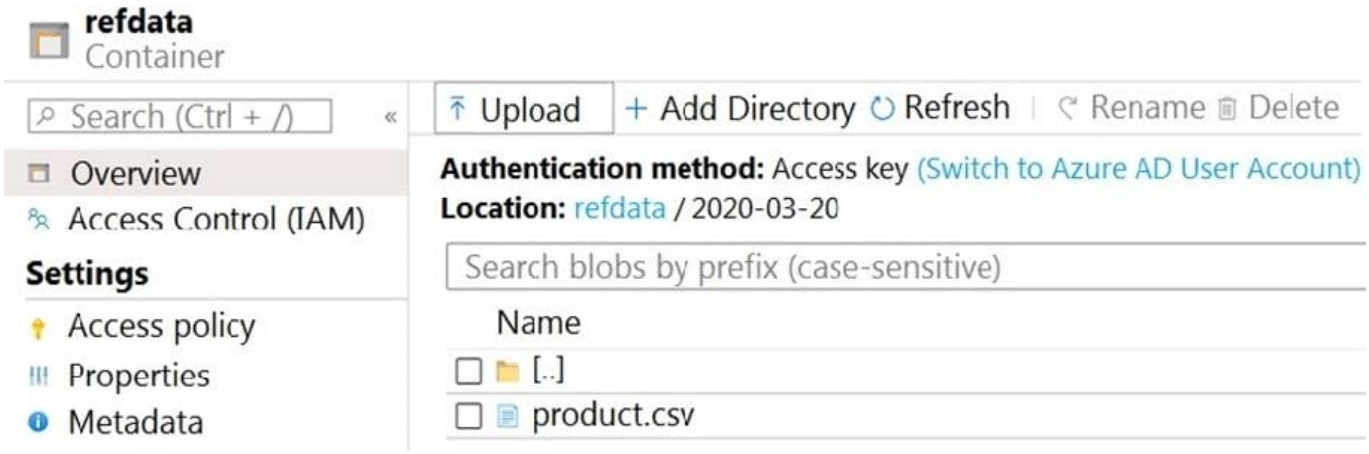
HOTSPOT

You are building an Azure Stream Analytics job that queries reference data from a product catalog file. The file is updated daily.

The reference data input details for the file are shown in the Input exhibit. (Click the Input tab.)



The storage account container view is shown in the Refdata exhibit. (Click the Refdata tab.)

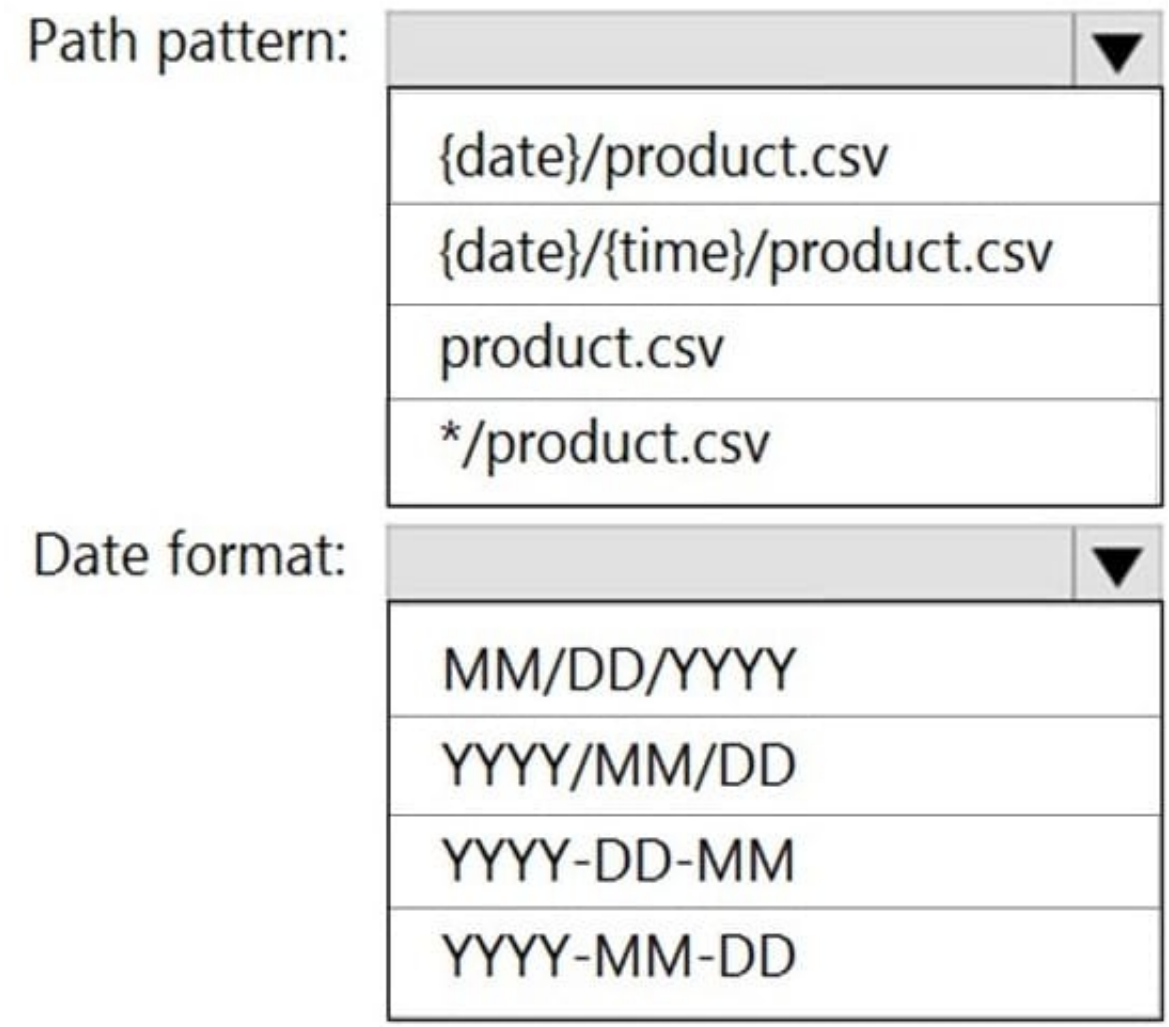


You need to configure the Stream Analytics job to pick up the new reference data.

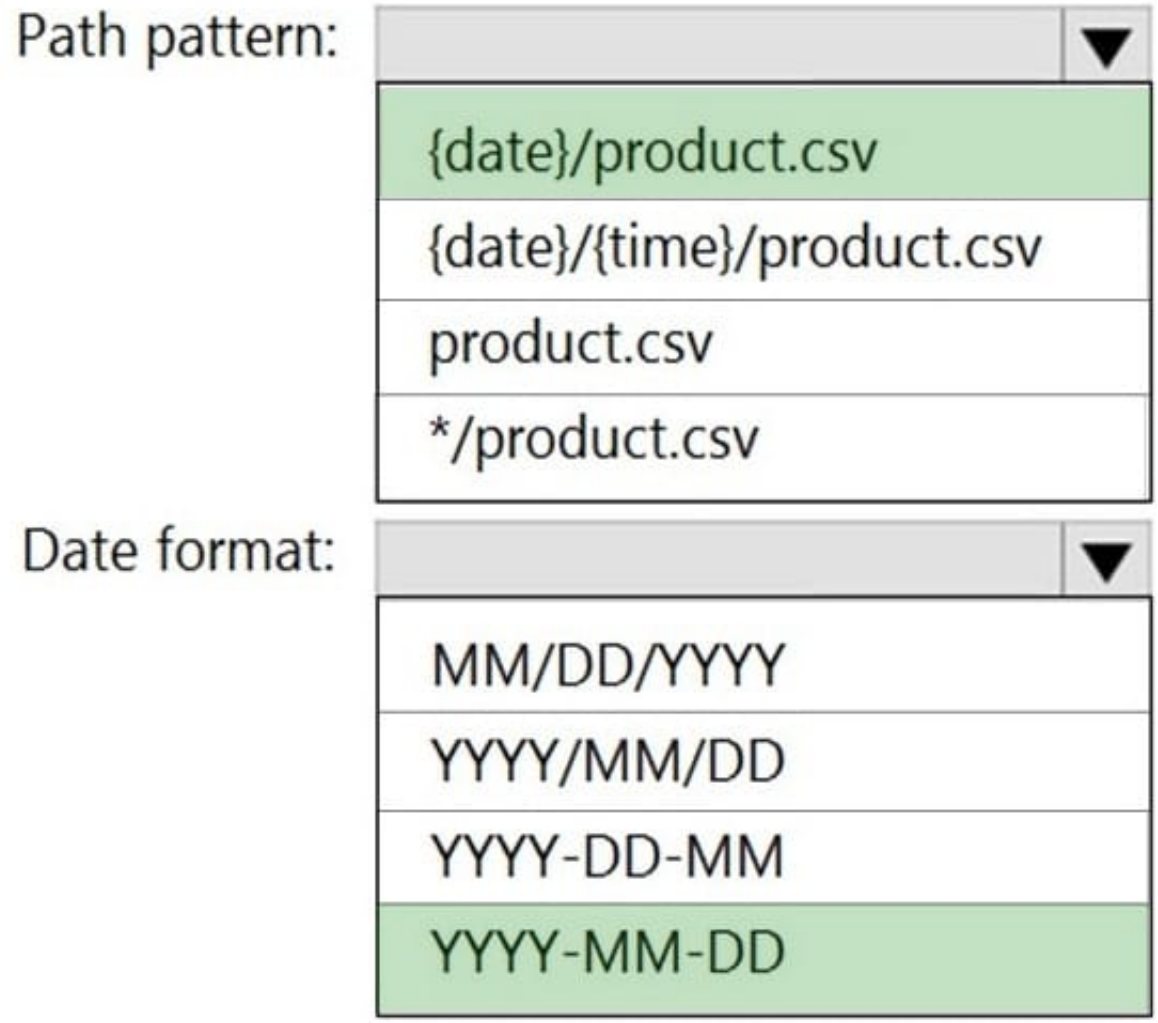
What should you configure? To answer, select the appropriate options in the answer area.

NOTE: Each correct selection is worth one point.

Hot Area:



Correct Answer:



Box 1: {date}/product.csv In the 2nd exhibit we see: Location: refdata / 2020-03-20 Note: Path Pattern: This is a required property that is used to locate your blobs within the specified container. Within the path, you may choose to specify one or more instances of the following 2 variables: {date}, {time} Example 1: products/{date}/{time}/product-list.csv Example 2: products/{date}/product-list.csv Example 3: product-list.csv

Box 2: YYYY-MM-DD

Note: Date Format [optional]: If you have used {date} within the Path Pattern that you specified, then you can select the date format in which your blobs are organized from the drop-down of supported formats. Example: YYYY/MM/DD, MM/DD/YYYY, etc.

QUESTION 6

You have an Azure Synapse Analytics dedicated SQL Pool1. Pool1 contains a partitioned fact table named dbo.Sales and a staging table named stg.Sales that has the matching table and partition definitions.

You need to overwrite the content of the first partition in dbo.Sales with the content of the same partition in stg.Sales. The solution must minimize load times.

What should you do?

A. Switch the first partition from dbo.Sales to stg.Sales.

B. Switch the first partition from stg.Sales to dbo. Sales.

C. Update dbo.Sales from stg.Sales.

D. Insert the data from stg.Sales into dbo.Sales.

Correct Answer: B

A way to eliminate rollbacks is to use Metadata Only operations like partition switching for data management. For example, rather than execute a DELETE statement to delete all rows in a table where the order_date was in October of 2001,

you could partition your data monthly. Then you can switch out the partition with data for an empty partition from another table

Note: Syntax:

SWITCH [PARTITION source_partition_number_expression] TO [schema_name.] target_table [PARTITION target_partition_number_expression]

Switches a block of data in one of the following ways:

1.

Reassigns all data of a table as a partition to an already-existing partitioned table.

2.

Switches a partition from one partitioned table to another.

3.

Reassigns all data in one partition of a partitioned table to an existing non-partitioned table.

Reference: https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

QUESTION 7

You have an Azure Data Factory pipeline named pipeline1.

You need to execute pipeline1 at 2 AM every day. The solution must ensure that if the trigger for pipeline1 stops, the next pipeline execution will occur at 2 AM, following a restart of the trigger.

Which type of trigger should you create?

A. schedule

B. tumbling

C. storage event

D. custom event

Correct Answer: A

Azure Data Factory introduced the three types of Triggers that specify why the pipeline will be fired; The Schedule trigger that allows you specify the date and time when the pipeline will be executed, the Tumbling window trigger in which the pipeline will be executed on a periodic interval, with the ability to save the pipeline state, and the Event-based trigger that will execute the pipeline as a response to a blob related event.

Reference: https://www.serverlessnotes.com/docs/schedule-azure-data-factory-pipeline-executions

QUESTION 8

You have an Azure Data Factory pipeline named Pipeline1!. Pipelinel contains a copy activity that sends data to an Azure Data Lake Storage Gen2 account. Pipeline 1 is executed by a schedule trigger.

You change the copy activity sink to a new storage account and merge the changes into the collaboration branch.

After Pipelinel executes, you discover that data is NOT copied to the new storage account.

You need to ensure that the data is copied to the new storage account.

What should you do?

A. Publish from the collaboration branch.

B. Configure the change feed of the new storage account.

C. Create a pull request.

D. Modify the schedule trigger.

Correct Answer: A

CI/CD lifecycle

1.

A development data factory is created and configured with Azure Repos Git. All developers should have permission to author Data Factory resources like pipelines and datasets.

2.

A developer creates a feature branch to make a change. They debug their pipeline runs with their most recent changes

3.

After a developer is satisfied with their changes, they create a pull request from their feature branch to the main or collaboration branch to get their changes reviewed by peers.

4.

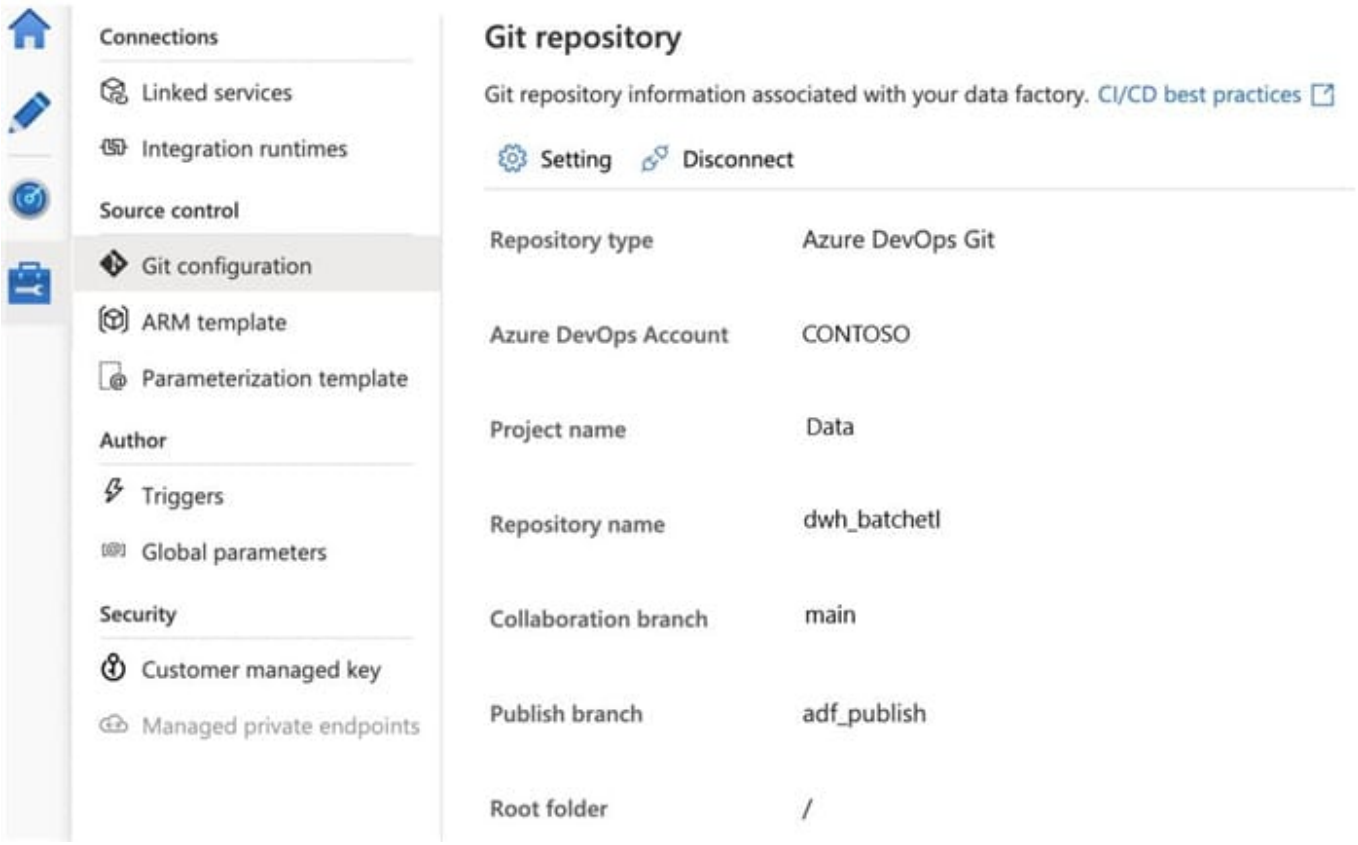
After a pull request is approved and changes are merged in the main branch, the changes get published to the development factory.

Reference: https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-delivery

QUESTION 9

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

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

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 10

You have an Azure Synapse Analytics dedicated SQL pool.

You need to Create a fact table named Table1 that will store sales data from the last three years. The solution must be

optimized for the following query operations:

1.

Show order counts by week.

2.

Calculate sales totals by region.

3.

Calculate sales totals by product.

4.

Find all the orders from a given month. Which data should you use to partition Table1?

A. region

B. product

C. week

D. month

Correct Answer: D

Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.

Benefits to queries Partitioning can also be used to improve query performance. A query that applies a filter to partitioned data can limit the scan to only the qualifying partitions. This method of filtering can avoid a full table scan and only scan a smaller subset of data. With the introduction of clustered columnstore indexes, the predicate elimination performance benefits are less beneficial, but in some cases there can be a benefit to queries.

For example, if the sales fact table is partitioned into 36 months using the sales date field, then queries that filter on the sale date can skip searching in partitions that don\\'t match the filter.

Note: Benefits to loads The primary benefit of partitioning in dedicated SQL pool is to improve the efficiency and performance of loading data by use of partition deletion, switching and merging. In most cases data is partitioned on a date column that is closely tied to the order in which the data is loaded into the SQL pool. One of the greatest benefits of using partitions to maintain data is the avoidance of transaction logging. While simply inserting, updating, or deleting data can be the most straightforward approach, with a little thought and effort, using partitioning during your load process can substantially improve performance.

Reference: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data- warehouse/sql-data- warehouse-tables-partition

---

QUESTION 11

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.

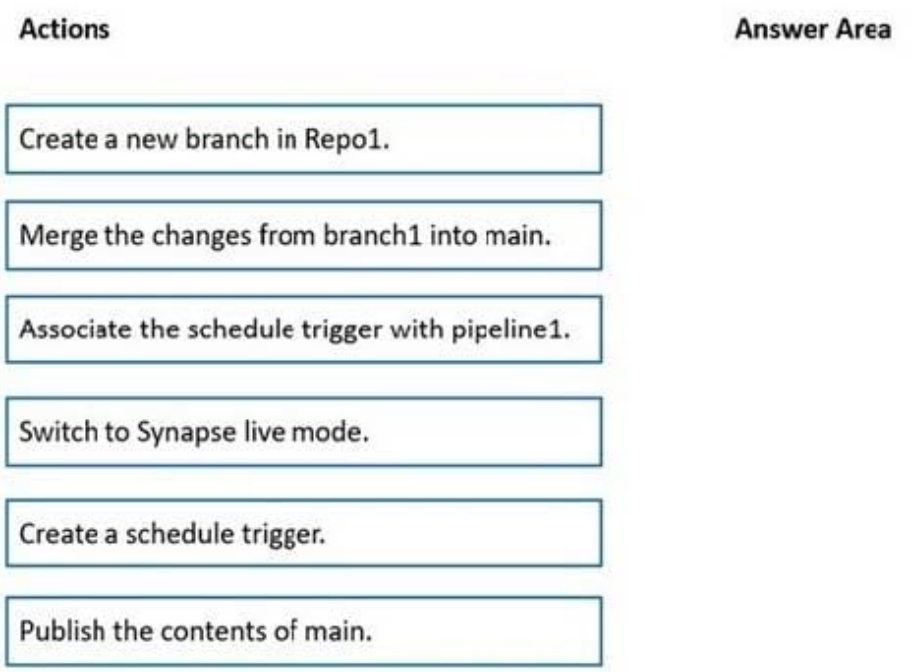
In workspace1, you complete testing of pipeline1.

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

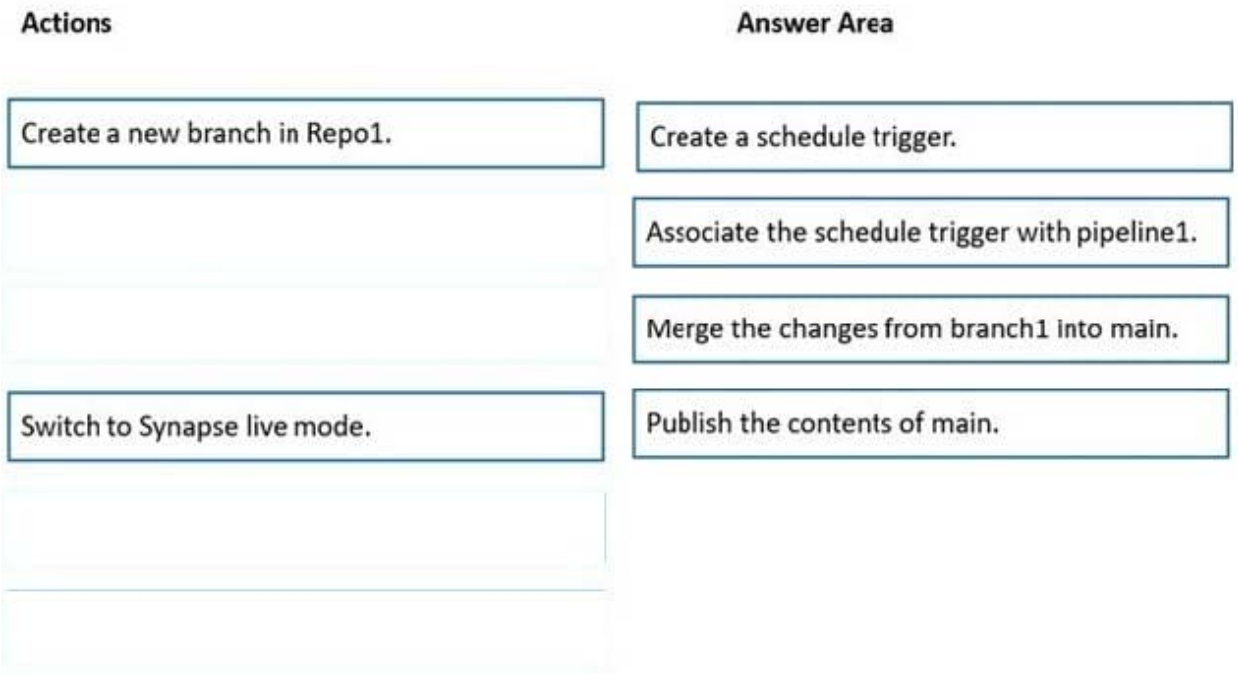
Which four actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

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:



Correct Answer:



QUESTION 12

You are creating an Apache Spark job in Azure Databricks that will ingest JSON-formatted data.

You need to convert a nested JSON string into a DataFrame that will contain multiple rows.

Which Spark SQL function should you use?

A. explode

B. filter

C. coalesce

D. extract

Correct Answer: A

Convert nested JSON to a flattened DataFrame

You can to flatten nested JSON, using only $"column.*" and explode methods.

Note: Extract and flatten

Use $"column.*" and explode methods to flatten the struct and array types before displaying the flattened DataFrame.

Scala

display(DF.select($"id" as "main_id",$"name",$"batters",$"ppu",explode($"topping")) // Exploding the topping column using explode as it is an array type withColumn("topping_id",$"col.id") // Extracting topping_id from col using DOT form

withColumn("topping_type",$"col.type") // Extracting topping_tytpe from col using DOT form drop($"col")

select($"*",$"batters.*") // Flattened the struct type batters tto array type which is batter drop($"batters")
select($"*",explode($"batter"))

drop($"batter")

withColumn("batter_id",$"col.id") // Extracting batter_id from col using DOT form withColumn ("battter_type",$"col.type") // Extracting battter_type from col using DOT form drop($"col"))

Reference: https://learn.microsoft.com/en-us/azure/databricks/kb/scala/flatten-nested- columns- dynamically

---

QUESTION 13

You are designing a streaming data solution that will ingest variable volumes of data. You need to ensure that you can change the partition count after creation.

Which service should you use to ingest the data?

A. Azure Event Hubs Dedicated

B. Azure Stream Analytics

C. Azure Data Factory

D. Azure Synapse Analytics

Correct Answer: A

You can\\'t change the partition count for an event hub after its creation except for the event hub in a dedicated cluster.

Reference: https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features

---

QUESTION 14

DRAG DROP

You have an Azure Data Lake Storage Gen 2 account named storage1.

You need to recommend a solution for accessing the content in storage1. The solution must meet the following requirements:

List and read permissions must be granted at the storage account level. Additional permissions can be applied to individual objects in storage1. Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must

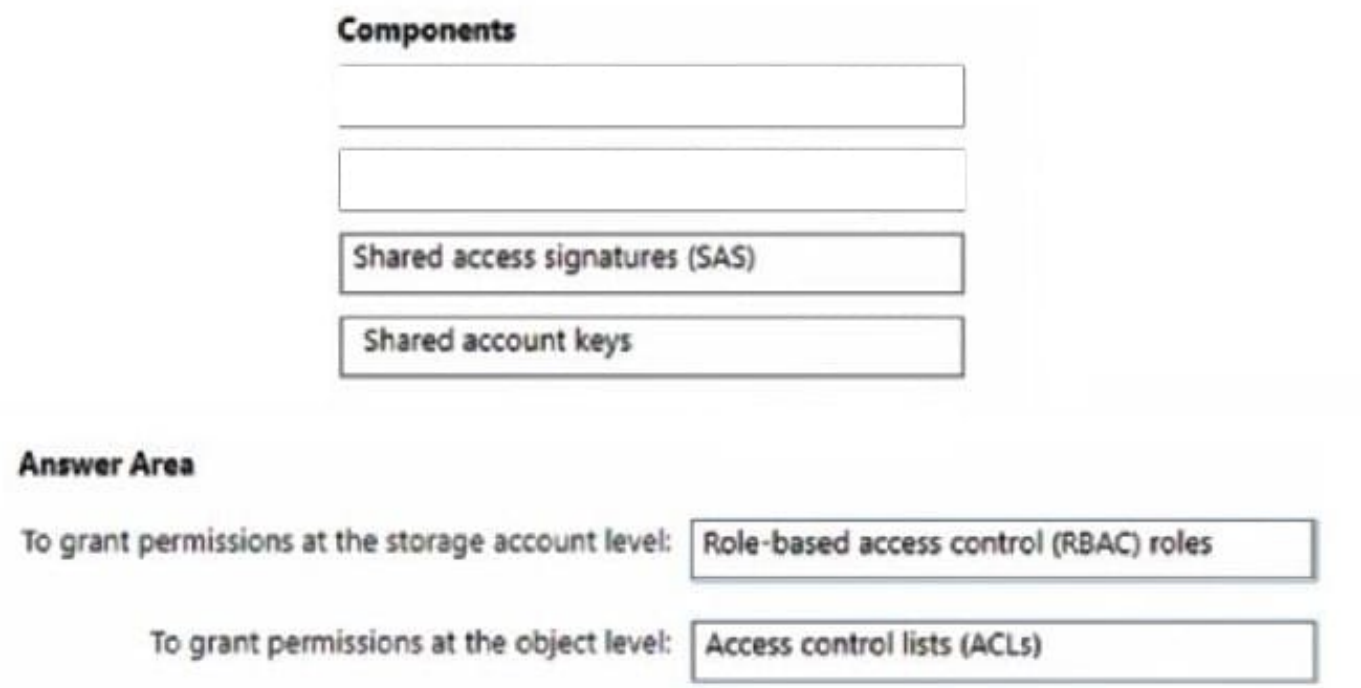
be used for authentication.

What should you use? To answer, drag the appropriate components to the correct requirements. Each component 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.

Select and Place:

Correct Answer:



QUESTION 15

You are designing an Azure Databricks cluster that runs user-defined local processes. You need to recommend a cluster configuration that meets the following requirements:

1.

Minimize query latency.

2.

Maximize the number of users that can run queries on the cluster at the same time.

3.

Reduce overall costs without compromising other requirements. Which cluster type should you recommend?

A. Standard with Auto Termination

B. High Concurrency with Autoscaling

C. High Concurrency with Auto Termination

D. Standard with Autoscaling

Correct Answer: B

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

Databricks chooses the appropriate number of workers required to run your job. This is referred to as autoscaling. Autoscaling makes it easier to achieve high cluster utilization, because you don\\'t need to provision the cluster to match a

workload.

Incorrect Answers:

C: The cluster configuration includes an auto terminate setting whose default value depends on cluster mode:

Standard and Single Node clusters terminate automatically after 120 minutes by default. High Concurrency clusters do not terminate automatically by default.

Reference: https://docs.microsoft.com/en-us/azure/databricks/clusters/configure