# Free Questions for Databricks-Certified-Data-Engineer-Associate

## Shared by Patrick on 27-03-2023

For More Free Questions and Preparation Resources

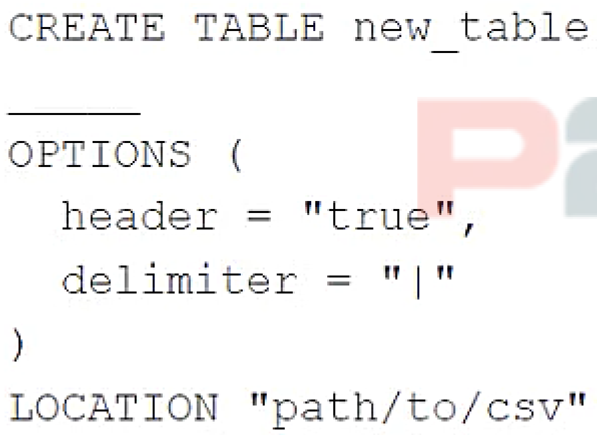
Check the Links on Last Page

# Question 1

Question Type: MultipleChoice

A data engineer needs to create a table in Databricks using data from a CSV file at location /path/to/csv.

They run the following command:



```
CREATE TABLE new_table
______
OPTIONS (
  header = "true",
  delimiter = "|"
)
LOCATION "path/to/csv"
```

Which of the following lines of code fills in the above blank to successfully complete the task?

## Options:

A- None of these lines of code are needed to successfully complete the task

B- USING CSV

C- FROM CSV

D- USING DELTA

E- FROM 'path/to/csv'

## Answer:

E

## Explanation:

A data lakehouse is a new paradigm that can be used to simplify and unify siloed data architectures that are specialized for specific use cases. A data lakehouse combines the best of both data lakes and data warehouses, providing a single platform that supports diverse data types, open standards, low-cost storage, high-performance queries, ACID transactions, schema enforcement, and governance. A data lakehouse enables data engineers to build reliable and scalable data pipelines that can serve various downstream applications and users, such as data science, machine learning, analytics, and reporting. A data lakehouse leverages the power of

Delta Lake, a storage layer that brings reliability and performance to data lakes.Reference:What is a data lakehouse?,Delta Lake,Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics

# Question 2

Question Type: MultipleChoice

A data engineer needs to create a table in Databricks using data from their organization's existing SQLite database. They run the following command:

```
CREATE TABLE jdbc_customer360

USING

OPTIONS (

url "jdbc:sqlite:/customers.db", dbtable "customer360"

)
```

Which line of code fills in the above blank to successfully complete the task?

## Options:

A- autoloader

B- org.apache.spark.sql.jdbc

C- sqlite

D- org.apache.spark.sql.sqlite

## Answer:

B

## Explanation:

To create a table in Databricks using data from an SQLite database, the correct syntax involves specifying the format of the data source. The format in the case of using JDBC (Java Database Connectivity) with SQLite is specified by the org.apache.spark.sql.jdbc format. This format allows Spark to interface with various relational databases through JDBC. Here is how the command should be structured:

```
CREATE TABLE jdbc_customer360
```

```
USING org.apache.spark.sql.jdbc

OPTIONS (

url 'jdbc:sqlite:/customers.db',

dbtable 'customer360'

)
```

The USING org.apache.spark.sql.jdbc line specifies that the JDBC data source is being used, enabling Spark to interact with the SQLite database via JDBC.

Reference: Databricks documentation on JDBC: Connecting to SQL Databases using JDBC

# Question 3

Question Type: MultipleChoice

A data engineer wants to create a new table containing the names of customers who live in France.

They have written the following command:

```
CREATE TABLE customersInFrance

_____ AS

SELECT id,

firstName,

lastName

FROM customerLocations

WHERE country = 'FRANCE';
```

A senior data engineer mentions that it is organization policy to include a table property indicating that the new table includes personally identifiable information (PII).

Which line of code fills in the above blank to successfully complete the task?

## Options:

A- COMMENT 'Contains PIT

B- 511

C- 'COMMENT PII'

D- TBLPROPERTIES PII

## Answer:

D

## Explanation:

To include a property indicating that a table contains personally identifiable information (PII), the TBLPROPERTIES keyword is used in SQL to add metadata to a table. The correct syntax to define a table property for PII is as follows:

CREATE TABLE customersInFrance

USING DELTA

TBLPROPERTIES ('PII' = 'true')

AS

SELECT id,

firstName,

lastName

FROM customerLocations

WHERE country = 'FRANCE';

The TBLPROPERTIES ('PII' = 'true') line correctly sets a table property that tags the table as containing personally identifiable information. This is in accordance with organizational policies for handling sensitive information.

Reference: Databricks documentation on Delta Lake: Delta Lake on Databricks

# Question 4

Question Type: MultipleChoice

A data engineer needs to use a Delta table as part of a data pipeline, but they do not know if they have the appropriate permissions.

In which of the following locations can the data engineer review their permissions on the table?

## Options:

A- Databricks Filesystem

B- Jobs

C- Dashboards

D- Repos

E- Data Explorer

## Answer:

E

## Explanation:

Data Explorer is a graphical interface that allows you to browse, create, and manage data objects such as databases, tables, and views in your workspace. You can also review and modify the permissions on these data objects using Data Explorer. To access Data Explorer, you can click on the Data icon in the sidebar, or use the %sql magic command in a notebook. You can then select a database and a table, and click on the Permissions tab to view and edit the access control lists (ACLs) for the table. You can also use SQL commands such as SHOW GRANT and GRANT to query and modify the permissions on a Delta table.Reference:

Data Explorer

Access control for Delta tables

SHOW GRANT

[GRANT]

# Question 5

Question Type: MultipleChoice

A data engineering team has noticed that their Databricks SQL queries are running too slowly when they are submitted to a non-running SQL endpoint. The data engineering team wants this issue to be resolved.

Which of the following approaches can the team use to reduce the time it takes to return results in this scenario?

## Options:

A- They can turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy to 'Reliability Optimized.'

B- They can turn on the Auto Stop feature for the SQL endpoint.

C- They can increase the cluster size of the SQL endpoint.

D- They can turn on the Serverless feature for the SQL endpoint.

E- They can increase the maximum bound of the SQL endpoint's scaling range

## Answer:

D

## Explanation:

Option D is the correct answer because it enables the Serverless feature for the SQL endpoint, which allows the endpoint to automatically scale up and down based on the query load. This way, the endpoint can handle more concurrent queries and reduce the time it takes to return results. The Serverless feature also reduces the cold start time of the endpoint, which is the time it takes to start the cluster when a query is submitted to a non-running endpoint. The Serverless feature is available for both AWS and Azure Databricks platforms.

# Question 6

Question Type: MultipleChoice

A Delta Live Table pipeline includes two datasets defined using streaming live table. Three datasets are defined against Delta Lake table sources using live table.

The table is configured to run in Production mode using the Continuous Pipeline Mode.

What is the expected outcome after clicking Start to update the pipeline assuming previously unprocessed data exists and all definitions are valid?

## Options:

A- All datasets will be updated once and the pipeline will shut down. The compute resources will be terminated.

B- All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing.

C- All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.

D- All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will be deployed for the update and terminated when the pipeline is stopped.

## Answer:

D

## Explanation:

In Delta Live Tables (DLT), when configured to run in Continuous Pipeline Mode, particularly in a production environment, the system is designed to continuously process and update data as it becomes available. This mode keeps the compute resources active to handle ongoing data processing and automatically updates all datasets defined in the pipeline at predefined intervals. Once the pipeline is manually stopped, the compute resources are terminated to conserve resources and reduce costs. This mode is suitable for production environments where datasets need to be kept up-to-date with the latest data.

Reference: Databricks documentation on Delta Live Tables: Delta Live Tables Guide

# Question 7

Question Type: MultipleChoice

Which of the following is a benefit of the Databricks Lakehouse Platform embracing open source technologies?

## Options:

A- Cloud-specific integrations
B- Simplified governance
C- Ability to scale storage
D- Ability to scale workloads
E- Avoiding vendor lock-in

## Answer:

E

## Explanation:

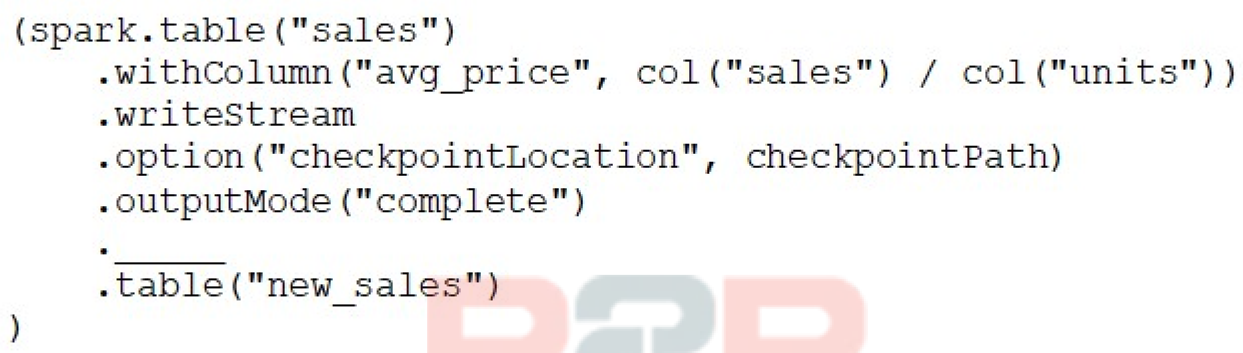
One of the benefits of the Databricks Lakehouse Platform embracing open source technologies is that it avoids vendor lock-in. This means that customers can use the same open source tools and frameworks across different cloud providers, and migrate their data and workloads without being tied to a specific vendor. The Databricks Lakehouse Platform is built on open source projects such as Apache Spark, Delta Lake, MLflow, and Redash, which are widely used and trusted by millions of developers. By supporting these open source technologies, the Databricks Lakehouse Platform enables customers to leverage the innovation and community of the open source ecosystem, and avoid the risk of being locked into proprietary or closed solutions. The other options are either not related to open source technologies (A, B, C, D), or not benefits of the Databricks Lakehouse Platform (A, B).Reference:Databricks Documentation - Built on open source,Databricks Documentation - What is the Lakehouse Platform?,Databricks Blog - Introducing the Databricks Lakehouse Platform.

# Question 8

Question Type: MultipleChoice

A data engineer has configured a Structured Streaming job to read from a table, manipulate the data, and then perform a streaming write into a new table.

The cade block used by the data engineer is below:



```
(spark.table("sales")
    .withColumn("avg_price", col("sales") / col("units"))
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("complete")
    .______
    .table("new_sales")
)
```

If the data engineer only wants the query to execute a micro-batch to process data every 5 seconds, which of the following lines of code should the data engineer use to fill in the blank?

## Options:

A- trigger('5 seconds')

B- trigger()

C- trigger(once='5 seconds')

D- trigger(processingTime='5 seconds')

E- trigger(continuous='5 seconds')

## Answer:

D

## Explanation:

The processingTime option specifies a time-based trigger interval for fixed interval micro-batches. This means that the query will execute a micro-batch to process data every 5 seconds, regardless of how much data is available. This option is suitable for near-real time processing workloads that require low latency and consistent processing frequency. The other options are either invalid syntax (A, C), default behavior (B), or experimental feature (E).Reference:Databricks Documentation - Configure Structured Streaming trigger intervals,Databricks Documentation - Trigger.