

# Databricks

## Exam Questions Databricks-Certified-Data-Engineer-Associate

Databricks Certified Data Engineer Associate Exam



**NEW QUESTION 1**

Which of the following commands will return the location of database customer360?

- A. DESCRIBE LOCATION customer360;
- B. DROP DATABASE customer360;
- C. DESCRIBE DATABASE customer360;
- D. ALTER DATABASE customer360 SET DBPROPERTIES ('location' = '/user');
- E. USE DATABASE customer360;

**Answer:** C

**Explanation:**

To retrieve the location of a database named "customer360" in a database management system like Hive or Databricks, you can use the DESCRIBE DATABASE command followed by the database name. This command will provide information about the database, including its location.

**NEW QUESTION 2**

Which of the following approaches should be used to send the Databricks Job owner an email in the case that the Job fails?

- A. Manually programming in an alert system in each cell of the Notebook
- B. Setting up an Alert in the Job page
- C. Setting up an Alert in the Notebook
- D. There is no way to notify the Job owner in the case of Job failure
- E. MLflow Model Registry Webhooks

**Answer:** B

**Explanation:**

<https://docs.databricks.com/en/workflows/jobs/job-notifications.html>

**NEW QUESTION 3**

A data engineering team has two tables. The first table march\_transactions is a collection of all retail transactions in the month of March. The second table april\_transactions is a collection of all retail transactions in the month of April. There are no duplicate records between the tables.

Which of the following commands should be run to create a new table all\_transactions that contains all records from march\_transactions and april\_transactions without duplicate records?

- A. CREATE TABLE all\_transactions AS SELECT \* FROM march\_transactions INNER JOIN SELECT \* FROM april\_transactions;
- B. CREATE TABLE all\_transactions AS SELECT \* FROM march\_transactions UNION SELECT \* FROM april\_transactions;
- C. CREATE TABLE all\_transactions AS SELECT \* FROM march\_transactions OUTER JOIN SELECT \* FROM april\_transactions;
- D. CREATE TABLE all\_transactions AS SELECT \* FROM march\_transactions INTERSECT SELECT \* FROM april\_transactions;
- E. CREATE TABLE all\_transactions AS SELECT \* FROM march\_transactions MERGE SELECT \* FROM april\_transactions;

**Answer:** B

**Explanation:**

To create a new table all\_transactions that contains all records from march\_transactions and april\_transactions without duplicate records, you should use the UNION operator, as shown in option B. This operator combines the result sets of the two tables while automatically removing duplicate records.

**NEW QUESTION 4**

A data analysis team has noticed that their Databricks SQL queries are running too slowly when connected to their always-on SQL endpoint. They claim that this issue is present when many members of the team are running small queries simultaneously. They ask the data engineering team for help. The data engineering team notices that each of the team's queries uses the same SQL endpoint.

Which of the following approaches can the data engineering team use to improve the latency of the team's queries?

- A. They can increase the cluster size of the SQL endpoint.
- B. They can increase the maximum bound of the SQL endpoint's scaling range.
- C. They can turn on the Auto Stop feature for the SQL endpoint.
- D. They can turn on the Serverless feature for the SQL endpoint.
- E. They can turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy to "Reliability Optimized."

**Answer:** A

**Explanation:**

When many users are running small queries simultaneously on a SQL endpoint, the database can become overloaded, causing slow query execution times. By increasing the cluster size of the SQL endpoint, the database can handle more simultaneous queries, resulting in faster query execution times.

**NEW QUESTION 5**

A data engineer wants to schedule their Databricks SQL dashboard to refresh every hour, but they only want the associated SQL endpoint to be running when it is necessary. The dashboard has multiple queries on multiple datasets associated with it. The data that feeds the dashboard is automatically processed using a Databricks Job.

Which of the following approaches can the data engineer use to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard?

- A. They can turn on the Auto Stop feature for the SQL endpoint.
- B. They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.
- C. They can reduce the cluster size of the SQL endpoint.
- D. They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.
- E. They can set up the dashboard's SQL endpoint to be serverless.

**Answer:** A

#### NEW QUESTION 6

A data engineer only wants to execute the final block of a Python program if the Python variable `day_of_week` is equal to 1 and the Python variable `review_period` is True.

Which of the following control flow statements should the data engineer use to begin this conditionally executed code block?

- A. if `day_of_week = 1` and `review_period`:
- B. if `day_of_week = 1` and `review_period = "True"`:
- C. if `day_of_week == 1` and `review_period == "True"`:
- D. if `day_of_week == 1` and `review_period`:
- E. if `day_of_week = 1 & review_period: = "True"`:

**Answer:** D

#### Explanation:

This statement will check if the variable `day_of_week` is equal to 1 and if the variable `review_period` evaluates to a truthy value. The use of the double equal sign (`==`) in the comparison of `day_of_week` is important, as a single equal sign (`=`) would be used to assign a value to the variable instead of checking its value. The use of a single ampersand (`&`) instead of the keyword `and` is not valid syntax in Python. The use of quotes around `True` in options B and C will result in a string comparison, which will not evaluate to `True` even if the value of `review_period` is `True`.

#### NEW QUESTION 7

Which of the following benefits is provided by the array functions from Spark SQL?

- A. An ability to work with data in a variety of types at once
- B. An ability to work with data within certain partitions and windows
- C. An ability to work with time-related data in specified intervals
- D. An ability to work with complex, nested data ingested from JSON files
- E. An ability to work with an array of tables for procedural automation

**Answer:** D

#### Explanation:

Array functions in Spark SQL are primarily used for working with arrays and complex, nested data structures, such as those often encountered when ingesting JSON files. These functions allow you to manipulate and query nested arrays and structures within your data, making it easier to extract and work with specific elements or values within complex data formats. While some of the other options (such as option A for working with different data types) are features of Spark SQL or SQL in general, array functions specifically excel at handling complex, nested data structures like those found in JSON files.

#### NEW QUESTION 8

Which of the following is hosted completely in the control plane of the classic Databricks architecture?

- A. Worker node
- B. JDBC data source
- C. Databricks web application
- D. Databricks Filesystem
- E. Driver node

**Answer:** C

#### Explanation:

In the classic Databricks architecture, the control plane includes components like the Databricks web application, the Databricks REST API, and the Databricks Workspace. These components are responsible for managing and controlling the Databricks environment, including cluster provisioning, notebook management, access control, and job scheduling. The other options, such as worker nodes, JDBC data sources, Databricks Filesystem (DBFS), and driver nodes, are typically part of the data plane or the execution environment, which is separate from the control plane. Worker nodes are responsible for executing tasks and computations, JDBC data sources are used to connect to external databases, DBFS is a distributed file system for data storage, and driver nodes are responsible for coordinating the execution of Spark jobs.

#### NEW QUESTION 9

A dataset has been defined using Delta Live Tables and includes an expectations clause:

```
CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION DROP ROW
```

What is the expected behavior when a batch of data containing data that violates these constraints is processed?

- A. Records that violate the expectation are dropped from the target dataset and loaded into a quarantine table.
- B. Records that violate the expectation are added to the target dataset and flagged as invalid in a field added to the target dataset.
- C. Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.
- D. Records that violate the expectation are added to the target dataset and recorded as invalid in the event log.
- E. Records that violate the expectation cause the job to fail.

**Answer:** C

#### Explanation:

With the defined constraint and expectation clause, when a batch of data is processed, any records that violate the expectation (in this case, where the timestamp is not greater than '2020-01-01') will be dropped from the target dataset. These dropped records will also be recorded as invalid in the event log, allowing for auditing and tracking of the data quality issues without causing the entire job to fail. <https://docs.databricks.com/en/delta-live-tables/expectations.html>

#### NEW QUESTION 10

Which of the following Git operations must be performed outside of Databricks Repos?

- A. Commit
- B. Pull
- C. Push
- D. Clone
- E. Merge

**Answer:** E

**Explanation:**

For following tasks, work in your Git provider:  
Create a pull request. Resolve merge conflicts. Merge or delete branches. Rebase a branch.  
<https://docs.databricks.com/repos/index.html>

**NEW QUESTION 10**

A data engineer has a Python variable `table_name` that they would like to use in a SQL query. They want to construct a Python code block that will run the query using `table_name`.

They have the following incomplete code block:

```
(f"SELECT customer_id, spend FROM {table_name}")
```

Which of the following can be used to fill in the blank to successfully complete the task?

- A. `spark.delta.sql`
- B. `spark.delta.table`
- C. `spark.table`
- D. `dbutils.sql`
- E. `spark.sql`

**Answer:** E

**NEW QUESTION 14**

A data engineer has been given a new record of data:

`id STRING = 'a1'`

`rank INTEGER = 6 rating FLOAT = 9.4`

Which of the following SQL commands can be used to append the new record to an existing Delta table `my_table`?

- A. `INSERT INTO my_table VALUES ('a1', 6, 9.4)`
- B. `my_table UNION VALUES ('a1', 6, 9.4)`
- C. `INSERT VALUES ( 'a1' , 6, 9.4) INTO my_table`
- D. `UPDATE my_table VALUES ('a1', 6, 9.4)`
- E. `UPDATE VALUES ('a1', 6, 9.4) my_table`

**Answer:** A

**NEW QUESTION 15**

A data engineer and data analyst are working together on a data pipeline. The data engineer is working on the raw, bronze, and silver layers of the pipeline using Python, and the data analyst is working on the gold layer of the pipeline using SQL. The raw source of the pipeline is a streaming input. They now want to migrate their pipeline to use Delta Live Tables.

Which of the following changes will need to be made to the pipeline when migrating to Delta Live Tables?

- A. None of these changes will need to be made
- B. The pipeline will need to stop using the medallion-based multi-hop architecture
- C. The pipeline will need to be written entirely in SQL
- D. The pipeline will need to use a batch source in place of a streaming source
- E. The pipeline will need to be written entirely in Python

**Answer:** A

**NEW QUESTION 18**

A data engineer that is new to using Python needs to create a Python function to add two integers together and return the sum?

Which of the following code blocks can the data engineer use to complete this task?

A)

```
function add_integers(x, y):  
    return x + y
```

B)

```
function add_integers(x, y):  
    x + y
```

C)

```
def add_integers(x, y):  
    print(x + y)
```

D)

```
def add_integers(x, y):  
    return x + y
```

E)

```
def add_integers(x, y):  
    x + y
```

- A. Option A
- B. Option B
- C. Option C
- D. Option D
- E. Option E

**Answer:** D

**Explanation:**

[https://www.w3schools.com/python/python\\_functions.asp](https://www.w3schools.com/python/python_functions.asp)

#### NEW QUESTION 19

Which of the following code blocks will remove the rows where the value in column age is greater than 25 from the existing Delta table my\_table and save the updated table?

- A. SELECT \* FROM my\_table WHERE age > 25;
- B. UPDATE my\_table WHERE age > 25;
- C. DELETE FROM my\_table WHERE age > 25;
- D. UPDATE my\_table WHERE age <= 25;
- E. DELETE FROM my\_table WHERE age <= 25;

**Answer:** C

#### NEW QUESTION 21

A data engineer has developed a data pipeline to ingest data from a JSON source using Auto Loader, but the engineer has not provided any type inference or schema hints in their pipeline. Upon reviewing the data, the data engineer has noticed that all of the columns in the target table are of the string type despite some of the fields only including float or boolean values.

Which of the following describes why Auto Loader inferred all of the columns to be of the string type?

- A. There was a type mismatch between the specific schema and the inferred schema
- B. JSON data is a text-based format
- C. Auto Loader only works with string data
- D. All of the fields had at least one null value
- E. Auto Loader cannot infer the schema of ingested data

**Answer:** B

**Explanation:**

JSON data is a text-based format that uses strings to represent all values. When Auto Loader infers the schema of JSON data, it assumes that all values are strings. This is because Auto Loader cannot determine the type of a value based on its string representation. <https://docs.databricks.com/en/ingestion/auto-loader/schema.html> Forexample, the following JSON string represents a value that is logically a boolean: JSON "true" Use code with caution. Learn more However, Auto Loader would infer that the type of this value is string. This is because Auto Loader cannot determine that the value is a boolean based on its string representation. In order to get Auto Loader to infer the correct types for columns, the data engineer can provide type inference or schema hints. Type inference hints can be used to specify the types of specific columns. Schema hints can be used to provide the entire schema of the data. Therefore, the correct answer is B. JSON data is a text-based format.

#### NEW QUESTION 24

Which of the following Structured Streaming queries is performing a hop from a Silver table to a Gold table?

A.

```
(spark.readStream.load(rawSalesLocation)  
    .writeStream  
    .option("checkpointLocation", checkpointPath)  
    .outputMode("append")  
    .table("newSales")  
)
```

B.

```
(spark.read.load(rawSalesLocation)  
    .writeStream  
    .option("checkpointLocation", checkpointPath)  
    .outputMode("append")  
    .table("newSales")  
)
```

C.



```
(spark.table("sales")
  .withColumn("avgPrice", col("sales") / col("units"))
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("append")
  .table("newSales")
)
```

D.

```
(spark.table("sales")
  .filter(col("units") > 0)
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("append")
  .table("newSales")
)
```

E.

```
(spark.table("sales")
  .groupBy("store")
  .agg(sum("sales"))
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("complete")
  .table("newSales")
)
```

A.

**Answer: E**

#### NEW QUESTION 29

.....

## Thank You for Trying Our Product

### We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questions and Answers in PDF Format

### Databricks-Certified-Data-Engineer-Associate Practice Exam Features:

- \* Databricks-Certified-Data-Engineer-Associate Questions and Answers Updated Frequently
- \* Databricks-Certified-Data-Engineer-Associate Practice Questions Verified by Expert Senior Certified Staff
- \* Databricks-Certified-Data-Engineer-Associate Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- \* Databricks-Certified-Data-Engineer-Associate Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

**100% Actual & Verified — Instant Download, Please Click**  
**[Order The Databricks-Certified-Data-Engineer-Associate Practice Test Here](#)**