



**Vendor:** Databricks

**Exam Code:** Databricks-Certified-Data-Engineer-Associate

**Exam Name:** Databricks Certified Data Engineer Associate

**Version:** DEMO

### QUESTION 1

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

- A. org.apache.spark.sql.jdbc
- B. autoloader
- C. DELTA
- D. sqlite
- E. org.apache.spark.sql.sqlite

**Answer: A**

### QUESTION 2

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**

### QUESTION 3

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**

#### QUESTION 4

A data engineer is attempting to drop a Spark SQL table my\_table. The data engineer wants to delete all table metadata and data.  
They run the following command:

```
DROP TABLE IF EXISTS my_table
```

While the object no longer appears when they run SHOW TABLES, the data files still exist.  
Which of the following describes why the data files still exist and the metadata files were deleted?

- A. The table's data was larger than 10 GB
- B. The table's data was smaller than 10 GB
- C. The table was external
- D. The table did not have a location
- E. The table was managed

**Answer: C**

#### QUESTION 5

A data engineer wants to create a data entity from a couple of tables. The data entity must be used by other data engineers in other sessions. It also must be saved to a physical location.  
Which of the following data entities should the data engineer create?

- A. Database
- B. Function
- C. View
- D. Temporary view
- E. Table

**Answer: E**

#### QUESTION 6

A data engineer is maintaining a data pipeline. Upon data ingestion, the data engineer notices that the source data is starting to have a lower level of quality. The data engineer would like to automate the process of monitoring the quality level.  
Which of the following tools can the data engineer use to solve this problem?

- A. Unity Catalog

- B. Data Explorer
- C. Delta Lake
- D. Delta Live Tables
- E. Auto Loader

**Answer: D**

#### QUESTION 7

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. Assuming previously unprocessed data exists and all definitions are valid, what is the expected outcome after clicking Start to update the pipeline?

- A. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing.
- B. All datasets will be updated once and the pipeline will persist without any processing. The compute resources will persist but go unused.
- C. 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.
- D. All datasets will be updated once and the pipeline will shut down. The compute resources will be terminated.
- E. All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.

**Answer: C**

#### QUESTION 8

In order for Structured Streaming to reliably track the exact progress of the processing so that it can handle any kind of failure by restarting and/or reprocessing, which of the following two approaches is used by Spark to record the offset range of the data being processed in each trigger?

- A. Checkpointing and Write-ahead Logs
- B. Structured Streaming cannot record the offset range of the data being processed in each trigger.
- C. Replayable Sources and Idempotent Sinks
- D. Write-ahead Logs and Idempotent Sinks
- E. Checkpointing and Idempotent Sinks

**Answer: A**

#### QUESTION 9

Which of the following describes the relationship between Gold tables and Silver tables?

- A. Gold tables are more likely to contain aggregations than Silver tables.
- B. Gold tables are more likely to contain valuable data than Silver tables.
- C. Gold tables are more likely to contain a less refined view of data than Silver tables.
- D. Gold tables are more likely to contain more data than Silver tables.
- E. Gold tables are more likely to contain truthful data than Silver tables.

**Answer: A**

**QUESTION 10**

Which of the following describes the relationship between Bronze tables and raw data?

- A. Bronze tables contain less data than raw data files.
- B. Bronze tables contain more truthful data than raw data.
- C. Bronze tables contain aggregates while raw data is unaggregated.
- D. Bronze tables contain a less refined view of data than raw data.
- E. Bronze tables contain raw data with a schema applied.

**Answer: E**

**QUESTION 11**

Which of the following tools is used by Auto Loader process data incrementally?

- A. Checkpointing
- B. Spark Structured Streaming
- C. Data Explorer
- D. Unity Catalog
- E. Databricks SQL

**Answer: B**

**QUESTION 12**

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 code block used by the data engineer is below:

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

- A. trigger("5 seconds")
- B. trigger()
- C. trigger(once="5 seconds")
- D. trigger(processingTime="5 seconds")
- E. trigger(continuous="5 seconds")

**Answer: D**

**QUESTION 13**

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**

**QUESTION 14**

Which of the following describes when to use the CREATE STREAMING LIVE TABLE (formerly CREATE INCREMENTAL LIVE TABLE) syntax over the CREATE LIVE TABLE syntax when creating Delta Live Tables (DLT) tables using SQL?

- A. CREATE STREAMING LIVE TABLE should be used when the subsequent step in the DLT pipeline is static.
- B. CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally.
- C. CREATE STREAMING LIVE TABLE is redundant for DLT and it does not need to be used.
- D. CREATE STREAMING LIVE TABLE should be used when data needs to be processed through complicated aggregations.
- E. CREATE STREAMING LIVE TABLE should be used when the previous step in the DLT pipeline is static.

**Answer: B**

## Thank You for Trying Our Product

### Lead2pass Certification Exam Features:

- ★ More than **99,900** Satisfied Customers Worldwide.
- ★ Average **99.9%** Success Rate.
- ★ **Free Update** to match latest and real exam scenarios.
- ★ **Instant Download** Access! No Setup required.
- ★ Questions & Answers are downloadable in **PDF** format and **VCE** test engine format.
- ★ Multi-Platform capabilities - **Windows, Laptop, Mac, Android, iPhone, iPod, iPad**.
- ★ **100%** Guaranteed Success or **100%** Money Back Guarantee.
- ★ **Fast**, helpful support **24x7**.



View list of all certification exams: <http://www.lead2pass.com/all-products.html>



Microsoft



ORACLE



CITRIX



JUNIPER  
NETWORKS



EMC<sup>2</sup>  
where information lives

**10% Discount Coupon Code: ASTR14**

