

Exam Questions Databricks-Certified-Data-Engineer-Associate

Databricks Certified Data Engineer Associate Exam

<https://www.2passeasy.com/dumps/Databricks-Certified-Data-Engineer-Associate/>



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

A data engineer has created a new database using the following command: CREATE DATABASE IF NOT EXISTS customer360;
In which of the following locations will the customer360 database be located?

- A. dbfs:/user/hive/database/customer360
- B. dbfs:/user/hive/warehouse
- C. dbfs:/user/hive/customer360
- D. More information is needed to determine the correct response

Answer: B

Explanation:

dbfs:/user/hive/warehouse - which is the default location

NEW QUESTION 3

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

Explanation:

The CREATE STREAMING LIVE TABLE syntax is used when you want to create Delta Live Tables (DLT) tables that are designed for processing data incrementally. This is typically used when your data pipeline involves streaming or incremental data updates, and you want the table to stay up to date as new data arrives. It allows you to define tables that can handle data changes incrementally without the need for full table refreshes.

NEW QUESTION 4

A data analyst has developed a query that runs against Delta table. They want help from the data engineering team to implement a series of tests to ensure the data returned by the query is clean. However, the data engineering team uses Python for its tests rather than SQL.

Which of the following operations could the data engineering team use to run the query and operate with the results in PySpark?

- A. SELECT * FROM sales
- B. spark.delta.table
- C. spark.sql
- D. There is no way to share data between PySpark and SQL.
- E. spark.table

Answer: C

Explanation:

```
from pyspark.sql import SparkSession spark = SparkSession.builder.getOrCreate()  
df = spark.sql("SELECT * FROM sales") print(df.count())
```

NEW QUESTION 5

A data engineer is attempting to drop a Spark SQL table my_table and runs the following command:

```
DROP TABLE IF EXISTS my_table;
```

After running this command, the engineer notices that the data files and metadata files have been deleted from the file system.

Which of the following describes why all of these files were deleted?

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

Answer: A

Explanation:

managed tables files and metadata are managed by metastore and will be deleted when the table is dropped. While external tables the metadata is stored in an external location. Hence when an external table is dropped you clear off only the metadata and the files (data) remain.

NEW QUESTION 6

Which of the following data lakehouse features results in improved data quality over a traditional data lake?

- A. A data lakehouse provides storage solutions for structured and unstructured data.
- B. A data lakehouse supports ACID-compliant transactions.
- C. A data lakehouse allows the use of SQL queries to examine data.
- D. A data lakehouse stores data in open formats.
- E. A data lakehouse enables machine learning and artificial intelligence workloads.

Answer: B

Explanation:

One of the key features of a data lakehouse that results in improved data quality over a traditional data lake is its support for ACID (Atomicity, Consistency, Isolation, Durability) transactions. ACID transactions provide data integrity and consistency guarantees, ensuring that operations on the data are reliable and that data is not left in an inconsistent state due to failures or concurrent access. In a traditional data lake, such transactional guarantees are often lacking, making it challenging to maintain data quality, especially in scenarios involving multiple data writes, updates, or complex transformations. A data lakehouse, by offering ACID compliance, helps maintain data quality by providing strong consistency and reliability, which is crucial for data pipelines and analytics.

NEW QUESTION 7

Which of the following statements regarding the relationship between Silver tables and Bronze tables is always true?

- A. Silver tables contain a less refined, less clean view of data than Bronze data.
- B. Silver tables contain aggregates while Bronze data is unaggregated.
- C. Silver tables contain more data than Bronze tables.
- D. Silver tables contain a more refined and cleaner view of data than Bronze tables.
- E. Silver tables contain less data than Bronze tables.

Answer: D

Explanation:

<https://www.databricks.com/glossary/medallion-architecture>

NEW QUESTION 8

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

Explanation:

The Bronze layer is where we land all the data from external source systems. The table structures in this layer correspond to the source system table structures "as-is," along with any additional metadata columns that capture the load date/time, process ID, etc. The focus in this layer is quick Change Data Capture and the ability to provide an historical archive of source (cold storage), data lineage, auditability, reprocessing if needed without rereading the data from the source system. <https://www.databricks.com/glossary/medallion-architecture#:~:text=Bronze%20layer%20%28raw%20data%29>

NEW QUESTION 9

A data engineer wants to create a new table containing the names of customers that 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 of the following lines of code fills in the above blank to successfully complete the task?

- A. There is no way to indicate whether a table contains PII.
- B. "COMMENT PII"
- C. TBLPROPERTIES PII

- D. COMMENT "Contains PII"
- E. PII

Answer: D

Explanation:

Ref: <https://www.databricks.com/discover/pages/data-quality-management> CREATE TABLE my_table (id INT COMMENT 'Unique Identification Number', name STRING COMMENT 'PII', age INT COMMENT 'PII') TBLPROPERTIES ('contains_pii'=True) COMMENT 'Contains PII';

NEW QUESTION 10

Which of the following benefits of using the Databricks Lakehouse Platform is provided by Delta Lake?

- A. The ability to manipulate the same data using a variety of languages
- B. The ability to collaborate in real time on a single notebook
- C. The ability to set up alerts for query failures
- D. The ability to support batch and streaming workloads
- E. The ability to distribute complex data operations

Answer: D

Explanation:

Delta Lake is a key component of the Databricks Lakehouse Platform that provides several benefits, and one of the most significant benefits is its ability to support both batch and streaming workloads seamlessly. Delta Lake allows you to process and analyze data in real-time (streaming) as well as in batch, making it a versatile choice for various data processing needs. While the other options may be benefits or capabilities of Databricks or the Lakehouse Platform in general, they are not specifically associated with Delta Lake.

NEW QUESTION 10

Which of the following commands can be used to write data into a Delta table while avoiding the writing of duplicate records?

- A. DROP
- B. IGNORE
- C. MERGE
- D. APPEND
- E. INSERT

Answer: C

Explanation:

To write data into a Delta table while avoiding the writing of duplicate records, you can use the MERGE command. The MERGE command in Delta Lake allows you to combine the ability to insert new records and update existing records in a single atomic operation. The MERGE command compares the data being written with the existing data in the Delta table based on specified matching criteria, typically using a primary key or unique identifier. It then performs conditional actions, such as inserting new records or updating existing records, depending on the comparison results. By using the MERGE command, you can handle the prevention of duplicate records in a more controlled and efficient manner. It allows you to synchronize and reconcile data from different sources while avoiding duplication and ensuring data integrity.

NEW QUESTION 11

A data engineer has a Python notebook in Databricks, but they need to use SQL to accomplish a specific task within a cell. They still want all of the other cells to use Python without making any changes to those cells.

Which of the following describes how the data engineer can use SQL within a cell of their Python notebook?

- A. It is not possible to use SQL in a Python notebook
- B. They can attach the cell to a SQL endpoint rather than a Databricks cluster
- C. They can simply write SQL syntax in the cell
- D. They can add %sql to the first line of the cell
- E. They can change the default language of the notebook to SQL

Answer: D

NEW QUESTION 15

A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to an ELT job. The ELT job has its Databricks SQL query that returns the number of input records containing unexpected NULL values. The data engineer wants their entire team to be notified via a messaging webhook whenever this value reaches 100.

Which of the following approaches can the data engineer use to notify their entire team via a messaging webhook whenever the number of NULL values reaches 100?

- A. They can set up an Alert with a custom template.
- B. They can set up an Alert with a new email alert destination.
- C. They can set up an Alert with a new webhook alert destination.
- D. They can set up an Alert with one-time notifications.
- E. They can set up an Alert without notifications.

Answer: C

Explanation:

To achieve this, the data engineer can set up an Alert in the Databricks workspace that triggers when the query results exceed the threshold of 100 NULL values. They can create a new webhook alert destination in the Alert's configuration settings and provide the necessary messaging webhook URL to receive notifications. When the Alert is triggered, it will send a message to the configured webhook URL, which will then notify the entire team of the issue.

NEW QUESTION 18

Which of the following describes a scenario in which a data engineer will want to use a single-node cluster?

- A. When they are working interactively with a small amount of data
- B. When they are running automated reports to be refreshed as quickly as possible
- C. When they are working with SQL within Databricks SQL
- D. When they are concerned about the ability to automatically scale with larger data
- E. When they are manually running reports with a large amount of data

Answer: A

Explanation:

A Single Node cluster is a cluster consisting of an Apache Spark driver and no Spark workers. A Single Node cluster supports Spark jobs and all Spark data sources, including Delta Lake. A Standard cluster requires a minimum of one Spark worker to run Spark jobs.

NEW QUESTION 19

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?

- 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: B

NEW QUESTION 21

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:

```
("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 24

A data engineer has left the organization. The data team needs to transfer ownership of the data engineer's Delta tables to a new data engineer. The new data engineer is the lead engineer on the data team.

Assuming the original data engineer no longer has access, which of the following individuals must be the one to transfer ownership of the Delta tables in Data Explorer?

- A. Databricks account representative
- B. This transfer is not possible
- C. Workspace administrator
- D. New lead data engineer
- E. Original data engineer

Answer: C

Explanation:

<https://docs.databricks.com/sql/admin/transfer-ownership.html>

NEW QUESTION 29

An engineering manager uses a Databricks SQL query to monitor ingestion latency for each data source. The manager checks the results of the query every day, but they are manually rerunning the query each day and waiting for the results.

Which of the following approaches can the manager use to ensure the results of the query are updated each day?

- A. They can schedule the query to refresh every 1 day from the SQL endpoint's page in Databricks SQL.
- B. They can schedule the query to refresh every 12 hours from the SQL endpoint's page in Databricks SQL.
- C. They can schedule the query to refresh every 1 day from the query's page in Databricks SQL.
- D. They can schedule the query to run every 1 day from the Jobs UI.
- E. They can schedule the query to run every 12 hours from the Jobs UI.

Answer: C

NEW QUESTION 30

A data engineer wants to create a relational object by pulling data from two tables. The relational object does not need to be used by other data engineers in other sessions. In order to save on storage costs, the data engineer wants to avoid copying and storing physical data. Which of the following relational objects should the data engineer create?

- A. Spark SQL Table
- B. View
- C. Database
- D. Temporary view
- E. Delta Table

Answer: D

Explanation:

Temp view : session based Create temp view view_name as query All these are termed as session ended: Opening a new notebook Detaching and reattaching a cluster Installing a python package Restarting a cluster

NEW QUESTION 32

In which of the following file formats is data from Delta Lake tables primarily stored?

- A. Delta
- B. CSV
- C. Parquet
- D. JSON
- E. A proprietary, optimized format specific to Databricks

Answer: C

Explanation:

<https://docs.delta.io/latest/delta-faq.html>

NEW QUESTION 35

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.readStream
  .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 process all of the available data in as many batches as required, which of the following lines of code should the data engineer use to fill in the blank?

- A. processingTime(1)
- B. trigger(availableNow=True)
- C. trigger(parallelBatch=True)
- D. trigger(processingTime="once")
- E. trigger(continuous="once")

Answer: B

Explanation:

<https://stackoverflow.com/questions/71061809/trigger-availablenow-for-delta-source-streaming-queries-in-pyspark-databricks>

NEW QUESTION 38

.....

THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual Databricks-Certified-Data-Engineer-Associate Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the Databricks-Certified-Data-Engineer-Associate Product From:

<https://www.2passeasy.com/dumps/Databricks-Certified-Data-Engineer-Associate/>

Money Back Guarantee

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