

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

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

Explanation:

```
CREATE TABLE new_employees_table USING JDBC
OPTIONS (
  url "<jdbc_url>",
  dbtable "<table_name>", user '<username>', password '<password>'
) AS
SELECT * FROM employees_table_vw https://docs.databricks.com/external-data/jdbc.html#language-sql
```

NEW QUESTION 3

A data organization leader is upset about the data analysis team's reports being different from the data engineering team's reports. The leader believes the siloed nature of their organization's data engineering and data analysis architectures is to blame.

Which of the following describes how a data lakehouse could alleviate this issue?

- A. Both teams would autoscale their work as data size evolves
- B. Both teams would use the same source of truth for their work
- C. Both teams would reorganize to report to the same department
- D. Both teams would be able to collaborate on projects in real-time
- E. Both teams would respond more quickly to ad-hoc requests

Answer: B

Explanation:

A data lakehouse is designed to unify the data engineering and data analysis architectures by integrating features of both data lakes and data warehouses. One of the key benefits of a data lakehouse is that it provides a common, centralized data repository (the "lake") that serves as a single source of truth for data storage and analysis. This allows both data engineering and data analysis teams to work with the same consistent data sets, reducing discrepancies and ensuring that the reports generated by both teams are based on the same underlying data.

NEW QUESTION 4

A data engineer is working with two tables. Each of these tables is displayed below in its entirety.

sales

customer_id	spend	units
a1	28.94	7
a3	874.12	23
a4	8.99	1

favorite_stores

customer_id	store_id
a1	s1
a2	s1
a4	s2

The data engineer runs the following query to join these tables together:

```
SELECT
    sales.customer_id,
    sales.spend,
    favorite_stores.store_id
FROM sales
LEFT JOIN favorite_stores
ON sales.customer_id = favorite_stores.customer_id;
```

Which of the following will be returned by the above query?

A.

customer_id	spend	store_id
a1	28.94	s1
a4	8.99	s2

B.

customer_id	spend	units	store_id
a1	28.94	7	s1
a4	8.99	1	s2

C.

customer_id	spend	store_id
a1	28.94	s1
a3	874.12	NULL
a4	8.99	s2

D.

customer_id	spend	store_id
a1	28.94	s1
a2	NULL	s1
a3	874.12	NULL
a4	8.99	s2

E.

customer_id	spend	store_id
a1	28.94	s1
a2	NULL	s1
a4	8.99	s2

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

Answer: C

NEW QUESTION 5

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 6

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 7

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

Explanation:

The Auto Loader process in Databricks is typically used in conjunction with Spark Structured Streaming to process data incrementally. Spark Structured Streaming is a real-time data processing framework that allows you to process data streams incrementally as new data arrives. The Auto Loader is a feature in Databricks that works with Structured Streaming to automatically detect and process new data files as they are added to a specified data source location. It allows for incremental data processing without the need for manual intervention.

How does Auto Loader track ingestion progress? As files are discovered, their metadata is persisted in a scalable key-value store (RocksDB) in the checkpoint location of your Auto Loader pipeline. This key-value store ensures that data is processed exactly once. In case of failures, Auto Loader can resume from where it left off by information stored in the checkpoint location and continue to provide exactly-once guarantees when writing data into Delta Lake. You don't need to maintain or manage any state yourself to achieve fault tolerance or exactly-once semantics.<https://docs.databricks.com/ingestion/auto-loader/index.html>

NEW QUESTION 8

A new data engineering team has been assigned to work on a project. The team will need access to database customers in order to see what tables already exist. The team has its own group team.

Which of the following commands can be used to grant the necessary permission on the entire database to the new team?

- A. GRANT VIEW ON CATALOG customers TO team;
- B. GRANT CREATE ON DATABASE customers TO team;
- C. GRANT USAGE ON CATALOG team TO customers;
- D. GRANT CREATE ON DATABASE team TO customers;
- E. GRANT USAGE ON DATABASE customers TO team;

Answer: E

Explanation:

The GRANT statement is used to grant privileges on a database, table, or view to a user or role. The ALL PRIVILEGES option grants all possible privileges on the specified object, such as CREATE, SELECT, MODIFY, and USAGE. The syntax of the GRANT statement is:

GRANT privilege_type ON object TO user_or_role;

Therefore, to grant full permissions on the database customers to the new data engineering team, the command should be:

GRANT ALL PRIVILEGES ON DATABASE customers TO team;

NEW QUESTION 9

A data engineer is using the following code block as part of a batch ingestion pipeline to read from a composable table:

```
transactions_df = (spark.read
    .schema(schema)
    .format("delta")
    .table("transactions")
)
```

Which of the following changes needs to be made so this code block will work when the transactions table is a stream source?

- A. Replace predict with a stream-friendly prediction function
- B. Replace schema(schema) with option ("maxFilesPerTrigger", 1)
- C. Replace "transactions" with the path to the location of the Delta table
- D. Replace format("delta") with format("stream")
- E. Replace spark.read with spark.readStream

Answer: E

Explanation:

<https://docs.databricks.com/en/structured-streaming/delta-lake.html>

NEW QUESTION 10

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

Explanation:

The engine uses checkpointing and write-ahead logs to record the offset range of the data being processed in each trigger. -- in the link search for "The engine uses " you'll find the answer. <https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html#:~:text=The%20engine%20uses%20checkpointing%20and,being%20processe d%20in%20each%20trigger.>

NEW QUESTION 10

A data engineer has three tables in a Delta Live Tables (DLT) pipeline. They have configured the pipeline to drop invalid records at each table. They notice that some data is being dropped due to quality concerns at some point in the DLT pipeline. They would like to determine at which table in their pipeline the data is being dropped.

Which of the following approaches can the data engineer take to identify the table that is dropping the records?

- A. They can set up separate expectations for each table when developing their DLT pipeline.
- B. They cannot determine which table is dropping the records.
- C. They can set up DLT to notify them via email when records are dropped.
- D. They can navigate to the DLT pipeline page, click on each table, and view the data quality statistics.
- E. They can navigate to the DLT pipeline page, click on the "Error" button, and review the present errors.

Answer: D

Explanation:

To identify the table in a Delta Live Tables (DLT) pipeline where data is being dropped due to quality concerns, the data engineer can navigate to the DLT pipeline page, click on each table in the pipeline, and view the data quality statistics. These statistics often include information about records dropped, violations of expectations, and other data quality metrics. By examining the data quality statistics for each table in the pipeline, the data engineer can determine at which table the data is being dropped.

NEW QUESTION 13

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 14

A data engineer wants to schedule their Databricks SQL dashboard to refresh once per day, but they only want the associated SQL endpoint to be running when it is necessary.

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

Answer: C

NEW QUESTION 15

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 19

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. Rebasing a branch.
<https://docs.databricks.com/repos/index.html>

NEW QUESTION 23

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 24

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

Explanation:

<https://docs.databricks.com/delta-live-tables/expectations.html>
Delta Live Tables is a tool provided by Databricks that can help data engineers automate the monitoring of data quality. It is designed for managing data pipelines, monitoring data quality, and automating workflows. With Delta Live Tables, you can set up data quality checks and alerts to detect issues and anomalies in your

data as it is ingested and processed in real-time. It provides a way to ensure that the data quality meets your desired standards and can trigger actions or notifications when issues are detected. While the other tools mentioned may have their own purposes in a data engineering environment, Delta Live Tables is specifically designed for data quality monitoring and automation within the Databricks ecosystem.

NEW QUESTION 29

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 31

A data engineer has a single-task Job that runs each morning before they begin working. After identifying an upstream data issue, they need to set up another task to run a new notebook prior to the original task.

Which of the following approaches can the data engineer use to set up the new task?

- A. They can clone the existing task in the existing Job and update it to run the new notebook.
- B. They can create a new task in the existing Job and then add it as a dependency of the original task.
- C. They can create a new task in the existing Job and then add the original task as a dependency of the new task.
- D. They can create a new job from scratch and add both tasks to run concurrently.
- E. They can clone the existing task to a new Job and then edit it to run the new notebook.

Answer: B

Explanation:

To set up the new task to run a new notebook prior to the original task in a single-task Job, the data engineer can use the following approach: In the existing Job, create a new task that corresponds to the new notebook that needs to be run. Set up the new task with the appropriate configuration, specifying the notebook to be executed and any necessary parameters or dependencies. Once the new task is created, designate it as a dependency of the original task in the Job configuration. This ensures that the new task is executed before the original task.

NEW QUESTION 35

A data engineer needs access to a table new_table, but they do not have the correct permissions. They can ask the table owner for permission, but they do not know who the table owner is.

Which of the following approaches can be used to identify the owner of new_table?

- A. Review the Permissions tab in the table's page in Data Explorer
- B. All of these options can be used to identify the owner of the table
- C. Review the Owner field in the table's page in Data Explorer
- D. Review the Owner field in the table's page in the cloud storage solution
- E. There is no way to identify the owner of the table

Answer: C

NEW QUESTION 36

Which of the following describes a benefit of creating an external table from Parquet rather than CSV when using a CREATE TABLE AS SELECT statement?

- A. Parquet files can be partitioned
- B. CREATE TABLE AS SELECT statements cannot be used on files
- C. Parquet files have a well-defined schema
- D. Parquet files have the ability to be optimized
- E. Parquet files will become Delta tables

Answer: C

Explanation:

<https://www.databricks.com/glossary/what-is-parquet#:~:text=Columnar%20storage%20like%20Apache%20Parquet,compared%20to%20row%2Doriented%20databases.> Columnar storage like Apache Parquet is designed to bring efficiency compared to row-based files like CSV. When querying, columnar storage you can skip over the non-relevant data very quickly. As a result, aggregation queries are less time-consuming compared to row-oriented databases.

NEW QUESTION 38

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 39

A data engineer has realized that the data files associated with a Delta table are incredibly small. They want to compact the small files to form larger files to improve performance.

Which of the following keywords can be used to compact the small files?

- A. REDUCE
- B. OPTIMIZE
- C. COMPACTION
- D. REPARTITION
- E. VACUUM

Answer: B

Explanation:

OPTIMIZE can be used to club small files into 1 and improve performance.

NEW QUESTION 41

A data engineer has joined an existing project and they see the following query in the project repository:

```
CREATE STREAMING LIVE TABLE loyal_customers AS SELECT customer_id -  
FROM STREAM(LIVE.customers) WHERE loyalty_level = 'high';
```

Which of the following describes why the STREAM function is included in the query?

- A. The STREAM function is not needed and will cause an error.
- B. The table being created is a live table.
- C. The customers table is a streaming live table.
- D. The customers table is a reference to a Structured Streaming query on a PySpark DataFrame.
- E. The data in the customers table has been updated since its last run.

Answer: C

Explanation:

<https://docs.databricks.com/en/sql/load-data-streaming-table.html> Load data into a streaming table

To create a streaming table from data in cloud object storage, paste the following into the query editor, and then click Run:

SQL

Copy to clipboardCopy

/* Load data from a volume */

```
CREATE OR REFRESH STREAMING TABLE <table-name> AS SELECT * FROM STREAM
```

```
read_files('/Volumes/<catalog>/<schema>/<volume>/<path>/<folder>')
```

/* Load data from an external location */

```
CREATE OR REFRESH STREAMING TABLE <table-name> AS
```

SELECT * FROM STREAM read_files('s3://<bucket>/<path>/<folder>')

NEW QUESTION 44

A data architect has determined that a table of the following format is necessary:

employeeId	startDate	avgRating
a1	2009-01-06	5.5
a2	2018-11-21	7.1
...

Which of the following code blocks uses SQL DDL commands to create an empty Delta table in the above format regardless of whether a table already exists with this name?

- ```
CREATE TABLE IF NOT EXISTS table_name (
 employeeId STRING,
 startDate DATE,
 avgRating FLOAT
)

CREATE OR REPLACE TABLE table_name AS
SELECT
 employeeId STRING,
 startDate DATE,
 avgRating FLOAT
USING DELTA

CREATE OR REPLACE TABLE table_name WITH COLUMNS (
 employeeId STRING,
 startDate DATE,
 avgRating FLOAT
) USING DELTA

CREATE TABLE table_name AS
SELECT
 employeeId STRING,
 startDate DATE,
 avgRating FLOAT

CREATE OR REPLACE TABLE table_name (
 employeeId STRING,
 startDate DATE,
 avgRating FLOAT
)
```
- A.

B.

C.

D.

E.

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

**Answer:** E

**NEW QUESTION 48**

.....

## 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](#)**