

Databricks

Exam Questions Databricks-Certified-Data-Engineer-Associate

Databricks Certified Data Engineer Associate Exam



NEW QUESTION 1

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 2

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 3

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 4

Which of the following must be specified when creating a new Delta Live Tables pipeline?

- A. A key-value pair configuration
- B. The preferred DBU/hour cost
- C. A path to cloud storage location for the written data
- D. A location of a target database for the written data
- E. At least one notebook library to be executed

Answer: E

Explanation:

<https://docs.databricks.com/en/delta-live-tables/tutorial-pipelines.html>

NEW QUESTION 5

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 6

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?

- 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: C

Explanation:

<https://www.databricks.com/blog/2022/03/10/top-5-databricks-performance-tips.html>

NEW QUESTION 7

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 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 has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to a data analytics dashboard for a retail use case. The job has a Databricks SQL query that returns the number of store-level records where sales is equal to zero. The data engineer wants their entire team to be notified via a messaging webhook whenever this value is greater than 0.

Which of the following approaches can the data engineer use to notify their entire team via a messaging webhook whenever the number of stores with \$0 in sales is greater than zero?

- 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 one-time notifications.
- D. They can set up an Alert with a new webhook alert destination.
- E. They can set up an Alert without notifications.

Answer: D

NEW QUESTION 10

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 10

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 15

A data engineer runs a statement every day to copy the previous day's sales into the table transactions. Each day's sales are in their own file in the location "/transactions/raw".

Today, the data engineer runs the following command to complete this task:

```
COPY INTO transactions
FROM "/transactions/raw"
FILEFORMAT = PARQUET;
```

After running the command today, the data engineer notices that the number of records in table transactions has not changed.

Which of the following describes why the statement might not have copied any new records into the table?

- A. The format of the files to be copied were not included with the FORMAT_OPTIONS keyword.
- B. The names of the files to be copied were not included with the FILES keyword.
- C. The previous day's file has already been copied into the table.
- D. The PARQUET file format does not support COPY INTO.
- E. The COPY INTO statement requires the table to be refreshed to view the copied rows.

Answer: C

Explanation:

<https://docs.databricks.com/en/ingestion/copy-into/index.html> The COPY

INTO SQL command lets you load data from a file location into a Delta table. This is a re- triable and idempotent operation; files in the source location that have already been loaded are skipped. if there are no new records, the only consistent choice is C no new files were loaded because already loaded files were skipped.

NEW QUESTION 20

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 22

A data engineer has realized that they made a mistake when making a daily update to a table. They need to use Delta time travel to restore the table to a version that is 3 days old. However, when the data engineer attempts to time travel to the older version, they are unable to restore the data because the data files have been deleted.

Which of the following explains why the data files are no longer present?

- A. The VACUUM command was run on the table
- B. The TIME TRAVEL command was run on the table
- C. The DELETE HISTORY command was run on the table

- D. The OPTIMIZE command was run on the table
- E. The HISTORY command was run on the table

Answer: A

Explanation:

The VACUUM command in Delta Lake is used to clean up and remove unnecessary data files that are no longer needed for time travel or query purposes. When you run VACUUM with certain retention settings, it can delete older data files, which might include versions of data that are older than the specified retention period. If the data engineer is unable to restore the table to a version that is 3 days old because the data files have been deleted, it's likely because the VACUUM command was run on the table, removing the older data files as part of data cleanup.

NEW QUESTION 27

Which of the following describes a scenario in which a data team will want to utilize cluster pools?

- A. An automated report needs to be refreshed as quickly as possible.
- B. An automated report needs to be made reproducible.
- C. An automated report needs to be tested to identify errors.
- D. An automated report needs to be version-controlled across multiple collaborators.
- E. An automated report needs to be runnable by all stakeholders.

Answer: A

Explanation:

Cluster pools are typically used in distributed computing environments, such as cloud-based data platforms like Databricks. They allow you to pre-allocate a set of compute resources (a cluster) for specific tasks or workloads. In this case, if an automated report needs to be refreshed as quickly as possible, you can allocate a cluster pool with sufficient resources to ensure fast data processing and report generation. This helps ensure that the report is generated with minimal latency and can be delivered to stakeholders in a timely manner. Cluster pools allow you to optimize resource allocation for high-demand, time-sensitive tasks like real-time report generation.

NEW QUESTION 31

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 34

An engineering manager wants to monitor the performance of a recent project using a Databricks SQL query. For the first week following the project's release, the manager wants the query results to be updated every minute. However, the manager is concerned that the compute resources used for the query will be left running and cost the organization a lot of money beyond the first week of the project's release.

Which of the following approaches can the engineering team use to ensure the query does not cost the organization any money beyond the first week of the project's release?

- A. They can set a limit to the number of DBUs that are consumed by the SQL Endpoint.
- B. They can set the query's refresh schedule to end after a certain number of refreshes.
- C. They cannot ensure the query does not cost the organization money beyond the first week of the project's release.
- D. They can set a limit to the number of individuals that are able to manage the query's refresh schedule.
- E. They can set the query's refresh schedule to end on a certain date in the query scheduler.

Answer: E

Explanation:

If a dashboard is configured for automatic updates, it has a Scheduled button at the top, rather than a Schedule button. To stop automatically updating the dashboard and remove its subscriptions:

Click Scheduled.

In the Refresh every drop-down, select Never.

Click Save. The Scheduled button label changes to Schedule. Source: <https://learn.microsoft.com/en-us/azure/databricks/sql/user/dashboards/>

NEW QUESTION 38

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

- G. All datasets will be updated once and the pipeline will shut dow
- H. The compute resources will be terminated.
- I. All datasets will be updated once and the pipeline will shut dow
- J. The compute resources will persist to allow for additional testing.

Answer: C

Explanation:

In a Delta Live Table pipeline running in Continuous Pipeline Mode, when you click Start to update the pipeline, the following outcome is expected: All datasets defined using STREAMING LIVE TABLE and LIVE TABLE against Delta Lake table sources will be updated at set intervals. The compute resources will be deployed for the update process and will be active during the execution of the pipeline. The compute resources will be terminated when the pipeline is stopped or shut down. This mode allows for continuous and periodic updates to the datasets as new data arrives or changes in the underlying Delta Lake tables occur. The compute resources are provisioned and utilized during the update intervals to process the data and perform the necessary operations.

NEW QUESTION 40

A single Job runs two notebooks as two separate tasks. A data engineer has noticed that one of the notebooks is running slowly in the Job's current run. The data engineer asks a tech lead for help in identifying why this might be the case.

Which of the following approaches can the tech lead use to identify why the notebook is running slowly as part of the Job?

- A. They can navigate to the Runs tab in the Jobs UI to immediately review the processing notebook.
- B. They can navigate to the Tasks tab in the Jobs UI and click on the active run to review the processing notebook.
- C. They can navigate to the Runs tab in the Jobs UI and click on the active run to review the processing notebook.
- D. There is no way to determine why a Job task is running slowly.
- E. They can navigate to the Tasks tab in the Jobs UI to immediately review the processing notebook.

Answer: C

Explanation:

The job run details page contains job output and links to logs, including information about the success or failure of each task in the job run. You can access job run details from the Runs tab for the job. To view job run details from the Runs tab, click the link for the run in the Start time column in the runs list view. To return to the Runs tab for the job, click the Job ID value.

If the job contains multiple tasks, click a task to view task run details, including: the cluster that ran the task
the Spark UI for the task logs for the task
metrics for the task

<https://docs.databricks.com/en/workflows/jobs/monitor-job-runs.html#job-run-details>

NEW QUESTION 44

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 49

Which of the following is stored in the Databricks customer's cloud account?

- A. Databricks web application
- B. Cluster management metadata
- C. Repos
- D. Data
- E. Notebooks

Answer: D

NEW QUESTION 50

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 53

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%20oriented%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 57

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 62

.....

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](#)