DATABRICKS-CERTIFIED-DATA-ENGINEER-ASSOCIATE Q&As

Databricks Certified Data Engineer Associate Exam

Pass Databricks DATABRICKS-CERTIFIED-DATA-ENGINEER-ASSOCIATE Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.certbus.com/databricks-certified-data-engineer-associate.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Databricks Official Exam Center

QUESTION 1

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

Correct 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

QUESTION 2

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.

Correct Answer: A

Explanation: In some data processing pipelines, especially those following a typical "Bronze-Silver-Gold" data lakehouse architecture, Silver tables are often considered a more refined version of the raw or Bronze data. Silver tables may include data cleansing, schema enforcement, and some initial transformations. Gold tables, on the other hand, typically represent a stage where data is further enriched, aggregated, and processed to provide valuable insights for analytical purposes. This could indeed involve more aggregations compared to Silver tables.

QUESTION 3

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

Correct Answer: C

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

QUESTION 4

A data engineer needs to apply custom logic to identify employees with more than 5 years of experience in array column employees in table stores. The custom logic should create a new column exp_employees that is an array of all of the employees with more than 5 years of experience for each row. In order to apply this custom logic at scale, the data engineer wants to use the FILTER higher-order function.

Which of the following code blocks successfully completes this task?

A.
```
SELECT
    store_id,
    employees,
    FILTER (employees, i -> i.years_exp > 5) AS exp_employees
FROM stores;
```

B.
```
SELECT
    store_id,
    employees,
    FILTER (exp_employees, years_exp > 5) AS exp_employees
FROM stores;
```

C.
```
SELECT
    store_id,
    employees,
    FILTER (employees, years_exp > 5) AS exp_employees
FROM stores;
```

D.
```
SELECT
    store_id,
    employees,
    CASE WHEN employees.years_exp > 5 THEN employees
         ELSE NULL
    END AS exp_employees
FROM stores;
```

E.
```
SELECT
    store_id,
    employees,
    FILTER (exp_employees, i -> i.years_exp > 5) AS exp_employees
FROM stores;
```

A. Option A

B. Option B

C. Option C

D. Option D

E. Option E

Correct Answer: A

QUESTION 5

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;

Correct 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.

QUESTION 6

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

Correct Answer: D

QUESTION 7

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.

Correct 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.

QUESTION 8

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

Correct 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.

QUESTION 9

Which of the following is a benefit of the Databricks Lakehouse Platform embracing open source technologies?

A. Cloud-specific integrations

B. Simplified governance

C. Ability to scale storage

D. Ability to scale workloads

E. Avoiding vendor lock-in

Correct Answer: E

Explanation: https://double.cloud/blog/posts/2023/01/break-free-from-vendor-lock-in-with- open-source-tech/

QUESTION 10

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

Correct Answer: C

Explanation: The reason why the data files still exist while the metadata files were deleted is because the table was external. When a table is external in Spark SQL (or in other database systems), it means that the table metadata (such as schema information and table structure) is managed externally, and Spark SQL assumes that the data is managed and maintained outside of the system. Therefore, when you execute a DROP TABLE statement for an external table, it removes only the table metadata from the catalog, leaving the data files intact. On the other hand, for managed tables (option E), Spark SQL manages both the metadata and the data files. When you drop a managed table, it deletes both the metadata and the associated data files, resulting in a complete removal of the table.

QUESTION 11

A data engineer needs to use a Delta table as part of a data pipeline, but they do not know if they have the appropriate permissions. In which of the following locations can the data engineer review their permissions on the table?

A. Databricks Filesystem

B. Jobs

C. Dashboards

D. Repos

E. Data Explorer

Correct Answer: E

QUESTION 12

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

Correct Answer: C

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

QUESTION 13

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.

Correct Answer: E

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/

QUESTION 14

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."

Correct 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.

QUESTION 15

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

Correct Answer: D

https://www.w3schools.com/python/python_functions.asp