Databricks

Exam Questions Databricks-Certified-Data-Engineer-Associate

Databricks Certified Data Engineer Associate Exam

About Exambible

Your Partner of IT Exam

Found in 1998

Exambible is a company specialized on providing high quality IT exam practice study materials, especially Cisco CCNA, CCDA, CCNP, CCIE, Checkpoint CCSE, CompTIA A+, Network+ certification practice exams and so on. We guarantee that the candidates will not only pass any IT exam at the first attempt but also get profound understanding about the certificates they have got. There are so many alike companies in this industry, however, Exambible has its unique advantages that other companies could not achieve.

Our Advances

* 99.9% Uptime

    All examinations will be up to date.

* 24/7 Quality Support

    We will provide service round the clock.

* 100% Pass Rate

    Our guarantee that you will pass the exam.

* Unique Gurantee

    If you do not pass the exam at the first time, we will not only arrange FULL REFUND for you, but also provide you another exam of your claim, ABSOLUTELY FREE!

NEW 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

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 2
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

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.

NEW QUESTION 3
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 a external location. hence when a external table is dropped you clear off only the metadata and the files (data) remain.

NEW QUESTION 4
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

Answer: E

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

NEW QUESTION 5
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 6
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 7
A data analyst has created a Delta table sales that is used by the entire data analysis team. They want help from the data engineering team to implement a series of tests to ensure the data is clean. However, the data engineering team uses Python for its tests rather than SQL.
Which of the following commands could the data engineering team use to access sales in PySpark?

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

Answer: E

Explanation:
https://spark.apache.org/docs/3.2.1/api/python/reference/api/pyspark.sql.SparkSession.tabl e.html

NEW QUESTION 8
A new data engineering team team has been assigned to an ELT project. The new data engineering team will need full privileges on the table sales to fully manage the project.
Which of the following commands can be used to grant full permissions on the database to the new data engineering team?

A. GRANT ALL PRIVILEGES ON TABLE sales TO team;
B. GRANT SELECT CREATE MODIFY ON TABLE sales TO team;
C. GRANT SELECT ON TABLE sales TO team;
D. GRANT USAGE ON TABLE sales TO team;
E. GRANT ALL PRIVILEGES ON TABLE team TO sales;

Answer: A

NEW QUESTION 9
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 " youll 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 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
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 12
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:
(f"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 14
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 16
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 17
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 18
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 20
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 23
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 24
......

Relate Links

100% Pass Your Databricks-Certified-Data-Engineer-Associate Exam with Exambible Prep Materials

https://www.exambible.com/Databricks-Certified-Data-Engineer-Associate-exam/

Contact us

We are proud of our high-quality customer service, which serves you around the clock 24/7.

Viste - https://www.exambible.com/