Free Questions for Databricks-Certified-Data-Engineer-Associate by certscare

Shared by Cook on 12-12-2023

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

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

Options:

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/

Question 2

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?

Options:

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

Question 3

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

Options:

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

Question 4

Which of the following commands will return the number of null values in the member_id column?

Options:

A- SELECT count(member_id) FROM my_table;

B- SELECT count(member_id) - count_null(member_id) FROM my_table;

C- SELECT count_if(member_id IS NULL) FROM my_table;

D- SELECT null(member_id) FROM my_table;

E- SELECT count_null(member_id) FROM my_table;

Answer:

C

Explanation:

https://docs.databricks.com/en/sql/language-manual/functions/count.html

Returns

A BIGINT.

If * is specified also counts row containing NULL values.

If expr are specified counts only rows for which all expr are not NULL.

If DISTINCT duplicate rows are not counted.

Question 5

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?

Options:

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

Question 6

Question Type: **MultipleChoice**

Which of the following can be used to simplify and unify siloed data architectures that are specialized for specific use cases?

Options:

A- None of these

B- Data lake

C- Data warehouse

D- All of these

E- Data lakehouse

Answer:

E

Question 7

Question Type: **MultipleChoice**

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?

Options:

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

Question 8

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

Options:

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.

Question 9

Question Type: MultipleChoice

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?

Options:

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())