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

Shared by Tyler on 15-04-2024

For More Free Questions and Preparation Resources

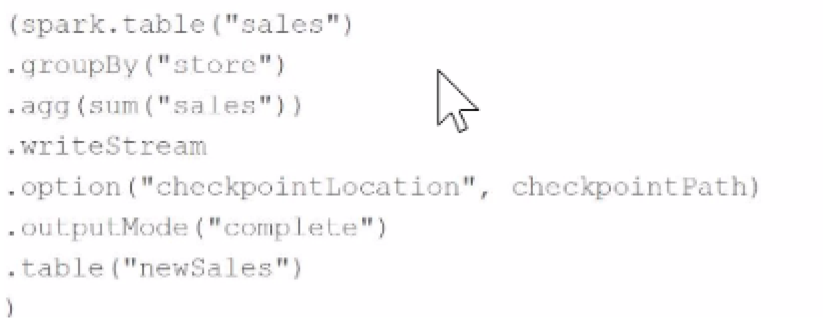
Check the Links on Last Page

Question 1

Question Type: MultipleChoice

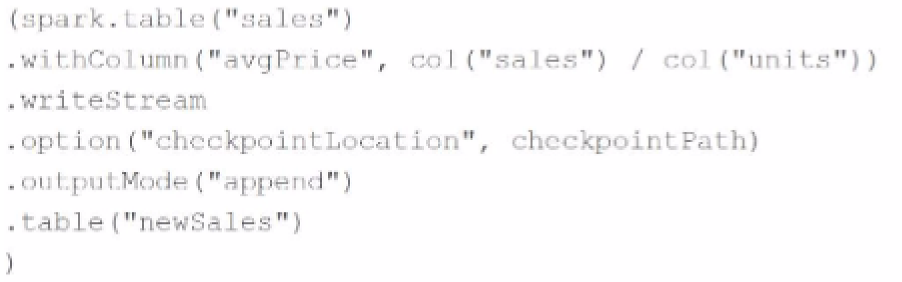
Which query is performing a streaming hop from raw data to a Bronze table?

A)



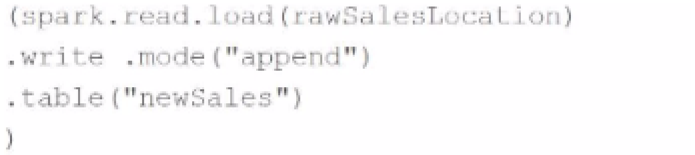
```
(spark.table("sales")
.groupBy("store")
.agg(sum("sales"))
.writeStream
.option("checkpointLocation", checkpointPath)
.outputMode("complete")
.table("newSales")
)
```

B)



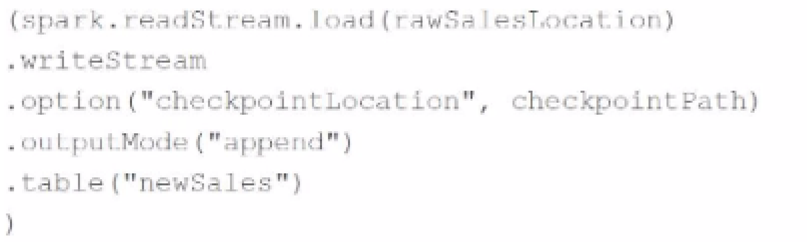
```
(spark.table("sales")
.withColumn("avgPrice", col("sales") / col("units"))
.writeStream
.option("checkpointLocation", checkpointPath)
.outputMode("append")
.table("newSales")
)
```

C)



```
(spark.read.load(rawSalesLocation)
.write .mode("append")
.table("newSales")
)
```

D)



```
(spark.readStream.load(rawSalesLocation)
.writeStream
.option("checkpointLocation", checkpointPath)
.outputMode("append")
.table("newSales")
)
```

Options:

A- Option A

B- Option B

C- Option C

D- Option D

Answer:

D

Question 2

Question Type: MultipleChoice

Which file format is used for storing Delta Lake Table?

Options:

A- Parquet

B- Delta

C- SV

D- JSON

Answer:

A

Question 3

Question Type: MultipleChoice

Which of the following describes the type of workloads that are always compatible with Auto Loader?

Options:

A- Dashboard workloads

B- Streaming workloads

C- Machine learning workloads

D- Serverless workloads

E- Batch workloads

Answer:

B

Explanation:

Auto Loader is a Structured Streaming source that incrementally and efficiently processes new data files as they arrive in cloud storage. It supports both Python and SQL in Delta Live Tables, which are ideal for building streaming data pipelines. Auto Loader can handle near real-time ingestion of millions of files per hour and provide exactly-once guarantees when writing data into Delta Lake. Auto Loader is not designed for dashboard, machine learning, serverless, or batch workloads, which have different requirements and characteristics.Reference:What is Auto Loader?,Delta Live Tables

Question 4

Question Type: **MultipleChoice**

Which of the following SQL keywords can be used to convert a table from a long format to a wide format?

Options:

A- PIVOT

B- CONVERT

C- WHERE

D- TRANSFORM

E- SUM

Answer:

A

Explanation:

The SQL keyword that can be used to convert a table from a long format to a wide format isPIVOT.The PIVOT clause is used to rotate the rows of a table into columns of a new table1.The PIVOT clause can aggregate the values of a column based on the distinct values of another column, and use those values as the column names of the new table1.The PIVOT clause can be useful for transforming data from a long format, where each row represents an observation with multiple attributes, to a wide format, where each row represents an observation with a single attribute and multiple values2.For example, the PIVOT clause can be used to convert a table that contains the sales of different products by different regions into a table that contains the sales of each product by each region as separate columns1.

The other options are not suitable for converting a table from a long format to a wide format.CONVERT is a function that can be used to change the data type of an expression3.WHERE is a clause that can be used to filter the rows of a table based on a condition4.TRANSFORM is a keyword that can be used to apply a user-defined function to a group of rows in a table5. SUM is a function that can be used to calculate the total of a numeric column.

1:PIVOT | Databricks on AWS

2:Reshaping Data - Long vs Wide Format | Databricks on AWS

3:CONVERT | Databricks on AWS

4:WHERE | Databricks on AWS

5:TRANSFORM | Databricks on AWS

: [SUM | Databricks on AWS]

Question 5

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?

Options:

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:

D

Explanation:

Option D is the correct answer because it enables the Serverless feature for the SQL endpoint, which allows the endpoint to automatically scale up and down based on the query load. This way, the endpoint can handle more concurrent queries and reduce the time it takes to return results. The Serverless feature also reduces the cold start time of the endpoint, which is the time it takes to start the cluster when a query is submitted to a non-running endpoint. The Serverless feature is available for both AWS and Azure Databricks platforms.

Question 6

Question Type: MultipleChoice

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?

Options:

A- Databricks Filesystem

B- Jobs

C- Dashboards

D- Repos

E- Data Explorer

Answer:

E

Explanation:

Data Explorer is a graphical interface that allows you to browse, create, and manage data objects such as databases, tables, and views in your workspace. You can also review and modify the permissions on these data objects using Data Explorer. To access Data Explorer, you can click on the Data icon in the sidebar, or use the %sql magic command in a notebook. You can then select a database and a table, and click on the Permissions tab to view and edit the access control lists (ACLs) for the table. You can also use SQL commands such as SHOW GRANT and GRANT to query and modify the permissions on a Delta table.Reference:

Data Explorer

Access control for Delta tables

SHOW GRANT

[GRANT]

Question 7

Question Type: MultipleChoice

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?

Options:

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:

The reason why all of the data files and metadata files were deleted from the file system after dropping the table is that the table was managed. A managed table is a table that is created and managed by Spark SQL. It stores both the data and the metadata in the default location specified by thespark.sql.warehouse.dirconfiguration property. When a managed table is dropped, both the data and the metadata are deleted from the file system.

Option B is not correct, as the size of the table's data does not affect the behavior of dropping the table. Whether the table's data is smaller or larger than 10 GB, the data files and metadata files will be deleted if the table is managed, and will be preserved if the table is external.

Option C is not correct, for the same reason as option B.

Option D is not correct, as an external table is a table that is created and managed by the user. It stores the data in a user-specified location, and only stores the metadata in the Spark SQL catalog. When an external table is dropped, only the metadata is deleted from the catalog, but the data files are preserved in the file system.

Option E is not correct, as a table must have a location to store the data. If the location is not specified by the user, it will use the default location for managed tables. Therefore, a table without a location is a managed table, and dropping it will delete both the data and the metadata.

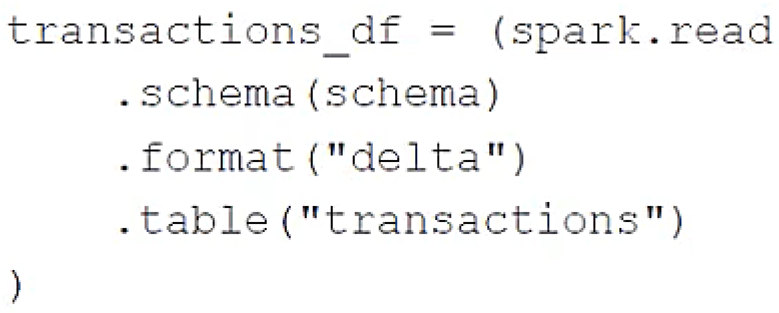
Managing Tables

[Databricks Data Engineer Professional Exam Guide]

Question 8

Question Type: MultipleChoice

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?

Options:

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:

: To read from a stream source, the data engineer needs to use the spark.readStream method instead of the spark.read method. The spark.readStream method returns a DataStreamReader object that can be used to specify the details of the input source, such as the format, the schema, the path, and the options. The spark.read method is only suitable for batch processing, not streaming processing. The other changes are not necessary or correct for reading from a stream source.Reference:Structured Streaming Programming Guide,Read a stream,Databricks Data Sources

Question 9

Question Type: **MultipleChoice**

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?

Options:

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

Answer:

B

Explanation:

JSON data is a text-based format that represents data as a collection of name-value pairs. By default, when Auto Loader infers the schema of JSON data, it treats all columns as strings. This is because JSON data can have varying data types for the same column across different files or records, and Auto Loader does not attempt to reconcile these differences. For example, a column named "age" may have integer values in some files, but string values in others. To avoid data loss or errors, Auto Loader infers the column as a string type. However, Auto Loader also provides an option to infer more precise column types based on the sample data. This option is called cloudFiles.inferColumnTypes and it can be set to true or false. When set to true, Auto Loader tries to infer the exact data types of the columns, such as integers, floats, booleans, or nested structures. When set to false, Auto Loader infers all columns as strings. The default value of this option is false.Reference:Configure schema inference and evolution in Auto Loader,Schema inference with auto loader (non-DLT and DLT),Using and Abusing Auto Loader's Inferred Schema,Explicit path to data or a defined schema required for Auto loader.

Question 10

Question Type: **MultipleChoice**

In which of the following scenarios should a data engineer use the MERGE INTO command instead of the INSERT INTO command?

Options:

A- When the location of the data needs to be changed

B- When the target table is an external table

C- When the source table can be deleted

D- When the target table cannot contain duplicate records

E- When the source is not a Delta table

Answer:

D

Explanation:

The MERGE INTO command is used to perform upserts, which are a combination of insertions and updates, based on a source table into a target Delta table1. The MERGE INTO command can handle scenarios where the target table cannot contain duplicate records, such as when there is a primary key or a unique constraint on the target table. The MERGE INTO command can match the source and target rows based on a merge condition and perform different actions depending on whether the rows are matched or not.For example, the MERGE INTO command can update the existing target rows with the new source values, insert the new source rows that do not exist in the target table, or delete the target rows that do not exist in the source table1.

The INSERT INTO command is used to append new rows to an existing table or create a new table from a query result2. The INSERT INTO command does not perform any updates or deletions on the existing target table rows.The INSERT INTO command can handle scenarios where the location of the data needs to be changed, such as when the data needs to be moved from one table to another, or when the data needs to be partitioned by a certain column2.The INSERT INTO command can also handle scenarios where the target table is an external table, such as when the data is stored in an external storage system like Amazon S3 or Azure Blob Storage3.The INSERT INTO command can also handle scenarios where the source table can be deleted, such as when the source table is a temporary table or a view4.The INSERT INTO command can also handle scenarios where the source is not a Delta table, such as when the source is a Parquet, CSV, JSON, or Avro file5.

1:MERGE INTO | Databricks on AWS

2: [INSERT INTO | Databricks on AWS]

3: [External tables | Databricks on AWS]

4: [Temporary views | Databricks on AWS]

5: [Data sources | Databricks on AWS]

20%
DISCOUNT