Databricks-Certified-Data-Engineer-Associate Practice Questions

Shared by Castaneda on 18-08-2026

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

A data engineer is designing a Bronze-to-Silver pipeline on the Databricks Data Intelligence Platform. The source system sends daily CSV files, and new optional columns are added over time.

The engineer needs a storage format and table capabilities that provide all of the following:

Writes that do not conform to the defined schema are rejected.

The schema can evolve to include new optional columns without manually recreating the table.

Previous table versions can be queried for debugging and auditing.

Which solution fulfills these requirements?

Options:

A- Use a Parquet table with Spark's default schema inference and rerun the job whenever the schema changes.
B- Use a Delta table with schema enforcement and recreate the table whenever new columns are added.
C- Use an external table with Auto Loader schema inference for the CSV files.
D- Use a Delta table with its native schema enforcement, schema evolution, and table-history capabilities.

Answer:

D

Question 2

Question Type: MultipleChoice

A data engineer is designing a data pipeline. The source system generates files in a shared directory that is also used by other processes. As a result, the files should be kept as is and will accumulate in the directory. The data engineer needs to identify which files are new since the previous run in the pipeline, and set up the pipeline to only ingest those new files with each run.

Which of the following tools can the data engineer use to solve this problem?

Options:

A- Unity Catalog
B- Delta Lake
C- Databricks SQL
D- Data Explorer
E- Auto Loader

Answer:

E

Explanation:

Auto Loader is a tool that can incrementally and efficiently process new data files as they arrive in cloud storage without any additional setup. Auto Loader provides a Structured Streaming source called cloudFiles, which automatically detects and processes new files in a given input directory path on the cloud file storage. Auto Loader also tracks the ingestion progress and ensures exactly-once semantics when writing data into Delta Lake. Auto Loader can ingest various file formats, such as JSON, CSV, XML, PARQUET, AVRO, ORC, TEXT, and BINARYFILE. Auto Loader has support for both Python and SQL in Delta Live Tables, which are a declarative way to build production-quality data pipelines with Databricks.Reference:What is Auto Loader?,Get started with Databricks Auto Loader,Auto Loader in Delta Live Tables

Question 3

Question Type: MultipleChoice

In Which option best 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]

Question 4

Question Type: MultipleChoice

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 option best control flow statements should the data engineer use to begin this conditionally executed code block?

Options:

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:

In Python, the==operator is used to compare the values of two variables, while the=operator is used to assign a value to a variable. Therefore, option A and E are incorrect, as they use the=operator for comparison. Option B and C are also incorrect, as they compare thereview_periodvariable to a string value'True', which is different from the boolean valueTrue. Option D is the correct answer, as it uses the==operator to compare theday_of_weekvariable to the integer value1, and theandoperator to check if both conditions are true. If both conditions are true, then the final block of the Python program will be executed.Reference: [Python Operators], [Python If ... Else]

Question 5

Question Type: MultipleChoice

A data analyst has a series of queries in a SQL program. The data analyst wants this program to run every day. They only want the final query in the program to run on Sundays. They ask for help from the data engineering team to complete this task.

Which of the following approaches could be used by the data engineering team to complete this task?

Options:

A- They could submit a feature request with Databricks to add this functionality.
B- They could wrap the queries using PySpark and use Python's control flow system to determine when to run the final query.
C- They could only run the entire program on Sundays.
D- They could automatically restrict access to the source table in the final query so that it is only accessible on Sundays.
E- They could redesign the data model to separate the data used in the final query into a new

table.

Answer:

B

Explanation:

This approach would allow the data engineering team to use the existing SQL program and add some logic to control the execution of the final query based on the day of the week. They could use thedatetimemodule in Python to get the current date and check if it is a Sunday. If so, they could run the final query, otherwise they could skip it. This way, they could schedule the program to run every day without changing the data model or the source table.Reference:PySpark SQL Module,Python datetime Module,Databricks Jobs

Question 6

Question Type: MultipleChoice

A data engineer has three tables in a Delta Live Tables (DLT) pipeline. They have configured the pipeline to drop invalid records at each table. They notice that some data is being dropped due to quality concerns at some point in the DLT pipeline. They would like to determine at which table in their pipeline the data is being dropped.

Which of the following approaches can the data engineer take to identify the table that is dropping the records?

Options:

A- They can set up separate expectations for each table when developing their DLT pipeline.
B- They cannot determine which table is dropping the records.
C- They can set up DLT to notify them via email when records are dropped.
D- They can navigate to the DLT pipeline page, click on each table, and view the data quality statistics.
E- They can navigate to the DLT pipeline page, click on the "Error" button, and review the present errors.

Answer:

D

Explanation:

One of the features of DLT is that it provides data quality metrics for each dataset in the pipeline, such as the number of records that pass or fail expectations, the number of records that are dropped, and the number of records that are written to the target. These metrics can be accessed from the DLT pipeline page, where the data engineer can click on each table and view the data quality statistics for the latest update or any previous update. This way, they can identify which table is dropping the records and why.Reference:

Monitor Delta Live Tables pipelines

Manage data quality with Delta Live Tables

Question 7

Question Type: MultipleChoice

Which SQL keyword can be used to convert a table from a long format to a wide format?

Options:

A- TRANSFORM
B- PIVOT
C- SUM
D- CONVERT

Answer:

B

Question 8

Question Type: MultipleChoice

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?

Options:

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:

An external table is a table that is defined in the metastore and points to an existing location in the storage system. When you drop an external table, only the metadata is deleted from the metastore, but the data files are not deleted from the storage system. This is because external tables are meant to be shared by multiple applications and users, and dropping them should not affect the data availability. On the other hand, a managed table is a table that is defined in the metastore and also managed by the metastore. When you drop a managed table, both the metadata and the data files are deleted from the metastore and the storage system, respectively. This is because managed tables are meant to be exclusive to the application or user that created them, and dropping them should free up the storage space. Therefore, the correct answer is C, because the table was external and only the metadata was deleted when the table was dropped.Reference:Databricks Documentation - Managed and External Tables,Databricks Documentation - Drop Table

Question 9

Question Type: MultipleChoice

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 option best commands can be used to grant full permissions on the database to the new data engineering team?

Options:

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

Explanation:

To grant full permissions on a table to a user or a group, you can use theGRANT ALL PRIVILEGES ON TABLEstatement. This statement will grant all the possible privileges on the table, such asSELECT,CREATE,MODIFY,DROP,ALTER, etc. Option A is the only code block that follows this syntax correctly. Option B is incorrect, as it does not grant all the possible privileges on the table, but only a subset of them. Option C is incorrect, as it only grants theSELECTprivilege on the table, which is not enough to fully manage the project. Option D is incorrect, as it grants theUSAGEprivilege on the table, which is not a valid privilege for tables. Option E is incorrect, as it grants all the privileges on the tableteamto the user or groupsales, which is the opposite of what the question asks.Reference:Grant privileges on a table using SQL | Databricks on AWS,Grant privileges on a table using SQL - Azure Databricks,SQL Privileges - Databricks

Question 10

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 location can the data engineer review their permissions on the table?

Options:

A- Jobs
B- Dashboards
C- Catalog Explorer
D- Repos

Answer:

C

Question 11

Question Type: MultipleChoice

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?

Options:

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

Answer:

C

Explanation:

: In Production mode, the pipeline runs continuously and updates the output tables whenever new data is available in the input sources. The compute resources are allocated on demand and released when the pipeline is stopped. This mode is suitable for production workloads that require high availability and reliability.Reference:Configure pipeline settings for Delta Live Tables,Tutorial: Run your first Delta Live Tables pipeline,Building Reliable Data Pipelines Using DataBricks' Delta Live Tables

Question 12

Question Type: MultipleChoice

A data engineer has realized that the data files associated with a Delta table are incredibly small. They want to compact the small files to form larger files to improve performance.

Which option best keywords can be used to compact the small files?

Options:

A- REDUCE
B- OPTIMIZE
C- COMPACTION
D- REPARTITION
E- VACUUM

Answer:

B

Explanation:

The keyword that can be used to compact the small files associated with a Delta table isOPTIMIZE.The OPTIMIZE command performs file compaction on a Delta table by rewriting a set of small files into a set of larger files1.This can improve the performance of queries that scan the table by reducing the number of files that need to be read and the amount of metadata that needs to be processed1.The OPTIMIZE command can also optionally sort the data within each file by a given set of columns, which can further improve the query performance by enabling data skipping and predicate pushdown1.The OPTIMIZE command can be applied to the whole table or to a specific partition of the table1.

The other keywords are not suitable for compacting the small files associated with a Delta table.REDUCE is a keyword used in the SQL syntax for aggregating data using a user-defined function2. COMPACTION is not a valid keyword in SQL or Python.REPARTITION is a keyword used in the Python syntax for changing the number of partitions of a DataFrame or an RDD3.VACUUM is a keyword used to remove files that are no longer referenced by a Delta table and are older than a retention threshold4.

1:OPTIMIZE | Databricks on AWS

2:REDUCE | Databricks on AWS

3:repartition | Databricks on AWS

4:VACUUM | Databricks on AWS

P2P exams