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

Databricks Certified Professional Data Engineer Exam

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

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

https://www.passapply.com/databricks-certified-professional-data-engineer.html

100% Passing Guarantee
100% Money Back Assurance

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

QUESTION 1

A junior data engineer seeks to leverage Delta Lake\\'s Change Data Feed functionality to create a Type 1 table representing all of the values that have ever been valid for all rows in abronzetable created with the propertydelta.enableChangeDataFeed = true. They plan to execute the following code as a daily job:

```
from pyspark.sql.functions import col

(spark.read.format("delta")
  .option("readChangeFeed", "true")
  .option("startingVersion", 0)
  .table("bronze")
  .filter(col("_change_type").isin(["update_postimage", "insert"]))
  .write
  .mode("append")
  .table("bronze_history_type1")
 )
```

Which statement describes the execution and results of running the above query multiple times?

A. Each time the job is executed, newly updated records will be merged into the target table, overwriting previous values with the same primary keys.

B. Each time the job is executed, the entire available history of inserted or updated records will be appended to the target table, resulting in many duplicate entries.

C. Each time the job is executed, the target table will be overwritten using the entire history of inserted or updated records, giving the desired result.

D. Each time the job is executed, the differences between the original and current versions are calculated; this may result in duplicate entries for some records.

E. Each time the job is executed, only those records that have been inserted or updated since the last execution will be appended to the target table giving the desired result.

Correct Answer: C

Explanation: This is the correct answer because it describes the execution and results of running the above query multiple times. The query uses the readChanges function to read all change events from a bronze table that has enabled change data feed. The readChanges function takes two arguments: version and options. The version argument specifies which version of the table to read changes from, and can be either a specific version number or -1 to indicate all versions. The options argument specifies additional options for reading changes, such as whether to include deletes or not. In this case, the query reads all changes from all versions of the bronze table and filters out delete events by setting includeDeletes to false. Then, it uses write.format("delta").mode("overwrite") to overwrite a target table using the entire history of inserted or updated records, giving the desired result of a Type 1 table representing all values that have ever been valid for all rows in the bronze table. Verified References: [Databricks Certified Data Engineer Professional], under "Delta Lake" section; Databricks Documentation, under "Read changes in batch queries" section.

QUESTION 2

What statement is true regarding the retention of job run history?

A. It is retained until you export or delete job run logs

B. It is retained for 30 days, during which time you can deliver job run logs to DBFS or S3

C. t is retained for 60 days, during which you can export notebook run results to HTML

D. It is retained for 60 days, after which logs are archived

E. It is retained for 90 days or until the run-id is re-used through custom run configuration

Correct Answer: B

Explanation: This is the correct answer because it is true regarding the retention of job run history. Job run history is the information about each run of a job, such as the start time, end time, status, logs, and output. Job run history is retained for 30 days by default, during which time you can view it in the Jobs UI or access it through the Jobs API. You can also deliver job run logs to DBFS or S3 using the Log Delivery feature, which allows you to specify a destination path and a delivery frequency for each job. By delivering job run logs to DBFS or S3, you can preserve them beyond the 30-day retention period and use them for further analysis or troubleshooting. Verified References: [Databricks Certified Data Engineer Professional], under "Databricks Jobs" section;Databricks Documentation, under "Job run history" section; Databricks Documentation, under "Log Delivery" section.

QUESTION 3

The security team is exploring whether or not the Databricks secrets module can be leveraged for connecting to an external database.

After testing the code with all Python variables being defined with strings, they upload the password to the secrets module and configure the correct permissions for the currently active user. They then modify their code to the following (leaving all other variables unchanged).

```
password = dbutils.secrets.get(scope="db_creds", key="jdbc_password")

print(password)

df = (spark
  .read
  .format("jdbc")
  .option("url", connection)
  .option("dbtable", tablename)
  .option("user", username)
  .option("password", password)
  )
```

Which statement describes what will happen when the above code is executed?

A. The connection to the external table will fail; the string "redacted" will be printed.

B. An interactive input box will appear in the notebook; if the right password is provided, the connection will succeed and the encoded password will be saved to DBFS.

C. An interactive input box will appear in the notebook; if the right password is provided, the connection will succeed and the password will be printed in plain text.

D. The connection to the external table will succeed; the string value of password will be printed in plain text.

E. The connection to the external table will succeed; the string "redacted" will be printed.

Correct Answer: E

Explanation: This is the correct answer because the code is using the dbutils.secrets.get method to retrieve the password from the secrets module and store it in a variable. The secrets module allows users to securely store and access sensitive information such as passwords, tokens, or API keys. The connection to the external table will succeed because the password variable will contain the actual password value. However, when printing the password variable, the string "redacted" will be displayed instead of the plain text password, as a security measure to prevent exposing sensitive information in notebooks. Verified References: [Databricks Certified Data Engineer Professional], under "Security and Governance" section; Databricks Documentation, under "Secrets" section.

QUESTION 4

The data engineering team has configured a job to process customer requests to be forgotten (have their data deleted). All user data that needs to be deleted is stored in Delta Lake tables using default table settings.

The team has decided to process all deletions from the previous week as a batch job at 1am each Sunday. The total duration of this job is less than one hour. Every Monday at 3am, a batch job executes a series ofVACUUMcommands on all Delta Lake tables throughout the organization.

The compliance officer has recently learned about Delta Lake\\'s time travel functionality. They are concerned that this might allow continued access to deleted data.

Assuming all delete logic is correctly implemented, which statement correctly addresses this concern?

A. Because the vacuum command permanently deletes all files containing deleted records, deleted records may be accessible with time travel for around 24 hours.

B. Because the default data retention threshold is 24 hours, data files containing deleted records will be retained until the vacuum job is run the following day.

C. Because Delta Lake time travel provides full access to the entire history of a table, deleted records can always be recreated by users with full admin privileges.

D. Because Delta Lake\\'s delete statements have ACID guarantees, deleted records will be permanently purged from all storage systems as soon as a delete job completes.

E. Because the default data retention threshold is 7 days, data files containing deleted records will be retained until the vacuum job is run 8 days later.

Correct Answer: A

Explanation: This is the correct answer because Delta Lake\\'s delete statements do not physically remove the data files that contain the deleted records, but only mark them as logically deleted in the transaction log. These files are still

accessible with time travel until they are permanently deleted by the vacuum command. The default data retention

threshold for vacuum is 7 days, but in this case it is overridden by setting it to 24 hours in each vacuum command. Therefore,

deleted records may be accessible with time travel for around 24 hours after they are deleted, until they are vacuumed. Verified References:

[Databricks Certified Data Engineer Professional], under "Delta Lake" section; [Databricks Documentation], under "Optimizations - Vacuum" section.

QUESTION 5

A Structured Streaming job deployed to production has been experiencing delays during peak hours of the day. At present, during normal execution, each microbatch of data is processed in less than 3 seconds. During peak hours of the day, execution time for each microbatch becomes very inconsistent, sometimes exceeding 30 seconds. The streaming write is currently configured with a trigger interval of 10 seconds.

Holding all other variables constant and assuming records need to be processed in less than 10 seconds, which adjustment will meet the requirement?

A. Decrease the trigger interval to 5 seconds; triggering batches more frequently allows idle executors to begin processing the next batch while longer running tasks from previous batches finish.

B. Increase the trigger interval to 30 seconds; setting the trigger interval near the maximum execution time observed for each batch is always best practice to ensure no records are dropped.

C. The trigger interval cannot be modified without modifying the checkpoint directory; to maintain the current stream state, increase the number of shuffle partitions to maximize parallelism.

D. Use the trigger once option and configure a Databricks job to execute the query every 10 seconds; this ensures all backlogged records are processed with each batch.

E. Decrease the trigger interval to 5 seconds; triggering batches more frequently may prevent records from backing up and large batches from causing spill.

Correct Answer: D

Explanation: This is the correct answer because it can meet the requirement of processing records in less than 10 seconds without modifying the checkpoint directory or dropping records. The trigger once option is a special type of trigger that runs the streaming query only once and terminates after processing all available data. This option can be useful for scenarios where you want to run streaming queries on demand or periodically, rather than continuously. By using the trigger once option and configuring a Databricks job to execute the query every 10 seconds, you can ensure that all backlogged records are processed with each batch and avoid inconsistent execution times. Verified References: [Databricks Certified Data Engineer Professional], under "Structured Streaming" section; Databricks Documentation, under "Trigger Once" section.