

[Google Discussions](#)

Exam Professional Data Engineer All Questions

View all questions & answers for the Professional Data Engineer exam

[Go to Exam](#)

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 271 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 271

Topic #: 1

[\[All Professional Data Engineer Questions\]](#)

You are monitoring your organization's data lake hosted on BigQuery. The ingestion pipelines read data from Pub/Sub and write the data into tables on BigQuery. After a new version of the ingestion pipelines is deployed, the daily stored data increased by 50%. The volumes of data in Pub/Sub remained the same and only some tables had their daily partition data size doubled. You need to investigate and fix the cause of the data increase. What should you do?

- A. 1. Check for duplicate rows in the BigQuery tables that have the daily partition data size doubled.
2. Schedule daily SQL jobs to deduplicate the affected tables.
3. Share the deduplication script with the other operational teams to reuse if this occurs to other tables.
- B. 1. Check for code errors in the deployed pipelines.
2. Check for multiple writing to pipeline BigQuery sink.
3. Check for errors in Cloud Logging during the day of the release of the new pipelines.
4. If no errors, restore the BigQuery tables to their content before the last release by using time travel.
- C. 1. Check for duplicate rows in the BigQuery tables that have the daily partition data size doubled.
2. Check the BigQuery Audit logs to find job IDs.
3. Use Cloud Monitoring to determine when the identified Dataflow jobs started and the pipeline code version.
4. When more than one pipeline ingests data into a table, stop all versions except the latest one.
- D. 1. Roll back the last deployment.
2. Restore the BigQuery tables to their content before the last release by using time travel.
3. Restart the Dataflow jobs and replay the messages by seeking the subscription to the timestamp of the release.

[Show Suggested Answer](#)

Comments

Type your comment...

Submit

  **raaad** Highly Voted  1 year, 4 months ago

Selected Answer: C

- Detailed Investigation of Logs and Jobs Checking for duplicate rows targets the potential immediate cause of the issue.
- Checking the BigQuery Audit logs helps identify which jobs might be contributing to the increased data volume.
- Using Cloud Monitoring to correlate job starts with pipeline versions helps identify if a specific version of the pipeline is responsible.
- Managing multiple versions of pipelines ensures that only the intended version is active, addressing any versioning errors that might have occurred during deployment.

=====

Why not B

While it addresses the symptom (excess data), it doesn't necessarily stop the problem from recurring. (The questions asked to investigate and fix)

   upvoted 12 times

  **mi_yulai** Most Recent  6 months ago

Why not D?


   upvoted 1 times

  **SamuelTsch** 6 months, 1 week ago

Selected Answer: B

No idea which one to choose. Option C miss a step - to restore the tables.

   upvoted 2 times

  **Ryannn23** 3 months, 1 week ago

" You need to investigate and fix the cause of the data increase. " - fixing the target tables was not required.


   upvoted 1 times

  **Matt_108** 1 year, 3 months ago

Selected Answer: C

Option C - agree with Raaad on the reasons

   upvoted 1 times

  **task_7** 1 year, 3 months ago

Selected Answer: B

B. Check for code errors in the deployed pipelines, multiple writing to pipeline BigQuery sink, errors in Cloud Logging, and if necessary, restore tables using time travel.

Check for code errors

Check for multiple writes

Check Cloud Logging

Restore tables if necessary:

   upvoted 2 times

  **RenePetersen** 1 year, 2 months ago

This does not fix the error, it basically assumes that the error is not really there.

   upvoted 3 times

Platform

- > Home
- > Examtopics PRO
- > All Exams
- > Training Courses

