← **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 176 DISCUSSION**

Actual exam question from Google's Professional Data Engineer

Question #: 176

Topic #: 1

[All Professional Data Engineer Questions]

You have uploaded 5 years of log data to Cloud Storage. A user reported that some data points in the log data are outside of their expected ranges, which indicates errors. You need to address this issue and be able to run the process again in the future while keeping the original data for compliance reasons. What should you do?

A. Import the data from Cloud Storage into BigQuery. Create a new BigQuery table, and skip the rows with errors.

B. Create a Compute Engine instance and create a new copy of the data in Cloud Storage. Skip the rows with errors.

C. Create a Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in Cloud Storage.

D. Create a Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to the same dataset in Cloud Storage.

Show Suggested Answer

by 👤 **PhuocT** at *Sept. 2, 2022, 7:15 p.m.*

Comments

Type your comment...

Submit

👤 **AWSandeep** `Highly Voted 👍` 2 years, 2 months ago

Selected Answer: C

C. Create a Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in Cloud Storage.

You can't filter out data using BQ load commands. You must imbed the logic to filter out data (i.e. time ranges) in another decoupled way (i.e. Dataflow, Cloud Functions, etc.). Therefore, A and B add additional complexity and deviates from the Data Lake design paradigm. D is wrong as the question strictly implies that the existing data set needs to be retained for compliance.

👍 🔄 🚩 upvoted 9 times

👤 **FP77** `Highly Voted 👍` 1 year, 2 months ago

Strange answers... Since when does cloud storage have datasets? Lol
Keeping this in mind, the answer must be C, but none is really correcg

👍 🔄 🚩 upvoted 5 times

👤 **ea2023** `Most Recent ⊘` 8 months, 3 weeks ago

why not D if the versioning is activated while creating your bucket ?

👍 🔄 🚩 upvoted 1 times

👤 **MaxNRG** 10 months, 2 weeks ago

Selected Answer: C

Option C is the best approach in this situation. Here is why:

Option A would remove data which may be needed for compliance reasons. Keeping the original data is preferred.
Option B makes a copy of the data but still removes potentially useful records. Additional storage costs would be incurred as well.
Option C uses Dataflow to clean the data by setting out of range values while keeping the original data intact. The fixed records are written to a new location for further analysis. This meets the requirements.
Option D writes the fixed data back to the original location, overwriting the original data. This would violate the compliance needs to keep the original data untouched.
So option C leverages Dataflow to properly clean the data while preserving the original data for compliance, at reasonable operational costs. This best achieves the stated requirements.

👍 🔄 🚩 upvoted 3 times

👤 **AzureDP900** 1 year, 10 months ago

C. Create a Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in Cloud Storage.

👍 🔄 🚩 upvoted 2 times

👤 **zellck** 1 year, 11 months ago

Selected Answer: C

C is the answer.

👍 🔄 🚩 upvoted 3 times

👤 **PhuocT** 2 years, 2 months ago

Selected Answer: C

C is correct

👍 🔄 🚩 upvoted 2 times