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 84 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 84

Topic #: 1

[All Professional Data Engineer Questions]

After migrating ETL jobs to run on BigQuery, you need to verify that the output of the migrated jobs is the same as the output of the original. You've loaded a table containing the output of the original job and want to compare the contents with output from the migrated job to show that they are identical. The tables do not contain a primary key column that would enable you to join them together for comparison.

What should you do?

A. Select random samples from the tables using the RAND() function and compare the samples.

B. Select random samples from the tables using the HASH() function and compare the samples.

C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sorting. Compare the hashes of each table.

D. Create stratified random samples using the OVER() function and compare equivalent samples from each table.

Show Suggested Answer

by rickywck at *March 17, 2020, 8:31 a.m.*

Comments

Type your comment...

Submit

rickywck Highly Voted 5 years, 1 month ago

C is the only way which all records will be compared.

upvoted 33 times

> **odacir** 2 years, 4 months ago
>
> Agree with your argument
>
> upvoted 2 times

[Removed] Highly Voted 5 years, 1 month ago

Answer: C
Description: Full comparison with this option, rest are comparison on sample which doesnot ensure all the data will be ok

upvoted 16 times

SamuelTsch Most Recent 6 months, 2 weeks ago

Selected Answer: C

Hash is always a good idea to compare the data

upvoted 1 times

midgoo 2 years, 2 months ago

In practice, I will do B. That means it may have error due to randomness. But that is how we normally do validation/QA in general, i.e. we test random samples

In this question, I will do C.

upvoted 3 times

musumusu 2 years, 2 months ago

key words here- Hash or collect value on "EACH table", after sorting the table.
Option C

upvoted 1 times

samdhimal 2 years, 3 months ago

C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sorting. Compare the hashes of each table. This approach will ensure that the data is read in a consistent order, and the hash function will provide a quick and efficient way to compare the contents of the tables and ensure that they are identical.

upvoted 1 times

> **samdhimal** 2 years, 3 months ago
>
> A. Selecting random samples from the tables using the RAND() function may not provide an accurate representation of the data and there is a risk that the comparison will not identify any differences between the tables.
>
> B. Selecting random samples from the tables using the HASH() function may not be an effective method for comparison, as the HASH() function may return different results for equivalent data.
>
> D. Creating stratified random samples using the OVER() function may not provide a comprehensive comparison between the tables as there is a risk that important differences could be missed in the sample data.
>
> upvoted 2 times

zellck 2 years, 5 months ago

Selected Answer: C

C is the answer.

upvoted 2 times

> **odacir** 2 years, 4 months ago
>
> All records need to be checked to be sure, so C is the answer
>
> upvoted 1 times

Leeeeee 2 years, 5 months ago

Selected Answer: C

All records

upvoted 1 times

hfuihe 2 years, 6 months ago

Selected Answer: B

B is the only way which all records will be compared.

upvoted 1 times

> **cloudmon** 2 years, 6 months ago

cloudmon 2 years, 6 months ago

You must have meant to say C

upvoted 2 times

medeis_jar 3 years, 3 months ago

Selected Answer: C

HASH() to compare data skipping dates and timestamps

upvoted 1 times

stefanop 3 years ago

The hash in answer C is used to select a sample of the table, not to compare them

upvoted 1 times

stefanop 3 years ago

Ignore my comment, it was about answer B.
I suggest you to go with answer C which is the only solution comparing all the rows/tables

upvoted 1 times

MaxNRG 3 years, 4 months ago

Selected Answer: C

options A B and D only will determine that it “might” be identical since is only a sample. HASH() can be helpful when doing bulk comparisons, but you still have to compare field by field to get the final answer.
The only one left is C which looks good to me

upvoted 2 times

JayZeeLee 3 years, 5 months ago

C.
The rest use RAND() at some point, which makes it hard to compare for consistency, unless there's a 'seed' option, which wasn't mentioned. So C.

upvoted 1 times

u_t_s 3 years, 7 months ago

Since there is no PK and it is possible that set of values is commons in some records which result in same hashkey for those records. But still Anwer is C

upvoted 3 times

sumanshu 3 years, 10 months ago

Vote for 'C"

upvoted 1 times

daghayeghi 4 years, 2 months ago

B:
Because said migrated to BigQuery, then we don't need Dataproc, and samples don't mean you don't compare all of data.

upvoted 3 times

yoshik 3 years, 7 months ago

a sample is a subset of data. then you should assure that the union of the samples contain the data set. Excessively complicated.
You migrate to BigQuery but need to check BigQuery output, that is why you should use another tool, Dataproc in this case.
Agree that then you should control Dataproc output but suppositions are becoming too many.

upvoted 1 times

atnafu2020 4 years, 8 months ago

C
Using Cloud Storage with big data

Cloud Storage is a key part of storing and working with Big Data on Google Cloud. Examples include:

Loading data into BigQuery.

Using Dataproc, which automatically installs the HDFS-compatible Cloud Storage connector, enabling the use of Cloud Storage buckets in parallel with HDFS.

Using a bucket to hold staging files and temporary data for Dataflow pipelines.

For Dataflow, a Cloud Storage bucket is required. For BigQuery and Dataproc, using a Cloud Storage bucket is optional but recommended.

gsutil is a command-line tool that enables you to work with Cloud Storage buckets and objects easily and robustly, in particular in big data scenarios. For example, with gsutil you can copy many files in parallel with a single command, copy large files efficiently, calculate checksums on your data, and measure performance from your local computer to Cloud Storage.

upvoted 3 times

haroldbenites 4 years, 8 months ago

C is correct

upvoted 4 times

haroldbenites 4 years, 8 months ago

It Says: "...that they are identical." , You must not use sample.

upvoted 3 times

Load full discussion...

Platform

- Home
- All Exams
- Examtopics PRO
- Training Courses

© 2024 ExamTopics