Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 99 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 99
Topic #: 1
[All Professional Data Engineer Questions]

You have a query that filters a BigQuery table using a WHERE clause on timestamp and ID columns. By using bq query `"-dry_run you learn that the query triggers a full scan of the table, even though the filter on timestamp and ID select a tiny fraction of the overall data. You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries. What should you do?

A. Create a separate table for each ID.

B. Use the LIMIT keyword to reduce the number of rows returned.

C. Recreate the table with a partitioning column and clustering column.

D. Use the bq query --maximum_bytes_billed flag to restrict the number of bytes billed.

Show Suggested Answer

by rickywck at *March 17, 2020, 10:05 a.m.*

Comments

Type your comment...

Submit

rickywck Highly Voted 3 years, 1 month ago

should be C:

https://cloud.google.com/bigquery/docs/best-practices-costs

upvoted 43 times

[Removed] Highly Voted 3 years, 1 month ago

Correct - C

upvoted 17 times

zellck Most Recent 5 months ago

Selected Answer: C

C is the answer.

https://cloud.google.com/bigquery/docs/partitioned-tables
A partitioned table is a special table that is divided into segments, called partitions, that make it easier to manage and query your data. By dividing a large table into smaller partitions, you can improve query performance, and you can control costs by reducing the number of bytes read by a query.

https://cloud.google.com/bigquery/docs/clustered-tables
lustered tables in BigQuery are tables that have a user-defined column sort order using clustered columns. Clustered tables can improve query performance and reduce query costs.

upvoted 4 times

Fezo 10 months ago

Selected Answer: C

C is the answer
https://cloud.google.com/bigquery/docs/best-practices-costs

upvoted 3 times

medeis_jar 1 year, 4 months ago

Selected Answer: C

C only make sense

upvoted 2 times

MaxNRG 1 year, 4 months ago

Selected Answer: C

https://cloud.google.com/bigquery/docs/best-practices-costs
Applying a LIMIT clause to a SELECT * query does not affect the amount of data read. You are billed for reading all bytes in the entire table, and the query counts against your free tier quota.
A and D doesnt make sense
Its C, when you want to select by a partition you should write something like:
CREATE TABLE `blablabla.partitioned`
PARTITION BY
DATE(timestamp)
CLUSTER BY id
AS
SELECT * FROM `blablabla`

upvoted 5 times

Anilgcp980 1 year, 4 months ago

this is a trap to make people fail by giving wrong answer as B.

upvoted 3 times

snadaf 1 year, 5 months ago

It's D, here is the link
https://cloud.google.com/bigquery/docs/best-practices-costs

upvoted 1 times

maurodipa 1 year, 5 months ago

Well, you mean C, isn't it?

upvoted 1 times

Crudgey 1 year, 5 months ago

Are they having a laugh at us by giving so many bad answers?

upvoted 5 times

tsoetan001 1 year, 6 months ago

Answer: B
Note: minimal change to sql

upvoted 1 times

szefco 1 year, 6 months ago

Not B. LIMIT will not reduce amount of data scanned - only limit the final output, but you will still be billed for scanning whole table.
C is correct. After applying partitioning ans clustering amount of bytes scanned will decrease

upvoted 3 times

Ysance_AGS 1 year, 7 months ago

"You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries" that doesn't mean that you can create or edit existing tables ! you only can edit the SQL query !!! so answer D is the correct one.

upvoted 2 times

szefco 1 year, 6 months ago

I don't agree. Question says "minimal changes to existing SQL queries" - if you recreate table with partitioning and clustering you don't need to change SQLs that read from that table.
C is correct answer here.

upvoted 1 times

squishy_fishy 1 year, 6 months ago

D would just block your query. The answer is C.

upvoted 1 times

nguyenmoon 1 year, 7 months ago

C - create partition table

upvoted 2 times

sumanshu 1 year, 10 months ago

Vote for C

upvoted 4 times

felixwtf 2 years, 4 months ago

LIMIT keyword is applied only at the end, i.e., only to limit the results already calculated. Therefore, a full table scan will have already happened. The where clause on the other hand would provide the desired filtering depending on the case. So, C is the correct answer.

upvoted 4 times

learnazureportal 2 years, 5 months ago

Not sure, why option C selected! The correct Answer is B. the question clearly says "minimal changes to existing SQL queries". who said that, recreate the table, with partitioning layout is minimal and is PART of SQL queries!

upvoted 2 times

hdmi_switch 1 year, 9 months ago

In addition to the previous reply, the LIMIT statement applies to the output (what you see in the UI), the full table scan will still happen. C is correct according to best practices.

upvoted 1 times

ceak 2 years, 5 months ago

recreating table will not affect existing sql queries as they will still be selecting the same table name, but the scan will hugely decrease. so, option C is the correct answer.

upvoted 5 times

squishy_fishy 1 year, 6 months ago

Recreating table is recommended by Google.

upvoted 1 times

arghya13 2 years, 5 months ago

should be C:

upvoted 2 times

gyclop 2 years, 7 months ago

Correct - C :
"Limit" keyword restricts the final dataset to "n" rows, but is not able to restrict full table scan

upvoted 3 times

Load full discussion...