## 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 53 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 53

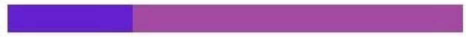
Topic #: 1

[All Professional Data Engineer Questions]

You are using Google BigQuery as your data warehouse. Your users report that the following simple query is running very slowly, no matter when they run the query:

SELECT country, state, city FROM [myproject:mydataset.mytable] GROUP BY country

You check the query plan for the query and see the following output in the Read section of Stage:1:



What is the most likely cause of the delay for this query?

A. Users are running too many concurrent queries in the system

B. The [myproject:mydataset.mytable] table has too many partitions

C. Either the state or the city columns in the [myproject:mydataset.mytable] table have too many NULL values

D. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew

Show Suggested Answer

by [deleted] at *March 21, 2020, 9:32 a.m.*

### Comments

Type your comment...

Submit

**[Removed]** Highly Voted 5 years, 1 month ago

Should be D

upvoted 26 times

**itche_scratche** Highly Voted 5 years ago

D; Purple is reading, Blue is writing. so majority is reading.

upvoted 25 times

> **squishy_fishy** 3 years, 6 months ago
>
> I have been looking for the color code descriptions for a while. Thank you!
>
> upvoted 1 times

**iooj** Most Recent 9 months, 1 week ago

Selected Answer: D

D - stands for Data Skew
The Read section of the query plan shows a heavy concentration of processing in one area (as indicated by the pink bar being much longer than the purple bar). This typically indicates data skew, where the majority of the data is processed by a small subset of nodes.

upvoted 1 times

**MaxNRG** 1 year, 4 months ago

Selected Answer: D

The most likely cause of the delay for this query is option D. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew.

Group by queries in BigQuery can run slowly when there is significant data skew on the grouped columns. Since the query is grouping by country, if most rows have the same country value, all that data will need to be shuffled to a single reducer to perform the aggregation. This can cause a data skew slowdown.

Options A and B might cause general slowness but are unlikely to affect this specific grouping query. Option C could also cause some slowness but not to the degree that heavy data skew on the grouped column could. So D is the most likely root cause. Optimizing the data distribution to reduce skew on the grouped column would likely speed up this query.

upvoted 2 times

**JOKKUNO** 1 year, 4 months ago

Data skew is when one or some partitions have significantly more data compared to other partitions. Data-skew is usually the result of operations that require re-partitioning the data, mostly join and grouping ( GroupBy ) operations. So D.

upvoted 1 times

**PolyMoe** 2 years, 3 months ago

Selected Answer: D

D. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew

Data skew occurs when one or more values in a column have a disproportionately large number of rows compared to other values in that column. This can cause performance issues when running queries that group by that column, like the one in the question. In this case, if most of the rows in the [myproject:mydataset.mytable] table have the same value in the country column, then the query will need to process a large number of rows with that value, which can cause significant delay.

upvoted 4 times

**AzureDP900** 2 years, 4 months ago

D is right

upvoted 3 times

**Krish6488** 2 years, 4 months ago

Selected Answer: D

data skewing causing imbalance in data distribution across slots. It also causes errors if the group by column has NULLS. Since option C does not call out the Group by column, D is a closer answer contextually

upvoted 2 times

**Jasar** 2 years, 5 months ago

Selected Answer: A

A is the best option becouse the color bar show the high number of reads and i think its not a skew becouse biguery was build to compute the data fast

upvoted 3 times

**arpitagrawal** 2 years, 8 months ago

The query would throw the error because you're using a group by clause on country but not aggregating city or state.

upvoted 11 times

**MisuLava** 2 years, 8 months ago

**Selected Answer: D**

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

upvoted 3 times

**Paul_Oprea** 3 years ago

BTW, how is the query even syntactically valid? It has non aggregated columns in the SELECT part of the query. That query will not run in the first place, unless I'm missing something.

upvoted 18 times

**Arkon88** 3 years, 2 months ago

**Selected Answer: D**

D
Image says that average(dark) and maximum(light) have difference in few times, this it is a skew

https://cloud.google.com/bigquery/query-plan-explanation
The color indicators show the relative timings for all steps across all stages. For example, the COMPUTE step of Stage 00 shows a bar whose shaded fraction is 21/30 since 30ms is the maximum time spent in a single step of any stage. The parallel input information shows that each stage required only a single worker, so there's no variance between average and slowest timings.

upvoted 1 times

**sraakesh95** 3 years, 3 months ago

**Selected Answer: D**

https://medium.com/slalom-build/using-bigquery-execution-plans-to-improve-query-performance-af141b0cc33d

upvoted 3 times

**sraakesh95** 3 years, 3 months ago

https://medium.com/slalom-build/using-bigquery-execution-plans-to-improve-query-performance-af141b0cc33d

upvoted 1 times

**medeis_jar** 3 years, 4 months ago

**Selected Answer: D**

D
Colors: Purple is reading, Blue is writing. so the majority is reading.
https://cloud.google.com/bigquery/docs/best-practices-performance-patterns

upvoted 5 times

**morpho4444** 3 years, 5 months ago

If you read this https://medium.com/slalom-build/using-bigquery-execution-plans-to-improve-query-performance-af141b0cc33d C can't be right because the skewness happen when the column you use for grouping contains lots of null values, here C mentions columns that aren't part of the grouping clause.

D, that's not how data get skewed, it gets skewed due to null values.

A is the only answer here.

upvoted 1 times

> **BigQuery** 3 years, 5 months ago
>
> A Cant be answer. Since users whenever running queries facing the problems.
>
> upvoted 1 times

Load full discussion...

- Examtopics PRO
- Training Courses