

[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 123 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 123

Topic #: 1

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

You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis.

Every hour, thousands of transactions are updated with a new status. The size of the initial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? (Choose two.)

- A. Denormalize the data as much as possible.
- B. Preserve the structure of the data as much as possible.
- C. Use BigQuery UPDATE to further reduce the size of the dataset.
- D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.
- E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro file. Use BigQuery's support for external data sources to query.

[Show Suggested Answer](#)

by  rickywck at March 20, 2020, 4:21 a.m.

Comments

[Submit](#)

rickywck Highly Voted 3 years, 7 months ago

I think AD is the answer. E will not improve performance.

upvoted 40 times

[Removed] Highly Voted 3 years, 7 months ago

Answer: A, D

Description: Denormalization will help in performance by reducing query time, update are not good with bigquery

upvoted 20 times

awssp12345 2 years, 3 months ago

My guess is append has better performance than update.

upvoted 3 times

midgoo Most Recent 7 months, 2 weeks ago

Selected Answer: BD

If we denormalize the data, the Data Science team will shout at us. Preserving it is the way to go

upvoted 3 times

vaga1 4 months, 4 weeks ago

Denormalization is just a best practice when using BQ.

upvoted 1 times

WillemHendr 5 months ago

Shouting data-science teams are not part of question, this is more about what is exam correct, not what it the best for your own situation

upvoted 5 times

odacir 11 months ago

Selected Answer: AD

A and D:

A- Improve performance

D- Is better for DS have all the history and not the last update...

upvoted 7 times

zelck 11 months ago

Selected Answer: AD

AD is the answer.

<https://cloud.google.com/bigquery/docs/best-practices-performance-nested>

Best practice: Use nested and repeated fields to denormalize data storage and increase query performance.

Denormalization is a common strategy for increasing read performance for relational datasets that were previously normalized. The recommended way to denormalize data in BigQuery is to use nested and repeated fields. It's best to use this strategy when the relationships are hierarchical and frequently queried together, such as in parent-child relationships.

upvoted 4 times

AzureDP900 10 months, 1 week ago

A, C is correct I agree

upvoted 1 times

NicolasN 1 year ago

The criteria for selecting a strategy are the performance and usability for the data science team. This team performs the analysis by querying stored data. So we don't care for performance related with data ingestion. According to this point of view:

A: YES - undisputedly favours query performance

B: YES - Keeping the structure unchanged promotes usability (the team won't need to update queries or ML models)

C: Questionable - Updating the status of a row instead of appending newer versions is keeping the size smaller. But does this affect significantly the analysis performance? Even if it does, creating materialized views to keep the most recent status per row eliminates it

D: NO - has nothing to do with DS team's tasks, affects ingestion performance

E: NO - demotes usability

upvoted 2 times

jkhong 10 months, 3 weeks ago

For B there is no mention that the current data structure is being used (...data science team WILL build machine learning models based on this data.) ... We're developing a new data model to be used by them in the future

👍 ↩ 🚩 upvoted 1 times

🗄 👤 **NicolasN** 1 year ago
(mistakenly voted AC instead of AB)

👍 ↩ 🚩 upvoted 1 times

🗄 👤 **DerickTW** 1 year, 1 month ago

Selected Answer: AC

The DML quota limit is removed since 2020, I think C is better than D now.

👍 ↩ 🚩 upvoted 1 times

🗄 👤 **devaid** 1 year ago

Is not about the quota. You should avoid using UPDATE because it makes a big scan of the table, and is not efficient or high performant. Usually prefer appends and merges instead, and using the optimized schema approach of Big Query that denormalizes the table to avoid joins and leverages nested and repeated fields.

👍 ↩ 🚩 upvoted 5 times

🗄 👤 **MaxNRG** 1 year, 10 months ago

Selected Answer: AD

A: Denormalization increases query speed for tables with billions of rows because BigQuery's performance degrades when doing JOINS on large tables, but with a denormalized data structure, you don't have to use JOINS, since all of the data has been combined into one table.

Denormalization also makes queries simpler because you do not have to use JOIN clauses.

https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data

D: BigQuery append

👍 ↩ 🚩 upvoted 4 times

🗄 👤 **medeis_jar** 1 year, 10 months ago

Selected Answer: AD

requirements are -> performance and usability.

Denormalization will help in performance by reducing query time, update is not good with big query.

And append has better performance than Update.

👍 ↩ 🚩 upvoted 3 times

🗄 👤 **doninakula** 1 year, 11 months ago

I think AD. E is not valid because it use external table which is not good for performance

👍 ↩ 🚩 upvoted 1 times

🗄 👤 **sumanshu** 2 years, 4 months ago

A - correct (denormalization will help)

B - data already heavily structured (no use and no impact)

C - more than 1500 Updates not possible

D - Not sure..(because appending will increase size and cost)

E - Does not look good (increase cost..also we are storing for all days....again for query we need to issue mutiple query for all days....)

So, A & D (left out of 5)

👍 ↩ 🚩 upvoted 5 times

🗄 👤 **Jeysolomon** 2 years, 4 months ago

Correct Answer: AE

A – Denormalisation helps improve performance.

B, C - Not helping to address the problem.

D – Append will increase the db size and cost involved for storage and also for large number of records to scan for queries by data science team which is costlier.

E - Addresses the problem of maximising the usability of the data science team and the data. They can analyse the data exported to cloud storage instead of reading from bigquery which is expensive and impact performance considerably.

👍 ↩ 🚩 upvoted 3 times

🗄 👤 **Chelseajcole** 2 years ago

It didn't mention cost is a concern

👍 ↩ 🚩 upvoted 1 times

🗄 👤 **retep007** 2 years, 1 month ago

E is wrong, you've been asked to use bigquery and reading files from storage in bq is significantly more time consuming

👍 ↩ 🚩 upvoted 2 times

🗄 👤 **daghayeghi** 2 years, 7 months ago

A D.

A, D.

Using BigQuery as an OLTP store is considered an anti-pattern. Because OLTP stores have a high volume of updates and deletes, they are a mismatch for the data warehouse use case. To decide which storage option best fits your use case, review the Cloud storage products table.

BigQuery is built for scale and can scale out as the size of the warehouse grows, so there is no need to delete older data. By keeping the entire history, you can deliver more insight on your business. If the storage cost is a concern, you can take advantage of BigQuery's long term storage pricing by archiving older data and using it for special analysis when the need arises. If you still have good reasons for dropping older data, you can use BigQuery's native support for date-partitioned tables and partition expiration. In other words, BigQuery can automatically delete older data.

https://cloud.google.com/solutions/bigquery-data-warehouse#handling_change

   upvoted 4 times

  **Hithesh** 2 years, 8 months ago

should be AC.. "Every hour, thousands of transactions are updated with a new status" if we append how we will handle the new status change..

   upvoted 2 times

  **sumanshu** 2 years, 4 months ago

C not possible, maximum 1500 updates possible in a day

   upvoted 1 times

  **raf2121** 2 years, 3 months ago

DML without limits now in BQ (below blog says March 2020, Not sure whether these questions were prepared before or after March 2020)

<https://cloud.google.com/blog/products/data-analytics/dml-without-limits-now-in-bigquery>

   upvoted 1 times

  **hdmi_switch** 2 years, 3 months ago

There is no more hard limit, but UPDATES are queued:

"BigQuery runs up to 2 of them concurrently, after which up to 20 are queued as PENDING. When a previously running job finishes, the next pending job is dequeued and run. Currently, queued mutating DML statements share a per-table queue with maximum length 20. Additional statements past the maximum queue length for each table fail."

With thousands of updates per hour, this doesn't seem feasible. I would assume the question is marked as outdated anyway or the answers are update in the actual exam.



   upvoted 4 times

  **daghayeghi** 2 years, 8 months ago

AC:

the problem is exactly about Updating and preserving size of database as much as possible, then denormalization and using UPDATE function from DML will address the issue. they don't want to update faster. then A & C is correct.

<https://cloud.google.com/solutions/bigquery-data-warehouse>

   upvoted 1 times

  **karthik89** 2 years, 8 months ago

you can update bigquery 1500 times in a day

   upvoted 3 times

  **daghayeghi** 2 years, 7 months ago

A, D:

it was my mistake, we should decrease update as Bigquery is not design for update.

https://cloud.google.com/solutions/bigquery-data-warehouse#handling_change

   upvoted 3 times

  **Nams_139** 2 years, 11 months ago

A,D Since the requirements are both performance and usability.

   upvoted 5 times

  **federicohi** 2 years, 11 months ago

i tink may be ita AC becuase appending its worst to increase dataset size. THe question seems to put like a problem the size of dataset and performance to datascience so inserting more rows decrease performance for them.

   upvoted 3 times

[Load full discussion...](#)



Platform

> Home

> Examtopics PRO

> All Exams

> Training Courses

