

[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 43 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 43

Topic #: 1

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

You work for a large fast food restaurant chain with over 400,000 employees. You store employee information in Google BigQuery in a Users table consisting of a FirstName field and a LastName field. A member of IT is building an application and asks you to modify the schema and data in BigQuery so the application can query a FullName field consisting of the value of the FirstName field concatenated with a space, followed by the value of the LastName field for each employee. How can you make that data available while minimizing cost?

- A. Create a view in BigQuery that concatenates the FirstName and LastName field values to produce the FullName.
- B. Add a new column called FullName to the Users table. Run an UPDATE statement that updates the FullName column for each user with the concatenation of the FirstName and LastName values.
- C. Create a Google Cloud Dataflow job that queries BigQuery for the entire Users table, concatenates the FirstName value and LastName value for each user, and loads the proper values for FirstName, LastName, and FullName into a new table in BigQuery.
- D. Use BigQuery to export the data for the table to a CSV file. Create a Google Cloud Dataproc job to process the CSV file and output a new CSV file containing the proper values for FirstName, LastName and FullName. Run a BigQuery load job to load the new CSV file into BigQuery.

[Show Suggested Answer](#)

by [rickywck](#) at March 17, 2020, 4:39 a.m.

## Comments

Type your comment...

Submit

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


Answer will be A because when you create View it does not store extra space and its a logical representation, for rest of the option you need to write large code and extra processing for dataflow/dataproc

   upvoted 68 times

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

Because views are not materialized, the query that defines the view is run each time the view is queried. Queries are billed according to the total amount of data in all table fields referenced directly or indirectly by the top-level query

   upvoted 12 times

  **cloud\_rider** 5 months, 1 week ago

This was try in Oracle era, BQ prunes the query before running, so having a view as an intermediate layer does not have any impact, unless there is a heavy filtering happening within the view definition.

   upvoted 1 times

  **Igdantas** 4 years, 9 months ago

Wouldn't "total amount of data in all table fields referenced directly or indirectly by the top-level query" be FirstName and LastName?

   upvoted 3 times

  **lollo1234** 3 years, 11 months ago



You're right, BigQuery bills on number of bytes processed, regardless of them being materialized. If you don't create a new column and use a view instead, you will probably have a small performance hit but query costs would be the same and storage cost wouldn't increase (unlike storing a new column)

   upvoted 4 times

  **yoshik** 3 years, 7 months ago

You are asked to modify the schema and data. By using a view, the underlined table remains intact.

   upvoted 12 times

  **HarshKothari21** 2 years, 7 months ago

good catch, yoshik.

  upvoted 1 times

  **alecuba16** 2 years, 12 months ago

Views are cached the same as regular tables are, so I don't get the point of billing. It will cost the same as query to a regular table.

   upvoted 3 times

  **ovokpus** 2 years, 5 months ago

the point of billing is extra storage costs for a new concatenated column

   upvoted 3 times

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

Can't be A

   upvoted 5 times

  **beowulf\_kat** 2 years, 6 months ago

I agree that A is correct. Also, I think B is wrong as the UPDATE statement is used to update values in existing columns, not to create a new column.

   upvoted 2 times

  **ovokpus** 2 years, 5 months ago


Of course, you use UPDATE after creating the new column, that is what the option said

   upvoted 2 times

  **[Removed]** 2 years, 2 months ago

What happen if there are new employees joining the company, update every single time?

   upvoted 1 times

  **funtoosh** 4 years, 2 months ago

cannot be 'A' as it clearly says that you need to change the schema and data.

commenter has clearly says that you need to change the schema and data.

👍 ↩ 🚩 upvoted 17 times

🗄 👤 **exnaniantwort** 3 years, 3 months ago

Your primary task is to "make data available".

Changing the schema is just the request from the member "A member of IT is building an application and \*\*\*asks you to modify the schema and data\*\*\* in BigQuery". You don't have to follow it if it does not make sense.

👍 ↩ 🚩 upvoted 15 times

🗄 👤 **YorelNation** 2 years, 8 months ago

A yes, That make a lot of sense and also if you update the table only once with UPDATE if there is a new employee it will not be up to date with the new column, if the app use a view it will be up to date every time it query. But in any case the cost will not be minimized.

👍 ↩ 🚩 upvoted 4 times

🗄 👤 **exnaniantwort** 3 years, 3 months ago

There is always different application requirement to use different format. That way you will just creating more and more redundant columns in different formats. That is tedious.

👍 ↩ 🚩 upvoted 6 times

🗄 👤 **BhupiSG** Highly Voted 4 years, 1 month ago

Correct: B

BigQuery has no quota on the DML statements. (Search Google - does bigquery have quota for update).

Why not C: This is a one time activity and SQL is the easiest way to program it. DataFlow is way overkill for this. You will need to find an engineer who can develop DataFlow pipelines. Whereas, SQL is so much more widely known and easier. One of the great features about BigQuery is its SQL interface. Even for BigQueryML services.

👍 ↩ 🚩 upvoted 47 times

🗄 👤 **lollo1234** 3 years, 11 months ago

DML statements don't increase costs, but storing a new column does. I see A is correct (also see my comment above)

👍 ↩ 🚩 upvoted 5 times

🗄 👤 **exnaniantwort** 3 years, 3 months ago

Exactly. Cost is the reason to reject B.

How come so many people vote for this wrong option?

👍 ↩ 🚩 upvoted 3 times

🗄 👤 **ler\_mp** 2 years, 4 months ago

Storage is cheap compared to computation

👍 ↩ 🚩 upvoted 8 times

🗄 👤 **DGames** 2 years, 4 months ago

But you need to maintain table means regularly you have to execute the update query whenever new data comes.

👍 ↩ 🚩 upvoted 2 times

🗄 👤 **lollo1234** 3 years, 11 months ago

I will also add that B would imply changing upstream workloads to write the new field every time a records gets added

👍 ↩ 🚩 upvoted 8 times

🗄 👤 **willyunger** Most Recent 1 month, 2 weeks ago

Selected Answer: A

Minimal cost: no extra space, no cost to set up, no need to write code, rest of applications see no change, no need to offload/reprocess/reload (although batch load is free).

👍 ↩ 🚩 upvoted 1 times

🗄 👤 **LP\_PDE** 3 months, 1 week ago

Selected Answer: B

I would say A but since it specifically says "modify" then the answer is B.

👍 ↩ 🚩 upvoted 1 times

🗄 👤 **iooj** 9 months, 1 week ago

E. Say to the IT specialist to take care of it on the app side...

B would work for historical data if we had an underlying change made to automate the concatenation for new records. It is not clear, so I would say A is a quick solution.

👍 ↩ 🚩 upvoted 2 times

🗄 👤 **Ramanaiah** 1 year ago

Selected Answer: B

Requirement is to be able to filter on full name. So, you would be querying all data unless you have materialized full name column.

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **philli1011** 1 year, 3 months ago

Definitely A

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **axantroff** 1 year, 5 months ago

**Selected Answer: A**

The question might be outdated, but I would like to offer my perspective:

1. Ideally, I would opt for a materialized view to avoid updating pipelines
2. In 2023, I see no concerns regarding the costs involved in storing denormalized data for analytical needs
3. Regarding this question I would choose option A, although the concern about extra costs due to recalculations is valid for me

👍 ↩ 🚩 upvoted 2 times

🗨️ 👤 **LaxmanTiwari** 1 year, 4 months ago

Did u pass the exam ?

👍 ↩ 🚩 upvoted 2 times

🗨️ 👤 **steghe** 1 year, 5 months ago

Answer should be A 'cos the First request is: make that data available.

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **alihabib** 1 year, 9 months ago

Its A ..... "asked to change schema" is a trick to test your skills. Better to make use of MV's if anyhow the application is gonna query repeatedly. MV's will rebuild itself, if query invalidates from cache results

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **nescafe7** 1 year, 9 months ago

**Selected Answer: A**

In the case of B, the data pipeline that adds new employee information must also be modified, which is not the correct answer in terms of cost minimization.

👍 ↩ 🚩 upvoted 2 times

🗨️ 👤 **Mathew106** 1 year, 9 months ago

**Selected Answer: A**

It's A. If you add a column to the table, you will be billed every time you query that new column. The same way you would be billed with the view created by A.

B,C and D create a new column. A does not create a new column. It just provides the interface for the application to access the data. B,C and D will have to be rerun to compute the column value of new customers.

A is done only once, costs 0 for storage, and is charged about the same as all the others when it comes to compute because even if you choose B C and D you would have to query the data in the end anyway.

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **autumn2005** 1 year, 9 months ago

**Selected Answer: C**

modify the schema

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **theseawillclaim** 1 year, 9 months ago

Can you code a script for a BQ Column? I don't think it's "B", it is pretty tricky

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **KC\_go\_reply** 1 year, 10 months ago

**Selected Answer: A**

Everything but A) new view is wrong.

B) sounds okay, but introduces a new column which means more storage, thus increasing cost.

C) Dataflow is obvious overkill for a simple task such as concatenating two strings.

D) Starting up a Dataproc cluster just for string concatenation is super overkill.

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **vaga1** 1 year, 11 months ago

**Selected Answer: A**

if a new field is only necessary for one project, and it is only the concatenation of two existing fields, it is ok to create a view that gets used for a specific task.

👍 ↻ 🚩 upvoted 1 times

🗨️ 👤 Jarek7 2 years ago

**Selected Answer: A**

I'd go for A.

The main issue with answers B,C,D is that they are just temporary solution. Whenever a new employee comes in (there are 400.000 of them at the moment, so we can expect every day a few new guys) we need to update the fullname table/field again. Additionally each of these answers need twice as much capacity (BigQuery stores data in a columnar format, so optimizing is not possible). Although the price for the needed capacity will be far below 0.01\$/month.

The main argument against A is that compute power costs more than capacity. Please look how BQ is priced:

[https://cloud.google.com/bigquery/pricing#query\\_pricing](https://cloud.google.com/bigquery/pricing#query_pricing)

In the default On-demand compute pricing it is charged for "the number of bytes processed by each query" so there will be no any difference in computing costs for any option.

Yeah, there is also this argument about modifying schema in the requirements. Lets be professional - it is not a requirement for OUR schema. If you can resolve the issue with 0 change to YOUR schema then you are more than ok. And anyway, from requestor point of view, the schema HE uses in his app will be modified as he needed.

👍 ↻ 🚩 upvoted 2 times

[Load full discussion...](#)



## Platform

> Home

> All Exams

> Examtopics PRO

> Training Courses



© 2024 ExamTopics