

[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 46 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 46

Topic #: 1

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

You work for an economic consulting firm that helps companies identify economic trends as they happen. As part of your analysis, you use Google BigQuery to correlate customer data with the average prices of the 100 most common goods sold, including bread, gasoline, milk, and others. The average prices of these goods are updated every 30 minutes. You want to make sure this data stays up to date so you can combine it with other data in BigQuery as cheaply as possible.

What should you do?

- A. Load the data every 30 minutes into a new partitioned table in BigQuery.
- B. Store and update the data in a regional Google Cloud Storage bucket and create a federated data source in BigQuery
- C. Store the data in Google Cloud Datastore. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Cloud Datastore
- D. Store the data in a file in a regional Google Cloud Storage bucket. Use Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.




[Show Suggested Answer](#)

by  mmarulli at March 11, 2020, 2:31 p.m.

## Comments


Type your comment...

[Submit](#)

  **mmarulli** Highly Voted  5 years, 1 month ago


this is one of the sample exam questions that google has on their website. The correct answer is B

   upvoted 43 times

  **nadavw** 8 months, 3 weeks ago

B - since it seems that not all data is in BigQuery but the analysis is done using BigQuery so federated query is the optimal approach

   upvoted 2 times

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

Answer: B

Description: B is correct because regional storage is cheaper than BigQuery storage.

   upvoted 13 times

  **funtoosh** 4 years, 2 months ago

it's not only cheaper but the requirement is that the data keep updating every 30 min and you need to combine the data in bigquery, use external tables to do that is the recommended practice

   upvoted 9 times

  **jatinbhatia2055** Most Recent  4 months, 3 weeks ago

**Selected Answer: A**

BigQuery is a powerful data warehouse designed for analyzing large datasets efficiently. Partitioning tables allows you to manage large datasets by splitting them into segments based on a key, such as time.

By creating a partitioned table and updating it every 30 minutes, you can load the new price data directly into the correct partitions. BigQuery's partitioned tables optimize both the storage and querying cost because BigQuery only scans the relevant partitions when querying, minimizing the amount of data read and hence reducing costs.

Partitioning by time (e.g., timestamp or date columns) is particularly effective for datasets with periodic updates (like price data) since each batch of data will be loaded into the corresponding partition.

   upvoted 2 times

  **SamuelTsch** 6 months, 2 weeks ago

**Selected Answer: B**

Actually, in this question, I think B is the most suitable. C, D are somehow overkill. A due to the minimum partition granularity. However, with B, the data could not be previewd also it is not possible to estimate the cost.

   upvoted 2 times

  **Pennepal** 1 year ago

D. Store the data in a file in a regional Google Cloud Storage bucket. Use Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

Here's why this approach is ideal:

**Cost-Effective Storage:** Cloud Storage offers regional storage classes that are cost-effective for frequently accessed data. Storing the price data in a regional Cloud Storage bucket keeps it readily available.

**Cloud Dataflow for Updates:** Cloud Dataflow is a managed service for building data pipelines. You can create a Dataflow job that runs every 30 minutes to:

Download the latest economic data file from Cloud Storage.

Process and potentially transform the data as needed.

Load the updated data into BigQuery.

**BigQuery Integration:** BigQuery seamlessly integrates with Cloud Dataflow. The Dataflow job can directly load the processed data into a BigQuery table for further analysis with your customer data.

   upvoted 2 times

  **TVH\_Data\_Engineer** 1 year, 4 months ago

**Selected Answer: A**

BigQuery supports partitioned tables, which allow for efficient querying and management of large datasets that are updated frequently. By loading the updated data into a new partition every 30 minutes, you can ensure that only relevant partitions are queried, reducing the amount of data processed and thereby minimizing costs.

What's wrong with B ? While creating a federated data source in BigQuery pointing to a Google Cloud Storage bucket is feasible, it might not be the most efficient for data that is updated every 30 minutes. Querying federated data sources can sometimes be more expensive and less performant than querying data stored directly in BigQuery.



   upvoted 3 times

  **Melampos** 2 years ago

**Selected Answer: D**

Federated queries let you send a query statement to Cloud Spanner or Cloud SQL databases not to cloud storage

   upvoted 1 times

  **sid\_is\_dis** 1 year, 10 months ago

Is you are right about "federated queries", but the option B says about "federated data source". These are different concepts

   upvoted 3 times

  **Abhilash\_pendyala** 2 years ago

ChatGPT says partitioned tables is the best approach, The answers here are quite contrasting with that answer, Even i thought it has to be option A, I am so confused now? Any proper straight forward answer ?

   upvoted 1 times

  **musumusu** 2 years, 2 months ago

Answer B:

Uploading data into staging tables/ external tables or federated source in BQ is the best approach.

Option A is also good approach, anyone can explain about his part what is wrong about this?

   upvoted 1 times

  **yoga9993** 2 years, 2 months ago

we can't implement A, it's because biquery partition table can only be done minimum in range 1 hour, the requirement said it must be update every 30 minutes, so A is impossible option as the minimum partition is in hour level

   upvoted 7 times

  **AzureDP900** 2 years, 4 months ago

B is right

   upvoted 1 times

  **Krish6488** 2 years, 4 months ago


**Selected Answer: B**

Discounting A due to limitations on partitions

Discounting C because datastore does not fit into the nature of data we are talking about and federation between BQ and datastore it an overkill

Between B and D, updating the price file on GCS and joining BQ tables and external tables sourcing data from GCS is most cost optimal way for this use case

   upvoted 2 times

  **ler\_mp** 2 years, 3 months ago

D is also overkill for this use case, so I'd pick B

   upvoted 1 times

  **jkhong** 2 years, 4 months ago

**Selected Answer: B**

Consideration: As cheaply as possible. Make sure data stays up to date.

Initially chose A. But in actuality there is no need to maintain or store past data so storage of past data and partitioning doesn't seem like a key requirement.

Instead we can connect just to a single Cloud Storage file, either by:

- replace previous prices with latest prices
- store previous prices in GCS if required to be retained



   upvoted 1 times

  **DGames** 2 years, 4 months ago

**Selected Answer: B**

B is most inexpensive approach.

   upvoted 1 times

  **odacir** 2 years, 4 months ago

**Selected Answer: B**


The technical requirement is having frequently access info to join with other BQ data, as cheap as possible. B fits perfectly.

Corner cases for external data sources:

- Avoiding duplicate data in BigQuery storage
- Queries that do not have strong performance requirements
- Small amount of frequently changing data to join with other tables in BigQuery

<https://cloud.google.com/blog/products/gcp/accessing-external-federated-data-sources-with-bigquerys-data-access-layer>

   upvoted 2 times

  **assU2** 2 years, 5 months ago

**Selected Answer: D**

I would say D, regional Google Cloud Storage bucket - cheap.

A - not cheap

B - NoSQL database for your web and mobile applications

C - Federated queries let you send a query statement to Cloud Spanner or Cloud SQL databases

And we need to combine data in DQ with data from bucket

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **MisuLava** 2 years, 6 months ago

according to this :

<https://cloud.google.com/bigquery/docs/external-data-sources>

Federated queries don't work with Cloud Storage.

how can it be B ?

👍 ↩ 🚩 upvoted 2 times

🗨️ 👤 **cloudmon** 2 years, 6 months ago

Correct, it cannot be B because BQ federated queries only work with Cloud SQL or Spanner

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **gudiking** 2 years, 5 months ago

It seems to me that they do: <https://cloud.google.com/bigquery/docs/external-data-cloud-storage>

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **ducc** 2 years, 8 months ago

**Selected Answer: B**

I voted for B

👍 ↩ 🚩 upvoted 2 times

[Load full discussion...](#)



## Platform

> Home

> All Exams

> Examtopics PRO

> Training Courses



© 2024 ExamTopics