

[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 72 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 72

Topic #: 1

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

You are designing storage for 20 TB of text files as part of deploying a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying aggregate values for multiple users who will query the data in Cloud Storage with multiple engines. Which storage service and schema design should you use?

- A. Use Cloud Bigtable for storage. Install the HBase shell on a Compute Engine instance to query the Cloud Bigtable data.
- B. Use Cloud Bigtable for storage. Link as permanent tables in BigQuery for query.
- C. Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.
- D. Use Cloud Storage for storage. Link as temporary tables in BigQuery for query.

[Show Suggested Answer](#)

by [rickywck](#) at March 17, 2020, 8:06 a.m.

## Comments

Type your comment...

[Submit](#)

[daghayeghi](#) [Highly Voted](#) 3 years, 1 month ago

answer C:

BigQuery can access data in external sources, known as federated sources. Instead of first

BigQuery can access data in external sources, known as *external sources*. Instead of first loading data into BigQuery, you can create a reference to an external source. External sources can be Cloud Bigtable, Cloud Storage, and Google Drive.

When accessing external data, you can create either permanent or temporary external tables. Permanent tables are those that are created in a dataset and linked to an external source. Dataset-level access controls can be applied to these tables. When you are using a temporary table, a table is created in a special dataset and will be available for approximately 24 hours. Temporary tables are useful for one-time operations, such as loading data into a data warehouse.

"Dan Sullivan" Book

   upvoted 54 times

  **[Removed]**  4 years, 1 month ago



Should be C

   upvoted 30 times

  **emmylou**  6 months ago

On so many of these questions, how do you actually know if you're correct. I said C but the correct answer was A. Honestly, it's driving me crazy.

   upvoted 1 times


  **Mathew106** 9 months, 2 weeks ago

**Selected Answer: C**

For the ones saying BigTable is cheaper, BigTable in eu-north1 costs \$0.748/hour per node. So if you were to run the node 24/7 you would pay more than 500\$ per month. Querying 1TB of data in BigQuery is 7.5\$. With smart querying and good database design you can minimize the bytes processed by BQ. So even though BigTable does not directly charge for querying, it charges for running the cluster and the overall price does not make sense. And as far as I know, it's not possible to spin up and shut down BigTable automatically.

Also, since the table is an external table to BigQuery, we incur no cost for storing that data in BigQuery and paying 300\$ per month for storage.

   upvoted 1 times

  **samdhimal** 1 year, 3 months ago

C. Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.

Cloud Storage is a highly durable and cost-effective object storage service that can be used to store large amounts of text files. By storing the input data in CSV format in Cloud Storage, you can minimize costs while still being able to query the data using BigQuery.

BigQuery is a fully-managed, highly-scalable data warehouse that allows you to perform fast SQL-like queries on large datasets. By linking the Cloud Storage data as permanent tables in BigQuery, you can enable multiple users to query the data using multiple engines without the need for additional compute resources. This approach would be the most cost-effective for querying aggregate values for multiple users, as BigQuery charges based on the amount of data scanned per query, so the more data you store in BigQuery the less you pay per query.

   upvoted 2 times

  **samdhimal** 1 year, 3 months ago

Option D, using Cloud Storage for storage and linking as temporary tables in BigQuery for query, would not be the best choice because temporary tables only exist for the duration of a user session or query and you would need to create and delete them each time a user queries the data, which would add additional cost and complexity to the process.

Option A, Using Cloud Bigtable for storage, and installing the HBase shell on a Compute Engine instance to query the data, is not a cost-effective solution as Cloud Bigtable is a managed NoSQL database service which is more expensive than storing in Cloud Storage and querying in BigQuery.

Option B, Using Cloud Bigtable for storage, and linking as permanent tables in BigQuery for query, is not a cost-effective solution as Cloud Bigtable is a managed NoSQL database service which is more expensive than storing in Cloud Storage and querying in BigQuery.

   upvoted 1 times

  **RoshanAshraf** 1 year, 3 months ago

**Selected Answer: C**

CSV files - Cloud Storage

BigQuery - Aggregate, multiple users

Permanent table - multiple users


External Tables is Easy to implement, cost effective

   upvoted 3 times

  **rivua** 1 year, 4 months ago

The 'correct' answers on this platform are ridiculous

   upvoted 7 times

 **zellick** 1 year, 5 months ago

**Selected Answer: C**

C is the answer.

[https://cloud.google.com/bigquery/docs/external-data-cloud-storage#create\\_a\\_permanent\\_external\\_table](https://cloud.google.com/bigquery/docs/external-data-cloud-storage#create_a_permanent_external_table)

   upvoted 2 times

 **VishalBule** 2 years, 2 months ago

Answer is C Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.

BigQuery can access data in external sources, known as federated sources. Instead of first loading data into BigQuery, you can create a reference to an external source. External sources can be Cloud Bigtable, Cloud Storage, and Google Drive.

When accessing external data, you can create either permanent or temporary external tables. Permanent tables are those that are created in a dataset and linked to an external source. Dataset-level access controls can be applied to these tables. When you are using a temporary table, a table is created in a special dataset and will be available for approximately 24 hours. Temporary tables are useful for one-time operations, such as loading data into a data warehouse

   upvoted 1 times

 **medeis\_jar** 2 years, 4 months ago

**Selected Answer: C**

Bigtable is expensive. So Cloud Storage for storing and BigQuery with permanent table for linking and querying.

   upvoted 3 times

 **MaxNRG** 2 years, 4 months ago

**Selected Answer: C**

Not A or B

Big table is expensive, the initial data is in csv format, besides, if others are going to query data with multiple engines...

GCS is the storage. Between c and D is all about permanent or temporary.

Permanent table is a table that is created in a dataset and is linked to your external data source. Because the table is permanent, you can use dataset-level access controls to share the table with others who also have access to the underlying external data source, and you can query the table at any time.

When you use a temporary table, you do not create a table in one of your BigQuery datasets. Because the table is not permanently stored in a dataset, it cannot be shared with others. Querying an external data source using a temporary table is useful for one-time, ad-hoc queries over external data, or for extract, transform, and load (ETL) processes.

I think is C.

   upvoted 5 times

 **MaxNRG** 2 years, 4 months ago

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

Permanent table—You create a table in a BigQuery dataset that is linked to your external data source. This allows you to use BigQuery dataset-level IAM roles to share the table with others who may have access to the underlying external data source. Use permanent tables when you need to share the table with others.

Temporary table—You submit a command that includes a query and creates a non-permanent table linked to the external data source. With this approach you do not create a table in one of your BigQuery datasets, so make sure to give consideration towards sharing the query or table. Consider using a temporary table for one-time, ad-hoc queries, or for one-time extract, transform, or load (ETL) workflows

   upvoted 3 times

 **maurodipa** 2 years, 5 months ago

Answer is A. While C seems the most reasonable answer there are 2 points to notice: a) load jobs are limited to 15 TB across all input files in BigQuery (<https://cloud.google.com/bigquery/quotas>); b) It is requested to minimize the cost of querying and queries in BigTable are free, while queries in BigQuery are charged per byte (<https://cloud.google.com/bigquery/pricing>)

   upvoted 2 times

 **Abhi16820** 2 years, 5 months ago

[https://cloud.google.com/bigquery/external-data-bigtable#:~:text=shared%20with%20others.-,Querying%20an%20external%20data%20source%20using%20a%20temporary%20table%20is%20useful%20for%20one%20time%20ad%20hoc%20queries%20over%20external%20data%2C%20or%20for%20extract%2C%20transform%2C%20and%20load%20\(ETL\)%20processes.,-Querying%20Cloud%20Bigtable](https://cloud.google.com/bigquery/external-data-bigtable#:~:text=shared%20with%20others.-,Querying%20an%20external%20data%20source%20using%20a%20temporary%20table%20is%20useful%20for%20one%20time%20ad%20hoc%20queries%20over%20external%20data%2C%20or%20for%20extract%2C%20transform%2C%20and%20load%20(ETL)%20processes.,-Querying%20Cloud%20Bigtable)

me%2C%20ad%20hoc%20queries%20over%20external%20data%2C%20or%20for%20extract%2C%20transform%2C%20and%20load%20(ETL)%20processes.,-Querying%20Cloud%20Bigtable

   upvoted 1 times

 **tsoetan001** 2 years, 6 months ago

C is the answer.

   upvoted 1 times

 **Ysance\_AGS** 2 years, 7 months ago

A is correct since the question asks "You want to minimize the cost of querying aggregate values" => Big Table is free when querying data.

   unvoted 3 times

... updated 6 times

  **nguyenmoon** 2 years, 7 months ago


Vote for C

   upvoted 1 times

  **gcp\_learner** 2 years, 10 months ago

Interesting options. For me, A & B ruled out because BigTable doesn't fit this use case, leaves us with C & D. C will incur additional cost of storing data in GCS & BigQuery because it mentions linking.

So I would go with D ie store the data in GCS and create external tables in BigQuery.

   upvoted 4 times

  **triipinbee** 2 years, 8 months ago

Storage cost of data for BQ is the same as standard cloud storage, actually less for long term storage as it automatically moves to nearline storage.

<https://cloud.google.com/bigquery/pricing#storage>

<https://cloud.google.com/storage#section-10>

   upvoted 1 times

  **Yiouk** 2 years, 9 months ago

[https://cloud.google.com/bigquery/docs/writing-results#temporary\\_and\\_permanent\\_tables](https://cloud.google.com/bigquery/docs/writing-results#temporary_and_permanent_tables)

   upvoted 1 times

[Load full discussion...](#)



## Platform

> Home

> All Exams

> Examtopics PRO

> Training Courses



© 2024 ExamTopics