

 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 70 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 70

Topic #: 1

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

You are designing storage for very large text files for a data pipeline on Google Cloud. You want to support ANSI SQL queries. You also want to support compression and parallel load from the input locations using Google recommended practices. What should you do?

- A. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.
- B. Transform text files to compressed Avro using Cloud Dataflow. Use Cloud Storage and BigQuery permanent linked tables for query.
- C. Compress text files to gzip using the Grid Computing Tools. Use BigQuery for storage and query.
- D. Compress text files to gzip using the Grid Computing Tools. Use Cloud Storage, and then import into Cloud Bigtable for query.

[Show Suggested Answer](#)

by  [Rajokkiyam](#) at March 15, 2020, 6:11 a.m.

Comments

[Submit](#)

🗄️ 👤 **Ganshank** Highly Voted 👍 5 years ago

B.

The question is focused on designing storage for very large files, with support for compression, ANSI SQL queries, and parallel loading from the input locations. This can be met using GCS for storage and Bigquery permanent tables with external data source in GCS.

👍 ↩️ 🚩 upvoted 59 times

🗄️ 👤 **atnafu2020** 4 years, 9 months ago

why GCS as external since Bigquery can be used as storage as well?

👍 ↩️ 🚩 upvoted 10 times

🗄️ 👤 **atnafu2020** 4 years, 9 months ago

A seems correct for me

👍 ↩️ 🚩 upvoted 11 times

🗄️ 👤 **atnafu2020** 4 years, 8 months ago

Since its best practice, i go by with B not A

👍 ↩️ 🚩 upvoted 4 times

🗄️ 👤 **gopinath_k** 4 years, 1 month ago

They want to store the files if you try with bq I think you will need to strike the word compression.

👍 ↩️ 🚩 upvoted 2 times

🗄️ 👤 **jkhong** 2 years, 4 months ago

The question focuses on "designing storage", rather than designing a data warehouse.

👍 ↩️ 🚩 upvoted 5 times

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

Should be A

👍 ↩️ 🚩 upvoted 15 times

🗄️ 👤 **tavva_prudhvi** 2 years, 10 months ago

Not A : Importing data into BigQuery may take more time compared to creating external tables on data. Additional storage costs by BigQuery is another issue which can be more expensive than Google Storage.

👍 ↩️ 🚩 upvoted 7 times

🗄️ 👤 **oussama7** Most Recent 🕒 1 month, 2 weeks ago

Selected Answer: A

Avro is a Google-recommended format for BigQuery because it supports schema evolution, efficient compression, and parallel processing. Using Cloud Dataflow ensures scalable transformation, and storing the data in BigQuery allows for optimized ANSI SQL queries.

👍 ↩️ 🚩 upvoted 1 times

🗄️ 👤 **Parandhaman_Margan** 1 month, 3 weeks ago

Selected Answer: A

Storing large text files with SQL support**

Avro in BigQuery (A) supports compression and efficient queries.

👍 ↩️ 🚩 upvoted 1 times

🗄️ 👤 **dcruzado** 2 months ago

Selected Answer: B

For me it would make more sense to store the data directly in BQ, but A does not make sense, because why compress to Avro if you dont store the avro files and directly saves the data to BQ?

You are not using the compression

👍 ↩️ 🚩 upvoted 1 times

🗄️ 👤 **imarri876** 2 months, 2 weeks ago

Selected Answer: A

BigQuery now has physical storage, which makes storage cost fairly cheap on BigQuery with compression. I would go with A.

👍 ↩️ 🚩 upvoted 1 times

🗄️ 👤 **deineiveu** 4 months, 3 weeks ago

Selected Answer: B

Gros fichier + sql = GCS + Bigquery

👍 ↩️ 🚩 upvoted 1 times

🗄️ 👤 **Nittin** 8 months, 1 week ago

Selected Answer: B

copy to gcs and use external tble in bq

   upvoted 1 times

  **carmltekai** 9 months, 2 weeks ago

Selected Answer: A

Should be A.

Check this link for the advantage of load Avro data to BigQuery https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-avro#advantages_of_avro

""The Avro binary format:

- * Is faster to load. The data can be read in parallel, even if the data blocks are compressed.

- * Doesn't require typing or serialization.

- * Is easier to parse because there are no encoding issues found in other formats such as ASCII.

When you load Avro files into BigQuery, the table schema is automatically retrieved from the self-describing source data.""

   upvoted 2 times

  **carmltekai** 9 months, 2 weeks ago

While option B can work, it introduces additional complexity by linking Cloud Storage with BigQuery. Directly storing data in BigQuery is more efficient for querying purposes.

There are no requirements about cost, So simple is better

   upvoted 1 times

  **SK1594** 1 year, 1 month ago

B makes sense

   upvoted 2 times


  **MaxNRG** 1 year, 4 months ago

Selected Answer: B

1. Store Avro files in GCS

2. Query them in BigQuery (federated tables)

   upvoted 3 times

  **forepick** 1 year, 11 months ago

Selected Answer: B

Answer is B.

The requirements are:

- storage for compressed text files
- parallel loads to SQL tool

AVRO is a compressed format for text files, which makes it possible to load chunks of a very large file in parallel to BigQuery.

gzip files are seamless in GCS though, but cannot load in parallel to BQ.

   upvoted 6 times

  **samdhimal** 2 years, 3 months ago

Correct Answer:

A. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.

This option offers several advantages:

- Transforming the text files to compressed Avro using Cloud Dataflow allows for parallel processing of the input data, improving the efficiency of the pipeline.

- Compressing the data in Avro format further reduces the storage space required and improves data transfer performance.

- Storing the data in BigQuery supports ANSI SQL queries and allows for easy querying of the data.

- BigQuery is a fully-managed data warehousing solution, it's scalable and can handle large datasets and concurrent queries, so it's suitable for large text files.

   upvoted 3 times

  **samdhimal** 2 years, 3 months ago

Option B is similar to option A but it's using a permanent linked table between Cloud Storage and BigQuery, this approach is not recommended as it's not efficient and could lead to data duplication, and it doesn't take advantage of the parallel processing capabilities of Cloud Dataflow.

Option C and D are incorrect because they don't take advantage of the parallel processing capabilities of Cloud Dataflow, and they don't use Avro format for compression which is more efficient and recommended by Google. Storing the data in

and they don't use Avro format for compression which is more efficient and recommended by Google. Storing the data in Cloud Bigtable also doesn't support ANSI SQL queries which is a requirement for this use case.

👍 ↩ 🚩 upvoted 1 times

🗄 👤 **jkhong** 2 years, 4 months ago

Selected Answer: B

Designing storage solution, not data warehousing -> So Cloud Storage.

Support compression -> just use Avro

Parallel load -> refers to upload from input locations, NOT download.

Load in parallel using -m flag for gsutil cp

<https://cloud.google.com/storage/docs/uploads-downloads#console>

👍 ↩ 🚩 upvoted 3 times

🗄 👤 **odacir** 2 years, 4 months ago

Selected Answer: B

C and D are discarded.

A and B are possible.

A is the best for query, but ... the sentence says: ou also want to support compression and parallel load from the input locations using Google recommended practices.

BigQuery only support parallel load from storage, storage support parallel load from CLI. So the only option is B.

👍 ↩ 🚩 upvoted 3 times

🗄 👤 **zelck** 2 years, 5 months ago

Selected Answer: B

B is the answer.

👍 ↩ 🚩 upvoted 1 times

🗄 👤 **nkit** 2 years, 5 months ago

Selected Answer: B

"Very large files" and "long term storage" are two key phrases- both of which indicate to pick cloud storage as option. Hence B is correct.

👍 ↩ 🚩 upvoted 1 times

[Load full discussion...](#)



Platform

> Home

> Examtopics PRO

> All Exams

> Training Courses

