

[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 225 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 225

Topic #: 1

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

Your organization stores customer data in an on-premises Apache Hadoop cluster in Apache Parquet format. Data is processed on a daily basis by Apache Spark jobs that run on the cluster. You are migrating the Spark jobs and Parquet data to Google Cloud. BigQuery will be used on future transformation pipelines so you need to ensure that your data is available in BigQuery. You want to use managed services, while minimizing ETL data processing changes and overhead costs. What should you do?

- A. Migrate your data to Cloud Storage and migrate the metadata to Dataproc Metastore (DPMS). Refactor Spark pipelines to write and read data on Cloud Storage, and run them on Dataproc Serverless.
- B. Migrate your data to Cloud Storage and register the bucket as a Dataplex asset. Refactor Spark pipelines to write and read data on Cloud Storage, and run them on Dataproc Serverless.
- C. Migrate your data to BigQuery. Refactor Spark pipelines to write and read data on BigQuery, and run them on Dataproc Serverless.
- D. Migrate your data to BigLake. Refactor Spark pipelines to write and read data on Cloud Storage, and run them on Dataproc on Compute Engine.

[Show Suggested Answer](#)

by [e70ea9e](#) at Dec. 30, 2023, 9:51 a.m.

Comments



Submit

 **raaad** Highly Voted  1 year, 4 months ago

Selected Answer: A

- This option involves moving Parquet files to Cloud Storage, which is a common and cost-effective storage solution for big data and is compatible with Spark jobs.
- Using Dataproc Metastore to manage metadata allows us to keep Hadoop ecosystem's structural information.
- Running Spark jobs on Dataproc Serverless takes advantage of managed Spark services without managing clusters.
- Once the data is in Cloud Storage, you can also easily load it into BigQuery for further analysis.


   upvoted 6 times

 **380e3c6** Most Recent  2 months, 2 weeks ago

Selected Answer: A

A is correct because it minimizes ETL changes, keeps Parquet data in Cloud Storage (cost-effective and Spark-compatible), and integrates with BigQuery via external tables. C is flawed** since moving directly to BigQuery requires refactoring Spark jobs, increasing complexity and costs. B adds unnecessary governance overhead, and D focuses on infrastructure instead of pipeline efficiency.

   upvoted 1 times

 **plum21** 2 months, 3 weeks ago

Selected Answer: D

The requirement:

"You want to use managed services"

excludes Dataproc Serverless.

Dataproc on Compute Engine remains.

Next requirement:

"BigQuery will be used on future transformation pipelines so you need to ensure that your data is available in BigQuery" -> BigLake


Next requirement:

"while minimizing ETL data processing changes and overhead costs" -> Refactor Spark pipelines to write and read data on Cloud Storage

Notes

1. Dataproc Metastore (DPMS) could be used on Dataproc to read data from BQ but not the other way round.

   upvoted 1 times

 **skhaire** 2 months, 4 weeks ago

Selected Answer: B

BigQuery Integration: The requirement is to make data available in BigQuery. Dataplex has built-in integration with BigQuery. It can automatically discover data in Cloud Storage and create external tables in BigQuery, making the data readily queryable. DPMS doesn't have this direct integration with BigQuery.


   upvoted 4 times

 **LP_PDE** 3 months, 1 week ago

Selected Answer: A

Both Spark and BigQuery can directly access data in Cloud Storage.

   upvoted 1 times

 **hrishi19** 5 months, 2 weeks ago

Selected Answer: C

The question states that the data should be available on BigQuery and only option C meets this requirement.

   upvoted 3 times

 **JamesKarianis** 8 months, 3 weeks ago

Selected Answer: A

A is correct

   upvoted 1 times

 **Anudeep58** 11 months ago

Selected Answer: A

Option B: Registering the bucket as a Dataplex asset adds an additional layer of data governance and management. While useful, it may not be necessary for your immediate migration needs and can introduce additional complexity.

Option C: Migrating data directly to BigQuery would require significant changes to your Spark pipelines since they would need to be refactored to read from and write to BigQuery instead of Parquet files. This approach could introduce higher costs due to BigQuery storage and querying.

Option D: Using BigLake and Dataproc on Compute Engine is more complex and requires more management compared to Dataproc Serverless. Additionally, it might not be as cost-effective as leveraging Cloud Storage and Dataproc Serverless.

Dataproc Serverless. Additionally, it might not be as cost-effective as leveraging Cloud Storage and Dataproc Serverless.

👍 ↩ 🚩 upvoted 3 times

🗨️ 👤 **aoifneofi_ef** 8 months, 2 weeks ago

Just adding further commentary on why A is correct while why other options are incorrect is explained above. Parquet files have schema engrained in them. Hence Spark pipelines on Hadoop Cluster may not have needed tables at all. Hence the simplest solution would be to move it to Cloud Storage instead of BigQuery and this way there would be minimal changes to the ETL pipelines - just change HDFS file system pointer to GCS file system for read writes and no need for any additional tables

👍 ↩ 🚩 upvoted 2 times

🗨️ 👤 **josech** 11 months, 3 weeks ago

Selected Answer: A

The question says "You want to use managed services, while minimizing ETL data processing changes and overhead costs". Dataproc is a managed service that doesn't need to refactor the data transformation Spark code you already have (you will have to refactor only the write and read code), and it has a Big Query connector for future use. <https://cloud.google.com/dataproc/docs/concepts/connectors/bigquery>

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **52ed0e5** 1 year, 1 month ago

Selected Answer: C

Migrate your data directly to BigQuery.
Refactor Spark pipelines to read from and write to BigQuery.
Run the Spark jobs on Dataproc Serverless.
The best choice for ensuring data availability in BigQuery. It allows seamless integration with BigQuery and minimizes ETL changes.

👍 ↩ 🚩 upvoted 3 times

🗨️ 👤 **Ramon98** 1 year, 2 months ago

Selected Answer: C

A tricky one, because of "you need to ensure that your data is available in BigQuery". The easiest and most straight forward migration seems answer A to me, and then you can use external tables to make the parquet data directly available in BigQuery. <https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-parquet>

However creating the external tables is an extra step? So therefore maybe C is the answer?

👍 ↩ 🚩 upvoted 4 times

🗨️ 👤 **Moss2011** 1 year, 2 months ago

Selected Answer: C

I think the key phrase here is "you need to ensure that your data is available in BigQuery" that's why I think C it's the best option

👍 ↩ 🚩 upvoted 1 times

🗨️ 👤 **JyoGCP** 1 year, 2 months ago

Selected Answer: C

I think it's C.

Dataproc can use BigQuery to read and write data.
Dataproc's BigQuery connector is a library that allows Spark and Hadoop applications to process and write data from BigQuery.

Here's how Dataproc can be used with BigQuery:
Process large datasets: Use Spark to process data stored in BigQuery.
Write results: Write the results back to BigQuery or other data storage for further analysis.
Read data: The BigQuery connector can read data from BigQuery into a Spark DataFrame.
Write data: The connector writes data to BigQuery by buffering all the data into a Cloud Storage temporary table.

👍 ↩ 🚩 upvoted 3 times

🗨️ 👤 **JyoGCP** 1 year, 2 months ago

As per question.. "BigQuery will be used on future transformation pipelines so you need to ensure that your data is available in BigQuery. You want to use managed services (DATAPROC), while minimizing ETL data processing changes and overhead costs."

👍 ↩ 🚩 upvoted 3 times

🗨️ 👤 **matijax** 1 year, 2 months ago

Selected Answer: B

I think its B and the reason is that registering the data as a Dataplex asset enables seamless integration with BigQuery later on. Dataplex simplifies data discovery and lineage tracking, making it easier to prepare your data for BigQuery transformations.

transformation.

   upvoted 3 times

  **saschak94** 1 year, 2 months ago

Why would I select A here? Why not moving the data to BigQuery and running Dataproc Serverless jobs accessing the data in BigQuery?

   upvoted 3 times

  **e70ea9e** 1 year, 4 months ago

Selected Answer: A

Managed Services: Leverages Dataproc Serverless for a fully managed Spark environment, reducing overhead and administrative tasks.

Minimal Data Processing Changes: Keeps Spark pipelines largely intact by working with Parquet files on Cloud Storage, minimizing refactoring efforts.

BigQuery Integration: Dataproc Serverless can directly access BigQuery, enabling future transformation pipelines without additional data movement.

Cost-Effective: Serverless model scales resources only when needed, optimizing costs for intermittent workloads.

   upvoted 3 times



Platform

> Home

> All Exams

> Examtopics PRO

> Training Courses



© 2024 ExamTopics