## EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 275 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 275
Topic #: 1
[All Professional Data Engineer Questions]

You created an analytics environment on Google Cloud so that your data scientist team can explore data without impacting the on-premises Apache Hadoop solution. The data in the on-premises Hadoop Distributed File System (HDFS) cluster is in Optimized Row Columnar (ORC) formatted files with multiple columns of Hive partitioning. The data scientist team needs to be able to explore the data in a similar way as they used the on-premises HDFS cluster with SQL on the Hive query engine. You need to choose the most cost-effective storage and processing solution. What should you do?

A. Import the ORC files to Bigtable tables for the data scientist team.

B. Import the ORC files to BigQuery tables for the data scientist team.

C. Copy the ORC files on Cloud Storage, then deploy a Dataproc cluster for the data scientist team.

D. Copy the ORC files on Cloud Storage, then create external BigQuery tables for the data scientist team.

Show Suggested Answer

by Smakyel79 at *Jan. 7, 2024, 5:17 p.m.*

### Comments

Type your comment...

Submit

**raaad** Highly Voted 1 year, 3 months ago

Selected Answer: D

- It leverages the strengths of BigQuery for SQL-based exploration while avoiding additional costs and complexity associated with data transformation or migration.
- The data remains in ORC format in Cloud Storage, and BigQuery's external tables feature allows direct querying of this data.

upvoted 8 times

**nadavw** 8 months, 1 week ago

There is a requirement to use a 'hive query engine", and BQ is using only the hive metastore and his own engine, so 'D' seems a better fit here.

upvoted 1 times

**kaisarfarel** Highly Voted 1 year, 1 month ago

I think C is the correct answer, DS want to explore the data in a "similar way as they used the on-premises HDFS cluster with SQL on the Hive query engine". Dataproc can help to create clusters quickly with the Hadoop cluster. CMIIW

upvoted 6 times

**apoio.certificacoes.closer** 4 months, 1 week ago

I think "Similar" is doing a lot of heavy lift on the confusion. If it was equal, I'd say C. Since it similar, it can be GoogleSQL (Bigquery).

upvoted 2 times

**Pime13** Most Recent 3 months, 4 weeks ago

Selected Answer: D

D. Copy the ORC files on Cloud Storage, then create external BigQuery tables for the data scientist team.

This approach allows you to leverage the scalability and cost-effectiveness of Cloud Storage while enabling your data scientists to query the data using BigQuery's powerful SQL engine without the need to move or transform the data. This setup also minimizes the need for additional infrastructure and maintenance, making it a practical choice for your analytics environment.

upvoted 1 times

**SamuelTsch** 6 months, 1 week ago

Selected Answer: B

using external tables have always limitations - affecting performance, no preview of the data and no cost estimation. So, why option D is correct?

upvoted 1 times

**hanoverquay** 1 year, 1 month ago

Selected Answer: D

option d

upvoted 1 times

**0725f1f** 1 year, 1 month ago

Selected Answer: C

it is talking about partition as well

upvoted 3 times

**JyoGCP** 1 year, 2 months ago

Selected Answer: D

Option D

upvoted 1 times

**Matt_108** 1 year, 3 months ago

Selected Answer: D

Option D - leverages BigQuery for SQL-based exploration on direct querying to cloud storage

upvoted 2 times

**Smakyel79** 1 year, 3 months ago

Selected Answer: D

This approach leverages BigQuery's powerful analytics capabilities without the overhead of data transformation or maintaining a separate cluster, while also allowing your team to use SQL for data exploration, similar to their experience with the on-premises Hadoop/Hive environment.

upvoted 3 times

## Platform

- Home
- Examtopics PRO
- All Exams
- Training Courses