

[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 167 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 167

Topic #: 1

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

Your company currently runs a large on-premises cluster using Spark, Hive, and HDFS in a colocation facility. The cluster is designed to accommodate peak usage on the system; however, many jobs are batch in nature, and usage of the cluster fluctuates quite dramatically. Your company is eager to move to the cloud to reduce the overhead associated with on-premises infrastructure and maintenance and to benefit from the cost savings. They are also hoping to modernize their existing infrastructure to use more serverless offerings in order to take advantage of the cloud. Because of the timing of their contract renewal with the colocation facility, they have only 2 months for their initial migration. How would you recommend they approach their upcoming migration strategy so they can maximize their cost savings in the cloud while still executing the migration in time?

- A. Migrate the workloads to Dataproc plus HDFS; modernize later.
- B. Migrate the workloads to Dataproc plus Cloud Storage; modernize later.
- C. Migrate the Spark workload to Dataproc plus HDFS, and modernize the Hive workload for BigQuery.
- D. Modernize the Spark workload for Dataflow and the Hive workload for BigQuery.

[Show Suggested Answer](#)

by [AWSandeep](#) at *Sept. 2, 2022, 6:57 p.m.*

Comments

Type your comment...

  **zellck** Highly Voted 1 year, 11 months ago

Selected Answer: B

B is the answer.

<https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#overview>

When you want to move your Apache Spark workloads from an on-premises environment to Google Cloud, we recommend using Dataproc to run Apache Spark/Apache Hadoop clusters. Dataproc is a fully managed, fully supported service offered by Google Cloud. It allows you to separate storage and compute, which helps you to manage your costs and be more flexible in scaling your workloads.

https://cloud.google.com/bigquery/docs/migration/hive#data_migration

Migrating Hive data from your on-premises or other cloud-based source cluster to BigQuery has two steps:

1. Copying data from a source cluster to Cloud Storage
2. Loading data from Cloud Storage into BigQuery

   upvoted 8 times

  **AzureDP900** 1 year, 10 months ago

B. Migrate the workloads to Dataproc plus Cloud Storage; modernize later.

   upvoted 1 times

  **MaxNRG** Most Recent 10 months, 2 weeks ago

Selected Answer: B

Based on the time constraint of 2 months and the goal to maximize cost savings, I would recommend option B - Migrate the workloads to Dataproc plus Cloud Storage; modernize later.

The key reasons are:

- Dataproc provides a fast, native migration path from on-prem Spark and Hive to the cloud. This allows meeting the 2 month timeline.
- Using Cloud Storage instead of HDFS avoids managing clusters for variable workloads and provides cost savings.
- Further optimizations and modernization to serverless (Dataflow, BigQuery) can happen incrementally later without time pressure.

   upvoted 2 times

  **MaxNRG** 10 months, 2 weeks ago

Option A still requires managing HDFS.

Option C and D require full modernization of workloads in 2 months which is likely infeasible.

Therefore, migrating to Dataproc with Cloud Storage fast tracks the migration within 2 months while realizing immediate cost savings, enabling the flexibility to iteratively modernize and optimize the workloads over time.

   upvoted 3 times

  **John_Pongthorn** 2 years, 1 month ago

Selected Answer: B

B is most likely

1. migrate job and infrastructure to dataproc on cloud
2. any data, move from hdfs on-premise to google cloud storage (one of them is Hive)

If you want to modernize Hive to Bigquery , you are need to move it into GCS(preceding step) first and load it into bigquery that is all.

<https://cloud.google.com/blog/products/data-analytics/apache-hive-to-bigquery>

<https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc>

<https://cloud.google.com/architecture/hadoop/hadoop-gcp-migration-data>

   upvoted 3 times

  **TNT87** 2 years, 1 month ago

Selected Answer: D

Answer D

   upvoted 1 times




  **dn_mohammed_data** 2 years, 1 month ago

you sould migrate spark to apache beam which is not the case here

   upvoted 1 times

  **TNT87** 2 years, 1 month ago

apache beam for what???

   upvoted 1 times

  **adarifian** 2 years ago

dataflow uses apache beam

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **TNT87** 1 year, 9 months ago

@adarifian Why use apache beam yet there is Dataflow an inhouse gcp solution to solve the problem? hence i said apache beam for what

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **ExamCtechs** 1 year ago

Dataflow IS apache beam, Dataflow is a Beam Runner.

If you go for that soulution you will need to modify your pipeline to use Beam

👍 🚩 upvoted 1 times

🗄️ 👤 **GyaneswarPanigrahi** 2 years, 1 month ago

D isn't feasible, within 2 months. Anyone who has worked in a Hadoop/ Big Data data warehousing or data lake project, knows how less time 2 months is, given the amount of data and associated complexities abound.

It should be B to begin with. And then gradually move towards D.

👍 ↩ 🚩 upvoted 3 times

🗄️ 👤 **TNT87** 2 years, 1 month ago

Selected Answer: B

Ans B

-cost saving

-time factor

-Spark -Data proc

👍 ↩ 🚩 upvoted 2 times

🗄️ 👤 **TNT87** 2 years, 1 month ago

Ans D is also relevant if you read this. Onthe other hand cloud storage isnt severless but bigquery is <https://cloud.google.com/hadoop-spark-migration>

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **damaldon** 2 years, 2 months ago

Ans.B as per the following link

<https://blog.devgenius.io/migrating-spark-jobs-to-google-cloud-file-event-sensor-to-dynamically-create-spark-cluster-7eff2c75423d>

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **YorelNation** 2 years, 2 months ago

Selected Answer: B

For the time window of two month I would recommend B and then start to implement D.

👍 ↩ 🚩 upvoted 2 times

🗄️ 👤 **ducc** 2 years, 2 months ago

It is B or D, still confusing

👍 ↩ 🚩 upvoted 2 times

🗄️ 👤 **AWSandeep** 2 years, 2 months ago

Selected Answer: D

D because the Apache Spark Runner can be used to execute Beam pipelines using Apache Spark. Also, Hive to BigQuery is not a difficult modernization/migration.

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **ExamCtechs** 1 year ago

Dataflow is a Runner of Beam it self

👍 ↩ 🚩 upvoted 1 times

> Home

> Examtopics PRO

> All Exams

> Training Courses

