

🔗 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 19 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 19

Topic #: 1

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

Your company's on-premises Apache Hadoop servers are approaching end-of-life, and IT has decided to migrate the cluster to Google Cloud Dataproc. A like-for-like migration of the cluster would require 50 TB of Google Persistent Disk per node. The CIO is concerned about the cost of using that much block storage. You want to minimize the storage cost of the migration. What should you do?

- A. Put the data into Google Cloud Storage.
- B. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.
- C. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.
- D. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.

[Show Suggested Answer](#)

by [deleted] at *March 17, 2020, 4:29 p.m.*

### Comments

Type your comment...

[Submit](#)

🗨️ **anji007** Highly Voted 7 months, 1 week ago

Ans: A

B: Wrong eVM wont solve the problem of larger storage prices.

C: May be, but nothing mentioned in terms of what to tune in the question, also this is like-for-like migration so tuning may not be part of the migration.

D: Again, this is like-for-like so need to define what is hot data and which is cold data, also persistent disk costlier than cloud storage.

   upvoted 8 times

  **fassil** Most Recent 3 weeks, 4 days ago

**Selected Answer: A**

A like-for-like migration to Cloud Dataproc that replicates on-premises Hadoop would require each node to have 50 TB of persistent disk, which is costly. Instead, you can minimize storage costs by leveraging Google Cloud Storage (GCS). Cloud Dataproc seamlessly integrates with GCS through the Hadoop connector, allowing you to store your data cost-effectively in Cloud Storage and run ephemeral clusters that read data directly from GCS. This approach eliminates the need for each node to carry 50 TB of expensive persistent disk storage while still supporting your Hadoop workload.

   upvoted 1 times

  **Parandhaman\_Margan** 1 month, 3 weeks ago

**Selected Answer: D**

. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.

Google Cloud Storage (GCS) is a cost-effective alternative to Persistent Disk for storing less frequently accessed ("cold") data.

Hot data that requires fast access can remain on Persistent Disk, reducing storage costs while maintaining performance.

Cloud Dataproc supports HDFS-to-GCS integration, allowing Hadoop jobs to access data in GCS seamlessly.D

   upvoted 1 times

  **Vullibabu** 1 year, 4 months ago

You are most of the people looking at like for like migration would require 50TB persistent storage but missing to look at CIO concern about cost of block storage...considering CIO concern the option here is cloud storage... moreover that is recommended as well ..

   upvoted 1 times

  **imran79** 1 year, 7 months ago

Option A: Put the data into Google Cloud Storage.

This is the best option. Google Cloud Dataproc is designed to work well with Google Cloud Storage. Using GCS instead of Persistent Disk can save money, and GCS offers advantages such as higher durability and the ability to share data across multiple clusters.

   upvoted 2 times

  **emmylou** 1 year, 7 months ago

I have seen this question in other places and I believe that you store the older data in Cloud Storage and retain processing data in persistent disk. D

   upvoted 1 times

  **hxy8** 1 year, 8 months ago

Answer: D

Question: A like-for- like migration of the cluster would require 50 TB of Google Persistent Disk per node. which means Persistent is still required.

   upvoted 1 times

  **suku2** 1 year, 7 months ago

Google Cloud Storage is designed for 11 9's availability. So it is also kind of persistent storage. Also, it is a Google product, hence recommended.



<https://cloud.google.com/storage/docs/availability-durability#key-concepts>

   upvoted 1 times

  **GHOST1985** 1 year, 8 months ago

the question is talking about block storage , GCS is object storage !

   upvoted 1 times

  **hjava** 1 year, 9 months ago

**Selected Answer: A**

GCS is cost-effective and also Google's recommendation!

   upvoted 1 times

  **bha11111** 2 years, 1 month ago

**Selected Answer: A**

Minimize cost then GCS

   upvoted 1 times

🗄️ 👤 **Nirca** 2 years, 3 months ago

**Selected Answer: A**

A - is the right answer.

👍 ↩️ 🚩 upvoted 1 times

🗄️ 👤 **DGames** 2 years, 4 months ago

**Selected Answer: A**

A - dataproc - storage - cost effective is cloud storage

👍 ↩️ 🚩 upvoted 1 times

🗄️ 👤 **devaid** 2 years, 7 months ago

**Selected Answer: A**

Cloud Storage

👍 ↩️ 🚩 upvoted 1 times

🗄️ 👤 **sankar\_s** 2 years, 11 months ago

**Selected Answer: A**

Cloud Storage is google recommended one

👍 ↩️ 🚩 upvoted 1 times

🗄️ 👤 **sumanshu** 3 years, 10 months ago

Vote for 'A'

👍 ↩️ 🚩 upvoted 2 times

🗄️ 👤 **sumanshu** 7 months, 1 week ago

A is correct because Google recommends using Cloud Storage instead of HDFS as it is much more cost effective especially when jobs aren't running.

B is not correct because this will decrease the compute cost but not the storage cost.

C is not correct because while this will reduce cost somewhat, it will not be as cost effective as using Cloud Storage.

D is not correct because while this will reduce cost somewhat, it will not be as cost effective as using Cloud Storage.

👍 ↩️ 🚩 upvoted 6 times

🗄️ 👤 **anudeepgupta42** 4 years ago

A, Moving the data to GCS will reduce the cost of running the dataproc clusters all the time \

👍 ↩️ 🚩 upvoted 1 times

🗄️ 👤 **naga** 4 years, 2 months ago

Correct A

👍 ↩️ 🚩 upvoted 2 times

[Load full discussion...](#)



## Platform

> [Home](#)

> [Examtopics PRO](#)

> [All Exams](#)

> [Training Courses](#)



