

🔗 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 156 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 156

Topic #: 1

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

You are planning to migrate your current on-premises Apache Hadoop deployment to the cloud. You need to ensure that the deployment is as fault-tolerant and cost-effective as possible for long-running batch jobs. You want to use a managed service. What should you do?

- A. Deploy a Dataproc cluster. Use a standard persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from `hdfs://` to `gs://`
- B. Deploy a Dataproc cluster. Use an SSD persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from `hdfs://` to `gs://`
- C. Install Hadoop and Spark on a 10-node Compute Engine instance group with standard instances. Install the Cloud Storage connector, and store the data in Cloud Storage. Change references in scripts from `hdfs://` to `gs://`
- D. Install Hadoop and Spark on a 10-node Compute Engine instance group with preemptible instances. Store data in HDFS. Change references in scripts from `hdfs://` to `gs://`

[Show Suggested Answer](#)

by [deleted] at *March 22, 2020, 7:31 a.m.*

## Comments

Type your comment...

  **[Removed]** Highly Voted  4 years, 7 months ago

Correct: A

Ask for cost effective so persistent disk are HDD which are cheaper in comparison to SSD.

   upvoted 33 times

  **[Removed]** Highly Voted  4 years, 7 months ago

Confused between A and B. For r/w intensive jobs need to use SSDs. But questions doesnt state anything about the nature of the jobs. So better to start with a default option.

Choose A

   upvoted 16 times

  **baubaumiaomiao** 2 years, 10 months ago

"You need to ensure that the deployment is as cost-effective as possible"  
hence, no SSD unless stated otherwise

   upvoted 3 times

  **mothkuri** Most Recent  8 months ago

Selected Answer: A

Options A is the right answer.

Option B using SSD persistent disk which will add more cost than default HDD

Option C & D are out of scope.

   upvoted 2 times

  **barnac1es** 1 year, 1 month ago

Selected Answer: A

Dataproc Managed Service: Dataproc is a fully managed service for running Apache Hadoop and Spark. It provides ease of management and automation.


Standard Persistent Disk: Using standard persistent disks for Dataproc workers ensures durability and is cost-effective compared to SSDs.

Preemptible Workers: By using 50% preemptible workers, you can significantly reduce costs while maintaining fault tolerance. Preemptible VMs are cheaper but can be preempted by Google, so having a mix of preemptible and non-preemptible workers provides cost savings with redundancy.

Storing Data in Cloud Storage: Storing data in Cloud Storage is highly durable, scalable, and cost-effective. It also makes data accessible to Dataproc clusters, and you can leverage native connectors for reading data from Cloud Storage.

Changing References to gs://: Updating your scripts to reference data in Cloud Storage using gs:// ensures that your jobs work seamlessly with the cloud storage infrastructure.

   upvoted 2 times

  **vaga1** 1 year, 5 months ago

Selected Answer: A

Apache Hadoop -> Dataproc or Compute Engine with proper SW installation

cost-effective -> use standard persistent disk + store data in Cloud Storage

batch -> Dataproc or Compute Engine with proper SW installation

managed service -> Dataproc

   upvoted 1 times

  **zellck** 1 year, 11 months ago

Selected Answer: A

A is the answer.

<https://cloud.google.com/dataproc/docs/concepts/overview>




Dataproc is a managed Spark and Hadoop service that lets you take advantage of open source data tools for batch processing, querying, streaming, and machine learning. Dataproc automation helps you create clusters quickly, manage them easily, and save money by turning clusters off when you don't need them. With less time and money spent on administration, you can focus on your jobs and your data.

   upvoted 2 times

  **MounicaN** 2 years, 1 month ago

Selected Answer: A

it says cost effective , hence no SSD

   upvoted 1 times

- 📄 👤 **JG123** 2 years, 11 months ago  
Correct: A  
👍 🔄 🚩 upvoted 2 times
- 📄 👤 **LORETOGOMEZ** 3 years, 3 months ago  
Correct : A  
Option B is usefull if you use HDFS, and in this case as you use preemptible machines it isn't worth use SSD disks.  
👍 🔄 🚩 upvoted 2 times
- 📄 👤 **ArunSingh1028** 3 years, 8 months ago  
Answer - B  
👍 🔄 🚩 upvoted 1 times
- 📄 👤 **StelSen** 3 years, 9 months ago  
Look at this link. <https://cloud.google.com/bigtable/docs/choosing-ssd-hdd>  
At the First look I chose Option-B as they mentioned SSD is cost-effective on most cases. But after reading the whole page, they also mentioned that for batch workloads, HDD is suggested as long as not heavy read. So I changed my mind to Option-A (I assumed this is not ready heavy process?).  
👍 🔄 🚩 upvoted 5 times
- 📄 👤 **NM1212** 2 years, 2 months ago  
Caution about the link you provided as reference. It's intendedfor BigTable which is GC's low-latency solution which is totally different requirement. Mentioning only because on first read I thought SSD is the obvious choice.  
Per below link, SSD may not be required unless there is a low-latency requirement or a high I/O requirement. Since the question does not specify anything like that, A looks correct.  
<https://cloud.google.com/solutions/migration/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc>  
👍 🔄 🚩 upvoted 1 times
- 📄 👤 **Alasmindas** 3 years, 11 months ago  
Option B - SSD disks, reasons:-  
The question asks "fault-tolerant and cost-effective as possible for long-running batch job".  
3 Key words are - fault tolerant / cost effective / long running batch jobs..  
  
The cost efficiency part mentioned in the question could be addressed by 50% preemptible disks and storing the data in cloud storage than HDFS.  
For long running batch jobs and as standard approach for Dataproc - we should always go with SSD disk types as per google recommendations.  
👍 🔄 🚩 upvoted 4 times
- 📄 👤 **beedle** 3 years, 11 months ago  
where is the proof...show me the link?  
👍 🔄 🚩 upvoted 2 times
- 📄 👤 **Raangs** 3 years, 8 months ago  
<https://cloud.google.com/solutions/migration/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc>  
As per this, SSD is only recommended if it is high IO intensive. In this question no where mentioned its high IO intensive, and asks for cost effective (as much as possible), so no need to use SSD.  
I will go with A.  
👍 🔄 🚩 upvoted 6 times
- 📄 👤 **Ravivarma4786** 4 years, 2 months ago  
Ans is B, for long running SDD suitable. HDD maintenance will be additional charge for long running jobs  
👍 🔄 🚩 upvoted 2 times
- 📄 👤 **Rajuuu** 4 years, 3 months ago  
Answer is A...Cloud Dataproc for Managed Cloud native application and HDD for cost-effective solution.  
👍 🔄 🚩 upvoted 7 times
- 📄 👤 **Rajokkiyam** 4 years, 7 months ago  
Answer A  
👍 🔄 🚩 upvoted 5 times

Platform

- > Home
- > Examtopics PRO
- > All Exams
- > Training Courses

