

[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 150 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 150

Topic #: 1

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

You want to build a managed Hadoop system as your data lake. The data transformation process is composed of a series of Hadoop jobs executed in sequence.

To accomplish the design of separating storage from compute, you decided to use the Cloud Storage connector to store all input data, output data, and intermediary data. However, you noticed that one Hadoop job runs very slowly with Cloud Dataproc, when compared with the on-premises bare-metal Hadoop environment (8-core nodes with 100-GB RAM). Analysis shows that this particular Hadoop job is disk I/O intensive. You want to resolve the issue. What should you do?

- A. Allocate sufficient memory to the Hadoop cluster, so that the intermediary data of that particular Hadoop job can be held in memory
- B. Allocate sufficient persistent disk space to the Hadoop cluster, and store the intermediate data of that particular Hadoop job on native HDFS
- C. Allocate more CPU cores of the virtual machine instances of the Hadoop cluster so that the networking bandwidth for each instance can scale up
- D. Allocate additional network interface card (NIC), and configure link aggregation in the operating system to use the combined throughput when working with Cloud Storage

[Show Suggested Answer](#)

by [rickywck](#) at March 17, 2020, 2:58 p.m.

## Comments

Type your comment...

Submit

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

Correct: B

Local HDFS storage is a good option if:

Your jobs require a lot of metadata operations—for example, you have thousands of partitions and directories, and each file size is relatively small.

You modify the HDFS data frequently or you rename directories. (Cloud Storage objects are immutable, so renaming a directory is an expensive operation because it consists of copying all objects to a new key and deleting them afterwards.)

You heavily use the append operation on HDFS files.

You have workloads that involve heavy I/O. For example, you have a lot of partitioned writes, such as the following:

```
spark.read().write.partitionBy(...).parquet("gs://")
```

You have I/O workloads that are especially sensitive to latency. For example, you require single-digit millisecond latency per storage operation.

   upvoted 39 times

  **Rajokkiyam** Highly Voted 4 years, 7 months ago

Answer B

Its google recommended approach to use LocalDisk/HDFS to store Intermediate result and use Cloud Storage for initial and final results.

   upvoted 15 times

  **Chelseajcole** 3 years, 1 month ago

Any link to support this recommended approach?

   upvoted 1 times

  **MaxNRG** Most Recent 10 months, 3 weeks ago

**Selected Answer: B**

Local HDFS storage is a good option if:

- You have workloads that involve heavy I/O. For example, you have a lot of partitioned writes such as the following:

```
spark.read().write.partitionBy(...).parquet("gs://")
```

- You have I/O workloads that are especially sensitive to latency. For example, you require single-digit millisecond latency per storage operation.

- Your jobs require a lot of metadata operations—for example, you have thousands of partitions and directories, and each file size is relatively small.

- You modify the HDFS data frequently or you rename directories. (Cloud Storage objects are immutable, so renaming a directory is an expensive operation because it consists of copying all objects to a new key and deleting them afterwards.)

- You heavily use the append operation on HDFS files.

   upvoted 1 times

  **MaxNRG** 10 months, 3 weeks ago


We recommend using Cloud Storage as the initial and final source of data in a big-data pipeline. For example, if a workflow contains five Spark jobs in series, the first job retrieves the initial data from Cloud Storage and then writes shuffle data and intermediate job output to HDFS. The final Spark job writes its results to Cloud Storage.

[https://cloud.google.com/solutions/migration/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#choose\\_storage\\_options](https://cloud.google.com/solutions/migration/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#choose_storage_options)

   upvoted 1 times

  **squishy\_fishy** 1 year ago

The correct answer is B.

   upvoted 1 times

  **barnac1es** 1 year, 1 month ago

**Selected Answer: B**

**Disk I/O Performance:** In a Cloud Dataproc cluster, the default setup uses local persistent disks for HDFS storage. These disks offer good disk I/O performance and are well-suited for storing intermediate data generated during Hadoop jobs.

**Data Locality:** Storing intermediate data on native HDFS allows for better data locality. This means that the data is stored on the same nodes where computation occurs, reducing the need for data transfer over the network. This can significantly improve the performance of disk I/O-intensive jobs.

**Scalability:** Cloud Dataproc clusters can be easily scaled up or down to meet the specific requirements of your jobs. You can

scalability. Cloud Dataproc clusters can be easily scaled up or down to meet the specific requirements of your jobs. You can allocate additional disk space as needed to accommodate the intermediate data generated by this particular Hadoop job.

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **DeepakVenkatachalam** 1 year, 1 month ago

Correct: A

I'd choose A as the doc states adding more SSDs are good for disk-intensive jobs especially those with many individual read and write operations

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

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **DeepakVenkatachalam** 1 year, 1 month ago

Typo Correct Answer is B. . Allocate sufficient persistent disk space to the Hadoop cluster, and store the intermediate data of that particular Hadoop job on native HDFS

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **arien\_chen** 1 year, 2 months ago

**Selected Answer: A**

I would choose A.

Google Storage is faster than HDFS in many cases.

<https://cloud.google.com/architecture/hadoop#:~:text=It%27s%20faster%20than%20HDFS%20in%20many%20cases.>

The question mention '(8-core nodes with 100-GB RAM)' on-premises Hadoop.

the problem may caused by insufficient memory, and does not mention cost would be an issue, so A 'memory' approach would be a better option.

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **vamgcp** 1 year, 3 months ago

**Selected Answer: B**

Best option is B. However allocating sufficient persistent disk space to the Hadoop cluster, and storing the intermediate data of that particular Hadoop job on native HDFS, would not improve the performance of the Hadoop job. In fact, it might even slow down the Hadoop job, as the data would have to be read and written to disk twice.

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **zellick** 1 year, 11 months ago

**Selected Answer: B**

B is the answer.

[https://cloud.google.com/architecture/hadoop/hadoop-gcp-migration-jobs#choosing\\_primary\\_disk\\_options](https://cloud.google.com/architecture/hadoop/hadoop-gcp-migration-jobs#choosing_primary_disk_options)

If your job is disk-intensive and is executing slowly on individual nodes, you can add more primary disk space. For particularly disk-intensive jobs, especially those with many individual read and write operations, you might be able to improve operation by adding local SSDs. Add enough SSDs to contain all of the space you need for local execution. Your local execution directories are spread across however many SSDs you add.

👍 ↩ 🚩 upvoted 4 times

🗄️ 👤 **John\_Pongthorn** 2 years, 1 month ago

**Selected Answer: B**

[https://cloud.google.com/architecture/hadoop/hadoop-gcp-migration-jobs#choosing\\_primary\\_disk\\_options](https://cloud.google.com/architecture/hadoop/hadoop-gcp-migration-jobs#choosing_primary_disk_options)

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **rrr000** 2 years, 2 months ago

B is not the right answer. The problem says that for intermediate data cloud storage is to be used, while B option says:

B ... the intermediate data of that particular Hadoop job on native HDFS

A is the right answer. If you have enough memory then the shuffle wont spill on the disk.

👍 ↩ 🚩 upvoted 3 times

🗄️ 👤 **rrr000** 2 years, 2 months ago

Further the question states that original on prem machines has 100gb ram.

8-core nodes with 100-GB RAM

👍 ↩ 🚩 upvoted 2 times

🗄️ 👤 **SoerenE** 2 years, 9 months ago

B should be the right answer: [https://cloud.google.com/compute/docs/disks/performance#optimize\\_disk\\_performance](https://cloud.google.com/compute/docs/disks/performance#optimize_disk_performance)

👍 ↩ 🚩 upvoted 1 times

🗄️ 👤 **medeis iar** 2 years, 10 months ago

— — **meade\_jar** 3 years, 10 months ago

**Selected Answer: B**

<https://cloud.google.com/solutions/migration/hadoop/hadoop-gcp-migration-jobs>

👍 ↩ 🚩 upvoted 3 times

☐ **JG123** 2 years, 11 months ago

Why there are so many wrong answers? Examtopics.com are you enjoying paid subscription by giving random answers from people?

Ans: B

👍 ↩ 🚩 upvoted 3 times

☐ **RT30** 3 years, 7 months ago

If your job is disk-intensive and is executing slowly on individual nodes, you can add more primary disk space. For particularly disk-intensive jobs, especially those with many individual read and write operations, you might be able to improve operation by adding local SSDs. Add enough SSDs to contain all of the space you need for local execution. Your local execution directories are spread across however many SSDs you add.

Its B

<https://cloud.google.com/solutions/migration/hadoop/hadoop-gcp-migration-jobs>

👍 ↩ 🚩 upvoted 3 times

☐ **ashuchip** 3 years, 10 months ago

yes B is correct

👍 ↩ 🚩 upvoted 2 times

☐ **Alasmindas** 3 years, 11 months ago

Correct Answer is Option B - Adding persistent disk space, reasons:-

- The question mentions that the particular job is "disk I/O intensive" - so the word "disk" is explicitly mentioned.
- Option B also mentions about local HDFS storage, which is ideally a good option of general I/O intensive work.

👍 ↩ 🚩 upvoted 5 times

[Load full discussion...](#)



## Platform

> Home

> All Exams

> Examtopics PRO

> Training Courses



© 2024 ExamTopics