

🔗 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 146 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 146

Topic #: 1

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

You want to migrate an on-premises Hadoop system to Cloud Dataproc. Hive is the primary tool in use, and the data format is Optimized Row Columnar (ORC).

All ORC files have been successfully copied to a Cloud Storage bucket. You need to replicate some data to the cluster's local Hadoop Distributed File System

(HDFS) to maximize performance. What are two ways to start using Hive in Cloud Dataproc? (Choose two.)

- A. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to HDFS. Mount the Hive tables locally.
- B. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to any node of the Dataproc cluster. Mount the Hive tables locally.
- C. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to the master node of the Dataproc cluster. Then run the Hadoop utility to copy them to HDFS. Mount the Hive tables from HDFS.
- D. Leverage Cloud Storage connector for Hadoop to mount the ORC files as external Hive tables. Replicate external Hive tables to the native ones.
- E. Load the ORC files into BigQuery. Leverage BigQuery connector for Hadoop to mount the BigQuery tables as external Hive tables. Replicate external Hive tables to the native ones.

[Show Suggested Answer](#)

by [deleted] at March 22, 2020, 8:31 a.m.

Comments

Comments

Type your comment...

Submit

  **Sid19** Highly Voted  3 years, 4 months ago

Answer is C and D 100%.

I know it says to transfer all the files but with the options provided c is the best choice.

Explanation

A and B cannot be true as gsutil can copy data to master node and then to HDFS from master node.

C -> works

D->works Recommended by google

E-> Will work but as the question says maximize performance this is not a case. As BigQuery Hadoop connector stores all the BQ data to GCS as temp and then processes it to HDFS. As data is already in GCS we don't need to load it to BQ and use a connector then unloads it back to GCS and then processes it.

   upvoted 26 times

  **DeepakD** 3 years, 3 months ago

How can master node store data? C is wrong.

   upvoted 2 times

  **KLei** 1 year, 6 months ago

must go to the master node first...

   upvoted 1 times

  **[Removed]** 3 years, 3 months ago

option D is mentioned in the new release 2019: <https://cloud.google.com/blog/products/data-analytics/new-release-of-cloud-storage-connector-for-hadoop-improving-performance-throughput-and-more>


   upvoted 2 times

  **WillemHendr** 1 year, 11 months ago

I feel indeed this question is testing if you understand, that gsutil cannot transfer to HDFS directly (eliminate A&B), and need an intermediate step, (making C doable, with a good result). D is found on official Google Docs. E doesn't have good end result.




   upvoted 3 times

[Load full discussion...](#)

  **[Removed]** Highly Voted  5 years, 1 month ago

Should be B C

   upvoted 17 times

  **shangning007** Most Recent  4 months, 2 weeks ago

Selected Answer: AD

<https://stackoverflow.com/questions/54429642/how-to-copy-a-file-from-a-gcs-bucket-in-dataproc-to-hdfs-using-google-cloud>
Based on here, you can copy a single file from Google Cloud Storage (GCS) to HDFS using the HDFS copy command. There is no need to copy to the master node first.

   upvoted 1 times

  **SamuelTsch** 6 months, 1 week ago

Selected Answer: CD

Actually I think A is correct as well.

   upvoted 1 times

  **patitonav** 1 year, 4 months ago

Selected Answer: DE

I think D and E are the best and easy way to go. For sure D, but I think that E can work too, the data can be loaded in BQ as an external table, so at the end the data will be always on the GCS.

   upvoted 2 times

  **barnac1es** 1 year, 7 months ago

Selected Answer: DE

D. Cloud Storage Connector for Hadoop: You can use the Cloud Storage connector for Hadoop to mount the ORC files stored in Cloud Storage as external Hive tables. This allows you to query the data without copying it to HDFS. You can replicate these external Hive tables to native Hive tables in Cloud Dataproc if needed.

E. Load ORC Files into BigQuery: Another approach is to load the ORC files into BigQuery, Google Cloud's data warehouse. Once the data is in BigQuery, you can use the BigQuery connector for Hadoop to mount the BigQuery tables as external

Hive tables in Cloud Dataproc. This leverages the power of BigQuery for analytics and allows you to replicate external Hive tables to native ones in Cloud Dataproc.

   upvoted 1 times

  **barnac1es** 1 year, 7 months ago

Selected Answer: DE

D. Cloud Storage Connector for Hadoop: You can use the Cloud Storage connector for Hadoop to mount the ORC files stored in Cloud Storage as external Hive tables. This allows you to query the data without copying it to HDFS. You can replicate these external Hive tables to native Hive tables in Cloud Dataproc if needed.

E. Load ORC Files into BigQuery: Another approach is to load the ORC files into BigQuery, Google Cloud's data warehouse. Once the data is in BigQuery, you can use the BigQuery connector for Hadoop to mount the BigQuery tables as external Hive tables in Cloud Dataproc. This leverages the power of BigQuery for analytics and allows you to replicate external Hive tables to native ones in Cloud Dataproc.

   upvoted 1 times

  **vamgcp** 1 year, 9 months ago

Selected Answer: AD

A is the most straightforward way to start using Hive in Cloud Dataproc. You can use the gsutil utility to transfer all ORC files from the Cloud Storage bucket to HDFS. Then, you can mount the Hive tables locally.

D is another option that you can use to start using Hive in Cloud Dataproc. You can leverage the Cloud Storage connector for Hadoop to mount the ORC files as external Hive tables. Then, you can replicate the external Hive tables to the native ones.

   upvoted 2 times

  **Qix** 1 year, 9 months ago

Selected Answer: BC

Answers are;


B. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to any node of the Dataproc cluster. Mount the Hive tables locally.

C. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to the master node of the Dataproc cluster. Then run the Hadoop utility to copy them to HDFS. Mount the Hive tables from HDFS.

You need to replicate some data to the cluster's local Hadoop Distributed File System (HDFS) to maximize performance. HDFS lies on datanode, data on master node needs to be copied on datanode.

B for managed hive table option, C for external hive table

   upvoted 3 times

  **izekc** 1 year, 12 months ago

Selected Answer: AD

AD is correct

   upvoted 1 times

  **Oleksandr0501** 2 years ago

Selected Answer: AD

i choose AD.

Searched in other w/s, read discussions here, and guess better AD.

   upvoted 1 times

  **Oleksandr0501** 2 years ago

gpt: Yes, that is correct. Option A is a valid way to transfer the ORC files to HDFS, and then mount the Hive tables locally. Option D is also valid, as it suggests using the Cloud Storage connector for Hadoop to mount the ORC files as external Hive tables and then replicating those external Hive tables to native ones.

Chatgpt agreed, after inserting question and variants, and said that AD are correct answers. And it agreed. It adds some confidence that these are good, but gpt can make mistakes

   upvoted 1 times

  **streeber** 2 years ago

Selected Answer: AD

A will copy to HDFS and so will D

   upvoted 1 times

  **hauhau** 2 years, 5 months ago

Selected Answer: AD



C: master node doesn't make sense

   upvoted 3 times

  **hauhau** 2 years, 5 months ago

B: from the Cloud Storage bucket to any node of the Dataproc cluster
-> still on cloud not maxize the speed

   upvoted 1 times

  **zellck** 2 years, 5 months ago

Selected Answer: CD

CD is the answer.

<https://cloud.google.com/dataproc/docs/concepts/connectors/cloud-storage>

The Cloud Storage connector is an open source Java library that lets you run Apache Hadoop or Apache Spark jobs directly on data in Cloud Storage, and offers a number of benefits over choosing the Hadoop Distributed File System (HDFS).

Connector Support. The Cloud Storage connector is supported by Google Cloud for use with Google Cloud products and use cases, and when used with Dataproc is supported at the same level as Dataproc.

   upvoted 4 times

  **tikki_boy** 2 years, 6 months ago

I'll go with DE

   upvoted 2 times

  **ducc** 2 years, 8 months ago

Selected Answer: CD

CD is correct

   upvoted 2 times

  **BigDataBB** 3 years, 2 months ago

Selected Answer: BC

You need to replicate some data to the cluster's local Hadoop Distributed File System (HDFS) to maximize performance.

HDFS lies on datanode, data on masternode needs to be copied on datanode.

B for managed hive table option, C for external hive table

   upvoted 1 times

[Load full discussion...](#)



Platform

> Home

> Examtopics PRO

> All Exams

> Training Courses

