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 17 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 17

Topic #: 1

[All Professional Data Engineer Questions]

Your company is migrating their 30-node Apache Hadoop cluster to the cloud. They want to re-use Hadoop jobs they have already created and minimize the management of the cluster as much as possible. They also want to be able to persist data beyond the life of the cluster. What should you do?

A. Create a Google Cloud Dataflow job to process the data.

B. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.

C. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.

D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.

E. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.

Show Suggested Answer

by [deleted] at *March 16, 2020, 11:26 a.m.*

Comments

Type your comment...

Submit

⊟ MaxNRG Highly Voted 3 years, 5 months ago

D is correct because it uses managed services, and also allows for the data to persist on GCS beyond the life of the cluster.
A is not correct because the goal is to re-use their Hadoop jobs and MapReduce and/or Spark jobs cannot simply be moved to Dataflow.
B is not correct because the goal is to persist the data beyond the life of the ephemeral clusters, and if HDFS is used as the primary attached storage mechanism, it will also disappear at the end of the cluster's life.
C is not correct because the goal is to use managed services as much as possible, and this is the opposite.
E is not correct because the goal is to use managed services as much as possible, and this is the opposite.

upvoted 12 times

certs4pk 5 months, 1 week ago

B is incorrect bcoz, it did not say 'off cluster' persistent HDFS discs???

upvoted 1 times

Radhika7983 Highly Voted 4 years, 6 months ago

The correct answer is D. Here is the explanation to why Data proc and why not Data flow.
When a company wants to move their existing Hadoop jobs on premise to cloud, we can simply move the jobs in cloud data prod and replace hdfs with gs:// which is google storage. This way you are keeping compute and storage separately. Hence the correct answer is D. However, if the company wants to complete create a new jobs and don't want to use the existing Hadoop jobs running on premise, the option is to create new data flow jobs.

upvoted 6 times

suku2 Most Recent 7 months, 1 week ago

Selected Answer: D

D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.
Dataproc clusters can be created to lift and shift existing Hadoop jobs
Data stored in Google Cloud Storage extends beyond the life of a Dataproc cluster.

upvoted 2 times

imran79 7 months, 1 week ago

D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.

Here's why:

Cloud Dataproc allows you to run Apache Hadoop jobs with minimal management. It is a managed Hadoop service.

Using the Google Cloud Storage (GCS) connector, Dataproc can access data stored in GCS, which allows data persistence beyond the life of the cluster. This means that even if the cluster is deleted, the data in GCS remains intact. Moreover, using GCS is often cheaper and more durable than using HDFS on persistent disks.

upvoted 1 times

certs4pk 5 months, 1 week ago

what if option B said, 'off cluster' persistent HDFS disks?

upvoted 1 times

rtcpost 7 months, 1 week ago

Selected Answer: D

D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.

Google Cloud Dataproc is a managed Hadoop and Spark service that allows you to easily create and manage Hadoop clusters in the cloud. By using the Google Cloud Storage connector, you can persist data in Google Cloud Storage, which provides durable storage beyond the cluster's lifecycle. This approach ensures data is retained even if the cluster is terminated, and it allows you to reuse your existing Hadoop jobs.

Option B (Creating a Dataproc cluster that uses persistent disks for HDFS) is another valid choice. However, using Google Cloud Storage for data storage and processing is often more cost-effective and scalable, especially when migrating to the cloud.

Options A, C, and E do not take full advantage of Google Cloud's services and the benefits of cloud-native data storage and processing with Google Cloud Storage and Dataproc.

upvoted 3 times

fahadminhas 10 months, 1 week ago

Option D is incorrect, as it would not provide persistent HDFS storage within cluster itself. Rather B should be the correct answer.

upvoted 1 times

kshehadyx 1 year, 7 months ago

Correct D

upvoted 1 times

bha11111 2 years, 1 month ago

Selected Answer: D

Hadoop --> Dataproc Persistent storage after the processing --> GCS

upvoted 2 times

samdhimal 2 years, 3 months ago

Selected Answer: D

D Seems right. Cloud storage can be used to achieve data storage even after the life of cluster.

upvoted 1 times

korntewin 2 years, 3 months ago

Selected Answer: D

The answer is D! Dataproc have no need for use to manage the infra and cloudstorage also no need for us to manage too!

upvoted 1 times

Nirca 2 years, 4 months ago

Selected Answer: D

D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.

upvoted 1 times

assU2 2 years, 6 months ago

Seems like it is D. https://cloud.google.com/dataproc/docs/concepts/dataproc-hdfs
Never saw they mentioned persistent disks, although they are not deleted with the clusters...

upvoted 1 times

assU2 2 years, 6 months ago

although:
By default, when no local SSDs are provided, HDFS data and intermediate shuffle data is stored on VM boot disks, which are Persistent Disks.

upvoted 1 times

assU2 2 years, 6 months ago

and it says that only VM Boot disks are deleted when the cluster is deleted.

upvoted 2 times

achafill 2 years, 6 months ago

Selected Answer: D

Correct Answer : D

upvoted 1 times

nkunwar 2 years, 7 months ago

Selected Answer: D

Dataproc cluster set up will be ephemeral to run HDFS Jobs and can be killed after Job execution killing persistent storage with cluster

upvoted 1 times

crisimenjivar 2 years, 8 months ago

Anwer: D

upvoted 1 times

Asheesh1909 2 years, 11 months ago

Isn't it A and D both dataflow for reusable jobs and gcs for data peraistance?

upvoted 1 times

kmaiti 3 years ago

Selected Answer: D

Two key points:
Managed hadoop cluster - dataproc
Persistent storage: GCS (dataproc uses gcs connector to connect to gcs)

upvoted 2 times

Load full discussion...

Platform

- Home
- All Exams
- Examtopics PRO
- Training Courses