Assignment and Exam Content

Cloud Dataproc

Cloud Gurus Seattle, USA
Training material

Always Delete your Cloud Resources to Avoid $$ Charges.

Cloud dataproc Lab

Cloud dataproc Lab Contains following topics

A Launch Cloud dataproc Instance

Understand basic concepts – Locations, Performance Read/ Write IOPS etc

B Launch Cloud Dataproc instance

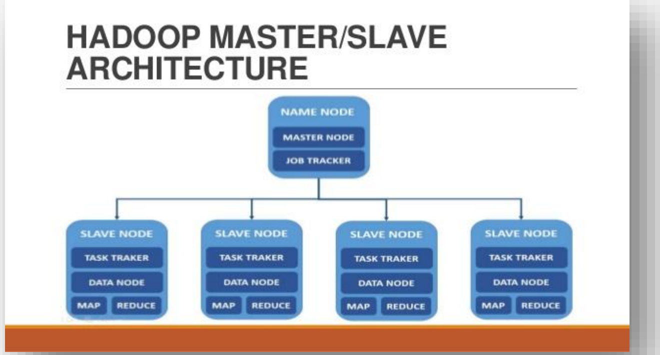
C Exam TIPS

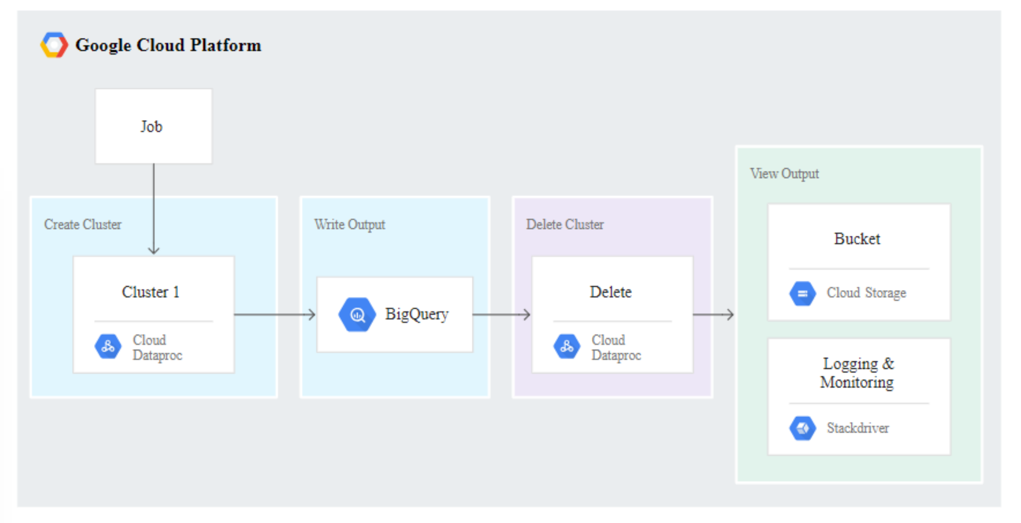
Cloud Dataproc is Hadoop and Spark clusters in GCP.

Cloud 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.

You can execute one of Jobs launching cluster OR

You can have continuous cluster running.



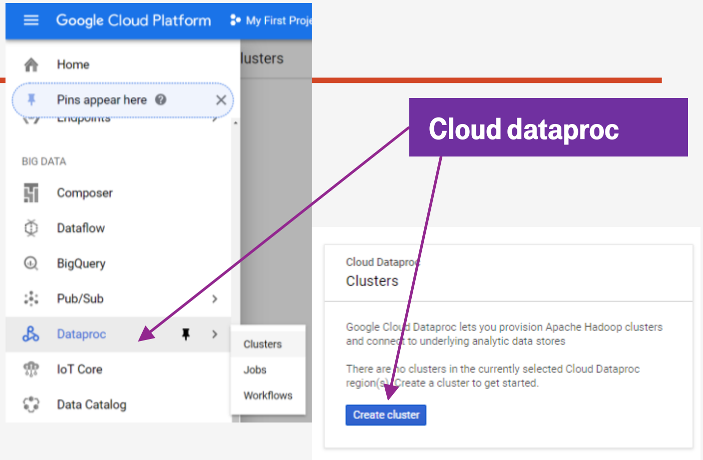


Always Delete your Cloud Resources to Avoid $$ Charges.

A

Create Cloud Dataproc Instance

1 Go To -> BIGDATA -> Dataproc -> Cluster -> Create Cluster



Google Cloud Dataproc lets you provision Apache Hadoop clusters and connect to underlying analytic data stores

There are no clusters in the currently selected Cloud Dataproc region(s). Create a cluster to get started.

Cloud dataproc

You will need to provide different parameter

2 Name : Name of cluster

Location : Could be Regional or Global .

If you select Region -> Zone is either selected by Dataproc or you can select you're your zone

If you select global -. You will need to specify which Zone the master should go

Master Node

Machine Type and other configuration – like Disk Type and Size

You can always customize CPU and RAM configuration based on your need

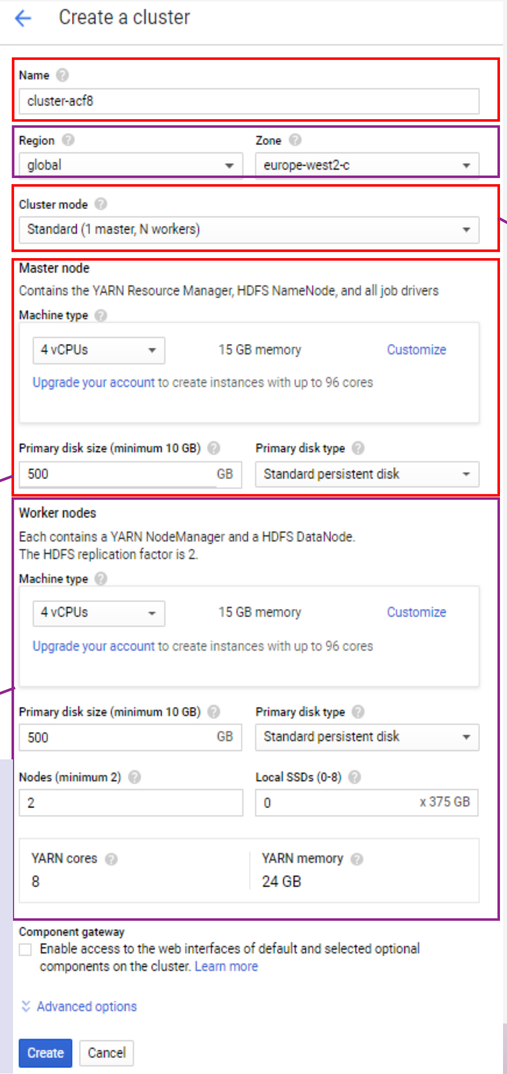
Worker Node

If its Standard or HA cluster – You can choose configuration for worker node.

With primary disk type and disk size.

You can specify number of Nodes and optionally attach local SSD's to worker nodes for data processing.

You can configure Yarn cores and Memory to run Yarn damain.



Create a cluster

Name: cluster-acf8

Region: global
Zone: europe-west2-c

Cluster mode: Standard (1 master, N workers)

Master node

Contains the YARN Resource Manager, HDFS NameNode, and all job drivers

Machine type: 4 vCPUs 15 GB memory Customize

Upgrade your account to create instances with up to 96 cores

Primary disk size (minimum 10 GB): 500 GB
Primary disk type: Standard persistent disk

Worker nodes

Each contains a YARN NodeManager and a HDFS DataNode.
The HDFS replication factor is 2.

Machine type: 4 vCPUs 15 GB memory Customize

Upgrade your account to create instances with up to 96 cores

Primary disk size (minimum 10 GB): 500 GB
Primary disk type: Standard persistent disk

Nodes (minimum 2): 2
Local SSDs (0-8): 0 x 375 GB

YARN cores: 8
YARN memory: 24 GB

Component gateway

□ Enable access to the web interfaces of default and selected optional components on the cluster. Learn more

Advanced options

Create Cancel

Cluster Mode

You can select different options

- Single Node –> Same Master and worker node
- Standard -> 1 Master and multiple Worker Node
- HA -> 3 Master (1 Active and two standby.) and multiple worker nodes

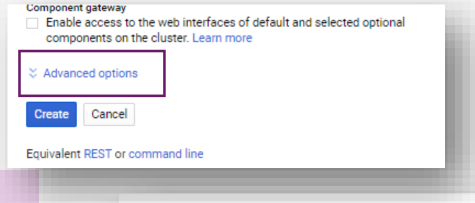
Depends on type of Cluster Mode you select – You will have choice of different configuration –

e.g. If its Single Node – Dataproc will not ask you worker node configuration.

1

Create Cloud Dataproc Instance

3 Advanced Options – Additional Configurations -> You can optionally supply additional parameters for Clusters, lets see what are those.



Component gateway
☐ Enable access to the web interfaces of default and selected optional components on the cluster. Learn more

Equivalent REST or command line

Component Gateway

You can enable web interface access – These are Hadoop and spark web interface comes along as open source Apache Hadoop and Apache Spark software. – By default they are not enabled.

Preemptible worker Nodes

Preemptible worker nodes can be used to reduce cost for worker node. Go through limitations for these nodes.

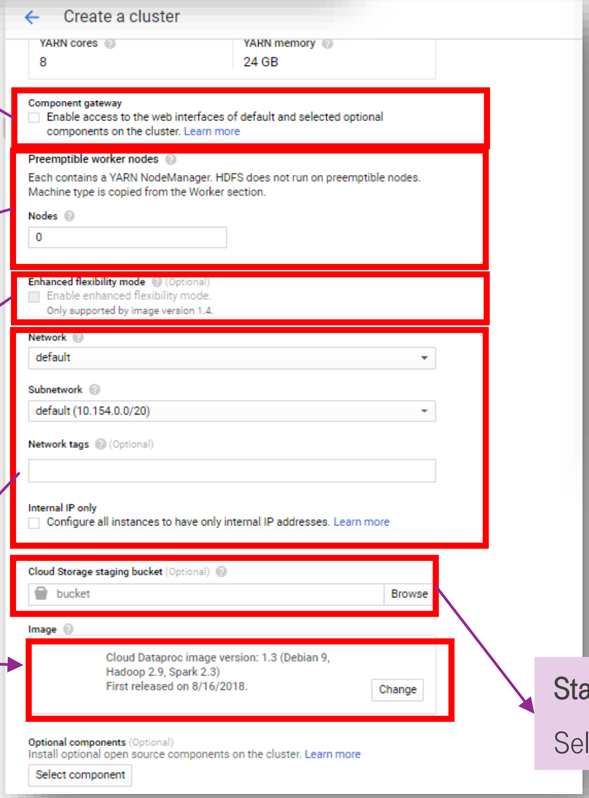
Flexible Mode

When a Cloud Dataproc node is removed, Cloud Dataproc Enhanced Flexibility Mode preserves stateful node data, such as mapreduce shuffle data – You will need to be latest image – version1.4 is required.

You can go to Image and change version here

Network

You can choose different networks here – Details in next network section. You can keep it default for now.



Create a cluster

YARN cores: 8
YARN memory: 24 GB

Component gateway
☐ Enable access to the web interfaces of default and selected optional components on the cluster. Learn more

Preemptible worker nodes
Each contains a YARN NodeManager. HDFS does not run on preemptible nodes. Machine type is copied from the Worker section.
Nodes: 0

Enhanced flexibility mode (Optional)
☐ Enable enhanced flexibility mode.
Only supported by image version 1.4.

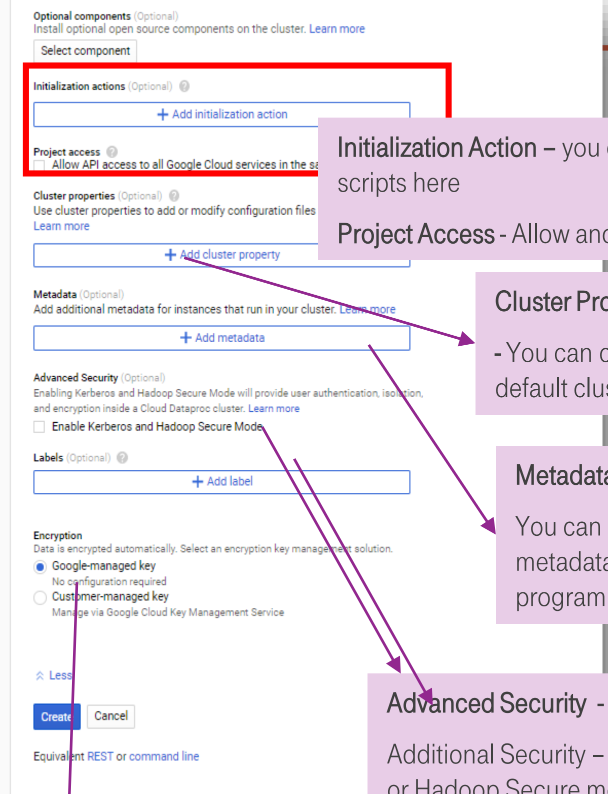
Network: default
Subnetwork: default (10.154.0.0/20)
Network tags (Optional)

Internal IP only
☐ Configure all instances to have only internal IP addresses. Learn more

Cloud Storage staging bucket (Optional)
bucket

Image
Cloud Dataproc image version: 1.3 (Debian 9, Hadoop 2.9, Spark 2.3)
First released on 8/16/2018.

Optional components (Optional)
Install optional open source components on the cluster. Learn more



Optional components (Optional)
Install optional open source components on the cluster. Learn more

Initialization actions (Optional)

Project access
☐ Allow API access to all Google Cloud services in the sa

Cluster properties (Optional)
Use cluster properties to add or modify configuration files
Learn more

Metadata (Optional)
Add additional metadata for instances that run in your cluster. Learn more

Advanced Security (Optional)
Enabling Kerberos and Hadoop Secure Mode will provide user authentication, isolation, and encryption inside a Cloud Dataproc cluster. Learn more
☐ Enable Kerberos and Hadoop Secure Mode

Labels (Optional)

Encryption
Data is encrypted automatically. Select an encryption key management solution.
● Google-managed key
No configuration required
○ Customer-managed key
Manage via Google Cloud Key Management Service

Equivalent REST or command line

Initialization Action – you can provide initializations scripts here

Project Access - Allow and disallow API access.

Cluster Properties

- You can change configuration of default cluster params

Metadata -

You can add additional instance metadata which can be retrieved at programmatically

Advanced Security -

Additional Security – Like Kerberos or Hadoop Secure mode.

Encryption – Google or your own key

Staging Area – You can use Cloud Storage bucket to store data.

Select bucket – or create bucket and select here.

1

Create Cloud Dataproc Instance

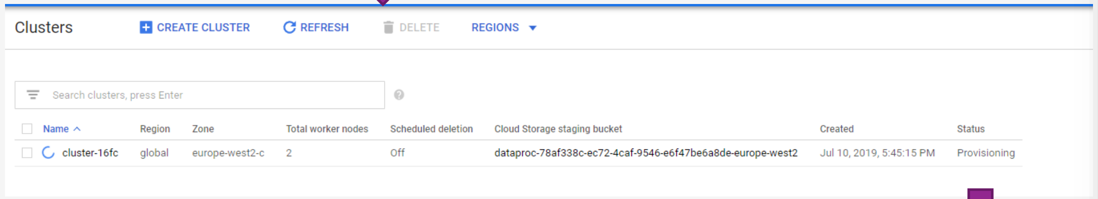
4

Create Standard node Cluster

Choose your own name and lowest machine type and other configuration as shown.

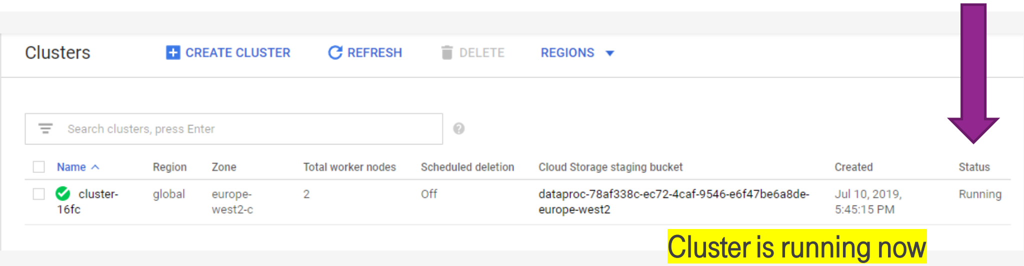
Hit Create button

Cluster is being created



Search clusters, press Enter

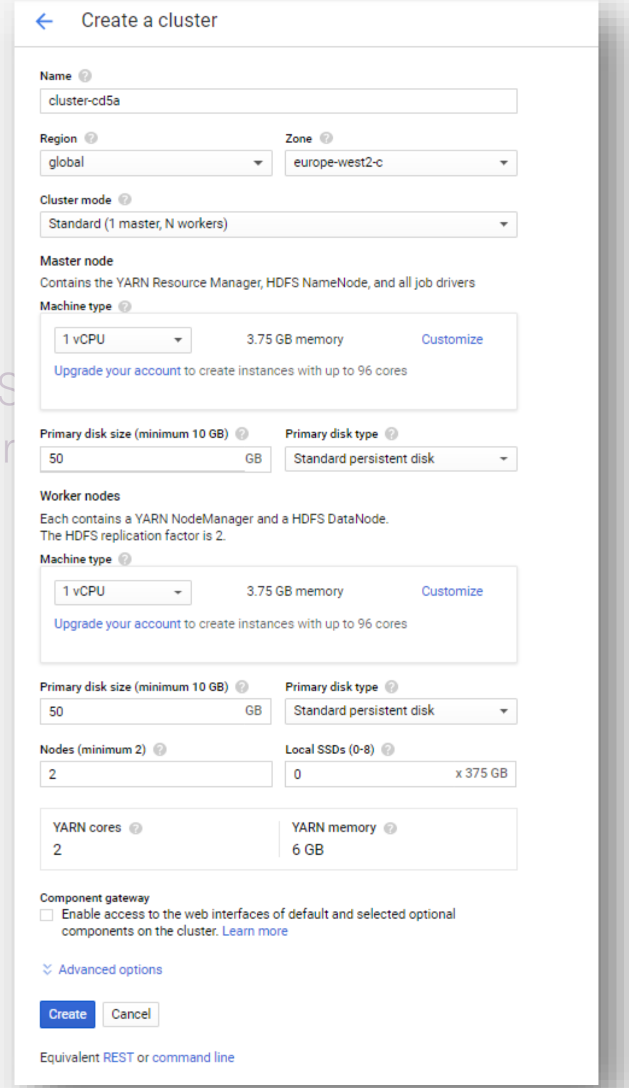
Name	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created	Status
cluster-16fc	global	europe-west2-c	2	Off	dataproc-78af338c-ec72-4caf-9546-e6f47be6a8de-europe-west2	Jul 10, 2019, 5:45:15 PM	Provisioning



Search clusters, press Enter

Name	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created	Status
cluster-16fc	global	europe-west2	2	Off	dataproc-78af338c-ec72-4caf-9546-e6f47be6a8de-europe-west2	Jul 10, 2019, 5:45:15 PM	Running

Cluster is running now



Create a cluster

Name: cluster-cd5a

Region: global

Zone: europe-west2-c

Cluster mode: Standard (1 master, N workers)

Master node

Contains the YARN Resource Manager, HDFS NameNode, and all job drivers

Machine type: 1 vCPU   3.75 GB memory   Customize

Upgrade your account to create instances with up to 96 cores

Primary disk size (minimum 10 GB): 50 GB

Primary disk type: Standard persistent disk

Worker nodes

Each contains a YARN NodeManager and a HDFS DataNode.
The HDFS replication factor is 2.

Machine type: 1 vCPU   3.75 GB memory   Customize

Upgrade your account to create instances with up to 96 cores

Primary disk size (minimum 10 GB): 50 GB

Primary disk type: Standard persistent disk

Nodes (minimum 2): 2

Local SSDs (0-8): 0 x 375 GB

YARN cores: 2

YARN memory: 6 GB

Component gateway

□ Enable access to the web interfaces of default and selected optional components on the cluster. Learn more

Advanced options

Equivalent REST or command line

1

Create Cloud Dataproc Instance

4 Explore the dashboard

Monitoring : see what's in dashboard, Click on Stackdriver Logging to see logs

Jobs : You can submit jobs here.
Please follow instructions using following link to submit jobs
https://cloud.google.com/dataproc/docs/guides/submit-job

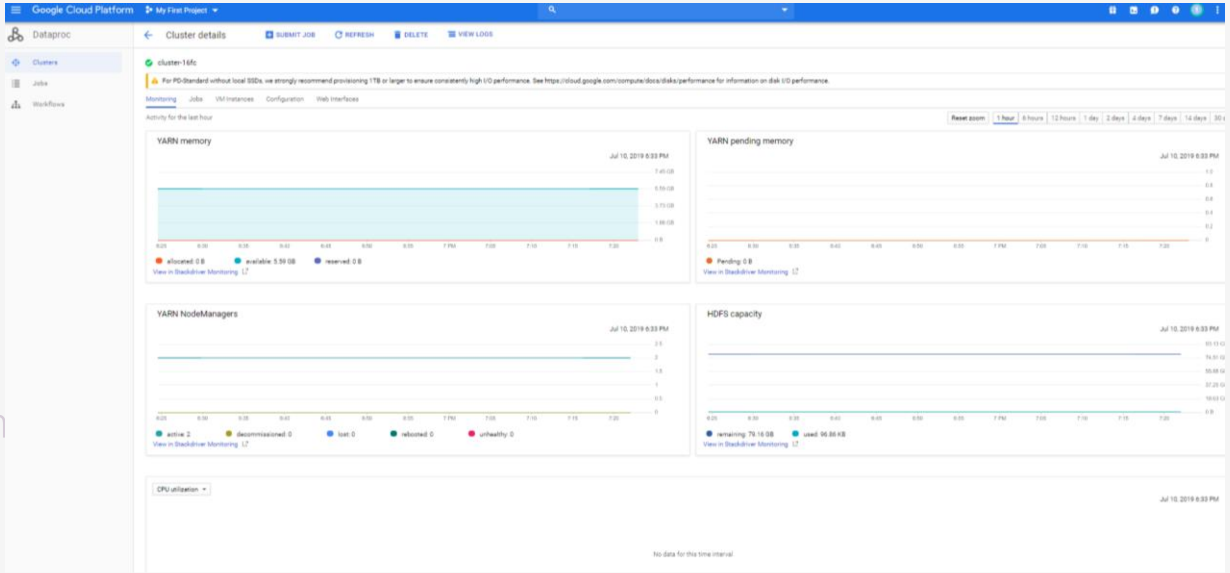
VM Instance : Observe Instances
You can ssh to master node by default. You can go to Compute -> VM instance and find these instances as well.
Now SSH to one of node

```
$ ps -aef | grep hadoop
```

```
$ps –aef  | grep yarn
```

Try same thing to Master node. – there are different processed running

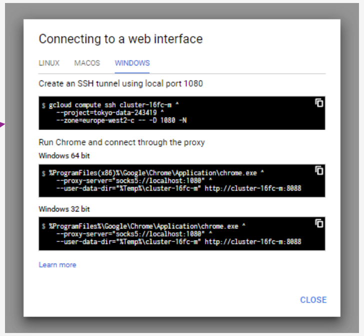
Configurations : Properties of cluster , including Hadoop/spark properties



For PD-Standard without local SSDs, we strongly recommend provisioning 1TB or larger to ensure consistently high I/O performance. See https://cloud.google.com/compute/docs/disks/performance for information on disk I/O performance.

Web Interface
- You need to create SSH tunnel for Web interface to work.
- Click on it and follow instructions

Click on View Logs – and see Stackdriver Logging opens Observe logs.



Connecting to a web interface

Create an SSH tunnel using local port 1080

```
$ gcloud compute ssh cluster-16fc-m ^
    --project=tokyo-data-243419 ^
    --zone=europe-west2-c -- -D 1080 -N
```

Run Chrome and connect through the proxy

Windows 64 bit

```
$ %ProgramFiles(x86)%\Google\Chrome\Application\chrome.exe ^
    --proxy-server="socks5://localhost:1080" ^
    --user-data-dir="%Temp%\cluster-16fc-m" http://cluster-16fc-m:8088
```

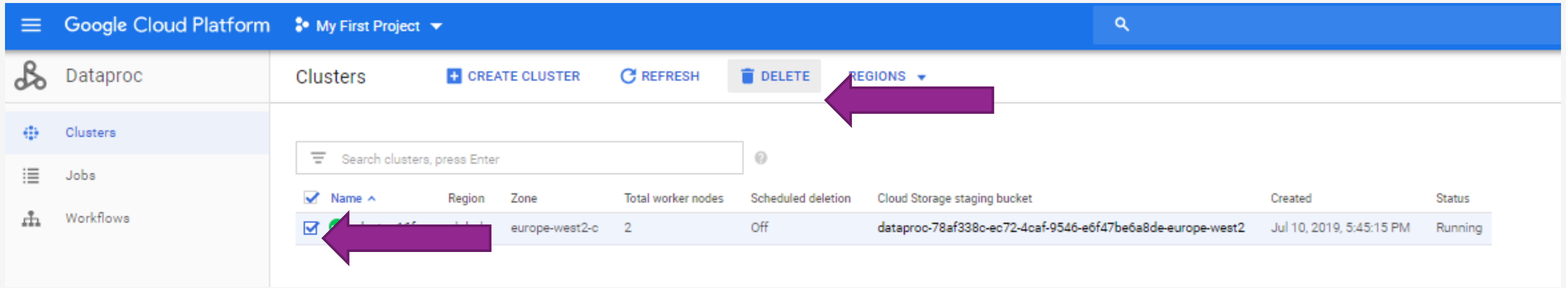
Windows 32 bit

```
$ %ProgramFiles%\Google\Chrome\Application\chrome.exe ^
    --proxy-server="socks5://localhost:1080" ^
    --user-data-dir="%Temp%\cluster-16fc-m" http://cluster-16fc-m:8088
```

2

Create Cloud Dataproc Instance

1



☑	Name	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created	Status
☑	[illegible]	[illegible]	europe-west2-c	2	Off	dataproc-78af338c-ec72-4caf-9546-e6f47be6a8de-europe-west2	Jul 10, 2019, 5:45:15 PM	Running

Before Creating next Cluster

- Delete Old cluster and Proceed further

Always Delete your Cloud Resources to Avoid $$ Charges.

Cloud Dataproc : Try Yourself

1 Create – Single Node cluster and observe the properties

2 Create HA Cluster and observe the properties – Multiple master nodes , go to Compute Engine and stop one of the master node and see behavior of cluster.

3 Exam Tips

Important concepts are

1. **Why it is used**
2. **Different between dataproc and dataflow**
3. **What kind of jobs does dataproc supports – Spark and Hadoop,**
4. **If you do not want to modify existing Hadoop/Soark cluster and want to take it to GCP**
5. **Use of Preemptible VM as node**
6. **How will you preserve intermediate data**

Cloud Dataflow

Cloud Dataflow is typically the preferred option for greenfield environments:

- Less operational overhead
- Unified approach to development of batch or streaming pipelines
- Uses Apache Beam
- Supports pipeline portability across Cloud Dataflow, Apache Spark, and Apache Flink as runtimes

Cloud Dataproc

Cloud Dataproc is good for environments dependent on specific components of the Apache big data ecosystem:

- Tools/packages
- Pipelines
- Skill sets of existing resources

gcloud dataproc command domains

gcloud

- Clusters
  - Create
  - Delete
  - Describe
  - List
  - ….
- Operations
  - Cancel
  - Delete
  - Describe
  - List
- Workflow-template
  - Add-job
  - Create
  - Delete
  - Export
  - Get-iam-policy
  - Import
  - Instantiate
  - List
  - Remove-job
  - Set and get iam policy
  - Set-managed-cluster
- Jobs
  - Delete
  - Describe
  - Get-iam-policy
  - Kill
  - List
  - Set-iam policy
  - Submit
    - Hadoop
    - Hive
    - Pig
    - Spark
    - Pyspark
    - Spark
    - Spark-sql
  - Update
  - Wait

Cloud Dataproc

End of Cloud Dataproc Assignment