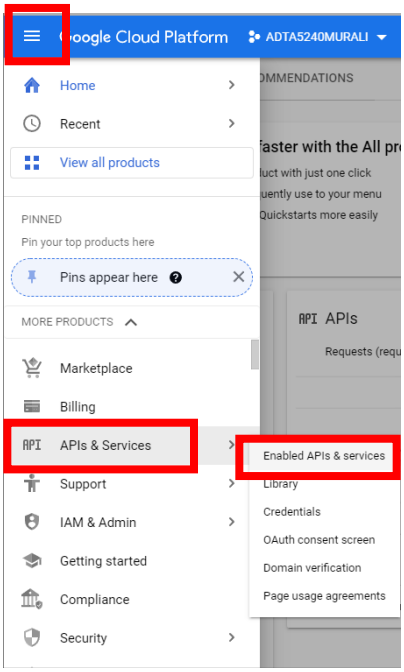
In this manual we will learn:

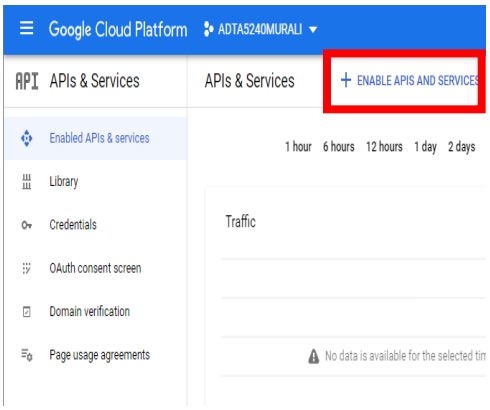
- To create a Hadoop and Spark cluster.

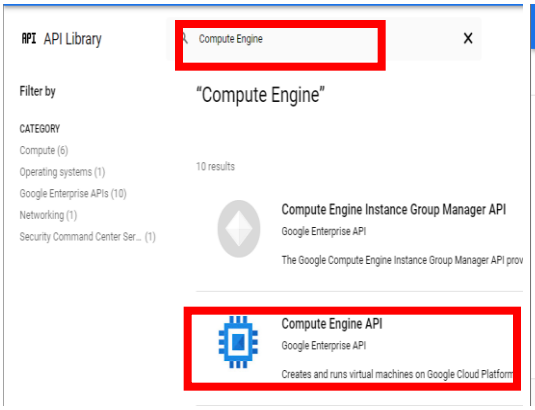
  Follow the steps to create 1 master node and 2 worker nodes

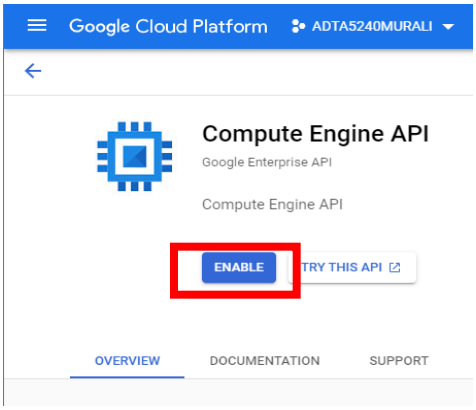
  ⇨ Log into your GCP account through your registered Gmail id.
  ⇨ Click on the Navigation panel (three horizontal lines on the top left corner).
  ⇨ Scroll to “APIs and Services” and select “Enabled API and Services”.



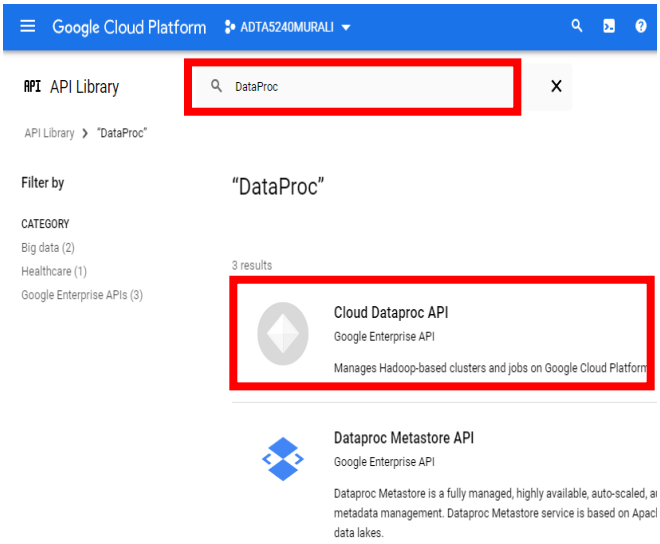
  ⇨ Select “ + Enable APIs and Services” as shown below and search for “Compute Engine API” and select “Enable”.

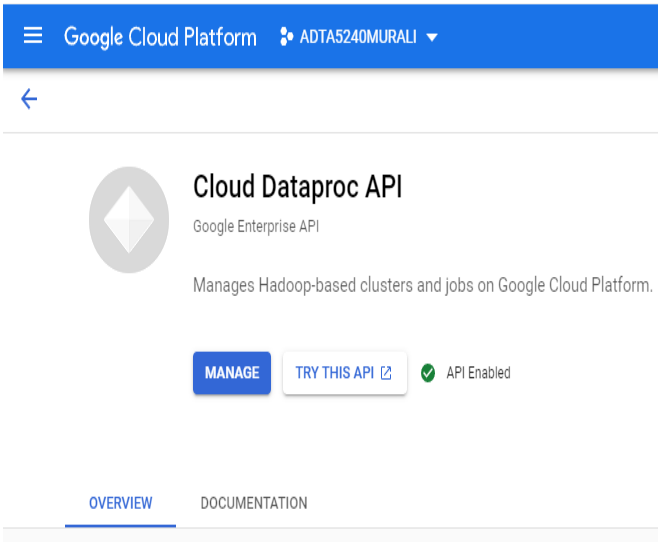






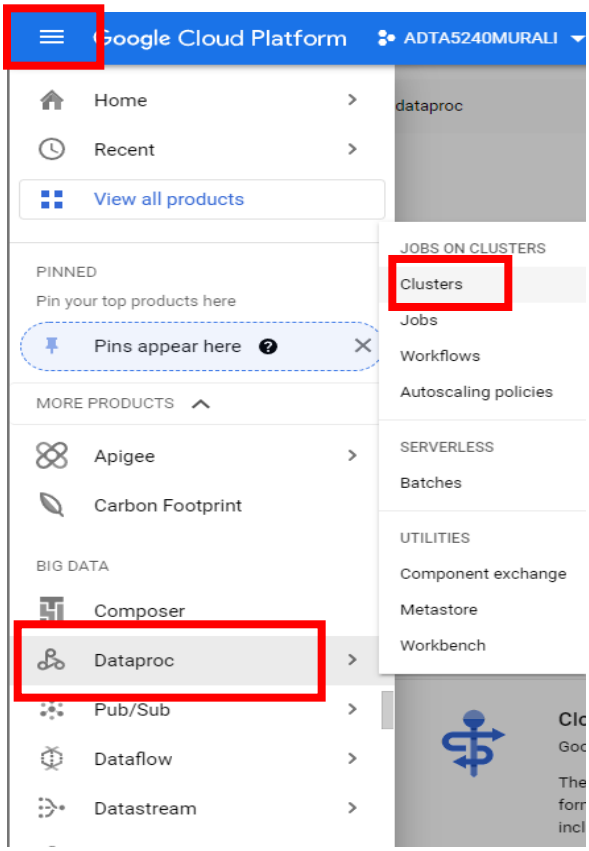
⇨ Next, we are going to download “Cloud DataProc API”. Search for “DataProc”. Select “Enable”. Since I already enabled it, it only gives me an option to manage it.

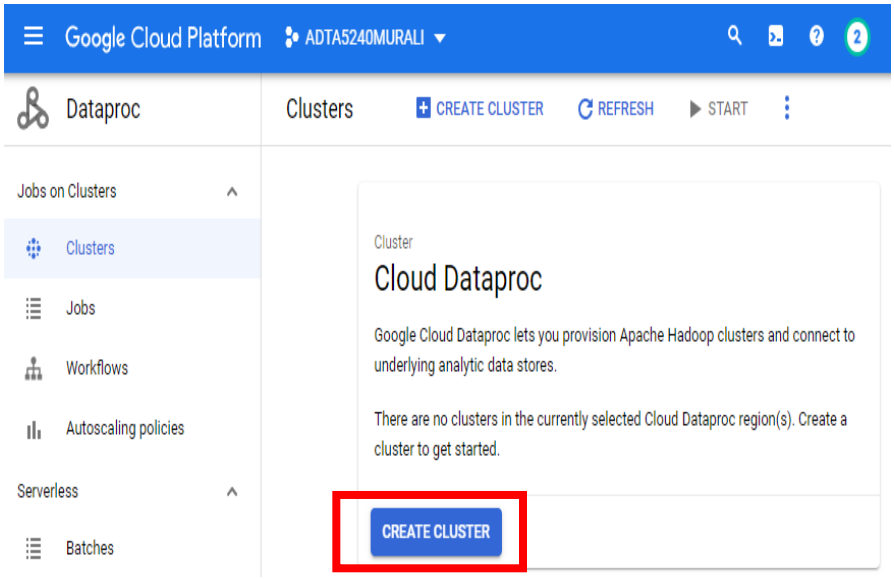




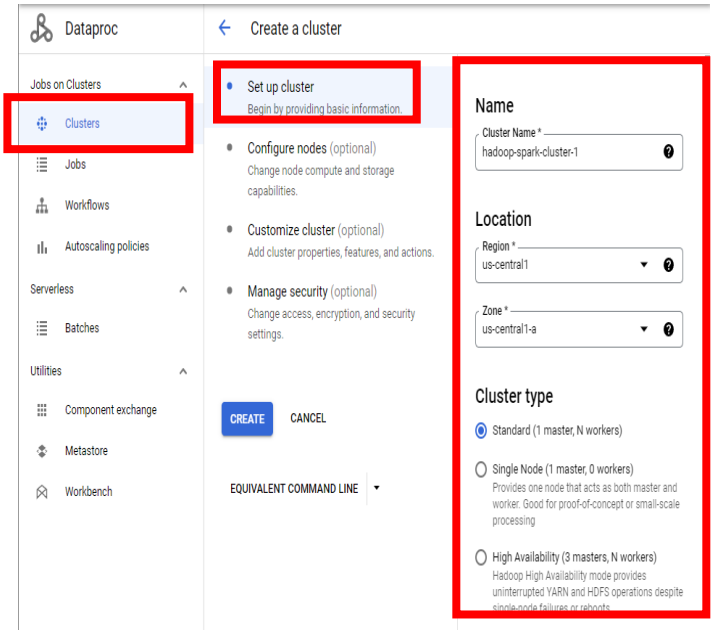
⇨ We will create clusters. To know more about clusters, click on the link below.
https://medium.com/@tudip/hadoop-ecosystem-in-google-cloud-platform-gcp-9d6eb70fc700

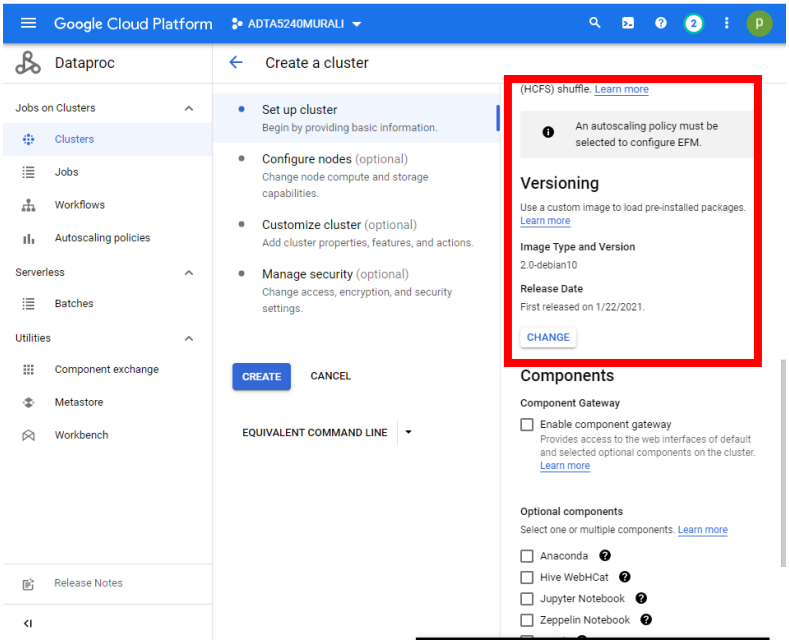
⇨ Click on the Navigation panel, scroll to the “Big Data” section, under “DataProc” option select “Clusters”.

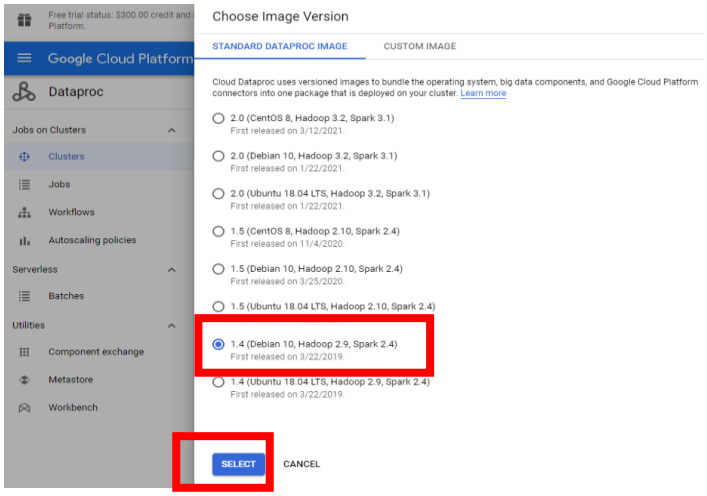




⇨ We are going to Set up cluster, configure nodes and Customize cluster.





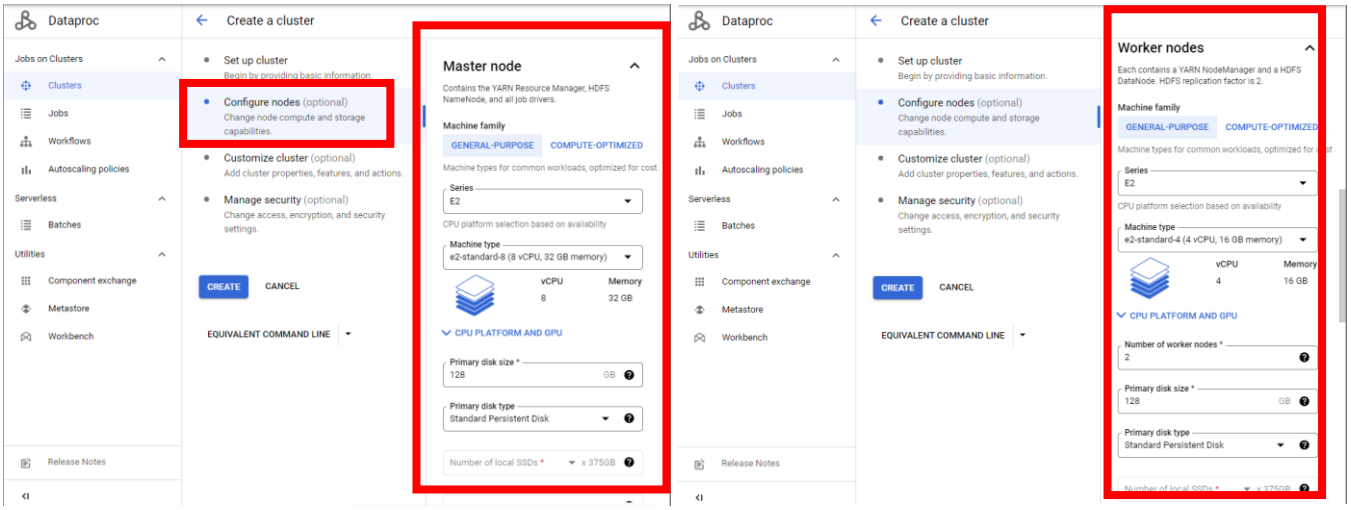


Cluster name: You can name it as you like but be sure to not use underscore.
Location: Region – us-central1, Zone – us-central1-a
Cluster type: Standard

DO NOT CLICK CREATE YET

Versioning – Click on Change and select 1.4 (Debian 10, Hadoop 2.9, Spark 2.4)

⇨ Configure nodes



Master node --- General purpose
Series: E2
Machine type: e2-standard-8(8 vCPUs, 32 GB memory)
Primary disk size: 128 GB
Primary disk size: Standard Persistent Disk

DO NOT CLICK CREATE YET

⇨ Scroll down to “worker nodes”

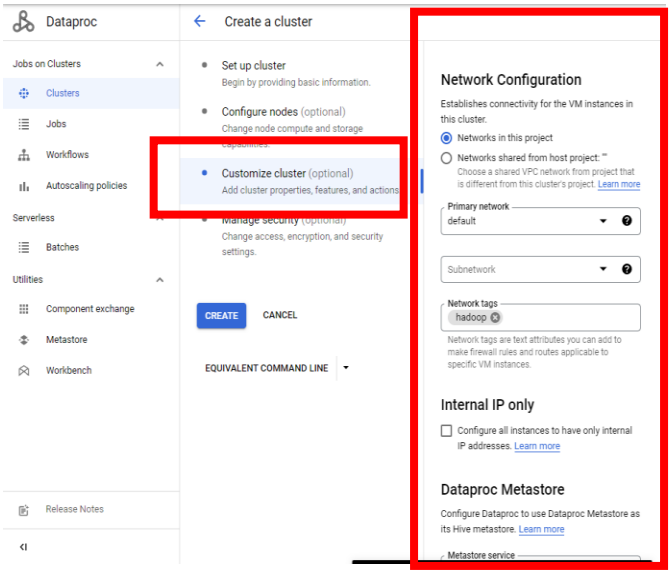
Worker nodes --- General purpose
Series: E2
Machine type: e2-standard-4(4 vCPUs, 16 GB memory)
Primary disk size: 128 GB
Primary disk size: Standard Persistent Disk

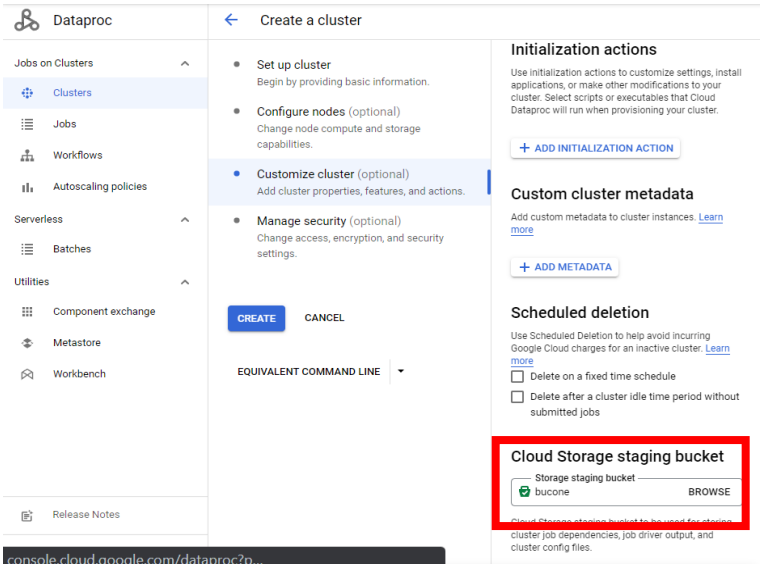
DO NOT CLICK CREATE YET

⇨ Customize cluster

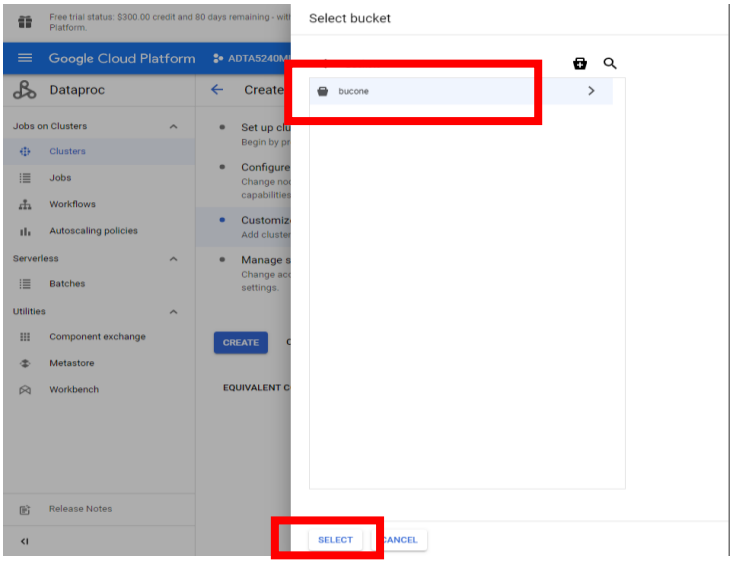
Network configuration: Primary network (default)
Network tags: Hadoop

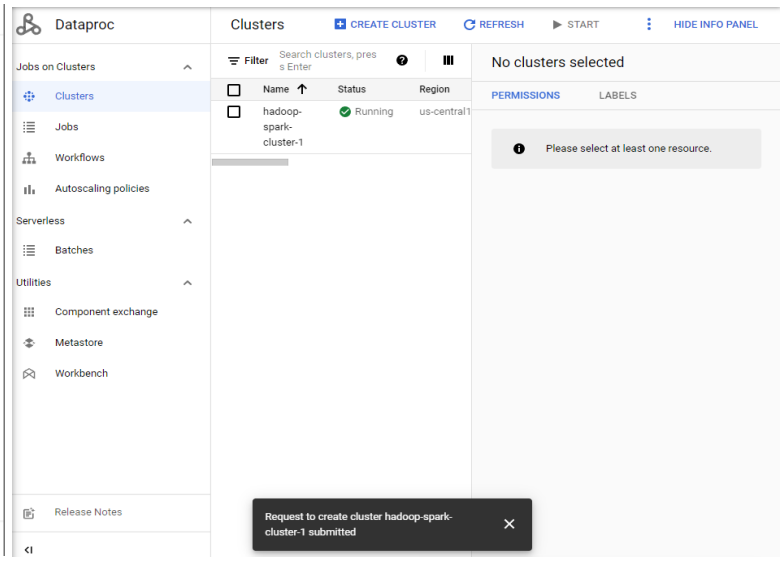
DO NOT CLICK CREATE YET



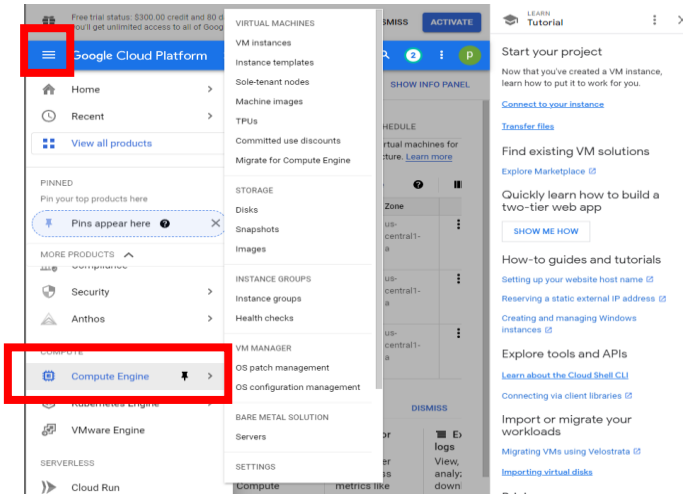


- ⇨ Load storage bucket.
- ⇨ Select BROWSE option from Cloud Storage Staging bucket.
- ⇨ Click the bucket where the data was already loaded.
- ⇨ Pick the bucket and then hit select.

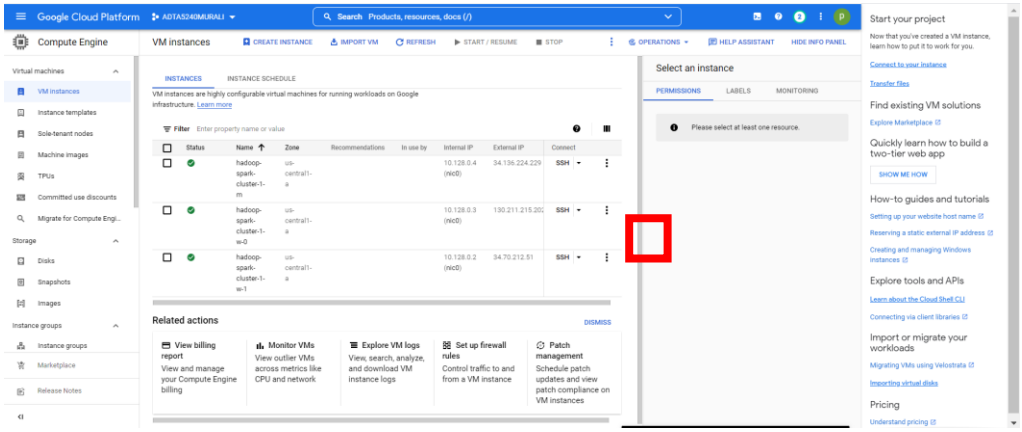




- ⇨ Click on Navigation pane (3 horizontal lines).
- ⇨ Under “Commute” section select “Compute Engine”.



⇨ This is what your console should look like if you followed the correct steps.



Before exiting the console make sure to stop all the clusters to avoid unnecessary charges. Click on the three-point option shown above and click STOP.