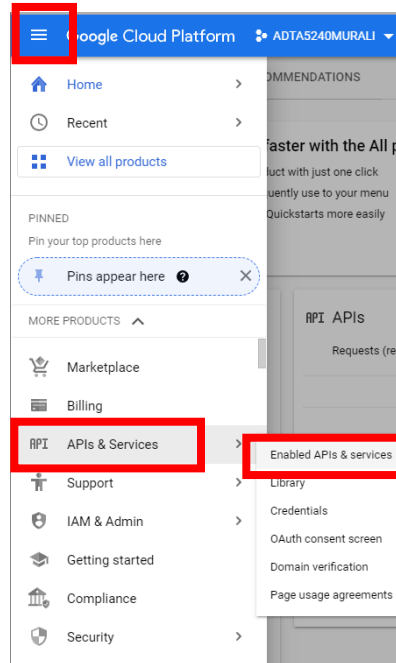


In this manual we will learn:

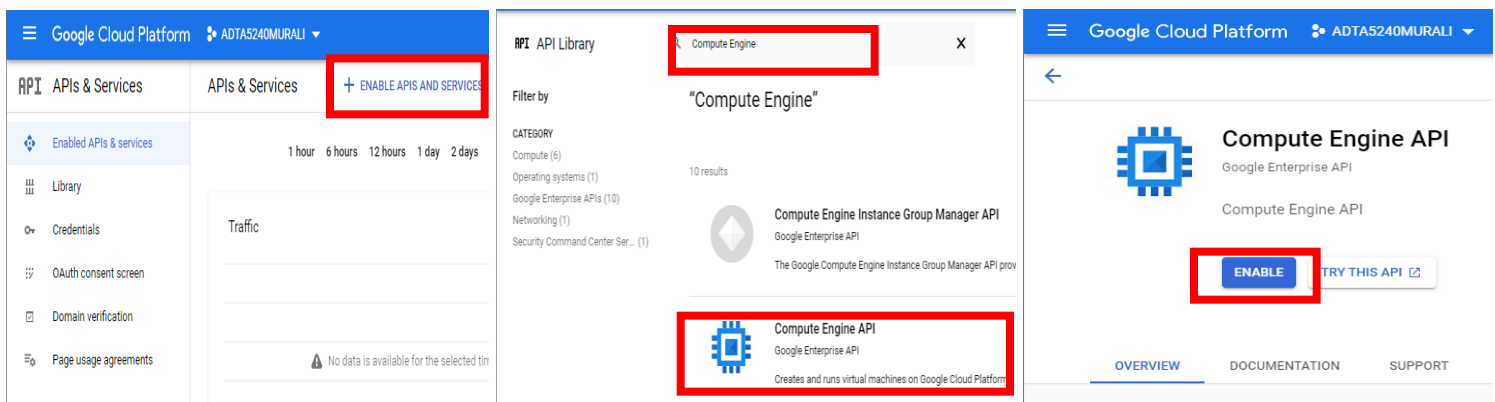
- To create a Hadoop and Spark cluster.

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

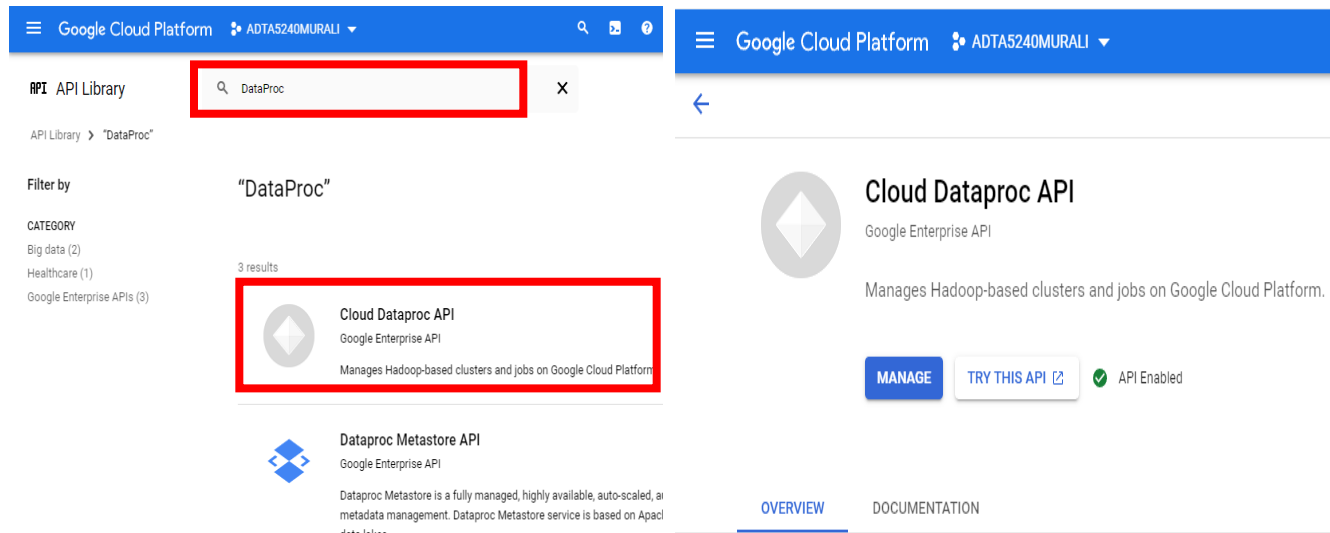
- ⇒ Log into your GCP account through your registered Gmail id.
- ⇒ Click on the Navigation panel (three horizontal lines on the top left corner).
- ⇒ Scroll to “APIs and Services” and select “Enabled API and Services”.



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



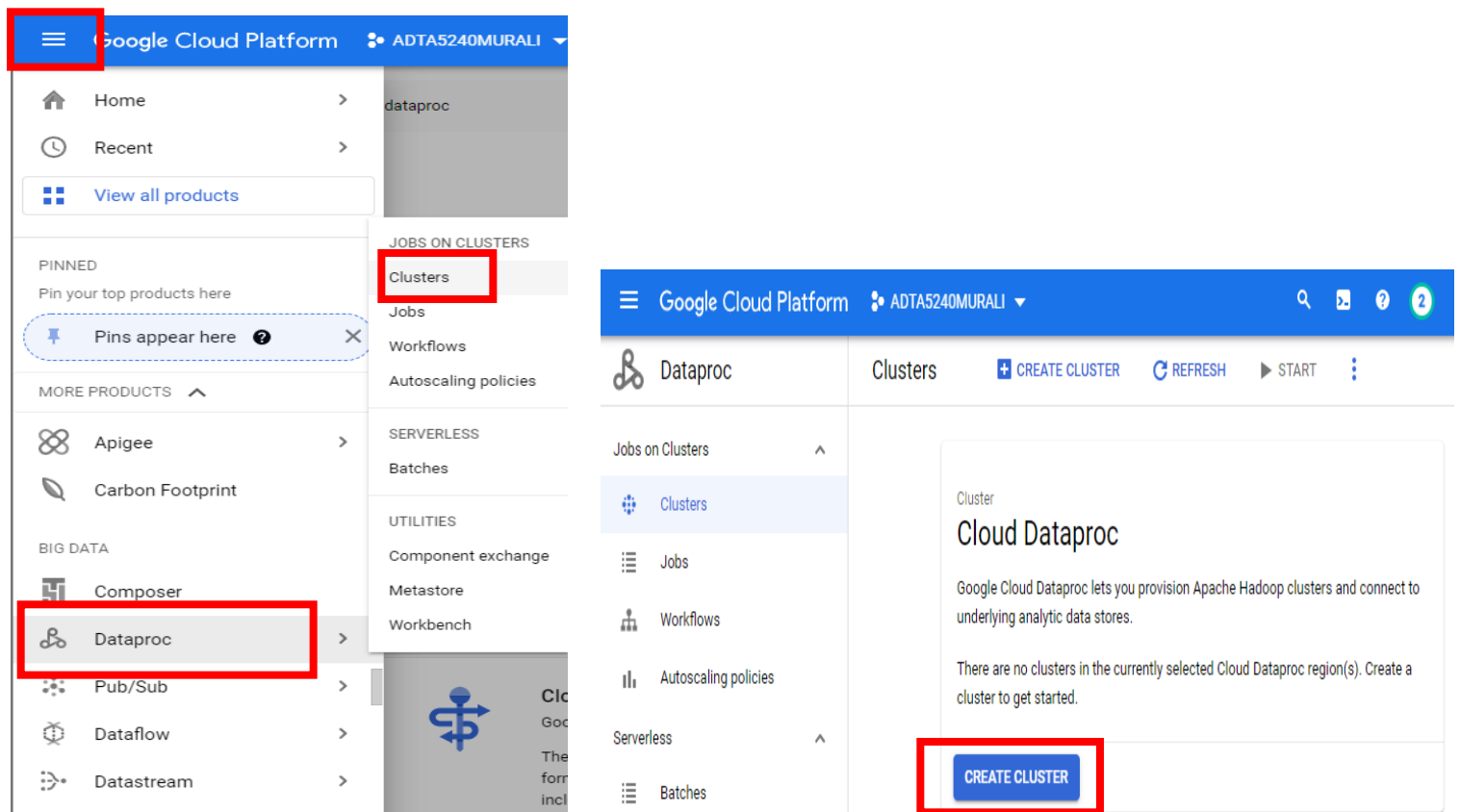
⇒ Next, we are going to download “Cloud DataProc API”. Search for “DataProc”. Select “Enable”. Since I already enabled it, it only gives me an option to manage it.



⇒ 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.

The image displays three screenshots of the Google Cloud Platform Dataproc 'Create a cluster' wizard.

First Screenshot: 'Set up cluster' step

- Name:** Cluster Name *
- Location:** Region * Zone *
- Cluster type:**
 - ☒ Standard (1 master, N workers)
 - ☐ Single Node (1 master, 0 workers)
 - ☐ High Availability (3 masters, N workers)

Second Screenshot: 'Versioning' step

- (HDFS) shuffle:** An autoscaling policy must be selected to configure EFM.
- Versioning:** Use a custom image to load pre-installed packages.
- Image Type and Version:** 2.0-debian10
- Release Date:** First released on 1/22/2021.
- Components:**
 - ☐ Enable component gateway
 - ☐ Anaconda
 - ☐ Hive WebHCat
 - ☐ Jupyter Notebook
 - ☐ Zeppelin Notebook

Third Screenshot: 'Choose Image Version' step

- STANDARD DATAPROC IMAGE**
- Choose Image Version:**
 - ☐ 2.0 (CentOS 8, Hadoop 3.2, Spark 3.1)
 - ☐ 2.0 (Debian 10, Hadoop 3.2, Spark 3.1)
 - ☐ 2.0 (Ubuntu 18.04 LTS, Hadoop 3.2, Spark 3.1)
 - ☐ 1.5 (CentOS 8, Hadoop 2.10, Spark 2.4)
 - ☐ 1.5 (Debian 10, Hadoop 2.10, Spark 2.4)
 - ☐ 1.5 (Ubuntu 18.04 LTS, Hadoop 2.10, Spark 2.4)
 - ☒ 1.4 (Debian 10, Hadoop 2.9, Spark 2.4)
 - ☐ 1.4 (Ubuntu 18.04 LTS, Hadoop 2.9, Spark 2.4)

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

The image shows two screenshots of the Google Cloud Dataproc 'Create a cluster' wizard. The left screenshot displays the 'Master node' configuration page. A red box highlights the 'Configure nodes (optional)' step in the left sidebar. The main content area shows the 'Machine family' section with 'GENERAL-PURPOSE' selected. The configuration includes: Series: E2, Machine type: e2-standard-8 (8 vCPU, 32 GB memory), vCPU: 8, Memory: 32 GB, Primary disk size: 128 GB, Primary disk type: Standard Persistent Disk, and Number of local SSDs: x 375GB. The right screenshot displays the 'Worker nodes' configuration page. A red box highlights the 'Machine family' section. The configuration includes: Series: E2, Machine type: e2-standard-4 (4 vCPU, 16 GB memory), vCPU: 4, Memory: 16 GB, Primary disk size: 128 GB, Primary disk type: Standard Persistent Disk, and Number of local SSDs: x 375GB.

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

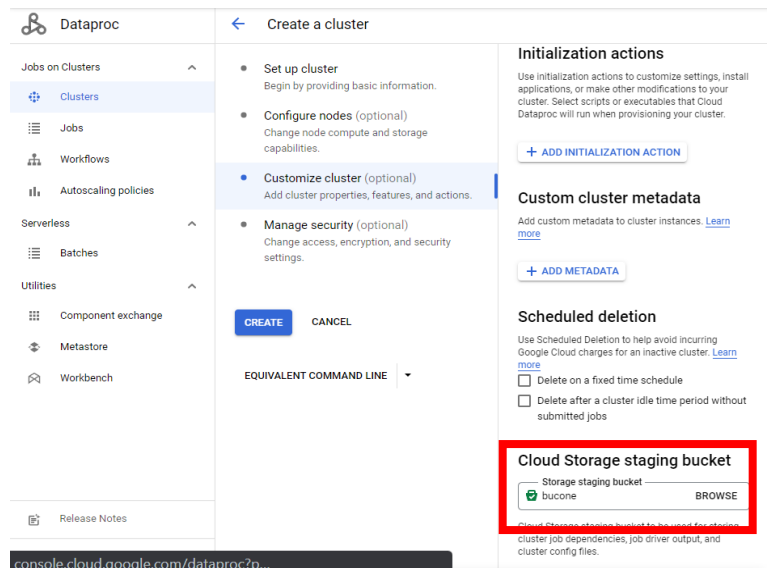
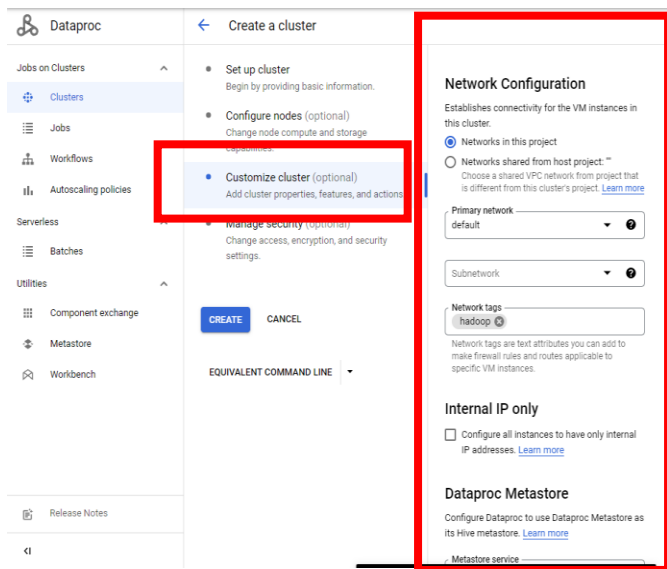
DO NOT CLICK CREATE YET

⇒ Customize cluster

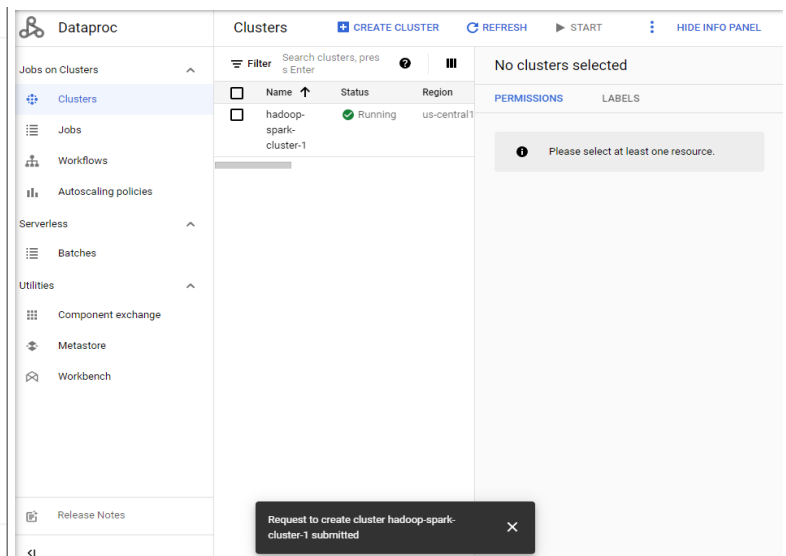
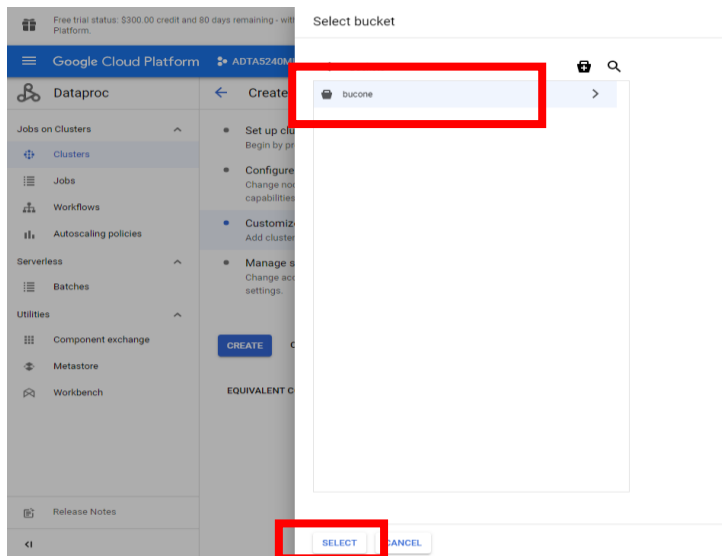
Network configuration: Primary network (default)

Network tags: Hadoop

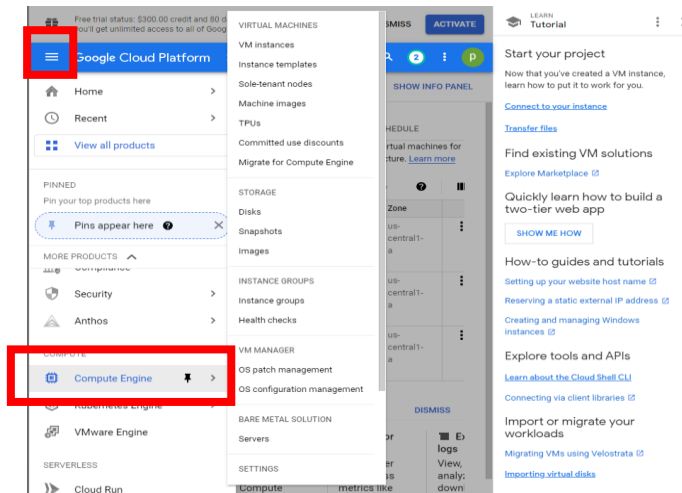
DO NOT CLICK CREATE YET



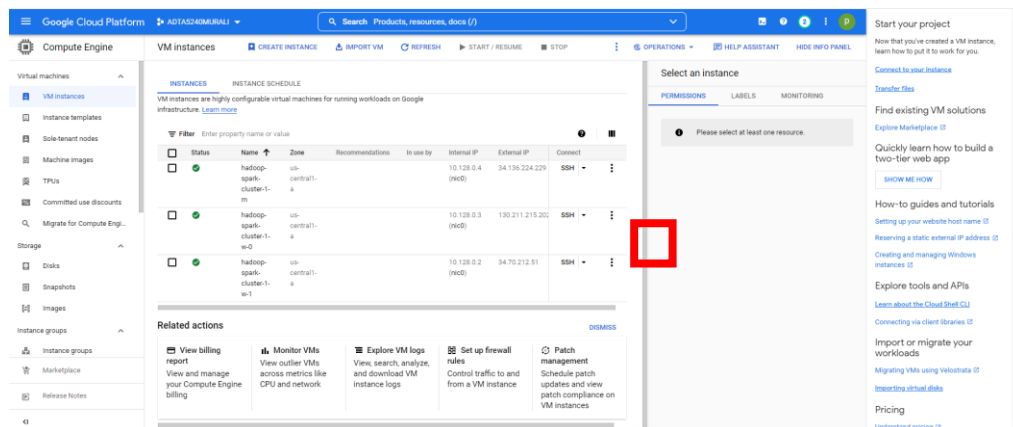
- ⇒ Load storage bucket.
- ⇒ Select BROWSE option from Cloud Storage Staging bucket.
- ⇒ Click the bucket where the data was already loaded.
- ⇒ Pick the bucket and then hit select.



- ⇒ 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.

