UNIVERSITY OF NORTH TEXAS, DENTON

Subject: INFO 5502
Subject Name: Principles and Techniques for Data Science
Instructor: Prof. Xiao Ting

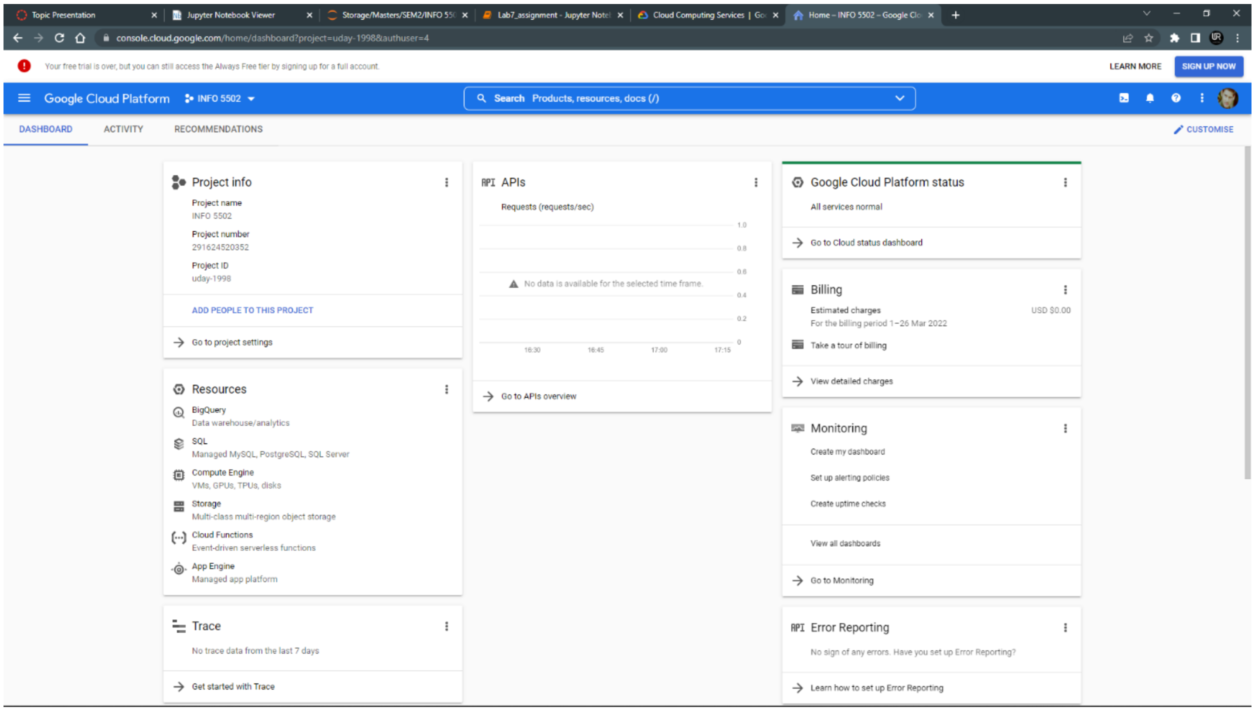
Assignment 7:
Spark for the Machine Learning

Team Name:
Uday Raj Suthapalli
Venkat Akshay Reddy Jaggavarapu
Vashisth Hareshkumar Sukhadiya

Create Storage Bucket in GCP

Step1: Create a New Storage Bucket.

- ➢ Open Google Cloud Platform.
- ➢ Go to GCP console.
- ➢ Click on the Navigation Pane on left side.
- ➢ Select Cloud Storage Open Browser tab.
- ➢ Click on the Create Bucket.



console.cloud.google.com/home/dashboard?project=uday-1998&authuser=4

Your free trial is over, but you can still access the Always Free tier by signing up for a full account.

Project name
INFO 5502

Project number
291624520352

Project ID
uday-1998

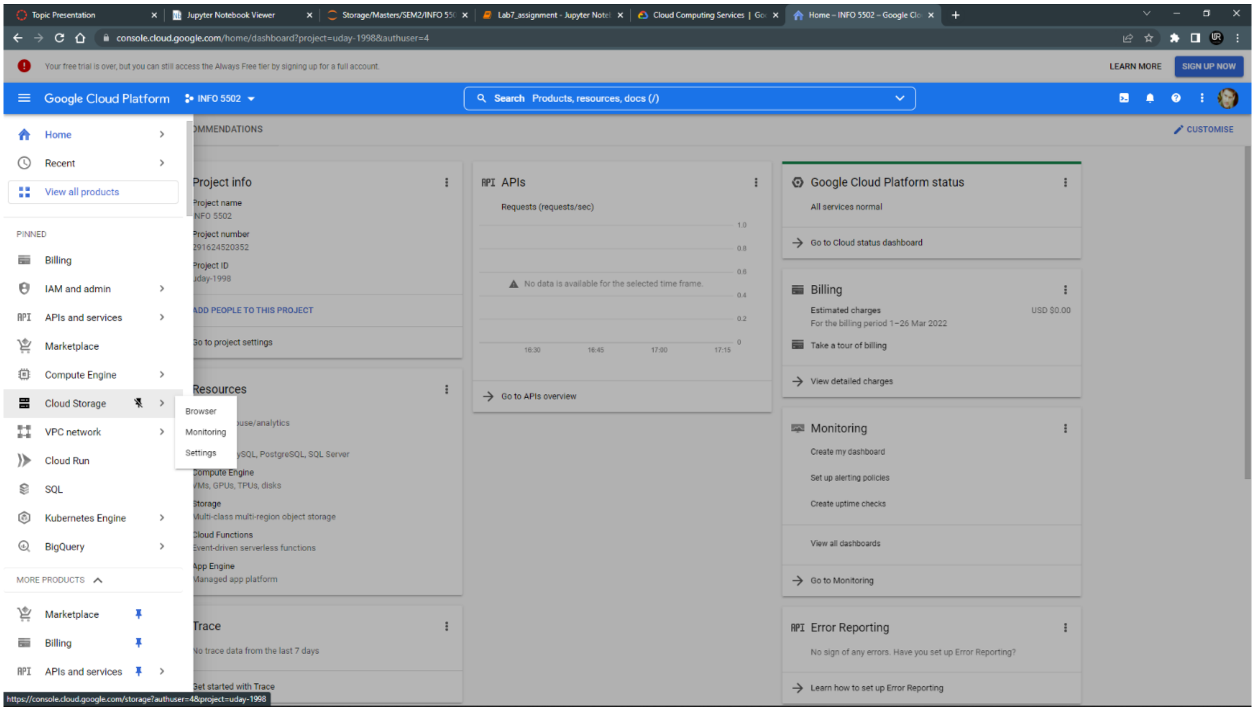
No data is available for the selected time frame.

All services normal

Estimated charges USD $0.00
For the billing period 1–26 Mar 2022

No trace data from the last 7 days

No sign of any errors. Have you set up Error Reporting?



Your free trial is over, but you can still access the Always Free tier by signing up for a full account.

Project info
Project name
INFO 5502
Project number
291624520352
Project ID
uday-1998

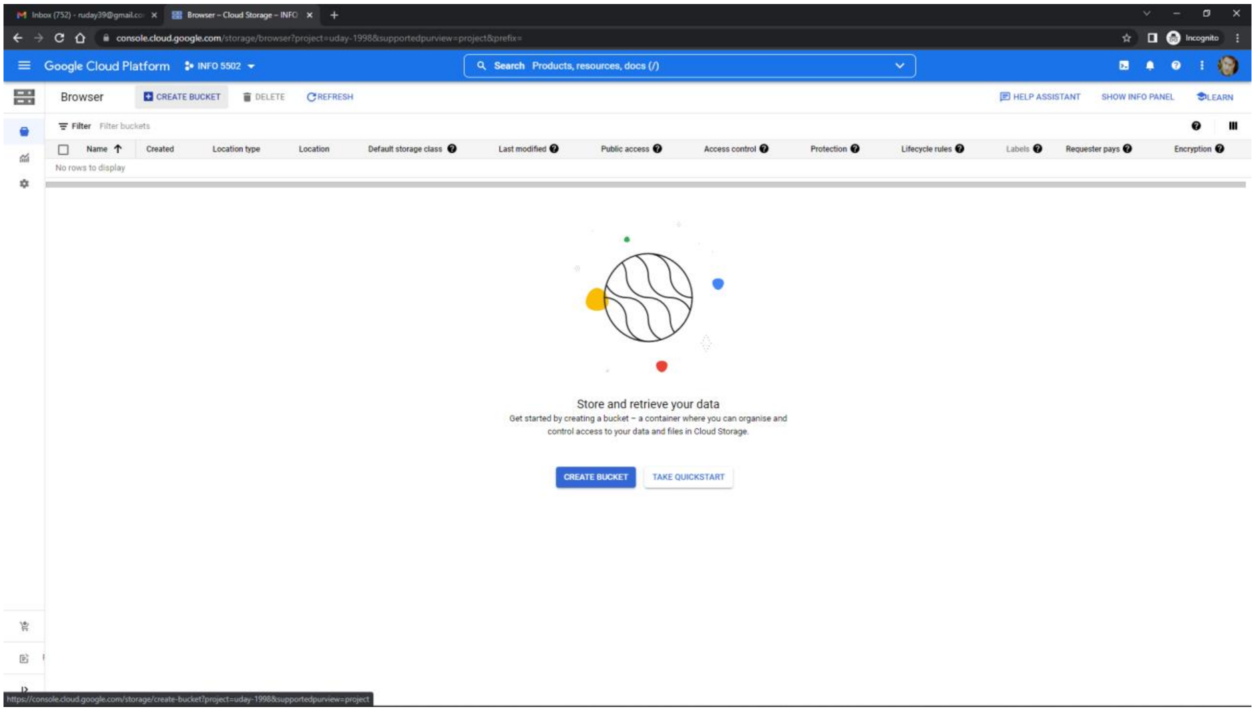
No data is available for the selected time frame.

Google Cloud Platform status
All services normal

Billing
Estimated charges USD $0.00
For the billing period 1 – 26 Mar 2022

No trace data from the last 7 days

No sign of any errors. Have you set up Error Reporting?



Store and retrieve your data
Get started by creating a bucket – a container where you can organise and control access to your data and files in Cloud Storage.

No rows to display

Step2: Create a GCP Bucket.

Step1:-

- Name your Bucket (Note: - Do not right 'Uppercase', 'no space' and not like that 'bah-vas-hkasj')

- Click on ‘Continue’ button.

Step2:-

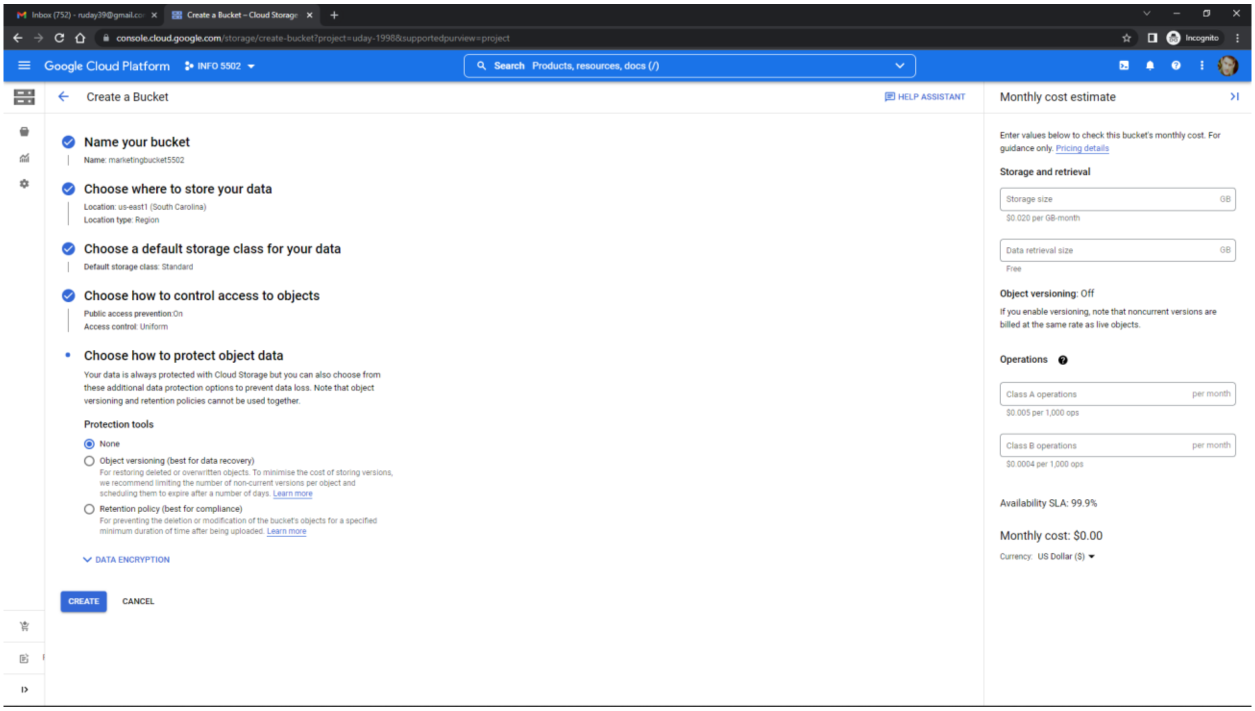
- Next step ‘Choose Where to Store Your Data’ as the defualt data (You can not cahange any information) Like, ‘Multi-Region’ and Location : US.

Step3:-

- ‘Choose your defualt storage class for your data’. Select Standard option.

Step4:-

- **‘**Choose how to control access to objects’. Select ‘Fine-grained’ means Set Object-level and Bucket-level: Permission can be granted at either level – bucket or folder.



Inbox (752) - ruday39@gmail.co
Create a Bucket – Cloud Storage
console.cloud.google.com/storage/create-bucket?project=uday-1998&supportedpurview=project
Incognito
Google Cloud Platform
INFO 5502
Search Products, resources, docs (/)
HELP ASSISTANT

Create a Bucket

Name your bucket
Name: marketingbucket5502

Choose where to store your data
Location: us-east1 (South Carolina)
Location type: Region

Choose a default storage class for your data
Default storage class: Standard

Choose how to control access to objects
Public access prevention:On
Access control: Uniform

Choose how to protect object data
Your data is always protected with Cloud Storage but you can also choose from these additional data protection options to prevent data loss. Note that object versioning and retention policies cannot be used together.

Protection tools

● None

○ Object versioning (best for data recovery)
For restoring deleted or overwritten objects. To minimise the cost of storing versions, we recommend limiting the number of non-current versions per object and scheduling them to expire after a number of days. Learn more

○ Retention policy (best for compliance)
For preventing the deletion or modification of the bucket's objects for a specified minimum duration of time after being uploaded. Learn more

Monthly cost estimate

Enter values below to check this bucket's monthly cost. For guidance only. Pricing details

Storage and retrieval

Storage size GB
$0.020 per GB-month

Data retrieval size GB
Free

Object versioning: Off
If you enable versioning, note that noncurrent versions are billed at the same rate as live objects.

Operations

Class A operations per month
$0.005 per 1,000 ops

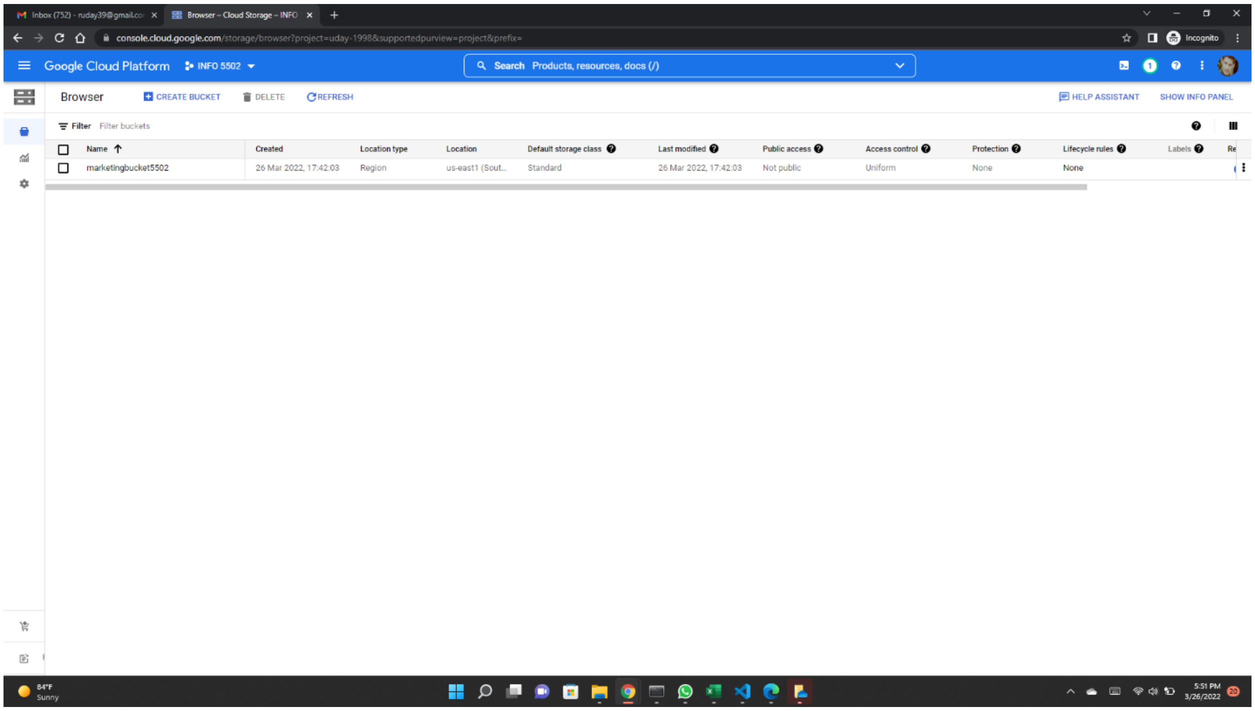
Class B operations per month
$0.0004 per 1,000 ops

Availability SLA: 99.9%

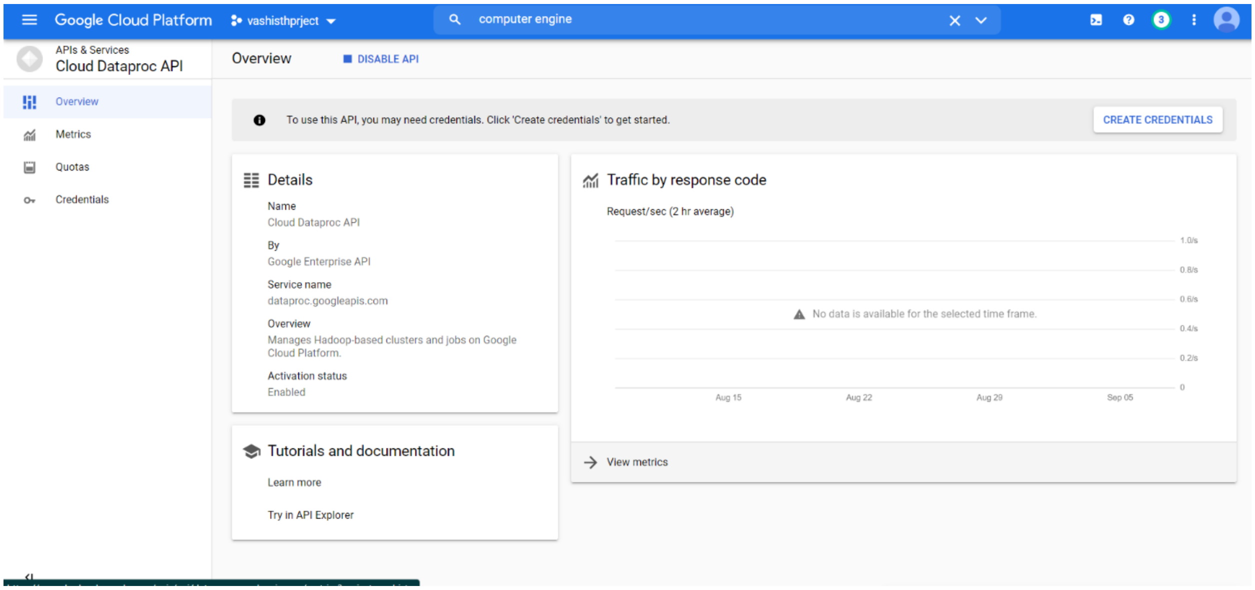
Monthly cost: $0.00
Currency: US Dollar ($)

Step5:-

- Advance Setting (Optional), Click on ‘Create’ Button.


Inbox (752) - ruday39@gmail.co
Browser – Cloud Storage – INFO
console.cloud.google.com/storage/browser?project=uday-1998&supportedpurview=project&prefix=
Incognito
Google Cloud Platform
INFO 5502
Search Products, resources, docs (/)
Browser
CREATE BUCKET
DELETE
REFRESH
HELP ASSISTANT
SHOW INFO PANEL
Filter Filter buckets
84°F
Sunny
5:51 PM
3/26/2022

Name	Created	Location type	Location	Default storage class	Last modified	Public access	Access control	Protection	Lifecycle rules	Labels	Re
marketingbucket5502	26 Mar 2022, 17:42:03	Region	us-east1 (Sout...	Standard	26 Mar 2022, 17:42:03	Not public	Uniform	None	None		



To use this API, you may need credentials. Click 'Create credentials' to get started.

Name
Cloud Dataproc API

By
Google Enterprise API

Service name
dataproc.googleapis.com

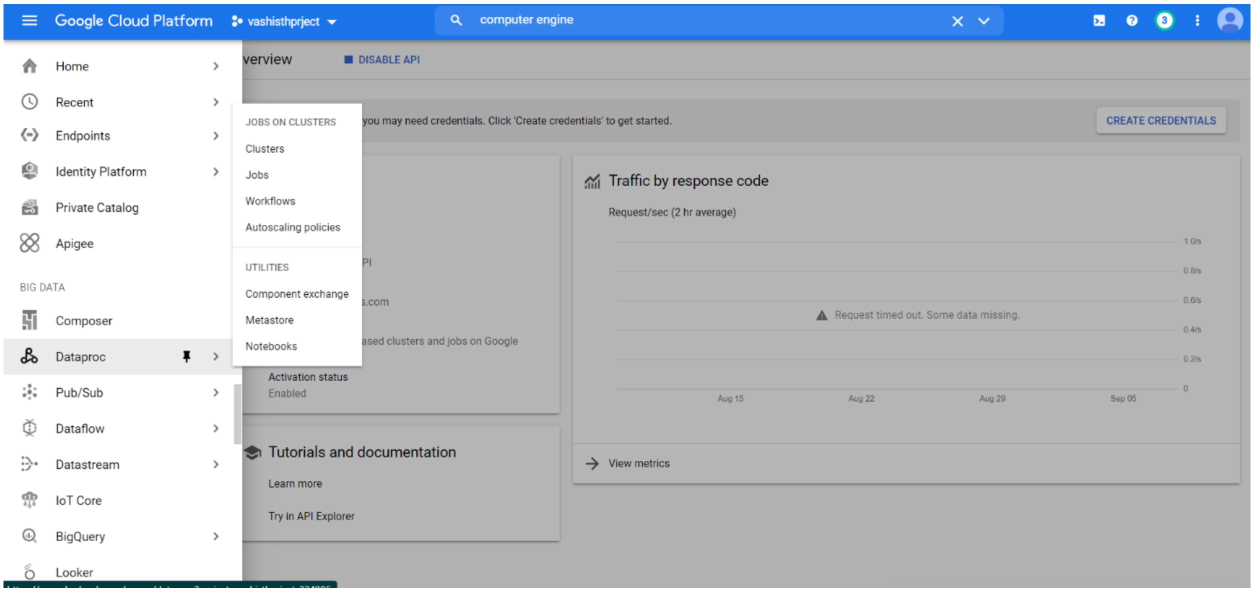
Overview
Manages Hadoop-based clusters and jobs on Google Cloud Platform.

Activation status
Enabled

Step3: - Create a GCP DataProc Cluster.

➢ Navigate to DataProc

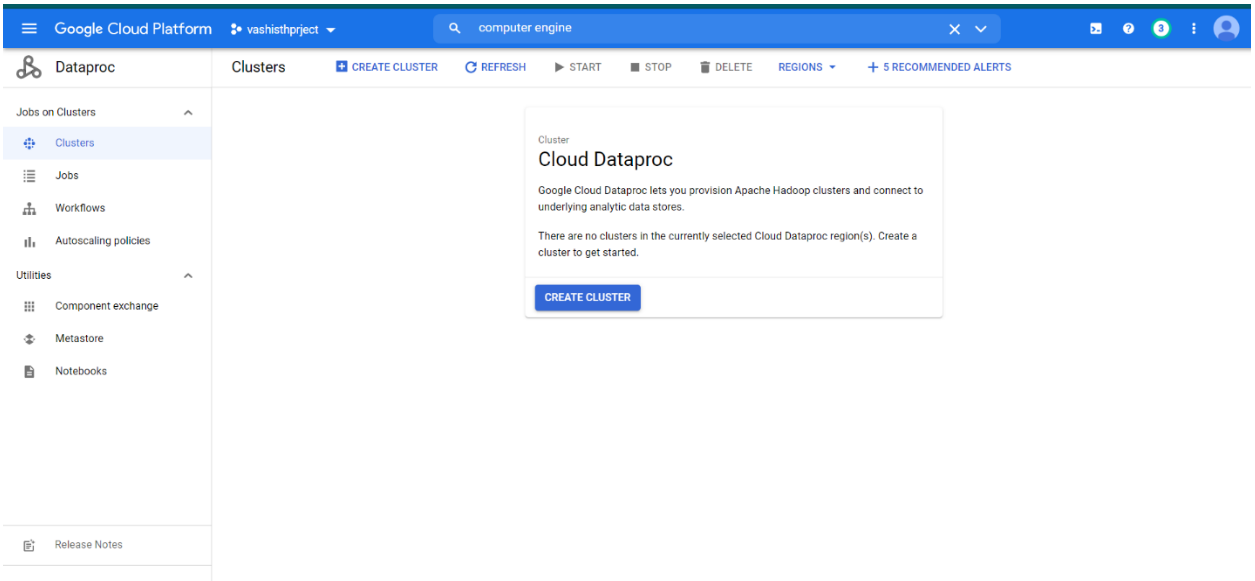
- Click the navigation menu to open the menu.
- Scroll over and go to the Bigdata, select the "DataProc".
- Click on "Cluster"



you may need credentials. Click 'Create credentials' to get started.

Request timed out. Some data missing.

➢ Click "Create Cluster" button.



Cloud Dataproc

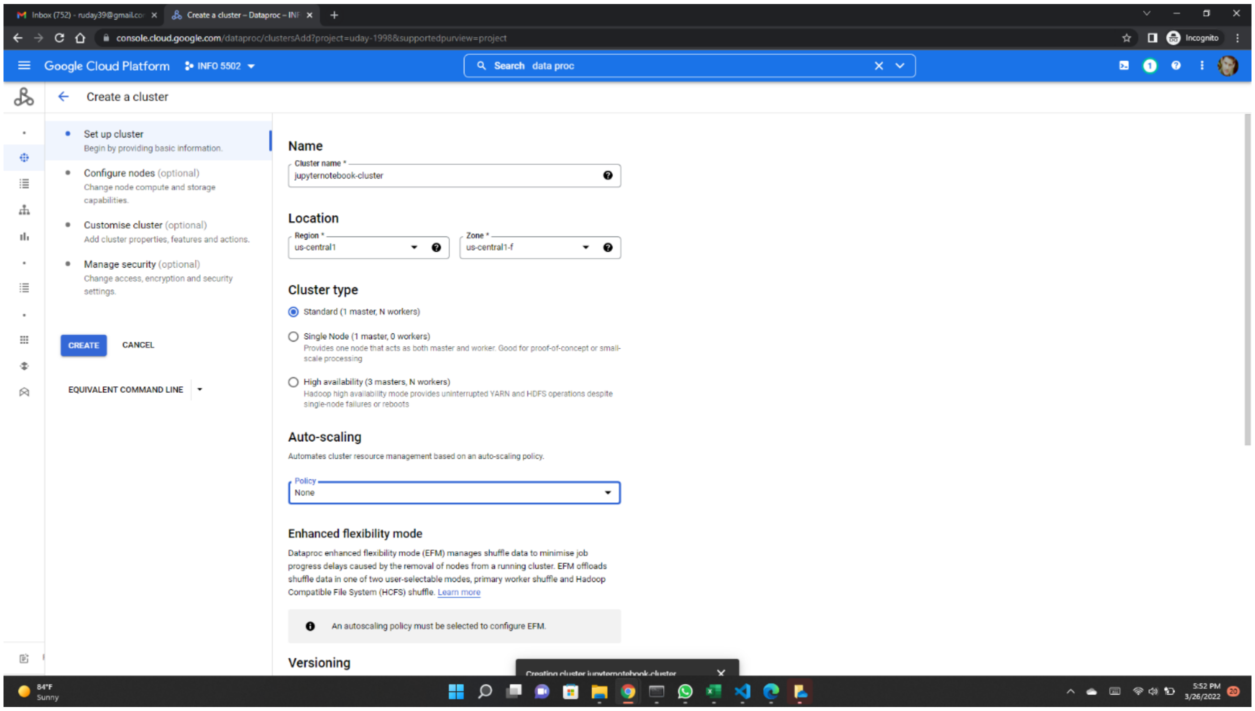
Google Cloud Dataproc lets you provision Apache Hadoop clusters and connect to underlying analytic data stores.

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

➢ Set up Cluster

- Enter "Name" as "jupyternotebook-cluster".
- Enter "Location" : Region as "us-central1" and Zone as "us-central1-f"
- Enter "Cluster Type" as "Standard (1 master, N workers)".

- Optional component select Jupyter notebook.
- Change system version.
  - Scroll down to “Versioning”
  - Image type and version 2.0-debian10
  - Components: - enable component gateway
  - Optional component we can select Jupyter Notebook
  - Here is the screen shot versioning change.


Inbox (752) - ruday39@gmail.cor
Create a cluster – Dataproc – INF
console.cloud.google.com/dataproc/clustersAdd?project=uday-1998&supportedpurview=project
Incognito
Google Cloud Platform
INFO 5502
Search data proc
84°F
Sunny
5:52 PM
3/26/2022

Create a cluster

- Set up cluster
  Begin by providing basic information.
- Configure nodes (optional)
  Change node compute and storage capabilities.
- Customise cluster (optional)
  Add cluster properties, features and actions.
- Manage security (optional)
  Change access, encryption and security settings.

CREATE CANCEL

EQUIVALENT COMMAND LINE

Name

Cluster name *: jupyternotebook-cluster

Location

Region *: us-central1
Zone *: us-central1-f

Cluster type

● Standard (1 master, N workers)

○ Single Node (1 master, 0 workers)
Provides one node that acts as both master and worker. Good for proof-of-concept or small-scale processing

○ High availability (3 masters, N workers)
Hadoop high availability mode provides uninterrupted YARN and HDFS operations despite single-node failures or reboots

Auto-scaling

Automates cluster resource management based on an auto-scaling policy.

Policy: None

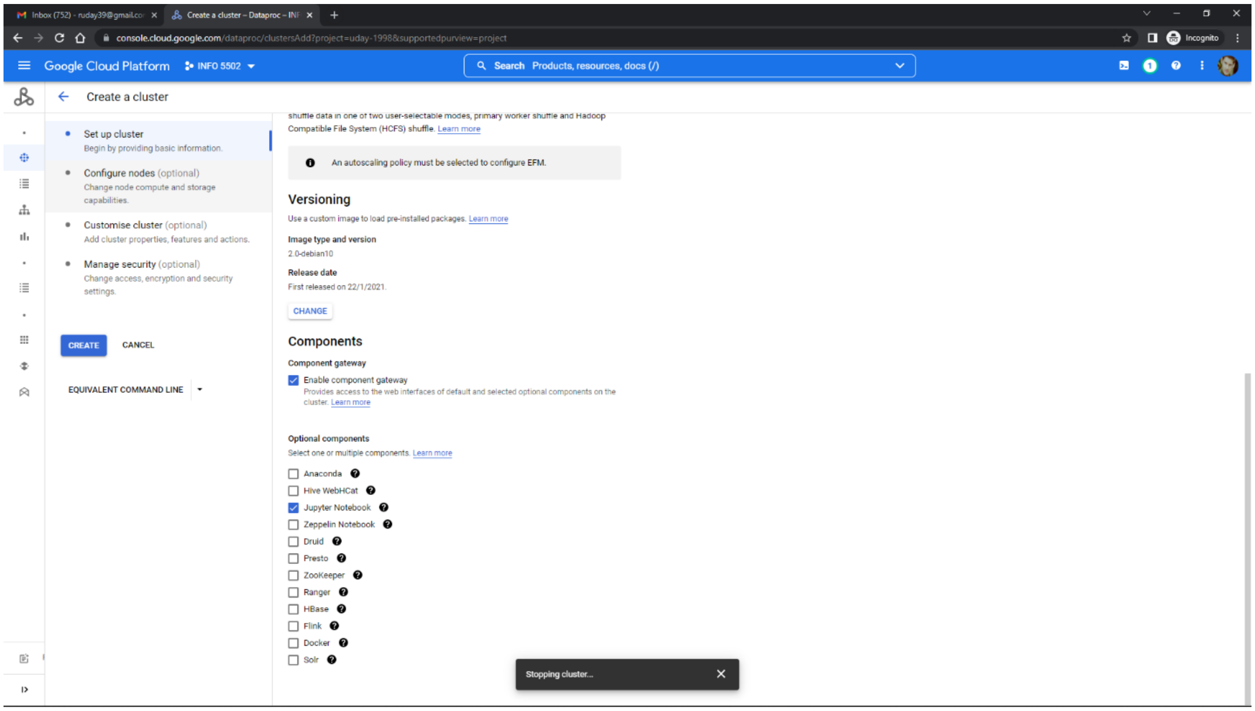
Enhanced flexibility mode

Dataproc enhanced flexibility mode (EFM) manages shuffle data to minimise job progress delays caused by the removal of nodes from a running cluster. EFM offloads shuffle data in one of two user-selectable modes, primary worker shuffle and Hadoop Compatible File System (HCFS) shuffle. Learn more

An autoscaling policy must be selected to configure EFM.

Versioning

Creating cluster jupyternotebook-cluster


Inbox (752) - ruday39@gmail.cor
Create a cluster – Dataproc – INF
console.cloud.google.com/dataproc/clustersAdd?project=uday-1998&supportedpurview=project
Google Cloud Platform
INFO 5502
Search Products, resources, docs (/)
Incognito
Create a cluster

- Set up cluster
  Begin by providing basic information.
- Configure nodes (optional)
  Change node compute and storage capabilities.
- Customise cluster (optional)
  Add cluster properties, features and actions.
- Manage security (optional)
  Change access, encryption and security settings.

CREATE
CANCEL
EQUIVALENT COMMAND LINE

shuffle data in one of two user-selectable modes, primary worker shuffle and Hadoop Compatible File System (HCFS) shuffle. Learn more

An autoscaling policy must be selected to configure EFM.

Versioning

Use a custom image to load pre-installed packages. Learn more

Image type and version
2.0-debian10

Release date
First released on 22/1/2021.

Components

Component gateway

☑ Enable component gateway
Provides access to the web interfaces of default and selected optional components on the cluster. Learn more

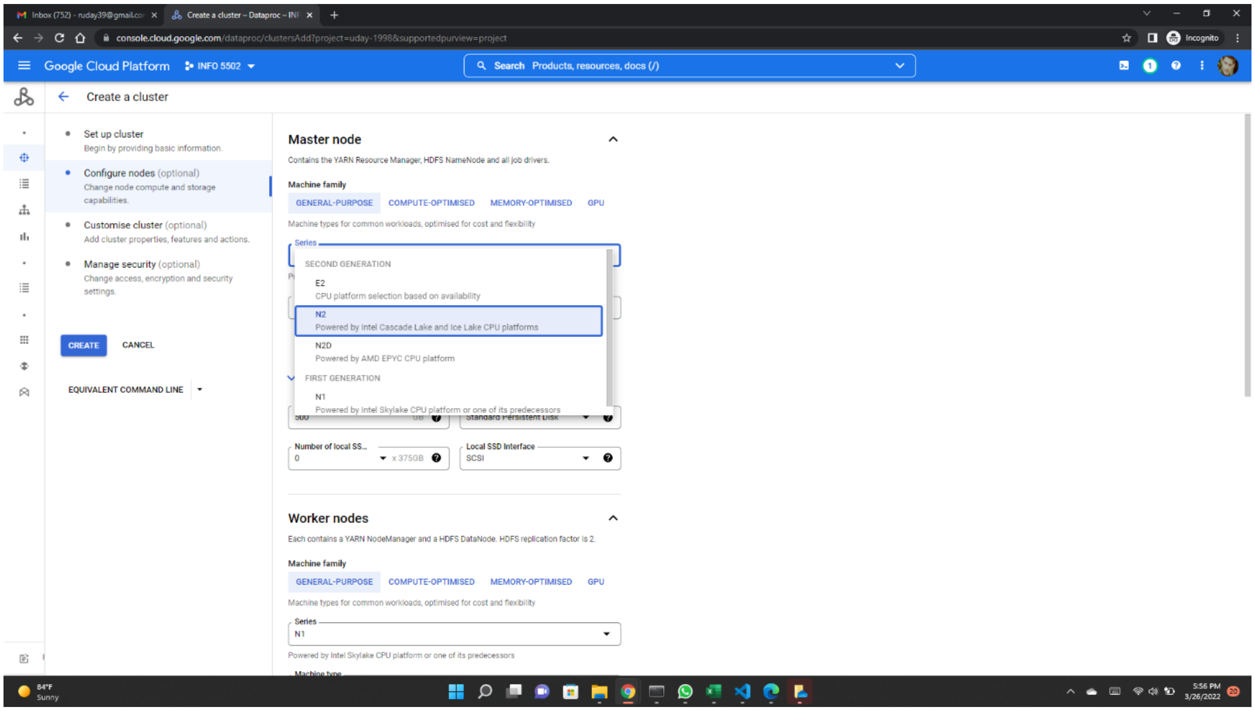
Optional components

Select one or multiple components. Learn more

☐ Anaconda
☐ Hive WebHCat
☑ Jupyter Notebook
☐ Zeppelin Notebook
☐ Druid
☐ Presto
☐ ZooKeeper
☐ Ranger
☐ HBase
☐ Flink
☐ Docker
☐ Solr

Stopping cluster...

➢ Configure master and work nodes.
- Click “Configure nodes”
- Select “GENERAL-PORPUSE” for the “Machine Family”
- Select “N2” for “Series”, Select “n2-standard-2 (2 vCPU, 8GB memory)” for the Machine Type”
- Select “500 GB” for the “Primary disk size (min 10GB)”
- Select “Standard Persistent Disk” for the “Primary disk type”
- Scroll down to “Worker nodes”.
- Select “GENERAL-PORPUSE” for the “Machine Family”
- Keep ”2” for the “Number of Worker Node”
- Select “N1” for “Series”, Select “n2-standard-2 (2 vCPU, 8GB memory)” for the Machine Type”
- Select “500 GB” for the “Primary disk size (min 10GB)”
- Select “Standard Persistent Disk” for the “Primary disk type”



console.cloud.google.com/dataproc/clustersAdd?project=uday-1998&supportedpurview=project

Google Cloud Platform · INFO 5502

Create a cluster

- Set up cluster — Begin by providing basic information.
- Configure nodes (optional) — Change node compute and storage capabilities.
- Customise cluster (optional) — Add cluster properties, features and actions.
- Manage security (optional) — Change access, encryption and security settings.

CREATE CANCEL

EQUIVALENT COMMAND LINE

Master node

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

Machine family: GENERAL-PURPOSE | COMPUTE-OPTIMISED | MEMORY-OPTIMISED | GPU

Machine types for common workloads, optimised for cost and flexibility

Series:

SECOND GENERATION
- E2 — CPU platform selection based on availability
- N2 — Powered by Intel Cascade Lake and Ice Lake CPU platforms
- N2D — Powered by AMD EPYC CPU platform

FIRST GENERATION
- N1 — Powered by Intel Skylake CPU platform or one of its predecessors

Number of local SS...: 0 x 375GB
Local SSD Interface: SCSI

Worker nodes

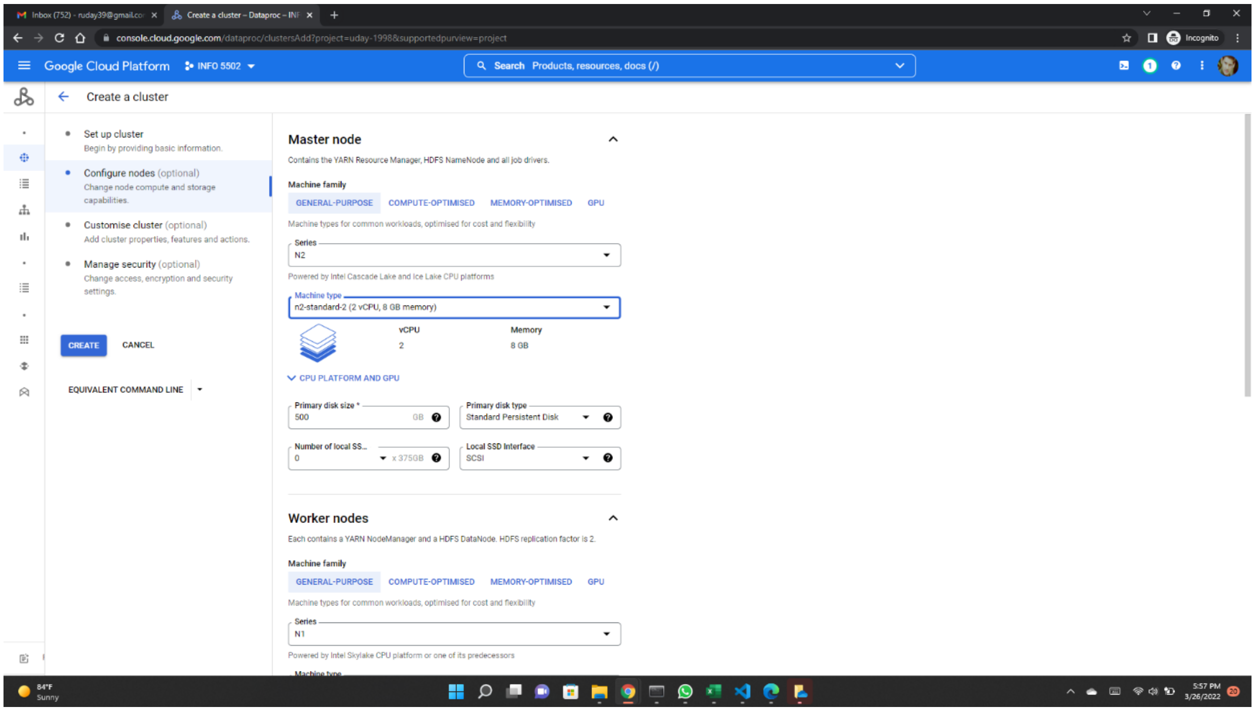
Each contains a YARN NodeManager and a HDFS DataNode. HDFS replication factor is 2.

Machine family: GENERAL-PURPOSE | COMPUTE-OPTIMISED | MEMORY-OPTIMISED | GPU

Machine types for common workloads, optimised for cost and flexibility

Series: N1

Powered by Intel Skylake CPU platform or one of its predecessors



console.cloud.google.com/dataproc/clustersAdd?project=uday-1998&supportedpurview=project

Google Cloud Platform · INFO 5502

Create a cluster

- Set up cluster — Begin by providing basic information.
- Configure nodes (optional) — Change node compute and storage capabilities.
- Customise cluster (optional) — Add cluster properties, features and actions.
- Manage security (optional) — Change access, encryption and security settings.

CREATE CANCEL

EQUIVALENT COMMAND LINE

Master node

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

Machine family: GENERAL-PURPOSE | COMPUTE-OPTIMISED | MEMORY-OPTIMISED | GPU

Machine types for common workloads, optimised for cost and flexibility

Series: N2

Powered by Intel Cascade Lake and Ice Lake CPU platforms

Machine type: n2-standard-2 (2 vCPU, 8 GB memory)

vCPU: 2 · Memory: 8 GB

CPU PLATFORM AND GPU

Primary disk size *: 500 GB
Primary disk type: Standard Persistent Disk
Number of local SS...: 0 x 375GB
Local SSD Interface: SCSI

Worker nodes

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

Machine family: GENERAL-PURPOSE | COMPUTE-OPTIMISED | MEMORY-OPTIMISED | GPU

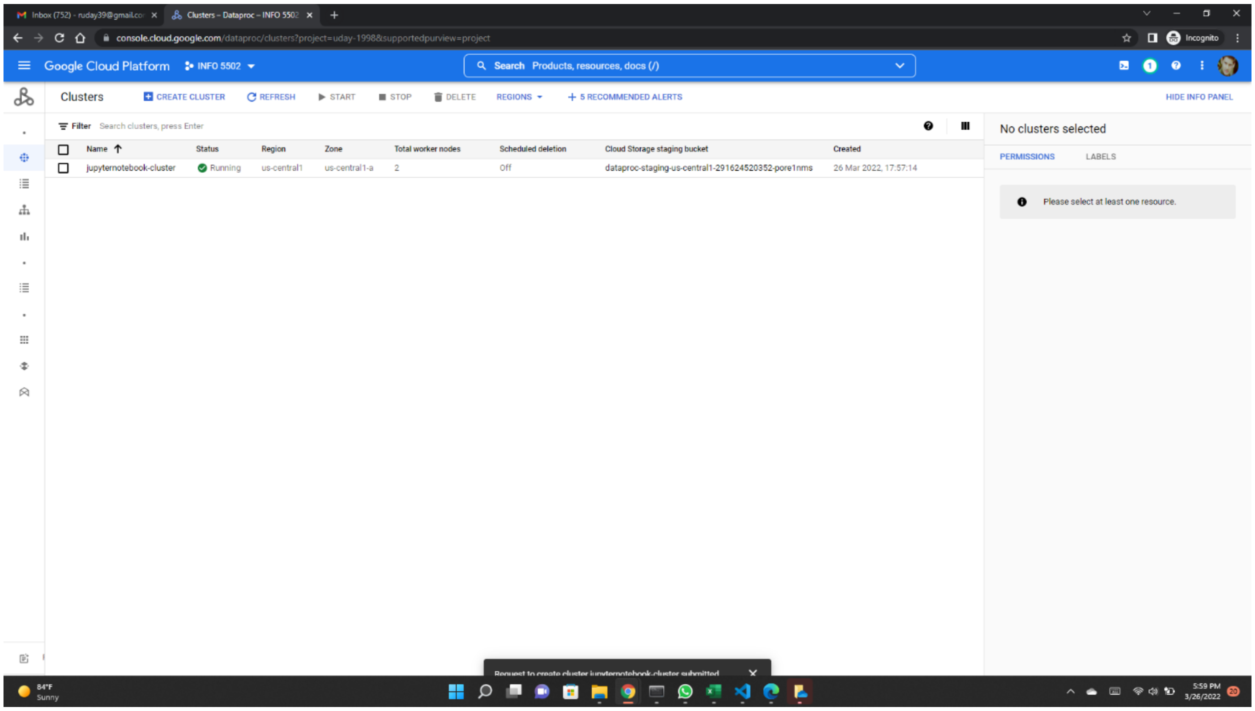
Machine types for common workloads, optimised for cost and flexibility

Series: N1

Powered by Intel Skylake CPU platform or one of its predecessors

- ➢ In Cloud Storage Staging Bucket, select browse and click on storage bucket Which is name as marketingbucket5502 .
- ➢ Click on select

- ➢ Now, Click on Create



Inbox (752) - ruday39@gmail.co
Clusters – Dataproc – INFO 5502
console.cloud.google.com/dataproc/clusters?project=uday-1998&supportedpurview=project
Incognito
Google Cloud Platform
INFO 5502
Search Products, resources, docs (/)
Clusters
CREATE CLUSTER
REFRESH
START
STOP
DELETE
REGIONS
5 RECOMMENDED ALERTS
HIDE INFO PANEL
Filter Search clusters, press Enter
PERMISSIONS
LABELS
84°F
Sunny
5:59 PM
3/26/2022

□	Name ↑	Status	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created
□	jupyternotebook-cluster	✓ Running	us-central1	us-central1-a	2	Off	dataproc-staging-us-central1-291624520352-pore1nms	26 Mar 2022, 17:57:14

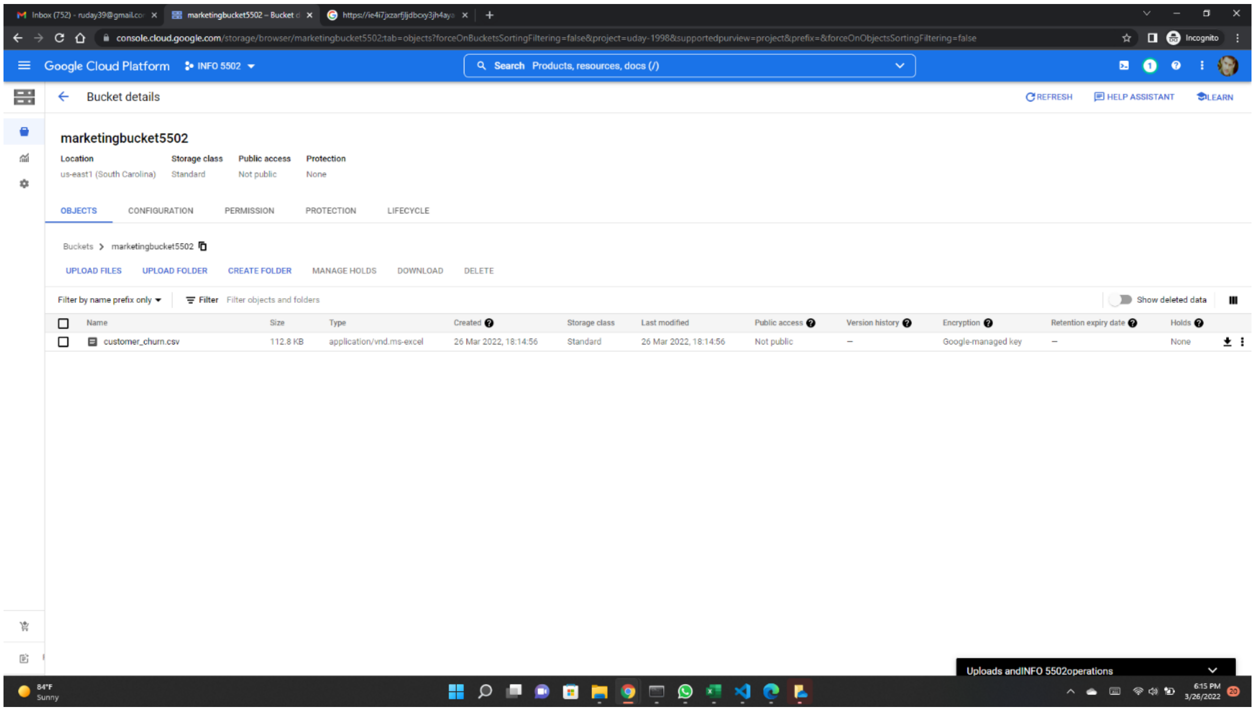
No clusters selected

Please select at least one resource.

Request to create cluster jupyternotebook-cluster submitted

Step4: - Upload Data to GCP Storage Buckets.

- ➢ Open the data folder by clicking on the folder name ‘data’.
- ➢ Click on “Upload files”.
- ➢ Browse for the file to be uploaded, highlight the files, and click “open”
- ➢ Upload files



console.cloud.google.com/storage/browser/marketingbucket5502;tab=objects?forceOnBucketsSortingFiltering=false&project=uday-1998&supportedpurview=project&prefix=&forceOnObjectsSortingFiltering=false

Google Cloud Platform · INFO 5502

Bucket details

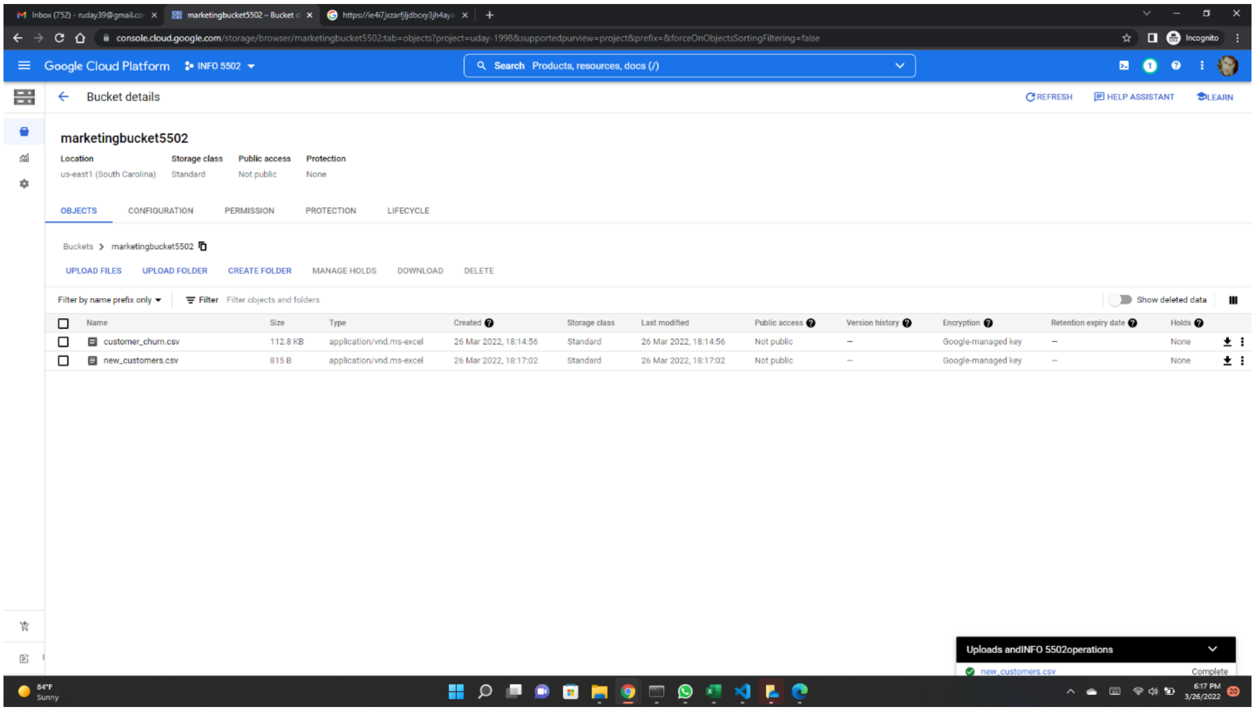
marketingbucket5502

Location	Storage class	Public access	Protection
us-east1 (South Carolina)	Standard	Not public	None

Buckets > marketingbucket5502

Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption	Retention expiry date	Holds
customer_churn.csv	112.8 KB	application/vnd.ms-excel	26 Mar 2022, 18:14:56	Standard	26 Mar 2022, 18:14:56	Not public	–	Google-managed key	–	None

Uploads and INFO 5502operations



console.cloud.google.com/storage/browser/marketingbucket5502;tab=objects?project=uday-1998&supportedpurview=project&prefix=&forceOnObjectsSortingFiltering=false

Google Cloud Platform · INFO 5502

Bucket details

marketingbucket5502

Location	Storage class	Public access	Protection
us-east1 (South Carolina)	Standard	Not public	None

Buckets > marketingbucket5502

Name	Size	Type	Created	Storage class	Last modified	Public access	Version history	Encryption	Retention expiry date	Holds
customer_churn.csv	112.8 KB	application/vnd.ms-excel	26 Mar 2022, 18:14:56	Standard	26 Mar 2022, 18:14:56	Not public	–	Google-managed key	–	None
new_customers.csv	815 B	application/vnd.ms-excel	26 Mar 2022, 18:17:02	Standard	26 Mar 2022, 18:17:02	Not public	–	Google-managed key	–	None

Uploads and INFO 5502operations

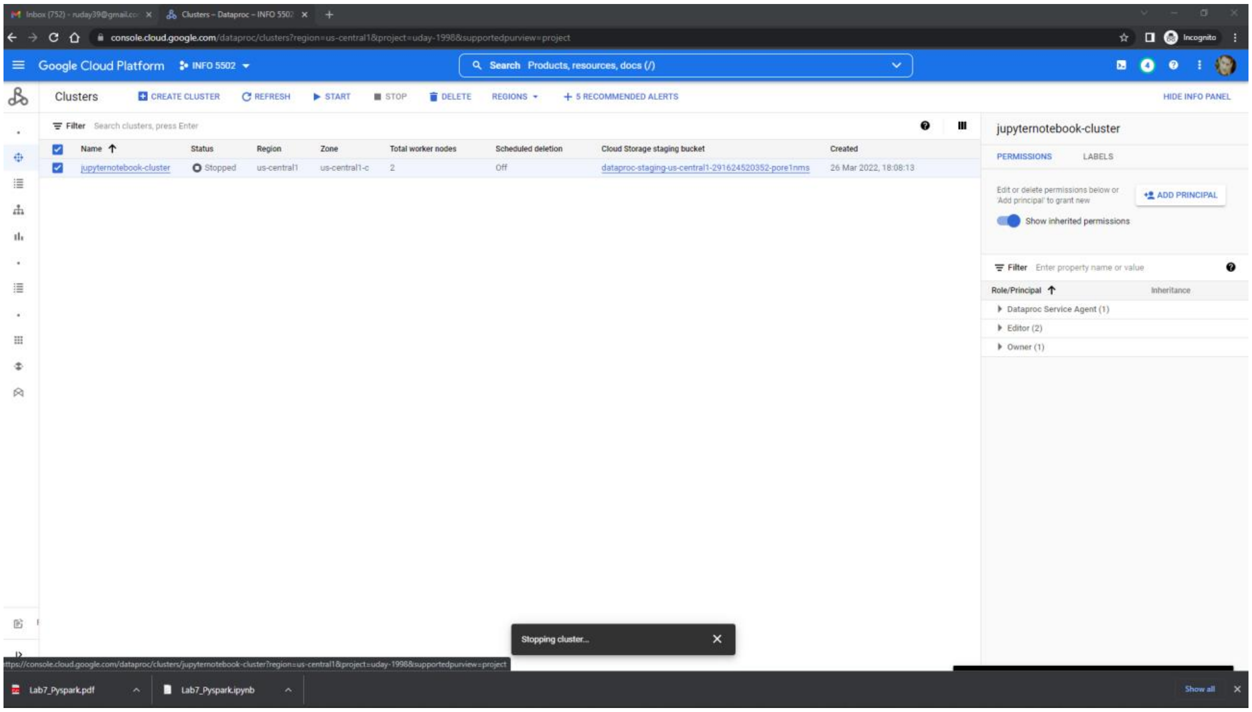
new_customers.csv Complete

- We have opened clusters.
- We have opened Jupyter Notebook from data proc cluster.
- Also, created a logistic ML model using pyspark.

How to Stop and delete Cluster Nodes and bucket in GCP

Step1: - Stop a Cluster Node in GCP

- Now, select three vertical dots near Connect and the click on the Stop.
- Then select Stop.
- As well as Delete cluster.


Inbox (752) - ruday39@gmail.co
Clusters – Dataproc – INFO 5502
console.cloud.google.com/dataproc/clusters?region=us-central1&project=uday-1998&supportedpurview=project
Incognito
Google Cloud Platform
INFO 5502
Search Products, resources, docs (/)
Clusters
CREATE CLUSTER
REFRESH
START
STOP
DELETE
REGIONS
5 RECOMMENDED ALERTS
HIDE INFO PANEL
Filter Search clusters, press Enter
PERMISSIONS
LABELS
ADD PRINCIPAL
Filter Enter property name or value
Lab7_Pyspark.pdf
Lab7_Pyspark.ipynb
Show all

☑	Name	Status	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created
☑	jupyternotebook-cluster	Stopped	us-central1	us-central1-c	2	Off	dataproc-staging-us-central1-291624520352-pore1nms	26 Mar 2022, 18:08:13

jupyternotebook-cluster

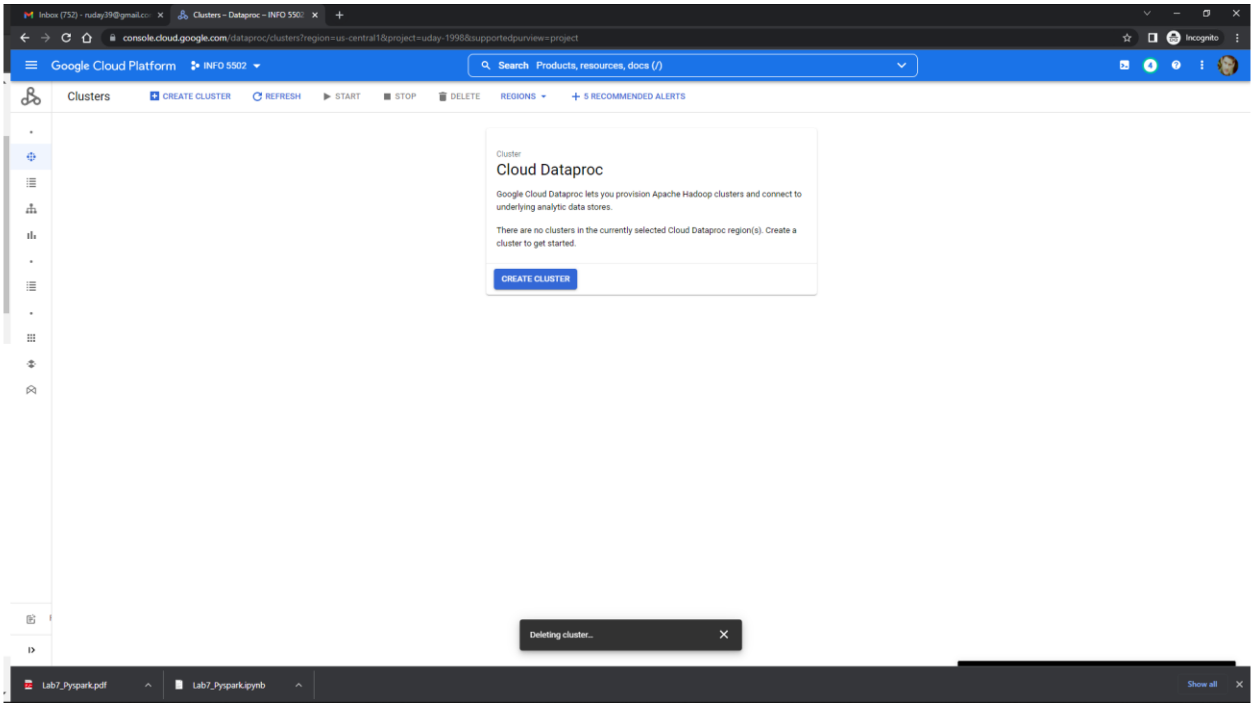
Edit or delete permissions below or 'Add principal' to grant new

Show inherited permissions

Role/Principal	Inheritance
Dataproc Service Agent (1)	
Editor (2)	
Owner (1)	

Stopping cluster...

https://console.cloud.google.com/dataproc/clusters/jupyternotebook-cluster?region=us-central1&project=uday-1998&supportedpurview=project



Cluster

Cloud Dataproc

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

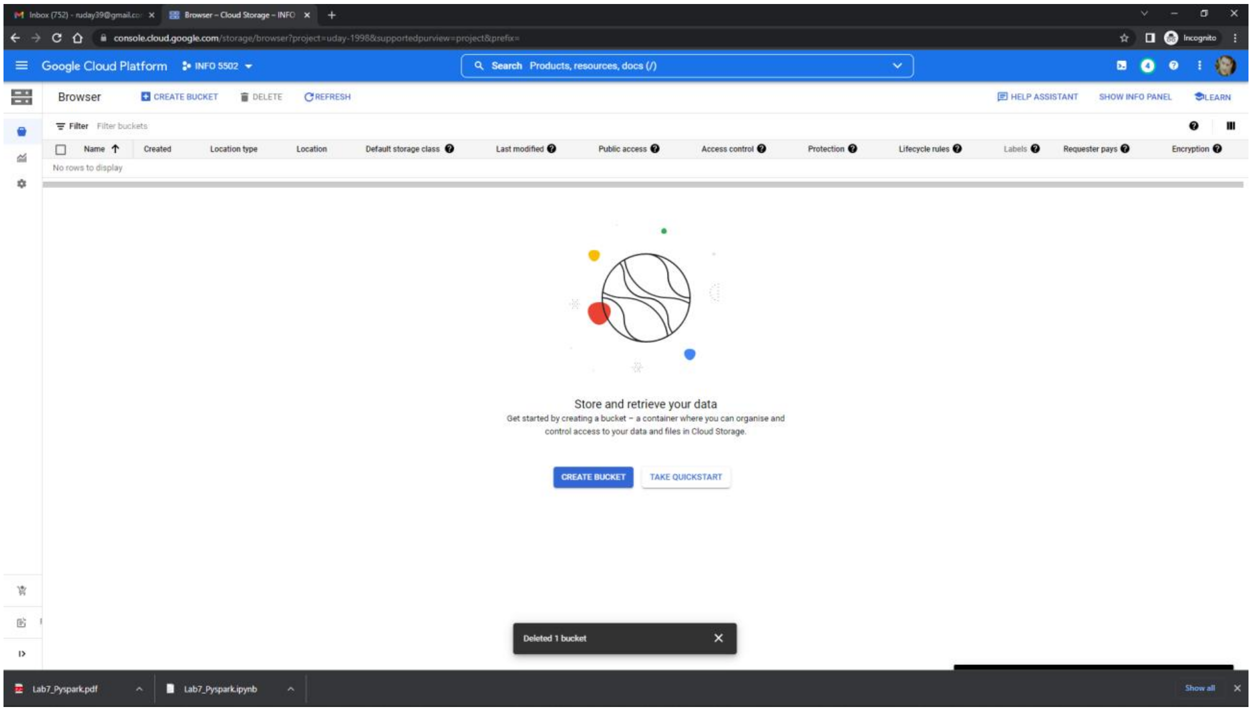
There are no clusters in the currently selected Cloud Dataproc region(s). Create a cluster to get started.

Deleting cluster...

Inbox (752) - ruday39@gmail.co
Clusters – Dataproc – INFO 5502
console.cloud.google.com/dataproc/clusters?region=us-central1&project=uday-1998&supportedpurview=project
Incognito
Google Cloud Platform
INFO 5502
Search Products, resources, docs (/)
Clusters
CREATE CLUSTER
REFRESH
START
STOP
DELETE
REGIONS
5 RECOMMENDED ALERTS
CREATE CLUSTER
Lab7_Pyspark.pdf
Lab7_Pyspark.ipynb
Show all

Step2: - Delete Bucket in GCP

- ➢ Delete Cluster which I was created.



Name ↑	Created	Location type	Location	Default storage class	Last modified	Public access	Access control	Protection	Lifecycle rules	Labels	Requester pays	Encryption
No rows to display												

Store and retrieve your data

Get started by creating a bucket – a container where you can organise and control access to your data and files in Cloud Storage.

Deleted 1 bucket