Instructions for using the Google Cloud Platform for TP3

Dear students, you will find below the instructions on how to use the Google Cloud Platform (GCP) required for the last part of TP3.

1. Obtaining GCP credits

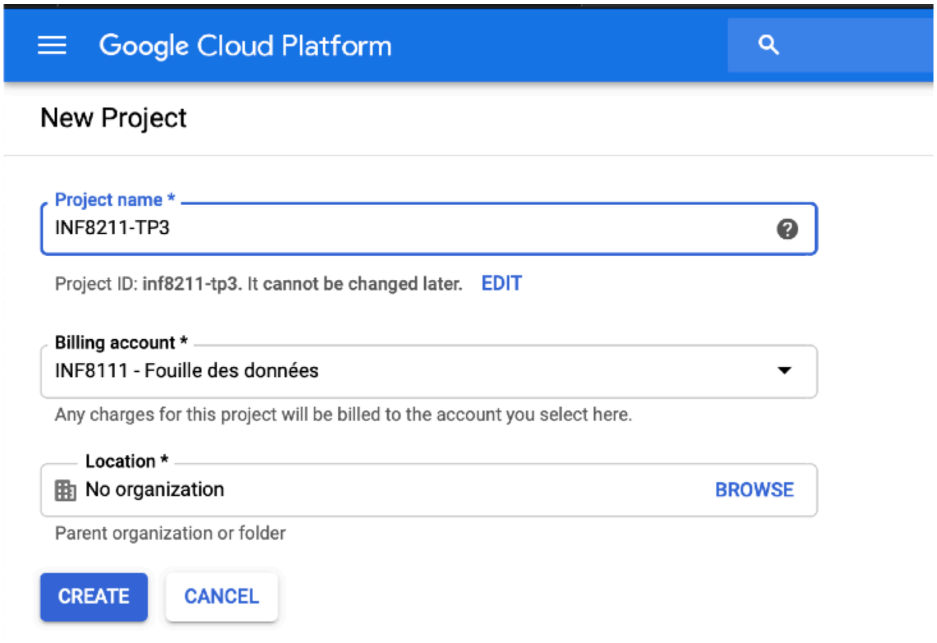
Here is the URL you will need to access in order to request a Google Cloud Platform coupon. You will be asked to provide your school email address and name. An email will be sent to you to confirm these details before a coupon is sent to you.

Student Coupon Retrieval Link

2. Creating a project

Once you have a billing account with your credits, go to your console and click to create a project.

Choose a name and the select the INF8111 - Fouille des données as your billing account linked to the project.



Project name *
INF8211-TP3

Project ID: inf8211-tp3. It cannot be changed later.

Billing account *
INF8111 - Fouille des données

Any charges for this project will be billed to the account you select here.

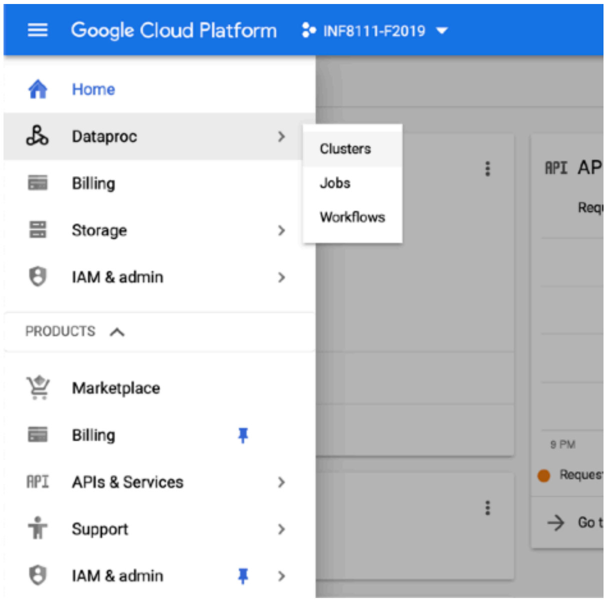
Location *
No organization

Parent organization or folder

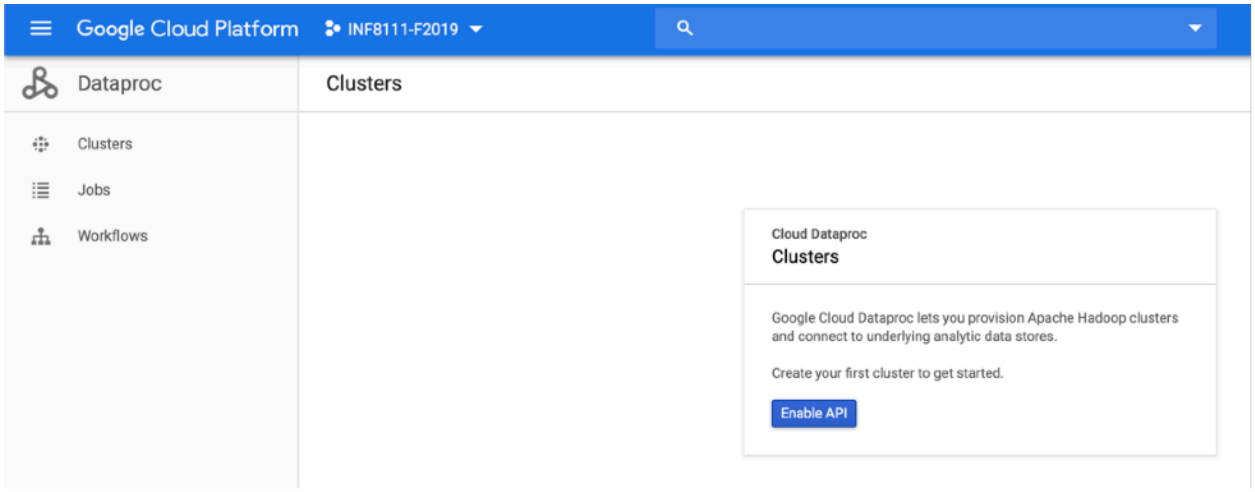
3. Enabling the required APIs.

To run our MBA algorithm, we will use the Dataproc service. However, first we need to Enable the APIs

On your console, click on the 3 lines on the top left and search for Dataproc -> Clusters



Next, click on Enable API. This process can take a few minutes.



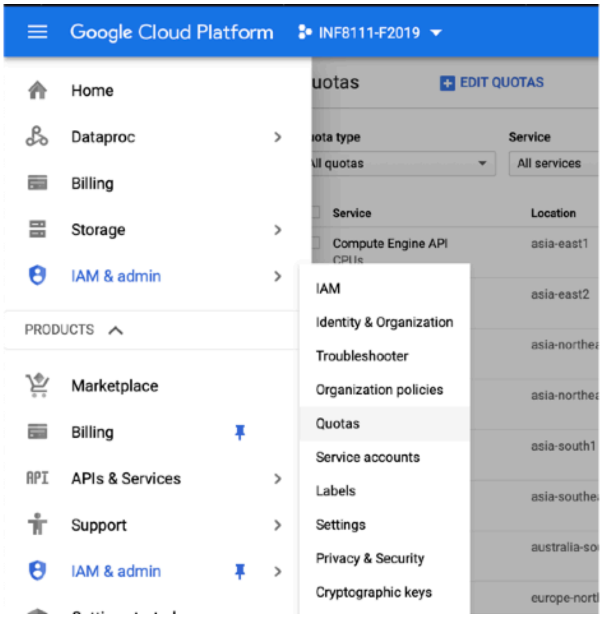
Google Cloud Dataproc lets you provision Apache Hadoop clusters and connect to underlying analytic data stores.

Create your first cluster to get started.

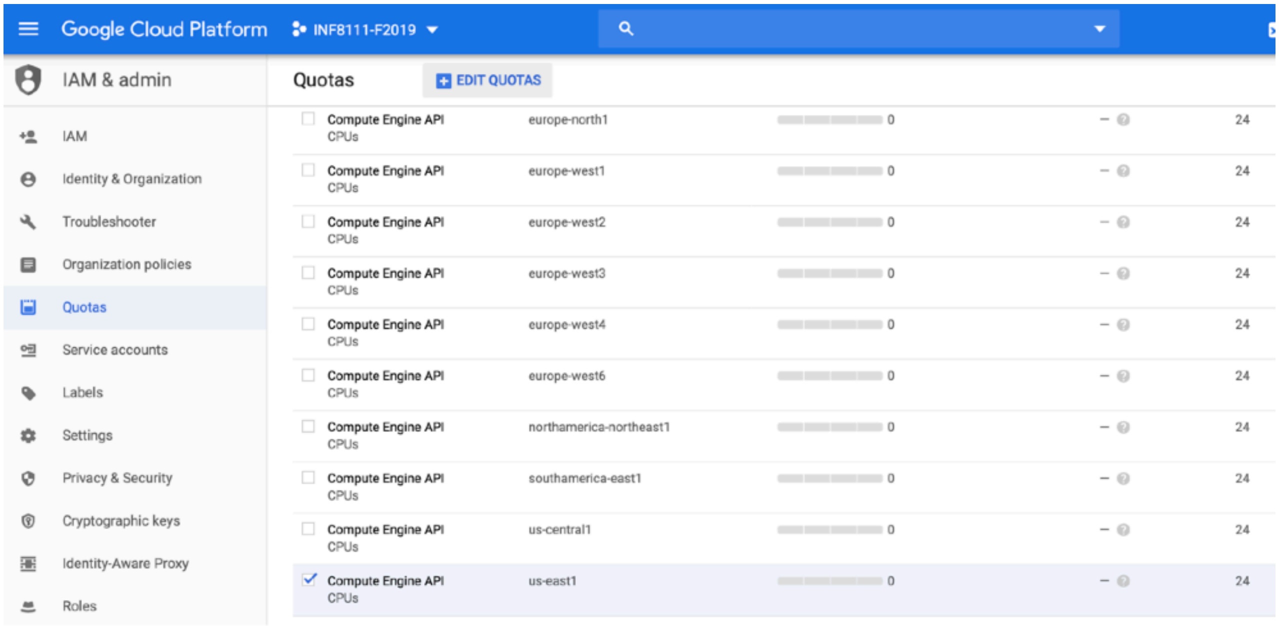
4. Requiring for more CPU cluster capacity

By default, the maximum number of CPUs allowed by GCP for this student credit account is 24, but we will need much more than that.

On your console, click on the 3 lines on the top left and search for IAM & admin -> Quotas

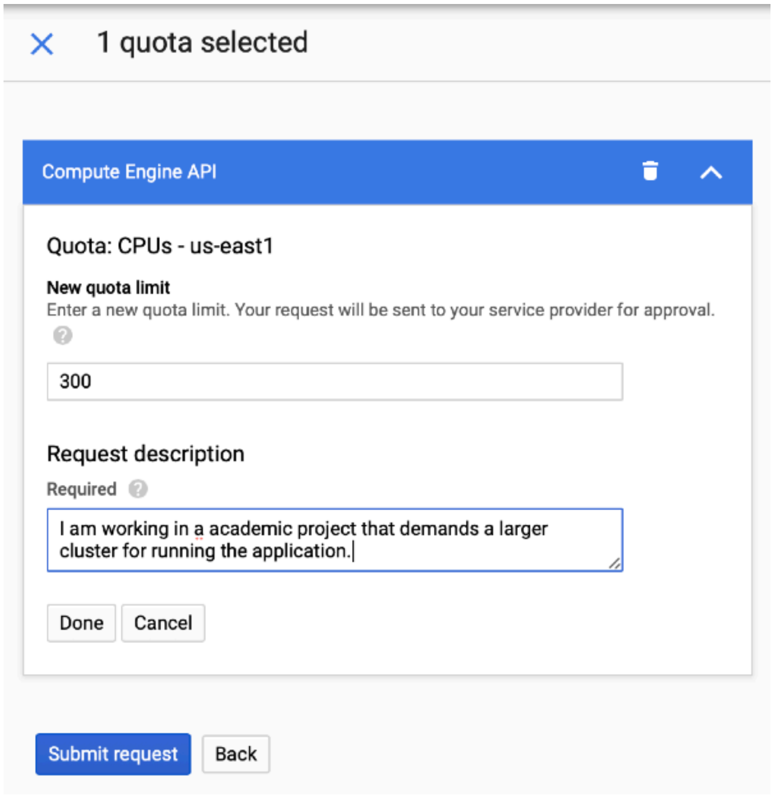


Once there, select “CPUs” under the **Metric** select box and look for “Compute Engine API” for the “us-east1” location. Select it and click on Edit Quotas.



□	Compute Engine API CPUs	europe-north1	0	—	24
□	Compute Engine API CPUs	europe-west1	0	—	24
□	Compute Engine API CPUs	europe-west2	0	—	24
□	Compute Engine API CPUs	europe-west3	0	—	24
□	Compute Engine API CPUs	europe-west4	0	—	24
□	Compute Engine API CPUs	europe-west6	0	—	24
□	Compute Engine API CPUs	northamerica-northeast1	0	—	24
□	Compute Engine API CPUs	southamerica-east1	0	—	24
□	Compute Engine API CPUs	us-central1	0	—	24
☑	Compute Engine API CPUs	us-east1	0	—	24

Once asked for the new quota limit, inform 300 and in the description box write something similar to the one showing the image below.



1 quota selected

Compute Engine API

Quota: CPUs - us-east1

New quota limit
Enter a new quota limit. Your request will be sent to your service provider for approval.

300

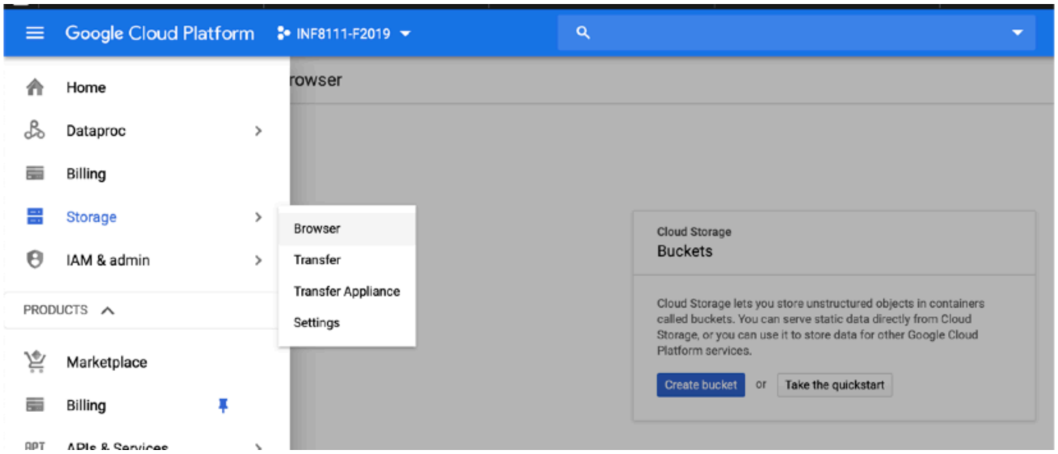
Request description
Required

I am working in a academic project that demands a larger cluster for running the application.

You will receive an email confirming your request. GCP usually takes between 30 minutes and a couple hours to process your request.

5. Creating a storage bucket

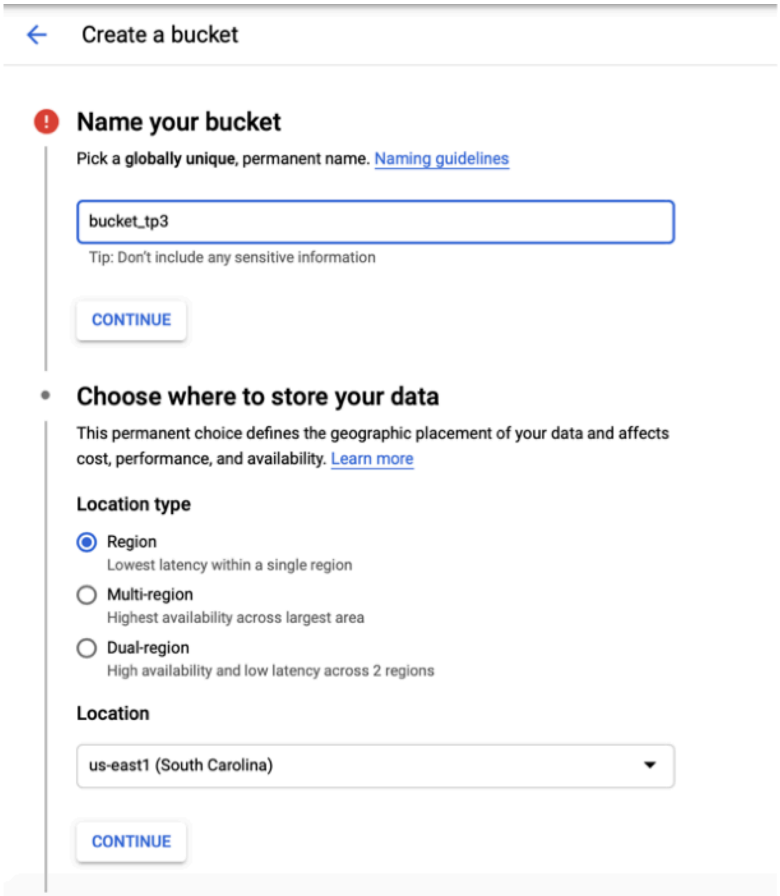
On your console, click on the 3 lines on the top left and search for Storage -> Browser and click in "Create bucket".



Cloud Storage
Buckets

Cloud Storage lets you store unstructured objects in containers called buckets. You can serve static data directly from Cloud Storage, or you can use it to store data for other Google Cloud Platform services.

Give your bucket and name and under the “Choose where to store your data”, select Region and search for us-east1 (same as the region that you ask for the quota increment) and press “Create”.



Create a bucket

Name your bucket

Pick a **globally unique**, permanent name. Naming guidelines

bucket_tp3

Tip: Don't include any sensitive information

Choose where to store your data

This permanent choice defines the geographic placement of your data and affects cost, performance, and availability. Learn more

Location type

● Region
Lowest latency within a single region

○ Multi-region
Highest availability across largest area

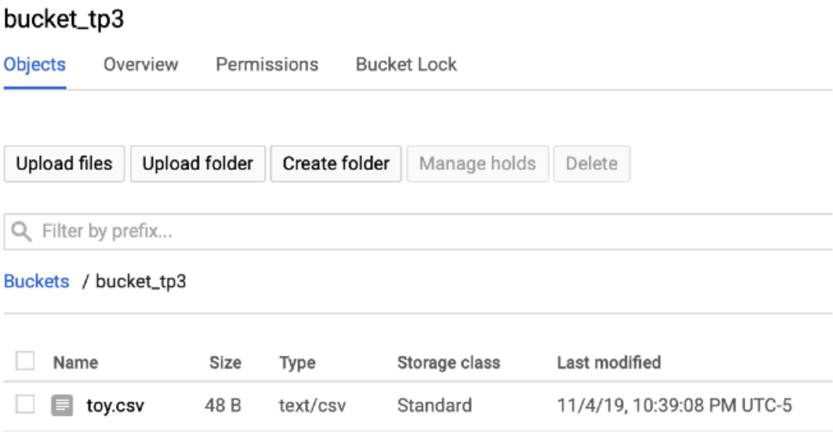
○ Dual-region
High availability and low latency across 2 regions

Location

us-east1 (South Carolina)

You will be redirected to your bucket page from where you starting uploading some files.

As an example, upload the toy.csv file to your bucket.

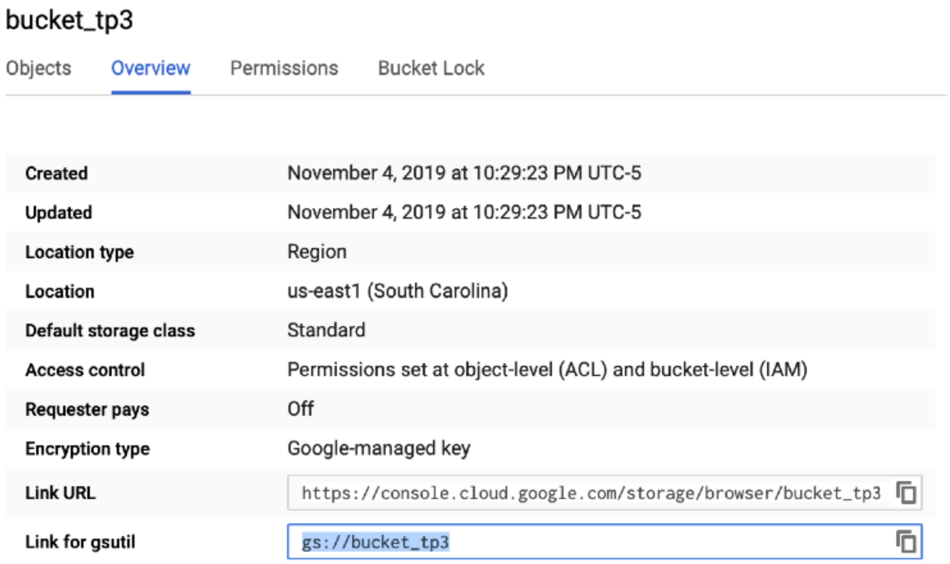


bucket_tp3

Buckets / bucket_tp3

□ Name	Size	Type	Storage class	Last modified
□ toy.csv	48 B	text/csv	Standard	11/4/19, 10:39:08 PM UTC-5

If you go the the Overview tab, the **Link for gsutil** gives you the address for your bucket. For example, to access my toy.csv file contained in my bucket, it path would be “gs://bucket_tp3/toy.csv".



Created	November 4, 2019 at 10:29:23 PM UTC-5
Updated	November 4, 2019 at 10:29:23 PM UTC-5
Location type	Region
Location	us-east1 (South Carolina)
Default storage class	Standard
Access control	Permissions set at object-level (ACL) and bucket-level (IAM)
Requester pays	Off
Encryption type	Google-managed key
Link URL	https://console.cloud.google.com/storage/browser/bucket_tp3
Link for gsutil	gs://bucket_tp3

6. Creating a computing cluster

Now everything is set for creating our cluster. Go again to the Dataproc -> Clusters and press Create Cluster.

You don’t have to change the name for the cluster, but it is necessary to specify the **Region.** Select us-east1 (or the region for which you requested a quota increase).

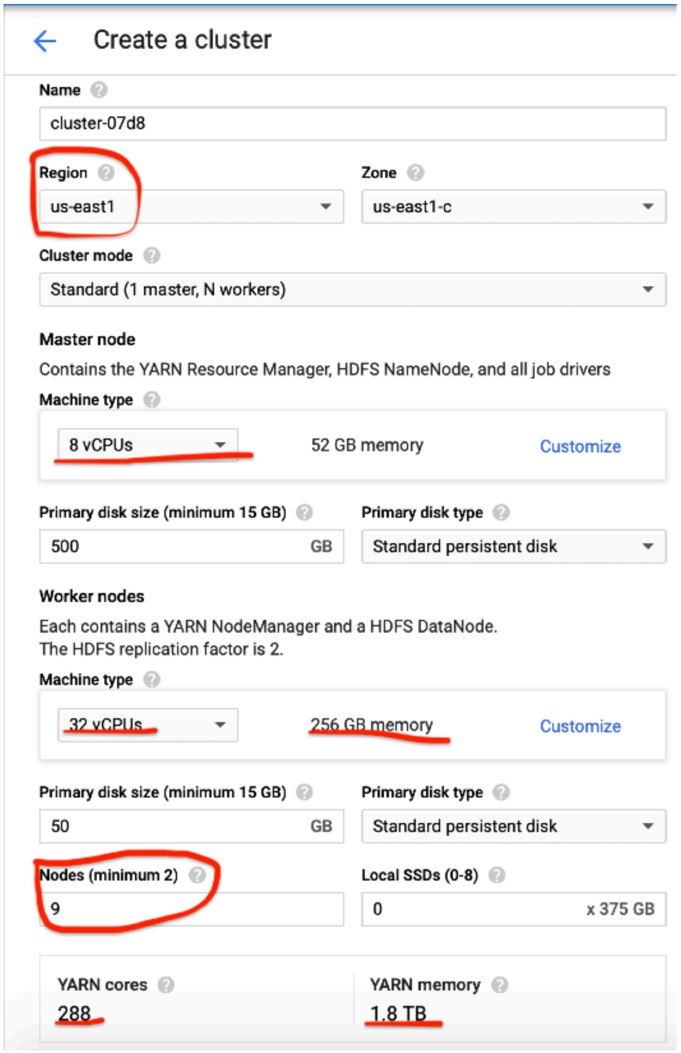
Now we have to set the number of CPUs that we will use in our cluster. The Cluster mode is the Standard(1 master, N workers)

In our application the most valuable resource is memory. Thus, both for the master node as for the workers nodes will will use machines from the type highmem.

- For the master node, select the 8vCPUs of type **n1-highmem-8.**

- For the worker nodes, select 9 nodes of 32vCPUs of type **n2-highmem-32.**

This will give your cluster an total of (8 + 288) vCPUs and 1.8 TB of memory.



Create a cluster

Name: cluster-07d8

Region: us-east1

Zone: us-east1-c

Cluster mode: Standard (1 master, N workers)

Master node

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

Machine type: 8 vCPUs — 52 GB memory — Customize

Primary disk size (minimum 15 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: 32 vCPUs — 256 GB memory — Customize

Primary disk size (minimum 15 GB): 50 GB

Primary disk type: Standard persistent disk

Nodes (minimum 2): 9

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

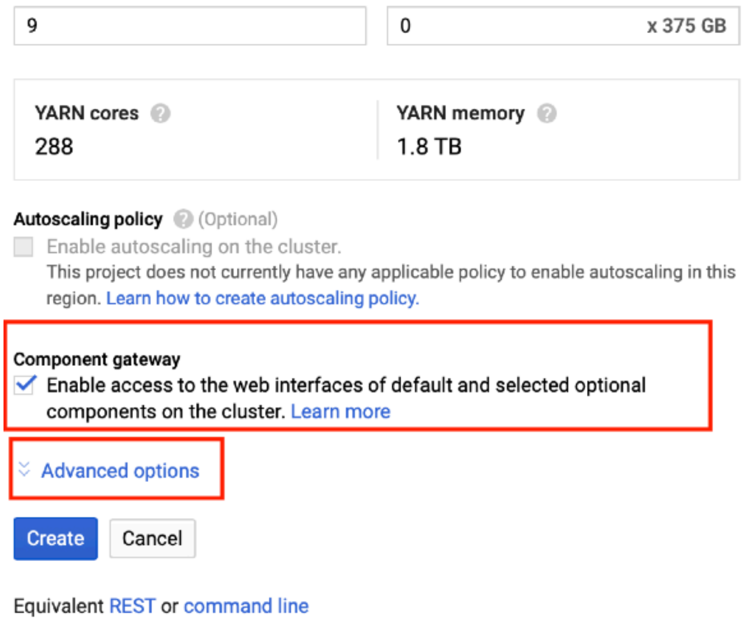
YARN cores: 288

YARN memory: 1.8 TB

Note: this cluster configuration is only a suggestion and may be advisable to try a smaller cluster in your first run. For example, you could first try to run the section 3.2 with a smaller cluster and then increase it to this configuration for running the application in 3.3. Also, learn how to calculate the price of a cluster, which can be done here**.**

VERY IMPORTANT:
There is still a crucial step in the cluster configuration to be done.

First, select the **Component gateway** option and click to expand the advance options:


9

0 x 375 GB

YARN cores 288

YARN memory 1.8 TB

Autoscaling policy (Optional)

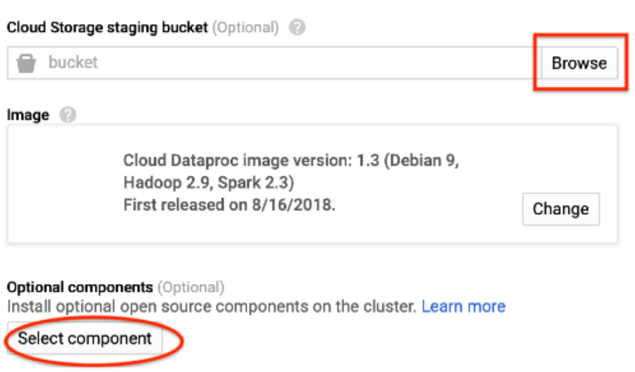
□ Enable autoscaling on the cluster.
This project does not currently have any applicable policy to enable autoscaling in this region. Learn how to create autoscaling policy.

Component gateway

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

Equivalent REST or command line

Look for **Cloud Storage staging bucket** and browser your bucket;
In **Optional components**, select ANACONDA and JUPYTER.

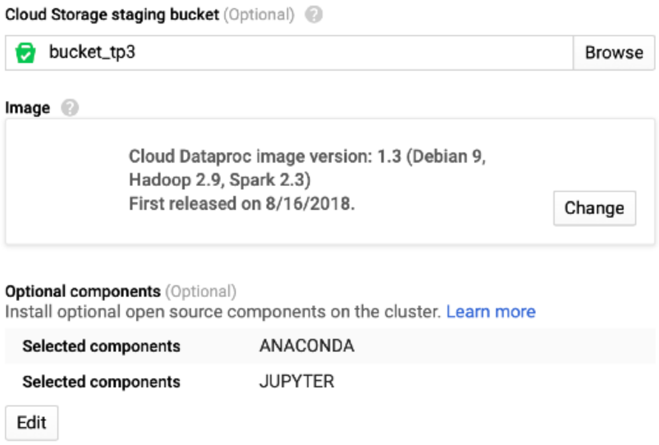

Cloud Storage staging bucket (Optional)

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


Cloud Storage staging bucket (Optional)

bucket_tp3

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

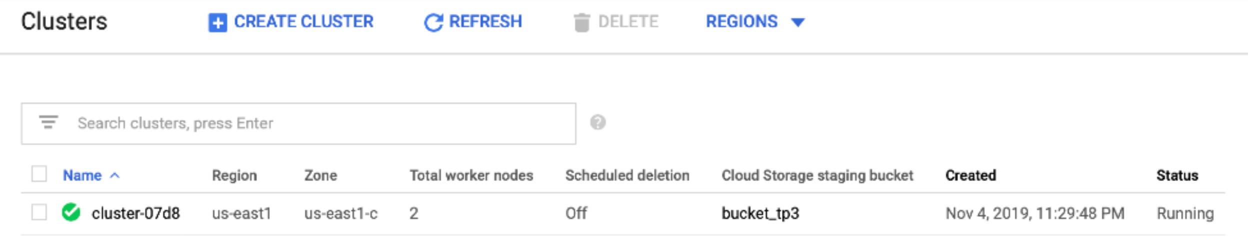
Selected components	ANACONDA
Selected components	JUPYTER

Warning: as we finish the configuration of our cluster and press create, GCP will start charging your billing account. Always remember to delete the cluster once you have finished your experiment.

Finally, press **Create** to create the cluster**.** It may take a few minutes until the cluster is created and ready to be used.

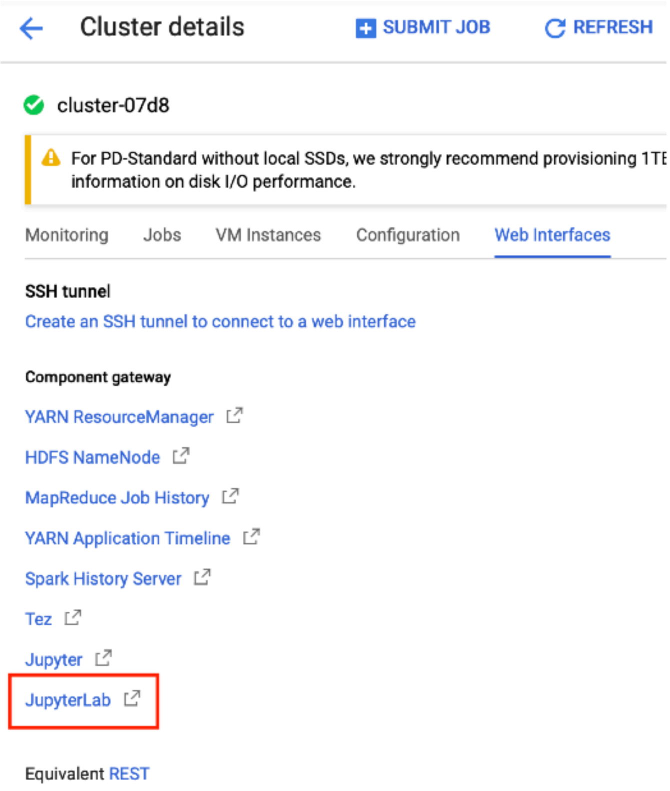
7. Using your cluster

Once your cluster is created, click to open it.



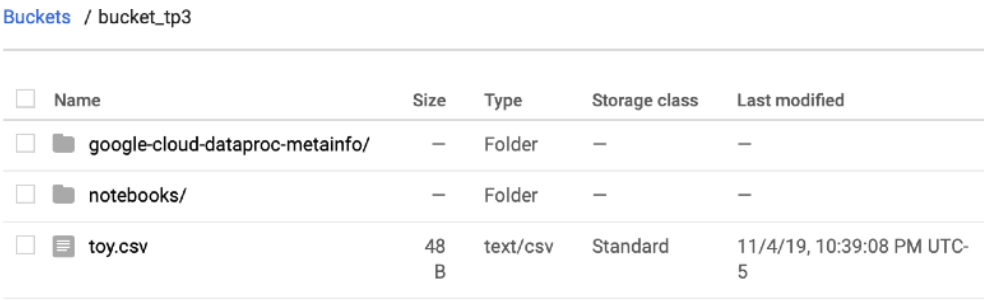
Name	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created	Status
cluster-07d8	us-east1	us-east1-c	2	Off	bucket_tp3	Nov 4, 2019, 11:29:48 PM	Running

Go to the **Web Interface** tab and click on JupyterLab



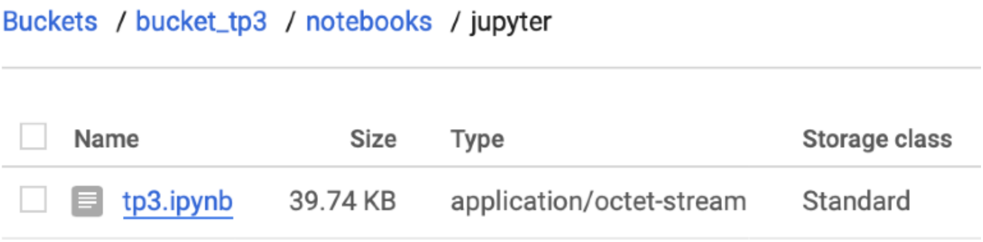
For PD-Standard without local SSDs, we strongly recommend provisioning 1TE information on disk I/O performance.

Now, go again to Storage -> Browser and open your bucket. We will see that now there is a notebooks folder.



Name	Size	Type	Storage class	Last modified
google-cloud-dataproc-metainfo/	—	Folder	—	—
notebooks/	—	Folder	—	—
toy.csv	48 B	text/csv	Standard	11/4/19, 10:39:08 PM UTC-5

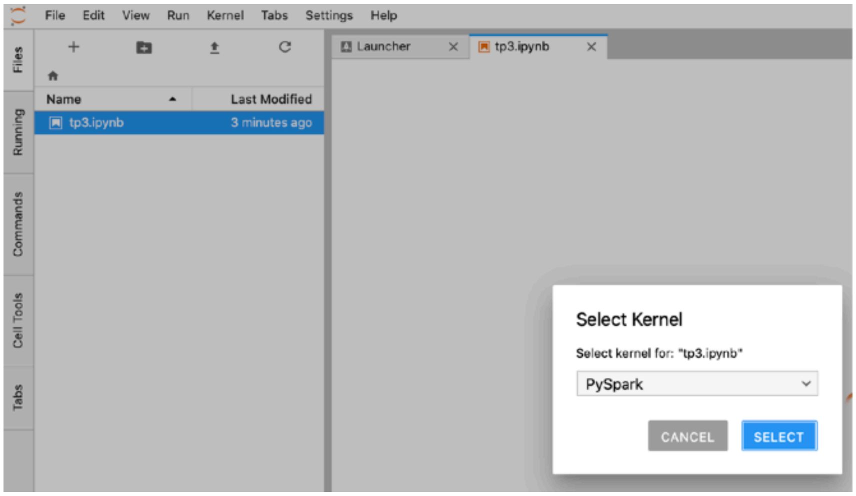
Go to the *notebooks/jupyter* folder and upload your .ipynb file.



Buckets / bucket_tp3 / notebooks / jupyter

□	Name	Size	Type	Storage class
□	tp3.ipynb	39.74 KB	application/octet-stream	Standard

The page that was open when you clicked in JupyterLab now should showing your Jupyter file. Open it and select the PySpark kernel.



Name	Last Modified
tp3.ipynb	3 minutes ago

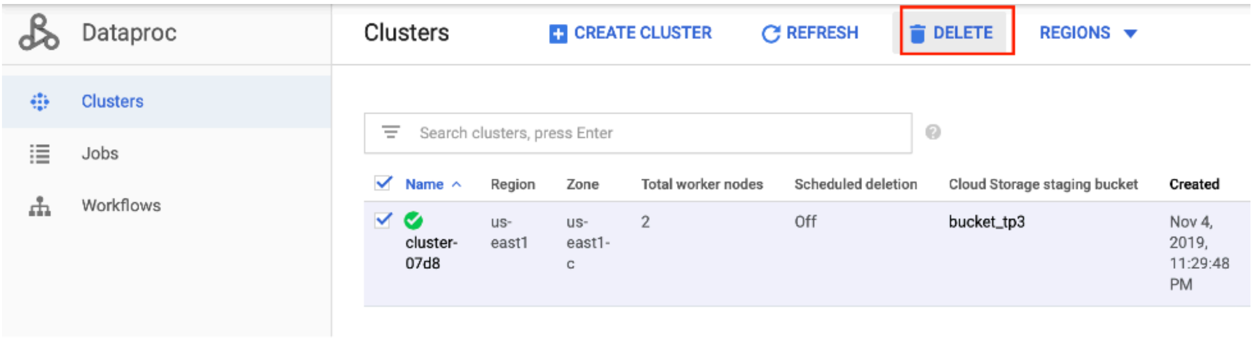
Select Kernel

Select kernel for: "tp3.ipynb"

PySpark

Just run your notebook as usual.

Once you have finished using the cluster, go to Dataproc -> clusters, select the cluster you desire to exclude and press **Delete.**



☑	Name	Region	Zone	Total worker nodes	Scheduled deletion	Cloud Storage staging bucket	Created
☑	✓ cluster-07d8	us-east1	us-east1-c	2	Off	bucket_tp3	Nov 4, 2019, 11:29:48 PM