Instructions for using the Google Cloud Platform for TP2

Dear students, you will find below the instructions on how to use the Google Cloud Platform (GCP) required for the last part of TP2.

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.

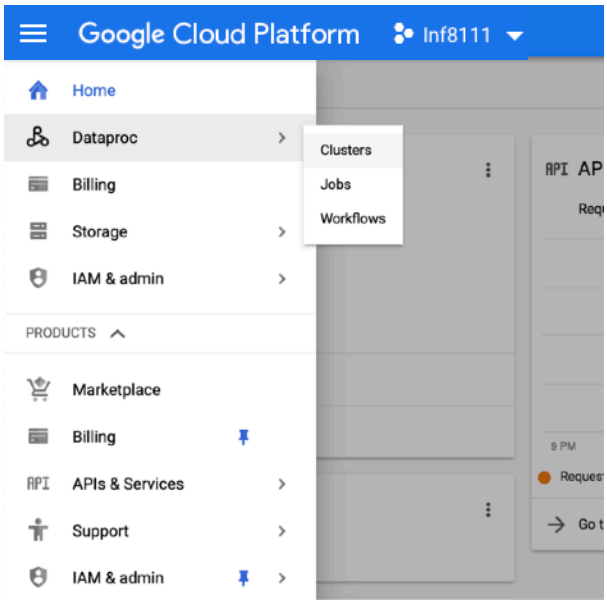
Student Coupon Retrieval Link

Once you have completed this step, you should have a project named **INF8111 - Fouille des données (Data mining).** This project is linked to a billing account with your credits of 50$.

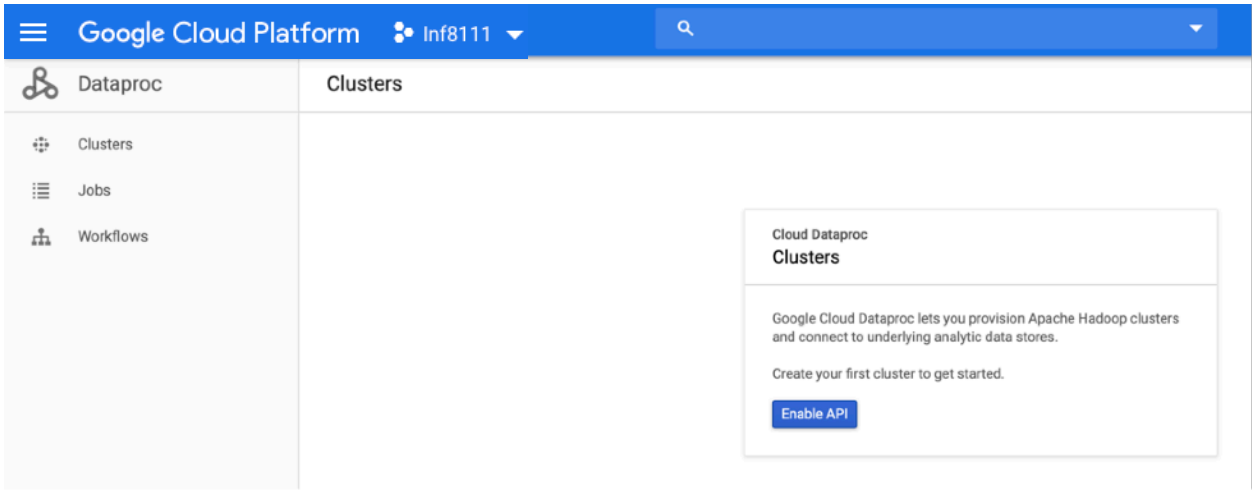
2. Enabling the required APIs.

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.

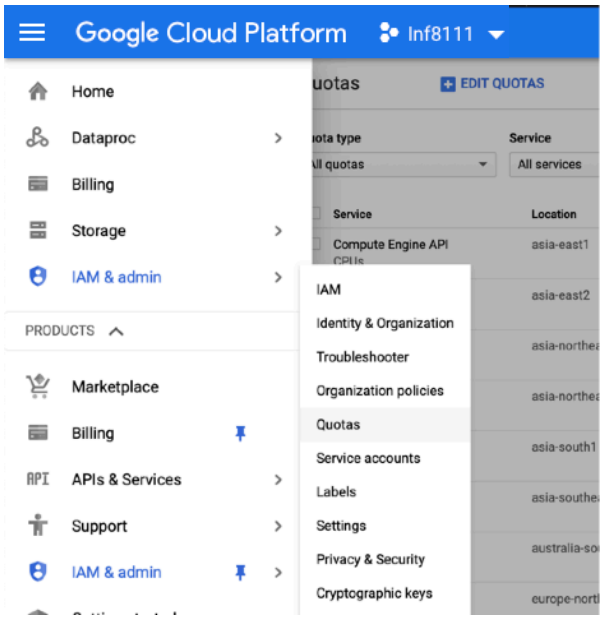


* this may be trigged automatically the first time you access this page.

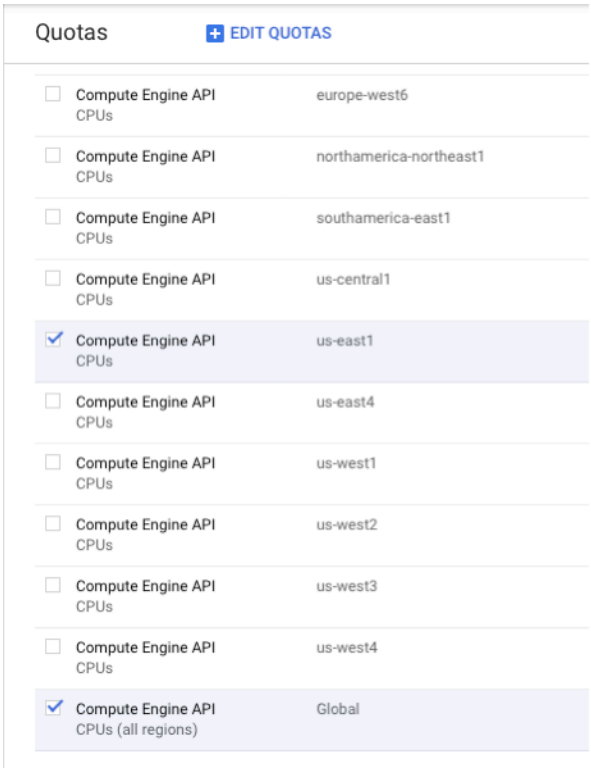
3. Requiring for more CPU cluster capacity

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.

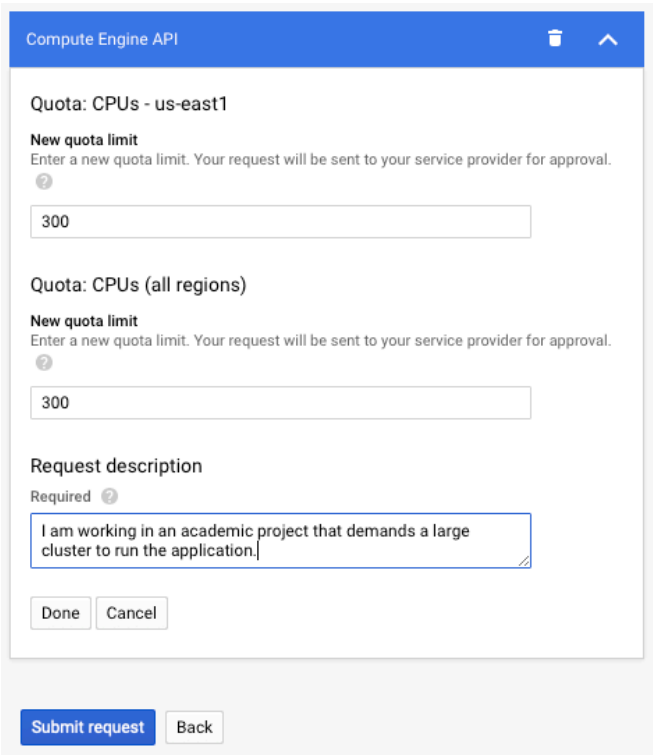
On your console, click on the 3 lines on the top left and search for IAM & admin -> Quotas



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



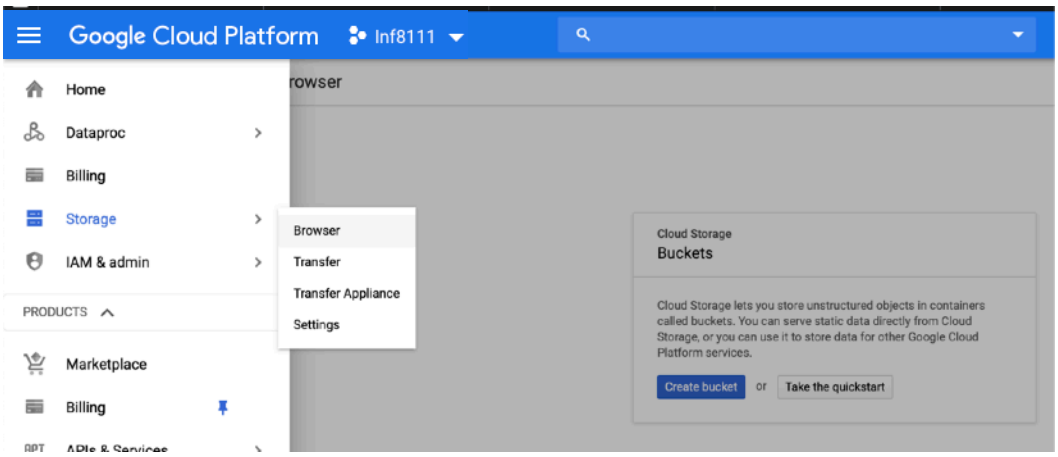
Once asked for the new quota limit, inform 300 and in the description box write something similar to the one showing the image below.



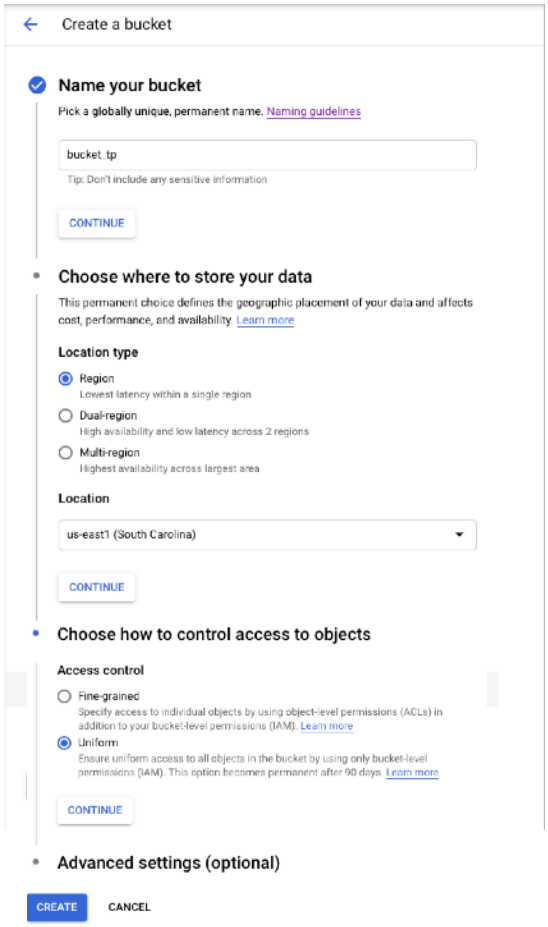
You will receive an email confirming your request. GCP usually takes between 30 minutes and a couple hours to process your request.

5. Creating a storage bucket

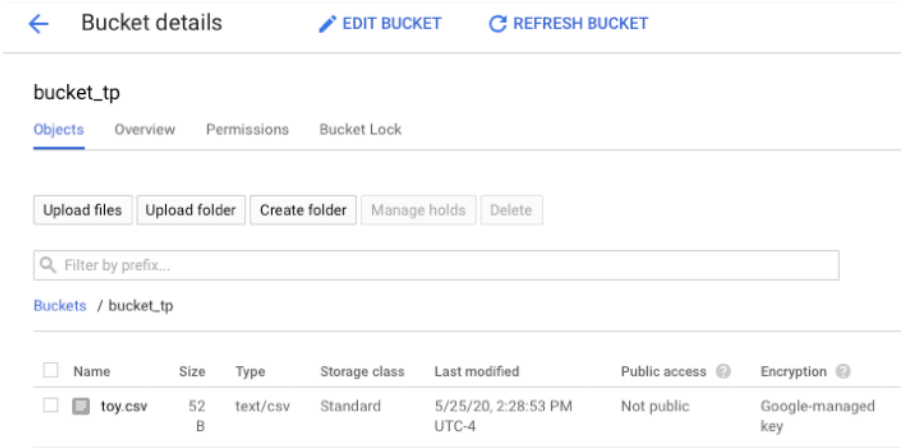
On your console, click on the 3 lines on the top left and search for Storage -> Browser and click in “Create bucket”.



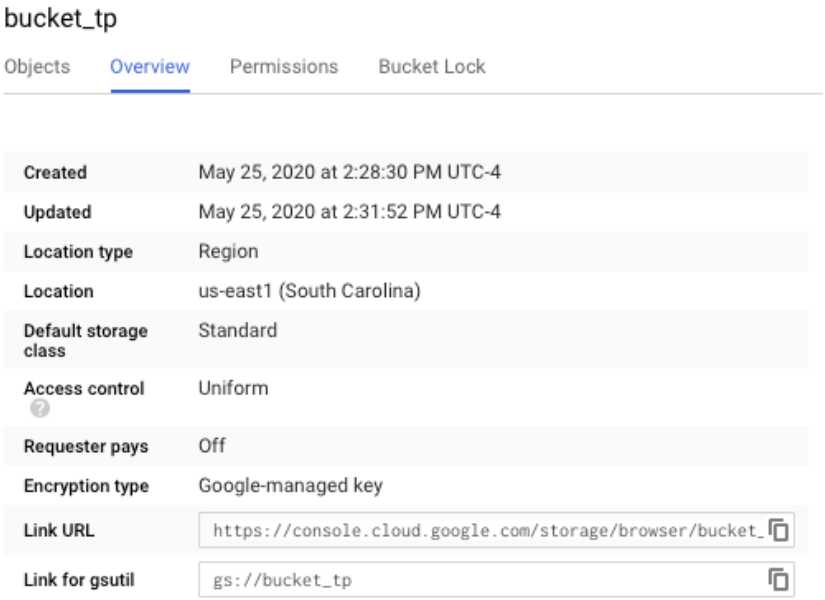
Give your bucket and name and under the “Choose where to store your data”, select Region and search for us-east1 (same as the region that you ask for the quota increment). Also set the access to objects as **uniform** and press “Create”.



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.



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_tp/toy.csv".



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)

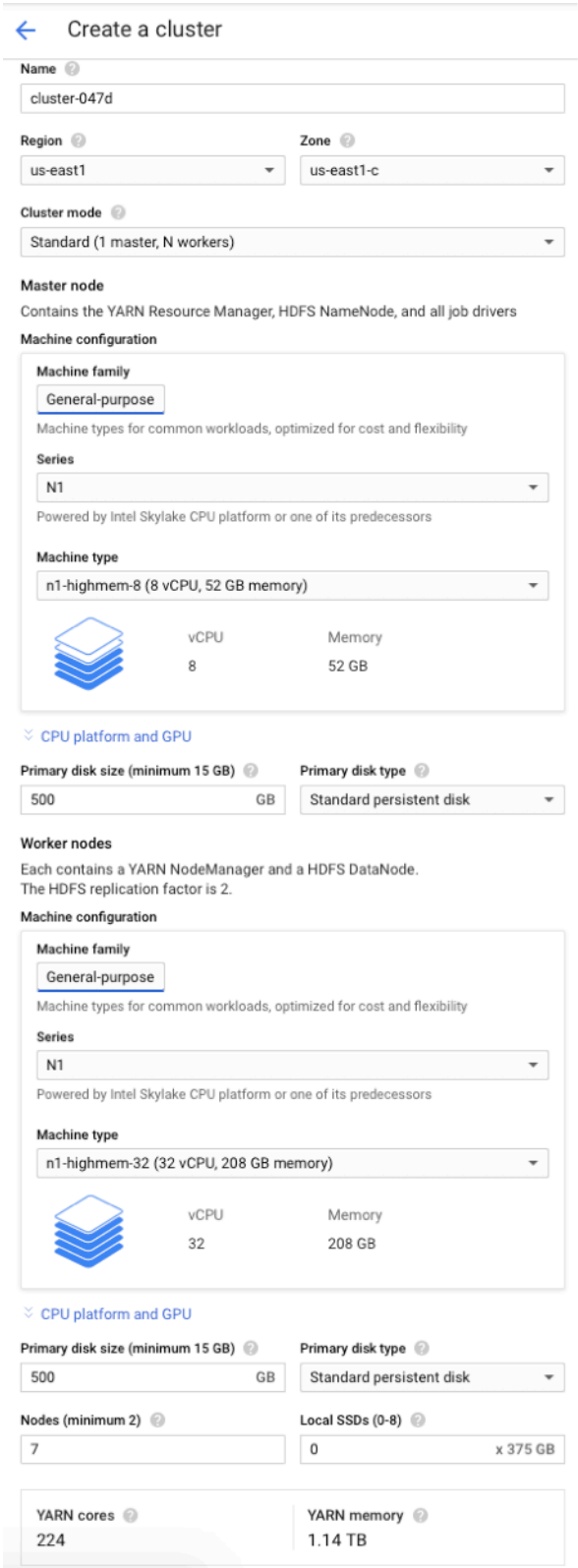
In our application the most valuable resource is memory. Thus, both for the master node as for the workers nodes will will use machines from the type highmem.

- For the master node, select the 32vCPUs of type **n1-highmem-32.**

- For the worker nodes, select 7 nodes of 32vCPUs of type **n2-highmem-32.**

This will give your cluster an total of 256 (32 +224) vCPUs and 1.14 TB of memory on the workers nodes.

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 up 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. For example, for this given configuration, we hourly price would be:

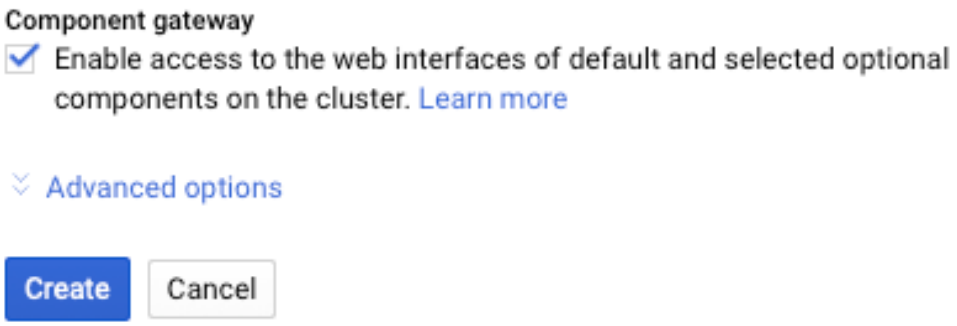


1*0.4 + 7*0.4 = $3.2 per hour.

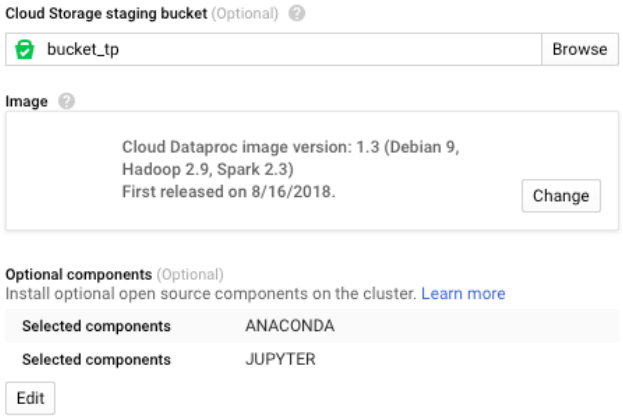
Alternatively, you can do the pricing calculation using this application**. There, navigate on the applications to find Cluster Dataproc and put the the cluster configuration that you want to estimate.**

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:



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

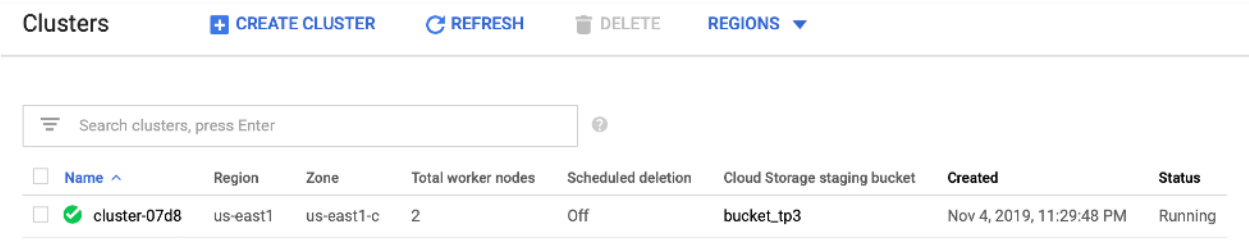


Warning: as you finish the configuration of your cluster and press create, GCP will start charging your billing account. Always remember to delete the cluster once you have finished your experiment.

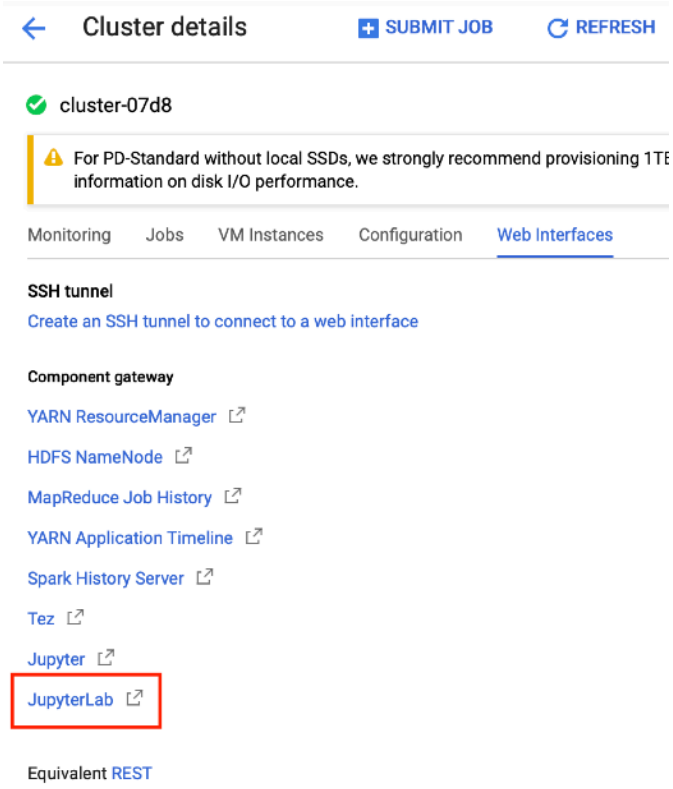
Finally, press **Create** to create the cluster**.** It may take a few minutes until the cluster is created and ready to be used.

7. Using your cluster

Once your cluster is created, click to open it.



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



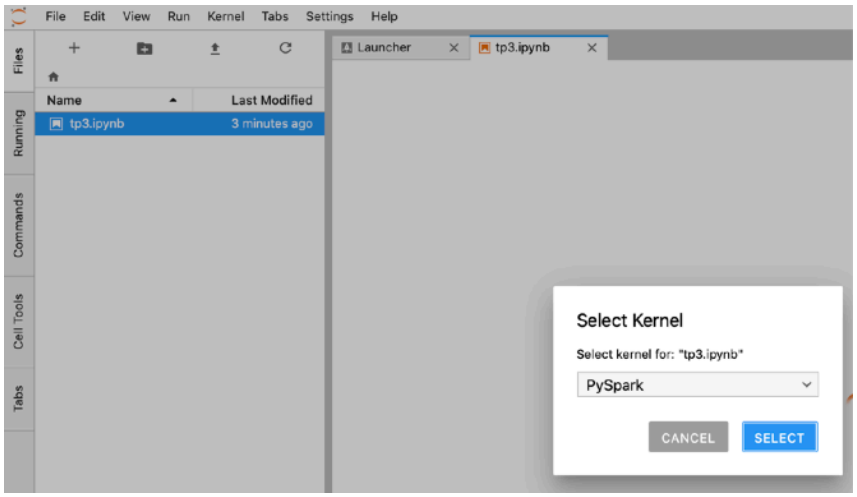
Now, go again to Storage -> Browser and open your bucket. We will see a notebooks folder.

Buckets / bucket_tp3

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.

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



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.**