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

- You will be asked for a name and email address, which needs to match your school domain. A confirmation email will be sent to you with a coupon code.
- You can only request ONE code per unique email address.

Please contact me if you have any questions or issues.

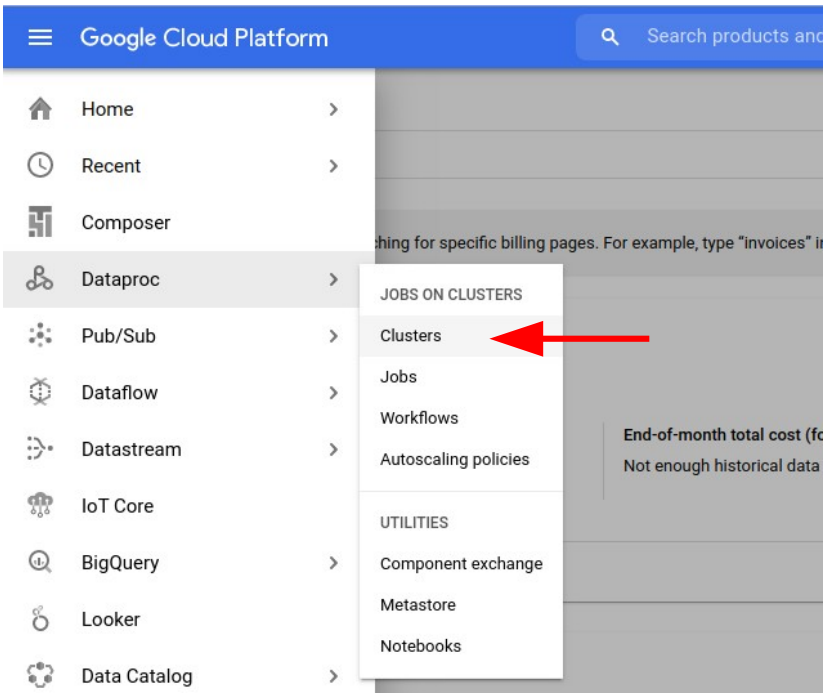
Once you have completed this step, you can see your billing account details with your credits of 50$ from the following link:

https://console.cloud.google.com/billing

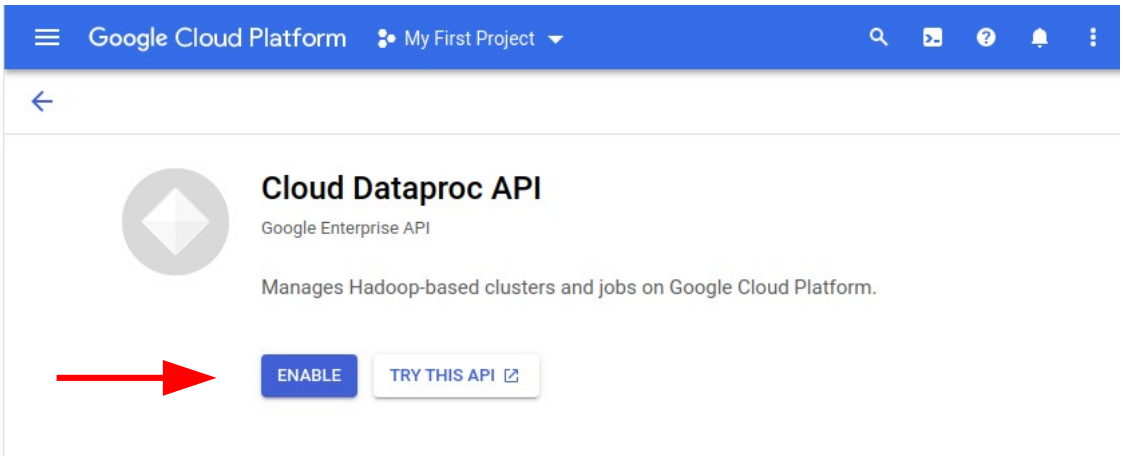
2. Enabling the required APIs

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** (Dataproc is located under the category of **big data**).



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

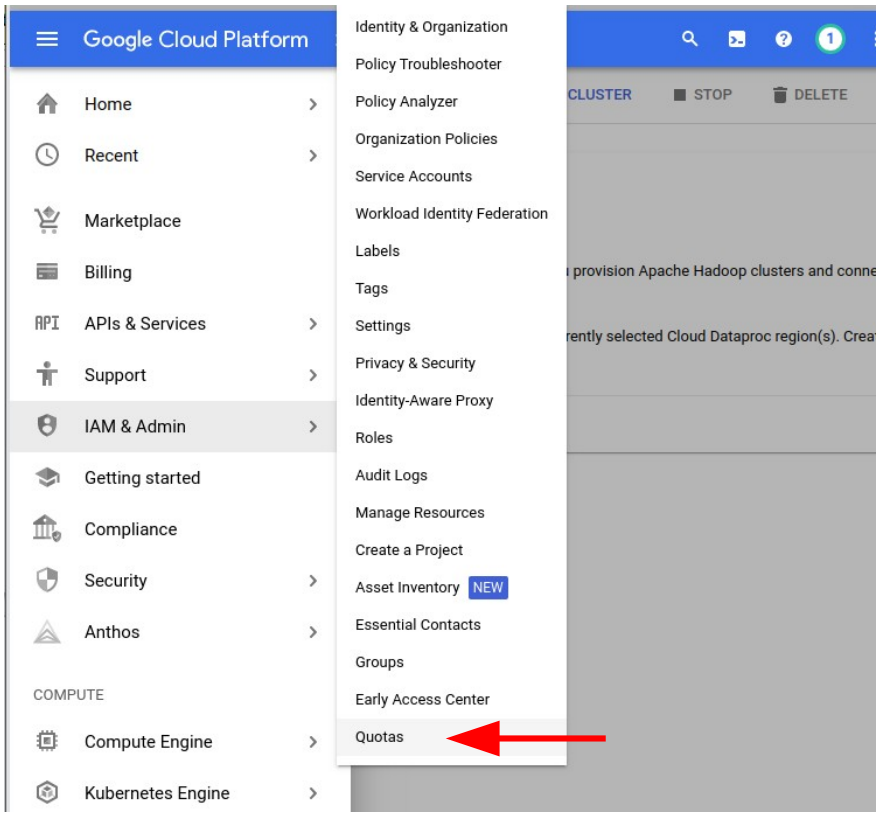


* this may be trigged automatically the first time you access this page.

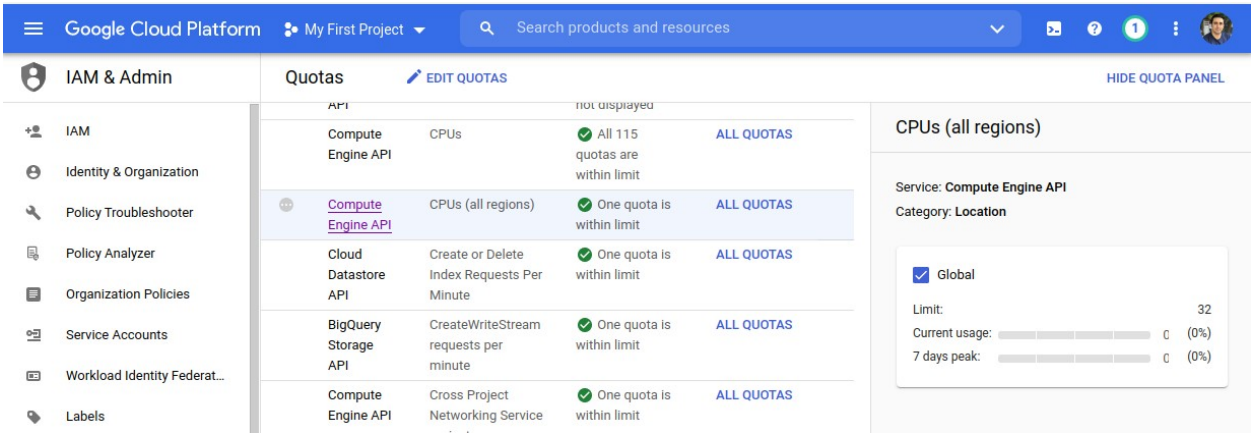
3. Requiring for more CPU cluster capacity

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.

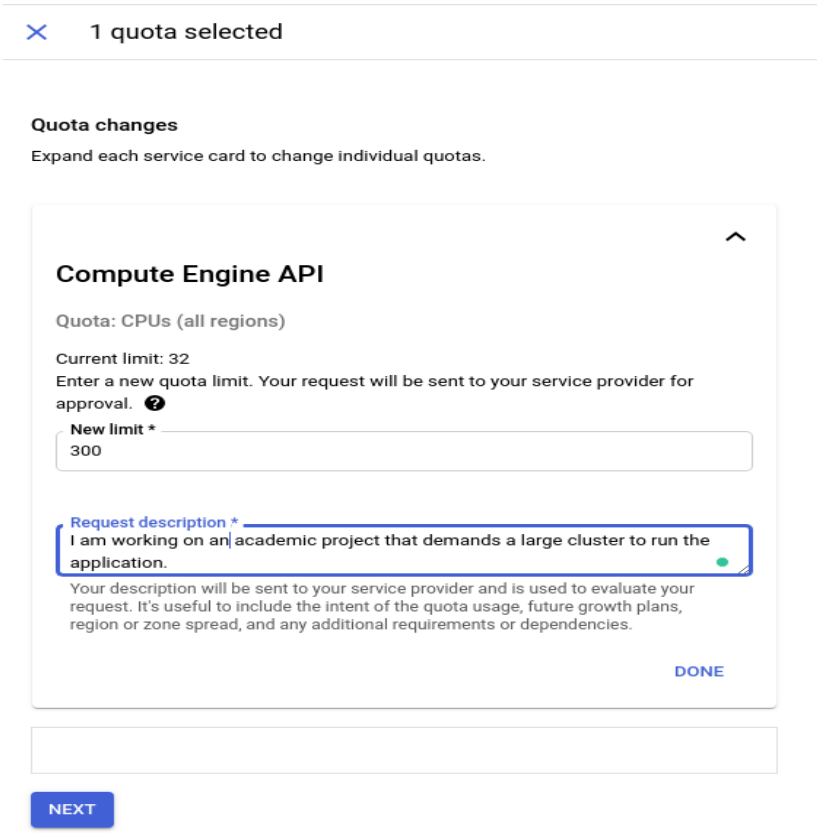
On your console, click on the 3 lines on the top left and search for **IAM & admin -> Quotas**



Once there, among the “Compute Engine API” services, look for the limit name of “CPUs (all regions)”. For more convenience, you can sort the table by limit names. Select it and click on Edit Quotas.



Then enter **128** in the **new limit** field and write something similar to the one showing the image below in the description box.



You will receive an email confirming your request. Your request usually takes between 30 minutes and 48 hours to be processed.

Then, among the “Compute Engine API” services, look for the limit name of “us-east1” and ask for a new quota limit (for example 128).

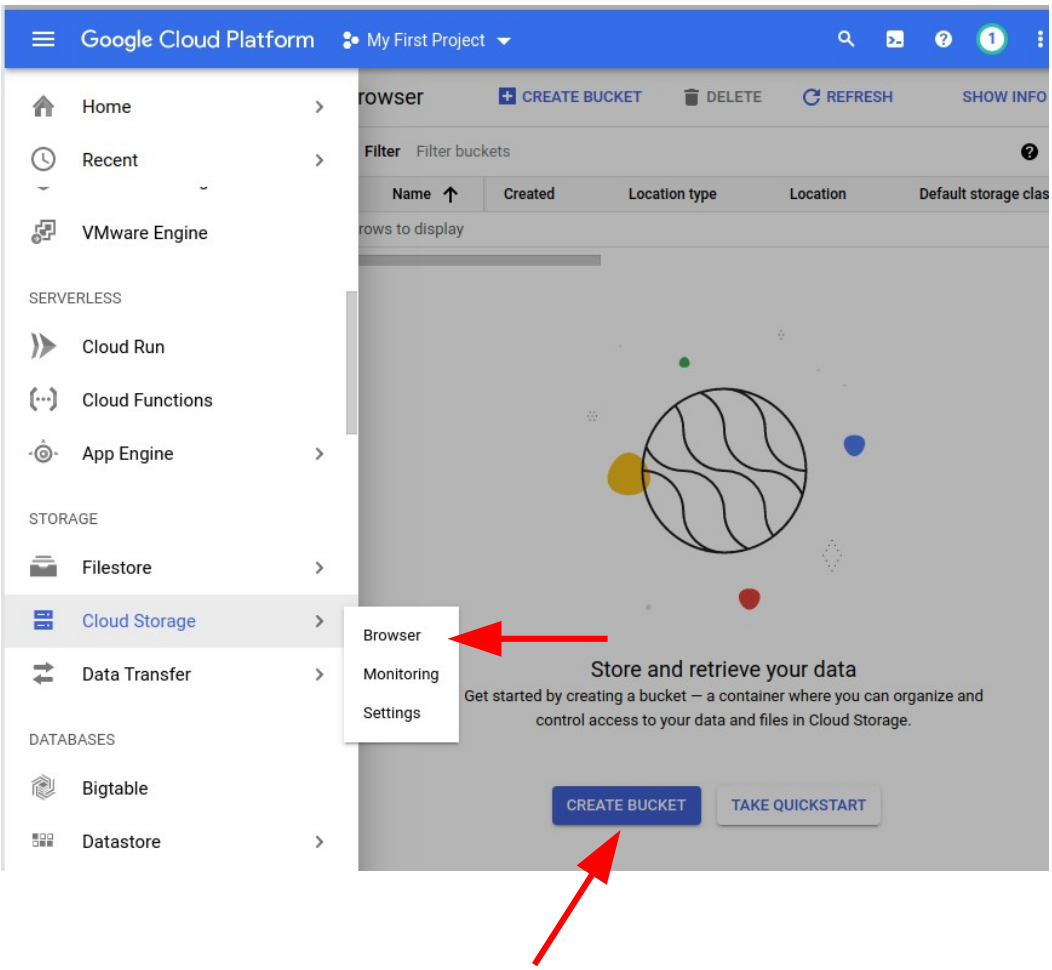
Note: Your request for CPUs (all regions) may be rejected and you may receive a message similar to Rear's message:

> “Unfortunately, we are unable to grant you additional quota at this time. If this is a new project please wait 48h until you resubmit the request or until your Billing account has additional history.”

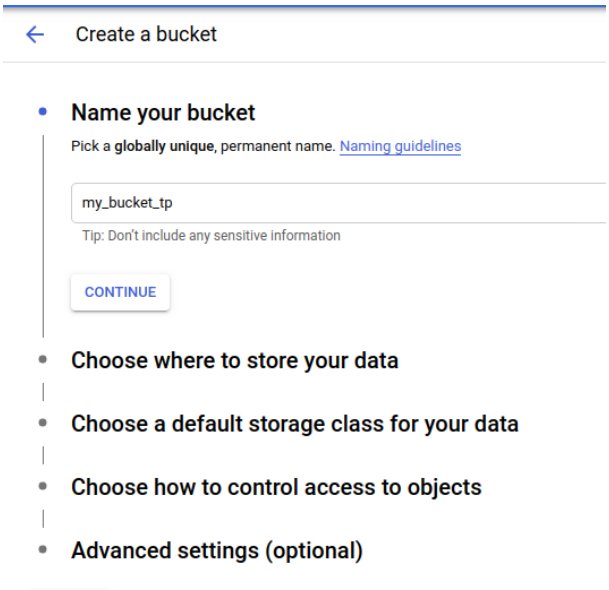
In this case, try to split your desired increase in quota into some steps that could be approved like 32 -> 64 -> 128. In this way, you can acquire 128 CPUs (all regions) or even more.

4. Creating a storage bucket

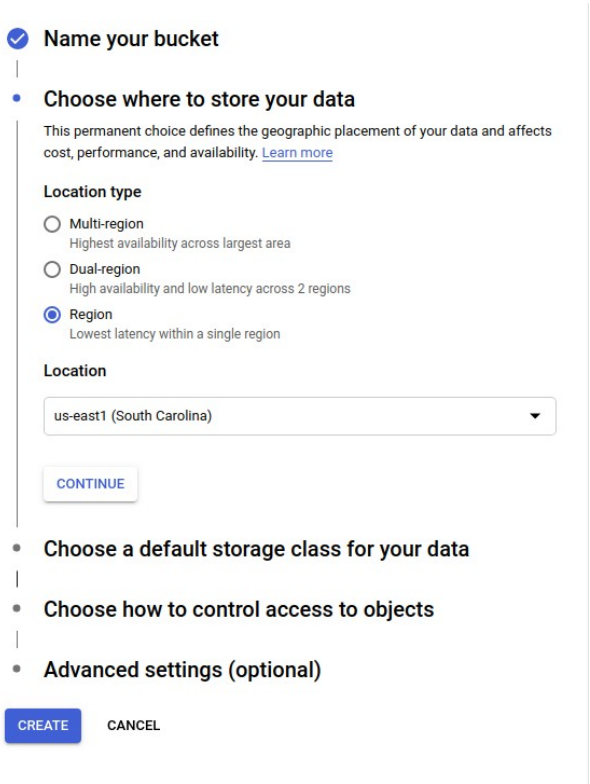
On your console, click on the 3 lines on the top left and search for **Cloud Storage -> Browser** and click in "**Create bucket**".



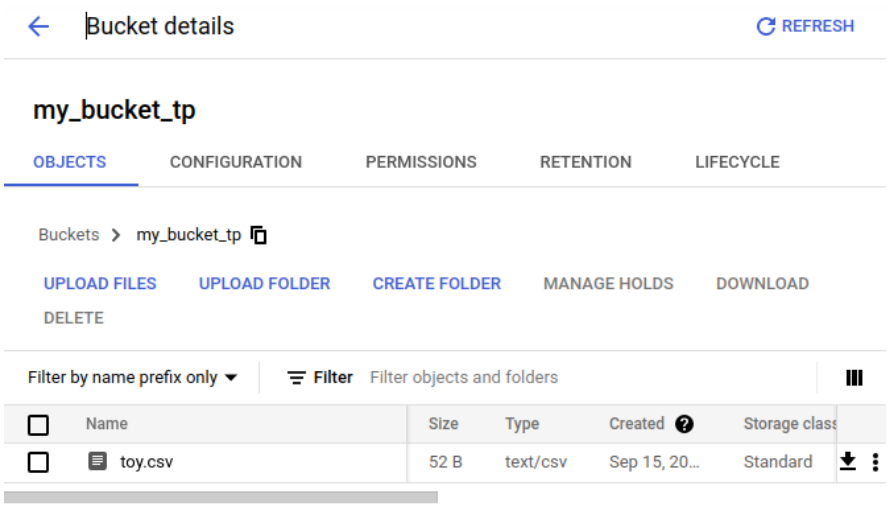
Then select a name for your bucket:



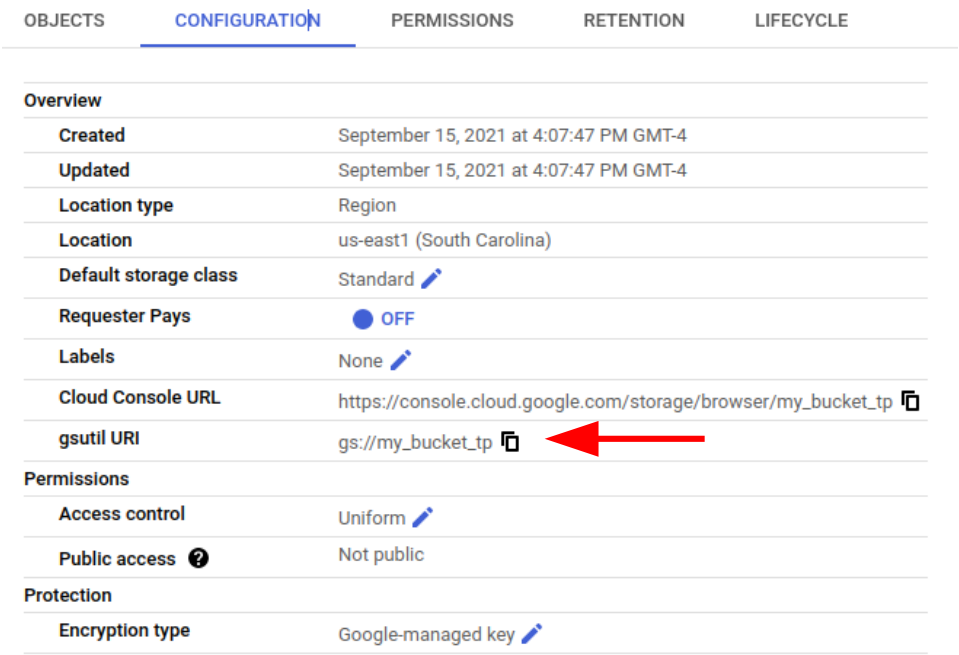
Under the "Choose where to store your data", select "**Region"** and search for "**us-east1**". Also set the access to objects as "**uniform"** and press "**Create**".



You will be redirected to your bucket page where you can upload your files. As an example, upload the **toy.csv** file to your bucket.

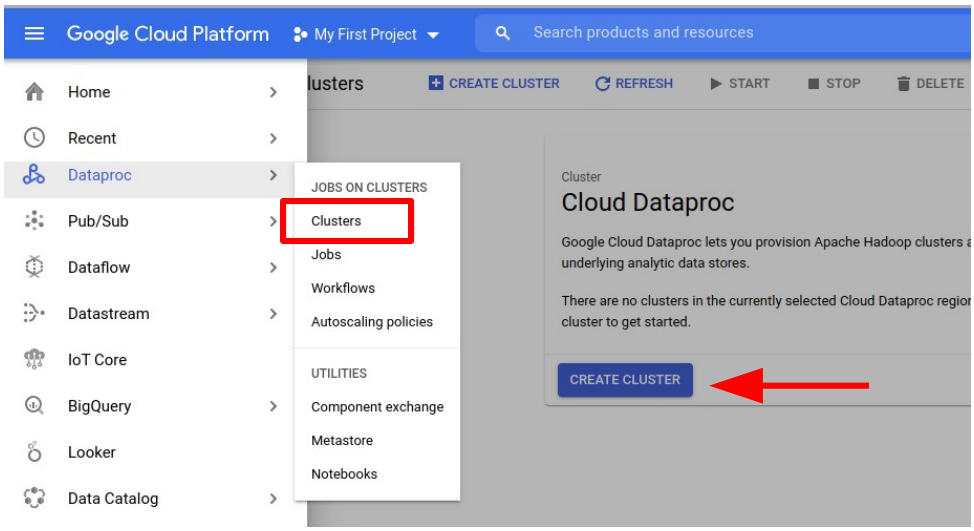


If you go the **CONFIGURATION** tab, the **gsutil URL** gives you the address for your bucket. For example, the path to access **toy.csv** file in this sample bucket would be: "**gs://my_bucket_tp/toy.csv**".

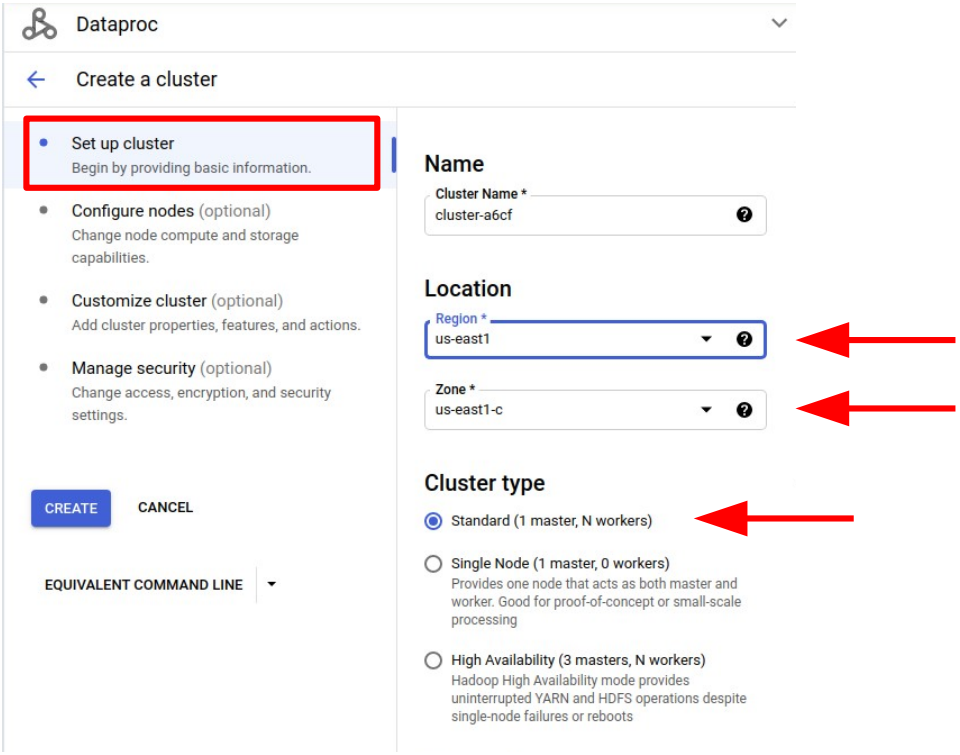


5. Creating a computing cluster

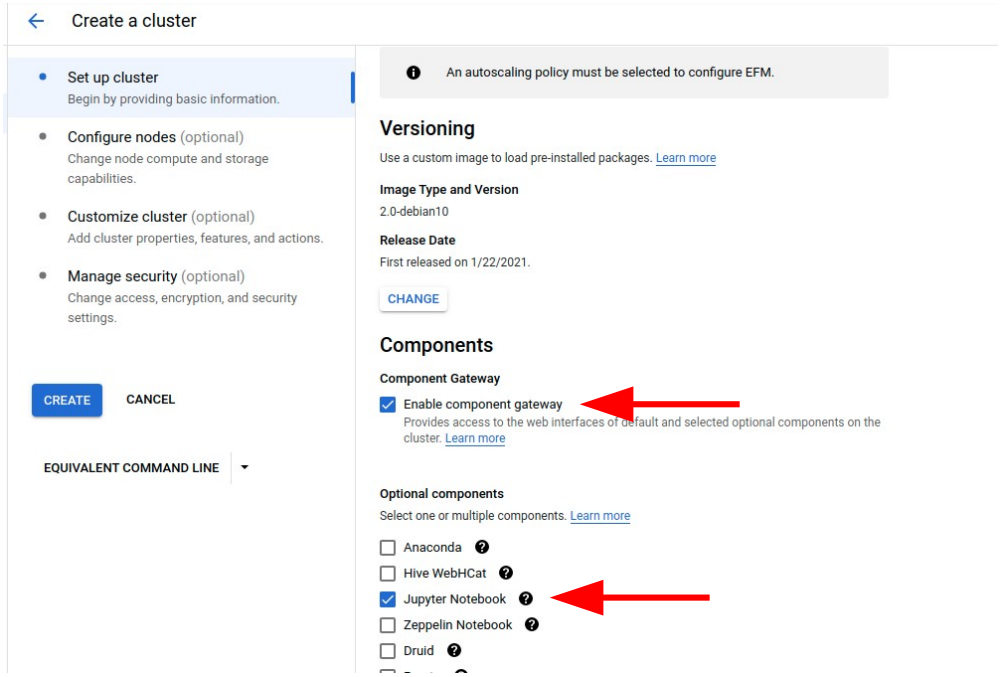
Now everything is set for creating our cluster. Go again to the **Dataproc -> Clusters** and press **Create Cluster**.



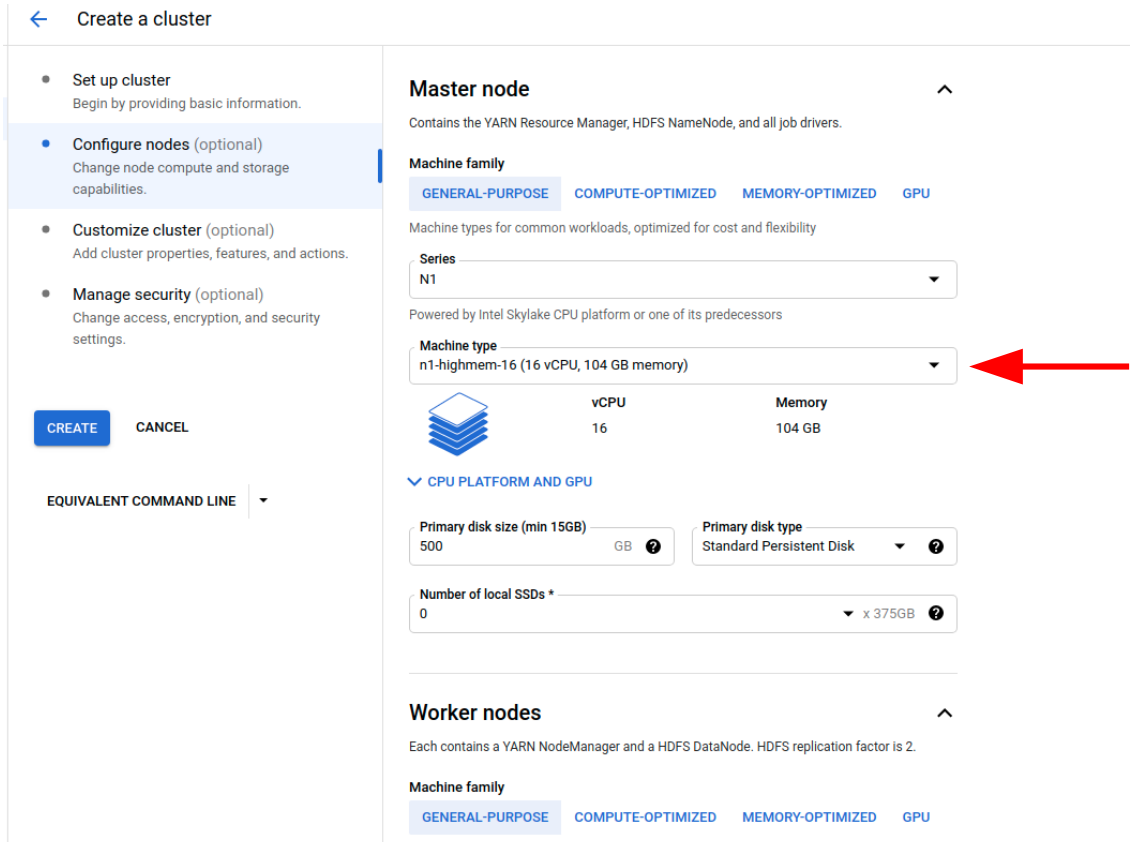
You don't need to change the name of the cluster, but it is necessary to specify the **Region.** Select **us-east1** (or the region for which you requested a quota increase). Set the **Cluster type** to **Standard (1 master, N workers)**.



VERY IMPORTANT: There is still a crucial step in the cluster configuration to be done. Select "**Enable component gateway"** and "**Jupyter Notebook"** options:

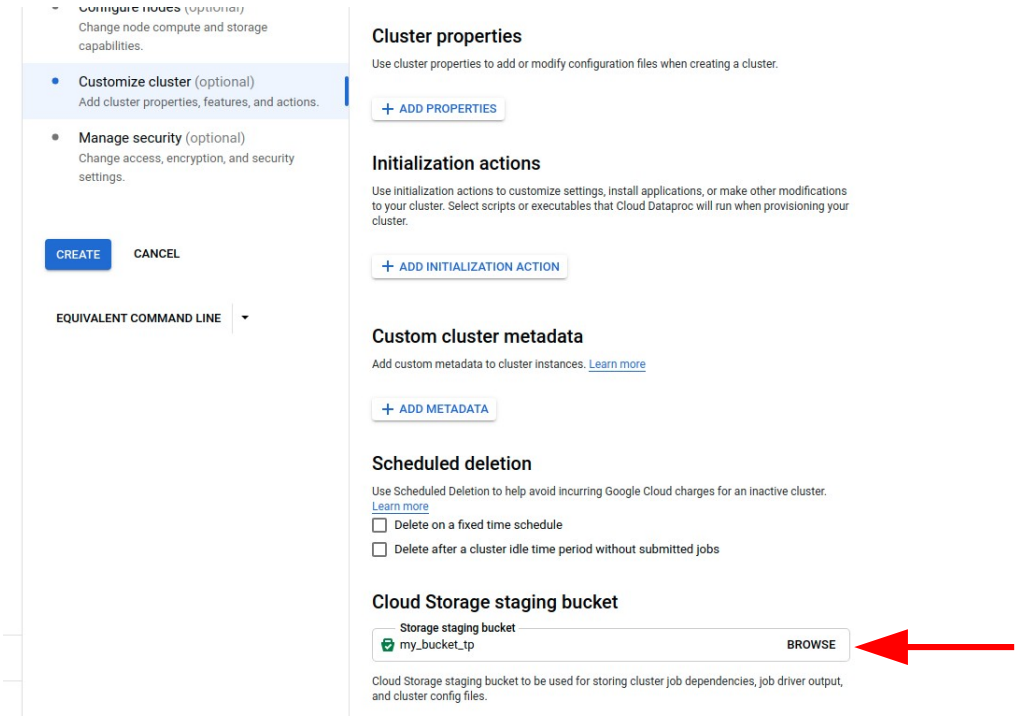


Now we have to configure nodes in the the cluster. The Cluster has 1 master and N workers. In our application the most valuable resource is memory. Thus, both for the master node as for the workers nodes we will use machines from the type **highmem**.



<u>Suggestion</u>: make a cluster of a master and 7 workers and select the type of n1-highmem-16 for all nodes. 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. 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.

Continue and look for "**Cloud Storage staging bucket"** in "**Customize cluster"** and browse your bucket:

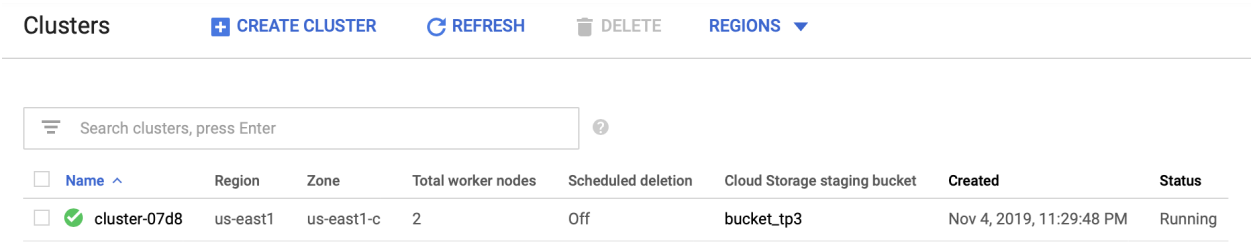


Warning: as you finish the configuration of your cluster and press create, GCP will start charging your billing account. Always remember to delete the cluster once you have finished your experiment.

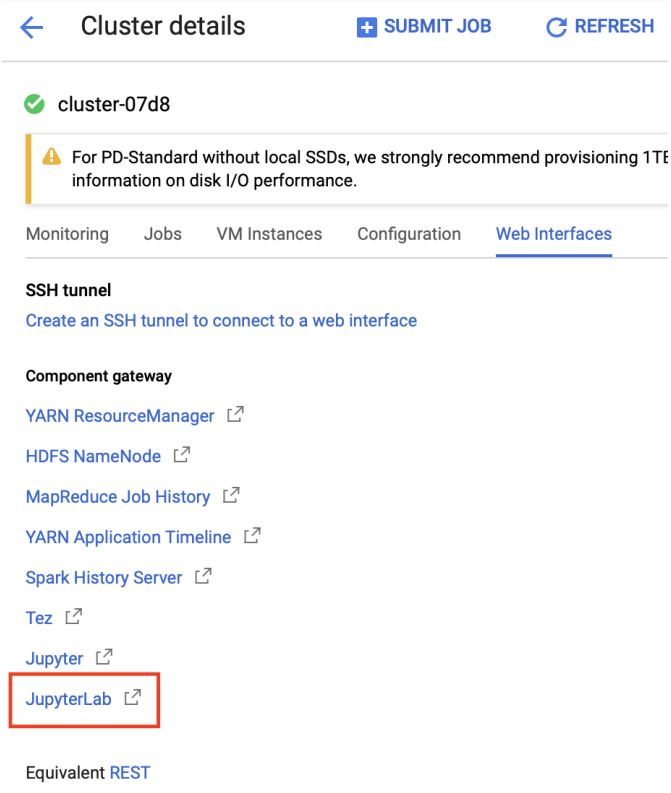
Finally, press **Create** to create the cluster**.** It may take a few minutes until the cluster is created and ready to be used.

6. Using your cluster

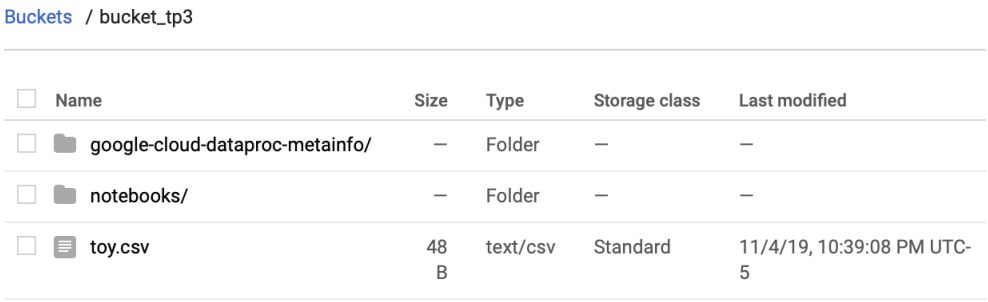
Once your cluster is created, click to open it.



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

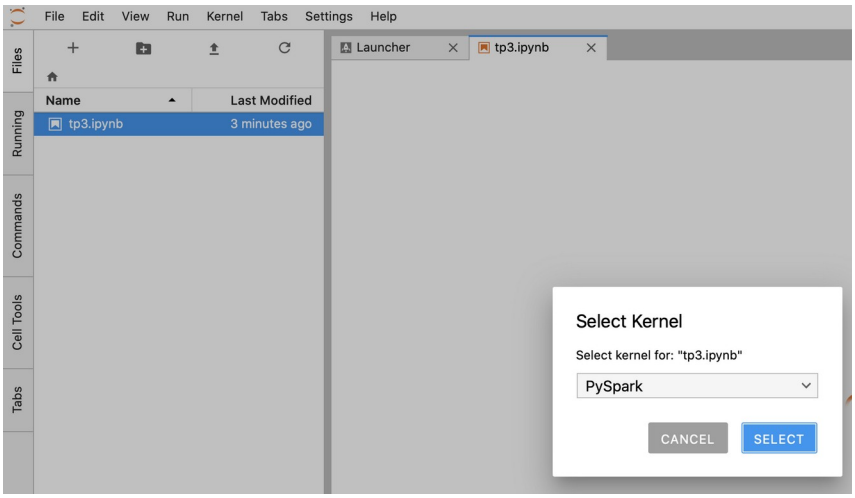


Now, go again to **Storage -> Browser** and open your bucket. We will see a notebooks folder.

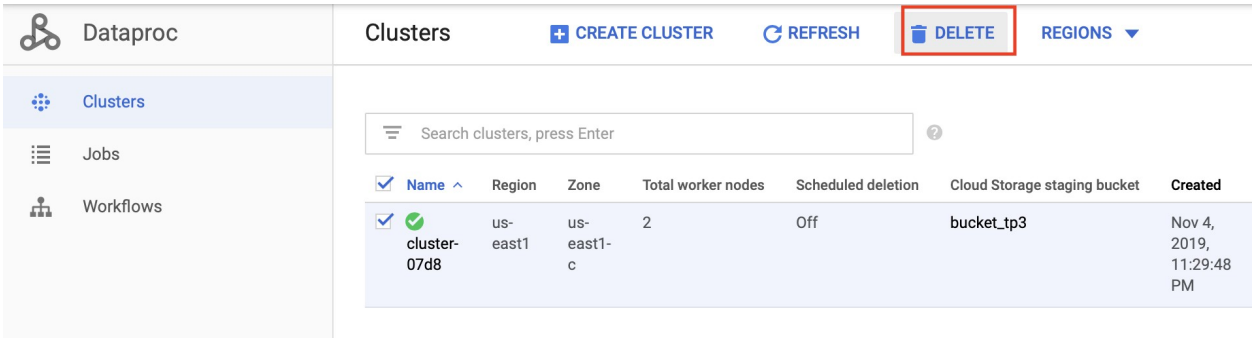


Go to the **notebooks/jupyter** folder and upload your .ipynb file.

The page that was open when you clicked in JupyterLab now should show 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.**