Google Cloud Platform

Fall 2020 Setup Guide [For Q4]

Getting Started

A video tutorial has been created to help walk through the steps of GCP setup.

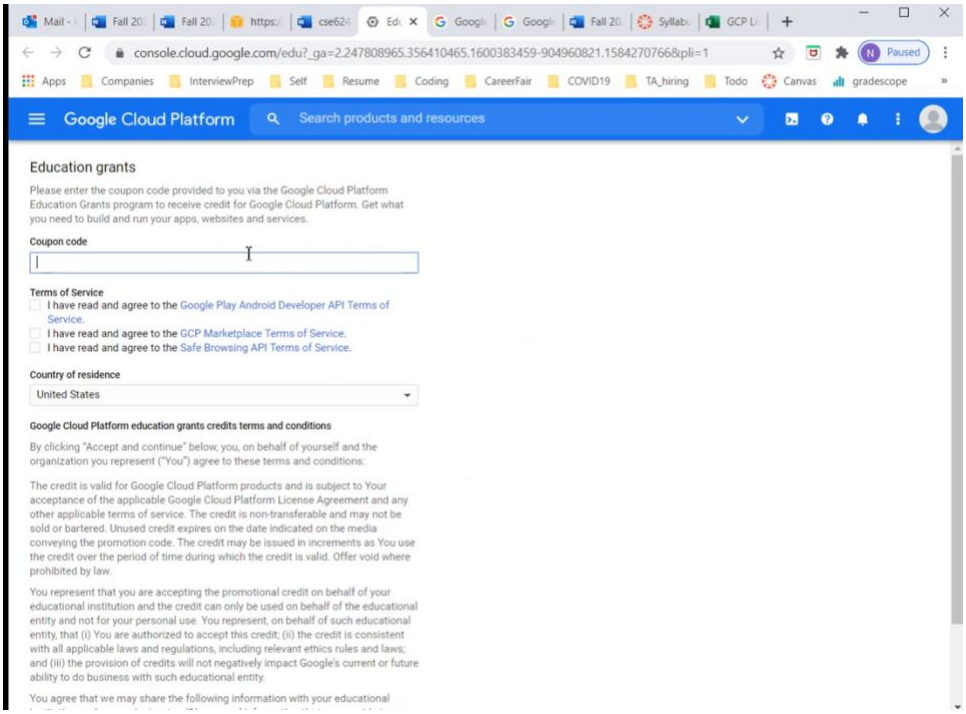
You can watch the video for:

- **GCP Storage setup here**
- **Dataproc Cluster here**

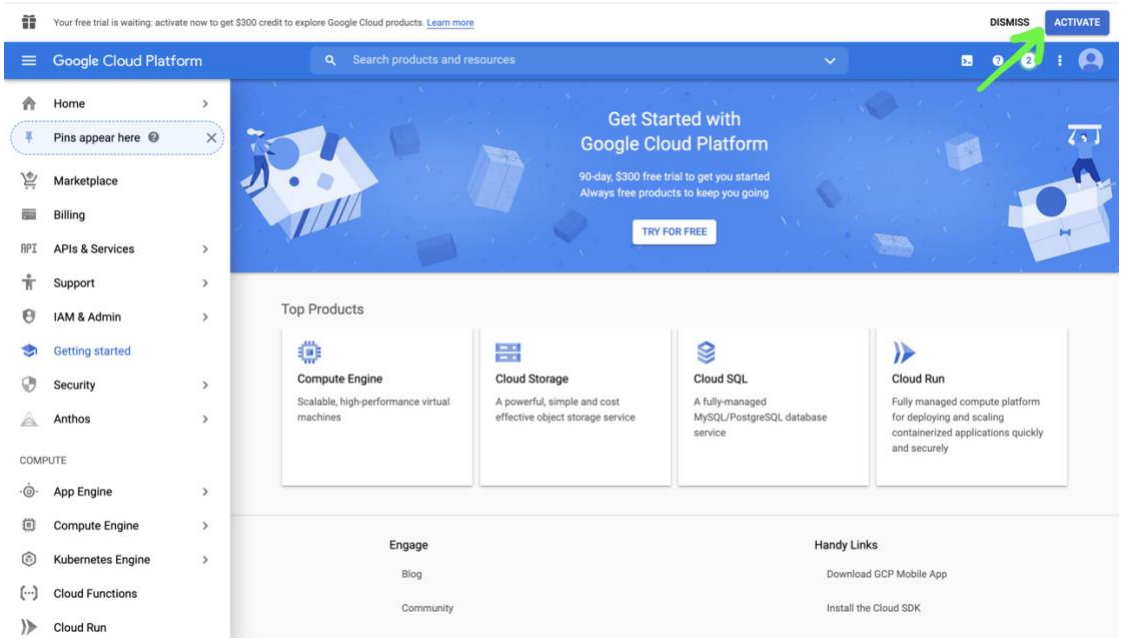
Instructions to setup GCP Storage and Dataproc Cluster are provided below:

Please make sure you have a Google account and if you don't have one, you would need to create a new account.

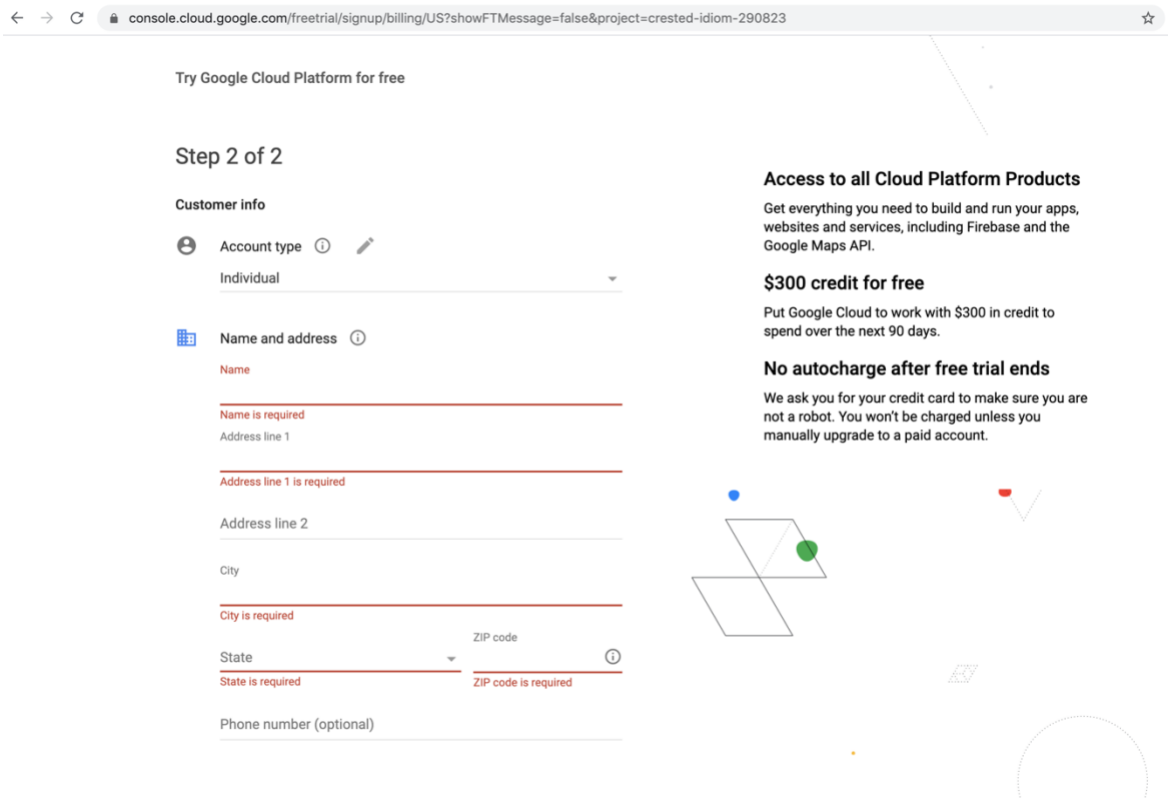
1. Login into your account using this link
2. Enter your free educational credits using this link.

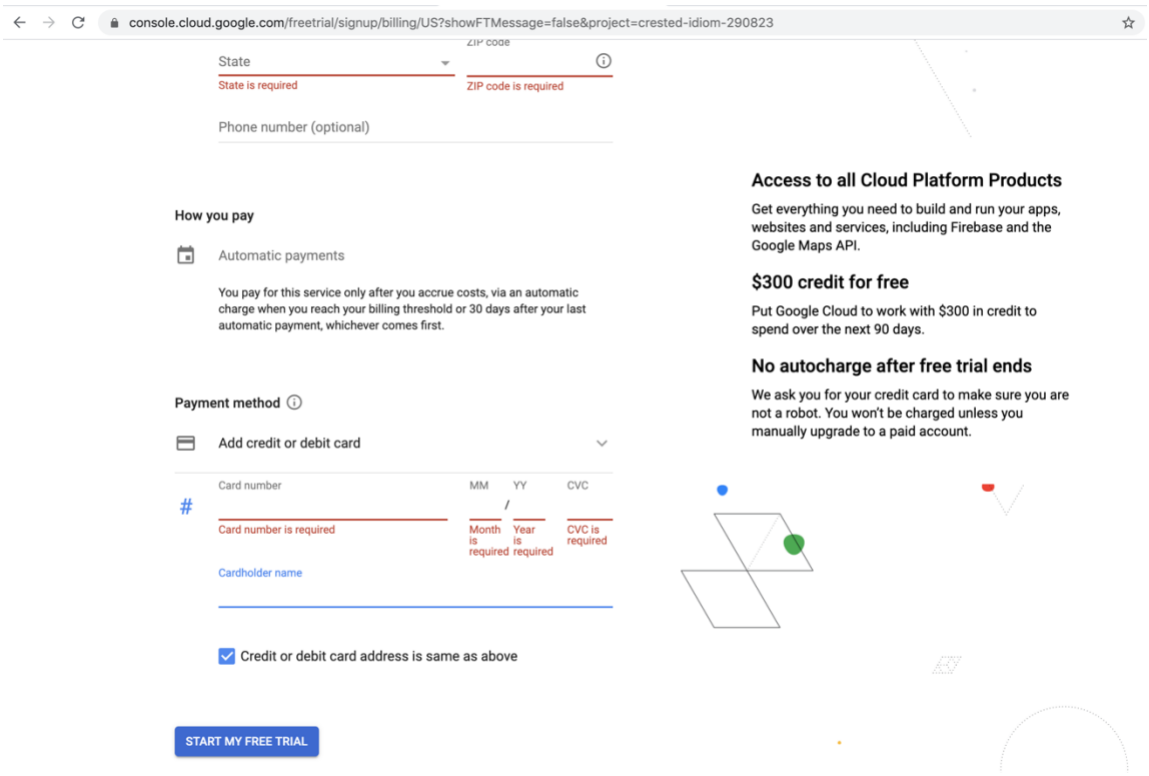


3. When you enter the educational credits, Google automatically creates a new billing account named "Data and Visual Analytics (DVA) pt1" and activate the "free trial option" at the top right corner of the console.



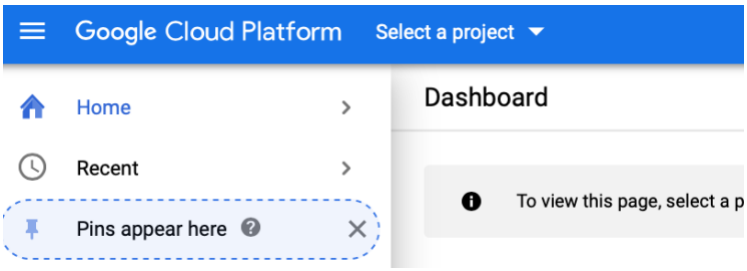
4. You will need to enter default billing credit card information if you are using GCP for first time. This is to confirm it is a human and you may need to give in the credit card details for billing (money would not be deducted from your card). Please make sure to use the billing account named “Data and Visual Analytics (DVA) pt1” is linked to a new project which will be used to create your storage and clusters (details in step 7). Don’t use your own personal billing account otherwise the free educational credits will not be applied.



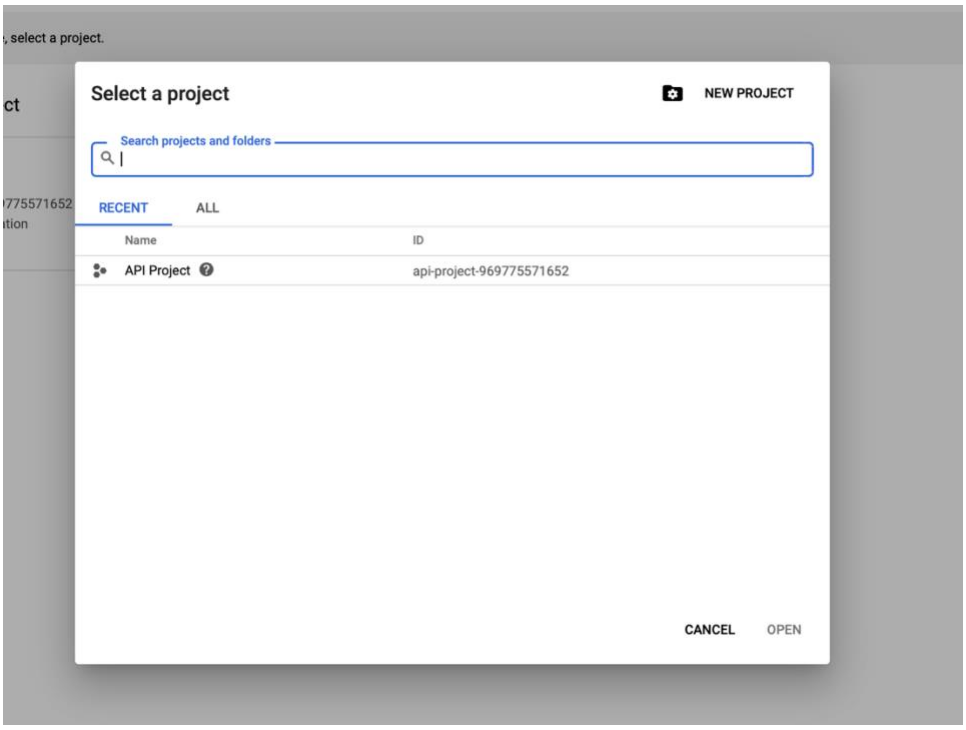


5. Go to [Google Console Home]

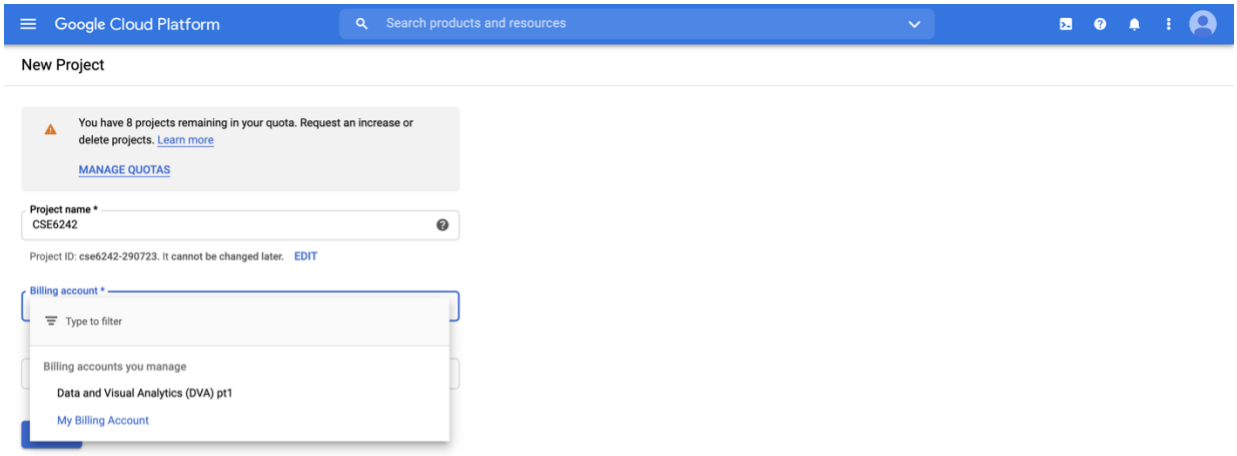
6. On the top left corner, click on dropdown to "**select a project" (or if you already have existing projects, it will list the latest project. Click on the drop down in this case as well)**

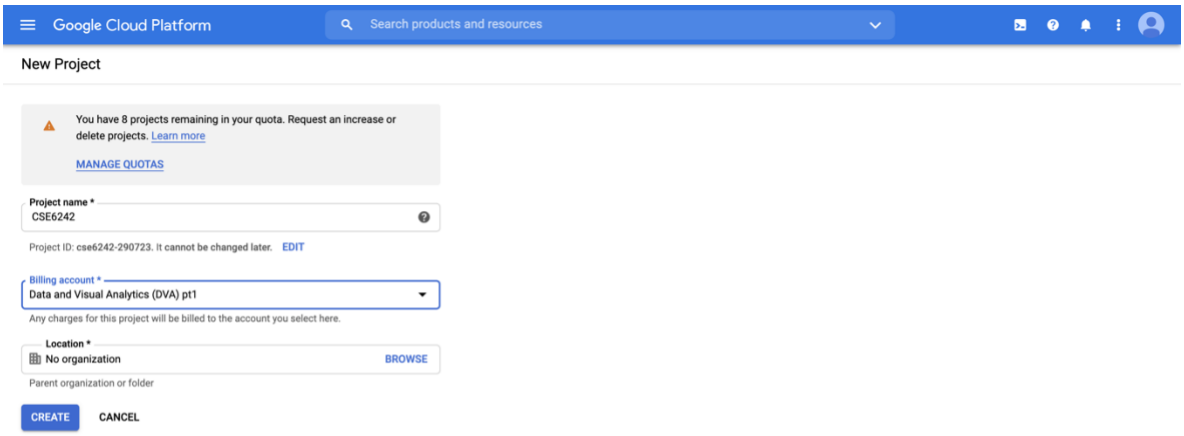


7. Click on New Project on the top right corner

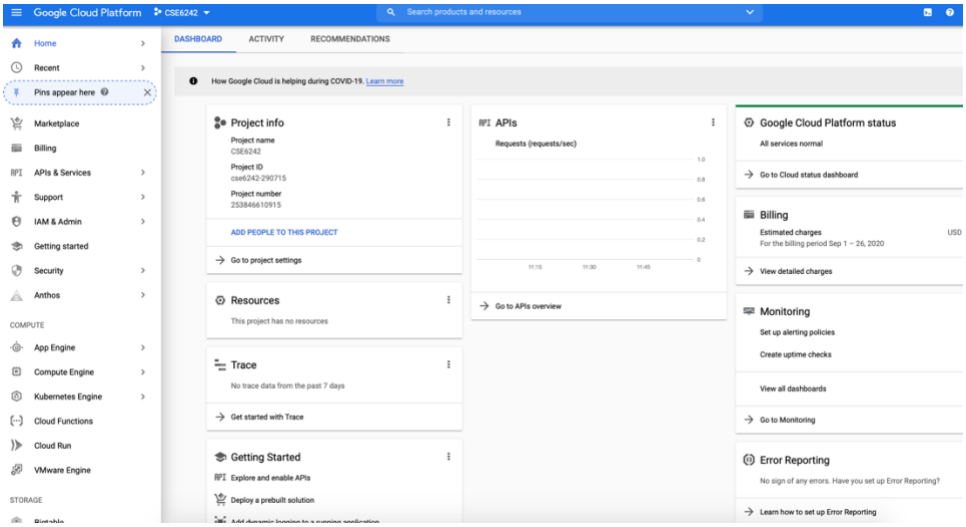


8. Enter CSE6242 (or any other default name) and create a new project





9. Wait till the new project is created



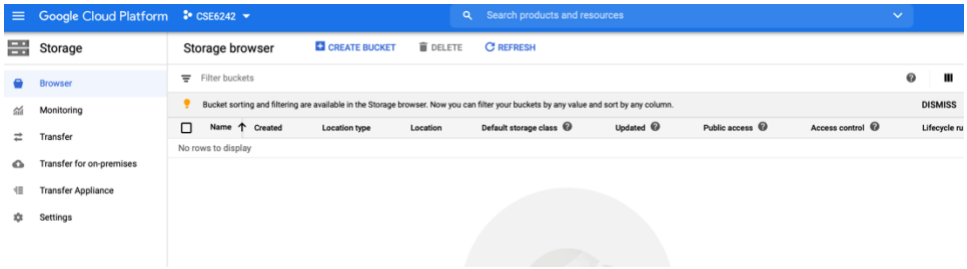
Note: The new Project CSE6242 should be linked to the Billing Account "Data and Visual Analytics (DVA) pt1". This step can be achieved while creating a new project in step 7 as listed above. If you have only one billing account (namely Billing Account "Data and Visual Analytics (DVA) pt1"), it will automatically link your project to Billing Account "Data and Visual Analytics (DVA) pt1".

Uploading data files to Spark compatible Google Cloud storage

We have listed the main steps from the documentation for uploading data file to your Google Cloud Platform:

Note: The data can be found on this dropbox link: yellow_tripdata_2019-01.csv

1. Go to Google Storage Home Page
2. On the top left corner, make sure the project is CSE6242



3. Click “CREATE BUCKET” to create a bucket.
    a. Use your GT Username as bucket name
    b. Choose Location Type as “Region - Low Latency with Single Region” (to minimize cost) and choose the closest location. For example, us-east4 (Northern Virginia) for students in east coast.
    c. Choose “Standard” storage class.
    d. Choose “Access Control” as “Fine Grained”
    e. Other settings - leave them as default.

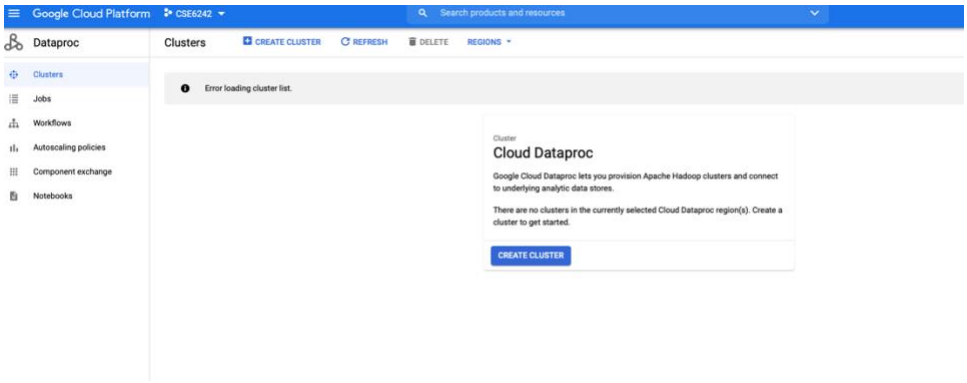
Once the bucket is created, click “UPLOAD FILES” and select ***yellow_tripdata_2019-01.csv*** from your local computer to upload to your bucket. The uploading process may take some time. You will see upload complete message and you will be able to view the yellow_tripdata_2019-01.csv file in the bucket.

Creating Clusters in Google DataProc (With Spark and Jupyter Notebook Components)

Google DataProc can be used to create an Apache Spark distribution cluster. This means that it handles large amounts of data on demand. The next step is to use Google’s DataProc web-based management tool to create a Linux cluster. Follow the recommended steps shown to create a new cluster (or see full documentation here on GCP site.

Follow the steps in the below link to create the cluster:

1. Go to Google Data Proc home page
2. On the top left corner, make sure the project is CSE6242



3. Click CREATE CLUSTER and enable API
4. Choose your GT Username as the cluster name (or you can use the default name)
5. Choose the closest region and location. For example, us-east4-a or us-east4-b or us-east4-c for students in east coast
6. If Nodes field defaults to 0; please change it to 2
7. Check the Component gateway checkbox (Enable access to the web interfaces of default and selected optional components on the cluster.)
8. You can use the provided defaults for the other options
9. Click on “Advanced Options”
10. If Nodes field defaults to 0; please change it to 2
11. In the “Cloud Storage staging bucket” field - Browse the name of the bucket you created in prior steps (only specify the name of the bucket which would be your GT Username). Your notebooks will be stored in Cloud Storage under gs://bucket-name/notebooks/jupyter
12. Under “Optional components”, click on “Select Component” button and select the "Anaconda" and "Jupyter Notebook" components
13. You can use the provided defaults for the other options
14. Click Create

It will take a few minutes for your Cluster to be up and running. Once the cluster is up and running:

1. Click on your Cluster Name to navigate into the cluster details screen
2. Click the Web Interfaces tab to display a list of Component Gateway links to the web interfaces of default and optional components installed on the cluster
3. Click the Jupyter link. The Jupyter notebook web UI opens in your local browser.
4. Upload HW3 Q4 skeleton notebook into your Jupter notebook web UI (under GCS folder and not under local folder)
5. Open the notebook, set the kernel to PySpark and you will be ready to start working on the Q4 section

Delete Storage and Clusters

After you complete the Q4 section of this homework, please remember to delete both storage and clusters. The deletion process is easy.

1. On the Storage home page, check bucket you want to delete and then click “DELETE.”
2. Wait till the bucket is deleted without any errors.
3. On the Cluster page, check cluster you want to delete and then click “DELETE.”
4. Wait till the cluster is deleted without any errors.