Spark-Lab 2

Basic instructions

This lab is to be done by each group, that means from Group A to Group H, so we will have 7-8 Members on each group. Due date would be 2nd of March 10 am. Any coursework submitted after this time will be discarded.

How to submit?
Generate a complete new email to me with a subject saying: GroupX-Spark-Lab2.pdf and the file attached with the same name format. The email must be sent by one member of the team and must content all the members of the team on cc. The names must be also displayed in the document.

Lab

In this lab, we will work on the idea of ephemeral clusters and how decoupling of storage and compute allows people to collaborate in big data projects without the costs needed to move big data sets from one store to another.

Another concept we will play with in this lab is the notebook concept. We will create two clusters with different notebooks attached to them.

1.- Create gs bucket and upload lab2Dataset.csv to your bucket.

lab2Dataset.csv can be downloaded from
https://storage.googleapis.com/mikele-ie-bucket/lab2Dataset.csv

You can upload the csv file either using command line tool gsutil or using the UI in the bucket window.

More info about creating buckets:

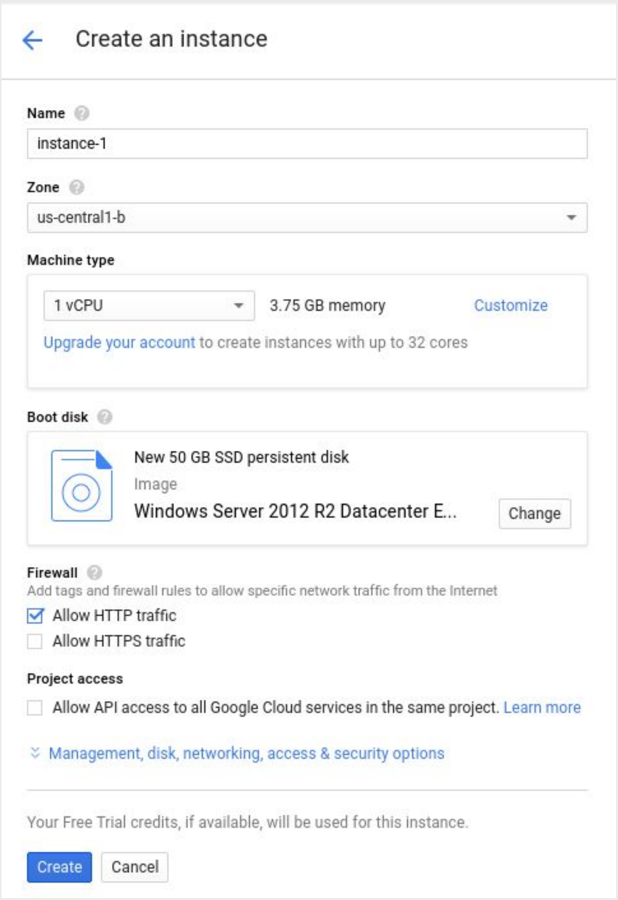
https://cloud.google.com/storage/docs/creating-buckets

2.- Create Windows vw as a bastion host or jump box. This machine will allow us to access notebooks through the browser.

1. In the GCP Console, go to the VM Instances page.

   GO TO THE VM INSTANCES PAGE

2. Click the **Create** button.
3. In the **Boot disk** section, click **Change** to begin configuring your boot disk.
4. In the **OS images** tab, choose **Windows Server 2012 R2**.
5. Click **Select**.
6. In the **Firewall** section, select **Allow HTTP traffic**.
7. Click the **Create** button to create the instance.



Create an instance

Name

instance-1

Zone

us-central1-b

Machine type

1 vCPU 3.75 GB memory Customize

Upgrade your account to create instances with up to 32 cores

Boot disk

New 50 GB SSD persistent disk

Image

Windows Server 2012 R2 Datacenter E...

Change

Firewall

Add tags and firewall rules to allow specific network traffic from the Internet

☑ Allow HTTP traffic

☐ Allow HTTPS traffic

Project access

☐ Allow API access to all Google Cloud services in the same project. Learn more

Management, disk, networking, access & security options

Your Free Trial credits, if available, will be used for this instance.

Create Cancel

1 vcpu would work but it will be slow, so choose 2 vcpus instead for better performance and double the memory. Choose the same zone where you are going to create your cluster.

Allow a short time for the instance to start up. Once ready, it will be listed on the VM Instances page with a green status icon.

Connect to your jump-box instance

1. Go to the VM Instances page in the Google Cloud Platform Console.

   GO TO THE VM INSTANCES PAGE

2. Under the **Name** column, click the name of your virtual machine instance.
3. Under the **Remote Access** section, click the **Set Windows Password** button.
4. Specify a username, then click **Set** to generate a new password for this Windows instance. Save the username and password so you can log into the instance.
5. Connect to your instance using RDP:
   - If you installed Chrome RDP by FusionLabs, click the RDP button under the **Remote Access** section.
   - If you're using a different RDP client (including Windows Remote Desktop Connection), click the RDP button's overflow menu and download the RDP file. Open the RDP file with your client. If you are using a Mac you can use Microsoft RDP client for Mac.

Due to security policies you will have to open the ports in the firewall before accessing the windows VM, in order to do this under cloudshell open RDP port:

```
gcloud compute firewall-rules create rdp1 --allow tcp:3389 --source-ranges 0.0.0.0/0
```

This will keep the connection open as long as we are logged in. if the connection is lost then we will need to open firewall port again. In production scenarios this is not needed as the network would have been set up to access the right networks and subnets.

Opening the port might last a few seconds, wait for a minute and then log onto the windows VM.

Install firefox within your windows vm by calling this url from the microsoft explorer:
https://storage.googleapis.com/mikele-download/Firefox%20Setup%2058.0.2.exe

If you experience some issues when calling the url, just add the url to the list of your trusted sites in Internet Explorer and try again.

Now we are going to create spark clusters from the command line. In the previous excercise we created them from the UI. The command line is just another option and usually the most popular.Let’s start with the first one. Open cloud shell and then follow the next steps.

Create spark cluster 1 with Jupyter Notebooks

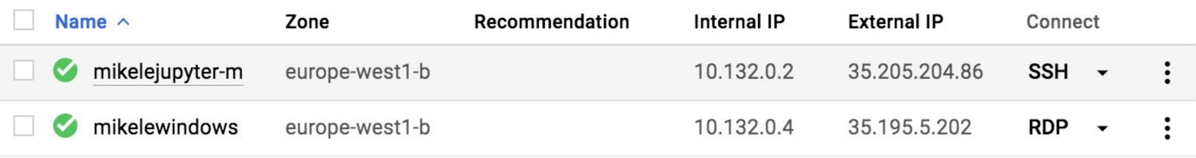
When executing the dataproc create command add --single-node as option, it should look like something like this:

```
gcloud dataproc clusters create myjupyterspark --project <myproject>
--bucket <mybucket> --single-node --master-machine-type n1-standard-1
--master-boot-disk-size 50 --initialization-actions
'gs://dataproc-initialization-actions/jupyter/jupyter.sh'
```

As you can see we are passing an initialization action to install jupyter in the master node. This is a transparent operation for the end user. Wait until the cluster is created.

Go back to your windows VM.

Check the internal ip of your cluster:

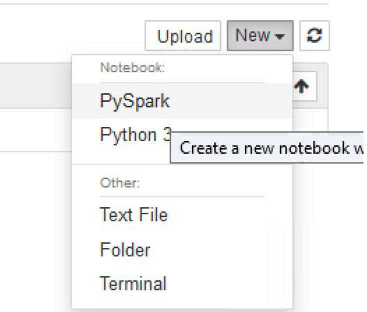


□	Name ^	Zone	Recommendation	Internal IP	External IP	Connect	
□	✓ mikelejupyter-m	europe-west1-b		10.132.0.2	35.205.204.86	SSH ▾	⋮
□	✓ mikelewindows	europe-west1-b		10.132.0.4	35.195.5.202	RDP ▾	⋮

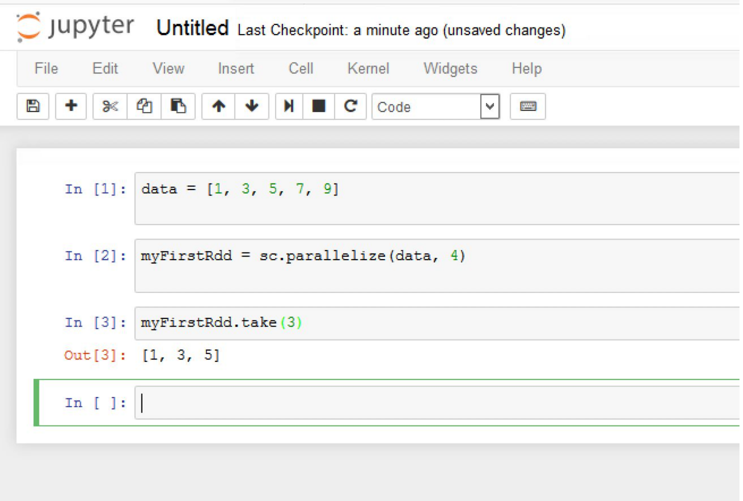
For example for mikelejupyter is 10.132.0.2.

Then, from within the windows machine open firefox and then you can call your jupyter notebooks with the internal IP and using port 8123.

Create a notebook buy clicking New and then PySpark in the right hand side of the screen:



Submit some basic RDD operations like you did in previous lab, for example:



```
In [1]: data = [1, 3, 5, 7, 9]

In [2]: myFirstRdd = sc.parallelize(data, 4)

In [3]: myFirstRdd.take(3)
Out[3]: [1, 3, 5]

In [ ]:
```

Remember to execute the different cells by pressing shift+enter

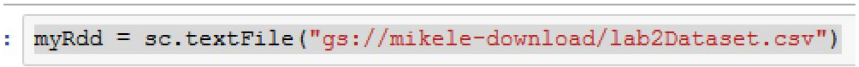
Then continue with the exercise.

In the recently created cluster 1:

Read file lab2Dataset.csv and create RDD. Remember that your file once in your bucket will have an absolute path like

gs://<yourbucket>/lab2Dataset.csv

For example from a bucket called mikele-download it would be something like:



```
myRdd = sc.textFile("gs://mikele-download/lab2Dataset.csv")
```

Let's code

Remove duplicates and write here how many duplicates were.

<ANSWER>
<Code here>
Add code for first cluster here

Number of duplicates:

Create new RDD with ids whose length is bigger than 3 .

Save RDD as text in gs bucket you created before. Choose a different file name.
</Code>
</ANSWER>

Delete cluster 1.

Create a new cluster, Cluster 2 with Datalab

Note:As we did with cluster 1 don't forget to add the --single-node option when using the create dataproc command as we did with the previous cluster. You must have deleted your cluster 1 before proceeding.

We are going to use a different notebook: datalab. The installation is very similar to what we did with jupyter.

In order to install datalab just use the datalab initialization action when creating the new cluster:

gs://dataproc-initialization-actions/datalab/datalab.sh
Datalab will be running on port 8080

Datalab is based on Jupyter so it will look very familiar.

Once the cluster is running
Read text file generated by cluster 1 and add it to rdd_main.
Get number of partitions and reduce the partitions to 2
Obtain the length of all the ids added together.

For example if ids are:
[alex, pepe, domin] then total would be 3 from alex plus 4 from pepe plus 5 from domin so 12

Sort alphabetically and display the first 10 ids (order from a to z)
Save the file to your bucket with the name sorted_rddids.txt

From rdd_main obtain the number of ids that share the first two characters.
For example if we have ids: aax , aat, aaron, bbt we would have aa,3 and bb,1

<ANSWER>
<code>
Add ALL your code here
</ code>
</ANSWER>