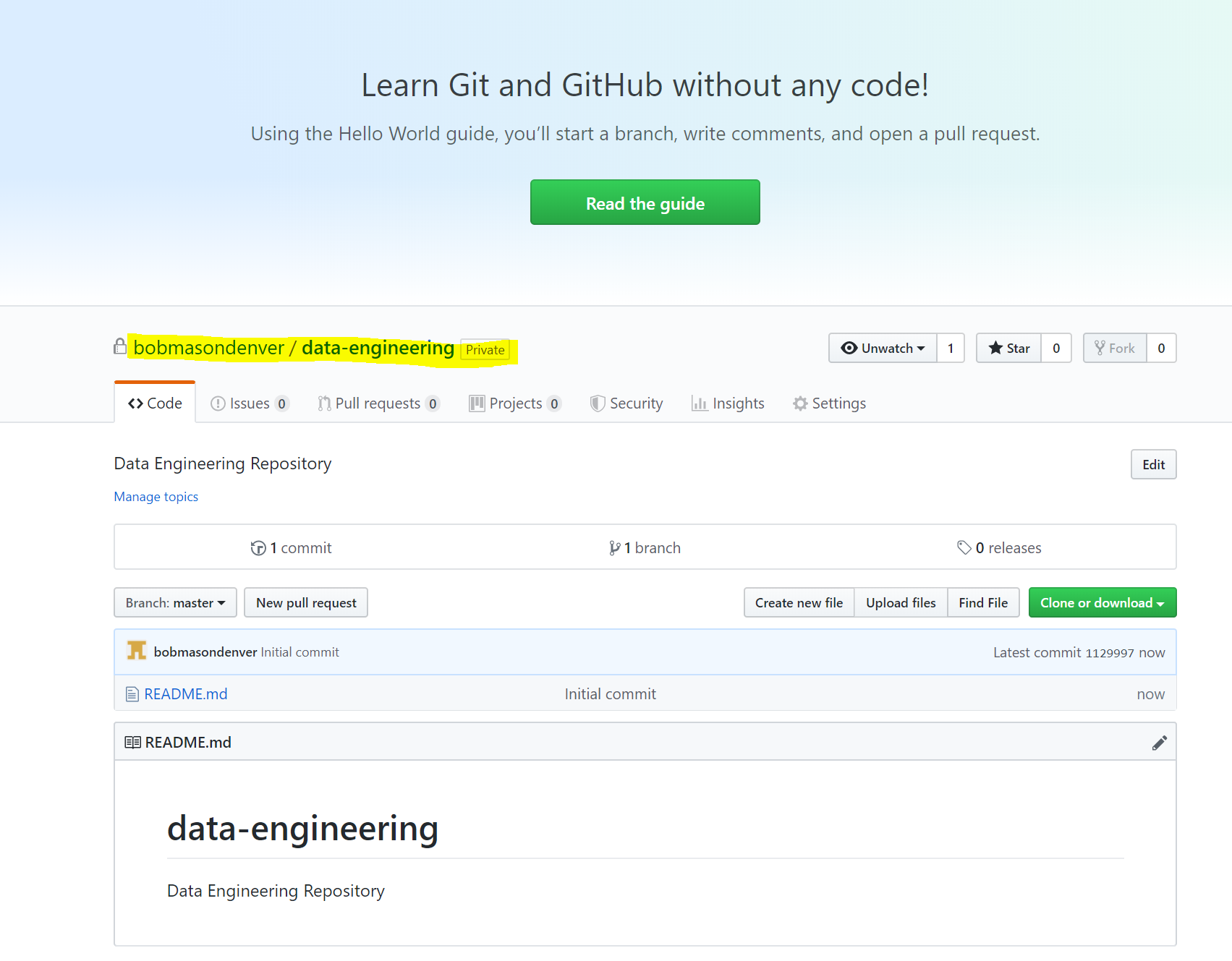
MSDS610 Data Engineering

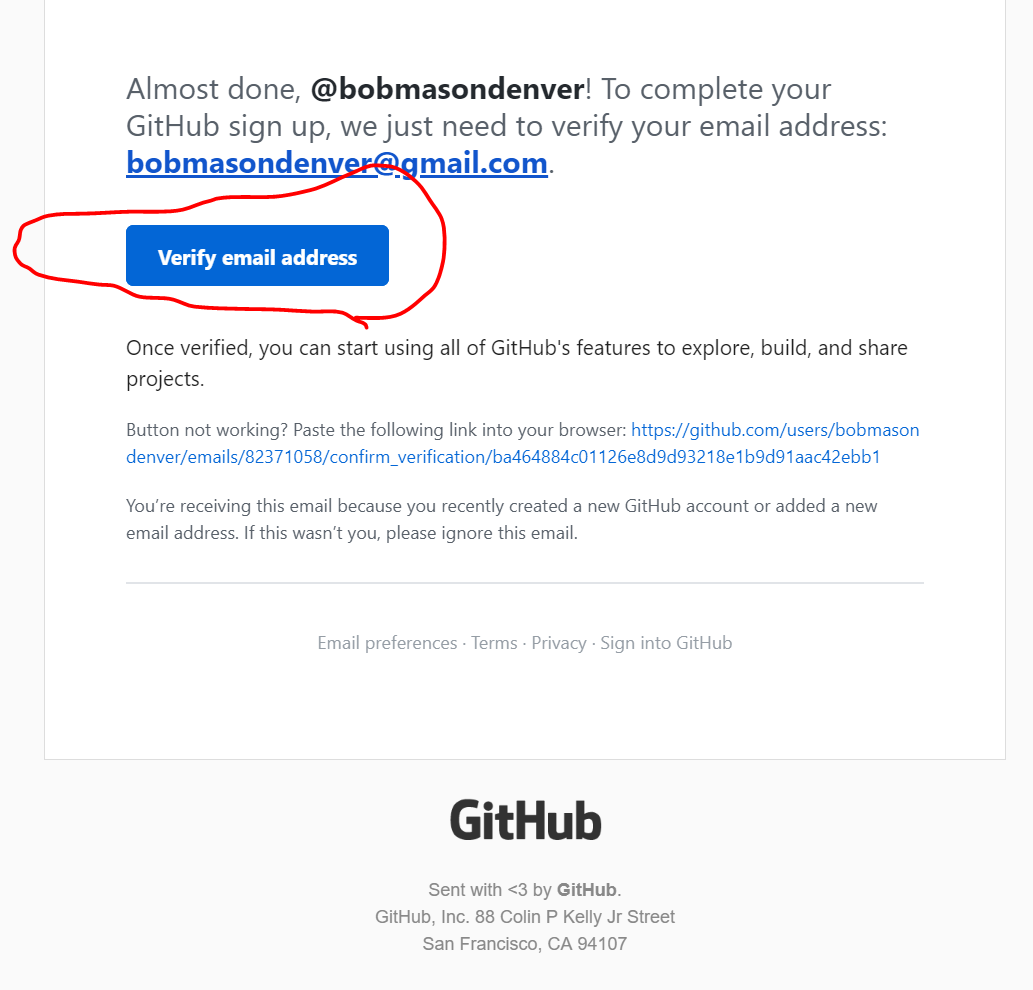
Week 2 Lab Supplemental – Part 1

as of 9/4/2019

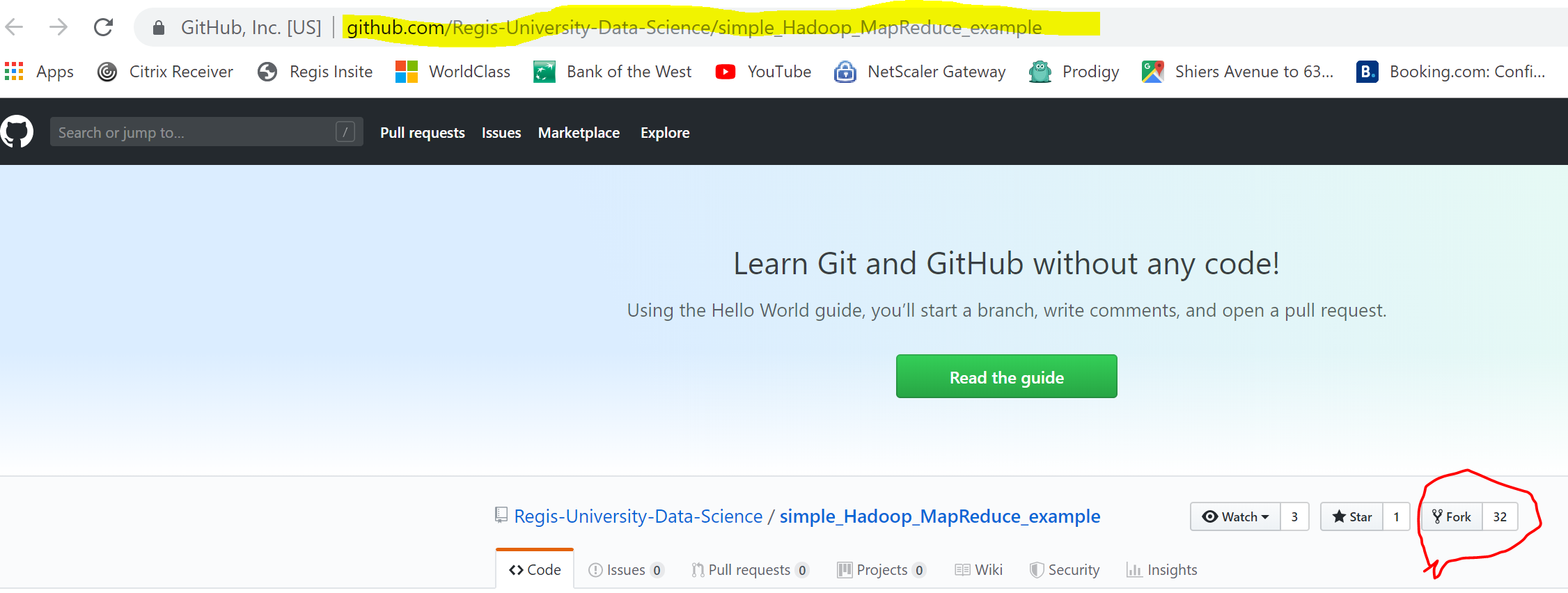
Go to github and create a new account. I created an account using the ID bobmasondenver, using the email address [bobmasondenver@gmail.com](mailto:bobmasondenver@gmail.com). I received an email message that asked me to verify my email address. I called the project data-engineering. I selected private as opposed to public, that way I can choose who has access to the files.



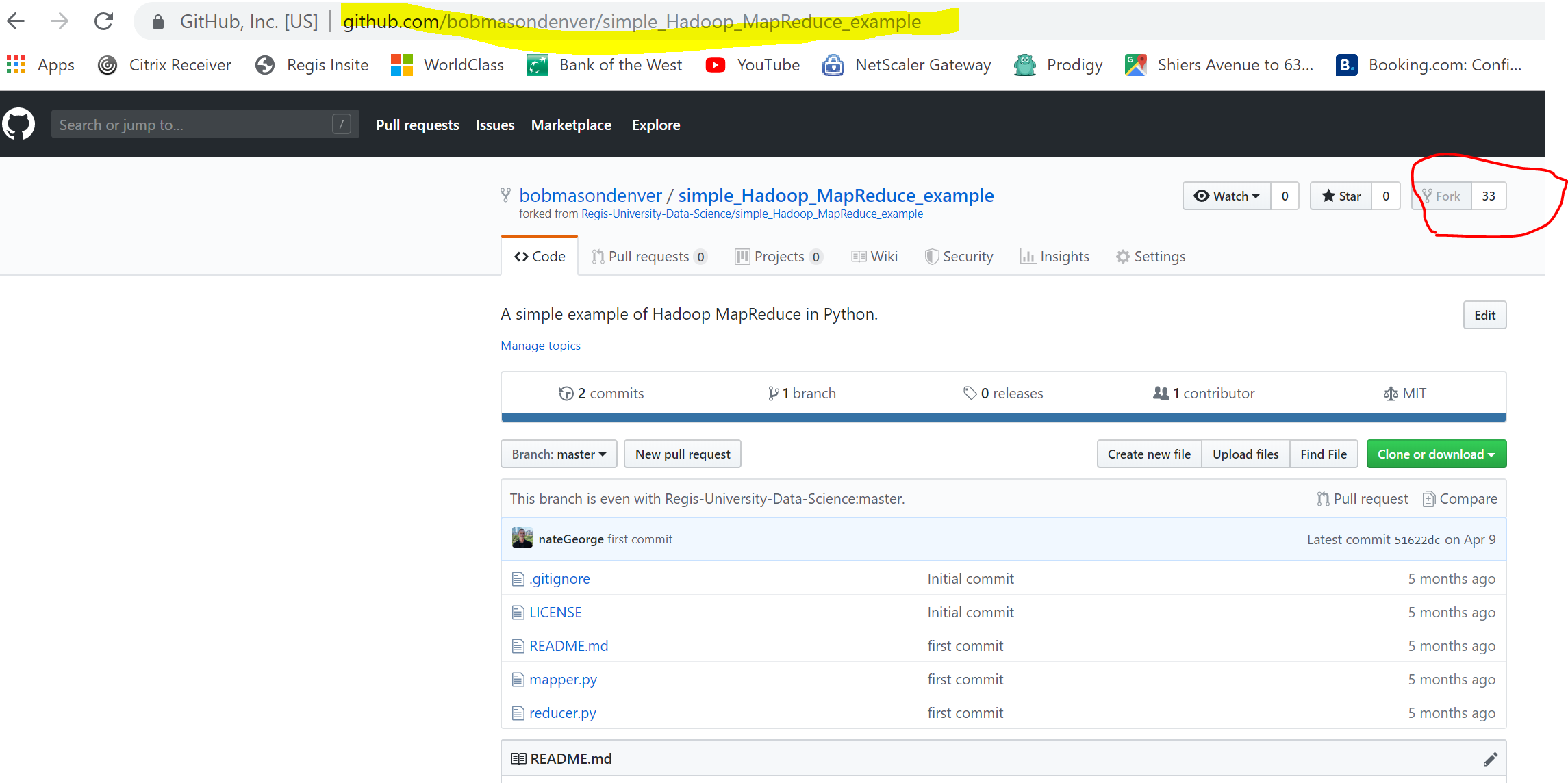
This is the email I received from Github to verify my email address. Click on the button below to verify your email address with Github.



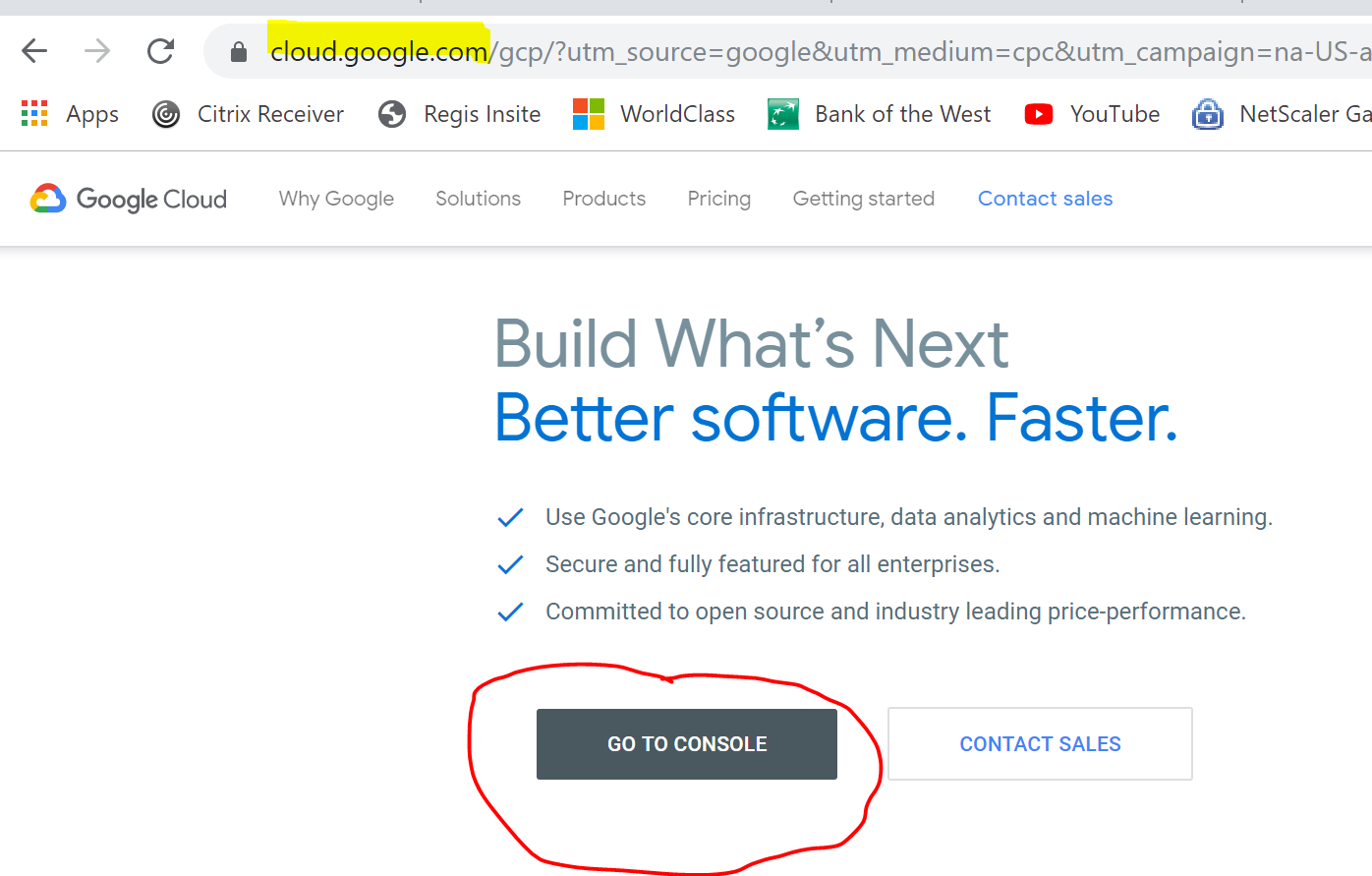
Click on the link from last week’s project and click on the Fork button.



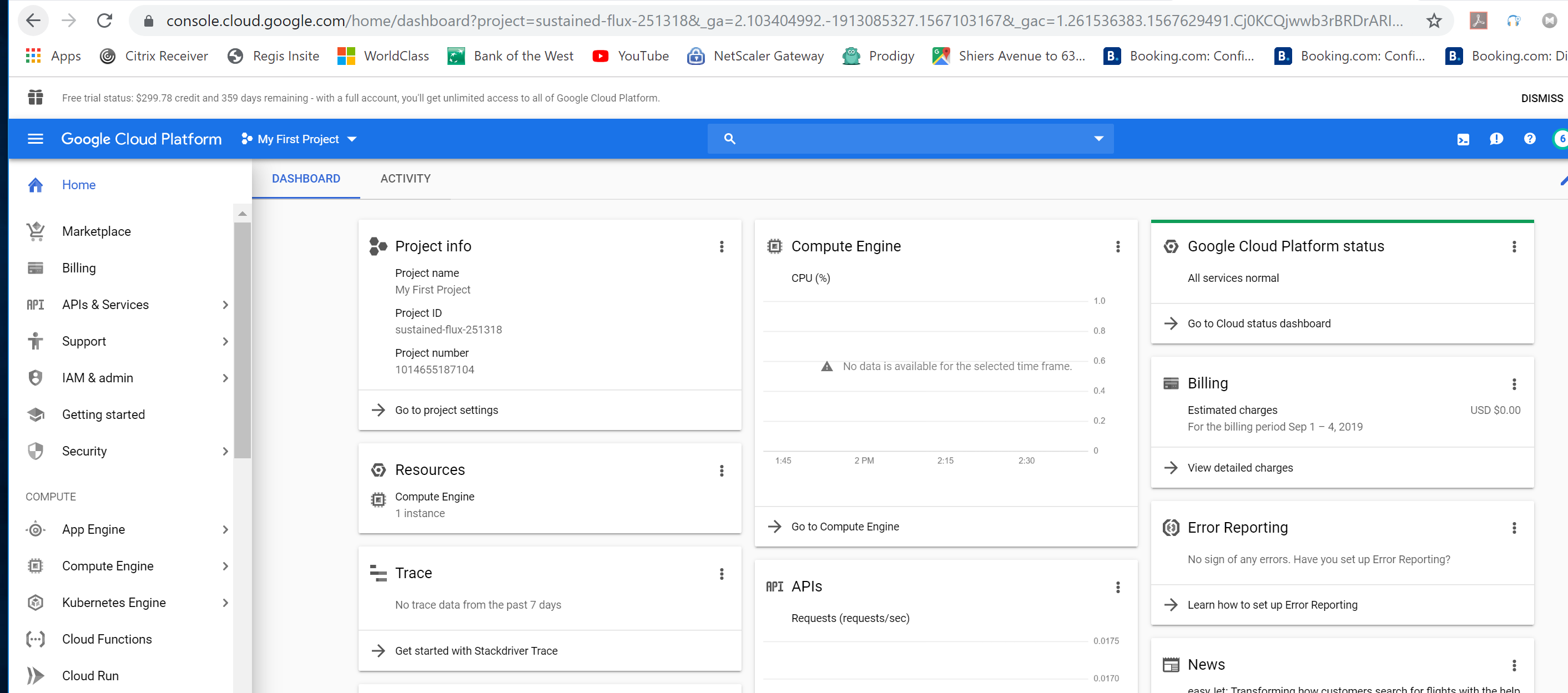
Notice that after I clicked the fork button (above), the project was copied to my new account bobmasondenver as shown below. The fork button is now gray.



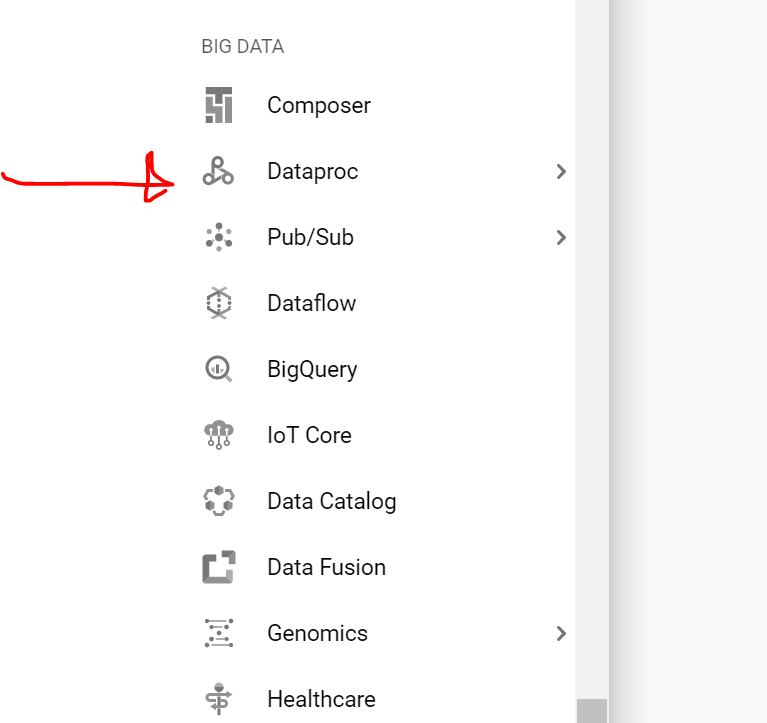
Go to our cloud.google.com and click on console. Login if you are not logged in already.



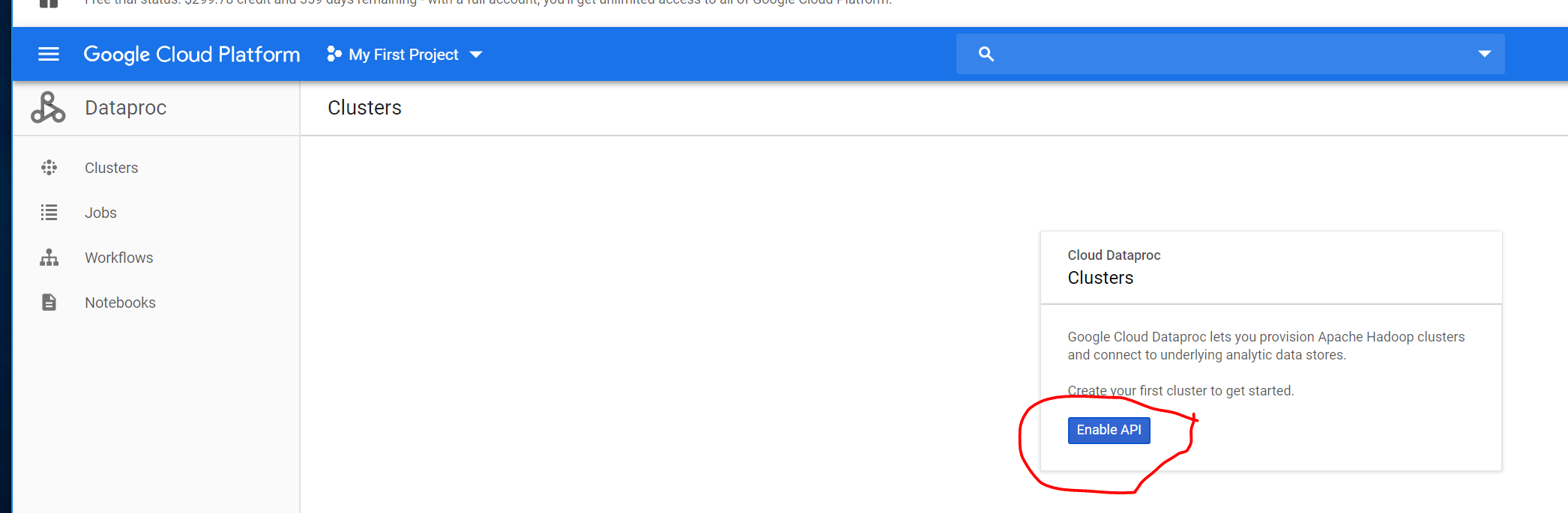
You may remember this console page from Lab 1.



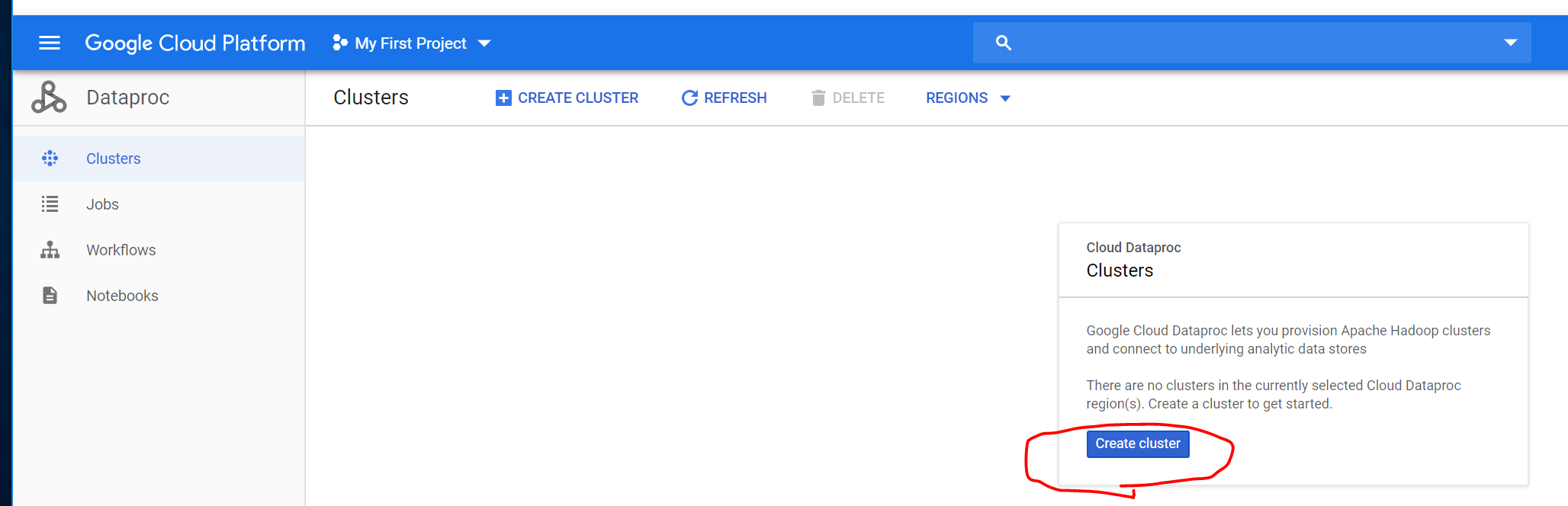
Scroll down on the left hand side of the window, Find and Click under Big Data, Data Proc, then Clusters



Click on the Enable API Button

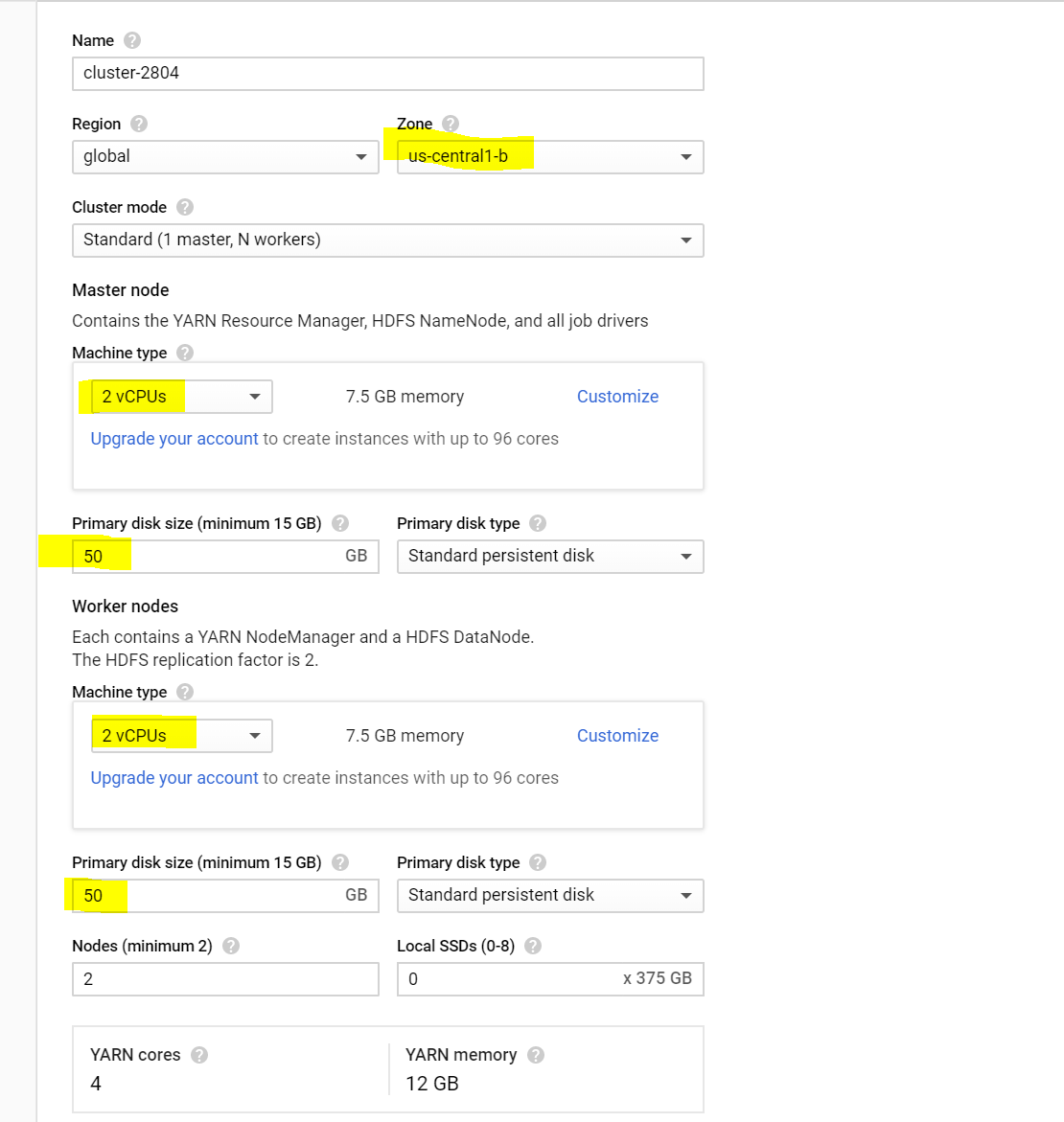


The option to create a Cluster will appear (as shown below). Click on the create cluster button.

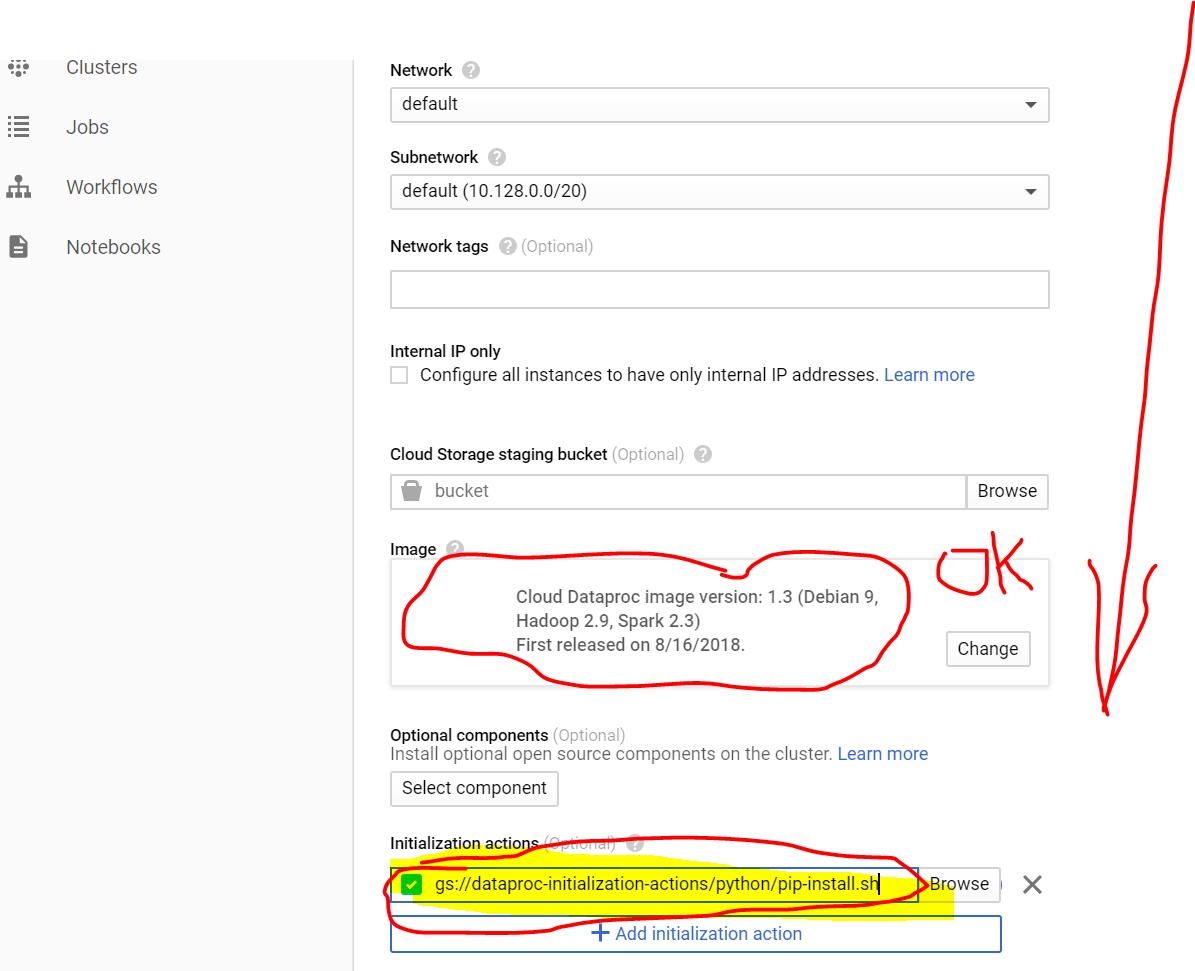


The create cluster window will appear (as shown below). We will leave many of the defaults. We will change the CPUs to 2 instead of 4 and change disk size (yellow) to 50 gig and change the zone to us-central1-b or other central zone to save on the cost.

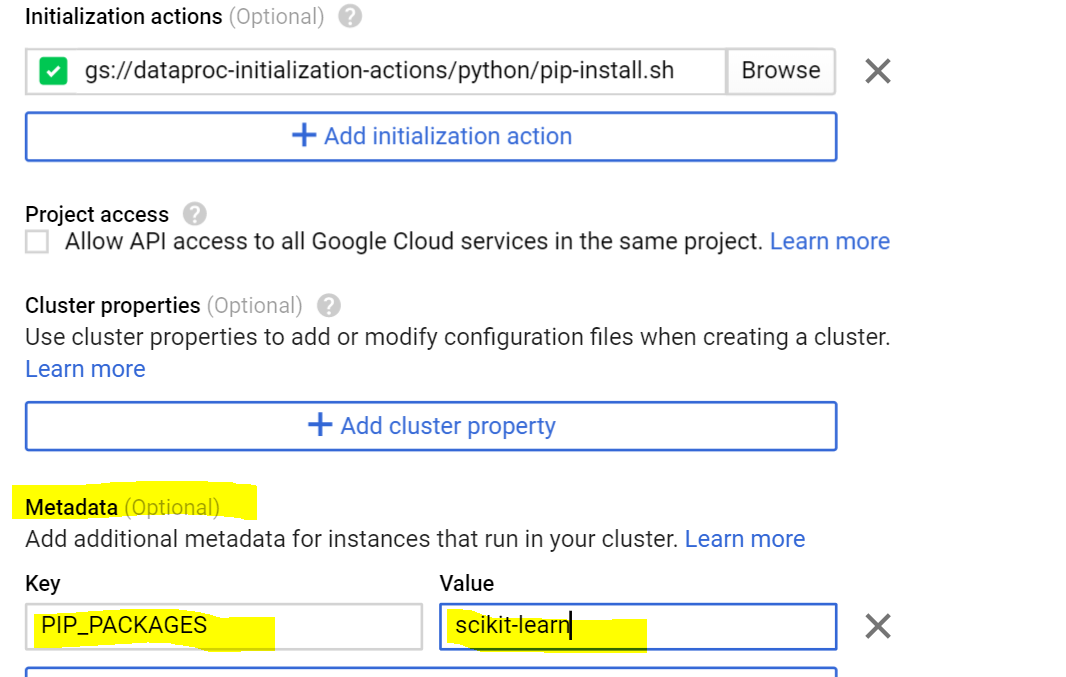
**Click on the advanced options to continue (DO NOT CLICK ON CREATE BUTTON YET)**



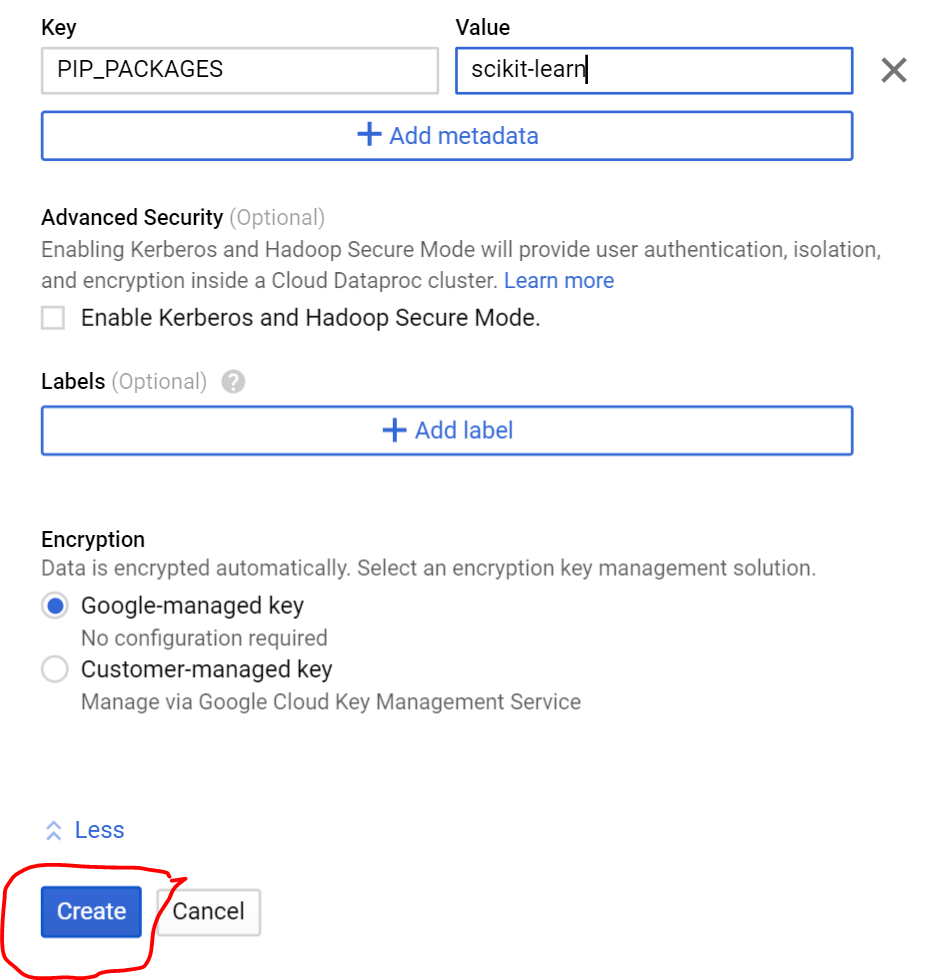
**Keep the defaults, notice the image, however don’t change it. Paste in the file name from the lab into the initialization action. After you move out of the box, you will see a green check mark.**



**Scroll down and then Add the metadata (as shown below in yellow)**

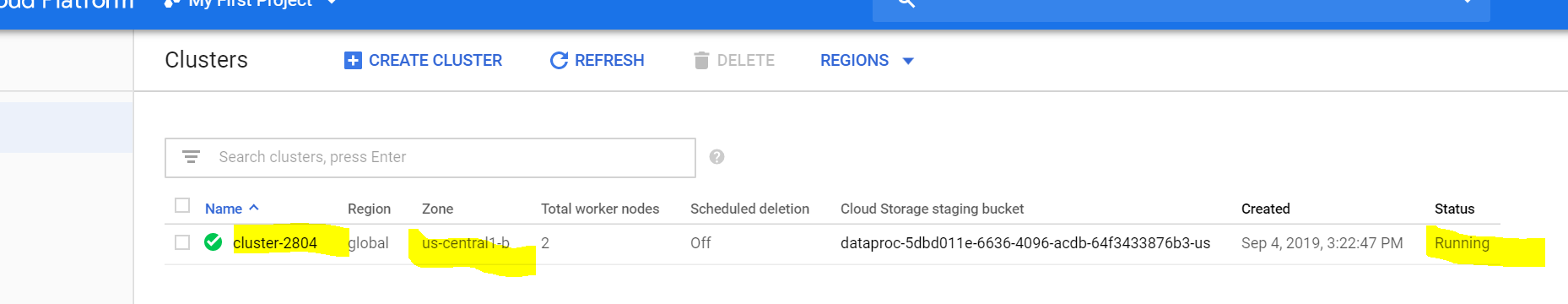


**Leave the remaining defaults, scroll to the bottom and click the create button.**

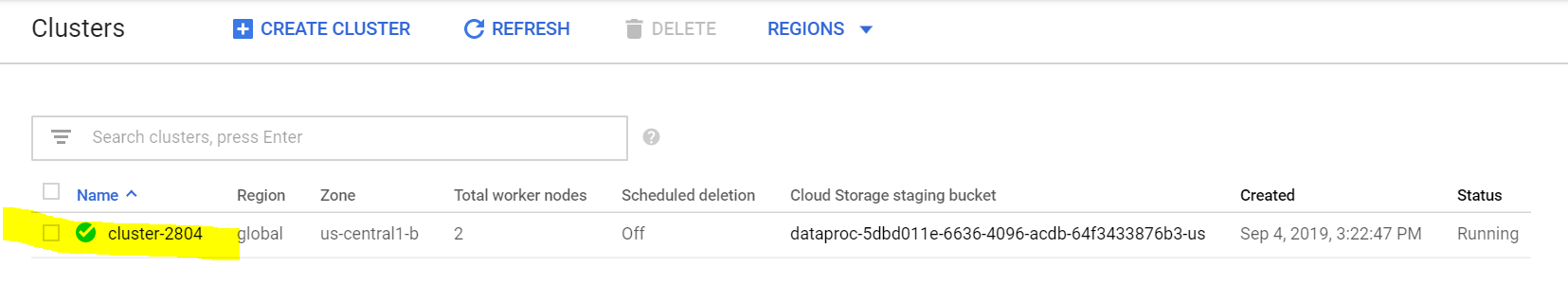


**The new cluster will be created and you will see that it is running.**

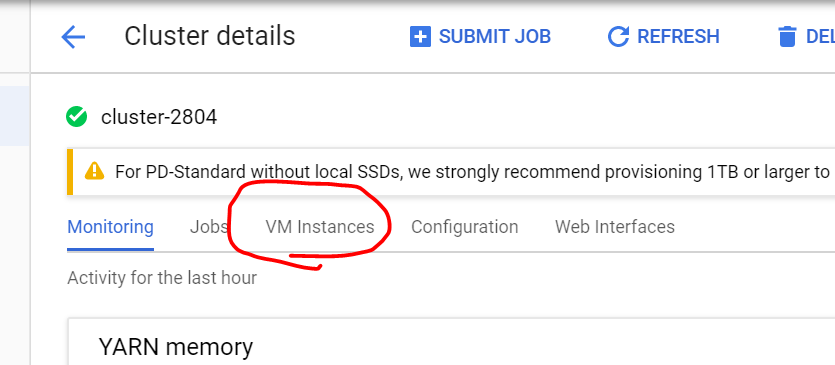
**Congratulations on creating a new cluster of three machines. Hadoop and other NoSQL databases are designed to run on clusters of commodity servers. In this case, we are using three machines for our new Google Cloud cluster.**



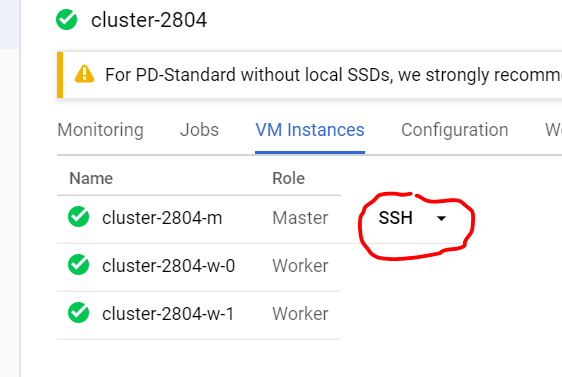
**Click on the cluster name**



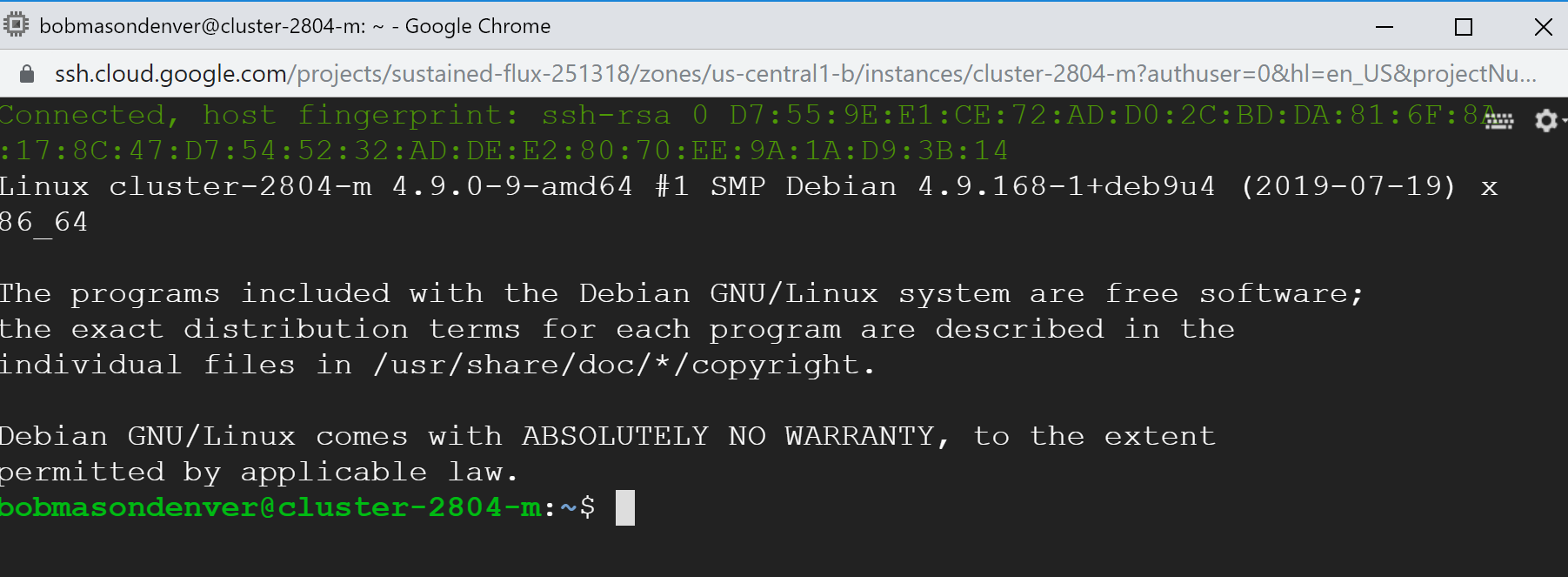
**Click on VM Instance**



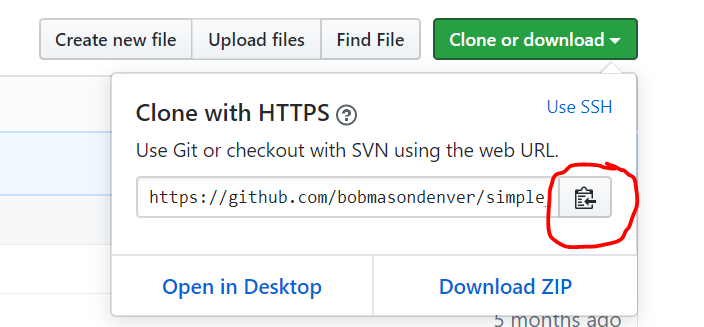
**Click on SSH to open an SSH window.**



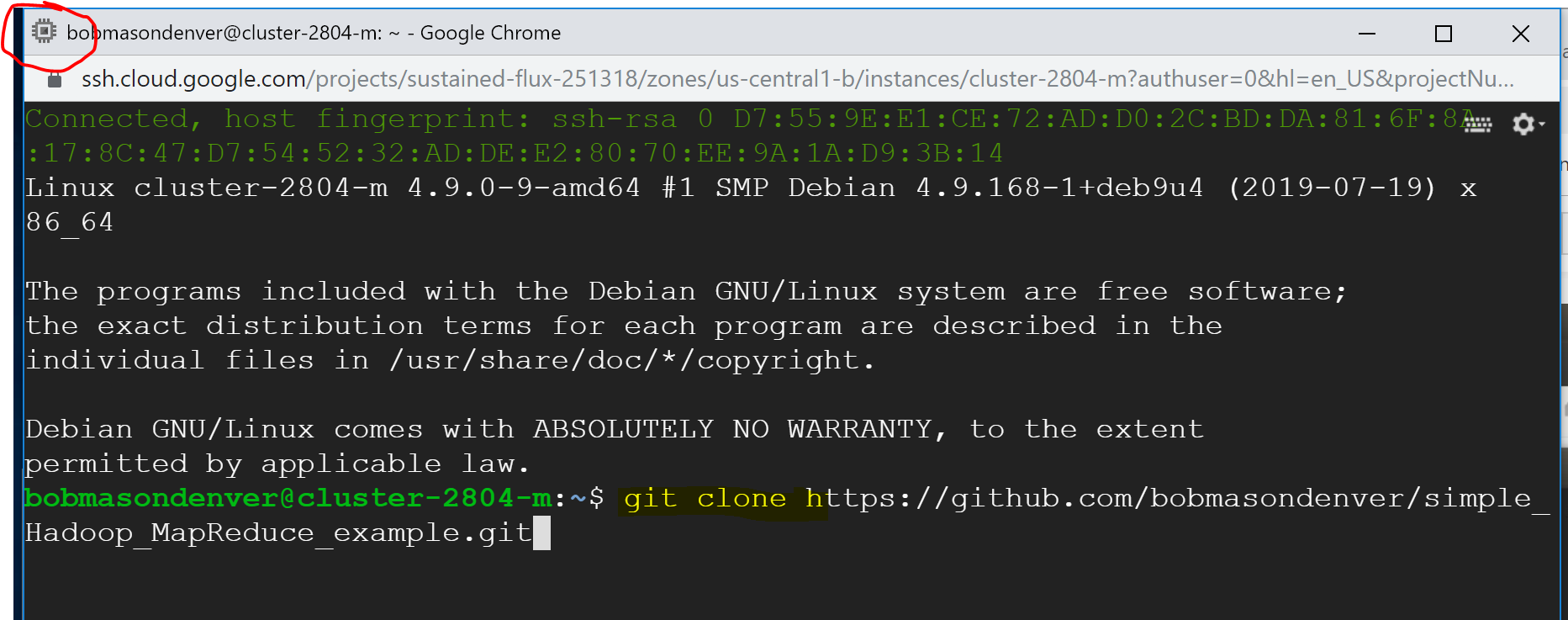
**An SSH terminal window will open, VERY NICE!**



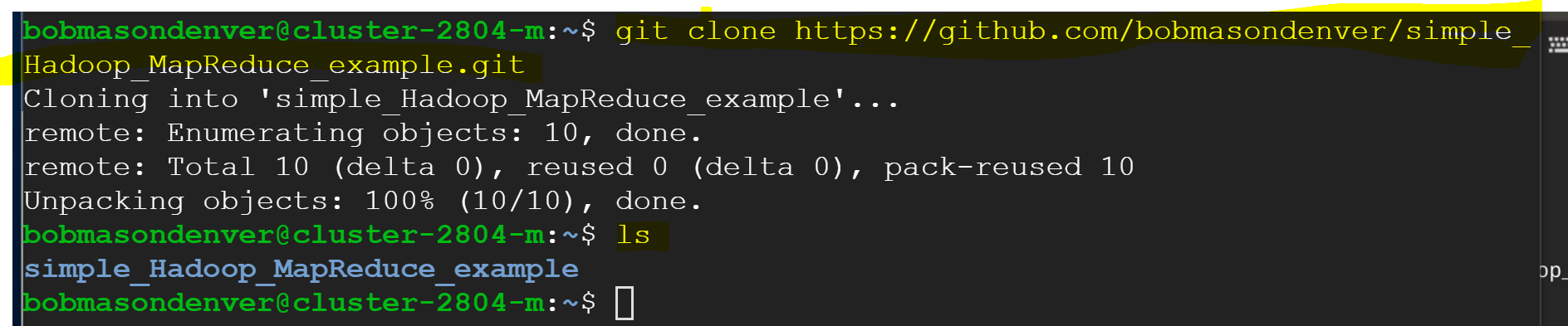
Go to your github repository and click on the clone or download button, then click on the copy button as shown below to copy the text to your PC clip board.



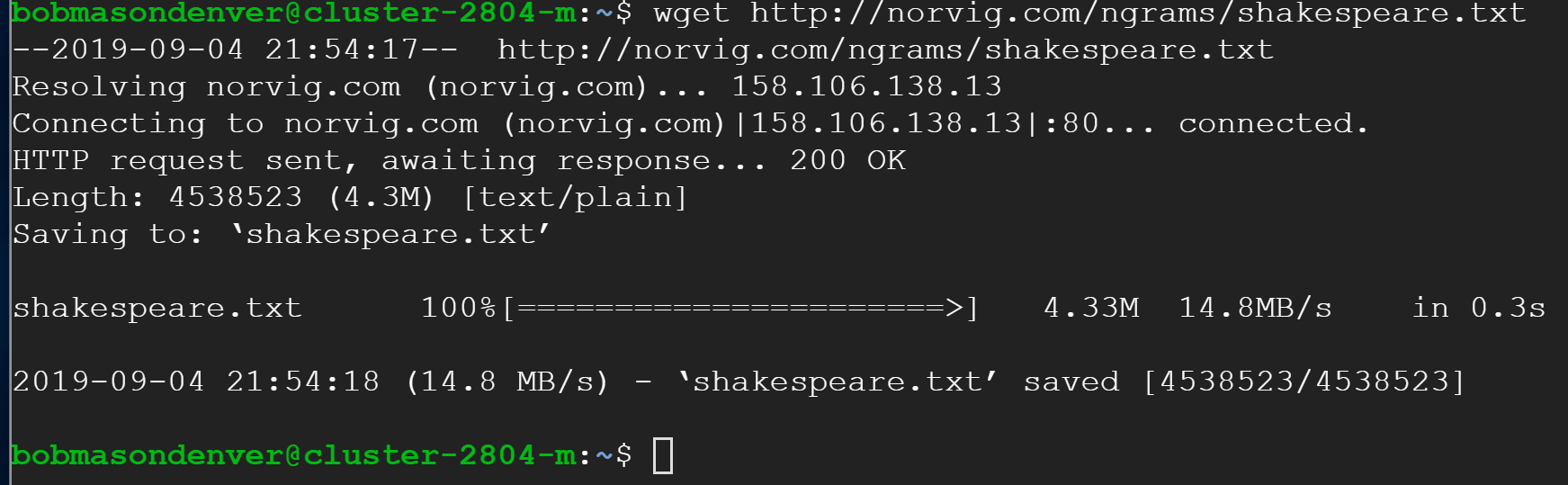
Go to the ssh terminal and type git clode, then you can use the icon on the top left of the window (circled in red below) to past in the rest of the string. Once your command line appears as shown below in the screen shot, you can press enter.



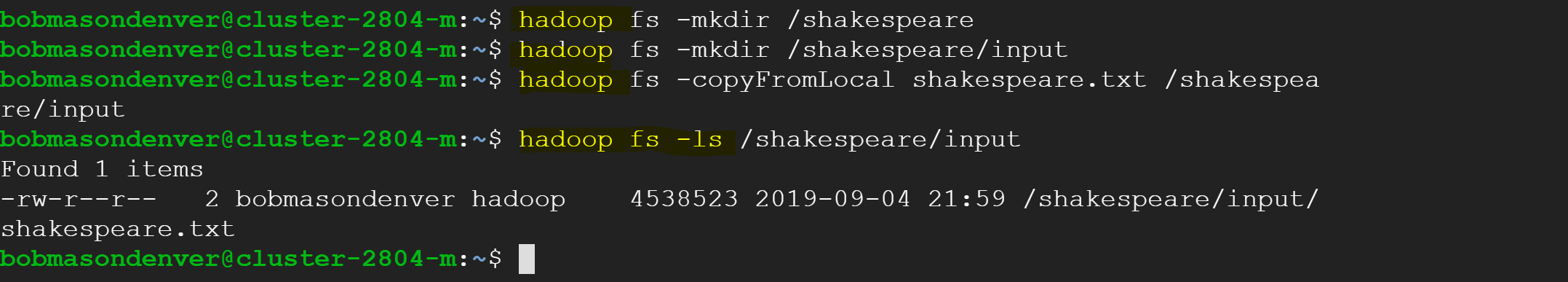
**You will receive a message that it was copied, then type a linux ls command to verify that the new directory exists (as shown below).**



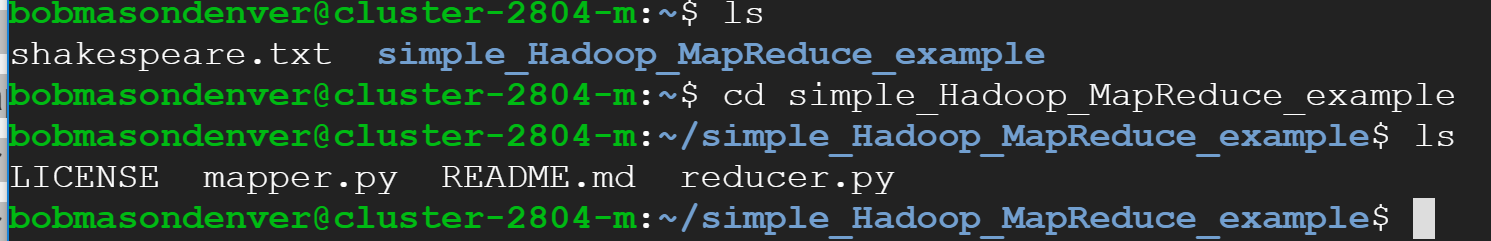
**Download the Shakespeare file that we used last week using the wget command from the lab 2.**



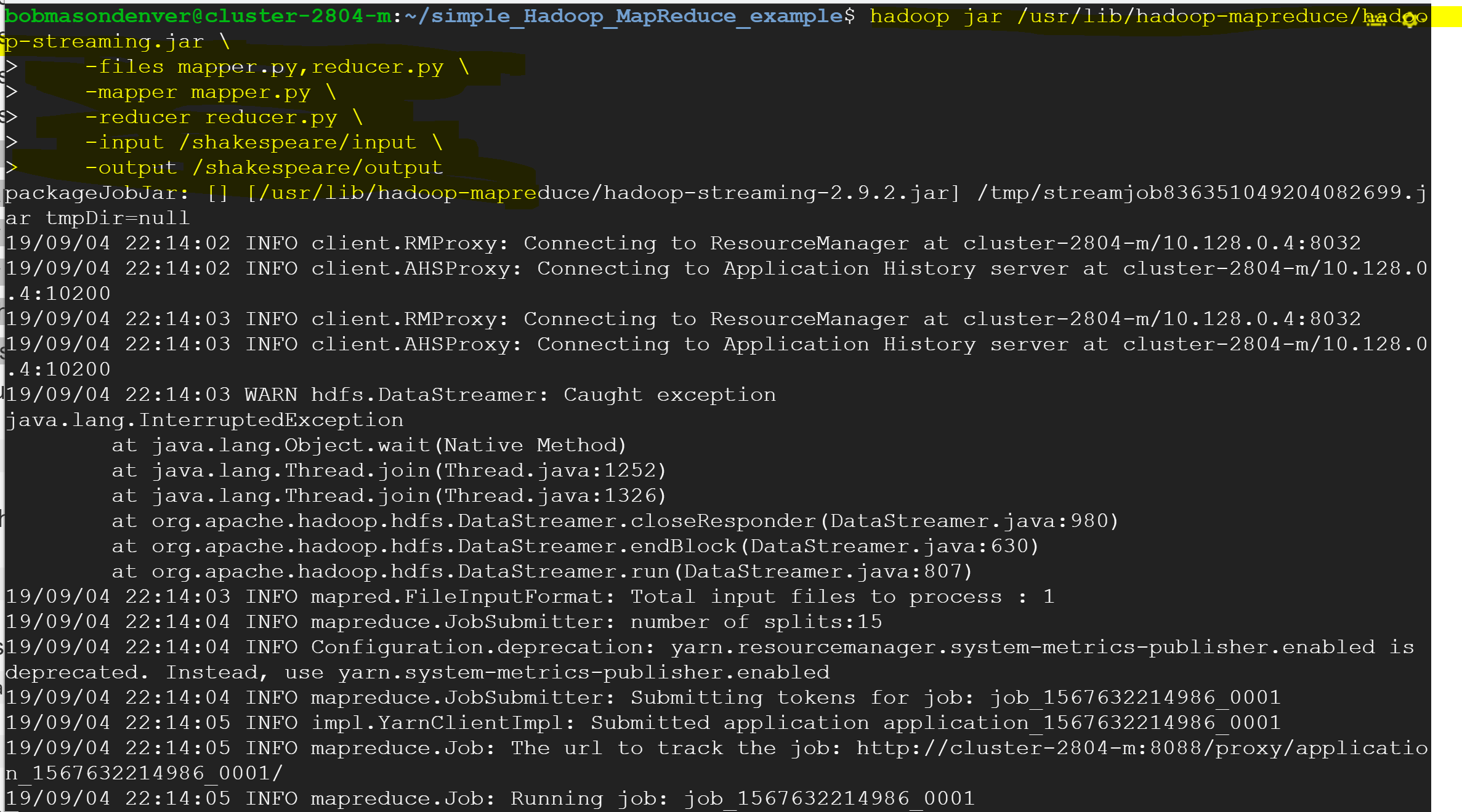
**You will copy the commands from lab 2 into the terminal window. To recap, we are using the hadoop fs command to make directories in hadoop (not our linux directory). Then, you copy the file to Hadoop and then verify that file was copied to the input directory.**



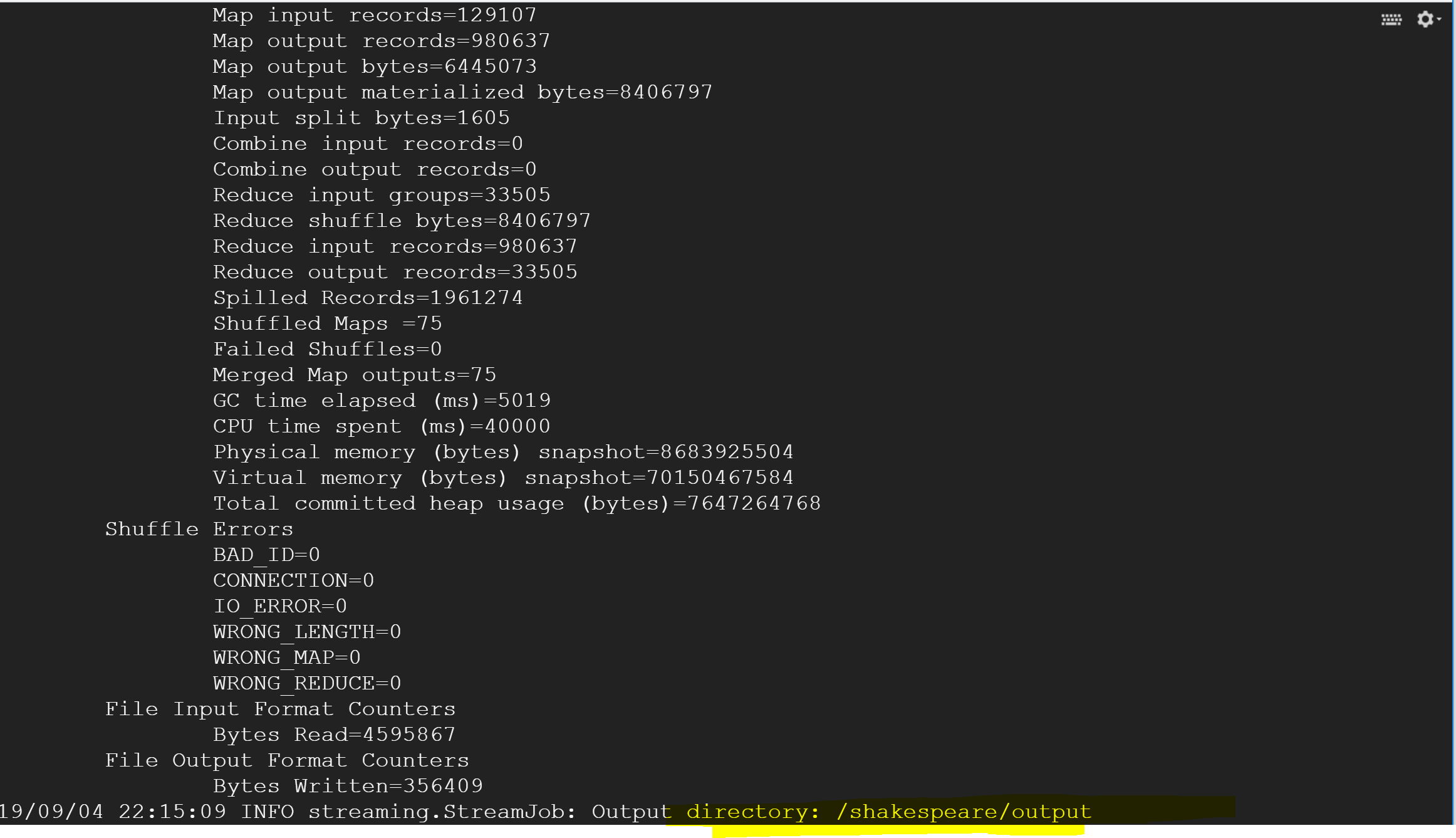
**Note: Change directories using the cd command into the simple\_Hadoop\_MapReduce\_example prior to running the map reduce script.**



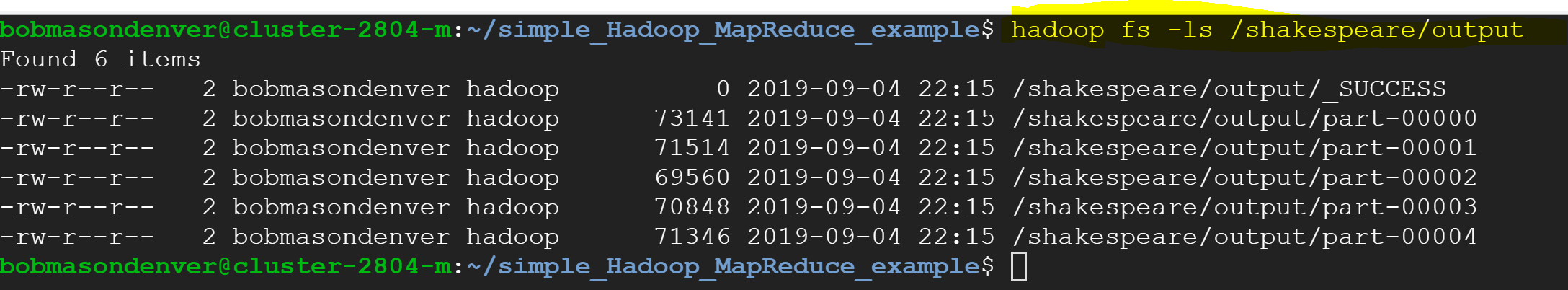
**Paste and then run the script from lab 2 to run the Hadoop job.**



**Notice that the output was written to the Shakespeare output directory.**



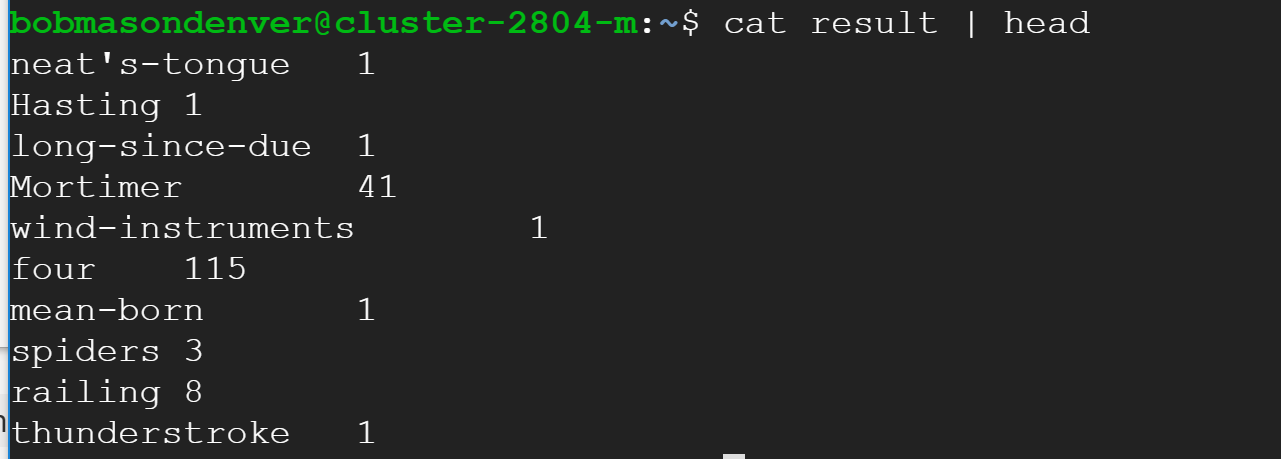
**Verify the output by using the Hadoop fs –ls command.**



**Copy the results using the merge command.**



**Then you can view the results using the Linux cat result | head command. The head extension will only display a limited number of rows at the top of the file.**



**This is the end of Lab 2 Part I Supplement with screen shots. Please see Lab 2 Part 2 Supplement for the rest of the screen shots for lab 2.**