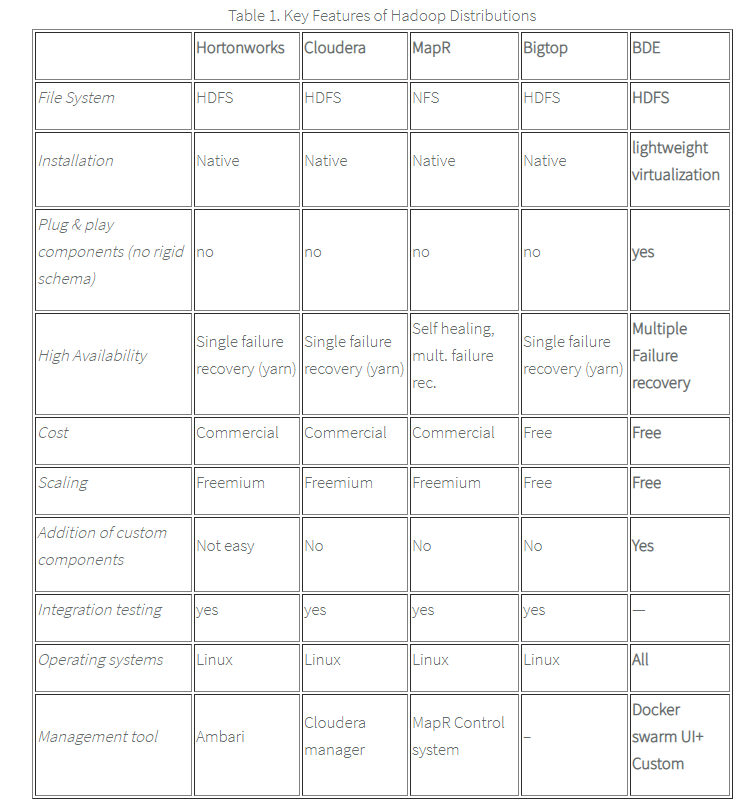
What is a Hadoop distribution?

Hadoop itself is open sources, as a result there has been a number of companies that have developed their own distributions. Simply put, the distributions provide improved code and functionality of the original Hadoop project. The distributions offer an ability to debug easily and offer software patches for an always up to date well-functioning package, these distributions also offer technical support, and tools for a more user friendly tool.

What is meaning of “Hadoop distribution” Retrieved from <https://stackoverflow.com/questions/35521709/what-is-meaning-of-hadoop-distribution>

What are some distributions of Hadoop?

From our readings this week the first notable distribution is BDE which stands for Big Data Europe, it is an open source platform that has developed interfaces for cluster creation and development, lessening the barrier to entry. Within the BDV vs. Other Hadoop Distributions article, the table below offered a great comparison across the various Hadoop distributions offered including MapR, Hortonworks, Cloudera, BigTop, and obviously BDE. This offers a compare and contrast across file systems, installation, plug and play components (which only BDE offers at this point in time), cost, scaling, custom components again BDE leading, integration testing, OS (BDE offering all vs. just Linux for the remaining distributions), and management tools. Since this is a pitch for BDE, obviously it seems like this offers the most functionality of the distribution options offered.



Ivan Ermilov. BDE vs. other Hadoop Distributions. <https://worldclass.regis.edu/d2l/le/content/245302/viewContent/3452021/View>

How does a Hadoop cluster work?

A Hadoop is a set of open source software utilities, utilizing a network of computers to provide a framework for distributed computing and storage by dividing a file into blocks and then storing it across a cluster of machines, thus the “cluster.” Hadoop will then create replicas of the blocks on the cluster, and distributes processing by dividing a job into various tasks that run in parallel over the computing cluster.

How Hadoop Works Internally – Inside Hadoop. Retrieved from <https://data-flair.training/blogs/how-hadoop-works-internally/>

Why does Hadoop create multiple output files?

In the map reduce process one output file is created for each reducer, the files that are created are named by the partition number for an example part r10000, r10001 and so on. The multiple output format class in MapReduce is utilized to format input data records into their appropriate sub directories based on the criteria that the data will be categorized by.

Hadoop Multiple Outputs Use Case. Retrieved from <http://hadooptutorial.info/category/interview-questions/mapreduce-interview-questions/>

How does this relate to the reducer step and number of compute nodes?

There are three phases to the map reduce process, shuffle, sort, and reduce. In the reduce phase the reduce task aggregates key value pairs. The multiple output process relates to the reducer step and the number of compute nodes because for each partition/compute notes an output file is built.

Hadoop Reducer – 3 Steps learning for MapReduce Reducer. Retrieved from: <https://data-flair.training/blogs/hadoop-reducer/>

Why should we use version control to store code for projects?

Version control is essential in storing code for projects because it allows you to step through the development process, create checkpoints by storing versions and being able to identify exactly what was changed or added from the prior updates. You can easily revert to a prior version if you blow up your code, and when you create updates you can note what you’ve done making it easy to debug. In addition, for collaborative work on large projects with various resources, you can work on your individual changes or development independently and you can merge in the pieces of development seamlessly, and again with the capability of reverting changes if there are issues.

What are pros/cons of using a cloud provider for a Hadoop cluster vs using our own in-house machines?

According to an article written by Shahebaz Shaik on Linked in there are a couple of factors to consider when running a Hadoop Cluster organizationally. The first consideration should be cost, if an organization wanted to build the infrastructure to run Hadoop on premise the overall cost of the numerous servers required, coupled with simply the space or real-estate needed, along with the support to run such hardware, and the power required ends up being a very large start up costs to scale for big data. This is why cloud computing is an easy decision especially for small organizations, or individuals like myself. The downside of utilizing cloud services is security, if data is stored locally it is much easier to monitor, keep the data locked down and know who has access etcetera. However, cloud providers for the most part have built strong data security measures. Users of cloud services do not offer the ability to control hardware a con, but offer software based managed services which is a pro.

“Hadoop On Cloud vs Hadoop On Premises” Shahebaz Shaik. January 2, 2017 Retrieved from <https://www.linkedin.com/pulse/hadoop-cloud-vs-premises-shahebaz-shaik/>,

What do our top word counts look like now, and what do they tell us?

At a very high level, today’s word counts can analyze sentiment of consumers or users of social media on different topics or consumer goods. They can count the number of instances that a positive sentiment or a negative sentiment about a topic or product to get a general feel for how consumers or the general public feel about certain things. An example is the sentiment of presidential candidates, utilizing twitter as a feed for analysis to get the publics feeling about a candidate by utilizing a word count algorithm to count the frequency of words such as: strong, presidential, appreciate, social justice etcetera. When it comes to consumer sentiment, companies can analyze the sentiment of a certain product by counting both positive associated or negatively associated words with a certain product or service to get indications of what consumers like or dislike about a particular service or good in the marketplace.

Explain what is going on in the download\_stopwords.sh file

The stopword.sh file provides our word count with a listing of words that are “throw-away words,” or words if counted would not provide any indication of what we are analyzing. In particular the stopword.sh file is the nltk list of English stopwords, that allows the algorithm to skip over words like: I, me, myself, and, or, but etcetera. Theses words provide nothing substantial to analyze with their counts.

What are some other use cases of Hadoop, and how we might accomplish such things in Python (e.g. pseudocode or a high-level description of the way to solve the problem with Python)?

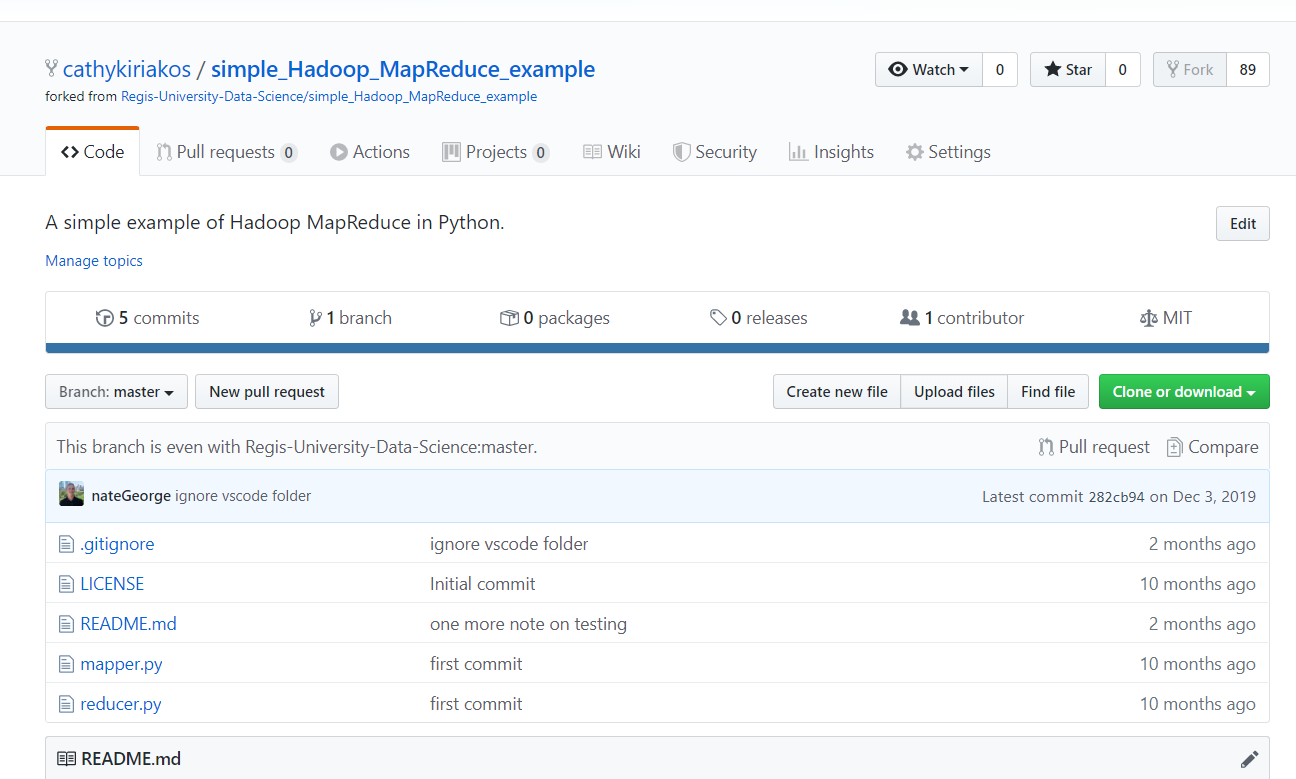
In an article written on Hackernoon, a handful of alternative use cases for Hadoop was listed. This included the following notable applications:

1. Converting site visitors: Hotels.com utilized Apache Cassandra and NoSQL databases to notify potential customers searching availability at locations that x number of other people are looking at this property at the moment.
2. Easy data analysis from multiple sources: Marks & Spencer utilized Cloudera Enterprise Data Hub Edition to better understand their consumer behaviors. By utilizing Hadoop the organization was able to use predicative analysis to keep their shelves fully stocked during times of high demand.
3. RBS utilized Cloudera to increase customer satisfaction by utilizing machine learning and predicative analysis to provide an in-depth knowledge of customer behaviors and needs. Allowing for the bank to foster strong relationships with end users.
4. Yahoo saved a ton on hardware costs for their 150 terabytes of machine data going that was moving through their data warehouse daily by using Hunk a Hadoop tool developed by Splunk that assisted them by utilizing real time data management. End users benefit by Yahoo identifying spam before it enters their inbox.

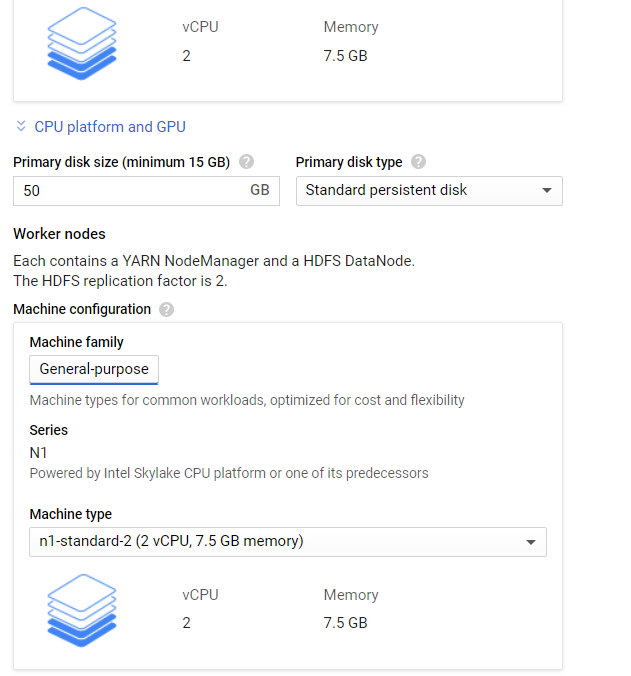
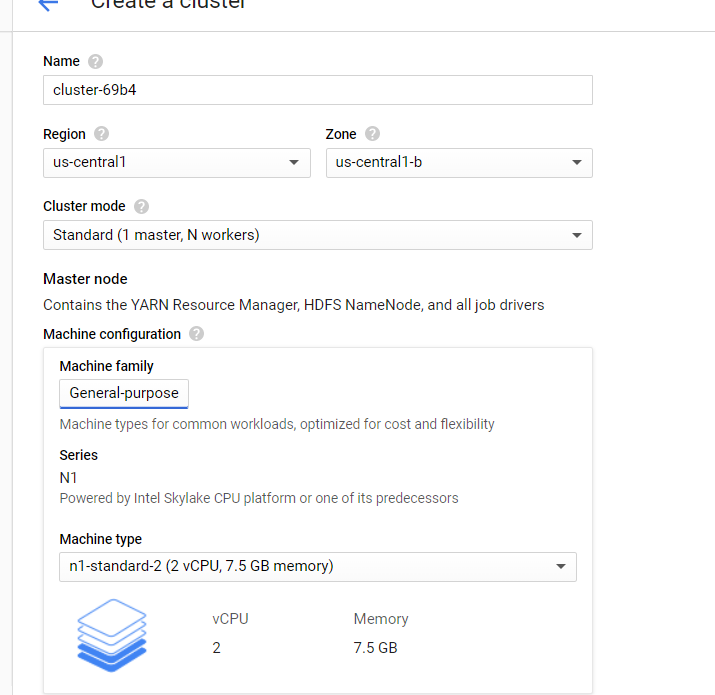
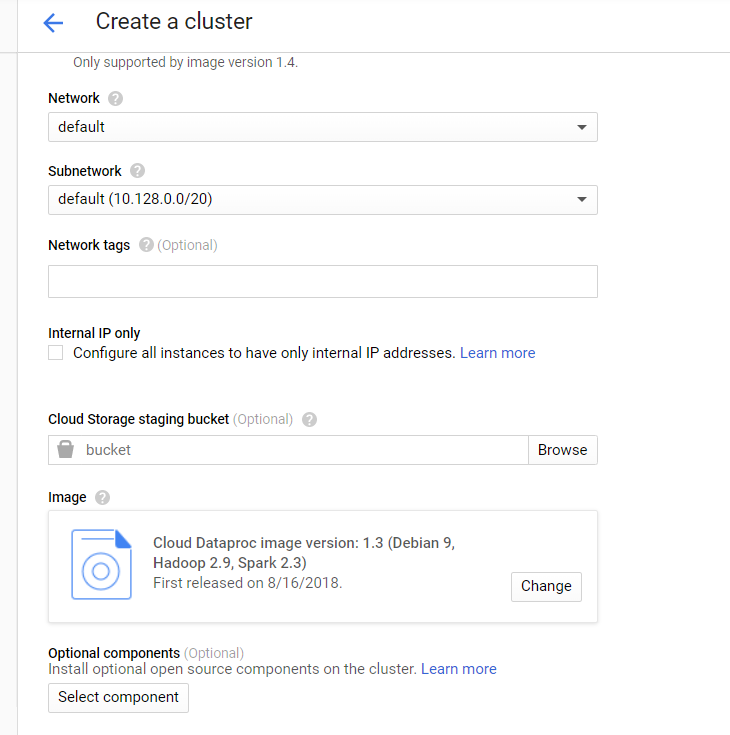
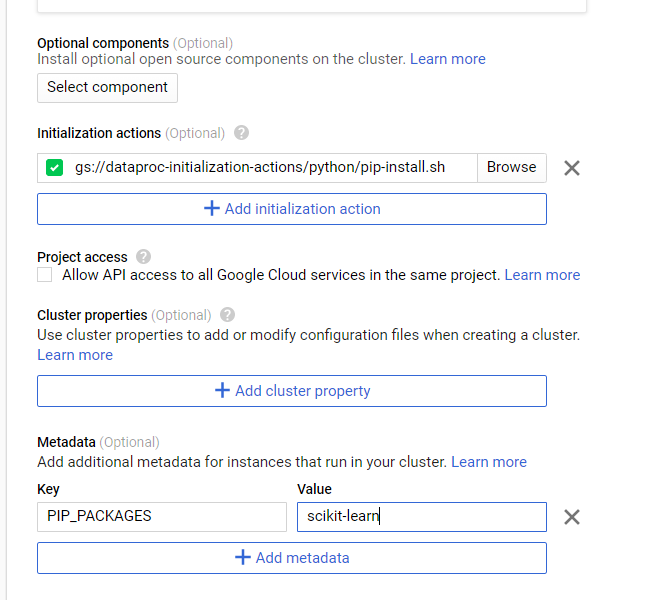
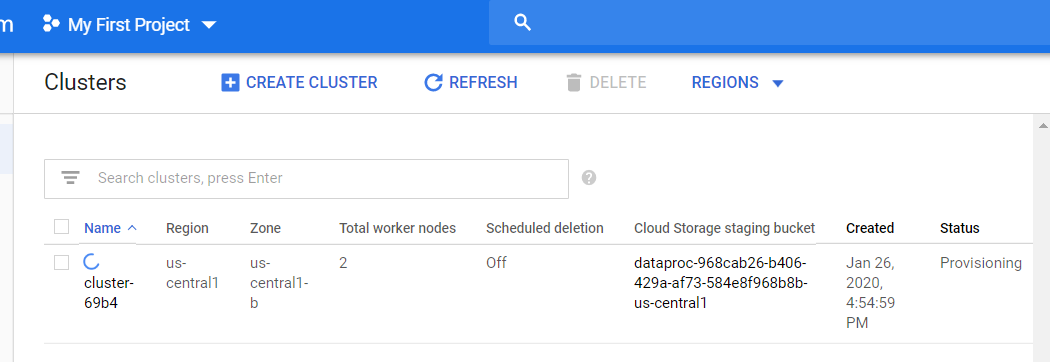
5 Use Cases of Hadoop that makes it an Asset to Enterprises. April 4, 2017. Retrieved from <https://hackernoon.com/5-use-cases-of-hadoop-that-makes-it-an-asset-to-enterprises-1486dedb409a>

**Technical Portion:**

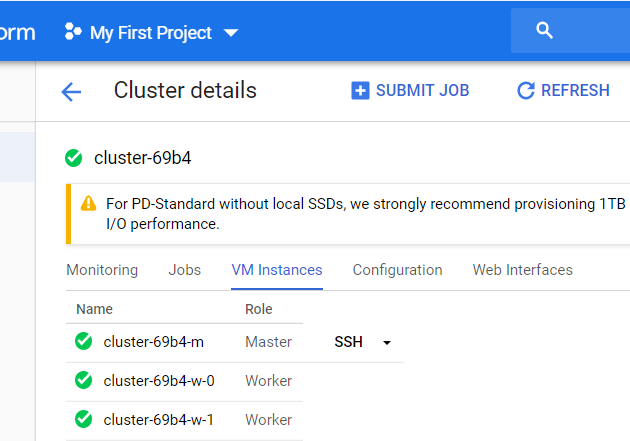
I created a new Data Engineering Repo in my GitHub profile, from here I created a fork to my profile as shown below:



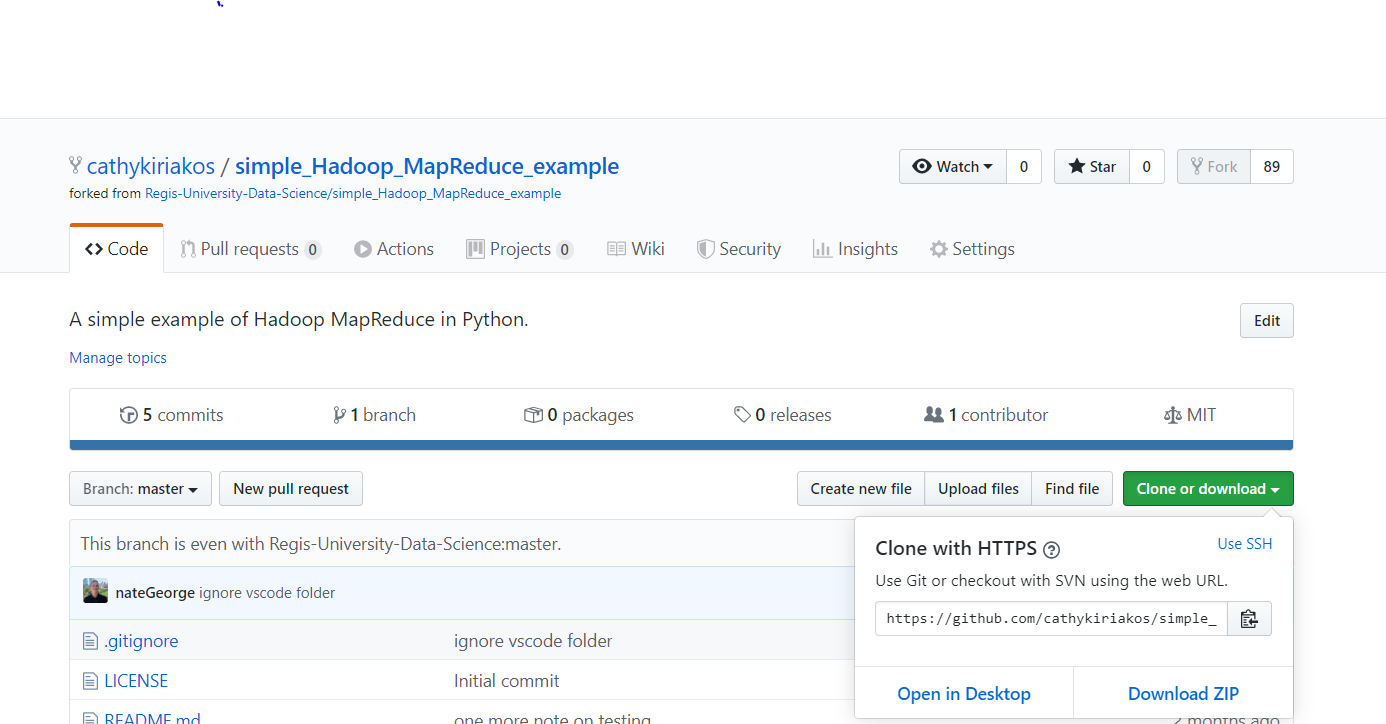
Then I navigated to my google cloud platform, from the home page I navigated to Big Data – Data Proc and enabled the api to create a cluster

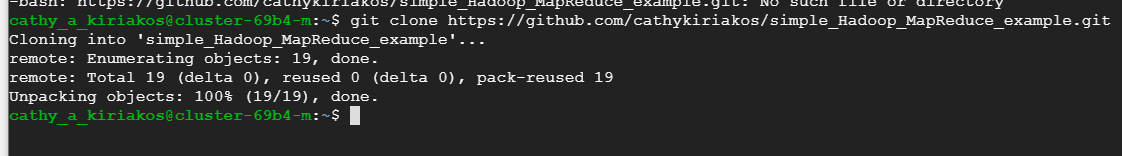
   Then I created my new cluster of three machines as shown below 

Then I clicked on my cluster name, within the cluster details I navigated to vm instances as show below, then clicked on ssh to open a terminal

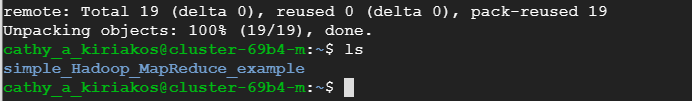


Then I Navigated back to my github page, and cloned the repository as shown below:

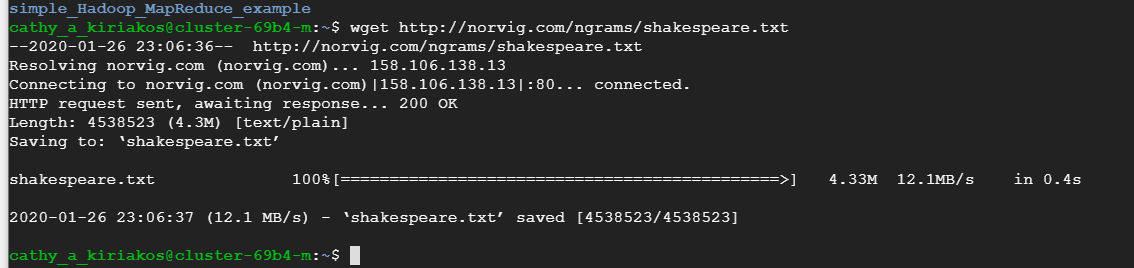




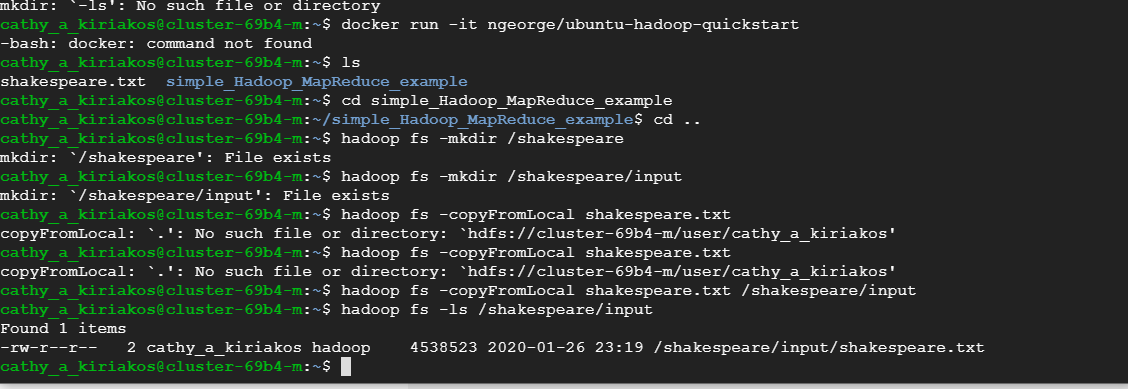
By calling ls I can see that the repository is there:



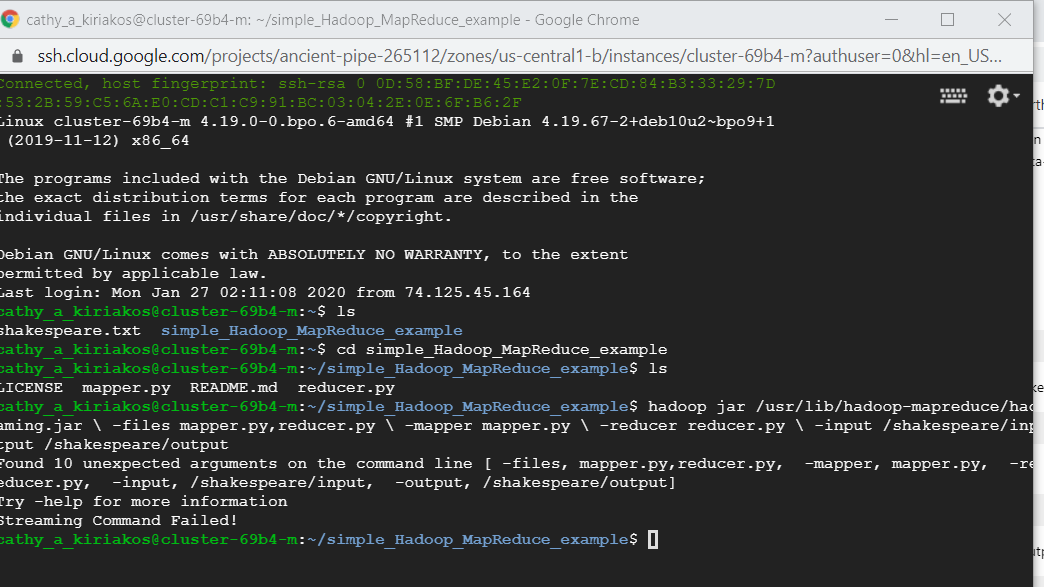
Then I downloaded the Shakespeare text from last week



Then I created the directories as I did last week just changing the calls to Hadoop fs



Then I navigated into the MapReduce Example directory by changing my directory, and then called the MapReduce scrips, where again I ran into streaming errors.

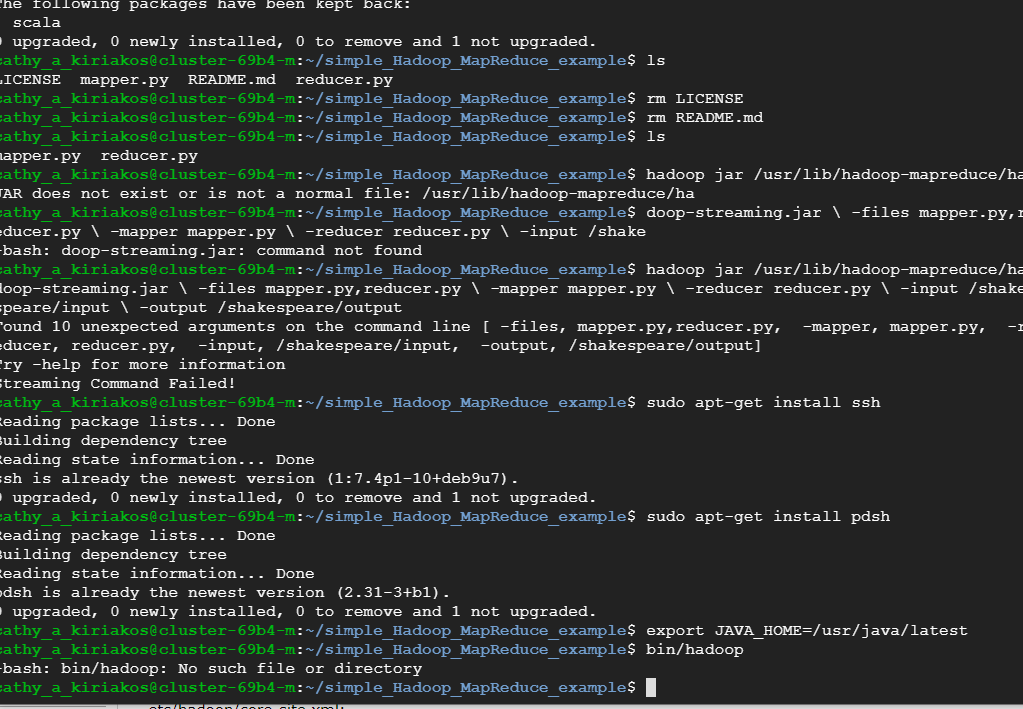


Initially I thought that it was the way that I was calling the map reduce because of my Hadoop version. But I can see that I’m on version 2.9.2 – confirming the way that I called the script.

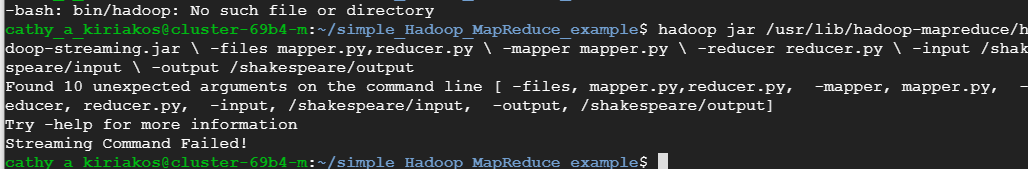
I did utilize my ubuntu partition and ensure that everything that I needed installed locally and within my vm was there, including java jdk version 8.

<https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html>

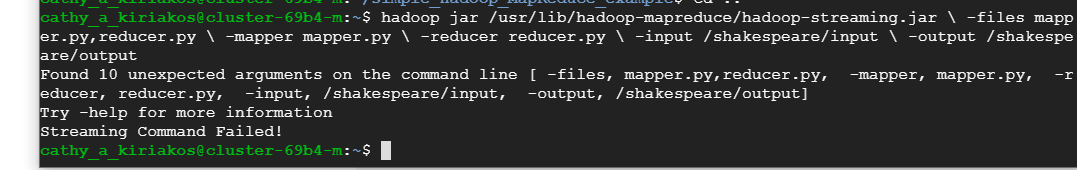
There were some noted prerequisites so I decided to go ahead and install them as well:



And tried to run the script again, getting the same error:



So, I thought maybe I’d change directories for the sake of it and see if I could get the script to run



With no luck. So I will continue to debug again this week.

I will move on to getting the stopwords, my nano terminal was empty versus the example shown in the supplemental. Making me think that there was an issue with python.

At this point since I’m a week behind as a result of my initial streaming issues. I will continue to debug and resubmit a completed lab once I can get everything working as it should.