Assignment 4

BIG DATA ANALYSIS

Competition

REPORT

TABLE OF CONTENTS

TEAM: = (3 MEMBERS)

Vidhan Dholakia (20151326)

SOFTWARE USED:

Microsoft Azure:

Cloud computing is when I can access computing services like servers, storage, networking, etc. over the Internet from a provider like Azure or AWS. In this assignment, I have used Microsoft Azure, which is a cloud computing platform. MS Azure is cost-effective, secure, reliable, flexible and durable as compared to the local or on-premises servers. I can directly access MS Azure using my authorized account. The advantage in such cloud-based platform is that I pay only for the things I use. Once I are done with the task, I must turn off the Virtual Machines and make sure every process working in the background is executed and completed. Failing to release resmyces in time would cost the user more everytime.

Hadoop / MapReduce:

Apache Hadoop is an open smyce framework for distributed storage and processing of large data on commodity hardware. The base Apache Hadoop framework is composed of several modules including Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN (Yet Another Resmyce Negotiator, and Hadoop MapReduce. Hadoop MapReduce is a programming model for large scale data processing.

Spark (PySpark):

PySpark is the Python API for Spark. PySpark is a great language to learn in order to create more scalable analyses and pipelines. It is used for performing exploratory data analysis at scale, building machine pipelines, etc. PySpark fully supports interactive use i.e. can be easily run. ./bin/pyspark to launch an interactive shell.

Installing and Configuration:

By default, PySpark requires python to be available on the system PATH and use it to run programs; an alternate Python executable may be specified by setting the PYSPARK_PYTHON environment variable in conf/spark-env.sh (or .cmd on Windows).

All of PySpark's library dependencies, including Py4J, are bundled with PySpark and automatically imported. Standalone PySpark applications should be run using the bin/pyspark script, which automatically configures the Java and Python environment using the settings in conf/spark-env.sh or.cmd. The script automatically adds the bin/pyspark package to the PYTHONPATH.

Hortonswork Data PlatformSandbox:

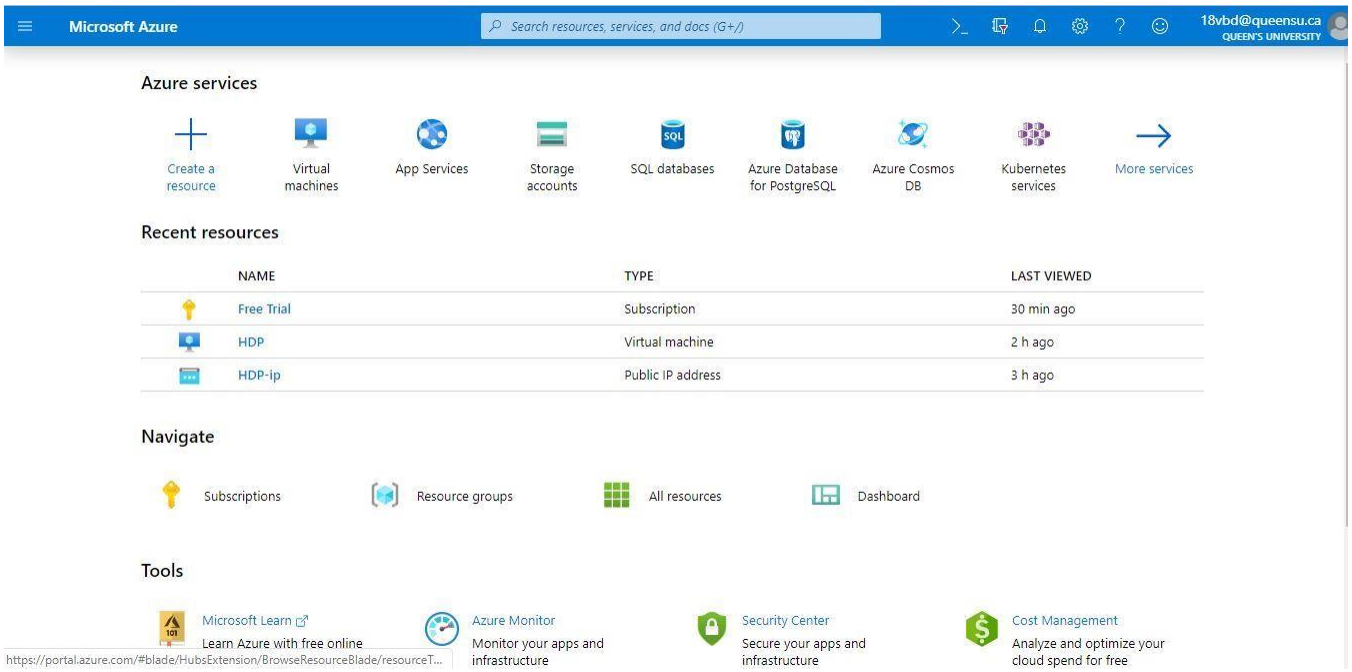
The HDP Sandbox makes it easy to get started with Apache Hadoop, Apache Spark, Apache Hive, Apache HBase, and Data Analytics Studio. The Hortonworks Sandbox is a single node implementation of HDP. It is packaged as a virtual machine to make evaluation and experimentation with HDP fast and easy. HDP is free to download and use in my enterprise and I can download it from https://www.cloudera.com/downloads.html.

How to Download the Data Set?

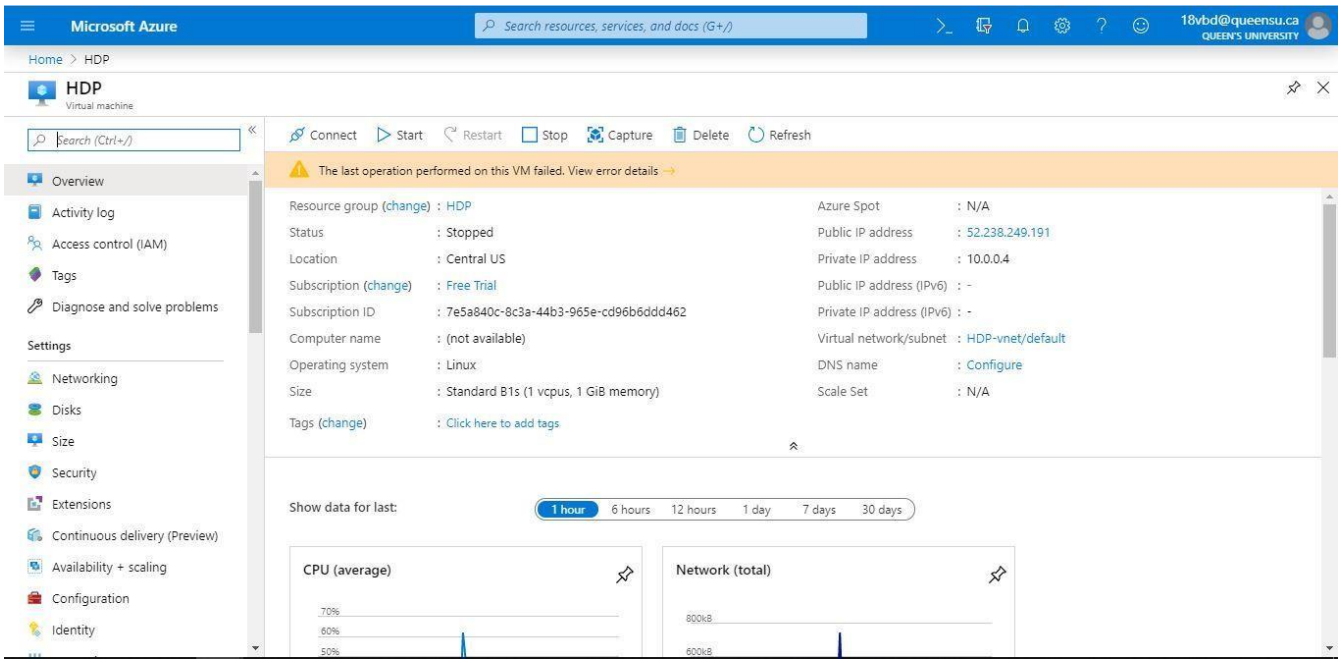
I have been given the list of both positive and negative words and also one sample data file. Other than this, I have a large data file to calculate the sentiment for each of the mentioned products.

ANALYTICS PROCESS:

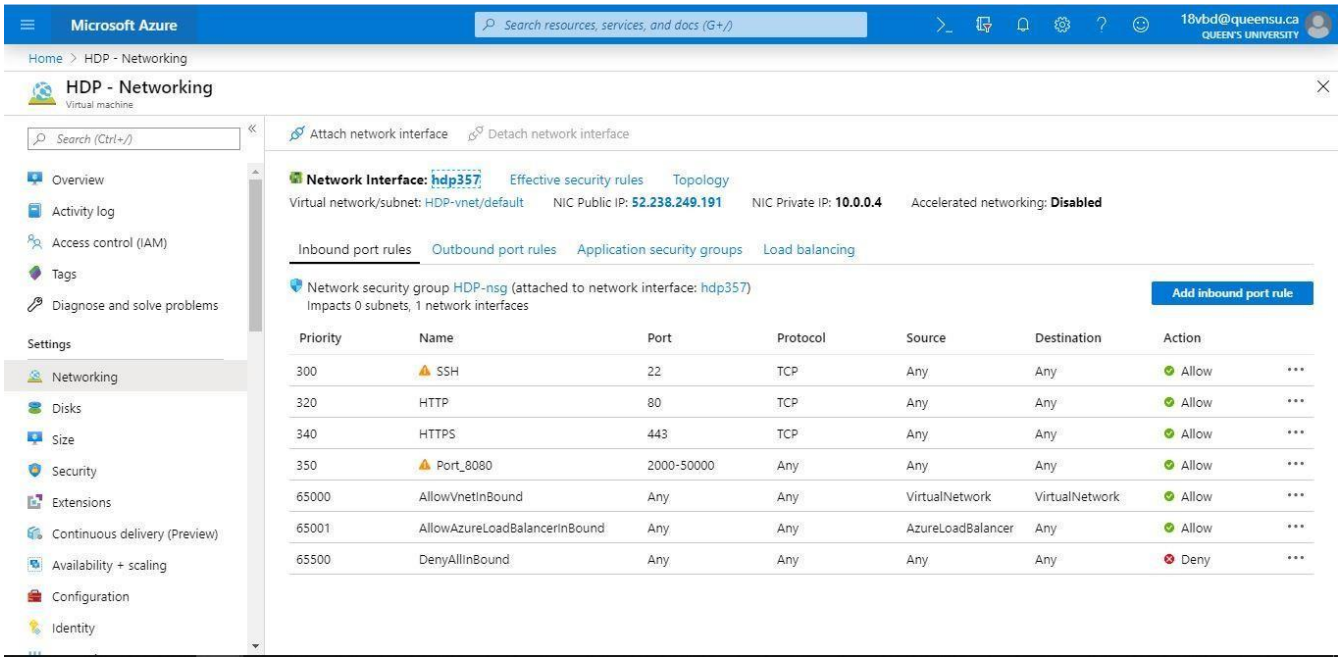
Logging into Microsoft Azure to create Virtual Machine for Hortonworks Data Platform.



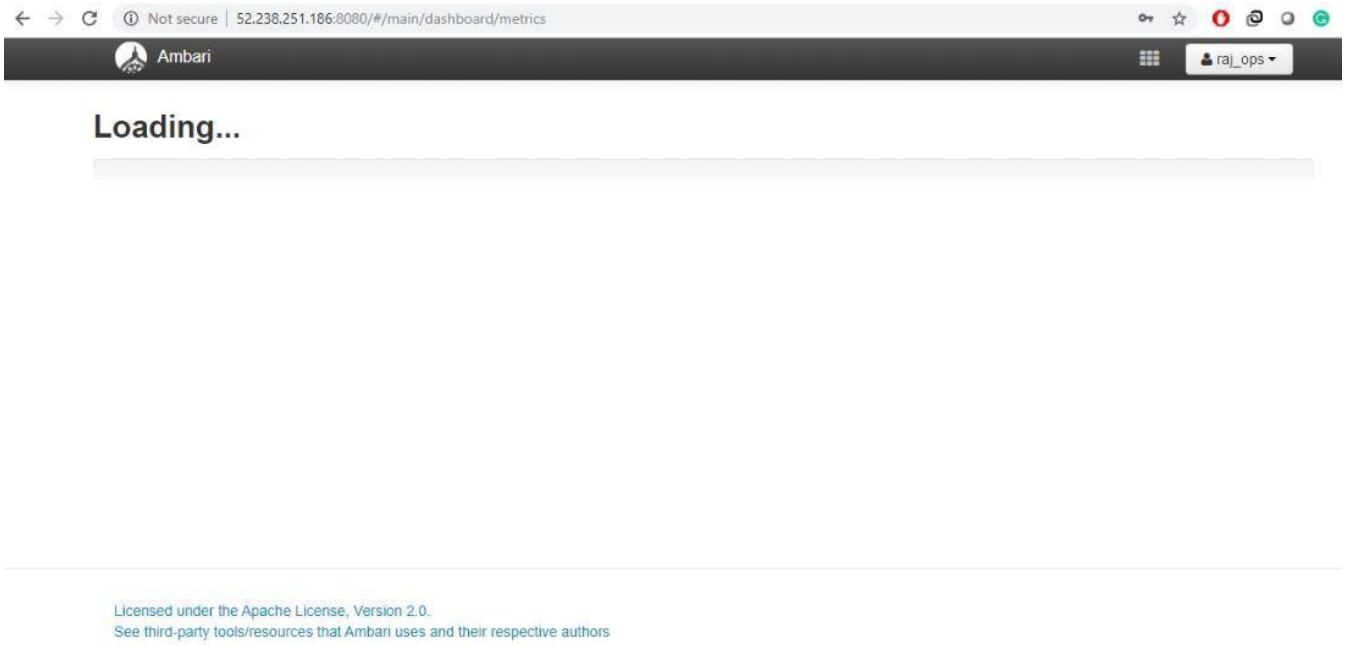
Overview of the Hortonworks Data Platform virtual machine.



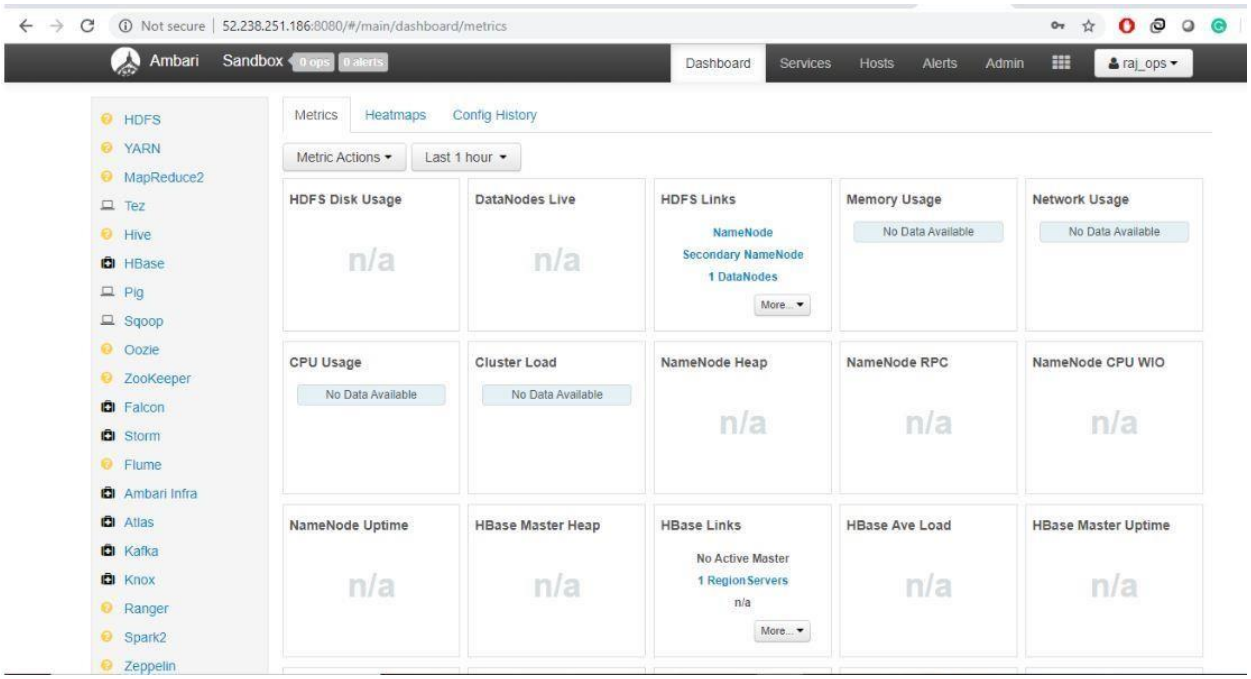
Add extra ports for HDP and Ambari server to communicate with the user. The ports for Ambari server is 8080, for SSH connection is 2222, for Zepplein notebook is 9995.



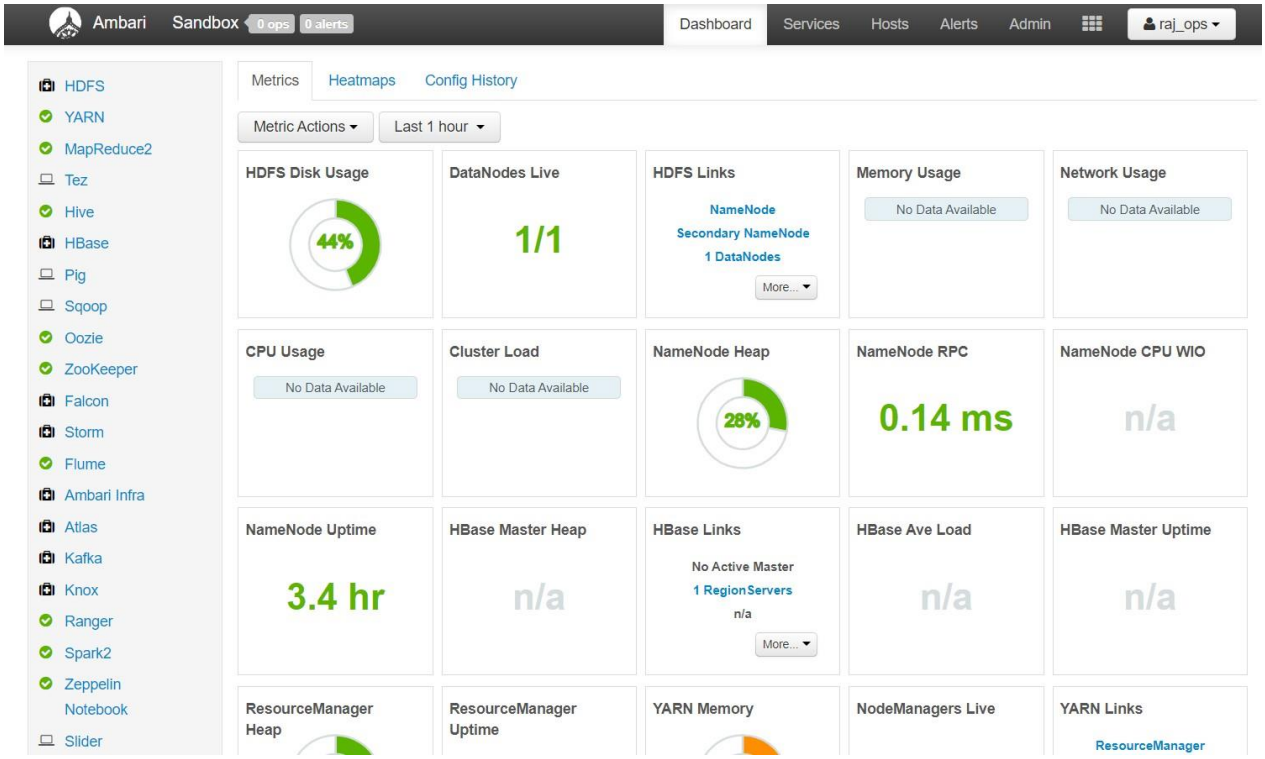
Ambari Server loading with credentials as mentioned on Hortonworks Data Platform page.



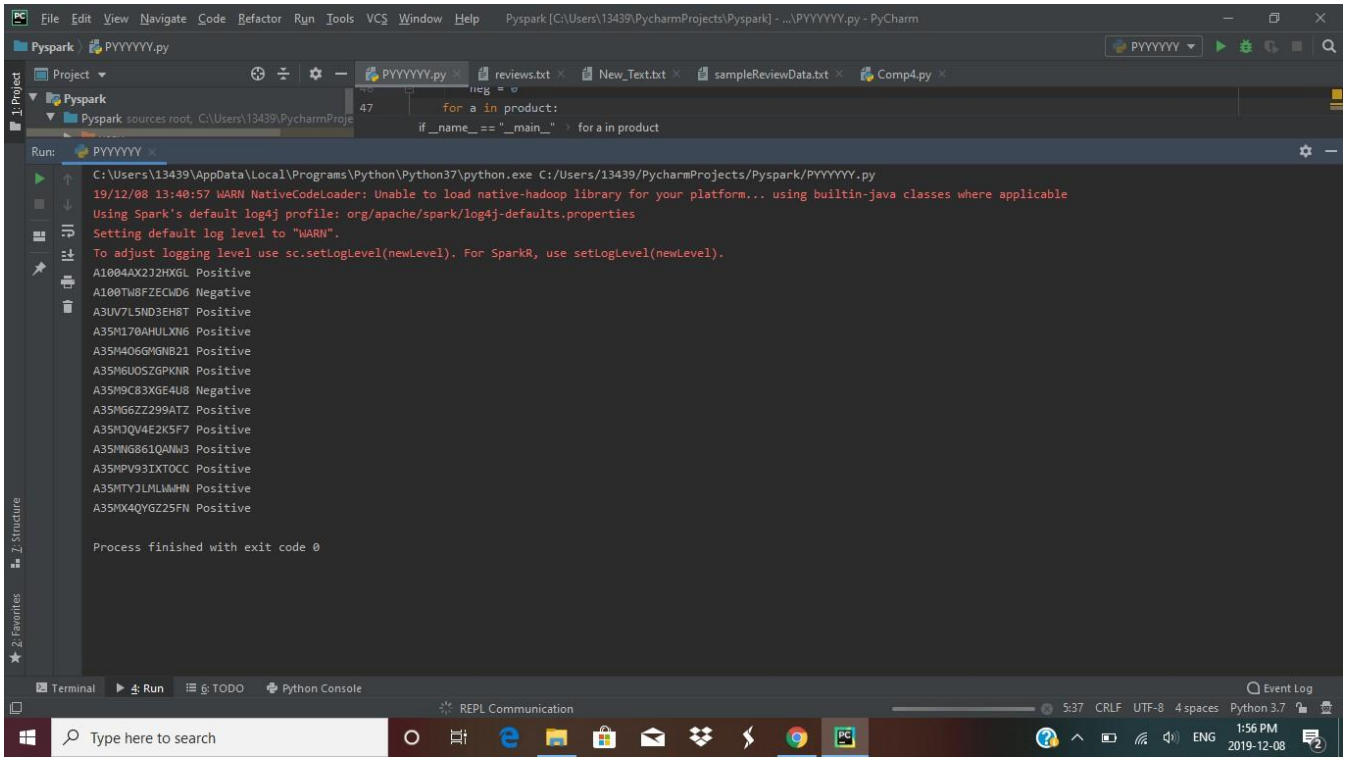
This is how Ambari server looks but the services aren't responding and hence I need to either restart all services or just the required services



With all services running, Ambari would look like this.



Pyspark output on local machine with set of data this is the output obtained for the code attached as a zip file in this competition

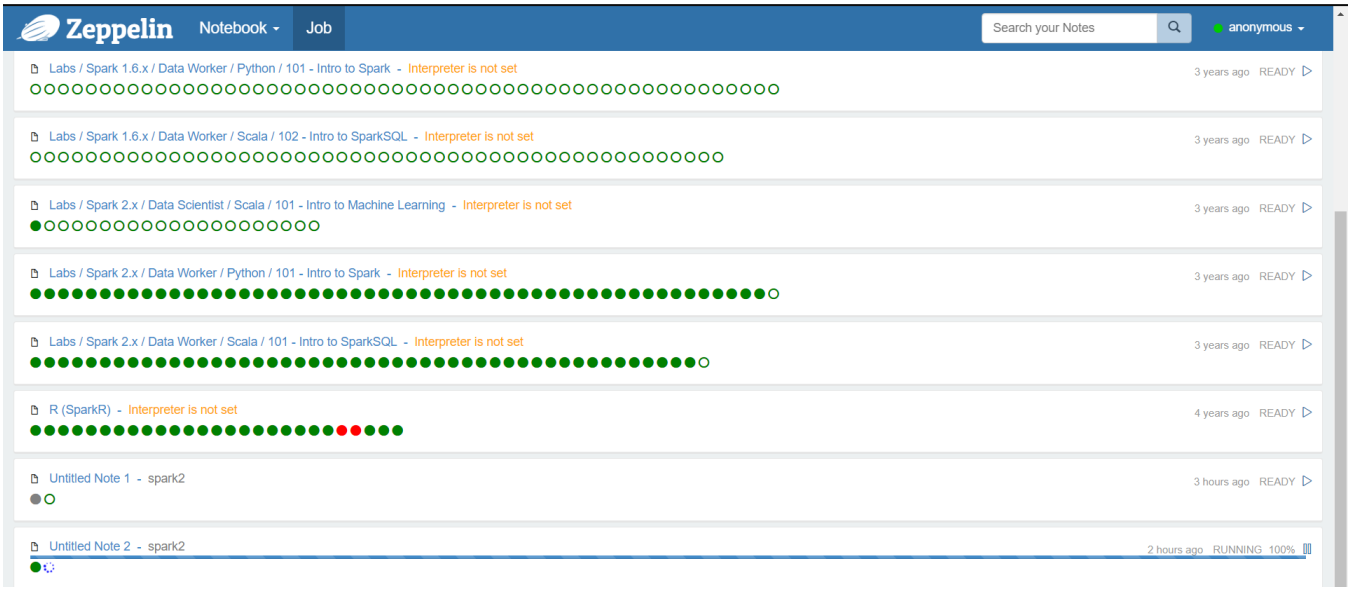


CHALLENGES:

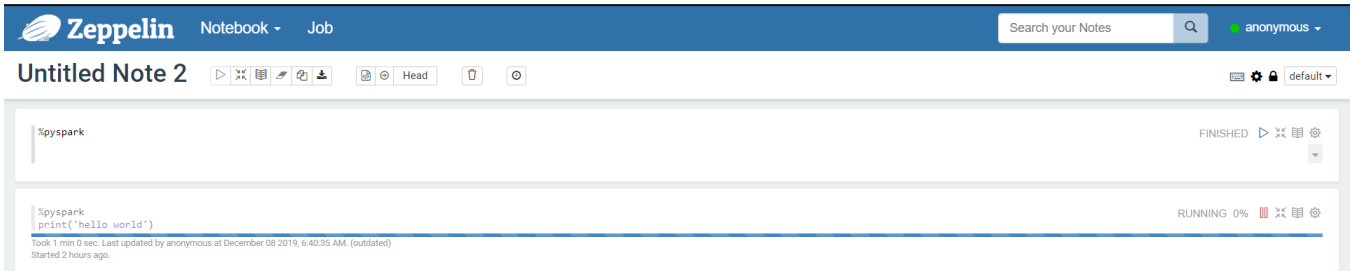
I came across many of the challenges as described:

- It was tough to set up the Microsoft Azure Virtual Machine which took most of the time in the competition.
- I started the work on this competition by using a physical computer to see what I are to deal with. I Ire able to get a desired output on the sample reviews.
- Large data was the problem to handle, uploading data onto cloud would take a lot of time.
- Azure virtual machine was not supportive enough with lack of documentation and very less time to pull the task I found it difficult to implement it on cloud.
- Availability of a lot of resmyces on the Azure cloud also creates problem as one should be aware of the choice of what resmyces should be used as it saves time and credits.

- Zeppelin Issues: The Azure lagged most of the time. As per the following screenshot, I Ire having trouble to work on PySpark on Zeppelin. Despite having the PySpark interpreter, I got the following output (Untitled Note 2 running from past 2 hmys and having no response):



Even a simple import function of PySpark Ire lagging a lot.



Another problem that I ran into was YARN out of memory error. I tried upgrading my storage as Ill as virtual memory but in vain. All services got stopped when the YARN memory gets filled and I had to restart my services which again takes around 1 hmy. Unfortunately, I had to complete my task with the memory problem.

IMPLEMENTATION ON PYSPARK:

- To run on local machine, I needed a *pyspark* setup first which requires *python* along with *JDK* 8 and *winutils*.
- I implemented my code on *PyCharm IDE* with all the available libraries of PySpark, I did this as above screenshots show I couldn't implement it on Zeppeline as it wasn't responsive most of the times.
- The spark code works in a pretty simple way I have 3 text files as input data viz. positive_data.txt, negative_data.txt and reviews.txt file.
- I convert this input txt's to RDD with Spark Context and use them to carry out map-reduce.
- With help of this map-reduce technique I predicted if the review/sentiment from a comment was positive or negative. For I mapped number of positive and negatives for each tag. Based on the number, I decide if it's a positive review or negative. Since, I not having a neutral feedback, I had to use greater than and less than function to deduce the positive and negative feedback.

REFERENCES:

[1] https://www.python.org/doc/

[2] https://docs.microsoft.com/en-us/azure/

[3] https://datasciencetoday.net/index.php/en-us/tutorials-top/163-sentiment-analysis using-python-part-i-machine-learning-model-comparison

[4] http://saptak.in/spark/001-configuring-hortonwork-sandbox-azure.html

[5] https://spark.apache.org/docs/latest/api/python/index.html

[6] https://www.cloudera.com/tutorials/getting-started-with-hdp-sandbox/1.html

[7] https://towardsdatascience.com/a-brief-introduction-to-pyspark-ff4284701873