**Principles of Big Data Management**

**Project Phase-1**

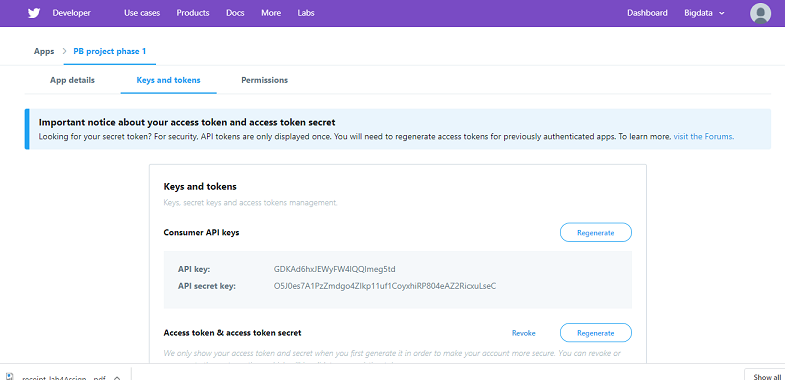
Harika Gurram – 16300892 – hghb6@mail. missouri.edu

Sumanth Medavarapu – 16295321 – smhqb@mail.umkc.edu

Ramya chennupati – 16297780 - scc9c@mail.missouri.edu

1. **Log in to Twitter developer account:**

Firstly, we need to create an account and then create an app in the Twitter developer’s account, user credentials(API keys, Access tokens) for accessing the Twitter streaming API will be generated.



1. **Importing Tweets:**

Tweepy and app secret keys are used for communication with Twitter platform and for importing the tweets.

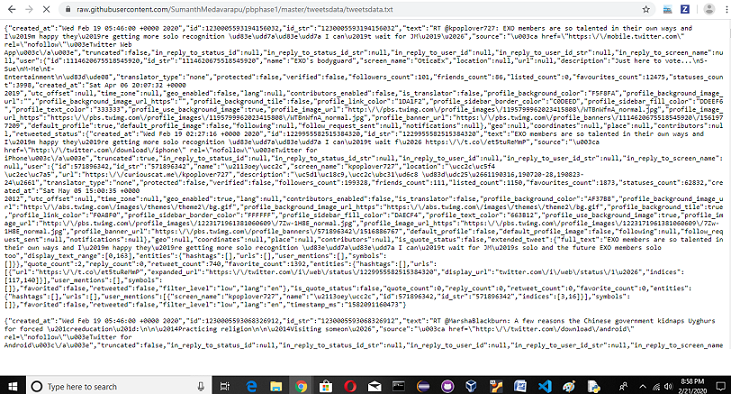
***No. of tweets collected:*** 1,00,000

***Source code****:* tweetsimport.py

<https://github.com/SumanthMedavarapu/pbphase1/blob/master/source/tweetsimport.py>

***Output***: tweetsdata.txt

<https://github.com/SumanthMedavarapu/pbphase1/blob/master/tweetsdata/tweetsdata.txt>



1. **Extracting Hashtags and URLs:**

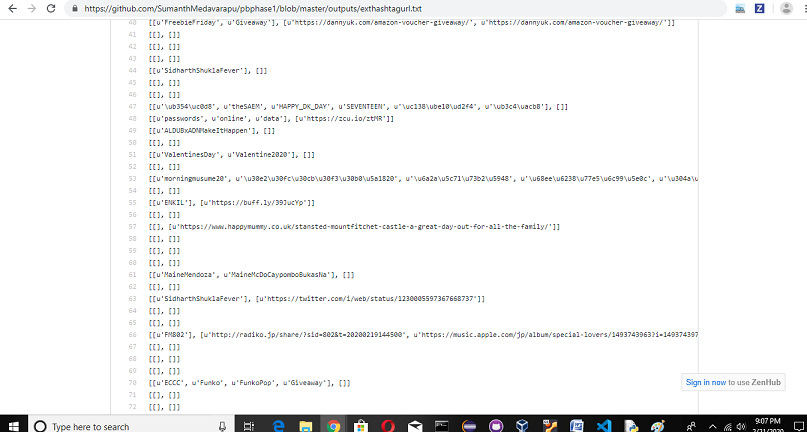
Hashtags and URLs are extracted from the imported tweets.

***Source code:*** extract\_url\_hash.py

<https://github.com/SumanthMedavarapu/pbphase1/blob/master/source/extract_url_hash.py>

***Output:*** exthashtagurl.txt

<https://github.com/SumanthMedavarapu/pbphase1/blob/master/outputs/exthashtagurl.txt>



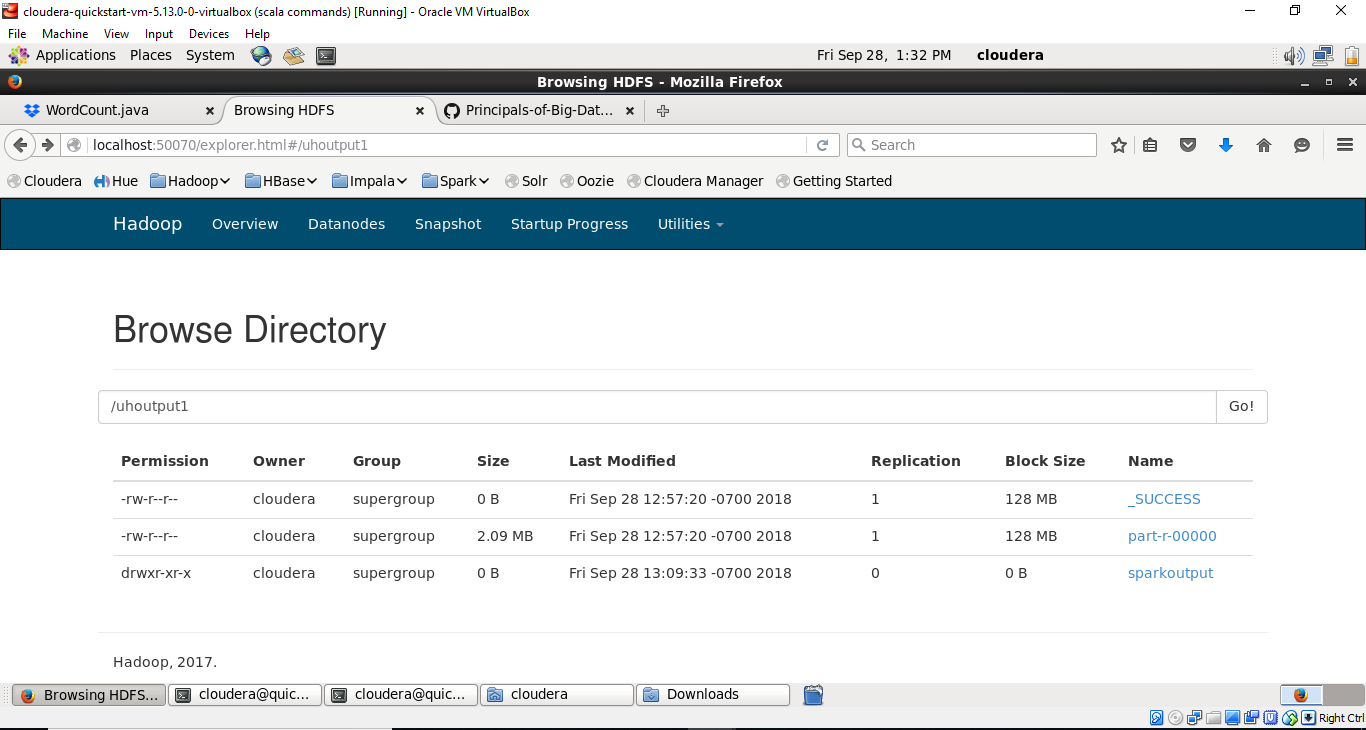
1. **Hadoop WordCount:**

We used hadoop-mapreduce-examples-2.9.0.jar file to run word count in hadoop . Before doing this , we need to create a directory inside hadoop(input) and then move your extracted hashtags file to hadoop by giving command ($HADOOP\_HOME/bin/hadoop fs –put /home/exthashtagurl.txt /input)

***Output:*** part-r-00000

<https://github.com/SumanthMedavarapu/pbphase1/blob/master/outputs/part-r-00000>

****

****

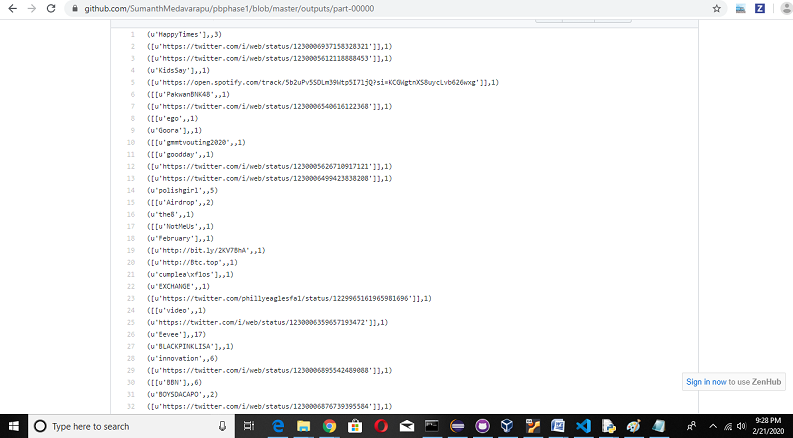
1. **Spark WordCount:**

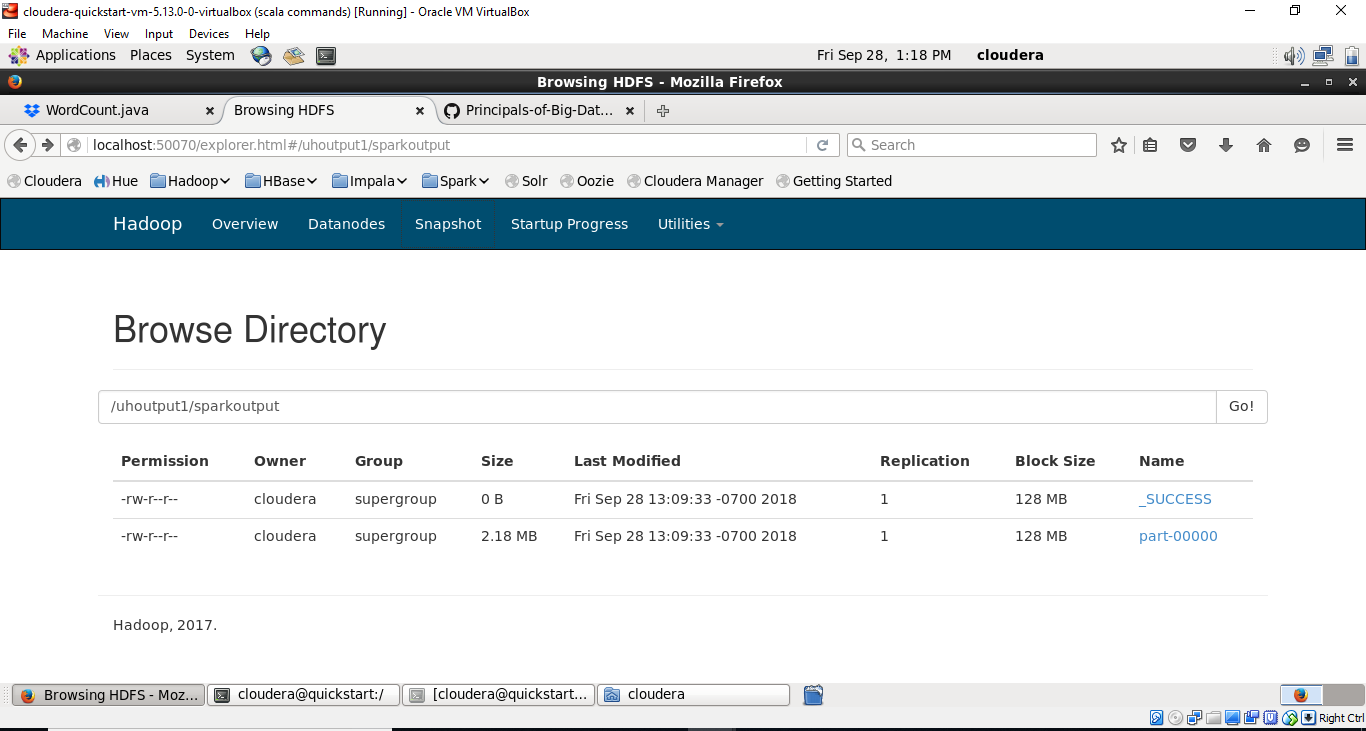
Extracted hashtags and URLs are given as input for this wordcount program.

***Input*:** exthashtagurl.txt

***Output:*** part-00000

<https://github.com/SumanthMedavarapu/pbphase1/blob/master/outputs/part-00000>

**

**

Github link: <https://github.com/sameer999/Principles-of-Big-Data-Management>

Tweets Output file: <https://drive.google.com/file/d/1DwNjNrKuOP9rZZn1tGaNc_rUD4t8SsRv/view?usp=sharing>

**References:**

1. <http://docs.tweepy.org>
2. <https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html>
3. <https://www.cloudera.com/documentation/enterprise/5-5-x/topics/spark_develop_run.html#wordcount__fig_png_bdj_h5>
4. <https://www.dezyre.com/apache-spark-tutorial/spark-tutorial>