**Principles of Big Data Management**

**Project Phase-1**

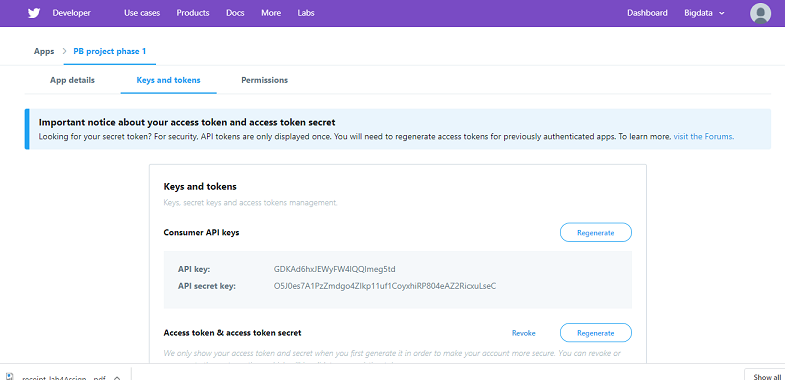
Harika Gurram – 16300892 – hghb6@mail.umkc.edu

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

Siva Sameer – 16297780 - sy9bd@mail.umkc.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.

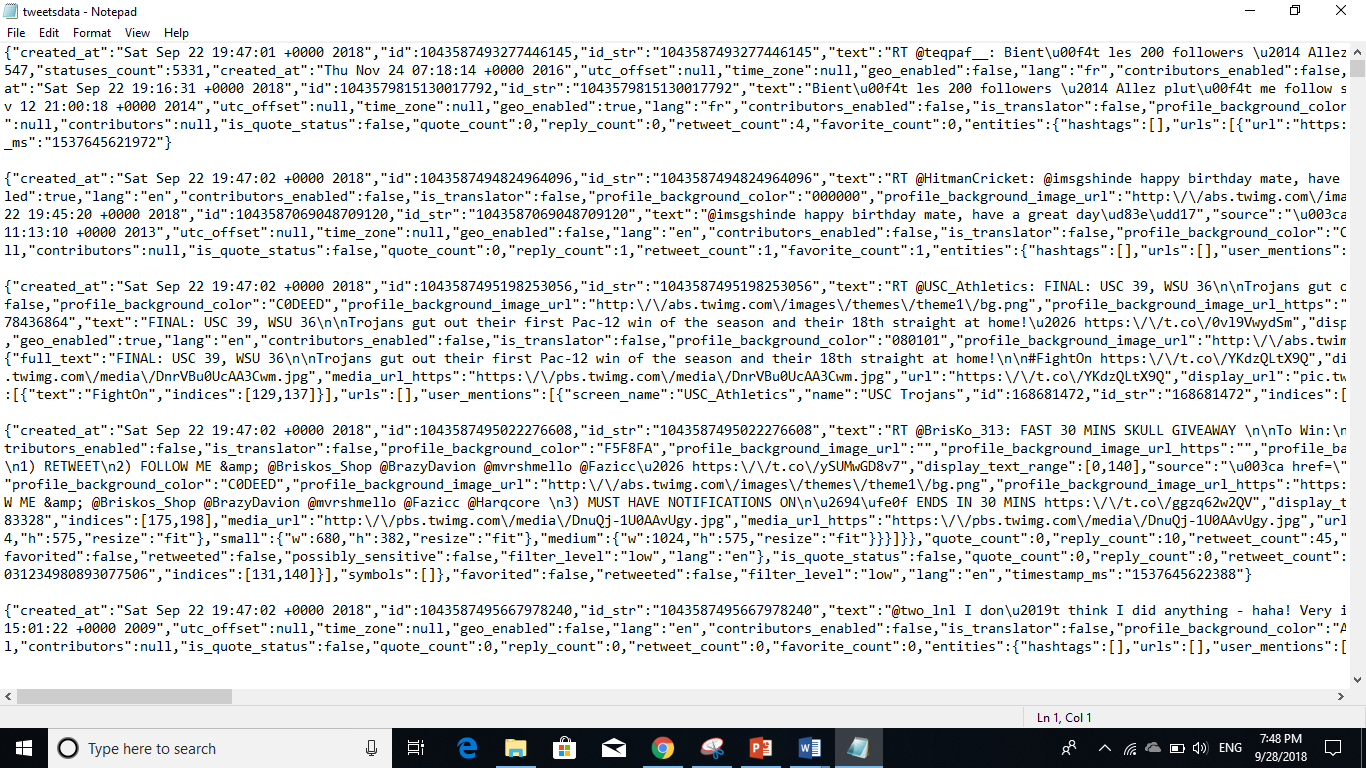
***No. of tweets collected:*** 1,45,000

***Source code****:* importing\_tweets.py

<https://github.com/sameer999/Principles-of-Big-Data-Management/blob/master/Source/importing_tweets.py>

***Output***: tweetsdata.txt

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



1. **Extracting Hashtags and URLs:**

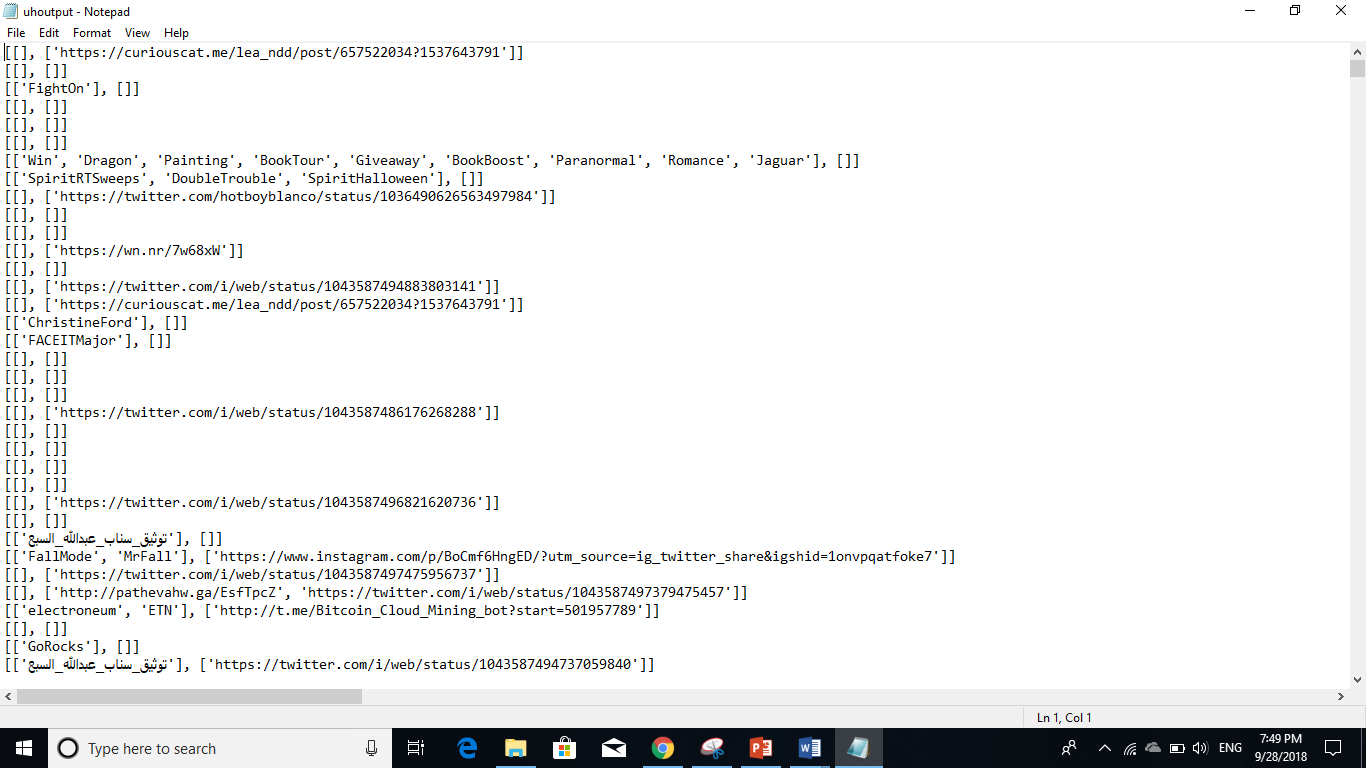
Hashtags and URLs are extracted from the imported tweets.

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

<https://github.com/sameer999/Principles-of-Big-Data-Management/blob/master/Source/extract_url_hash.py>

***Output:*** uhoutput.txt

<https://raw.githubusercontent.com/sameer999/Principles-of-Big-Data-Management/master/Outputs/uhoutput.txt>



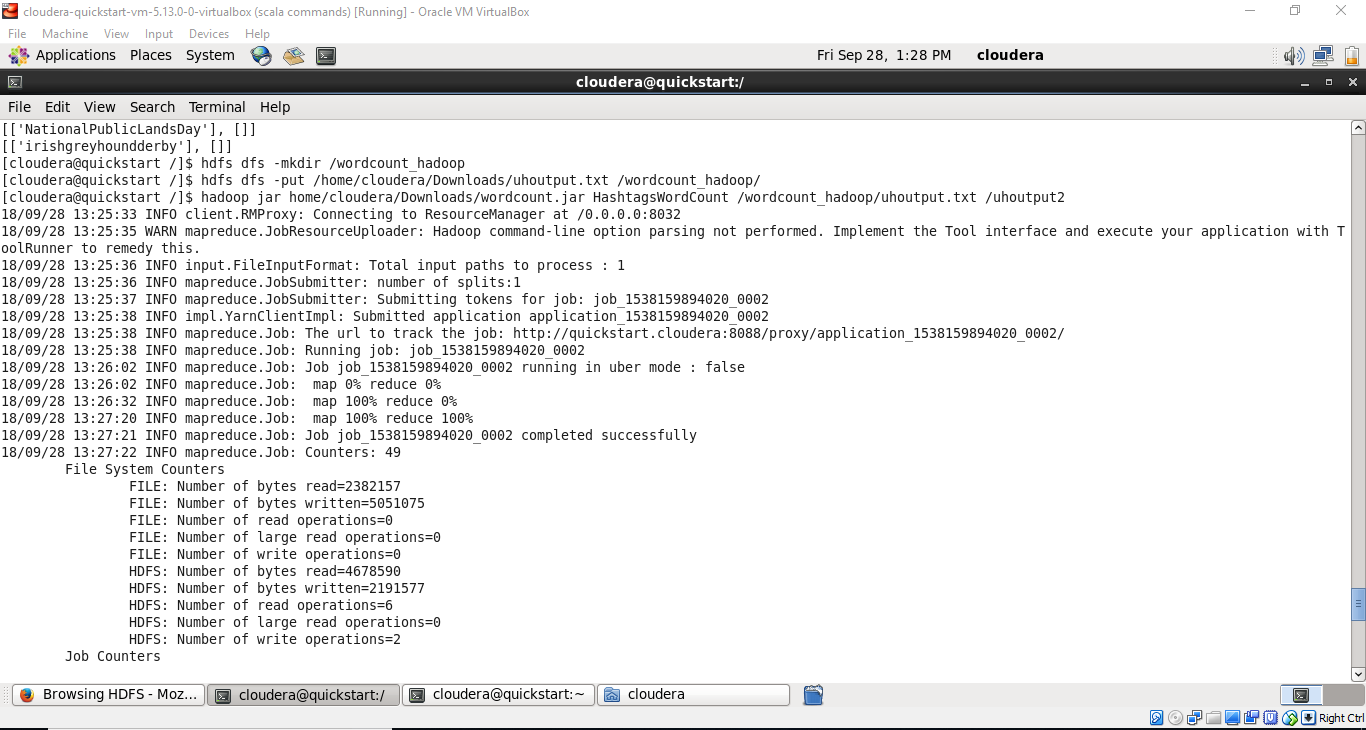
1. **Hadoop WordCount:**

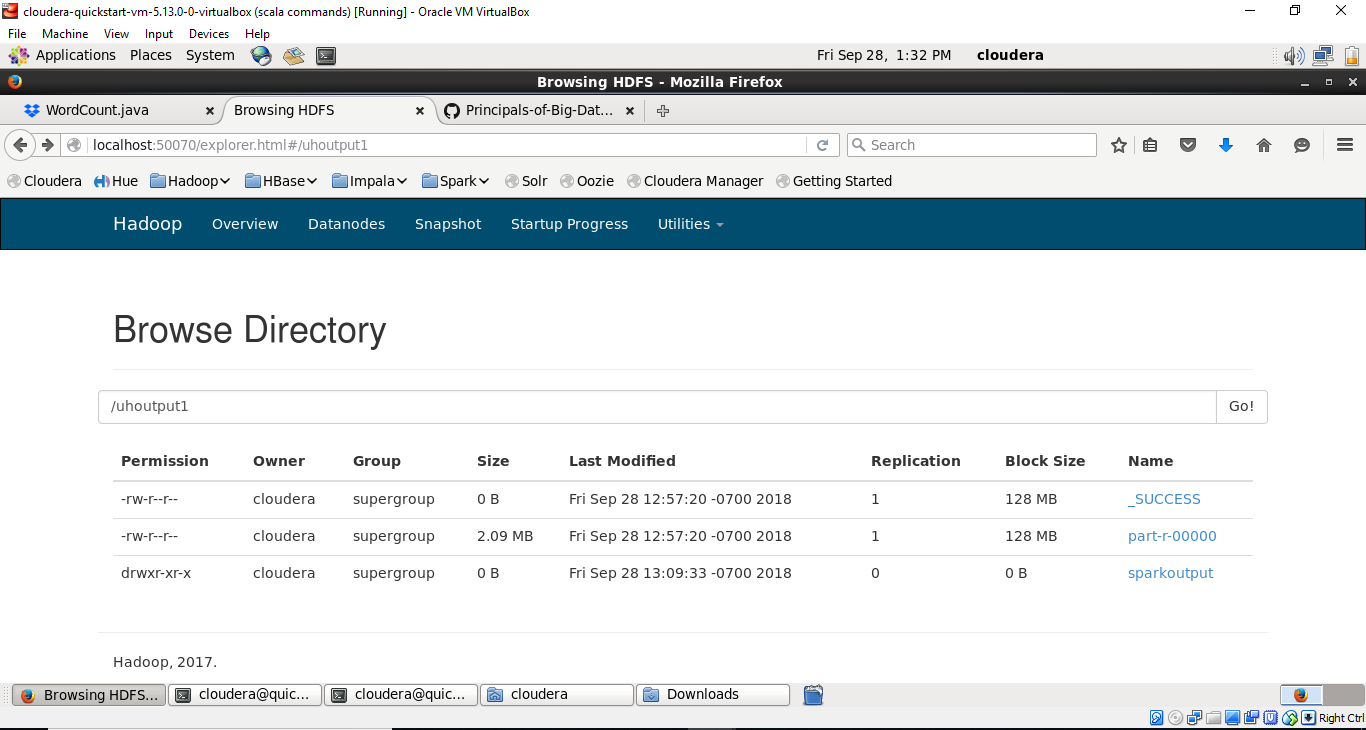
Wordcount.java is used for generating wordcount.jar to execute wordcount program on the extracted Hashtags and URLs.

<https://github.com/sameer999/Principles-of-Big-Data-Management/blob/master/Source/wordcount.java>

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

<https://raw.githubusercontent.com/sameer999/Principles-of-Big-Data-Management/master/Outputs/part-r-00000.txt>

****

****

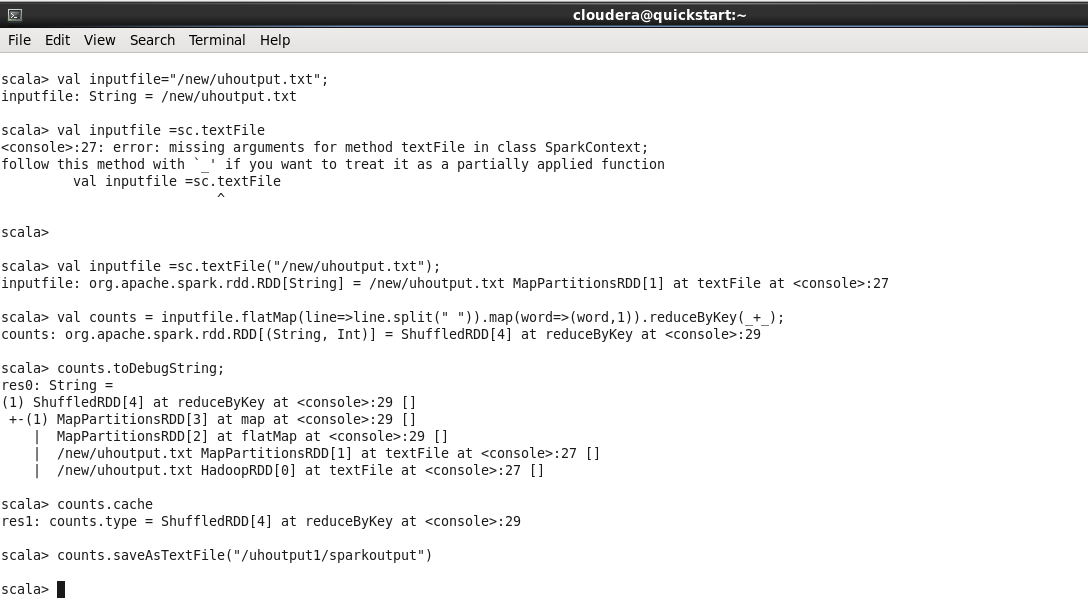
1. **Spark WordCount:**

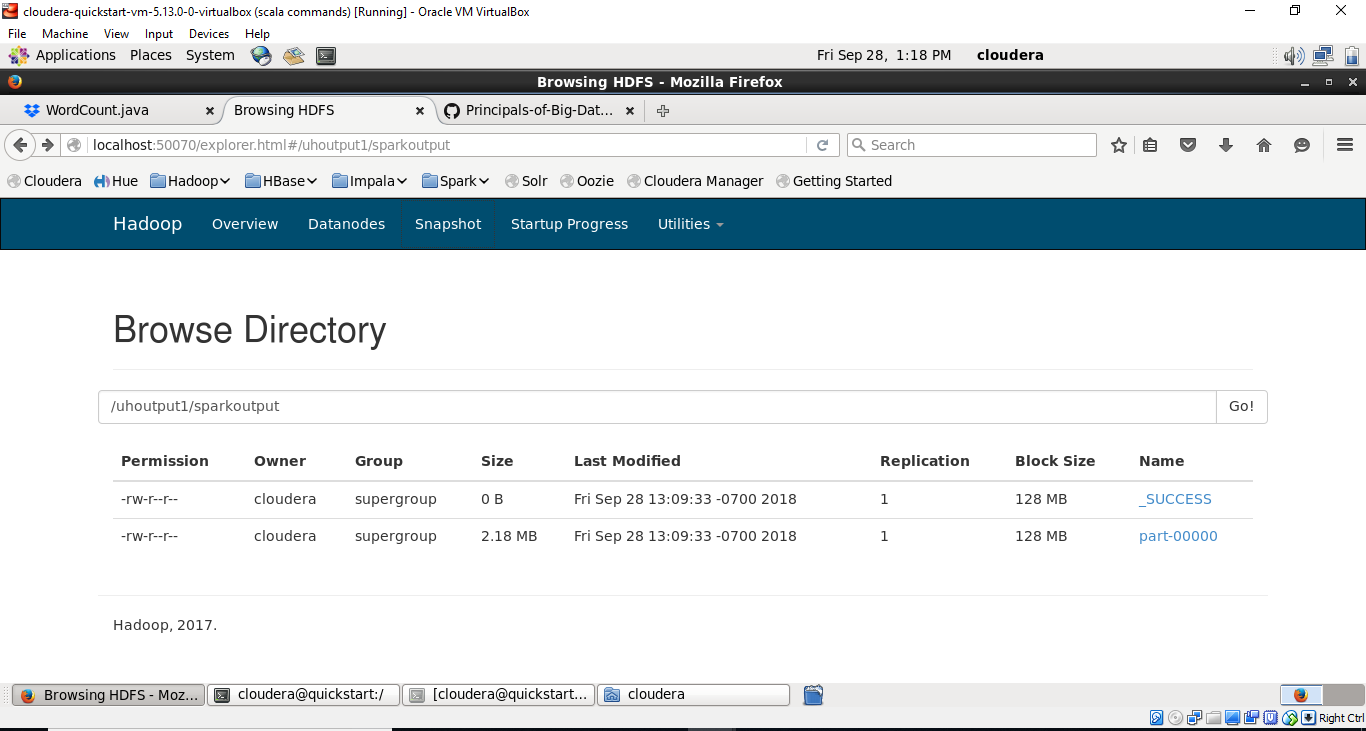
Extracted hashtags and URLs are given as input for this wordcount program.

***Input*:** uhoutput.txt

***Output:*** part-00000

[*https://raw.githubusercontent.com/sameer999/Principles-of-Big-Data-Management/master/Outputs/part-00000.txt*](https://raw.githubusercontent.com/sameer999/Principles-of-Big-Data-Management/master/Outputs/part-00000.txt)

**

**

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>