**Principles of Big Data Management**

**Project Phase-1**

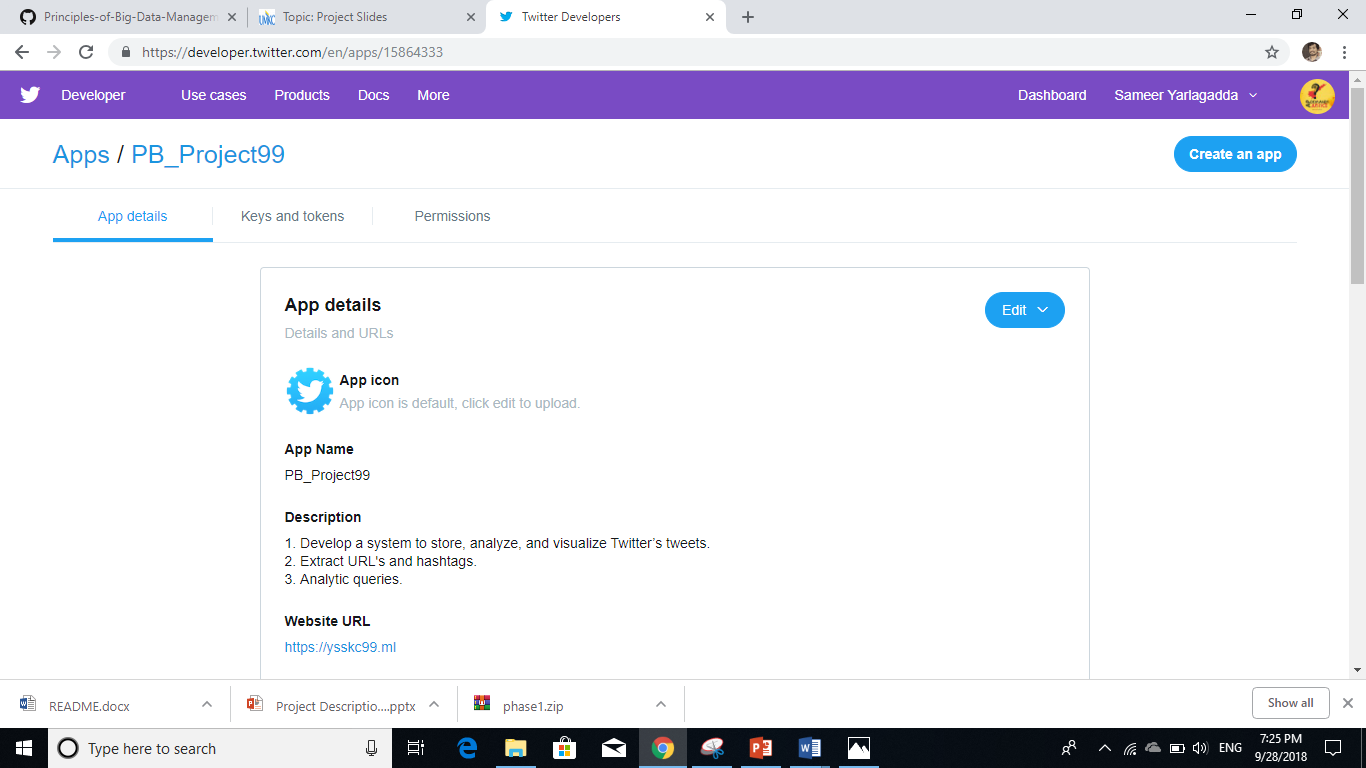
Adisekhar Raju Uppalapati – 16274435

Mahesh Mandava – 16274437

Siva Sameer Krishna Yarlagadda – 16271011

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

After creating an app in the Twitter developer’s account, user credentials for accessing the Twitter streaming API will be generated.



1. **Importing Tweets:**

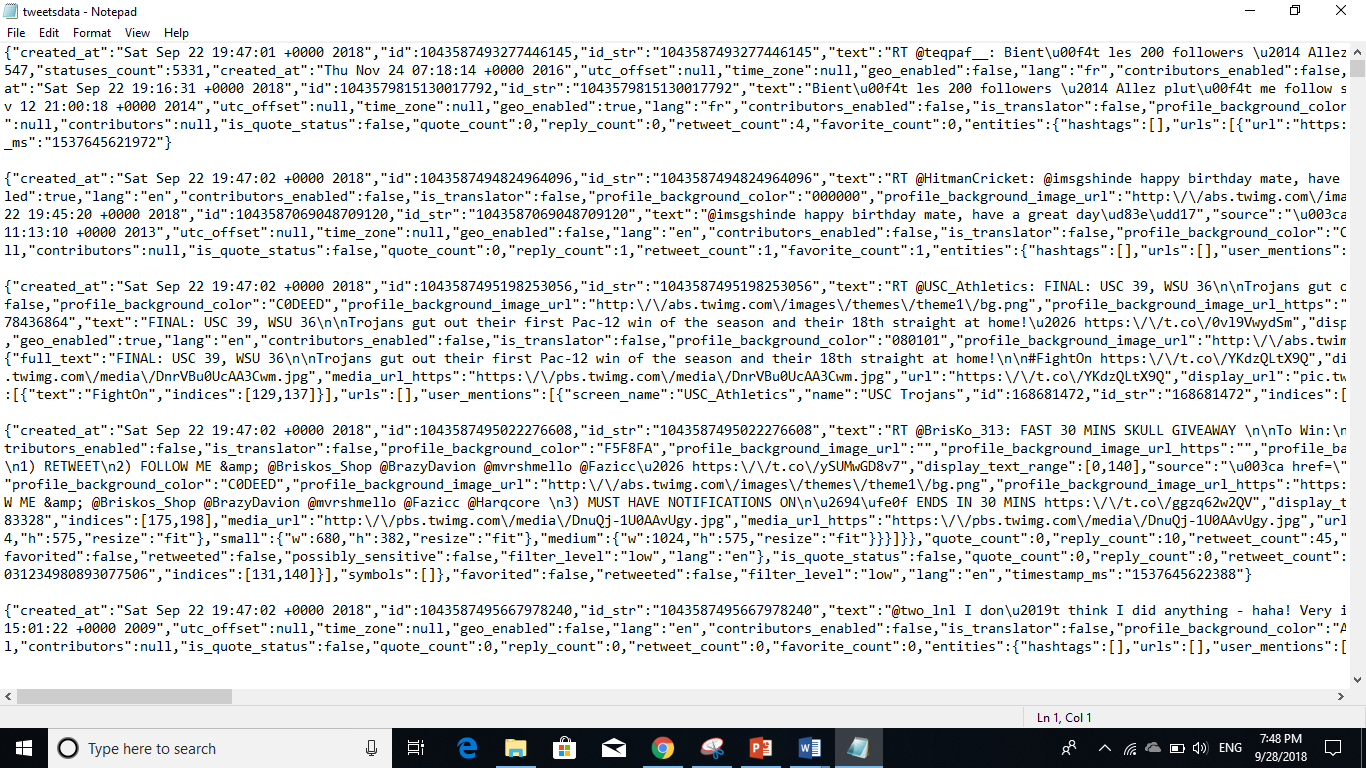
Tweepy and app secret keys are used for communication with Twitter platform and for importing the tweets.

***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



1. **Extracting Hashtags and URLs:**

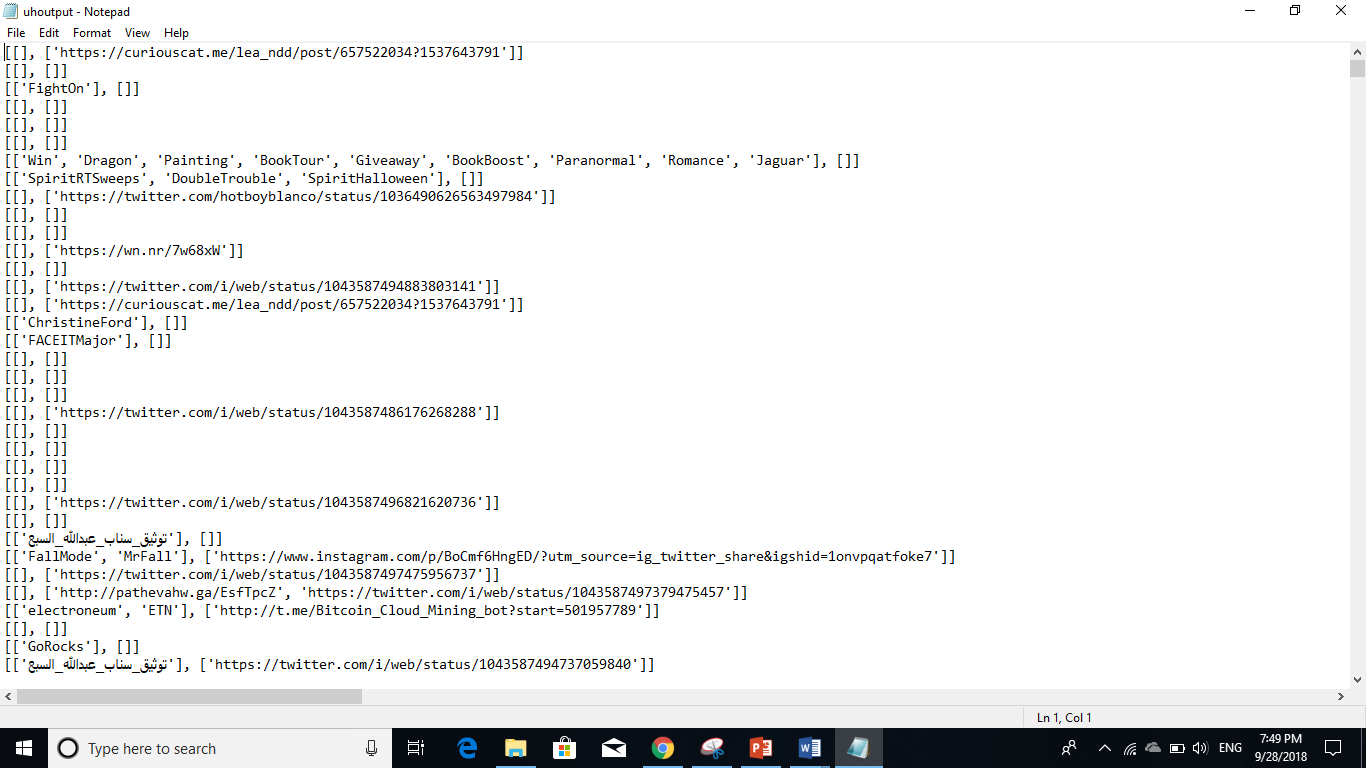
Hashtags and URLs are extracted from the imported tweets.

***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>



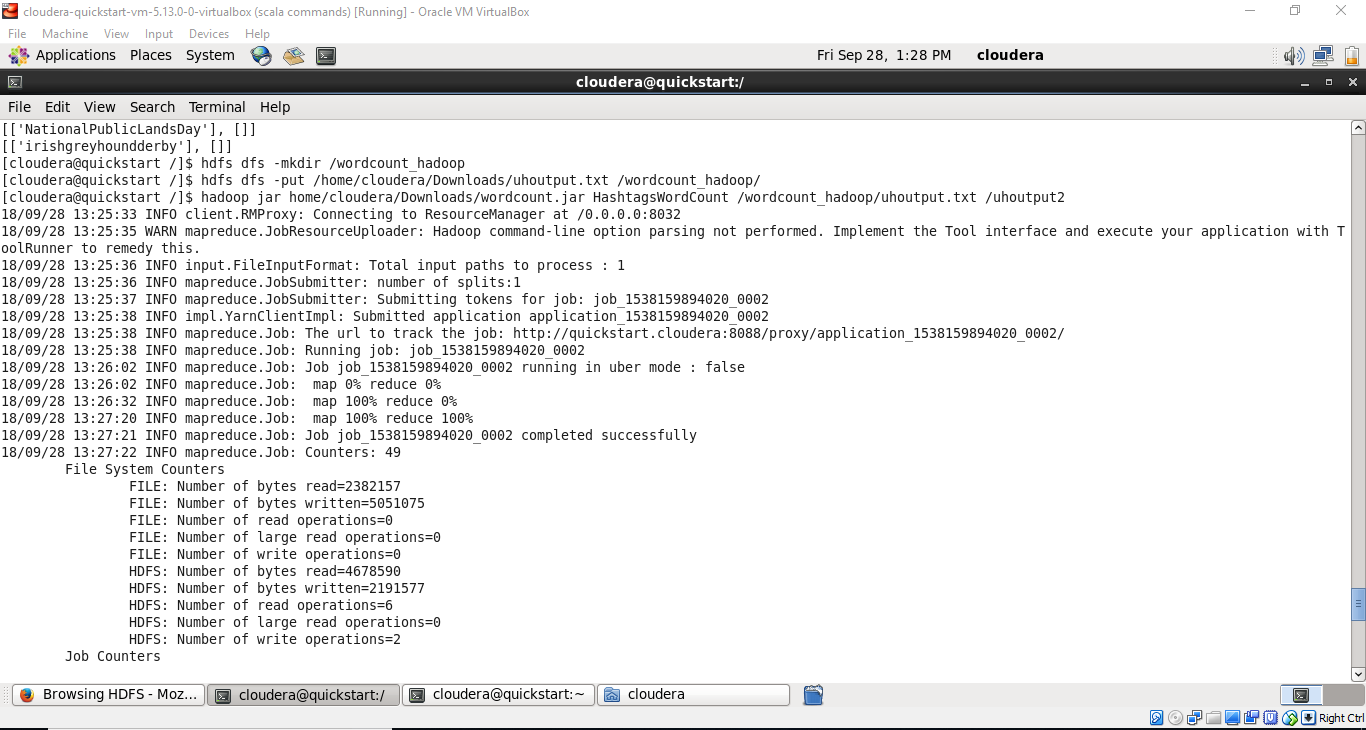
1. **Hadoop WordCount:**

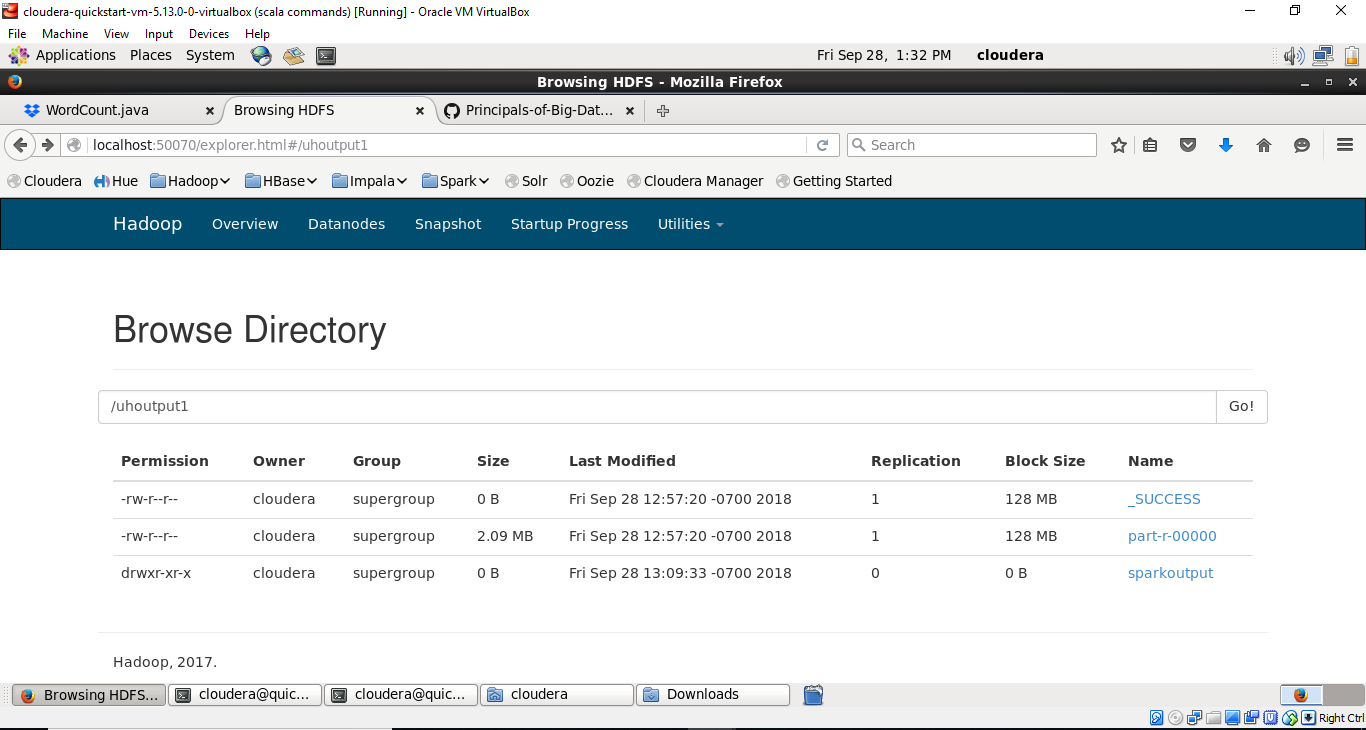
Wordcount.java is used for generating wordcount.jar to execute wordcount program.

<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>

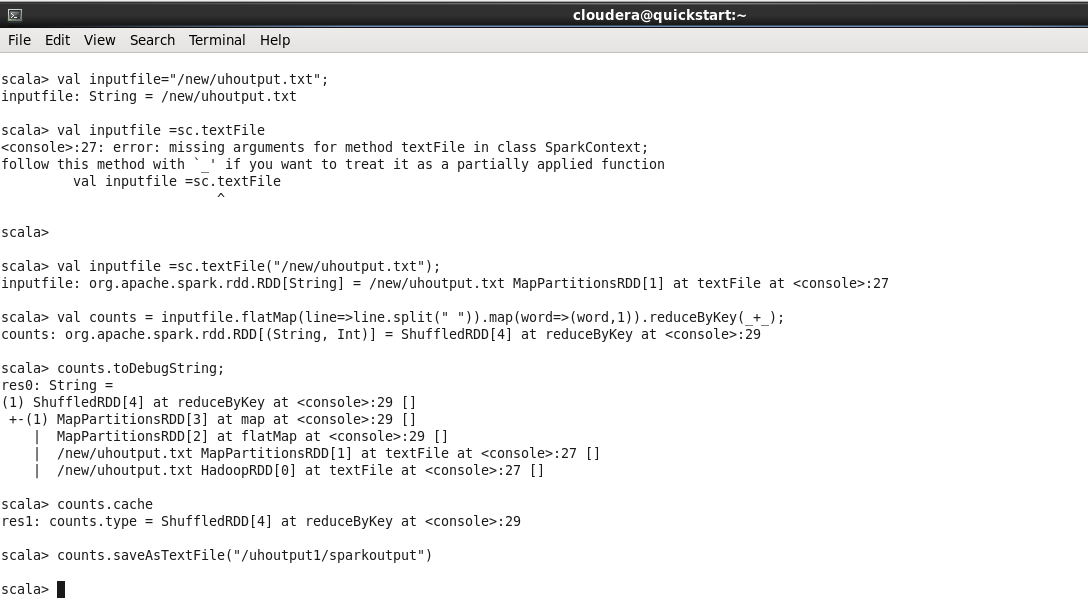
****

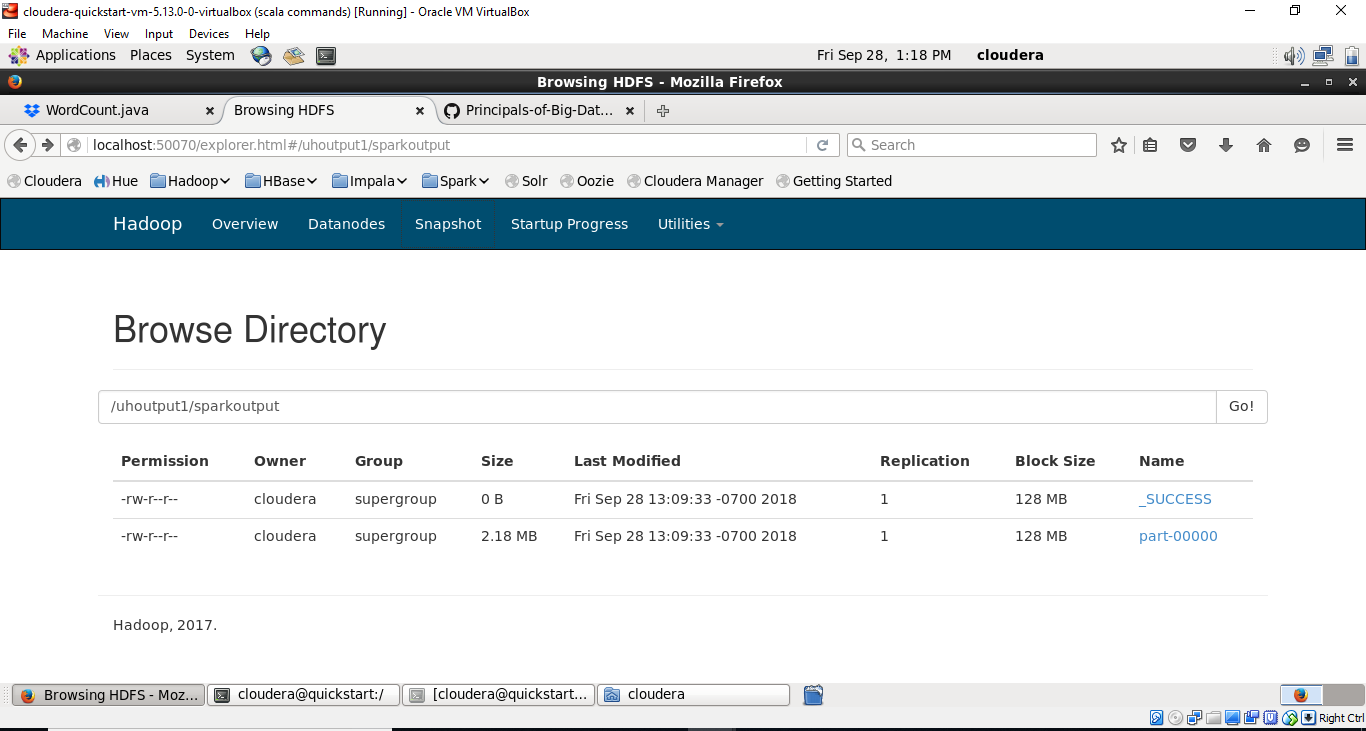
****

1. **Spark WordCount:**

***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://mailmissouri-my.sharepoint.com/:t:/g/personal/sy9bd_mail_umkc_edu/EXGKs9wSx9VMpD2B8v-aujAB6m8rLJubiXq38_-YV1163g?e=c0bI3I>