

**Principles of Big Data Management**

**Phase-1**

Gente Harun Sai Kumar – 16294902 – hgkfw@mail.missouri.edu

Bichinepally Rahul Babu – 16300459 – rbkpb@mail.missouri.edu

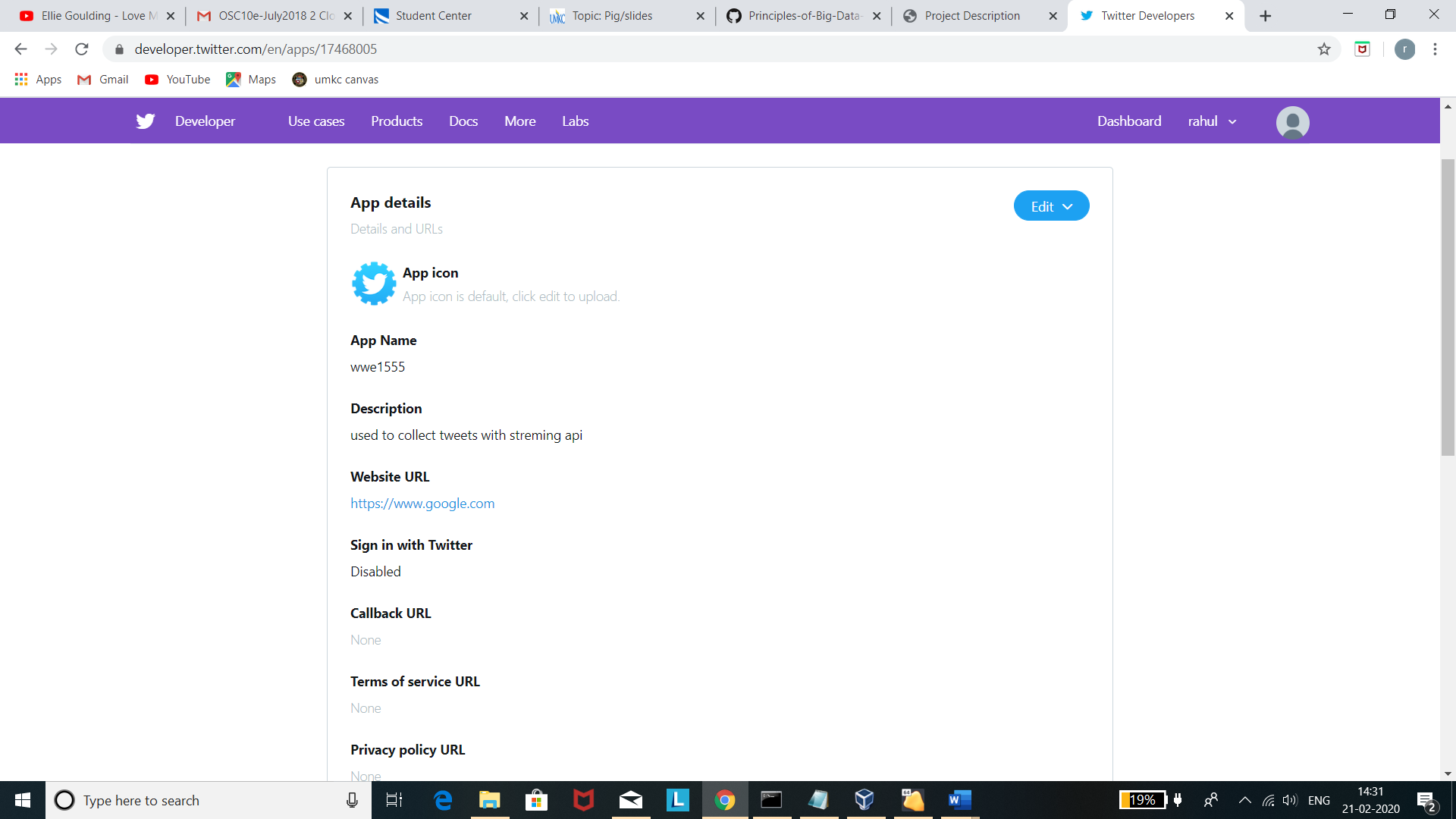
Sarika Reddy Kota -16302630-skrqr@mail.missouri.edu

**APPROACH :**

**Extraction Of Tweets:**

**1.Create and Log in to Twitter Developer account:**

Create a twitter developer account, User credentials to access the Twitter streaming API will be generated.



**2.Tweet extraction**

With the help of generated tokens and tweepy library we have extracted the tweets from twitter database.

No.of tweets collected: 146465

Source code: tweet\_extraction.py

Output: <https://github.com/Harun2703/Principles-of-Biggdata/blob/master/Phase%201/tweet_extraction.py.py>

**3.Extraction of Hashtags and URLs:**

Imported tweets data is used to extract the Hashtags and URLs.

Source Command for hashtags: jq .entities.hashtags[].text /home/osc/PB/wwe.json>hashtags\_wwe.txt

Source Command for URL’s : jq .entities.url[].url /home/osc/PB/wwe.JSON>urls\_wwe.txt

Output: <https://github.com/Harun2703/Principles-of-Biggdata/tree/master/Phase%201/Output/hashtags%20and%20URLs%20extracts>

**4.Hadoop Wordcount:**

With Hadoop installed, we have run the wordcount on the extracted Hashtags and URL’s

Command :

1)hdfs dfs -mkdir /wordcount\_hadoop

2)hdfs dfs -put /home/osc/PB/hashtags\_wwe.txt /wordcount\_hadoop/

3)Hadoop jar $HADOOP\_HOME/share/hadoop/mapreduce/Hadoop-mapreduce-examples-2.8.5.jar wordcount /wordcount\_hadoop/hashtags\_wwe.txt /output\_v2

[same for URL wordcount]

Output: <https://github.com/Harun2703/Principles-of-Biggdata/tree/master/Phase%201/Output/hadoop_wordcount>

**5.Spark Wordcount:**

Hashtags and URLs which are extracted from the tweets are given as inputs to Spark Wordcount.

**Command** :

1. $SPARK\_HOME/bin/spark-submit run-example JavaWordCount /home/osc/PB/urls\_wwe.txt
2. $SPARK\_HOME/bin/spark-submit run-example JavaWordCount /home/osc/PB/hashtags\_wwe.txt

**Output:**

<https://github.com/Harun2703/Principles-of-Biggdata/tree/master/Phase%201/Output/spark_wordcount>

**REFERENCES:**

1. <https://www.guru99.com/how-to-install-hadoop.html>
2. <https://data-flair.training/blogs/install-spark-ubuntu/>