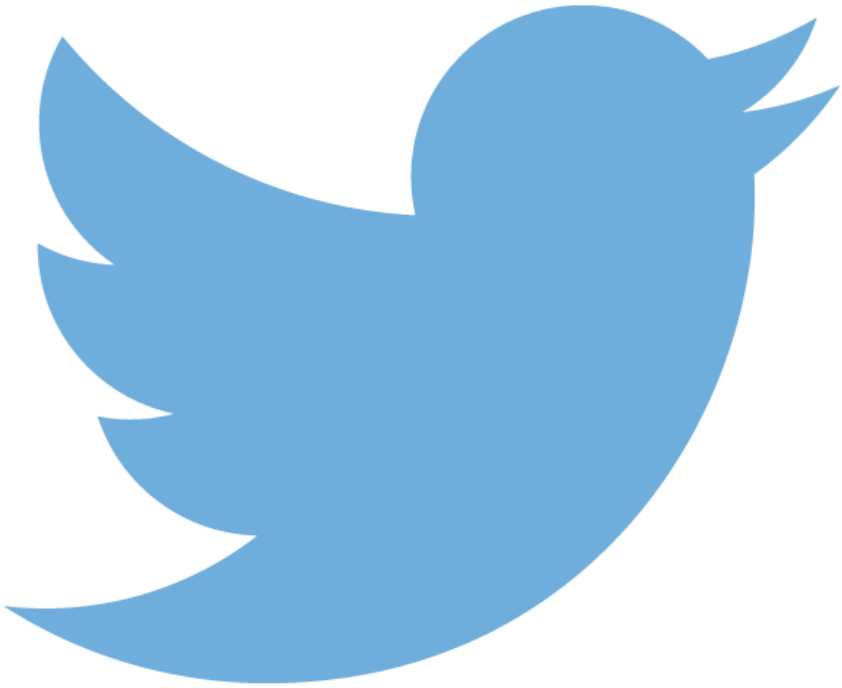
Principles of Big Data Management

Project 2



Team

1. Saketh Garuda (sg7kf)
2. Mudunuri Sri Sai Sarat Chandra Varma (smyx4)
3. Yalamanchili Sowmya (syb7c)
4. Nandanamudi Sreelakshmi (snhnc)

1.Introduction

The main objective of this project is analyzing the twitter data about 'Politics' category and analyze the data using Hadoop MapReduce. After collecting, the necessary tweets data we have used python code to retrieve text from the tweets. Based on the text we have calculated the list of words in the tweets text file which includes the list of duplicates and unique words in the tweets text file using MapReduce program and also calculated the ratio of number of unique words and to the number of duplicate words. We also implemented MapReduce program to return the top ten best times to post a tweet on twitter.

The entire document gives the walk through and scope of the working environment of the project.

2.System Requirements

Software Requirements:

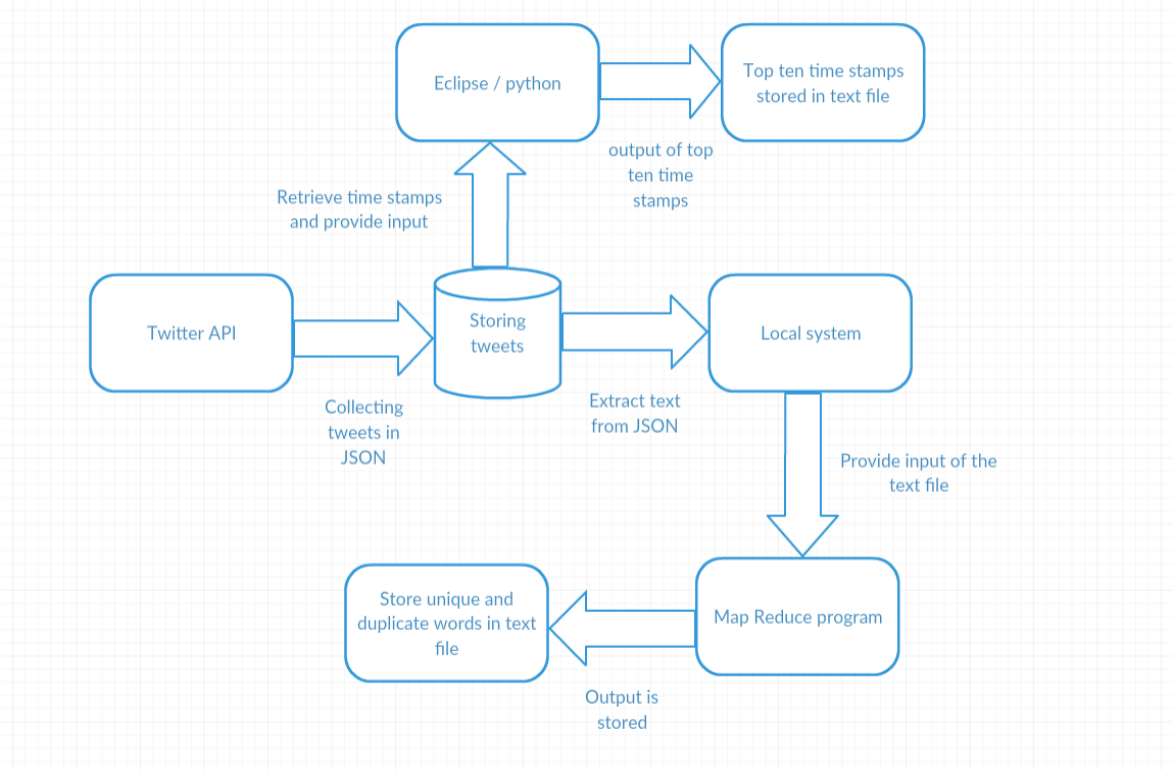
- Eclipse
- Net Beans
- Apache Hadoop V 2.6.0-cdh5.8.0
- JDK 1.8

Programming Languages:

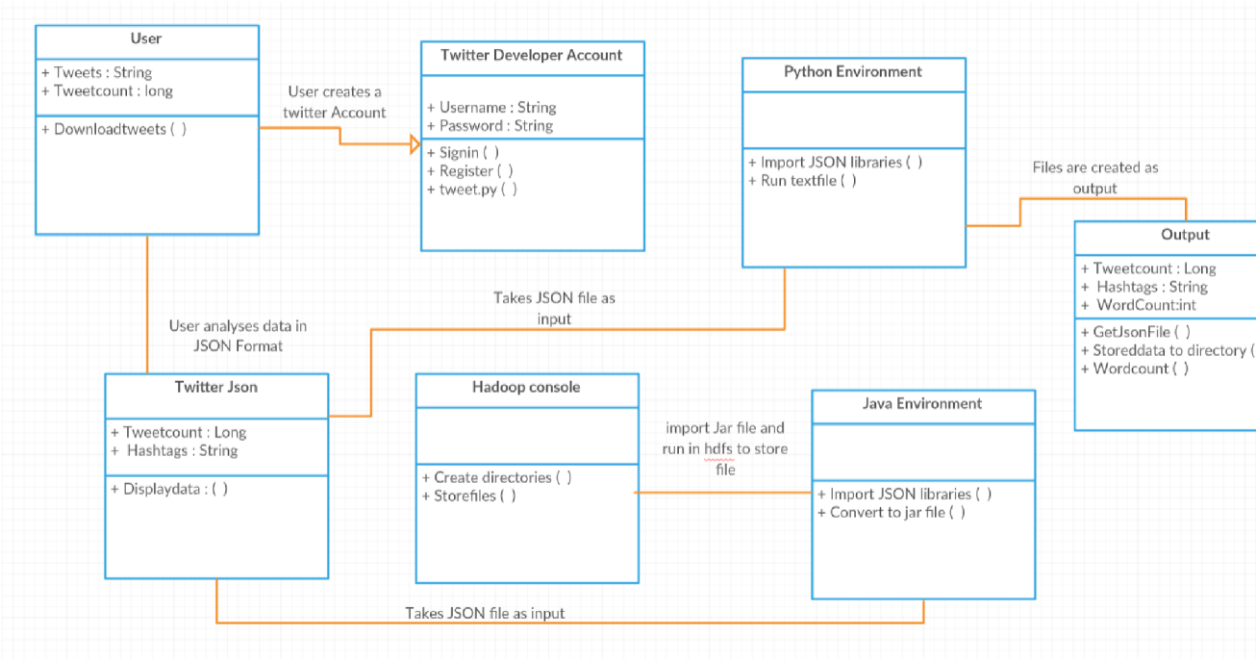
- Java
- Python

3.Software Architecture and Design

- Twitter tweets are extracted using the API by running python program using the twitter tokens. The collected tweets are stored in JSON format in the local system.



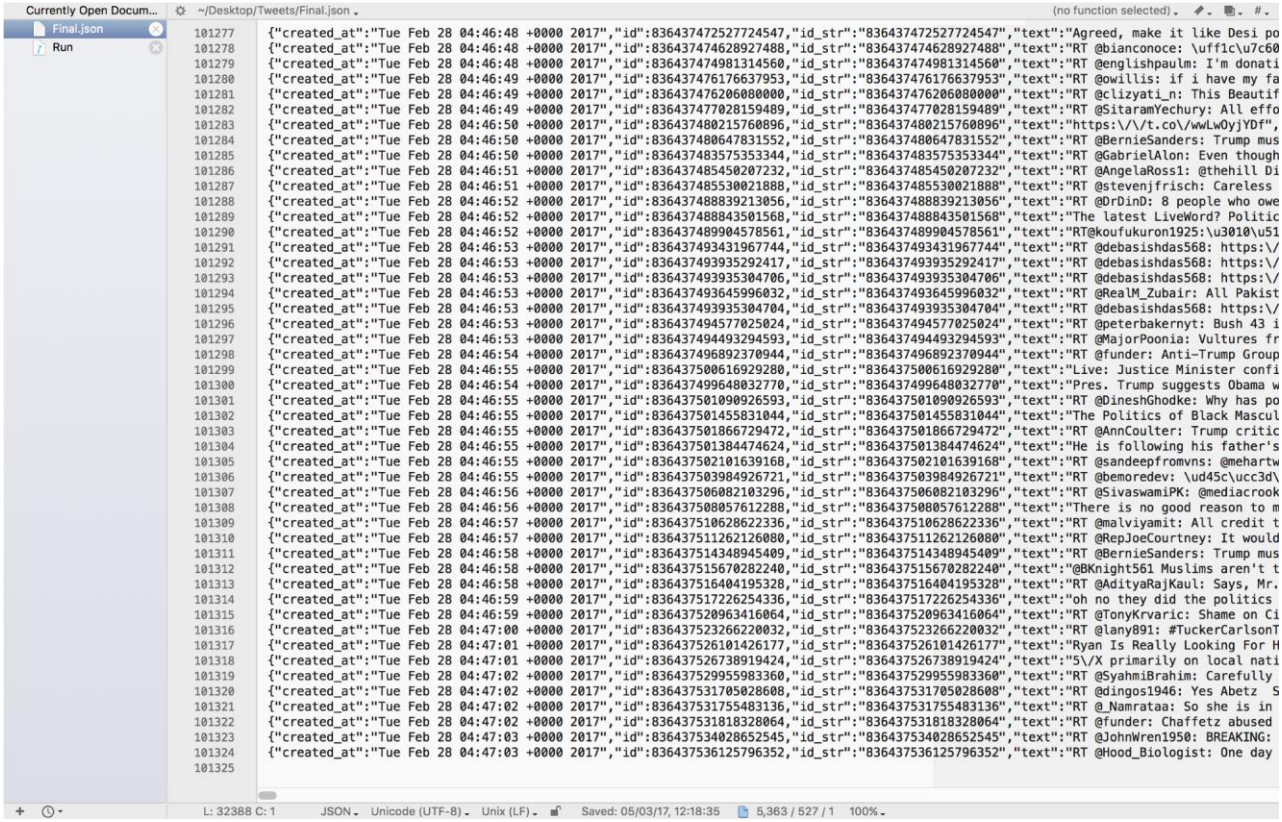
- Here is the class diagram for the project



4.Main Requirements

Collect tweets in JavaScript Object Notation

- We've collected 100k+ tweets related to category 'Politics' into a JSON file using the twitter API. JSON file is used for future work in retrieving the text and top 10 best times from the entire list of tweets. Here is the screenshot of JSON file,



- After collecting the 100k tweets, we've written a python code for separating text from tweets file and it is stored in tweet text file.

- Below screenshot indicates the python code for retrieving text from the tweets file. We have given the input **"Tweets.txt"** file and parsed the JSON data with **"tweet['text']"** and stored it an output file.

TweetText.py - C:\Users\sreel\Desktop\PBProject2\Main Requirements\TweetText.py (3.6.0)

File Edit Format Run Options Window Help

```
# Import the necessary package to process data in JSON format
try:
    import json
except ImportError:
    import simplejson as json

# We use the file saved from last step as example
tweets_filename = 'Tweets.txt'
tweets_file = open(tweets_filename, "r")

for line in tweets_file:
    try:
        # Read in one line of the file, convert it into a json object
        tweet = json.loads(line.strip())
        if 'text' in tweet: # only messages contains 'text' field is a tweet

            print (tweet['text']) # content of the tweet

    except:
        # read in a line is not in JSON format (sometimes error occured)
        continue
```

This is the input tweets file

This is the criteria to retrieve the content of the tweet

Ln: 1 Col: 0

- Here is the screenshot of output TweetsText file

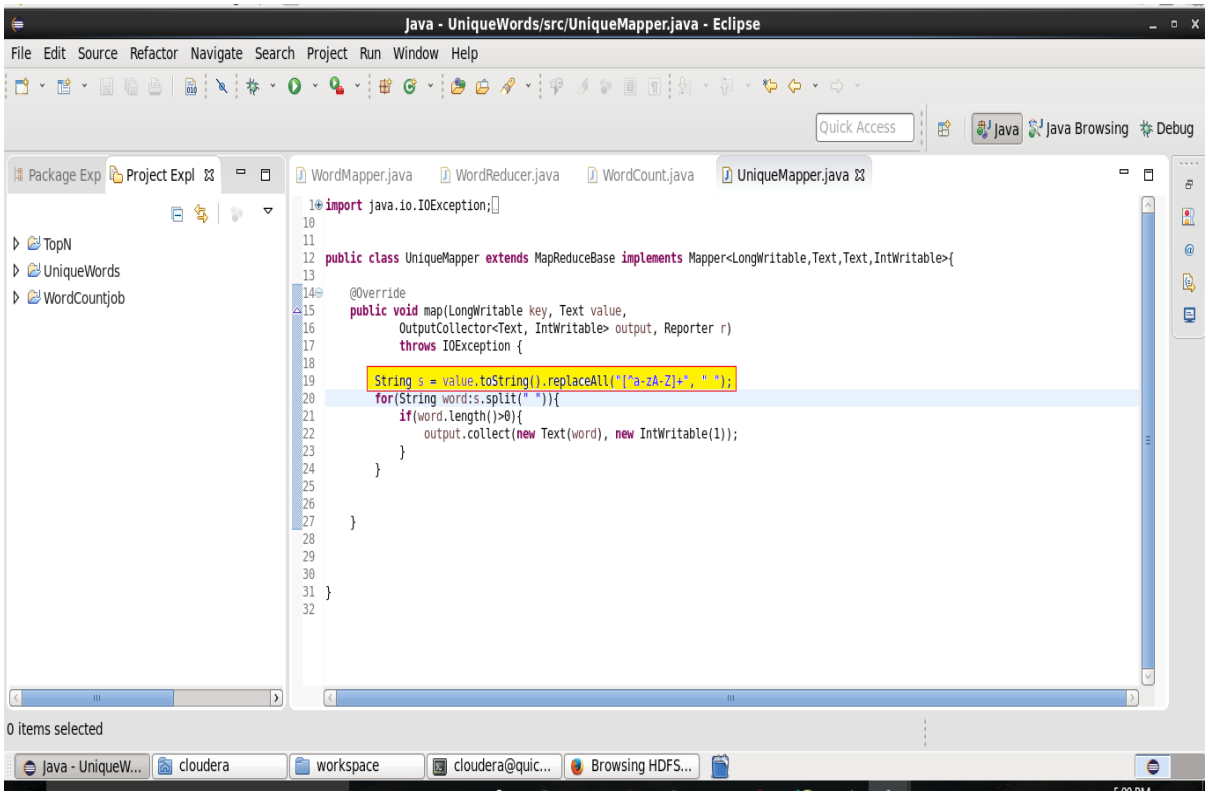
tweettext - Notepad

File Edit Format View Help

```
RT @Del_ivered: John Boothman, Johnny-come-lately to @theSNP has form in bullying claims,so ideal for SNP,his politics have changed. https:…
"I'm sorry, I don't want to see him"-father of #NavySeal killed in #Yemen re: Trump as son's casket arrived in U.S.… https://t.co/Ad2IA9tDCE
@HackettTerence @gregsteube @NRA https://t.co/3AAXMWSL8w
RT @mikiebarb: "He is shocked that he is not in control of the press." Must read from @GlennThrush &amp; @grynbaum https://t.co/NoNtDp7UFj
RT @ALT_DOJ: Everyone keep an eye on @ossoff if you live in GA get out and help if you can! https://t.co/AGltnfMxn3
RT @iSupportPTI: @DunyaNews @betterpakistan @WajSKhan Panama Money Laundring Network telling Imran Khan how to do politics.We are sick of S…
RT @MarcusC22973194: BREAKING AND URGENT
JASON CHAFFETZ NOT ONLY TAKES RUSSIAN MONEY, HE IS ALSO LINKED TO TRUMP THROUGH CHINA…
RT @KeithOlbermann: This Senator has been an amoral jackass from day one. https://t.co/q6Cqo2oAxx
RT @camboviet: Chelsea Clinton Attends ‘I am Muslim Too’ Protest in NYC https://t.co/aiiy2Txnyk
RT @ShaunKing: Will Republicans investigate the failed Yemen raid like they did Benghazi?

Family of slain SEAL is demanding it.

https://…
RT @ShaunKing: Will Republicans investigate the failed Yemen raid like they did Benghazi?

Family of slain SEAL is demanding it.

https://…
RT @funder: Chaffetz abused his power to try &amp; reverse a law—in WASHINGTON DC. HE NEEDS TO RESIGN.

#russiagate #trumprussia #rt https://t.…
RT @yashar: Trump now has his own version of a Benghazi mom. The father of the Navy Seal killed in Yemen wants answers.…
@JoeNBC Let's see Sanders campaign emails. Politics is unfair &amp; not for the faint hearted-Sanders did nada in 36 years to form 3rd party
Farage SLAMS May for not showing support to Le Pen https://t.co/doz43ueIc2
RT @ShaunKing: 1. William Owens was so disgusted w/ Trump's Yemen raid that killed his son, Ryan, that he refused to meet w/ Trump

https:/…
RT @ZekeJMiller: Slain SEAL’s dad wants answers: ‘Don’t hide behind my son’s death’ https://t.co/wFEHNfZagR
RT @DoctorRobin: Donald Trump advisor Sheriff David A Clarke was in Russia while Mike Flynn dined with Putin https://t.co/33eDtmA3y7 #Trump…
RT @tongs_ya_bass: The most divisive part of Scottish politics is when English politicians travel north to tell us we're divisive! #SadiqKh…
RT @jaketapper: Father of SEAL killed in Yemen raid spurned meeting with Trump, wants answers | Miami Herald https://t.co/7mmbWfmu1d
Slain SEAL’s dad wants answers: ‘Don’t hide behind my son’s death’ https://t.co/oQbBc5PAEt
https://t.co/dlDICcQwCK
RT @europa_nexus: Hedge-fund billionaire and Donald Trump backer 'played key role in Brexit campaign' https://t.co/Q5PfMoSNNW
How fascist facebook account data helped the right wing know they would win the election:

https://t.co/hfsJEZbzBP
RT @SenCortezMasto: 21 female U.S. senators is not enough. Women need more seats at the table. Period.

My interview w/ @ELLEmagazine: ht…
RT @KeithOlbermann: This Senator has been an amoral jackass from day one. https://t.co/q6Cqo2oAxx
```

- We have implemented the Hadoop MapReduce in java code to find the list of duplicate and unique words. The words that occurred more than once are considered as duplicate words and the words which occurred once are considered as unique.

- We have filtered words in the text file using the below code which removes the special characters and split them with whitespace delimition. The below screenshot refers the same:



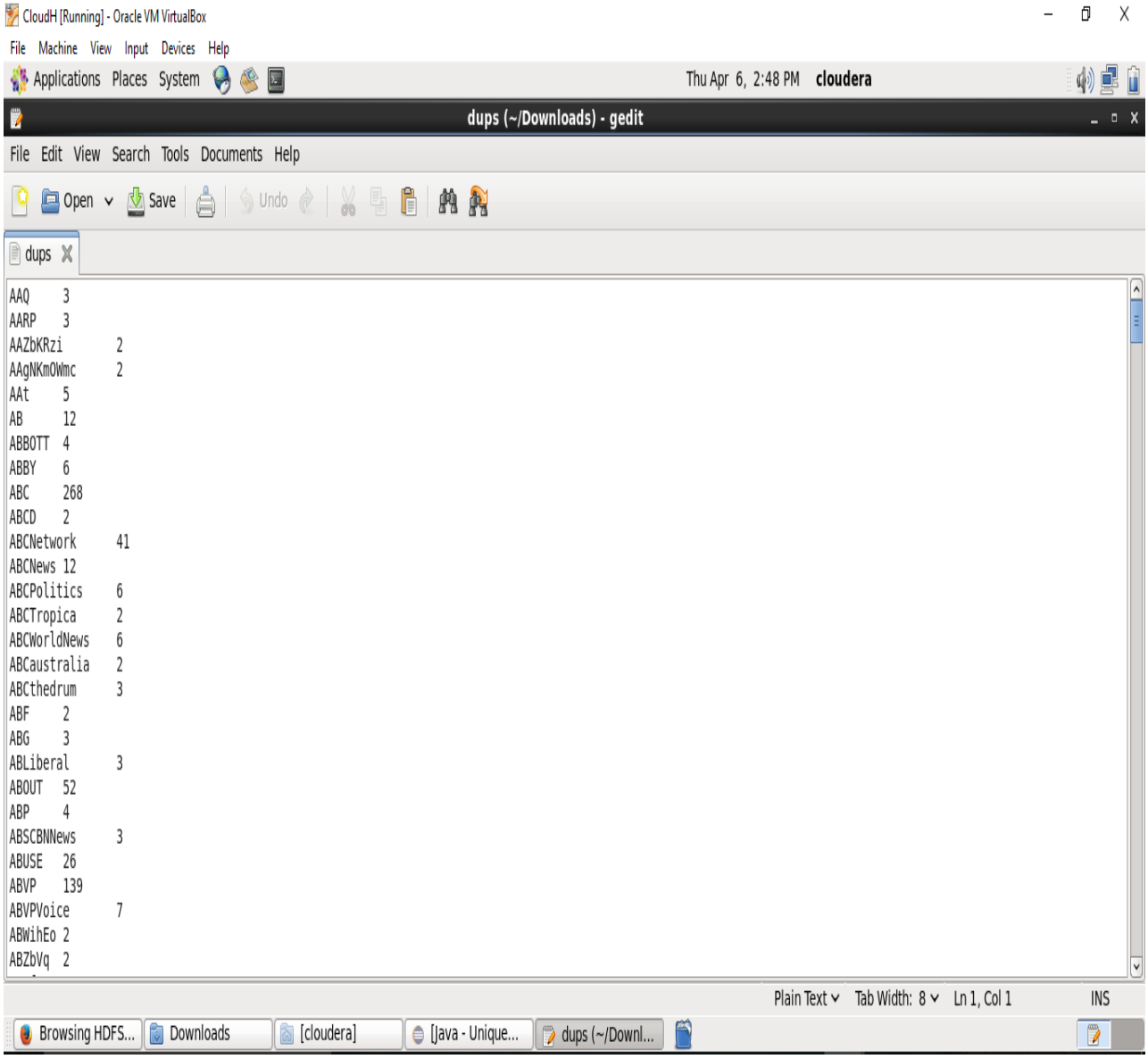
For Duplicate Words:

We have exported the java MapReduce program to jar file and run in the Hadoop environment to get the list of duplicate words. By using the command:

hadoop jar Duplicates.jar WordCount /user/cloudera/test/tweettext.txt /user/cloudera/test/dups.txt

The above hadoop command takes the Duplicates jar file with the class name that has main method i.e **WordCount** with input path **"/user/cloudera/test/tweettext.txt"** and output path. **"/user/cloudera/test/dups.txt".**

Below screenshot shows the list of duplicate words which are stored in the dups text file the count on right of the words shows that they are duplicated as they occurred more than once.

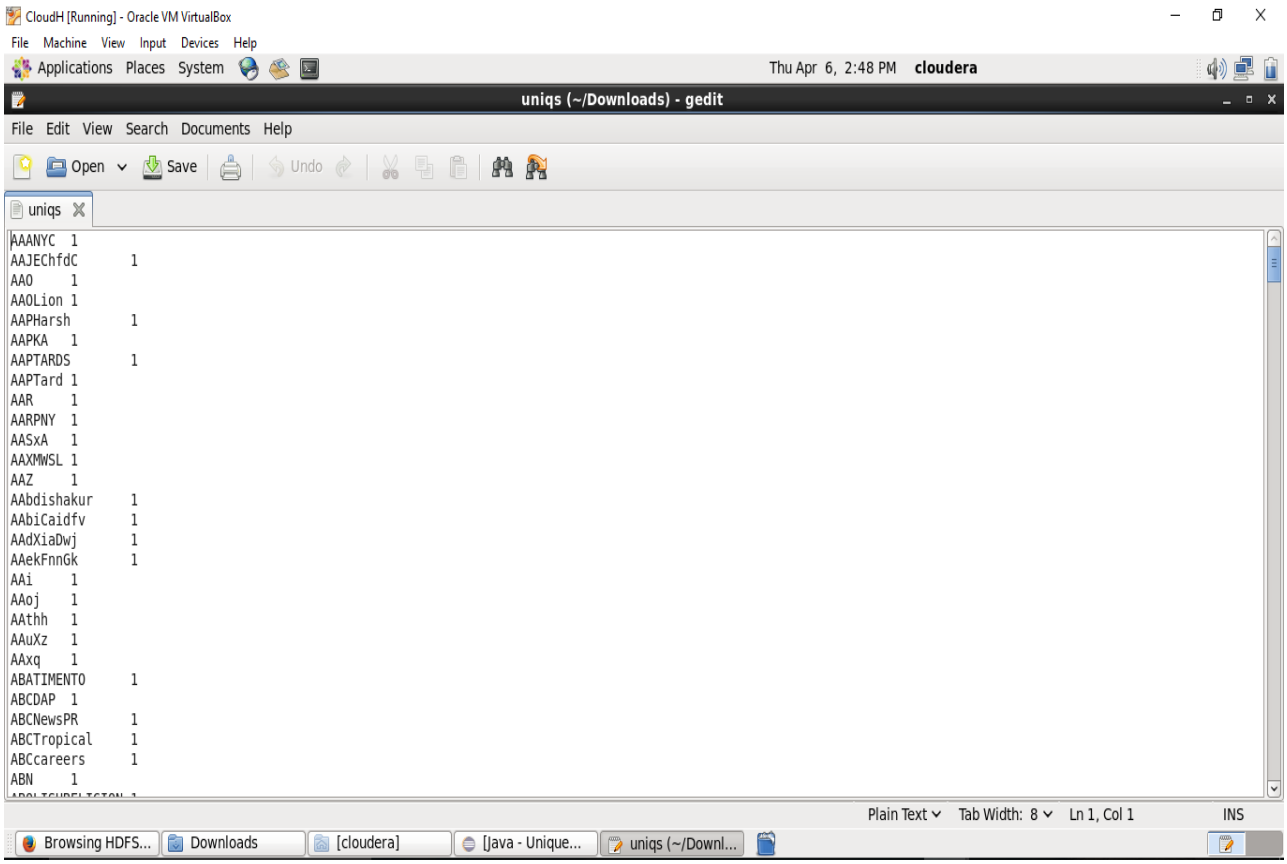


For Unique Words:

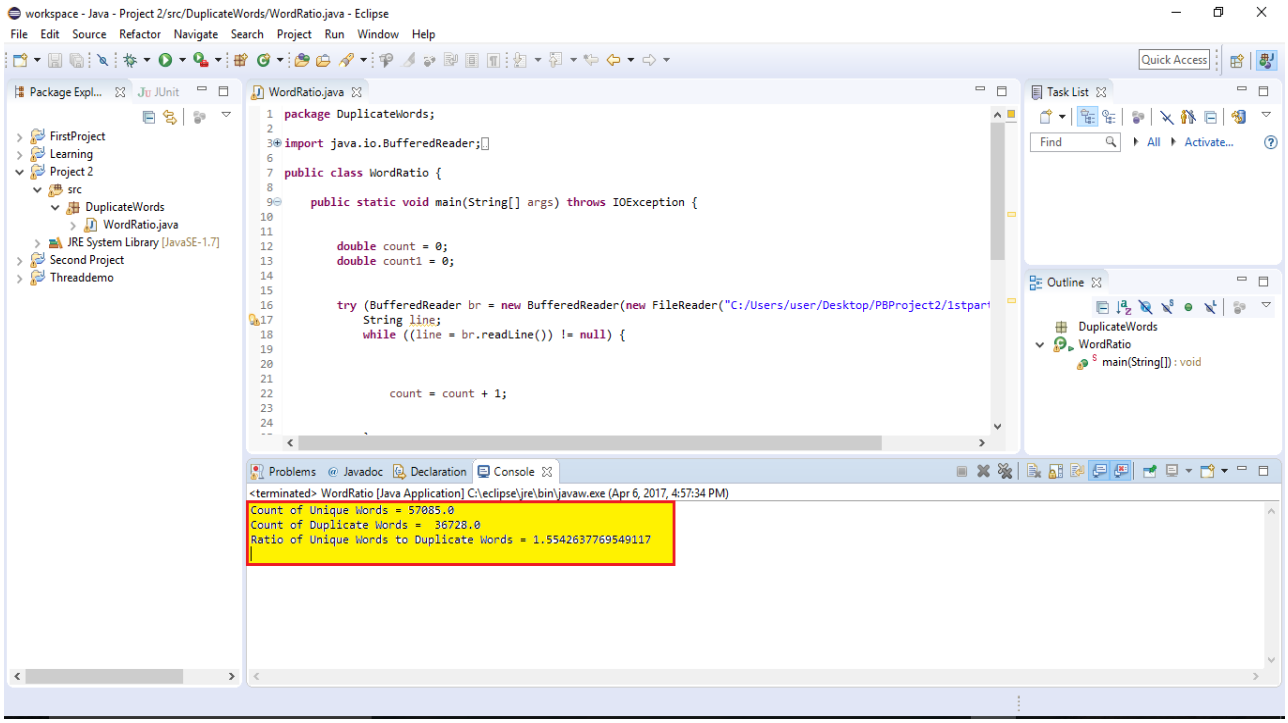
We have exported the java MapReduce program to jar file and run in the Hadoop environment to get the list of duplicate words. By using the command:

hadoop jar Uniques.jar UniqueCount /user/cloudera/test/tweettext.txt /user/cloudera/test/uniqs.txt

Below screenshot shows the list of unique words which are stored in the uniqs text file and the count on the right side of the words shows that the words are unique as they occurred only once.



We have used java code to calculate the ratio of the number of duplicate words and unique words, by taking uniqs and dups text files as input and calculated the **Unique Words: 57085.0** and **Duplicate Words 36728.0** and also the **Ratio of Unique Words to Duplicate Words 1.554** and the screenshot below depicts the same.



5. EXTRA REQUIREMENT:

We have also done the extra requirement to calculate the top ten best times to post a tweet on a twitter using java MapReduce program.

As first step we have retrieved the timestamp from tweets using python code and the output is stored in the TimeStamp text file

The below python code shows “Tweets.txt” is the input file and it prints out the timestamp of the tweets creation and stores it in the TimeStamp text file.

TweetTime.py - C:\Users\user\Desktop\PBProject2\Extra Requirements\TweetTime.py (3.6.0)

File Edit Format Run Options Window Help

```
# Import the necessary package to process data in JSON format
try:
    import json
except ImportError:
    import simplejson as json

# We use the file saved from last step as example
tweets_filename = 'Tweets.txt'
tweets_file = open(tweets_filename, "r")

for line in tweets_file:
    try:
        # Read in one line of the file, convert it into a json object
        tweet = json.loads(line.strip())
        if 'text' in tweet: # only messages contains 'text' field is a tweet

            print (tweet['created_at']) # when the tweet posted

    except:
        # read in a line is not in JSON format (sometimes error occured)
        continue
```

Ln: 1 Col: 0

TimeStamp Text file:

TimeStamp - Notepad

File Edit Format View Help

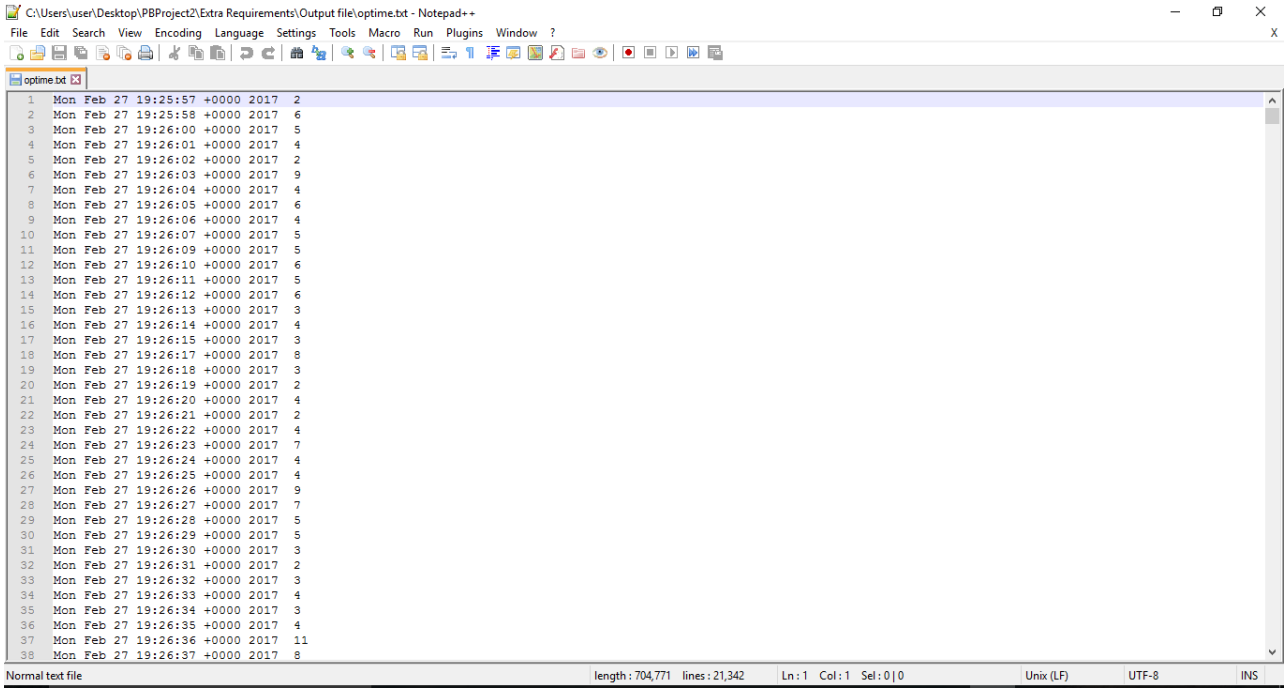
```
Sun Feb 26 18:47:47 +0000 2017
Sun Feb 26 18:47:47 +0000 2017
Sun Feb 26 18:47:48 +0000 2017
Sun Feb 26 18:47:48 +0000 2017
Sun Feb 26 18:47:49 +0000 2017
Sun Feb 26 18:47:49 +0000 2017
Sun Feb 26 18:47:49 +0000 2017
Sun Feb 26 18:47:49 +0000 2017
Sun Feb 26 18:47:49 +0000 2017
Sun Feb 26 18:47:49 +0000 2017
Sun Feb 26 18:47:50 +0000 2017
Sun Feb 26 18:47:50 +0000 2017
Sun Feb 26 18:47:50 +0000 2017
Sun Feb 26 18:47:50 +0000 2017
Sun Feb 26 18:47:50 +0000 2017
Sun Feb 26 18:47:50 +0000 2017
Sun Feb 26 18:47:51 +0000 2017
Sun Feb 26 18:47:52 +0000 2017
Sun Feb 26 18:47:52 +0000 2017
Sun Feb 26 18:47:53 +0000 2017
Sun Feb 26 18:47:53 +0000 2017
Sun Feb 26 18:47:54 +0000 2017
Sun Feb 26 18:47:54 +0000 2017
Sun Feb 26 18:47:54 +0000 2017
Sun Feb 26 18:47:54 +0000 2017
Sun Feb 26 18:47:54 +0000 2017
Sun Feb 26 18:47:54 +0000 2017
Sun Feb 26 18:47:55 +0000 2017
Sun Feb 26 18:47:55 +0000 2017
Sun Feb 26 18:47:55 +0000 2017
Sun Feb 26 18:47:55 +0000 2017
Sun Feb 26 18:47:55 +0000 2017
Sun Feb 26 18:47:56 +0000 2017
Sun Feb 26 18:47:56 +0000 2017
Sun Feb 26 18:47:56 +0000 2017
Sun Feb 26 18:47:56 +0000 2017
Sun Feb 26 18:47:56 +0000 2017
Sun Feb 26 18:47:57 +0000 2017
Sun Feb 26 18:47:57 +0000 2017
Sun Feb 26 18:47:57 +0000 2017
Sun Feb 26 18:47:57 +0000 2017
Sun Feb 26 18:47:58 +0000 2017
```

To find the number of tweets created at the particular time:

We have exported the java MapReduce program to jar file and run in the Hadoop environment to get the list of the timestamps. By using the below hadoop command the java map reduce program will run and gives the below output file with count of the timestamps.

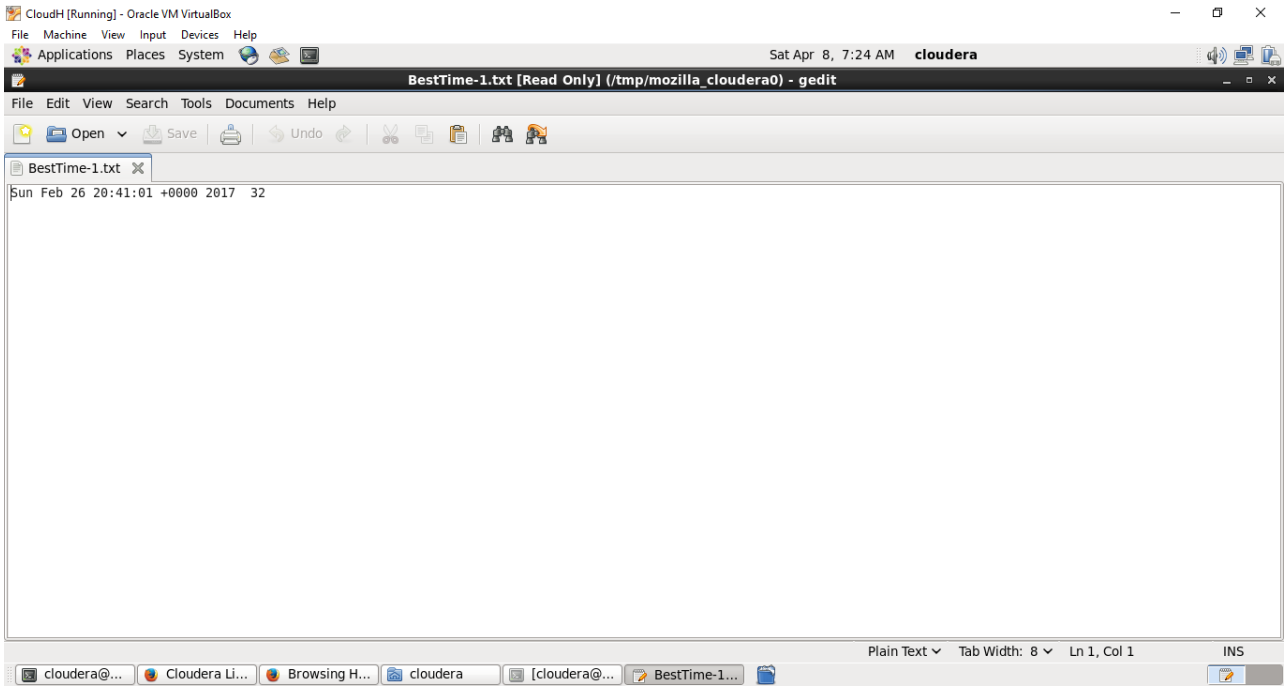
hadoop jar TimeSort.jar TimeCount /user/cloudera/test/TimeStamp.txt /user/cloudera/test/TimeStampOutput1.txt

The below screen shots shows the output after running the above code and it gives number of tweets that occurred at that particular time. The count on the right side of the timestamp shows the same.



Best Time to post tweet:

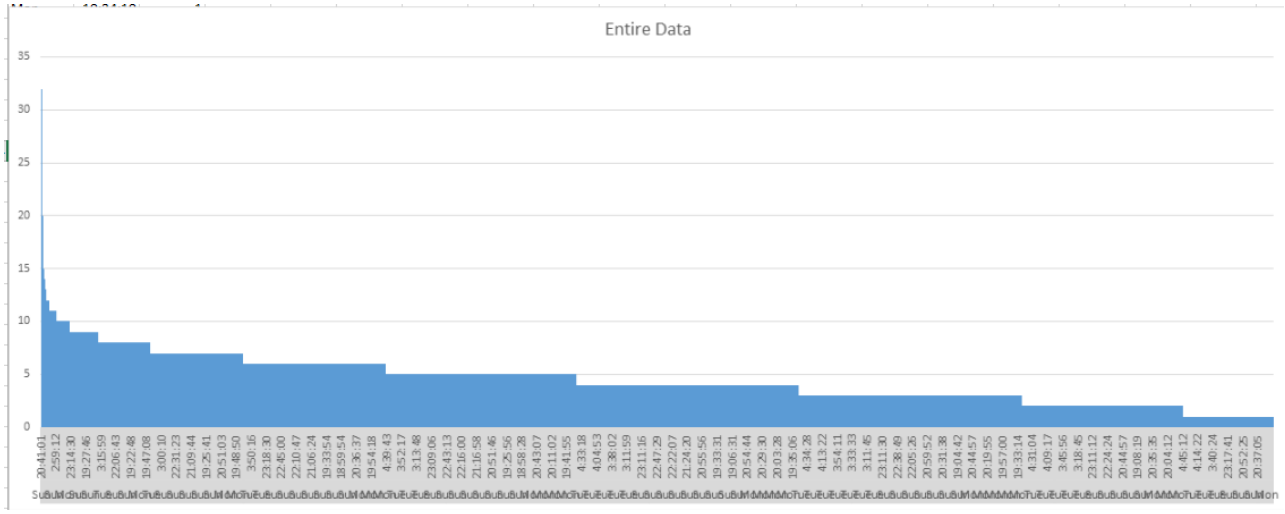
The below screenshot shows that "**Sun Feb 26 20:41:01 +0000 2017**" is the best time to post a tweet. It is the time that occurred 32 times more than any timestamp of tweet creation. This means most of the tweets are created at this time.



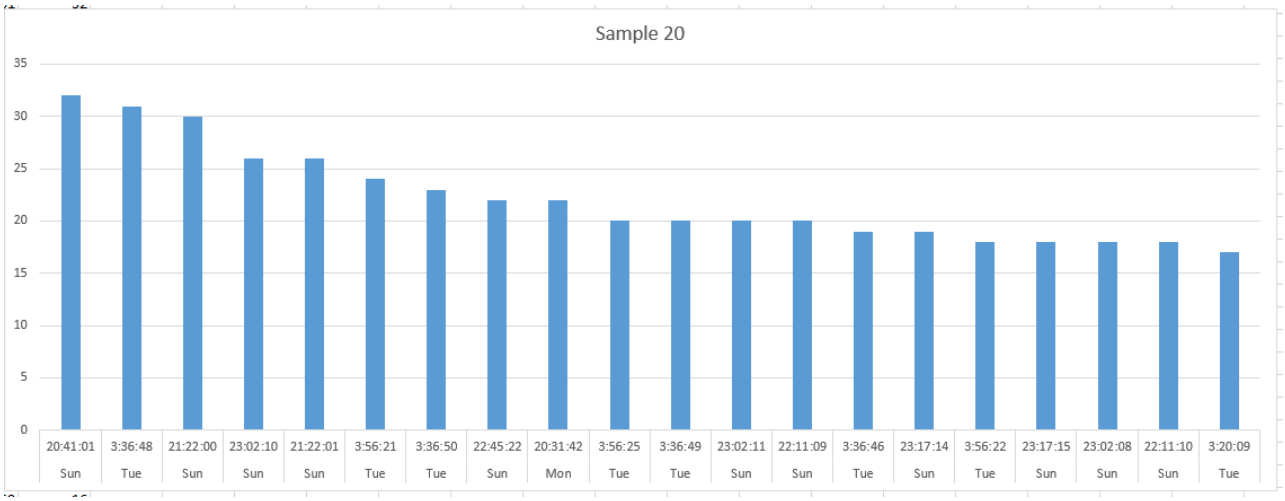
Metric Criterion:

We have considered TimeStamp of the tweet to propose a metric criterion. TimeStamp refers to the time of the tweet creation. As we can see from the top 10 best times output most of the tweets are created during **Sunday** and the best time to post a tweet is also on Sunday.

Below is the graph for the entire data.



The above graph is not clear as it is the graph for over 1Lakh data. So to clearly explain we have took a sample of 20 timestamps from the entire data. Below is the graph for it:



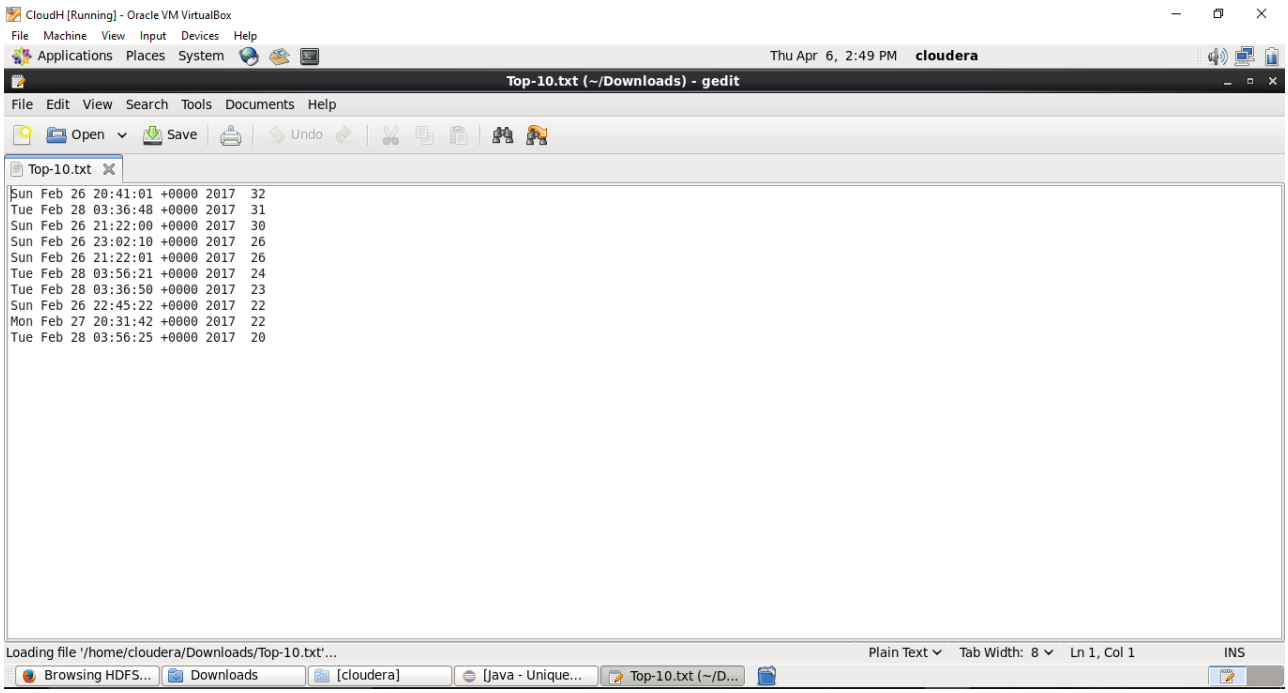
In the above graph X-axis represents the time and day the tweet was created and Y-axis shows the number of tweets created at that particular time. From the graph we can see that at timestamp **"Sun, 20:41:01"** most of the tweets are created and the lowest in this sample 20 graph is seen at **"Tue, 3:20:09".** In this sample data out of 20 days; 11 days are Sundays i.e tweets are created or posted mostly on Sundays. From the data we can also assume that best day to post a tweet is on Sunday. And another 8 days the tweets are posted on Tuesdays and remaining one day the tweets are posted on Monday. Also, if we consider timings most of the tweets are collected during night time. From this we can conclude that best time or most tweets are posted during night time on Sundays.

Top Ten Times:

To get the **Top Ten** timestamps from the TimeStampOutput1 file from above by using the command:

```
hadoop fs -cat /user/cloudera/test/TimeStampOutput1.txt/part* | sort -n -k7 -r | head -n10 >
        Top-10.txt
```

The above hadoop command takes the timestamp output from the above and sorts the value in the key 7 that is the count of the timestamps when the tweets are collected and gives the top 10 timestamps and stores in the Top-10.txt file.



5.References

- http://stackoverflow.com
- http://www.tweepy.org/
- https://www.cloudera.com/
- https://www.youtube.com/watch?v=TWYd0TD8Ops
- http://www.aegissofttech.com/Articles/how-to-get-top-n-words-count-using-big-data-hadoop-mapreduce-paradigm-with-developers-assistance.html