BIG DATA PROJECT REPORT

Nihad Guluzade, 17011903

Rahmi Cemre Unal, 14011052

The goal of this project is to develop several Map/Reduce programs to analyze provided dataset. The dataset includes tweets containing Bitcoin or BTC from 01/01/2016 to 29/03/2019. The format for the input of a Mapreduce job is as follows: "User, fullname, tweet-id,timestamp, url, likes,replies,retweets, text, html"

The objective of this project is to implement a Hadoop MapReduce environment on a given dataset to analyze the following:

1. The users who have the most likes.
2. The users who have the most retweets.
3. The users who have the most replies.
4. Number of tweets that were posted on a specific day.
5. Number of tweets that have specific number of characters.

In this project, 5 hadoop job implemented. For each job, there is a mapper class and a reducer class.

Most Likes, Most retweets, Most replies

The mapper class for this job maps all the users and like numbers of their tweet's like (key,value). It also adds the given number of people that has most likes to a tree map.

The reducer class for this job adds all the likes number for each user and put them in a tree map like (sum,user). Classes with same functionalities exist for most retweets and most replies jobs too.

Additional parameter can be specified on the GUI to generate the top numbers of users by their number of likes, retweets, or replies.

# of tweets for each day

The mapper class for this job maps all the tweets and their posted dates in a format which is (key,value).

The reducer class for this job adds all the individual time stamps which are posted dates for tweets and the sums of them are the results for number of tweets posted for each days.

For other purposes, the exact date can be picked from the GUI to generate the tweets that were posted on the chosen date.

# of tweets has # of characters

The mapper class for this job maps all the tweets by the number of their length.

The reducer class for this job counts the number of tweets that have same number of characters.

Exact length can be specified to generate the tweets.

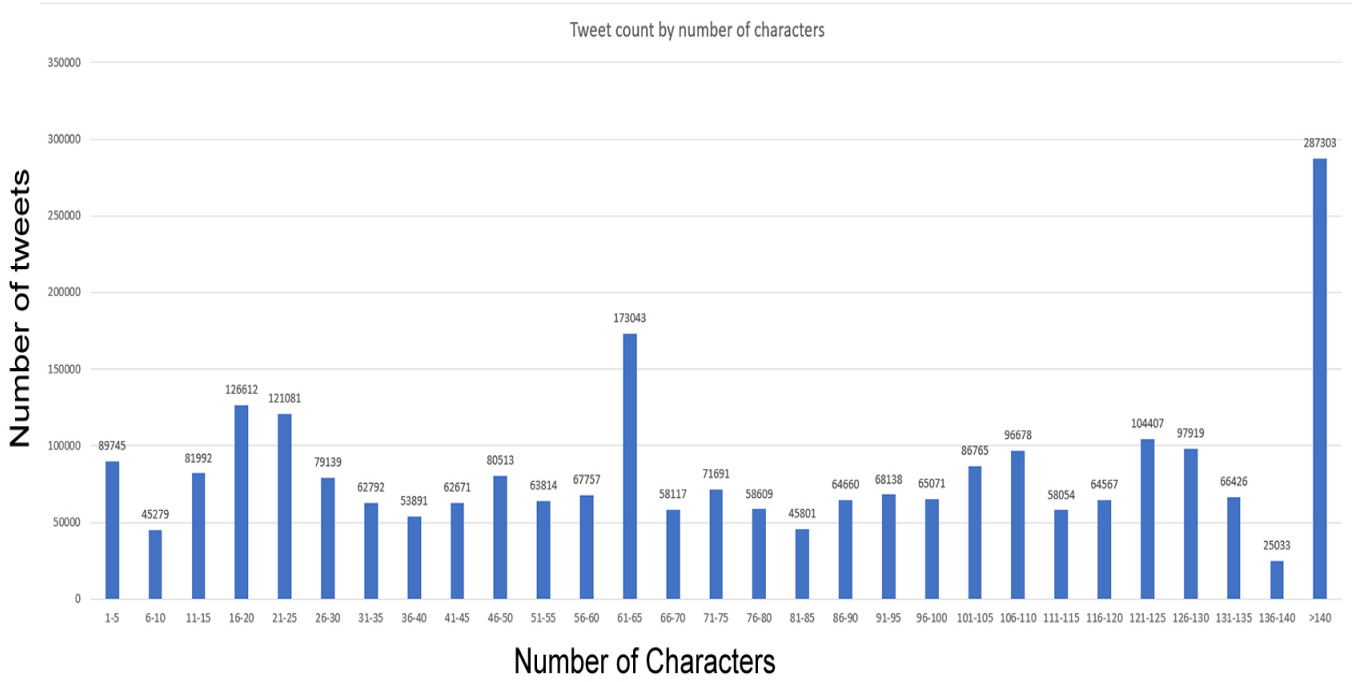
The difficulties encountered

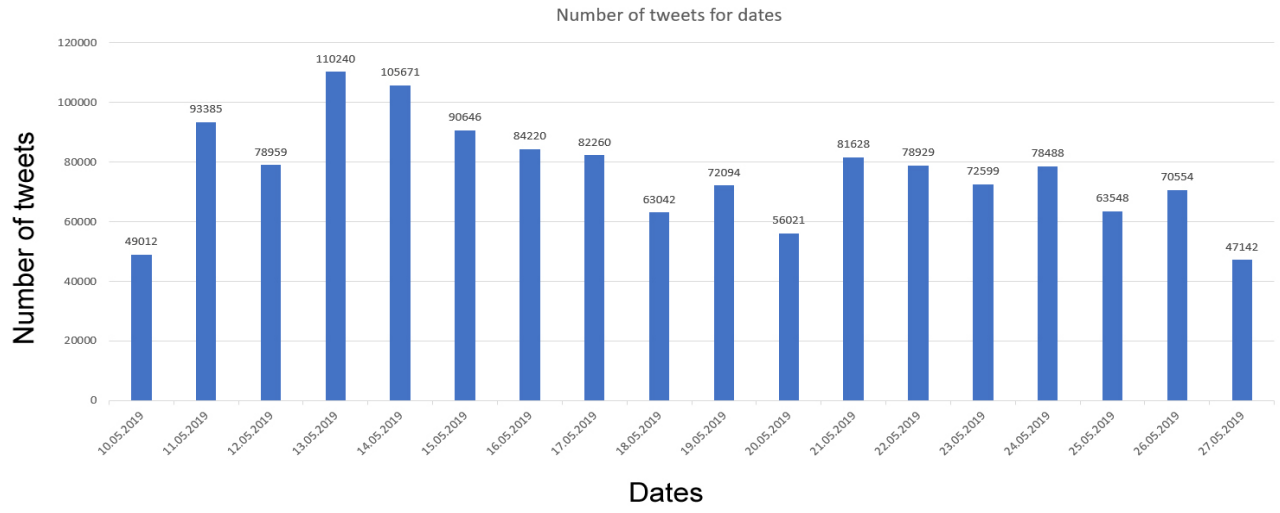
One of the main challanges for starting to this project was Hadoop installation. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. Since we were dealing with a big data project for the first time, the environment to be worked on had to be prepared first and this process took a lot of time.

Appropriate java jdk is installed on the device where the Hadoop will be installed and environment variables are set. In order for Hadoop to perform distributed operations, the files were made changes with the descriptions on the site.

Another challenge was the size of the data set. Since the dataset was so big to fit our personal computer's memories, we decided to take a part of it and use that part for the jobs.

RESULTS


Tweet count by number of characters
Number of tweets
350000
300000
250000
200000
150000
100000
50000
0
89745
45279
81992
126612
121081
79139
62792
53891
62671
80513
63814
67757
173043
58117
71691
58609
45801
64660
68138
65071
86765
96678
58054
64567
104407
97919
66426
25033
287303
1-5
6-10
11-15
16-20
21-25
26-30
31-35
36-40
41-45
46-50
51-55
56-60
61-65
66-70
71-75
76-80
81-85
86-90
91-95
96-100
101-105
106-110
111-115
116-120
121-125
126-130
131-135
136-140
>140
Number of Characters


Number of tweets for dates
Number of tweets
120000
100000
80000
60000
40000
20000
0
49012
93385
78959
110240
105671
90646
84220
82260
63042
72094
56021
81628
78929
72599
78488
63548
70554
47142
10.05.2019
11.05.2019
12.05.2019
13.05.2019
14.05.2019
15.05.2019
16.05.2019
17.05.2019
18.05.2019
19.05.2019
20.05.2019
21.05.2019
22.05.2019
23.05.2019
24.05.2019
25.05.2019
26.05.2019
27.05.2019
Dates

USE CASE DIAGRAM