**Targeted Tweets- TWT**

Contributed by

Vignesh Sethuraman - Vxs144030

Balaji Ayyadurai - Bxa140230

Bhargavi Ravula - Bxr140530

Sanjay Gandhi - Sxg141730

Avinash Nagasundaram - Axn143730

Shriram Vellaichamy - Sxv144730

Vijay Ramanathan - Vxr141530

# Abstract:

In the current Social media world, Content based marketing and targeted advertising are playing a key role. Twitter is one the largest Social media forum with millions of users logging in and actively participating every single day. Every organization has a huge community of followers retweeting their tweets every single day. Would it be awesome to know when your followers are online so that your messages can reach them instantly?

# Objective:

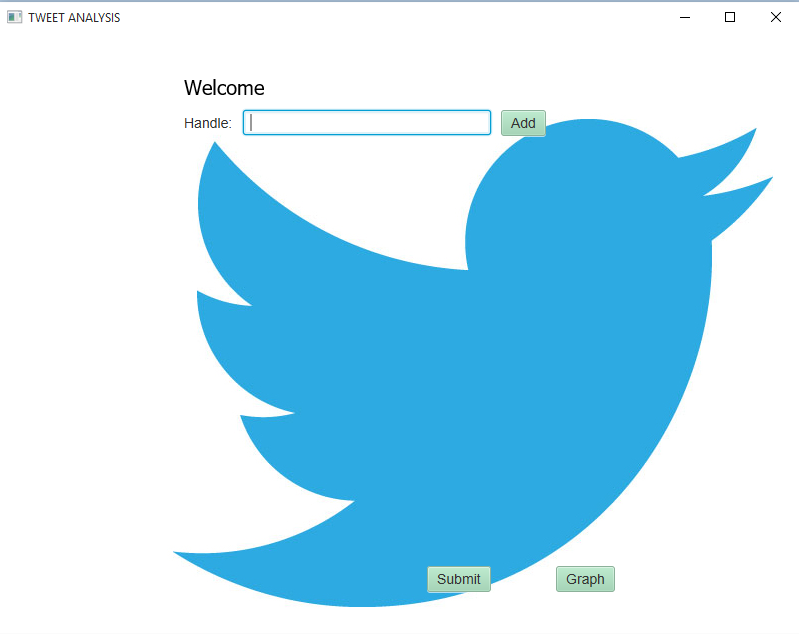
To analyze twitter data of an organization or a celebrity to identify the time window when most of their active followers are online. Every users’ tweeting time during the 24 hour period is identified and clustered in decreasing fashion of the most active window so that we can publish tweets to target maximum exposure.

# Software Requirements:

* Web user interface for input(twitter handle to extract)
* HDFS for storing and clustering of data
* Map-Reduce for selecting and gathering twitter data.

# Algorithms and Testing methods used:

* A twitter handle is selected for testing. For example CNN, which would be entered in the below given user interface



* The UI provide the input either to handle a single tweeter handle or multiple handle to analyze
* After the tweeter handle is submitted for processing, all the tweets from the start of its creation till date are analyzed for tweets that have the most number of retweets
* The top 1500 of those retweets are selected and the followers who retweeted them are identified
* For each of those 1500 followers, all of their tweets starting from the beginning of timeline till date are clustered with respect to their posted time. Convergence of the clustering process will result in finding out their most active tweeting time

## Hadoop Map-Reduce Algorithm

* To process the tweets from the twitter handle, twitter4j API is used and the first of the two Hadoop-map-reduce is used to reduce the tweets and retweets
* Based on retweets which provide the top 20% followers re-tweeting the max post(ranking based on the tweet is sorted) would be produced as a result of the first Hadoop-map-reduce job
* Now the second Hadoop Map-reduce is used to pull out the dump of all the top tweeter’s tweets(1500 tweeters with each pulling out 3200+ tweets)
* This dump is them processed by the clustering algorithm for user frequency analysis.

## K-means Clustering Algorithm

* K-means clustering algorithm is used for clustering tweets based on time.

|  |  |
| --- | --- |
| 1. Initialize the center of the clusters | μi= [some value](http://www.onmyphd.com/?p=k-means.clustering#h3_init) ,i=1,...,k |
| 2. Attribute the closest cluster to each data point | ci={j:d(xj,μi)≤d(xj,μl),l≠i,j=1,...,n} |
| 3. Set the position of each cluster to the mean of all data points belonging to that cluster | μi=1|ci|∑j∈cixj,∀i |
| 4. Repeat steps 2-3 until convergence |  |
| Notation | |c|= number of elements in c |

K =6

Initial set of Centroids: 2, 6, 10, 14, 18, 22.

# Code Snippet:

Figure 1 : K Means - Clustering Algorithm using map reduce to arrive at an optimal time line break down of active user window

Figure 2: K Means – Clustering Mapping of user data for time stamp mapping.

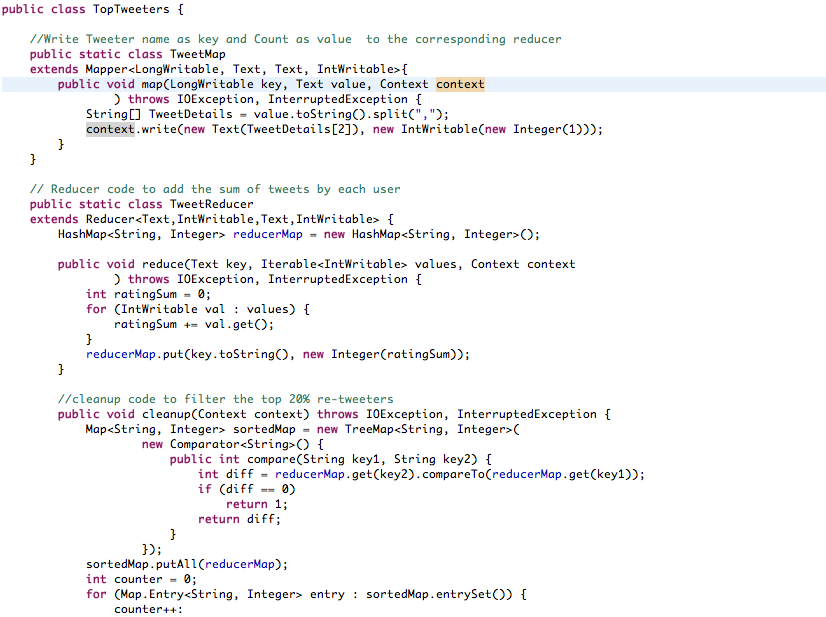


Figure 3: Tweet Reader – Map – Reduce Code segment

Figure 4: User tweets extracted based on time for clustering process 

Figure 5: Retweet user extraction code for Map-reduce job

# Result Analysis:

Below shown is the graph plot the output from the clustering data.

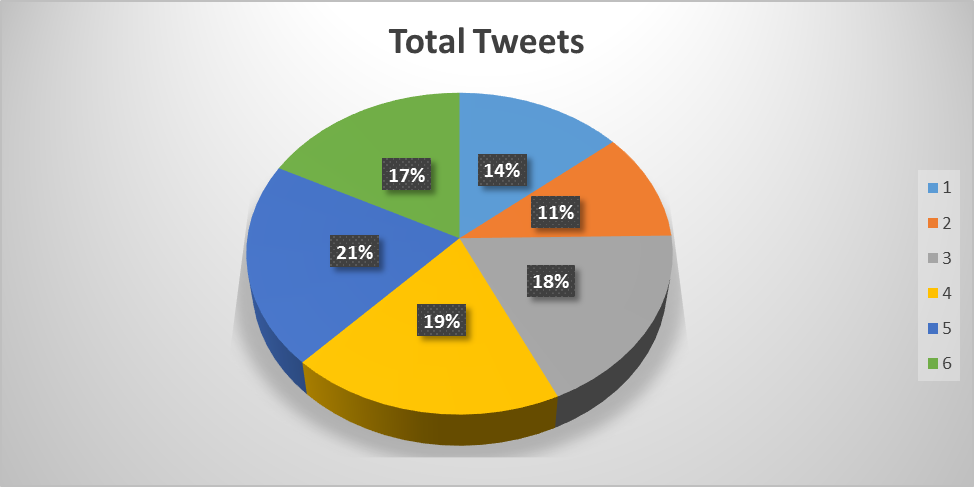


Figure 6: Overall split up over a 24 hour period

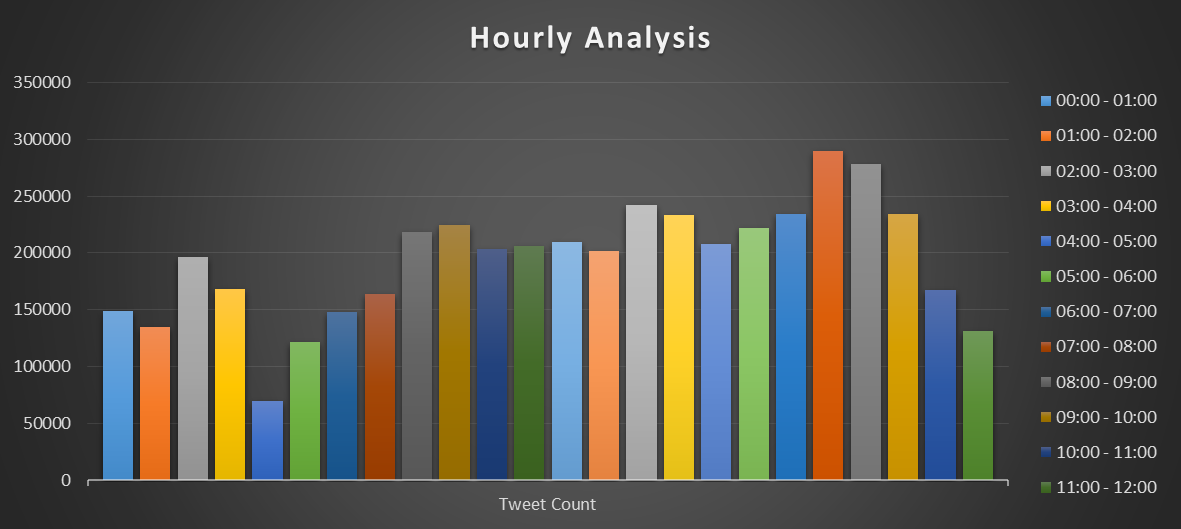


Figure 7: Hourly analysis of the tweet count

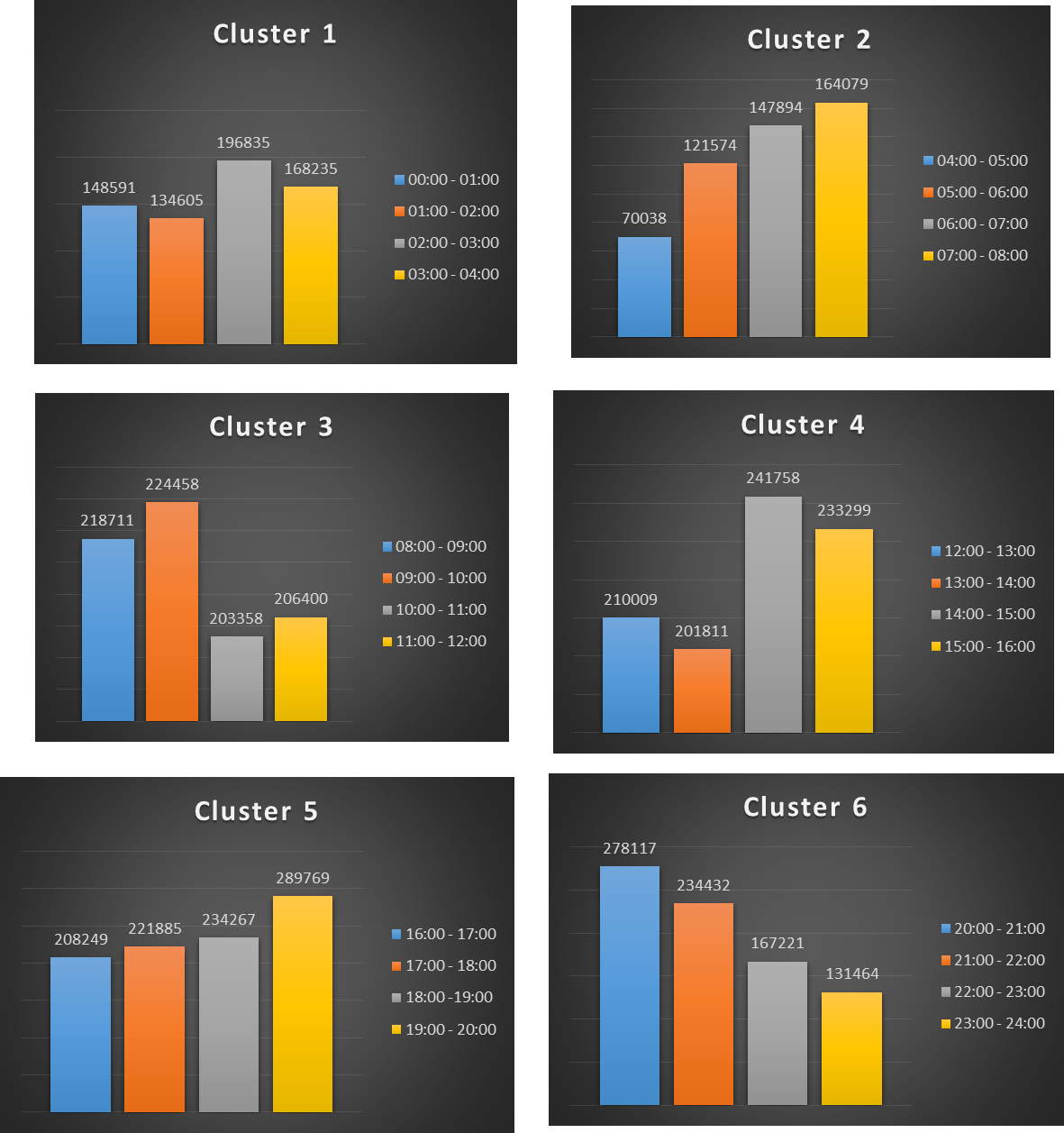


Figure 8: Cluster Split up Analysis

# Future Scope:

Since the data mined from the above mentioned Hadoop-map-reduce jobs contains various information like geography and demography of the users/target audience, further analysis to yield the sentiment analysis and word frequency extraction (content based advertisement) can be performed on the json objects collected which contains the details of the users.

Also it’s observed that there are a number of intermediate files written to the filesystem between the Hadoop-map-reduce jobs, hence a suitable adaptation of the above procedure in Spark-Scala system would be yielding better results in terms of performance of the data analysis.

# Conclusion:

Thus from the above mentioned case of twitter data analysis for CNN tweets and CNN followers tweet, we arrive at a conclusion that during a 24hr windows an optimal time frame for targeted maximum exposure would be between 19:00 and 20:00 and second optimal target would be 20:00 and 21:00. Also a further breakdown analysis for each window is also listed above in the result analysis section where the maximum contribution arrives from the cluster 5 which is window of 4 hours between 16:00 and 20:00. Thus we have successfully implemented big data concepts and machine learning concepts to analyze huge data and arrive at a meaningful analysis.