International Journal of Pure and Applied Mathematics
Volume 118 No. 7 2018, 523-530
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Twitter Data Analysis for Live Streaming By Using Flume Technology

A.Jaya Mabel Rani
Jeppiaar Maamallan Engineering College, Sriperumpudur, India
Email:jayamabelrani@yahoo.com

Abstract— In order to collect and process the streaming data from various streaming sites and produce an analytical report that helps to get a clear pictorial representation of events, the assets of streaming process generates massive volume of real time data mainly referred to the term "Big Data". This system proposes to produce a graphical visualization (bar chart, pie-chart, etc.,) for sentimental analysis using NLP, analyzing particular event on particular period of time and WORD CLOUD in which each independent word acts as hyperlink. Those hyperlinks are being redirected to browser for geographical map with location tagging. The generated results will be useful for either government or private firms (News reports/Consultancies/Research). In order to aggregate, store and analyse the streaming data that are being generated Day-By-Day we get into the concept of Hadoop And Flume Technologies. Products of apache foundation been used as an open source tools for data analytics. With the help of these tools and technologies process of predicting and analyzing data has been made much easier and efficient. This proposed system designed by an API that helps to collect data from Twitter/ other streaming sites by using '#' tag/Keywords. Tweets by the News channel and retweets by the public are being collected. Those generated data serves as a input for analytical operation. Purpose of generating word cloud from the tweets which brings out the emotions of the country and tell the importance of the particular event along with the location by using the tool "Flume". Tableau is connected to the cluster that helps in analysing and generating the word cloud.

***Index Terms*—Big data , Hadoop and flume, tweets, tableau, cloud, NLP, streaming, hyperlink, geographical data.**

I. INTRODUCTION

Word clouds are one of the simplest and most intuitive ways of visualizing text data. The data collected of last two week, from most of all public tweets related to what the people are talking around the globe The word cloud are generated through the following process. First, a computer program takes a text and counts how frequent each word is., for example if there was a character named "Cat" as well as generic animals referred to as "cat". Also remove stop words[1,2], which add little to the final visualization in many cases. Second, creation of word frequency list and incorporating these changes, then the program puts them into the ready queue and starts to print them. Then the word most frequently appears is placed as the highest position .

In this paper used, very simple state, with other basic concepts along with AntConc and Voyant Tools. The generation of Word clouds for a corpse of text can provide as a initial point for the in depth analysis. For occurrence, these are help to judge whether a presented text is associated to a exact information required or not..

This work, discover the potential of using word clouds at the very core of text analysis and urbanized the Word Cloud surveyor a classical system that uses word clouds as its main revelation and communication hub with sophisticated natural language processing[3], sophisticated communication possibilities, and a top level of power for users to present support for various kinds of text analysis responsibilities.

The key aid of this work is: A text analytics move towards on word clouds that present a wide range of communication and logical features; and easy to use influential and extremely configurable implementation of the approach. A qualitative consumer study that yields additional imminent into the approach.

There are two various forms to examine the word cloud. First, the largest font of the four words of the PBC that came up time after time. Second, each of these core values displayed in different color of font

II. LITERATURE SURVEY

In the existing system the Social networking sites have word cloud. Word cloud which have commercial frequencies terms by the reviews and summary of the reviews with the live connection and the data's have to store the query result and then import it into the tool and it will take time to process the dataset.

The study on word clouds cataract in one of two categories; work that examine the efficiency and visual insight of word

clouds, and work that develops improvements and extensions to the word cloud hallucination. There have been more than a few attempts to examine the efficiency and awareness of word clouds. Bateman conducted a user study in which they thoroughly varied nine illustration properties of word clouds. Here originate that the properties with the major effect on the consumers' concentration are size of font, weight, color and also experiential a strong outcome of font size in their consumer' study. The terms in the heart of the cloud obtain more concentration on regular than terms near the borders.

Word clouds have been compared to un-weighted lists and other user interfaces in a numeral studies. The outcome point out that consumers are on average more efficient in spotting a exact term in an alphabetically prearranged un-weighted list than in an alphabetically arranged word cloud. However, often used vocabulary are found more rapidly in word clouds due to their bigger font sizes. Similar Tag Clouds come together [4,5,6] the ideas of word clouds and parallel coordinates to permit for a straight similarity of word frequencies at various points in time or from different data sources. Tree Clouds combine word clouds with trees to envisage the semantic relatedness of terms. Prefix Tag Clouds use prefix trees to cluster various word forms and envisage the sub trees as Tag Clouds. Then finally, all 3D types of Word Clouds, such as WP-Cumulus which provides the rotating in three-dimensional sphere of words. whereas a part of these extensions has been intended for exact application contexts, others can be used more commonly. Here adopted a few of these thoughts in our loom, such as the circular word cloud layout or the interactive importance of word relatives.

This mechanism mainly deliberate rectangular word clouds[7] with a sequential line-by-line arrangement. Still, several improvements and extensions to this basic arrangement have been projected in the last couple of years. For illustration, use slicing trees, nested tables, and rectangle packing to optimize the allocation of space in HTML-based word clouds. It present a associated algorithm for white space optimization, which can cope with a variety of shaped word clouds. It places words with a spherical fashion with the most common ones at the heart and those with lower occurrence towards the limits.

There are a number of word analysis systems that build by use of word clouds. Examples can be found in domains such as patent analysis[8], investigative analysis, opinion mining, etc. In most of these systems, Word Clouds are used in a fixed way to visually review word documents. An interactive word cloud variation has been implemented in the VisGets system. The of time bar chart and a geographical map are used to emphasize the related words and other views.

After a word file is overloaded, the system performs a linguistic analysis of its stuffing. So here used the Stanford CoreNLP tools for this reason and execute several pre-processing procedure like, tokenization, sentence splitting, part-of-speech classification, lemmatization, and named-entity recognition. Based on the results of the part-of-speech tagger, additionally implemented a detector for ostensible multiword language expression. It joins all incessant sequences of proper nouns that arise in the similar verdict.

In this modern achievement, two terms are linked if they co-occur within the same verdict. The substitute implementations might calculate the co-occurrences on superior word segments, such as paragraphs or whole documents. In this loom the associated words are denoted by a yellow color box which dissemination associated to the qualified co-occurrence frequency. Showing the complete frequency values to users lets them simply identify correct false impersonation.

III. PROPOSED SYSTEM

News channel data's are collected from Twitter by using '#' tag, tweets by the News channel and retweets by the public. Purpose of News channel word cloud which brings out the emotions of the country and tell the importance of the nations along with the location by using the tool. Using Flume, Tableau is connected to the cluster will help to extracting data to Tableau without having to store the query result.

One of the main characteristics of the proposed system is the dynamic data's from one security profile to another. The proposed system is more isolated than the existing system and it prevents the data in all other applications to isolation.

The proposed system can generate the location details of the profile user to update in database. System configuration files are analyzed in the proposed system. The System can vary the data from different storage areas. The more importance of the proposed system is to automatic activation of Security profile depending on the context, in which the device is being used. They are very speedy. Poring over date to expand them from research takes time. They are appealing, visual depiction of data tends to have an contact and generates attention amongst the end users. It is easily understandable by viewers. Many of the social networks or any business websites can use word cloud to attract the public interests and make them to understand. The architecture of Flume is given in the following figure 1

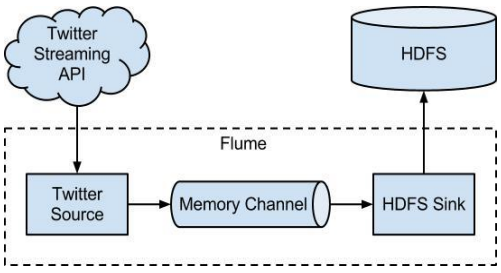

Figure 1 Flume Architecture

IV SYSTEM IMPLEMENTATION

Big data is data that exceeds the dispensation ability of conservative database systems. If the information data is

very big, moves very fast too, or does not athletic the structures of conventional database architectures. Every day of our life create millions and millions bytes of data. More than 90% of the information in the cloud has been created in the last three to four years only. This data is getting from everywhere like sensors used to gather weather information, posts to social media sites, digital images and videos, transaction records, and call data records, GPS signals, etc. Social media information provided that outstanding insights to companies on consumer performance and sentiment [8] that can be incorporated with CRM data for investigation, with 230 million tweets posted on Twitter per day, 2.7 billion likes, comments and more than sixty hours of video per minute uploaded in face book and you tube. Because of unstructured nature of big data, it is very difficult to manage, process, protect, and for privacy and security. NoSQL systems can supply insights into patterns and trends based on real-time data with least coding and without the need for data expert and extra infrastructure.

HADOOP

The HDFS (Hadoop Distributed File System) [9] divides large files into various nodes in the group. Additionally each portion is replicated across various machines So that, if any one of the machine failure does not effect in any data being unavailable.

Hadoop Streaming

Hadoop provides an application program interface(API) to MapReduce that allows to write mapping and reducing functions in languages other than Java. Hadoop Streaming uses Unix standard streams for the interface between Hadoop and program.

Streaming data access

Hadoop Distributed File System (HDFS) is built based on the idea of most efficient data processing format of write-once, read-many-times pattern.

Java MapReduce

There are three things need for Java Map Reduce i. a map function,ii. a reduce function, iii. some code to run the job. The map function is denoted by the Mapper Interface, using map() method. The Mapper interface define with four formal type parameters, such as 1.the input key, 2.input value, 3.output key, 4. output value types. The map() function passed a key, a value and also provides an instance of OutputCollector for writing the output.

A JobConf is an object, which forms the measurement of the job. It gives the power to run the job. After construction of a JobConf object, specify the input and output of the paths. Next, specify the map and reduce types by using setMapperClass() and setReducerClass() methods. The setOutputKeyClass() and setOutputValueClass() methods are used to control the output types. The static runJob() method is used to submits the job and taking information about its tolerance.

JOB TRACKER

There are six detailed levels in workflows. They are:

1. Submission of job. 2. Initialization of job. 3. Job assignment. 4. Execution of task. 5. Task progression. 6. Status updating.

APACHE FLUME

Apache Flume is a scattered, consistent, and available service for efficiently collecting, aggregating, and shifting large amounts of log data. Apache Flume is a tool/service/data ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log files, events, etc… from various sources to a centralized data store. Flume is used to move the log data generated by application servers into HDFS at a higher speed.

Advantages of Flume

Flume[10], can get the data from multiple servers directly into Hadoop. Along with the log files, Flume is also used to importance huge volumes of event data formed by social networking sites like Facebook and Twitter, and e-commerce websites like Amazon and Flipkart, etc. Flume ropes for a large set of sources and destinations types.

JSON

JSON means that JavaScript Object Notation. It is an self-governing data format and is the best substitute for XML. This explains how to parse the JSON file and extract necessary information from it. There are three parts are necessary for this. The first part does the HTTP call. The next one converts the stream into a sequence o characters(string). Then the third part converts the string to a jsonobject JSON eElements.

Using JSON in Hadoop

There are two type of operation: 1.Convert java class to JSON data (Serialization). 2.Parse JSON data and create java classes (Deserialization).

PRE-PROCESSING

The files that were provided by Yelp, were in JSON format. This was converted to “csv” for the ease of programming. During the conversion of the Review data file the following issues were faced:

Each record was treated as an individual file and multiple records could not be combined. The format of the JSON file had to be changed in order for the software JSON

Buddy to convert JSON file into one single CSV file[11]. Once the JSON file was converted to a CSV file format, the Text (Review) field with commas (,) was split into multiple columns.

For the system implementation this paper used;i3processor, 4GB and more RAM, 200GB and more hard disk, Ubuntu-16.04 amd64 operating system, Python, Java programming language, Technique Apache-Flume and Tableau tool.

IV TWITTER APPLICATION CREATION

For getting tweets from Twitter, there is in need to create a Twitter application. Twitter application creation steps are given below.

Step 1

First click with the link of https://apps.twitter.com/ and sign in the twitter account and do some work with twitter Application window where you are able to create, delete, and manage Twitter Apps.

Step 2

In the next step click on the Create New App button. Then you will be going to a window where you will get an application form to fill your detail information.

Step 3

Then the new App will be created. New app is used to create Consumer Key, Access Key, and Access Token Key. This will be used to edit in the Flume.conf file. While fetching data's from Twitter these Consumer Key, Access Key, Access Token Key is used to fetch data's which is lively tweeting in the account.

Step 4

These keys are used to Access Tokens tab and it can observe a button with the name of Create my access token. By Clicking this we can generate the access token.

Step 5

Finally, click on the right side top Test OAuth button of the page, Which display the Consumer key, secret, Access token. These are usedl to configure the agent in Flume.

Step 6

Consumer key, Secret, Access token are used to configure the Flume agent.

4.4 CONFIGURING FLUME

Twitter 1% Firehose Source

Only a single percentage (1%) of sample twitter fire hosr is experimented by streaming Application program interface(API) for converting them to Avro format, and sends to a downstream Flume sink. The jar files source can be located in the lib folder.

Setting the classpath

Then set the class path in the lib folder of Flume by exporting,

CLASSPATH=$CLASSPATH:/FLUME_HOME/lib/*

While configuring this source, you have to provide values to the Source type, consumerKey, consumerSecret, accessToken, accessTokenSecret, maxBatchSize by default value is 1000.

SETTING UP HADOOP CLUSTER

- Download and setup java-jdk.
- Download hadoop package format "hadoop-2.7.2.tar.gz" and unzip it.
- Setup up HADOOP_HOME environment.
- Configure the hadoop files inside conf directory.
- Once all the configuration is done run the terminal command "jps" to view all the nodes are working.

Now open the browser and enter into localhost:50070 to view the hadoop.

LOADING DATA INTO HDFS

HDFS sink comes stock with Apache Flume. Easily separates files by creation time. Data's will be loaded into this hdfs location : - hdfs://localhost:8020/user/flume/news

Interpreting the Output

The final output of tableau consists of 3 parts:

A Geo Location Map created based on the State and City of each business. This is represented on the lower right corner of the visualization output. The Business Name and its Location (City and State) is displayed on the lower left corner. The Top half of the visualization displays the Word Cloud containing the keywords of the business. Darker the color and larger the size of the word greater the count. This is an interactive visualization, when the Business Name is clicked, the location of the Business is displayed on the map and the word cloud for the business is generated on the top.

Verifying the Output

1. Breakdown: Initially only a single record was given as input to the MR, this helped in verifying if the output produced by the MR was accurate. This also helped in verifying if the output was in the correct format. 2. MR Unit Tests: MR unit was used to test if the output generated was right. 3. Finally after the entire output was generated random records were manually verified.

TERM FILTER AND SEARCH

By clicking the terms we can select the trems and colored the selected terms in red. This can be added and removed at any time and in any order from the term filter

The functionality of the is used to focus entire co-occurrences of data by removing all terms from the word cloud. It is a main trait of the approach and provides a supple focus and context technique for the users.

V PERFORMANCE ANALYSIS ANALYSIS BASED ON SERVICE

The following Figure 2 shows the framework for the computer Performance Measurement in Cloud Computing In this system establishes a set of performance criteria based on characteristics

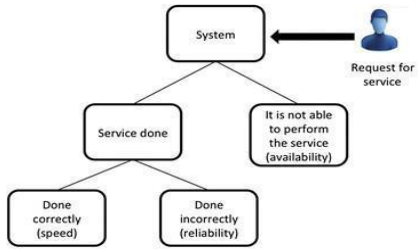

Figure 2. Performance analysis of a service request to a system

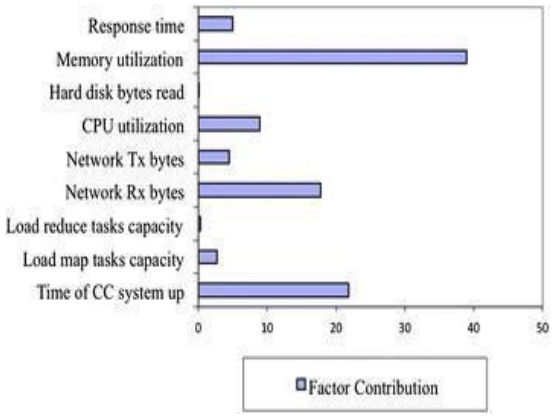

Figure 3 Percentage contribution factors based on processing time.

CONTRIBUTION OF FACTORS

In the about testing about 40% of Memory utilization has the highest influence on the processing time. Network Rx bytes is third highest influence on the processing time. Time of CC system up is the second is the factor with the least influence on the dispensation time in the cluster.

Figure 3 shows the percentage contribution for each factors wth dispensation time output which is related to system configurations.

RESULTS

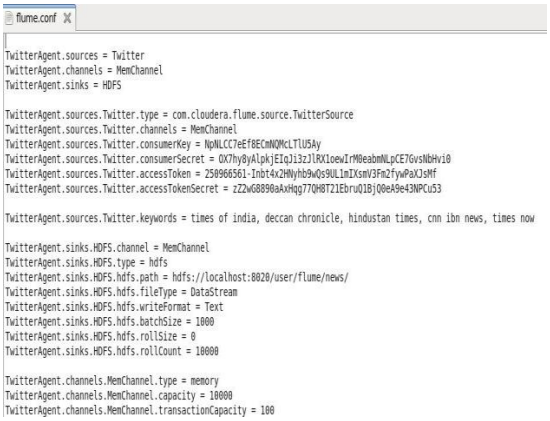

Figure 4.Twitter Flume.conf file to fetch data

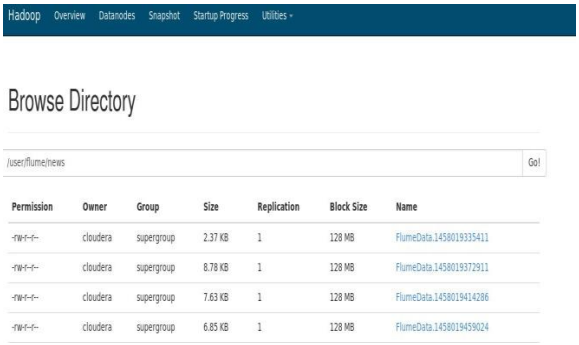

Figure 5. Fetched data's from Twitter and sinked into HDFS

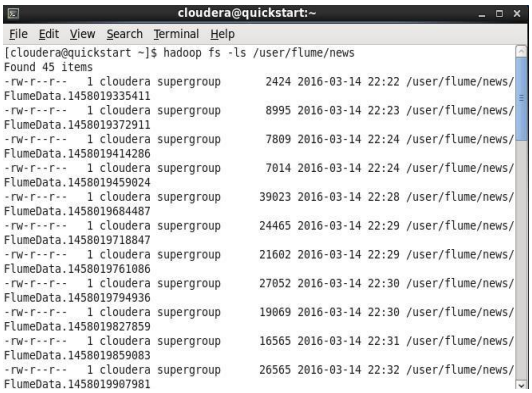

Figure 6. Hadoop fs –ls command to list the hdfs files

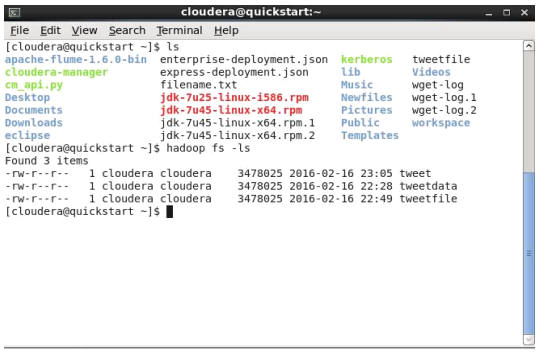

Figure 7. List of files in HDFS

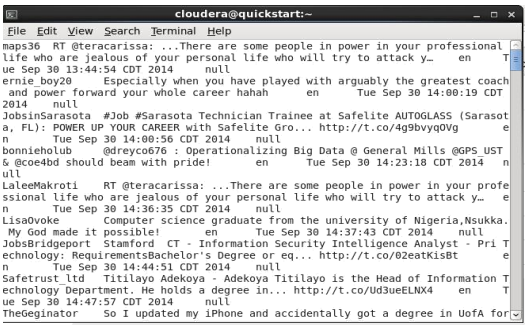

Figure 8. Data fetching from Twitter Agent

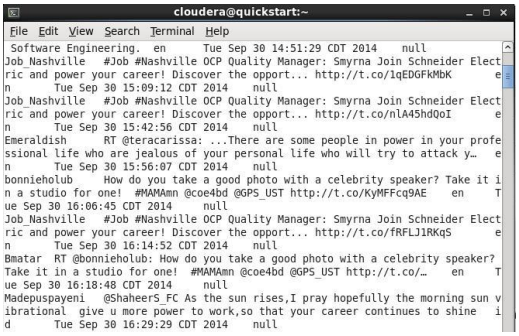

Figure 9. Tweet files

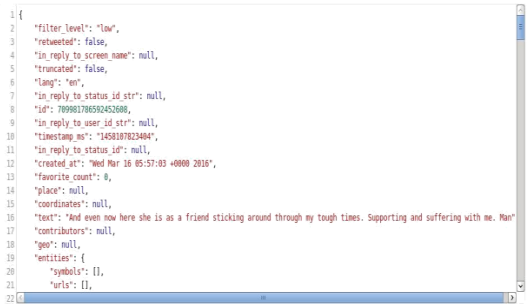

Figure 10. JSON validator

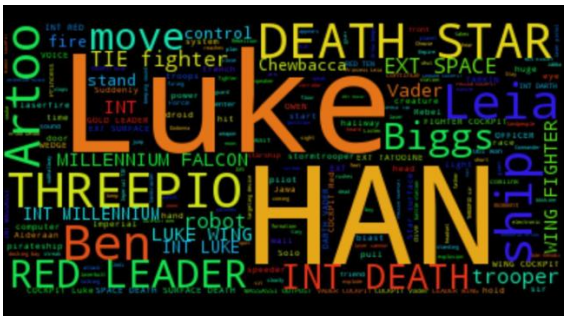

Figure 11. Word Cloud Output

The above figures 4 to 11 shows various output results like Twitter Flume.conf file to fetch data, Fetched data's from Twitter and sinked into HDFS, Hadoop fs –ls command to list the hdfs files, List of files in HDFS, Data fetching from TwitterAgent, Tweet files, JSON validator, Word Cloud Output.

CONCLUSION

The addition of the basic word cloud apparition with extra information and interactive features to convert it into a great tool for text analytics. Word clouds as its inner apparition method and integrate several interactive features into one reliable framework for interactive text analysis. The study outcome point out that word clouds are certainly a successful tool for text analysis if equipped with further information.

In upcoming work, plan to address the handling and comparison of multiple documents. An motivating question in this admiration is how to enlarge the word cloud view to permit for the comparison of various documents at a, parallel time. This could, be done using various colors the word cloud module of the Many Eyes website. Another choice could be Parallel Tag Clouds or by using the matrix form of word cloud. Generally aim to incorporate the offered approach with related work to permit for even more wide-ranging text analysis.

6.2 LIMITATIONS

It mainly concentrates on the news data from streaming process from the social network. Based on the flume technique, the data's from the twitter will be extracted.

Tableau was connected on the cluster, this helped in efficiently extracting data to Tableau.

The connection was live, it almost immediately reflected on Tableau without having to story the query result and then import into the tool.

6.2 FUTURE ENHANCEMENT

- The proposed system can also be improved in several conditions. For instance, to make the rules

specification process easier, a solution could be more specific, when defining by the developer side.

- As an enhancement of this process, the another tool or technique should access the data from the twitter. The administrative part can be improvised as it is in the proposed system. So that enhanced report will be generated from the administrative part.

REFERENCES

[1] M. Burch, S. Lohmann, D. Pompe, and D. Weiskopf. Prefix tag clouds. In Proc. 17th Int. Conf. Information Visualisation, IV '13, pages 45–50. IEEE, 2013.

[2] Christopher D.Manning, Prabhakar Raghavan, Hinrich Schutze, "Introduction to Information Retrieval", Cambridge university press, 2016, pp 100- 122.

[3] Dena Demner,fushman,Wendy W.Chapman,Clement c Donald, "What can natural language processing do for clinical decision support" , Journal of biomedical information, Elsevier, Volume 42,Issue 5, October 2009,pp 760 -772.

[4] C. Collins, F. B. Viegas, and M.Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In Proc. of IEEE Symp. Visual Analytics Science and Technology, VAST '09, pages 91–98. IEEE, 2009.

[5] S. Bateman, C. Gutwin, and M. Nacenta. Seeing things in the clouds: the effect of visual features on tag cloud selections. In Proc. 19th ACM Conf. Hypertext and Hypermedia, HT '08, pages 193–202.

[6] Y.-X. Chen, R. Santamar´ıa, A. Butz, and R. Ther´on. Tag-Clusters: Semantic aggregation of collaborative tags beyond tagclouds. In Proc. 10th Int. Symp. Smart Graphics, SG '09, pages 56–67. Springer, 2009.

[7] Healey, Ramaswamy, Visualizing twitter sentiment, https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/, 2016

[8] http://www.hadoopadmin.co.in/sources-of-bigdata/.

[9] Yelakala Pragna, High Performance Fault-Tolerant Hadoop Distributed File System, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 5 Issue: 5 1137 – 1145

[10] Munesh Kataria1 , Ms. Pooja Mittal, International Journal of Computer Science and Mobile Computing, Big Data and Hadoop with Components like Flume, Pig, Hive and Jaql, ISSN 2320–088X, IJCSMC, Vol. 3, Issue. 7, July 2014, pg.759 – 765.

[11] http://libguides.library.kent.edu/SPSS/ImportData

[12] Duong, Van minh, Sentiment and Influence Analysis of Twitter Tweets, United States (12) Patent Application Publication (10) Pub. N0.: US 2013/0103667 A1 Minh (43) Pub. Date: Apr. 25, 2013.

[13] J. Schrammel, M. Leitner, and M. Tscheligi. Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches. In Proc. SIGCHI Conf. Human Factors in Computing Systems, CHI '09, pages 2037–2040. ACM, 2009.

[14] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. OpinionSeer: Interactive visualization of hotel customer feedback. IEEE Trans. Vis. Comput. Graphics, 16(6):1109–1118, 2010.

[15] S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of tag cloud layouts: Task-related performance and visual exploration. In Proc. 12th IFIP TC 13 Int. Conf. Human-Computer Interaction: Part I, INTERACT '09, pages 392–404. Springer, 2009.

[16] Johnston L A, Vikram Krishnamurthy. An Improvement to the Interacting Multiple Model (IMM) Algorithm. IEEE Trans. on Signal Processing, 2001, 49: 2909-2923.

[17] ZHU Hong-yan, HAN Chong-zhao, HAN Hong, et al. Study on Approaches for Track Initiation[J]. Acta Aeronautica et Astronautica Sinica, 2004, 25(3)：284-288.