International Journal of Engineering Research ISSN: 2319-6890)(online),2347-5013(print)
Volume No.5 Issue: Special 5, pp: 992-1128 20 May 2016

EPH-Enhancement of Parallel Mining using Hadoop

Neha Mangla
Assosiate Professor, A.I.T.,Bangalore
Apj.neha@gmail.com

Sushma K
MTech Scholar, A.I.T., Bangalore
sushmakrishna2@gmail.com

ABSTRACT

Data in this era is generating at tremendous rate so now it is need of today to handle the data to gain useful insight, this data can be useful for researcher and accommodation to do analysis. As we know traditional system cannot handle more than terabytes of data since it affects performance and also storage is very costly. Bigdata is a innovative technique analyze, store, manage, distribute and capture datasets. To achieve compressed storage in this implement a parallel mining algorithm called as enhancement of parallel mining using Hadoop. Hadoop is a platform which enables the distributing processing using mapreduce programming. This help in getting result at very fast rate as result in less time help in competing for growth of business. For the analysis in this paper unstructured datasets from real-time is taken and converted to structured format and process in mapreduce. It is found in literature existing mining algorithm for real time datasets lacks in fault tolerance, load balancing, data distribution and automatic parallelization. To overcome these disadvantages we implement mapreduce for association analysis. In EPH we improve performance by distributing load across the computing nodes .In our proposed solution we use real-world celestial spectral data .The graphical representation of traditional system comparison with Hadoop is shown in this paper.

Keywords: Bigdata, Hadoop, Mapreduce, Parallel Mining, Association analysis, Enhancement of parallel mining using Hadoop(EPH).

I. INTRODUCTION

Information Technology is growing rapidly and volume of data is increasing such as social media, balck box so on and it is reaching petabytes of data threshold and as increase in data also increases computational requirements which includes fault tolerance, load balancing, data distribution and automatic parallelization. In terms of academics and business bigdata is become the key role. Here efficient parallel mining algorithm techniques are used to easy and fast processing of data. In this we consider a data mining tool called as R tool to compare with the proposed system where we process our unstructured data perform the association analysis on the datasets and represent using graphical presentation in R tool the time required to process the datasets is more it recursively process the datasets and it cant process large amount of datasets which is great drawback.

Association analysis cannot be done on the real time datasets so this one of drawback .The real time data can be processed in Mapreduce. Firstly we generate realistic data by creating developer account and by streaming the data in the flume and the unstructured data is taken to hue and particular data can be searched using the solaries. We can also change colour, bolds, and highlights and so on in solaries only. Now the data as to be converted to structured data using Hive and this structured data is fed into Mapreduce is a programming model which consists of two phase:-

1. Firstly, Mapper phase which is each separate line and produces a key value pair.
2. Secondly, to do the association analysis and it is represented using graphical representation. After the all these task performed we compare the two system speed, processing speed, availability, an time taken for the execution of data and many more criteria.

The contributions of this paper are:-

I made complete overview about parallel mining of realistic datasets and converted to structured datasets.

I developed a parallel mining method using Map reduce programming model I also gave complete overview about existing traditional system R tool and did the association analysis using the datasets.I

also show the load balancing and how data is being distributed in the clustering nodes and processing is done.

The comparison of both the system is showed and it is measured in turn of processing speed, scalability ,availability, performance and what kind of synthetic and real world datasets that can be processed in these tools.

The paper contains as follows in Section II it describes background knowledge how the sentiment analysis of twitter data Section III The Overview about the existing system and how they are processing the Twitter data Section IV Overview about EPH Section V Implementation EPH Algorithm Section VI EPH over existing system Section VII Conclusion.

II. PRELIMINARY

Nowadays Wide area network plays a important role in the world this tend me to work on Twitter data. In this section we are going to learn about how we do Sentimental analysis on Twitter data from the mapreduce and R Tool. A sentimental analysis is nothing but retrieving the tweets from the Twitter API and calculating the score for that particular data. Firstly, we explain the sentimental analysis for here in existing system we are using the corpus based approach for the sentimental analysis which also include Natural language processing and machine learning processing.

There are some key characteristics of this tweets data:-

- *Message Length:* The maximum length of a Twitter message is 140 characters. This is different from previous sentiment classification research that focused on classifying longer texts, such as product and movie reviews.
- *Writing technique:* The occurrence of incorrect spellings and cyber slang in tweets is more often in comparison with other domains. As the messages are quick and short, people use acronyms, misspell, and use emoticons and other characters that convey special meanings.

Availability: The amount of data available is immense. More people tweet in the public domain as compared to Facebook (as Facebook has many privacy settings) thus making data more readily available. The Twitter API facilitates collection of tweets for training.

Topics: Twitter users post messages about a range of topics unlike other sites which are designed for a specific topic. This differs from a large fraction of past research, which focused on specific domains such as movie reviews.

Real time: Blogs are updated at longer intervals of time as blogs characteristically are longer in nature and writing them takes time. Tweets on the other hand being limited to 140 letters and are updated very often. This gives a more real time feel and represents the first reactions to events

Secondly to process the data in the map reduce the sentimental analysis is done in the Flume.

Flume is a reliable, distributed and available service for efficiently aggregating, collecting and moving the large amount of celestial data and log data.

Flume components interact in the following way:

Client starts the flow of the flume

The transmission of the Event is done by the client to a source operating with the Agent

This Event is received by the Source and delivered to one or more channels.

Channels can be drained due the one more sink with the same Agent.

The ingestion rate from the drain rate can be calculated by Channels using producer-consumer model of data exchange

When spikes in client side activity cause data to be generated faster than can be handled by the provisioned destination capacity can handle, the **Channel** size increases. This allows sources to continue normal operation for the duration of the spike.

The **Sink** of one **Agent** can be chained to the **Source** of another **Agent**. This chaining enables the creation of complex data flow topologies.

Because Flume's distributed architecture requires no central coordination point. Each agent runs independently of others with no inherent single point of failure, and Flume can easily scale horizontally

Initially I create a gmail account and followed by a Twitter account in this and start with our developer account and go for the option manage your app and generate the keys and access token.

Figure 1. Snapshot of creation of developer account

International Journal of Engineering Research
Volume No.5 Issue: Special 5, pp: 992-1128
ISSN: 2319-6890)(online),2347-5013(print)
20 May 2016

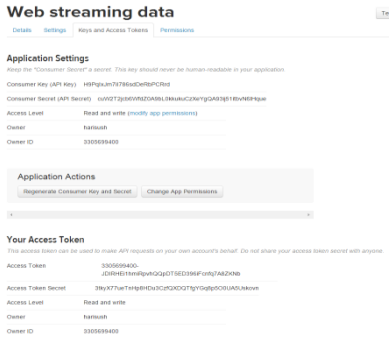


Figure 2. I have developer account with name WebStreaming data with keys and access tokens.

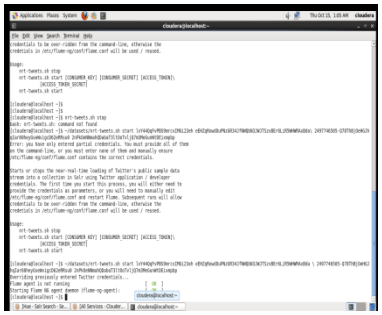


Figure 3. Coding in Hadoop to connect twiiter with Flume

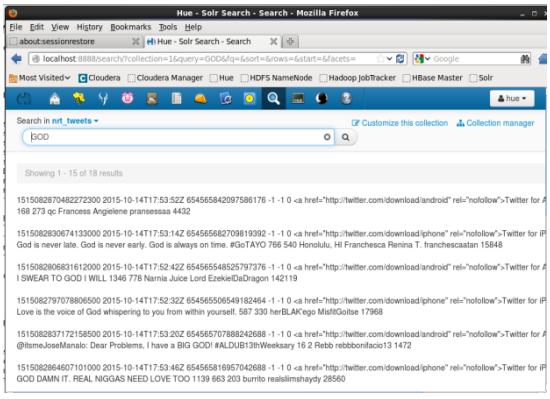


Figure 4.Twitter data retrieved in flume.

Figure 5 Query done in Hive platform

The data which is present now is unstructured data shown in Figure 5 which has to be converted to the structured format where this job is done by the hive platform only after convertion it can be processed to the mapreduce and it is processed for the further analysis.

III. EXISTING SYSTEM

Here I am going to learn about the how the tweets can be extracted using the code in the R tool with the called as twitter R[1][2], we need to install this package first. Later the tweets are converted to dataframe then corpus. The code is given below:

```
hive> use harish_db;
OK
Time taken: 0.046 seconds
hive> create table senti(id int,text string)
    > row format delimited
    > fields terminated by ',';
OK
Time taken: 0.812 seconds
hive> load data local inpath '/home/training/Desktop/senti.txt' overwrite into table senti;
Copying data from file:/home/training/Desktop/senti.txt
Copying file: file:/home/training/Desktop/senti.txt
Loading data to table harish_db.senti
Deleted hdfs://localhost/user/hive/warehouse/harish_db.db/senti
OK
Time taken: 0.347 seconds
hive> CREATE EXTERNAL TABLE dictionary (
    >     type string,
    >     length int,
    >     word string,
    >     pos string,
    >     stemmed string,
    >     polarity string
    > )
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'
    > STORED AS TEXTFILE
    > LOCATION '/home/training/Desktop/dictionary';
OK
Time taken: 0.081 seconds
hive>
```

```
# convert tweets to a data frame
df <- do.call("rbind", lapply(rdmTweets, as.data.frame))
dim(df)
library(tm)
# build a corpus, and specify the source to be character vectors
myCorpus <- Corpus(VectorSource(df$text))
```

Next step is used for stemmed words to retrieve their radicals from the twitter data. I need to install few packages they are Snowball, RWeka, Rjava and RWekajars. Now we have to convert this unstructured to structure in terms of matrix. Where here row means terms column means entity we build a term-document matrix from the above processed corpus with function TermDocumentMatrix()[7] After the completion of this stage a row will be generated.

It is now list Frequent term and association. findFreqTerms() this function is list the number of frequent less than 10(10 is just a example).Now I will plot the bar graph for the twitter data which is having the maximum tweets and retweets show in Figure 7.

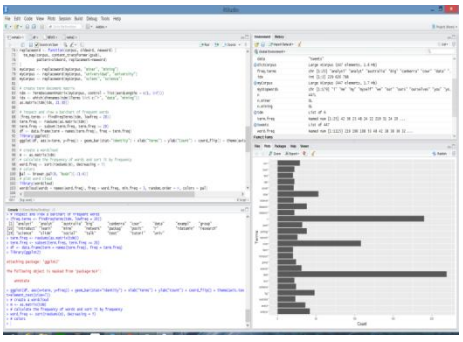


Figure 6. It shows the bar graph the maximum tweets that is happened.

I can also generate the word cloud for the particular datasets. I can calculate the maximum frequency and I can generate the word cloud for those maximum words. I can generate the networks of the terms for the particular data. a term-document matrix, term DocMatrix, is loaded into R, later it is converted to adjacency matrix. This matrix contains the maximum frequencies and it is in the form of table. Figure 7 shows the graph of the maximum number of words repeated in

ISSN: 2319-6890)(online),2347-5013(print)
20 May 2016

the table and we can build a graph with graph.adjacency()[6] from package igraph.

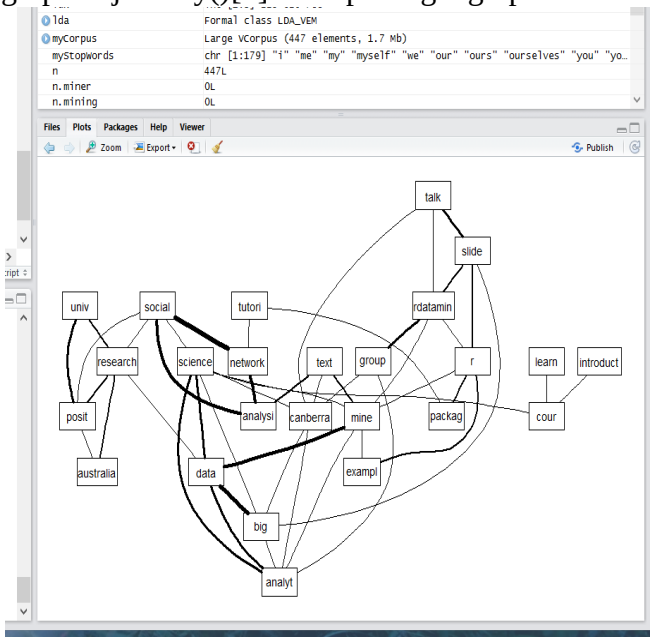


Figure 7. Represents the graph the maximum number of repeated words

IV. OVERVIEW OF EPH

In this proposed system I will process the structured data which we had retrieved from the flume platform and it should be processed into the mapreduce for the getting the maximum of tweets and retweets and pictorial presentation of it. The complete architecture about how it is working as shown in Figure 8 It is giving the complete as I discussed in Section II how the data is extracted from the flume the unstructured data later it is converted structured data using the hive platform and it is saved in hadoop distributed file system (HDFS)[11] and later it is processed to the mapreduce job as show in the Figure 8.In the mapper stage it process the structured data and the maximum of tweets and retweets done is saved and after the mapreduce job is completed successfully then it is saved in hdfs and it is sent to the Intelligent graph system for the result. I have already discussed about fluming of the data in the Section II itself.

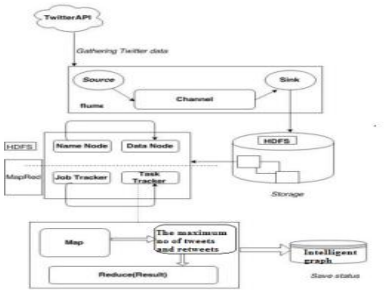


Figure 8 The complete overview about EPH mechanism

MapReduce Model

This model is used in parallel data paradigm. This model consists of 2 stages:- one is the map phase and other is the reducer phase.The data structure which is present in mapreduce[7][8] model is Tuple. The data which is sent sent inside the mapreduce task is converted to the tuple format. Each line is taken as the separate chunk into the mapreduce task. The map function is a list of <k1,v1> pairs.This is again processed to the reduce stage here it accepts only key value pair. The output from the reduce function is in the form of <k2,v3> pairs. The equation for this process can be represented as :

$$Map<k1,v1> = list <k2,v2>$$

$$Reduce(<k2,list(v2)>) = <k2,v3>$$

To conclude their will 3 stage in mapreduce one is splitting stage, Mapper stage and later is reduction stage:-

Splitting stage: Here the input is taken from the mapreduce and each line is split into multiple chunks and can be processed parallel[6][7]. I can also fix the of the chunk if necessary and their exists a function splitter where I can define my own split rules for the given datsets but by default Hadoop only completes this step.

Mapping stage: In this stage it will read the data chunk and convert the data into tuples. As I told for the text the input for map function will be the each single line in this file. This will emit out in tuple format and here it is shuffled and sorted and stored in the local file and processed to the reducer stage.

Reduction stage: After all the map tasks done all aggregation happen in this stage only List of values is accepted with the same key in the reduce function. We define our own process for the given value. The output is obtained which is in tuple format.

This model can be processed for large amount of datasets and here both map function and reduce function can parallel on the cluster without draining the performance of the system I am going to measure and discuss in Section VI.

V. IMPLEMENTATION OF EPH

Here in Mapreduce model I develop the code for EPH Program and in this there are three function are implemented and these algorithm are written in java.EPH is build a mechanism for data distribution, load balancing and parallelization for

 doi : 10.17950/ijer/v5i5/004

ISSN: 2319-6890)(online),2347-5013(print)
20 May 2016

the large number of datasets. To improve the storage efficiency of the system also.

Here the aggregation stage where in this process the given dataset is sent to mapper stage for processing and the association analysis is taken where it accepts the data key value pair in the given system it count the maximum of all tweets present in the table and later it is processed in intelligent graph to obtain the excepted result

Now I will the processing in the mapreduce before I start with mapreduce I should all the services using the command called as start-all.sh. I can check in the weather services are started or not using jps command and I can also check in hadoop administration shown in Figure 10 and Figure 11.I can create a new file in the HDFS here I have created a file called as fi. Using the command show in the Figure 12 I will copy the twitter data in the file fi and I can weather the data is copied or not in hadoop adminstartion as show in Figure 13.

Now it is time to start the mapreduce process and in the mapreduce process the for each instace of how much the mapreduce shows the data their only and wait the reduce to finish 100% as shown in the Figure 14.If the mapreduce successful then the fie called success will be created in fi inside the output file as shown in Figure 15.If the process is unsuccessful then it gives error message then and their only.

The obtained output must be put into the Intelligent graph which is written in php it is processed as shown in the Figure 16 The result obtained by the mapreduce is shown in term of line graph, bar graph, pie graph as shown in the Figure 17,18 and 19 .Even other types can also be implemented.

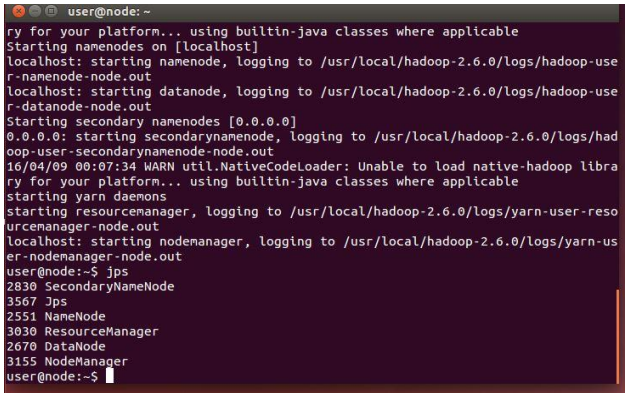


Figure 10. Start all the services

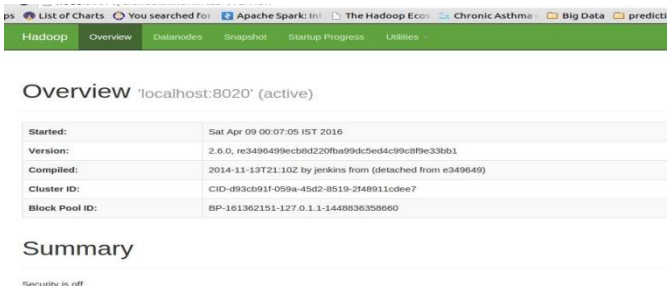


Figure 11. Checking in the hadoop administration services are started or not.

.

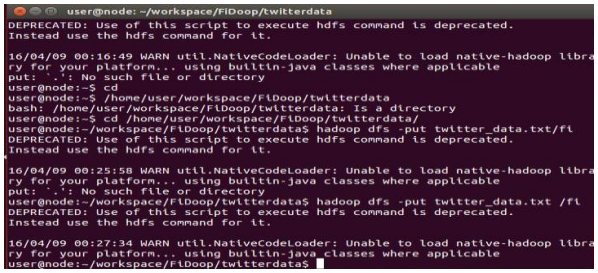


Figure 12 The data is copied to the fi file.

Figure 13 Shows the copied file in the hadoop administartion

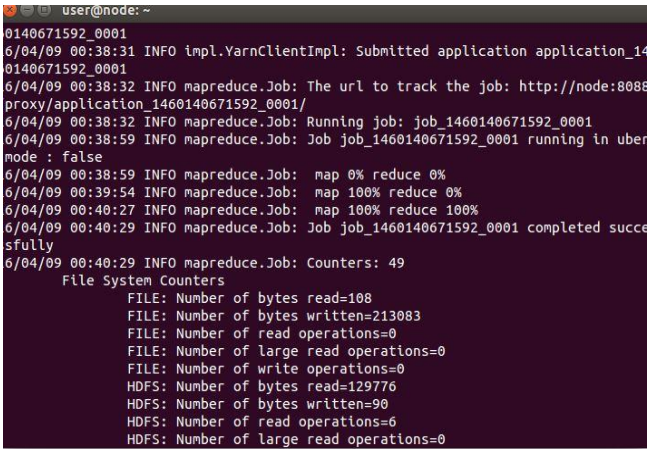


Figure 14. The mapreduce process for the file

ISSN: 2319-6890)(online),2347-5013(print)
20 May 2016

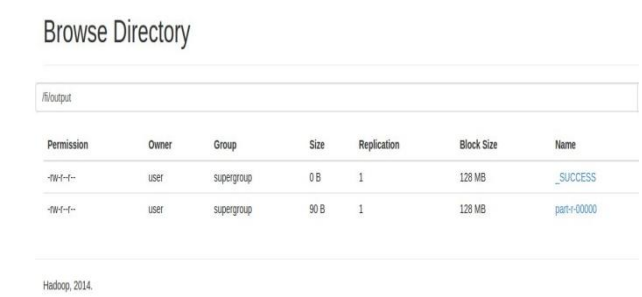


Figure 15 The success file created in fi file if mapreduce process complete successfully

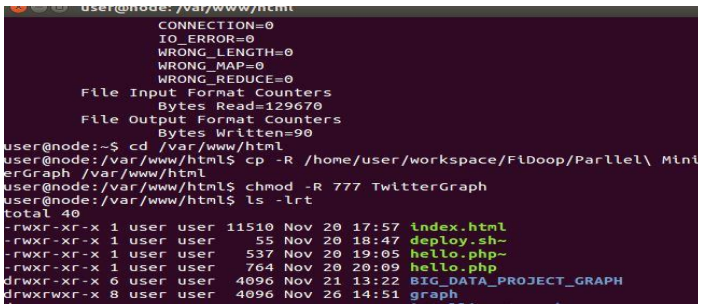


Figure 16 Initializing to the twitter graph and processing in it

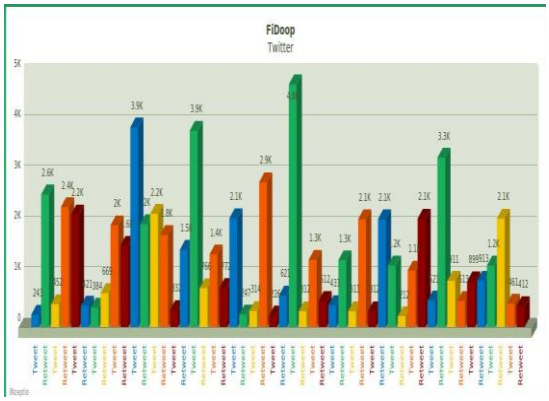
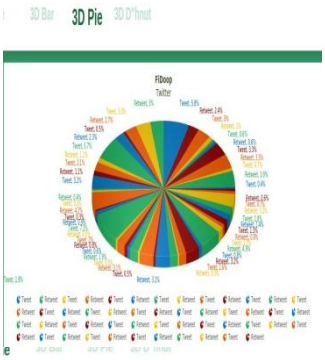


Figure 17 The excepted result obtained in bar graph



.

Figure 18 The result obtained in pie graph

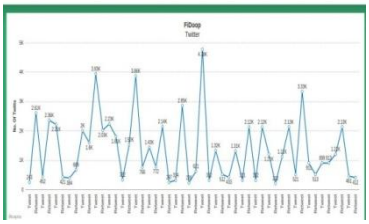


Figure 19 The result obtained in line graph

VI. EPH OVER EXSISTING TOOL

In this thesis I have work on the traditional system that is R tool now in this section. According to the programs done by me I have taken few factors for the comparison with the system.Few factors are:-

- Parallelism: The mapreduce supports the parallelism processing data parallel we can process even 1TB of data in just minute (minimum of 4 nodes) but when consider in the R tool it doesn't provide any parallelism in its structure.
- Datasets: As the data increases we can also include nodes in the cluster which increase the performance of system and so large datasets are processed in the system but In the increase in data the R tool performance is drained.
- Fault tolerance: In mapreduce iif some task is being interrupted then task tracker sends the heartbeat signal to the job tracker and the re-execution is done automatically but in R tool no such facilities.
- CPU Time: This is a little harder to understand, because this time will not change if a job does not change. This is the time consumed by CPU, which is used for processing instructions of a computer program. It is only related to the number of instructions in a program, which means if we parallelize a program, the total execution time might be the third of the original, but the CPU time would not change because the amount of instructions does not change but are executed in parallel.
- Availability : when we store the data in HDFS if there is 1 cluster present but by default 3 replication is created in the HDFS.So if one datasets is duplicated or deleted we can retrieve the data but when compared with the traditional system there is no such advantage is provided.
- Scalable: Hadoop is highly storage platform distributes the data across the many nodes it is not necessary that we install expensive servers but any kind of servers can be installed by this way the storage can be expanded .This is not provided in datamining tool
- Load balancing: The datasets on the cluster is automatically scattered between the node in the cluster without any user instruction if there is more load on single load then it is automatically transferred to the other node where this cannot be done in the traditional tool.

doi : 10.17950/ijer/v5i5/004

- Data distribution :By default the namenode distribute the data between the cluster in the node so user did not any kind instruction for the distribution of the data and in the data mining tool it is not possible.

VII. CONCLUSION

In the paper I applied and worked both on data mining and hadoop as per the result obtained Datamining tool which applied could not manage loadbalancing,datadistrubution ,fault tolerance and many more problems is faced in the datamining tool so this solved by hadoop platform in the mapreduce when compared with the data mining tool mapreduce is advantageous in all the terms. EPH incorporates the parallel mechanism for the twitter data where I can achieve compressed storage and it is necessity not to build any conditional pattern. Firstly in this processes I retrieved the data using the flume which is in unstructured format and later stored in the HDFS and from the HDFS it is then processed to the hive platform and converted to the structured data.

I performed the analysis in the data mining tool I retrieved the data from the twitter data obtained in flume I processed the data in the r tool and we converted the data unstructured to structured data and processed the data to obtain the bar graph where the maximum frequency of words where it was obtained after using a package called as igraph and later the same data was used to obtain the tree graph.

In the hadoop system all services are started the data present in hdfs which is obtained from flume.The important stage of the paper is the mapreduce algorithm which was obtained where their three algorithm implemented mapper,reducer and driver config code was implemented where the data format specified in the for inputting the data is taken and the splitting of the data is done and in map stage it is converted to key value pair and it is sent reduce stage in the reduce stage the aggregation is done calculating the frequency of the tweets data present then later it is processed to the Driver function and aggregation is done. Later I run the mapreduce process as the process gets completed a success file gets created in the HDFS and later we connect the result with the intelligent graph system and result is obtained in term of bar graph,line graph and 3D graph and so on having the maimum tweet and retweet in the table.

VIII. REFERENCES

1. Mangla N and Khola R.K " Optimization of IP Routing with Content Delivery Network" published in "Networking and Information Technology (ICNIT), 2010 InternationalConference "of IEEE Explore,June11-122010, E-ISBN : 978-1-4244-7578-0 pp.424-428 .
2. Apoorv Agarwal, Fadi Biadsy, and Kathleen Mckeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009), pages 24–32, March.
3. Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44.
4. Neha Mangla and Dr. R.K.Khola "A Way to Implement BGP with Geographic Information"published in International Journal of Electronics Engineering, 2 (2), 2010, pp.349–353.
5. NehaMangla and Dr.R.K.Khola," Application Based Route Optimization", IOSR Journal of Engineering (IOSRJEN) ISSN: 2250-3021 Volume 2, Issue 8 (August 2012), PP 78-82.
6. Adam Bermingham and Alan Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage is brevity an advantage? ACM, pages 1833–1836. C. Fellbaum. 1998. Wordnet, an electronic lexicaldatabase. MIT Press.
7. Lammel, R.: Google's MapReduce Programming Model - Revisited. Science of Computer Programming 70, 1–30 (2008)