iJCEM
International Journal of Core Engineering & Management
ISSN: 2348 9510
International Journal Of Core Engineering & Management (IJCEM)
Volume 1, Issue 11, February 2015

Tweet Analysis: Twitter Data processing Using Apache Hadoop

Manoj Kumar Danthala (Author)
Dept. Computer Science Engineering
Keshav Memorial Institute of Technology (KMIT)
Hyderabad, India
manojdanthala@gmail.com

Abstract

'BIG DATA' has been getting much importance in different industries over the last year or two, on a scale that has generated lots of data every day. Big Data is a term applied to data sets of very large size such that the traditional databases are unable to process their operations in a reasonable amount of time. It has tremendous potential to transform business and power in several ways. Here the challenge is not only storing the data, but also accessing and analyzing the required data in specified amount of time. One of the popular implementation to solve the above challenges of big data is using Hadoop. Hadoop is well-known open-source implementation of the MapReduce programming model for processing big data in parallel of data-intensive jobs on clusters of commodity servers. It is highly scalable compute platform. Hadoop enables users to store and process bulk amount which is not possible while using less scalable techniques.

Twitter, one of the largest social media site receives tweets in millions every day in the range of Zettabyte per year. This huge amount of raw data can be used for industrial or business purpose by organizing according to our requirement and processing. This paper provides a way of analyzing of big data such as twitter data using Apache Hadoop which will process and analyze the tweets on a Hadoop clusters. This also includes visualizing the results into a pictorial representations of twitter users and their tweets.

Index Terms— BigData , Hadoop, MapReduce

I. INTRODUCTION

Over past ten years, industries and organizations doesn't need to store and perform much operations and analytics on data of the customers. But around from 2005, the need to transform everything into data is much entertained to satisfy the requirements of the people. So Big data came into picture in the real time business analysis of processing data.

The term big data refers to the data that is generating around us everyday life. It is generally exceeds the capacity of normal conventional traditional databases. For example by combining a large number of signals from the user's actions and those of their friends, Facebook developed the large network area to the users to share their views, ideas and lot many things.

The value of big data to an organizations falls into two categories: analytical use and enabling new products based on the existing ones. Big data can reveal the issues hidden by data that is too costly to process and perform the analytics such as user's transactions, social and geographical data issues faced by the industry.

The major characteristics and challenges of big data are Volume, Velocity, and Variety. These are called as 3V's of big data which are used to characterize different aspects of big data.

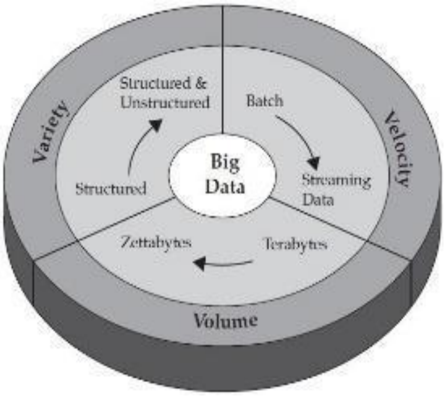


Fig1 : BigData Challenges

Velocity

The Velocity is defined as the speed at which the data is created, modified, retrieved and processed. This is one of the major challenges of big data because the amount of time for processing and performing different operations are considered a lot when dealing with massive amounts of data.

If we use traditional databases for such types of data, then it is useless and doesn't give full satisfaction to the customers in the organizations and industries.

So processing rate of such data is taken into account considerably when talking about big data. Hence volume is one of the big challenges in dealing with big data.

Volume

The second challenge is the volume i.e amount of data which is processed. There are many business areas where we are uploading and processing the data in very high rate in terms of terabytes and zeta bytes. If we take present statistics, every day we are uploading around 25 terabytes of data into facebook, 12 terabytes of data in twitter and around 10 terabytes of data from various devices like RFID, telecommunications and networking.

Here the storing of these huge amounts of data will require high clusters and large servers with high bandwidth. And here the problem is not only storing the information but also the processing at much higher speed. This became the major issue nowadays in most of the companies.

Variety

In the distributed environment there may be the chances of presenting various types of data. This is known as variety of data. These can be categorized as structured, semi structured and unstructured data. The process of analysis and performing operations are varying from one and another. Social media like Facebook posts or Tweets can give different insights, such as sentiment analysis on your brand, while sensory data will give you information about how a product is used and what the mistakes are. So this is the major issue to process information from different sets of data.

Veracity

Big data Veracity refers to the biases, noise and abnormally in data. It is the data that is being stored, and mined meaningful to the problem being analysed. In other words, Veracity can be treated as the uncertainty of data due to data inconsistency and incompleteness, ambiguities, latency, model approximations in the process of analysing data.

II. HADOOP- MAPREDUCE

In the distributed environment, the data should be available to users and capable of performing different analysis from databases in a specified amount of time. If we use the normal approaches, it will be difficult to achieve big data challenges. So these types of data required a specialized platform to deal in such cases.

Hadoop is the one of the solution to solve big data problems. Hadoop is the open source flexible infrastructure for large scale computation and data processing on a network commodity of hardware systems. Here it deals with both structured and unstructured data and various challenges of big data are solved. The main components of Hadoop are commodity hardware and MapReduce.

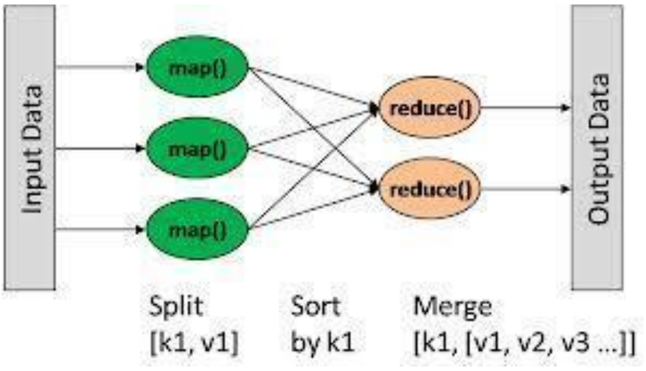


Fig 2: MapReduce Technique

Here MapReduce is the heart of Hadoop and each job is performed by this technique only. It is this programming paradigm that allows for massive scalability across hundreds or thousands of servers in a Hadoop cluster. The term MapReduce refers to two tasks in Hadoop i.e. Map and Reduce. In the first step, it takes the data and converts it into another set of data. Here the each word is referred as key and the number of occurrences is treated as value. So MapReduce tuple consists of key value pairs. Then second step consists of reduce operation where it takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job.

III. PROPOSED SYSTEM

The major issues involved in big data are the following:

The first challenge faced is storing and accessing the information from the large huge amount of data sets from the clusters. We need a standard computing platform to manage large data since the data is growing, and data stores in different data storage locations in a centralized system, which will scale down the huge data into sizable data for computing.

The second challenge is retrieving the data from the large social media data sets. In the scenarios where the data is growing daily, it’s somewhat difficult to accessing the data from the large networks if we want to do specific action to be performed.

The third challenge concentrates on the algorithm design for handling the problems raised by the huge data volume and the dynamic data characteristics.

This paper proposes three modules for finding and performing operation on social media data sets.

The main scope of the project is to analyzing and fetching the Twitter IDs of those users whose statuses have been retweeted the most by the user whose tweets are being analyzed.

First the system involves collecting the tweets from the social network using the twitter API's. Then second, this consists of standard platform as Hadoop to solve the challenges of big data through MapReduce framework where the complete data is mapped to frequent datasets and reduced to smaller sizable data to ease of handling. And finally includes analysing the collected tweets and fetching the Twitter IDs of those users whose statuses have been retweeted the most by the user whose tweets are being analysed.

IV. SYSTEM ARCHITECTURE

The following figure shows the system architecture of the proposed system:

In the first step I collected the Twitter data (Tweets) using API streaming of tokens and apache flume. Then I uploaded the tweets into Hadoop Files Systems by hdfs commands. It includes moving of complete tweets of different users to file systems. And Finally I applied MapReduce Technique to find out the Twitter ID's of the most tweeted people. After performing the MapReduce job, it retrieves the resulted data with the ids.

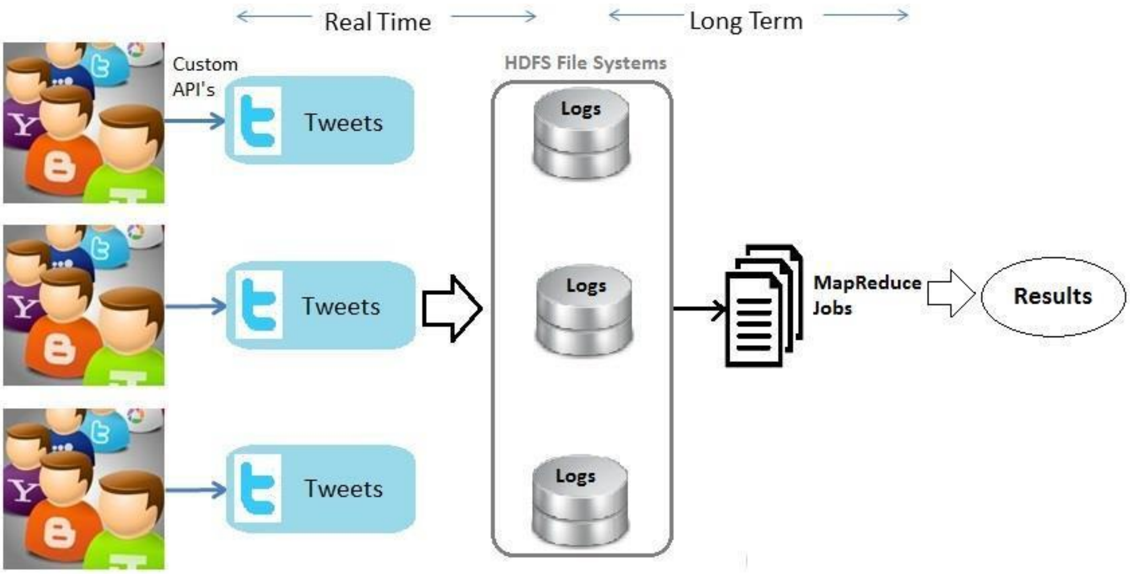


Fig 3: System Architecture

V. APPLICATIONS AND SCOPE

The major applications of the big data are

- Sentiment Analysis: Sentiment data is unstructured data that represents opinions, emotions, and attitudes contained in sources such as social media posts, blogs, online product reviews, and customer support interactions.

  Different companies and Organizations use social media analysis to understand how the public feels about something at a particular moment in time, and also to track how those opinions change over time.
- Text Analytics: It is the process of deriving the high quality information from the raw data such as unstructured data and predicting the analysis.
- Volume Trending: Here volume is estimated in terms of amount of data to process a job. Volume trending is a big issue nowadays. Day by day it has been increasing in a much higher rate in the organizations and social media sites etc.
- Predictive Analytics: Predictive analysis gives the predictive scores to the organizations to help in making the smart decisions and improve business solutions. It optimizes marketing campaigns and website behavior to increase customer responses in business, conversions and meetings, and to decrease the burden on the people. Each customer's predictive score informs actions to be taken with that customer.
- Massively Scalable Architectures:
- Social Media Data: With Hadoop, we can mine Twitter, Facebook and other social media conversations for sentiment data about people and used it to make targeted, real time decisions that increase market share.
- Web Clickstream data: Hadoop makes easy to track customers and their activities in different issues like products purchasing and viewing etc. It makes analyzers to know the behavior and interest of the customers and can able to visualize similar type of products to the customers.

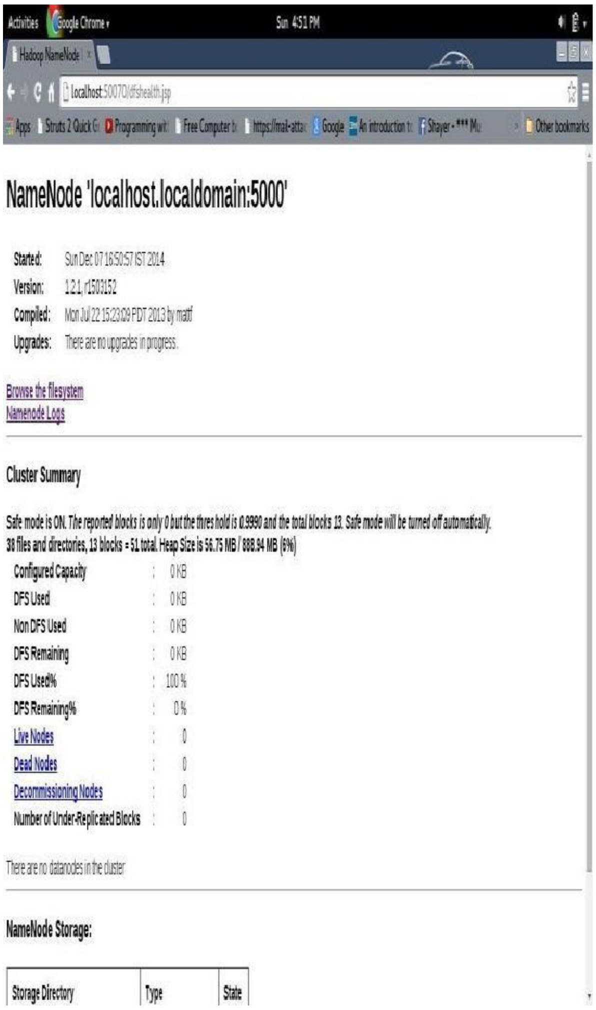
A. Scope

The Proposed system can finds the most popular information about the people, organizations and can be used in the field of analytics.

Applications

- ✓ Finding out the twitter id’s of those persons whose tweets are retweeted number of times.
- ✓ Finds the most number of follows in the social networking sites.
- ✓ This system can be useful to track the business analysis of the organizations.
- ✓ Allows researchers to retrieve and analyze the data easily from large datasets.

VI. SCREENSHOTS



NameNode 'localhost.localdomain:5000'

Started:	Sun Dec 07 16:50:57 IST 2014
Version:	1.2.1, r1503152
Compiled:	Mon Jul 22 15:23:09 PDT 2013 by mattf
Upgrades:	There are no upgrades in progress.

Browse the filesystem
Namenode Logs

Cluster Summary

Safe mode is ON. *The reported blocks is only 0 but the thres hold is 0.9990 and the total blocks 13. Safe mode will be turned off automatically.*
38 files and directories, 13 blocks = 51 total. Heap Size is 56.75 MB / 888.94 MB (6%)

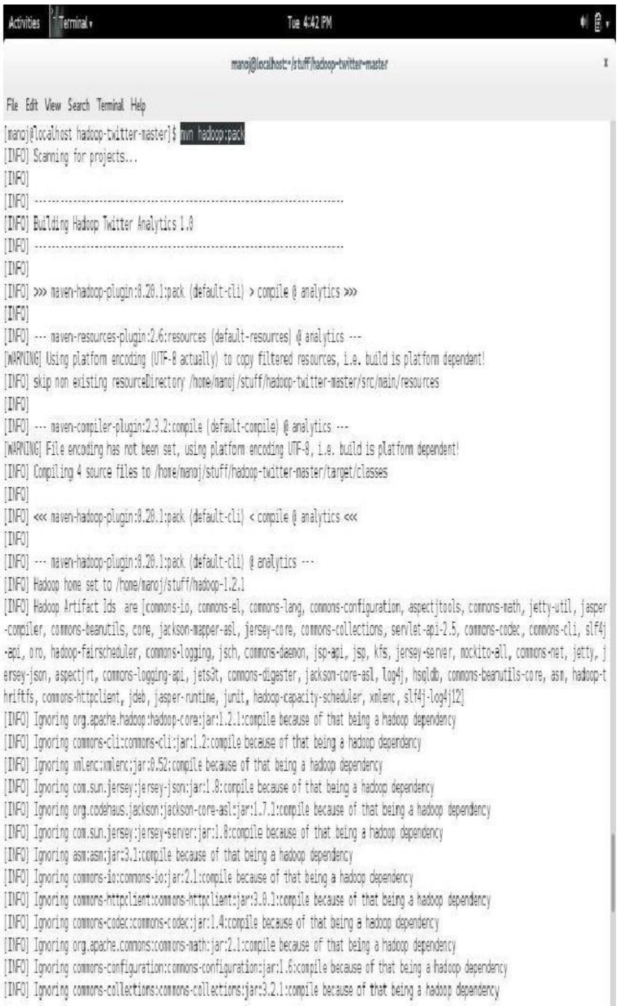
Configured Capacity	:	0 KB
DFS Used	:	0 KB
Non DFS Used	:	0 KB
DFS Remaining	:	0 KB
DFS Used%	:	100 %
DFS Remaining%	:	0 %
Live Nodes	:	0
Dead Nodes	:	0
Decommissioning Nodes	:	0
Number of Under-Replicated Blocks	:	0

There are no datanodes in the cluster

NameNode Storage:

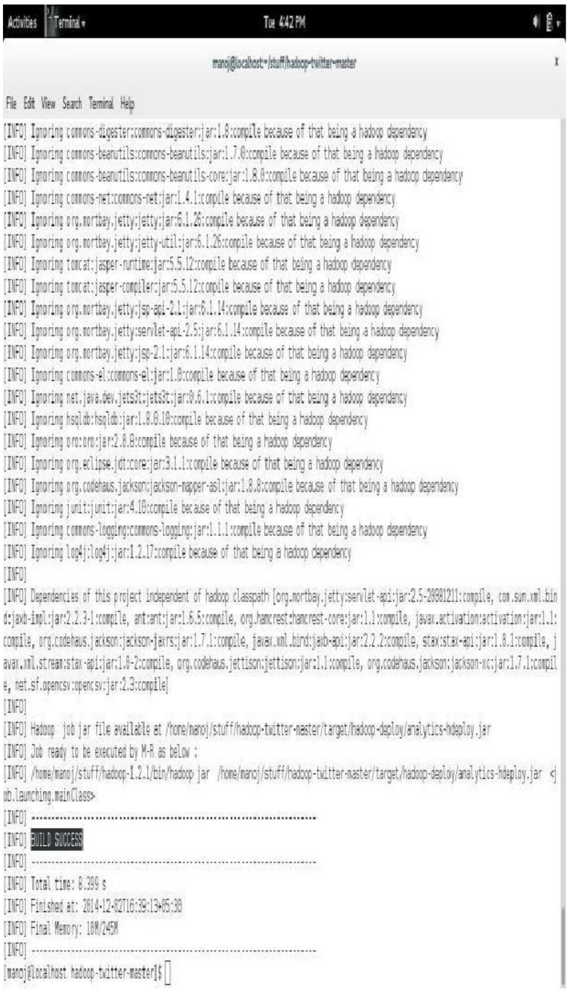
Storage Directory	Type	State

A. Hadoop Cluster Setup


Activities
Terminal
Tue 4:42 PM
manoj@localhost:~/stuff/hadoop-twitter-master
File Edit View Search Terminal Help

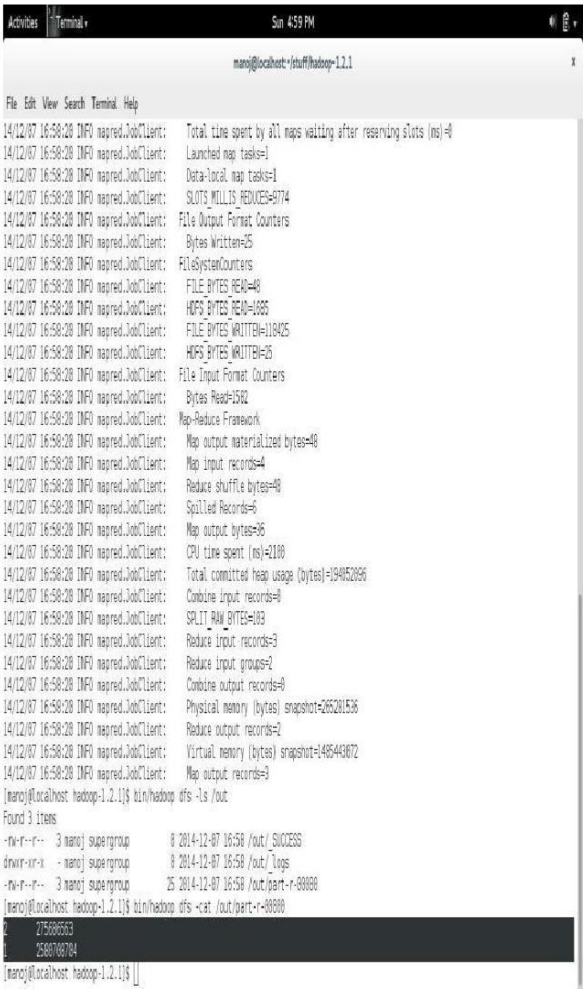
```
[manoj@localhost hadoop-twitter-master]$ mvn hadoop:pack
[INFO] Scanning for projects...
[INFO]
[INFO] ------------------------------------------------------------------------
[INFO] Building Hadoop Twitter Analytics 1.0
[INFO] ------------------------------------------------------------------------
[INFO]
[INFO] >>> maven-hadoop-plugin:0.20.1:pack (default-cli) > compile @ analytics >>>
[INFO]
[INFO] --- maven-resources-plugin:2.6:resources (default-resources) @ analytics ---
[WARNING] Using platform encoding (UTF-8 actually) to copy filtered resources, i.e. build is platform dependent!
[INFO] skip non existing resourceDirectory /home/manoj/stuff/hadoop-twitter-master/src/main/resources
[INFO]
[INFO] --- maven-compiler-plugin:2.3.2:compile (default-compile) @ analytics ---
[WARNING] File encoding has not been set, using platform encoding UTF-8, i.e. build is platform dependent!
[INFO] Compiling 4 source files to /home/manoj/stuff/hadoop-twitter-master/target/classes
[INFO]
[INFO] <<< maven-hadoop-plugin:0.20.1:pack (default-cli) < compile @ analytics <<<
[INFO]
[INFO] --- maven-hadoop-plugin:0.20.1:pack (default-cli) @ analytics ---
[INFO] Hadoop home set to /home/manoj/stuff/hadoop-1.2.1
[INFO] Hadoop Artifact Ids  are [commons-io, commons-el, commons-lang, commons-configuration, aspectjtools, commons-math, jetty-util, jasper
-compiler, commons-beanutils, core, jackson-mapper-asl, jersey-core, commons-collections, servlet-api-2.5, commons-codec, commons-cli, slf4j
-api, oro, hadoop-fairscheduler, commons-logging, jsch, commons-daemon, jsp-api, jsp, kfs, jersey-server, mockito-all, commons-net, jetty, j
ersey-json, aspectjrt, commons-logging-api, jets3t, commons-digester, jackson-core-asl, log4j, hsqldb, commons-beanutils-core, asm, hadoop-t
hriftfs, commons-httpclient, jdeb, jasper-runtime, junit, hadoop-capacity-scheduler, xmlenc, slf4j-log4j12]
[INFO] Ignoring org.apache.hadoop:hadoop-core:jar:1.2.1:compile because of that being a hadoop dependency
[INFO] Ignoring commons-cli:commons-cli:jar:1.2:compile because of that being a hadoop dependency
[INFO] Ignoring xmlenc:xmlenc:jar:0.52:compile because of that being a hadoop dependency
[INFO] Ignoring com.sun.jersey:jersey-json:jar:1.8:compile because of that being a hadoop dependency
[INFO] Ignoring org.codehaus.jackson:jackson-core-asl:jar:1.7.1:compile because of that being a hadoop dependency
[INFO] Ignoring com.sun.jersey:jersey-server:jar:1.8:compile because of that being a hadoop dependency
[INFO] Ignoring asm:asm:jar:3.1:compile because of that being a hadoop dependency
[INFO] Ignoring commons-io:commons-io:jar:2.1:compile because of that being a hadoop dependency
[INFO] Ignoring commons-httpclient:commons-httpclient:jar:3.0.1:compile because of that being a hadoop dependency
[INFO] Ignoring commons-codec:commons-codec:jar:1.4:compile because of that being a hadoop dependency
[INFO] Ignoring org.apache.commons:commons-math:jar:2.1:compile because of that being a hadoop dependency
[INFO] Ignoring commons-configuration:commons-configuration:jar:1.6:compile because of that being a hadoop dependency
[INFO] Ignoring commons-collections:commons-collections:jar:3.2.1:compile because of that being a hadoop dependency
```

B. Installing Package


Activities
Terminal
Tue 4:42 PM
manoj@localhost:~/stuff/hadoop-twitter-master
File Edit View Search Terminal Help

```
[INFO] Ignoring commons-digester:commons-digester:jar:1.8:compile because of that being a hadoop dependency
[INFO] Ignoring commons-beanutils:commons-beanutils:jar:1.7.0:compile because of that being a hadoop dependency
[INFO] Ignoring commons-beanutils:commons-beanutils-core:jar:1.8.0:compile because of that being a hadoop dependency
[INFO] Ignoring commons-net:commons-net:jar:1.4.1:compile because of that being a hadoop dependency
[INFO] Ignoring org.mortbay.jetty:jetty:jar:6.1.26:compile because of that being a hadoop dependency
[INFO] Ignoring org.mortbay.jetty:jetty-util:jar:6.1.26:compile because of that being a hadoop dependency
[INFO] Ignoring tomcat:jasper-runtime:jar:5.5.12:compile because of that being a hadoop dependency
[INFO] Ignoring tomcat:jasper-compiler:jar:5.5.12:compile because of that being a hadoop dependency
[INFO] Ignoring org.mortbay.jetty:jsp-api-2.1:jar:6.1.14:compile because of that being a hadoop dependency
[INFO] Ignoring org.mortbay.jetty:servlet-api-2.5:jar:6.1.14:compile because of that being a hadoop dependency
[INFO] Ignoring org.mortbay.jetty:jsp-2.1:jar:6.1.14:compile because of that being a hadoop dependency
[INFO] Ignoring commons-el:commons-el:jar:1.0:compile because of that being a hadoop dependency
[INFO] Ignoring net.java.dev.jets3t:jets3t:jar:0.6.1:compile because of that being a hadoop dependency
[INFO] Ignoring hsqldb:hsqldb:jar:1.8.0.10:compile because of that being a hadoop dependency
[INFO] Ignoring oro:oro:jar:2.0.8:compile because of that being a hadoop dependency
[INFO] Ignoring org.eclipse.jdt:core:jar:3.1.1:compile because of that being a hadoop dependency
[INFO] Ignoring org.codehaus.jackson:jackson-mapper-asl:jar:1.8.8:compile because of that being a hadoop dependency
[INFO] Ignoring junit:junit:jar:4.10:compile because of that being a hadoop dependency
[INFO] Ignoring commons-logging:commons-logging:jar:1.1.1:compile because of that being a hadoop dependency
[INFO] Ignoring log4j:log4j:jar:1.2.17:compile because of that being a hadoop dependency
[INFO]
[INFO] Dependencies of this project independent of hadoop classpath [org.mortbay.jetty:servlet-api:jar:2.5-20081211:compile, com.sun.xml.bin
d:jaxb-impl:jar:2.2.3-1:compile, ant:ant:jar:1.6.5:compile, org.hamcrest:hamcrest-core:jar:1.1:compile, javax.activation:activation:jar:1.1:
compile, org.codehaus.jackson:jackson-jaxrs:jar:1.7.1:compile, javax.xml.bind:jaxb-api:jar:2.2.2:compile, stax:stax-api:jar:1.0.1:compile, j
avax.xml.stream:stax-api:jar:1.0-2:compile, org.codehaus.jettison:jettison:jar:1.1:compile, org.codehaus.jackson:jackson-xc:jar:1.7.1:compil
e, net.sf.opencsv:opencsv:jar:2.3:compile]
[INFO]
[INFO] Hadoop  job jar file available at /home/manoj/stuff/hadoop-twitter-master/target/hadoop-deploy/analytics-hdeploy.jar
[INFO] Job ready to be executed by M-R as below :
[INFO] /home/manoj/stuff/hadoop-1.2.1/bin/hadoop jar  /home/manoj/stuff/hadoop-twitter-master/target/hadoop-deploy/analytics-hdeploy.jar  <j
ob.launching.mainClass>
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time: 8.399 s
[INFO] Finished at: 2014-12-02T16:39:13+05:30
[INFO] Final Memory: 10M/245M
[INFO] ------------------------------------------------------------------------
[manoj@localhost hadoop-twitter-master]$ 
```

C. Twitter Setup Packages


Activities
Terminal
Sun 4:59 PM
manoj@localhost:~/stuff/hadoop-1.2.1
File Edit View Search Terminal Help

```
14/12/07 16:58:28 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
14/12/07 16:58:28 INFO mapred.JobClient:     Launched map tasks=1
14/12/07 16:58:28 INFO mapred.JobClient:     Data-local map tasks=1
14/12/07 16:58:28 INFO mapred.JobClient:     SLOTS_MILLIS_REDUCES=9774
14/12/07 16:58:28 INFO mapred.JobClient:   File Output Format Counters
14/12/07 16:58:28 INFO mapred.JobClient:     Bytes Written=25
14/12/07 16:58:28 INFO mapred.JobClient:   FileSystemCounters
14/12/07 16:58:28 INFO mapred.JobClient:     FILE_BYTES_READ=48
14/12/07 16:58:28 INFO mapred.JobClient:     HDFS_BYTES_READ=1685
14/12/07 16:58:28 INFO mapred.JobClient:     FILE_BYTES_WRITTEN=118425
14/12/07 16:58:28 INFO mapred.JobClient:     HDFS_BYTES_WRITTEN=25
14/12/07 16:58:28 INFO mapred.JobClient:   File Input Format Counters
14/12/07 16:58:28 INFO mapred.JobClient:     Bytes Read=1582
14/12/07 16:58:28 INFO mapred.JobClient:   Map-Reduce Framework
14/12/07 16:58:28 INFO mapred.JobClient:     Map output materialized bytes=48
14/12/07 16:58:28 INFO mapred.JobClient:     Map input records=4
14/12/07 16:58:28 INFO mapred.JobClient:     Reduce shuffle bytes=48
14/12/07 16:58:28 INFO mapred.JobClient:     Spilled Records=6
14/12/07 16:58:28 INFO mapred.JobClient:     Map output bytes=36
14/12/07 16:58:28 INFO mapred.JobClient:     CPU time spent (ms)=2100
14/12/07 16:58:28 INFO mapred.JobClient:     Total committed heap usage (bytes)=194052096
14/12/07 16:58:28 INFO mapred.JobClient:     Combine input records=0
14/12/07 16:58:28 INFO mapred.JobClient:     SPLIT_RAW_BYTES=103
14/12/07 16:58:28 INFO mapred.JobClient:     Reduce input records=3
14/12/07 16:58:28 INFO mapred.JobClient:     Reduce input groups=2
14/12/07 16:58:28 INFO mapred.JobClient:     Combine output records=0
14/12/07 16:58:28 INFO mapred.JobClient:     Physical memory (bytes) snapshot=265281536
14/12/07 16:58:28 INFO mapred.JobClient:     Reduce output records=2
14/12/07 16:58:28 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=1485443072
14/12/07 16:58:28 INFO mapred.JobClient:     Map output records=3
[manoj@localhost hadoop-1.2.1]$ bin/hadoop dfs -ls /out
Found 3 items
-rw-r--r--   3 manoj supergroup          0 2014-12-07 16:58 /out/_SUCCESS
drwxr-xr-x   - manoj supergroup          0 2014-12-07 16:58 /out/_logs
-rw-r--r--   3 manoj supergroup         25 2014-12-07 16:58 /out/part-r-00000
[manoj@localhost hadoop-1.2.1]$ bin/hadoop dfs -cat /out/part-r-00000
2       275686563
1       2500708704
[manoj@localhost hadoop-1.2.1]$ 
```

D. Results of the System

VII. FUTURE WORK

Nowadays big data has become the buzzword in IT industry organizations. The need of analysing and processing of information has grown a lot. This paper implemented the analysing of big data (tweets) only for text. Further analysis can be done to images and all types of multimedia files based on index support. The result of Text mining and data analysis would help in suggesting related pages based on different types of data. So that industries make the data easily available to people who is using and trying accessing such type of data.

VIII. CONCLUSION

Traditional Enterprise Data Warehouses do not have the ability to keep up with rapidly increasing social media data. With this system, one can build a dashboard to monitor the sentiment of Twitter traffic around any given topic in near real-time (that is, with a delay of 1-2 minutes), allowing you or your users to take advantage of near real-time Twitter sentiment for business insights or any other purpose.

There are several ways to define and analyse the social media data such as facebook, Twitter etc. Here anyone can perform different operations queries in these type of data. But the problem arises when dealing with bigdata of several types of unstructured data. Here it is solved by using Hadoop and its packages. And we have done some analysis on the tweets and the most number of tweet ids.

So it is concluded that processing time and retrieving capabilities are made very easy when compared to other processing and analysing techniques for large amounts of data.

IX REFERENCES

[1] Paul C. Zikopoulos, Chris Eaton, Dirk deRoos "Understanding Big Data", ISBN 978-07179053-6.

[2] Penchalaiah.C, Murali.GSuresh Babu.A, Effective Sentiment Analysis on Twitter Data using: Apache Flume and Hive, Computer Science and EngineeringDept, JNTUACEP, Pulivendula, Vol. 1 Issue 8, October 2014.

[3] Mr. Swapnil A. Kale, Prof. Sangram S.Dandge, Understanding the Big Data problems and their solutions using Hadoop MapReduce, ISSN 2319 – 4847,Volume 3.

[4] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in the 26th IEEE Symposium on Mass Storage Systems and Technologies, pp. 1-10, May 2010.

[5] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Communications of the ACM, Vol. 51, Iss. 1, pp. 107-113, January 2008.

[6] T. White, "The Hadoop Distributed Filesystem," Hadoop: The Definitive Guide, pp. 41-73, GravensteinHighwaNorth, Sebastopol: O'Reilly Media, Inc., 2010.

[7] Chansup Byun, William Arcand, David Bestor, Bill Bergeron, Matthew Hubbell, Jeremy Kepner, Andrew McCabe, Peter Michaleas, Julie Mullen, David O'Gwynn, Andrew Prout, Albert Reuther, Antonio Rosa, Charles Yee, " Driving Big Data With Big Compute", MIT Lincoln Laboratory, Lexington, MA, U.S.A.

[8](OnlineResource) http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[9](OnlineResource)http://hadoop.apache.org/docs/r2.5.0/hadoop -project-dist/hadoop- common/SingleCluster.html