Usage of Hadoop on Twitter Data Analysis on Natural Disaster Management System

M.V.Sangameswar, Dr. M.Nagabhushana Rao, N.S.Murthy

Abstract The occurrence of the natural disaster or national event emanates plenty of information through social media. The data generated by various social media is mostly vocal, which will not be useful for providing relief for the victims of the natural calamity. The vocal data prevents the victims from getting relief help. One of the most popular social media Twitter receives everyday data to the tune of zettabytes per annum. Meticulous use of such huge data helps for the developmental program of the country in different areas namely is Business, Industry, Education, Medicine etc. Data analysis of data emanating from twitter is HADOOP since it works for BIGDATA. An attempt is made in this paper to discuss the utility of flume and HDFS on the analysis of twitter data. Real time twitter data is extracted to HDFS through FLUME. A query language synnonium to the Hive is utilized for extraction and analysis of some data.

Index Terms — Big Data, Apache Hadoop, MapReduce, HDFS, FLUME, HIVE and Twitter.

I. Introduction

The Internet transmits textual vales. Many products manufactures use internet textual data to elicit views of their products. An automatic tabulation can handle large datasets for the analysis of subjective data. Nowadays users have a choice to express their views and opinions on any topic on social media websites. These media also get the comments and grades on the response. Twitter data is used by the websites to make blogs and forums and product review. Generally twitter messages are brief observations of status messages and product reviews. The popularity of the messages displayed is rank and compare in the author's opinion.

Data arriving in tweets with high frequency requires analysis in view of the limitations of storage. Sentiment analysis uses twitter data as a source to classify the messages with positive or negative feelings since manual classification is costly and time-consuming. The time and authors identification are provided in twitter data. To express the level of sentiments smileys or emoticons provide the emotion state. Those referred tweets provide the basis for sentiment analysis.

Sentimental analysis on twitter data is not in an easy job, in comparison to review data. Generally, tweets are brief, power loaded slangs, emoticons with usual jargons of twitter. Twitter provides Application Programming Interface (API) providing the developer analysis of 1% tweets with any particular keyword. The API does mining of response with respect to the keyword. Twitter data is not structured and has abbreviations and emoticons providing a view on author’s response. The availability of the user's location in the tweets with help to compare the response trends in the different geographic area.

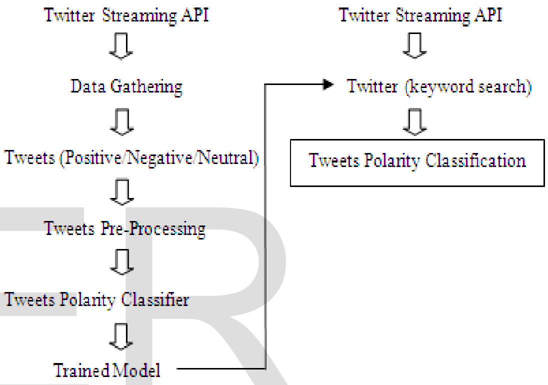


Fig 1: Twitter Analysis Steps

Twitter data analysis consists of

a) Collection of data in local HDFS using FLUME
b) Remove noises and meaningless symbols
c) Extract feature vector utilizing unigram or Ngram
d) Use HIVE for post analysis of twitter.

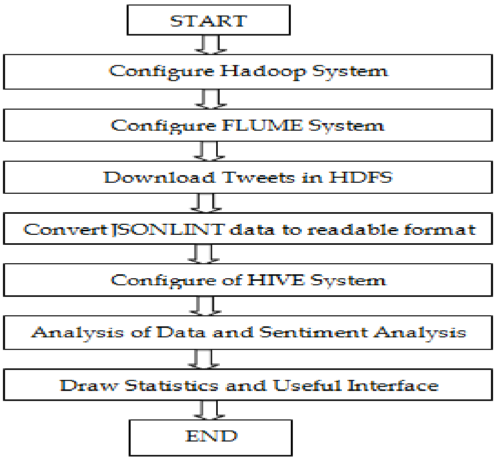

START
Configure Hadoop System
Configure FLUME System
Download Tweets in HDFS
Convert JSONLINT data to readable format
Configure of HIVE System
Analysis of Data and Sentiment Analysis
Draw Statistics and Useful Interface
END

Fig 1: Framework for Twitter Analysis

1) M.V.Sangameswar, Research Scholar, Rayalaseema University, Kurnool, Andhra Pradesh, India, sangamrjy@gmail.com
2) Dr. M.Nagabhushana Rao, Professor in CSE Department, K.L.University-Vijayawada, Andhra Pradesh, India,mnraosir@gmail.com
3) N.S.Murthy, Professor in HBS Department, Godavari Institute of Engineering & Technology, Rajahmundry, Andhra Pradesh, India, nsmurthy50@gmail.com

II. Introduction Big Data and Hadoop

Data sets having huge volume, high speed and varied data which increases daily are called BIGDATA. Usually, known data management techniques are difficult to apply for BIGDATA. HADOOP data analysis pack introduced by Apache will solve BIGDATA. The accuracy of the data analysis will be high if more data is used as it will enhance the confidence on the conclusion drawn. Further, it will increase the efficiency in various operations and reduced cost and time. A good instrument for twitter analysis is apache HADOOP for a large volume of data. For processing the large volume of data in distributed mode HADOOP is a vital tool.

Characteristics

1. Optimum handling of the huge quantity of unstructured data using inexpensive hardware.
2. Uses relatively less expensive computers
3. Data is replicated across multiple computers for the ease of computation on any machine

Computing data in a distributed mode is multifaceted. The features of HADOOP are given below.

- Analysis on a number of machines with computing facility from the cloud.
- It runs on fewer expenses Hardware and overcomes the nonfunctioning of it.
- Can handle large volumes of data with the addition of more systems to the group.
- Simplicity and accessibility of HADOOP provides a superior order for writing and running large programs.

Hadoop is optimized to handle massive quantities of structured, semi-structured, and unstructured data. Hadoop clones the data in multiple systems. In case of one system goes down data is retrieved from other system.

Another data processing model is MapReduce. Data processing primitives are Mappers and Reducers in MapReduce. Dividing data processing applications into Mappers and Reducers is not easy in data processing. The advantage of MapReduce form is in scaling. The program can be run over hundreds of machines in a group by changing the merely configuration. Many programmers prefer to MapReduce model for scalability.

The components of are represented in the following diagram. We are using HIVE and FLUME for twitter analysis.

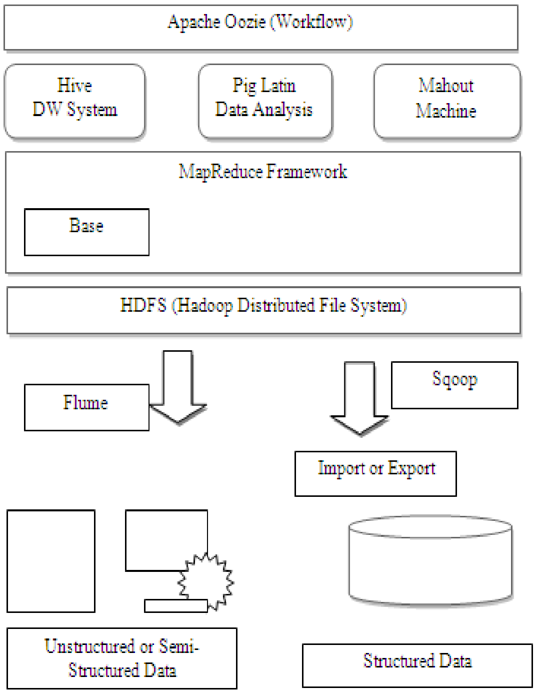


Fig 3: Hadoop Eco-Components

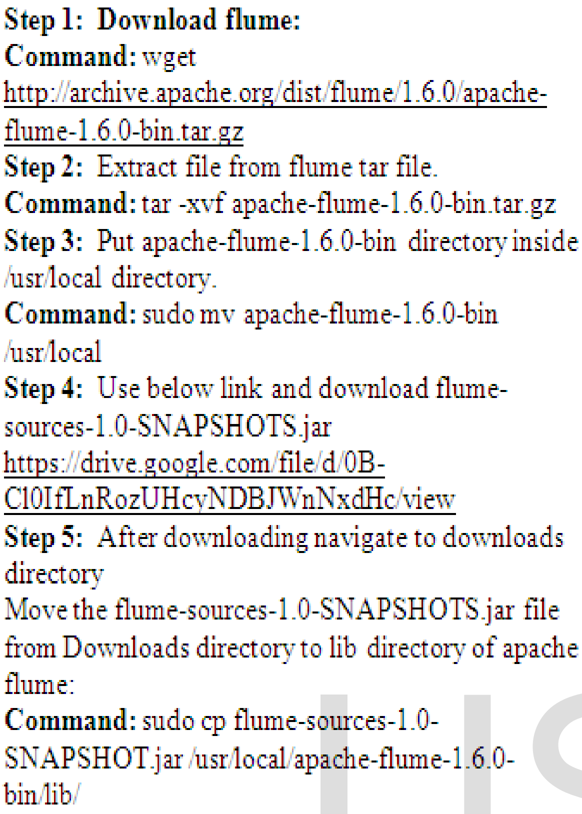
Basically, HADOOP file system makes use of Google file system. Further, it produces a file system of distributed nature for running and a group of computers providing reliable and consistence results.

The utility of masters/slave architecture is employed in HDFS. The file system metadata is managed by NameNode and actual data is stored by slave DataNodes. HADOOP has a special advantage in processing and storage in distributed Storage. Data in HADOOP is secure, reliable, efficient, speedy, scalable and data accessible. The use of HADOOP for tweet processing is popular due to the above qualities.

III. FLUME

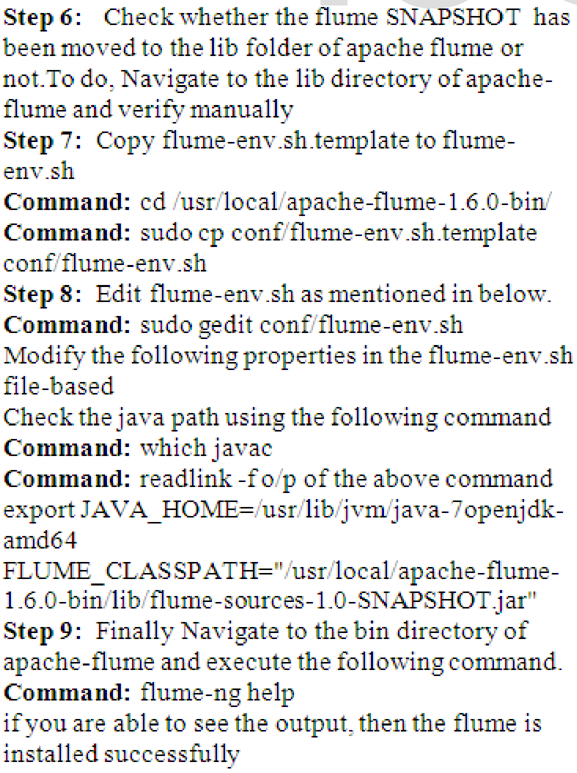
Collecting, aggregating and transporting a large amount of incoming data done by Apache Flume dealing with large files from different sources of central data storage. Primary Apache Flume is designed to copy log data from different web servers to HDFS. The package is reliable, distributed and configurable tool. In addition to transportation network traffic data of social media, emails and any other data from any source Flume can transport large quantitative of event data.

For a single node, the FLUME is installed after installing VMware and HADOOP.



Step 1: Download flume:
Command: wget http://archive.apache.org/dist/flume/1.6.0/apache-flume-1.6.0-bin.tar.gz
Step 2: Extract file from flume tar file.
Command: tar -xvf apache-flume-1.6.0-bin.tar.gz
Step 3: Put apache-flume-1.6.0-bin directory inside /usr/local directory.
Command: sudo mv apache-flume-1.6.0-bin /usr/local
Step 4: Use below link and download flume-sources-1.0-SNAPSHOTS.jar https://drive.google.com/file/d/0B-Cl0IfLnRozUHcyNDBJWnNxdHc/view
Step 5: After downloading navigate to downloads directory
Move the flume-sources-1.0-SNAPSHOTS.jar file from Downloads directory to lib directory of apache flume:
Command: sudo cp flume-sources-1.0-SNAPSHOT.jar /usr/local/apache-flume-1.6.0-bin/lib/

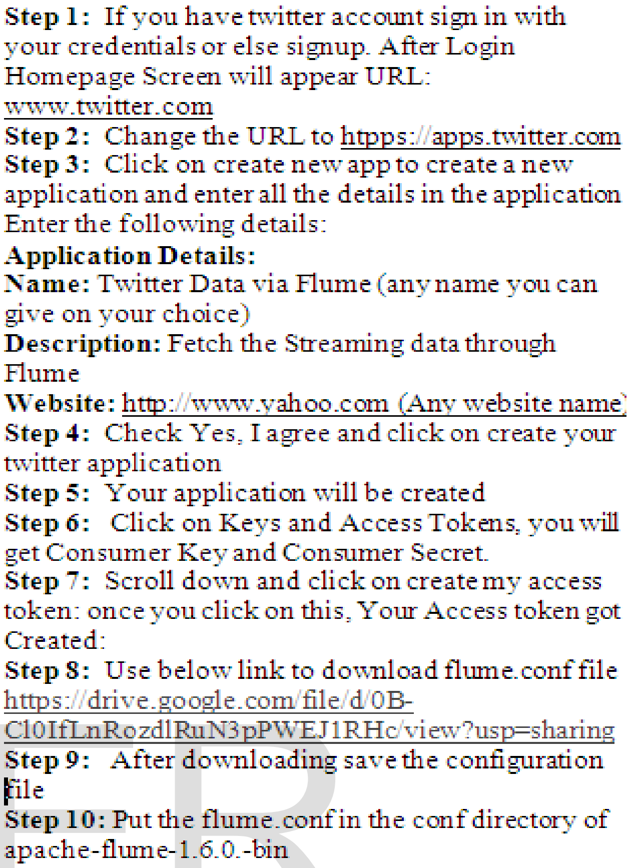
Fig 4(A): Flume Instalation



Step 6: Check whether the flume SNAPSHOT has been moved to the lib folder of apache flume or not.To do, Navigate to the lib directory of apache-flume and verify manually
Step 7: Copy flume-env.sh.template to flume-env.sh
Command: cd /usr/local/apache-flume-1.6.0-bin/
Command: sudo cp conf/flume-env.sh.template conf/flume-env.sh
Step 8: Edit flume-env.sh as mentioned in below.
Command: sudo gedit conf/flume-env.sh
Modify the following properties in the flume-env.sh file-based
Check the java path using the following command
Command: which javac
Command: readlink -f o/p of the above command
export JAVA_HOME=/usr/lib/jvm/java-7openjdk-amd64
FLUME_CLASSPATH="/usr/local/apache-flume-1.6.0-bin/lib/flume-sources-1.0-SNAPSHOT.jar"
Step 9: Finally Navigate to the bin directory of apache-flume and execute the following command.
Command: flume-ng help
if you are able to see the output, then the flume is installed successfully

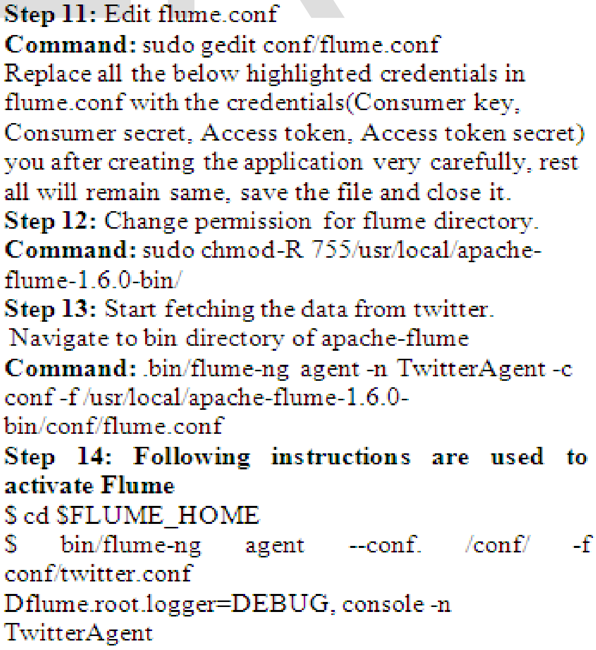
Fig 4(B): Flume Instalation

STREAMING TWITTER DATA:



Step 1: If you have twitter account sign in with your credentials or else signup. After Login Homepage Screen will appear URL: www.twitter.com
Step 2: Change the URL to htpps://apps.twitter.com
Step 3: Click on create new app to create a new application and enter all the details in the application Enter the following details:
Application Details:
Name: Twitter Data via Flume (any name you can give on your choice)
Description: Fetch the Streaming data through Flume
Website: http://www.yahoo.com (Any website name)
Step 4: Check Yes, I agree and click on create your twitter application
Step 5: Your application will be created
Step 6: Click on Keys and Access Tokens, you will get Consumer Key and Consumer Secret.
Step 7: Scroll down and click on create my access token: once you click on this, Your Access token got Created:
Step 8: Use below link to download flume.conf file https://drive.google.com/file/d/0B-Cl0IfLnRozdlRuN3pPWEJ1RHc/view?usp=sharing
Step 9: After downloading save the configuration file
Step 10: Put the flume.conf in the conf directory of apache-flume-1.6.0.-bin

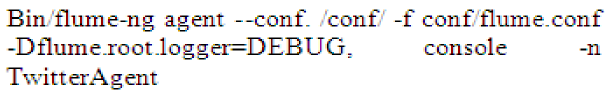
Fig 5(A) : Streaming Twitter Data



Step 11: Edit flume.conf
Command: sudo gedit conf/flume.conf
Replace all the below highlighted credentials in flume.conf with the credentials(Consumer key, Consumer secret, Access token, Access token secret) you after creating the application very carefully, rest all will remain same, save the file and close it.
Step 12: Change permission for flume directory.
Command: sudo chmod-R 755/usr/local/apache-flume-1.6.0-bin/
Step 13: Start fetching the data from twitter.
Navigate to bin directory of apache-flume
Command: .bin/flume-ng agent -n TwitterAgent -c conf -f /usr/local/apache-flume-1.6.0-bin/conf/flume.conf
Step 14: Following instructions are used to activate Flume
$ cd $FLUME_HOME
$ bin/flume-ng agent --conf. /conf/ -f conf/twitter.conf Dflume.root.logger=DEBUG, console -n TwitterAgent

Fig 5(B) : Streaming Twitter Data

Login twitter creates a new application after logging. The sink is configured as HDFS and its searches to store tweets. After execution it follows a specified path for downloading HDFS. To initiate this following commands are to be executed



```
Bin/flume-ng agent --conf. /conf/ -f conf/flume.conf
-Dflume.root.logger=DEBUG,      console      -n
TwitterAgent
```

Fig 6: Start flume using the above command

Twitter appears in HDFC after some time. Refresh if tweets are not downloaded in the prescribed part. Data remains temporarily in container/channel and tweets starts downloading in HDFS. Convert JSON data to the readable format through jsonserde.jar.

IV. HIVE

Hive is a data warehousing tool. Hive is used to query structured data built on top of Hadoop. Facebook created hive components to manage their ever-growing volumes of log data hive makes HDFS for storage, MapReduce for execution and stores metadata in an RDBMS. Hive provides HQL (High Query Language) which is similar to SQL. HQL is easy to code. These queries are implied by Hive with the help of MapReduce and stores in Hadoop. Hive provides extensive data type functions and formats and data summarization and analysis. Hive supports rich data types such as structs, lists, and maps. Hive supports SQL filters, group-by, and order-by clauses. Hourly log data can be stored directly into HDFS and then data cleansing is performed on the log file. Finally, hive table(s) can be created to query the log file.

HIVE DATA UNITS:

1. Databases: the namespace for tables.
2. Tables: a set of records that have a similar schema.
3. Partitions: logically separations of data based on the classification of given information as per specific at -tributes. Once the hive has partitioned the data based on a specified key, it starts to assemble the records into specific folders as and when the records are inserted
4. Buckets (or clusters): similar to partitions but uses hash functions to segregate data and determines the cluster or bucket into which the record should be placed.

HIVE FILE FORMAT:

TEXT FILE: The default file format is textfile. In this format, each record is a line in the file. In the text file, different control characters are used as delimiters. The delimiters are ^A (octal 001, separates all fields), ^B (octal 002, separates the elements in the array), ^C (octal 003, separates key value pair), and \n. The term field is used then overriding the default delimiter. The supported text files are CSV and TSV, JSON or XML documents too can be specified as text file

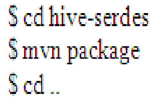
SEQUENTIAL FILE: sequential files are flat files that store binary key-value pairs. It includes compression support which reduces the CPU, I/O requirement.

Record Columnar file (RCFILE): RCFile stores the data in the column-oriented manner which ensures that Aggregation operations, not an expensive operation.

V. JSON SERDE

In order to interpret JSON data properly make sure the HIVE table coordinates the job for a query. HIVE process input data in a row format whereas it is getting the twitter in data in Jason format. This makes Hive not able to perform this job. DeSerializer interfaces HIVE to transmit the data into HIVE for the process.

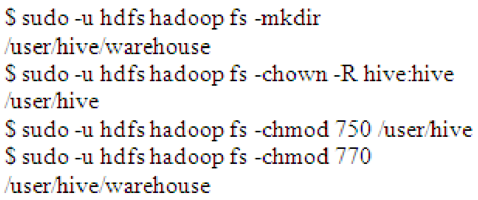
To construct the hive-serdes JAR, beyond the root regarding the git repository



```
$ cd hive-serdes
$ mvn package
$ cd ..
```

Fig 7: Build hive-serde Jar

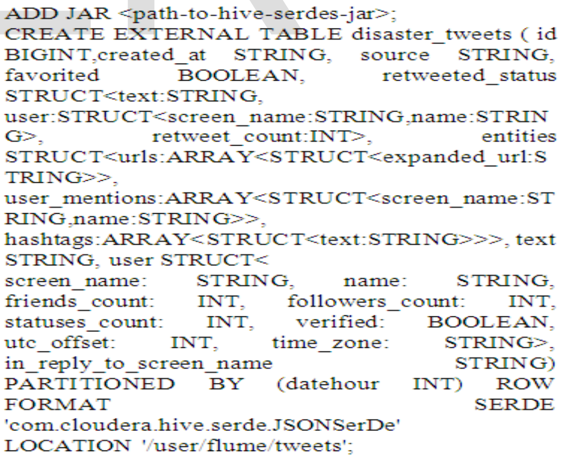
This will grow a file for consideration referred to as hive-serdes-1.0-SNAPSHOT.jar in the goal directory. After that Create the Hive directory hierarchy the use of the say the word below:



```
$ sudo -u hdfs hadoop fs -mkdir
/user/hive/warehouse
$ sudo -u hdfs hadoop fs -chown -R hive:hive
/user/hive
$ sudo -u hdfs hadoop fs -chmod 750 /user/hive
$ sudo -u hdfs hadoop fs -chmod 770
/user/hive/warehouse
```

Fig 8: Create Hive Directory Hierarchy

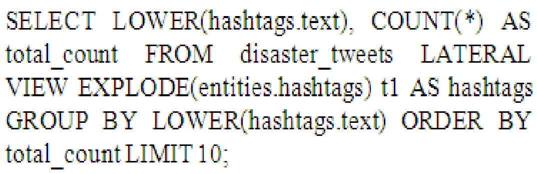
The configuration of hive metastore is to use MySQL. Install MySQL, JDBC Driver in the location /var/lib/hive/lib. After creating tweet table run Hive and perform the subsequent instructions.



```
ADD JAR <path-to-hive-serdes-jar>;
CREATE EXTERNAL TABLE disaster_tweets ( id
BIGINT,created_at   STRING,   source   STRING,
favorited       BOOLEAN,       retweeted_status
STRUCT<text:STRING,
user:STRUCT<screen_name:STRING,name:STRIN
G>,        retweet_count:INT>,          entities
STRUCT<urls:ARRAY<STRUCT<expanded_url:S
TRING>>,
user_mentions:ARRAY<STRUCT<screen_name:ST
RING,name:STRING>>,
hashtags:ARRAY<STRUCT<text:STRING>>>, text
STRING, user STRUCT<
screen_name:    STRING,    name:    STRING,
friends_count:   INT,   followers_count:   INT,
statuses_count:   INT,   verified:   BOOLEAN,
utc_offset:    INT,    time_zone:    STRING>,
in_reply_to_screen_name                 STRING)
PARTITIONED    BY    (datehour    INT)    ROW
FORMAT                                   SERDE
'com.cloudera.hive.serde.JSONSerDe'
LOCATION '/user/flume/tweets';
```

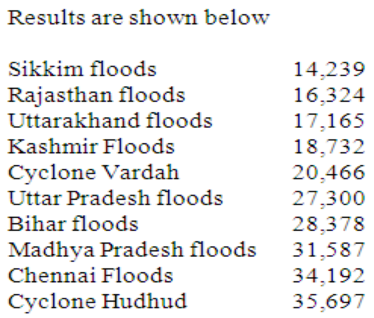
Fig 9: Create an External Table

Data on Natural Disasters, Chennai Floods, and HudHud cyclone are tweeted in tweets table. Ten common hashtags are accessed on the data to execute this query was asked.



```
SELECT  LOWER(hashtags.text),  COUNT(*)  AS
total_count   FROM   disaster_tweets   LATERAL
VIEW EXPLODE(entities.hashtags) t1 AS hashtags
GROUP BY LOWER(hashtags.text) ORDER BY
total_count LIMIT 10;
```

Fig 10: Common Hastags



Results are shown below

Sikkim floods	14,239
Rajasthan floods	16,324
Uttarakhand floods	17,165
Kashmir Floods	18,732
Cyclone Vardah	20,466
Uttar Pradesh floods	27,300
Bihar floods	28,378
Madhya Pradesh floods	31,587
Chennai Floods	34,192
Cyclone Hudhud	35,697

Fig 11: The popularity of tweets on different topics is shown in an increasing order

VI. Conclusion:

Twitter tool divulge different opinions on a number of issues and topics. It provides keen insight the topic. It may be a good area for analysis for taking decisions in different areas. For performing twitter post analysis HADOOP is an important option. Performing analysis on diverse topics by changing the keywords in query after loading of Flume and Hive. The analysis helps in finding people's response to the natural disaster. For that, it helps in strategy planning finding the polarity of tweets collected.

REFERENCES

1. [1] Sunil B. Mane , Sunil B. Mane, Yashwant Sawant, Saif Kazi, Vaibhav Shinde , “Real Time Sentiment Analysis of Twitter Data Using Hadoop”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3098 – 3100 , ISSN:0975-9646.
2. [2] Mahalakshmi R, Suseela S , “Big-SoSA:Social Sentiment Analysis and Data Visualization on Big Data”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 4, April 2015 , pp 304-306, ISSN : 2278-1021.
3. [3] Matthew Koehler, Spencer Greenhalgh, Andrea Zellner, Michigan State University, United States , “Potential Applications of Sentiment Analysis in Educational Research and Practice – Is SITE the Friendliest Conference?”, Mar 02, 2015 in Las Vegas, NV, United States ISBN 978-1-939797-13-1 Publisher: Association for the Advancement of Computing in Education (AACE).
4. [4] Ramesh R, Divya G, Divya D, Merin K Kurian , “Big Data Sentiment Analysis using Hadoop “, (IJIRST)International Journal for Innovative Research in Science & Technology,Volume 1 , Issue 11 , April 2015 ISSN : 2349-6010.
5. [5] Peiman Barnaghi, Parsa Ghaffari, John G. Breslin , “Text Analysis and Sentiment Polarity on FIFA World Cup 2014 Tweets” , Conference ACM SIGKDD’15, August 10-13, 2015, Sydney, Australia. Copyright 2015 ACM 1-58113-000-0/08/2015.
6. [6] “Mining Data from Twitter” from AbhishangaUpadhyay, Luis Mao, Malavika Goda Krishna (PDF)
7. [7] G.Vinodhini , RM.Chandrasekaran, “Sentiment Analysis and Opinion Mining: A Survey” , International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012 ISSN: 2277 128X.
8. [8] Michael G. Noll, Applied Research, Big Data, Distributed Systems, Open Source, "Running Hadoop on Ubuntu Linux (Single-Node Cluster)", [online], available at http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/
9. [9] "Install Apache Hadoop 2.6.0 in Ubuntu (Multi node/Cluster setup)", [online], available at http://pingax.com/install-apache-hadoop-ubuntu-cluster-setup/
10. [10] Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce",6-8 Dec. 2012.
11. [11] https://blog.cloudera.com/blog/ 2012/11/analyz ing-twitter-data-with-hadoop-part-3-querying-semi-structured-data-with-hive/
12. [12] "Application Programming Interface."Wikipedia . Wikimedia Foundation, 23 Oct. 2014. Web. 24 Oct. 2014.
13. [13] "Twitter's API --HowStuffWorks.˝HowStuffWorks. N.p., n.d. Web. 24 Oct. 2014.
14. [14] Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More-Matthew A. Russell
15. [15] "The Streaming APIs."Twitter Developers. N.p., n.d. Web. 23 Oct. 2014.
16. [16] Sangeeta, Twitter Data Analysis Using FLUME & HIVE on Hadoop FrameWorkSpecial Issue on International Journal of Recent Advances in Engineering & Technology (IJRAET) V-4 I-2 For National Conference on Recent Innovations in Science, Technology & Management (NCRISTM), 26th to 27th February 2016.
17. [17] A. Bifet and E. Frank, "Sentiment knowledge discovery in twitter streaming data," in Discovery Science, 2010, pp. 1-15.
18. [18] T. Blog, "Insights into the #WorldCup conversation on Twitter," in Twitter Blog, ed, 2014.
19. [19] S. Sinha, C. Dyer, K. Gimpel, and N. A. Smith, "Predicting the NFL using Twitter,"
20. arXiv preprint arXiv:1310.6998, 2013.
21. [20] P. Priyanthan, B. Gokulakrishnan, T. Ragavan, N. Prasath, and A. S. Perera, "Opinion mining and sentiment analysis on a twitter data stream," ICTer 2012, 2012.
22. [21] D. Terrana, A. Augello, and G. Pilato, "Automatic Unsupervised Polarity Detection on a Twitter Data Stream," in Semantic Computing (ICSC), 2014 IEEE International Conference on , 2014, pp. 128 -134.
23. [22] L. Zhang, "Sentiment analysis on Twitter with stock price and significant keyword correlation," 2013.
24. [23] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," ACM SIGKDD explorations newsletter, vol. 11, pp. 10 -18, 2009.