ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Trend Analysis through Hashtags Popularity Level Using Hadoop with Hive

Deepak Ranjan, Dr. Tripti Arjariya, Dr. Mohit Gangwar

Department of Computer Science & Engineering, Bhabha Engineering Research Institute, Bhopal, India

Prof. & Head, Department of Computer Science & Engineering, Bhabha Engineering Research Institute, Bhopal, India

Principal, Department of Computer Science & Engineering, Bhabha Engineering Research Institute, Bhopal, India

ABSTRACT*:* Social media generates massive amounts of data every minute, which is caused by its mainstream adoption over the past years. Innovations in the industry have enabled new ways of communications between people and created many business opportunities. Big Data in social media require effective and advanced processing technologies. Purpose of data mining analyses is to find valuable patterns and insights from Twitter data. Thus, analysis for twitter data is meaningful for both individuals and organizations to make decisions. Due to the huge amount of data generated by twitter every day, a system which can store and process big data is becoming a problem. In this paper, we present a method to collect twitter data sets, and store and analyze the data sets on Hadoop platform. Here we can fetch real time twitter data and store it into the HDFS and then we develop an trend analysis mechanism which is hive web interface to analyze the twitter hashtag keywords along with its frequency through which we can get the most trendy keywords right now on the twitter through hashtag popularity level [11][13].

KEYWORDS*:* Hadoop, data mining, data analysis, social datas, flume, hive.

I. INTRODUCTION

Over the past years, the amounts of data generated from the Internet services have increased significantly. Innovations in the field of ICT have enabled new business opportunities for creating services capable of handling vast data volumes. Technology reached the level, where people are interconnected with social media on daily basis and are able to share their life over social networks. Social networking over Internet has become popular in the last years, which is also justified with the increased data volumes. New challenges appeared in relation to data storage architectures with scalability features and effective processing algorithms.

Data mining analysis has a great potential in finding meaningful insights within social networks data. Twitter social network is a service developed in order to enable communication between people by sending short messages. According to research [1], Twitter as the second largest social media platform, right behind Facebook, generates around 350, 000 tweets each minute, or 21 million per hour. Such volumes of data present challenges for engineers to develop innovative solution for effective data architecture and processing capabilities in order to apply data mining. The importance of Big Data implementations within enterprises in various sectors, such as health industry, retail, telecom or social networks plays crucial scenario in optimizing business processes and creating new value propositions for revenue streams.

According to Accenture research [1], 87% of enterprises believe that Big Data will reshape the industry in the next years. Therefore, the early adoption of this trend may create additional value to enterprises in sense of data mining of customer's opinions about company in the Twitter use case. Knowing what customers think about the services or how services can be innovated in the future gives companies insights and strategies for development. Furthermore, survey also found that 89% of respondents said that companies who do not adapt to this Big Data trend would risk losing market share.

Start by searching Twitter. Check replies to your association's tweets and the accounts of other leading industry organizations to learn what hashtags [13] they are using. You can also search common industry terms. Compare hashtags by using Topsy (topsy.com/analytics). This site keeps you up to date with what's trending and what people are following [11].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Gathering Data with Apache Flume

To automate the movement of tweets from the API to HDFS, without our manual intervention, Flume is used. Apache Flume is a reliable and distributed system for effectively gathering and moving large amounts of data from various sources to a common storage area. Major components of flume are source, memory channel and the sink.

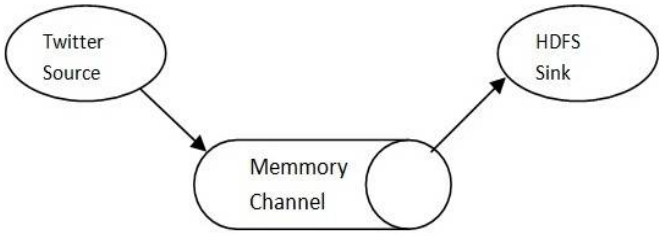

Figure1: Flume Components

Twitter source is an event-driven source that uses Twitter4j library for accessing streaming API. Tweets are collected and aggregated into fundamental units of data called as an event. An event incorporates a byte payload and an optional header. The coordination of event-flows from the streaming API to HDFS is undertaken by Agent. The acquired tweets are stored into one or more memory channels. A memory channel is a temporary storage that uses an in-memory queue to retain events until they are ingested by the sink. Using memory channel, tweets are processed in batches that can be configured to hold a constant number of tweets. To procure tweets for a given keyword filter query is used. Sink writes events to a pre configured location. This system makes use of the HDFS-sink that deposits tweets into HDFS.

HADOOP

The Apache Hadoop [9] project develops open-source software for scalable, reliable, distributed computing. The Apache Hadoop library is a framework that allows for the distributed processing of large data sets beyond clusters of computers using thousands of computational independent computers and large amount (terabytes, petabytes) of data. Hadoop was derived from Google File System (GFS) and Google's Map Reduce. Apache Hadoop is good choice for twitter analysis as it works for distributed huge data. Apache Hadoop is an open source framework for distributed storage and large scale distributed processing of data-sets on clusters. Hadoop runs applications using the MapReduce algorithm, where the data is processed in parallel on different clusters nodes. In short, Hadoop framework is able enough to develop applications able of running on clusters of computers and they could perform complete statistical analysis for a huge amounts of data. Hadoop MapReduce is a software framework [8] for easily writing applications which process big amounts of data in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

Hive

After congregating the tweets into HDFS they are analyzed by queries using Hive. Apache Hive data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. In opinion mining system, hive is used to query out interested part of the tweets which can be an opinion, comments related to a specific topic or a trending hash tag. Twitter API loads the HDFS with tweets which are represented as JSON blobs. Processing twitter data in relational database such as SQL requires significant transformations due to nested data structures. Hive facilitates an interface that provides easy access to tweets using HiveQL that supports nested data structures. Hive compiler converts the HiveQL queries into map reduce jobs. Partition feature in hive allows tweet tables to split into different directories. By constructing queries that includes partitions, hive can determine the partition comprising the result. The location of twitter tables are explicitly specified in “Hive External Table” which are partitioned. Hive uses SerDe (Serializer- Deserializer) interface in determining record processing steps. Deserializer interface intakes string of

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

tweets and translates it into a Java Object that Hive can manipulate on. The Serializer interface intakes a java Object that Hive has worked on and converts it into required data to be written on HDFS.

II. LITERATURE REVIEW

In [1] This paper author investigates the problem of predicting Twitter hashtags popularity level. A data set of more than 18 million tweets containing 748 thousand hashtags has been prepared by using Twitter's rest API. Early adoption properties including profile of tweet authors and adoption time series are used to predict a tag's later popularity level. The followers count and tweets count are two such characteristics related to adopters' profile. On the other hand, two types of frequency domain analyses are used to augment the simple mean and standard deviation characteristics of the adoption time series. Fourier transform (FT) spectrum and wavelet transform (WT) spectrum are considered in this study. Experimental results show that WT spectrum improves the prediction result of viral hashtags while FT spectrum does not.

Online social networks provide communication channels to spread an idea, behavior, style or usage throughout the world village. Twitter is a special online service that provides both social network and microblog functions. Posting tweets through devices from desktop to mobile is the main activity of the microblog function, while following and retweeting offer the social network function. Users post tweets by encoding topics in the form of hashtags, which are summarized by Twitter to make a list of current trending tags.

In [2], the author describes that Big data analytics has attracted intense interest from all academia and industry recently for its attempt to extract knowledge, information and wisdom form big data. Big data and cloud computing, two of the most important trends that are defining the new emerging analytical tools. Big data analytical capabilities using cloud delivery models could ease adoption for many industry, and most important thinking to cost saving, it could simplify useful insights that could providing them with different kinds of competitive advantage. Many companies to provide online Big Data analytical tools some of the top most companies like Amazon Big data Analytics Platform ,HIVE web based Interface, SAP Big data Analytics, IBM InfoSphere BigInsights, TERADATA Big Data Analytics, 1010data Big Data Platform, Cloudera Big Data Solution etc. Those companies analyze huge amount of data with help of different type of tools and also provide easy or simple user interface for analyzing data.

III. OBSERVATION

Hadoop and bigdata analytical tools, for getting raw data from the Social Network, we may use Hadoop online streaming tool such as Apache Flume, apache kafka. By utilizing this tool only, we are going to configure everything, which we wanted to get (data) from the Social Network. Mainly we want to set the configuration model and also want to define what information that we want to collect form Social Network. All these will be stored into our HDFS (Hadoop Distributed File System) in our own prescribed format. From this unrefined data we are going to refined the data using analytical tools and than start analysing these social data to predict or to help in decision making.

IV. PROBLEM DEFINITION

The project focuses on using Twitter, the most popular micro blogging platform, for the task of sentiment analysis. The tweets are important for analysis because data arrive at a high frequency and algorithms that process them must do so under very strict constraints of storage and time. Because these tweets generates the huge information related to different area like government, election, etc. millions of tweets is generated every day and which is very useful in decision making because every one is share their view and opinions on issues or variety of topics. The analysis of twitter data gives real view on trends or different user opinions regarding what they think and to analysis these data provide a better way for making any decision. But the twitter site generates petabyte of data per day which is difficult to handle with traditional tools and technique for this we need a new powerful technique to handle bigdata[10].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

V. PROPOSED WORK

For analysing these large and complex data required a power tool, we are using hadoop[6] which is a open source implementation of mapreduce, a powerful tool designed for deep analysis and transformation of very large data.

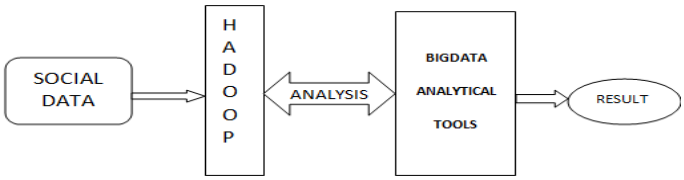

Figure 2: Workflow Diagram

This paper we design algorithm for handling the problems raised by the larger data volume and the dynamic data characteristics for finding and performing operation on social media data sets. For analysing first we used standard platform as hadoop on single node ubuntu machine 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 ,After that we can use bigdata analytical tools to refine such unstructured data and analyse the social data using bigdata analytical tools.

VI. PROPOSED METHODOLOGY

Our Steps or Algorithm Steps will follow:

1. First we can create a social account and than we can use social API's[12] for fetching real time social data and store it into HDFS.
2. For fetching social data we can use bigdata tools such as apache flume through which we can authenticate our keys and start fetching fetching data from social sites.
3. After fetching data, the data is store into HDFS (Hadoop Distributed File System), which is very reliable for storing such huge amount of data.
4. After storing data into HDFS, we can pre-process the data because from the social sites an unstructured raw data is coming, which is very difficult to analyze such kind of unstructured data, so we can first pre-process the data and convert it into some structure form.
5. After pre-processing we can start analyzing such huge amount of social data.

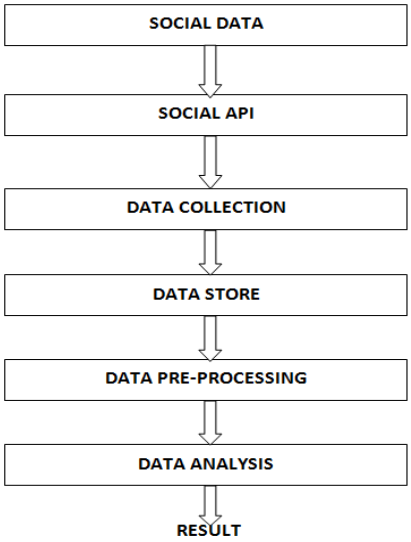

Figure 3: Analysis Steps

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

VII. EXPERIMENTAL & RESULT ANALYSIS

All the experiments were performed using an i5-2410M CPU @ 2.30 GHz processor and 4 GB of RAM running ubuntu 14 [10]. As we have seen the procedure how to overcome the problem that we are facing in the existing problem that is shown clearly in the proposed system. So, to achieve this we are going to follow the following methods:

- Creating Twitter Application.
- Getting data using Flume.
- Querying using Hive Query Language (HQL)

Creating Twitter Application

First of all if we want to do analysis on Twitter data we want to get Twitter data first so to get it we want to create an account in Twitter developer and create an application by clicking on the new application button provided by them shown in fig. 4 After creating a new application just create the access tokens so that we no need to provide our authentication details there and also after creating application it will be having one consumer keys to access that application for getting Twitter data. The following is the figure that show clearly how the application data looks after creating the application and here it's self we can see the consumer details and also the access token details. We want to take this keys and token details and want to set in the Flume configuration file such that we can get the required data from the Twitter in the form of tweets.

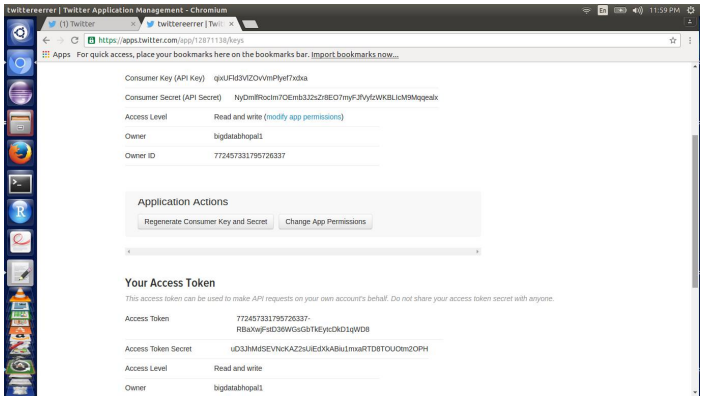

Figure 4: Creating twitter app and Generation access token keys

The figure 4 show clearly the application keys that are generated after creating application and in this keys we can see the top two keys are the API key and API secret. And coming to the reaming two keys it is nothing but know as the access tokens that we want to generate it by ourselves by clicking the generate access token. After clicking that we can get the two keys that are our account access token and coming to that one is Access token and the other one is the Access token secret.

Getting data using Flume

After creating an application in the Twitter developer site we want to use the consumer key and secret along with the access token and secret values. By which we can access the Twitter and we can get the information that what we want exactly here we will get everything in JSON format and this is stored in the HDFS that we have given the location where to save all the data that comes from the Twitter. The following is the configuration file that we want to use to get the Twitter data from the Twitter. All the details we have to fill in the flume-twitter. conf file shown in the figure.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS
Â
TwitterAgent.sources.Twitter.type =
com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = FlRx3d0n8duIQ0UvGeGtTA
TwitterAgent.sources.Twitter.consumerSecret =
DS7TTbxhmQ7oCUlDntpQQRqQllFFOiyNoOMEDD0lA
TwitterAgent.sources.Twitter.accessToken = 1643982224-
xTfNpLrARoWKxRh9KtFqc7aoB8KAAHkCcfC5vDk
TwitterAgent.sources.Twitter.accessTokenSecret =
PqkbuBqF3AVskgx1OKgXKOZzV7EMWRmRG0p8hvLQYKs
Â
TwitterAgent.sources.Twitter.keywords = hadoop, big data,
analytics, bigdata, cloudera, data science, data scientiest,
business intelligence, mapreduce, data warehouse, data
warehousing, mahout, hbase, nosql, newsql, businessintelligence,
cloudcomputing
Â
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path =
hdfs://localhost:9000/user/flume/tweets/
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000
â
```

Querying using Hive Query Language

(HQL)After running the Flume by setting the above configuration then the Twitter data will automatically will save into HDFS where we have the set the path storage to save the Twitter data that was taken by using Flume. The following is the figure that shows clearly how the data is stored in the HDFS in a documented format and the raw data that we got form the Twitter is also in the JSON format .

From these data first we want to create a table named as raw where the filtered data want to set into a formatted structured such that by which we can say clearly that we have converted the unstructured data into structured format. For this we want to use some custom serde concepts. These concepts are nothing but how we are going to read the data that is in the form of JSON format for that we are using the custom serde for JSON so that our hive can read the JSONdata and can create a table in our prescribed format the data are shown below.

```
WARNING: org.apache.hadoop.metrics.jvm.EventCounter is deprecated. Please use org.apache.hadoop.log.metrics.EventCounter in all t
he log4j.properties files.
Logging initialized using configuration in jar:file:/home/abhi/work/hive-0.10.0/lib/hive-common-0.10.0.jar!/hive-log4j.properties
Hive history file=/tmp/abhi/hive_job_log_abhi_201704260036_1286465063.txt
hive> create database beri;
OK
Time taken: 3.474 seconds
hive> use beri;
OK
Time taken: 0.056 seconds
hive> add jar file:///home/abhi/work/hive-0.10.0/lib/hive-serdes-1.0-SNAPSHOT.jar;
Added file:///home/abhi/work/hive-0.10.0/lib/hive-serdes-1.0-SNAPSHOT.jar to class path
Added resource: file:///home/abhi/work/hive-0.10.0/lib/hive-serdes-1.0-SNAPSHOT.jar
hive> CREATE EXTERNAL TABLE tweets (id BIGINT,entities STRUCT<hashtags:ARRAY<STRUCT<text:STRING>>>) ROW FORMAT SERDE 'com.clouder
a.hive.serde.JSONSerDe' LOCATION '/home/beri/project';
OK
Time taken: 1.386 seconds
hive>
```

Before we can query the data, we need to ensure that the Hive table can properly interpret the JSON data. By default, Hive expects that input files use a delimited row format, but our Twitter data is in a JSON format, which will not work with the defaults. And we can use the Hive SerDe interface to specify how to interpret what we've loaded. SerDe stands for Serializer and Deserializer, which are interfaces that tell Hive how it should translate the data into something that Hive can process. For these we added a jar file by command

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

ADD JAR <path-to-hive-serdes-jar>;

Using these jar file and custom serde we can store the unstructured data into hive table tweets which has been created above in a structure manner. The figure 5 shows the structure data stored in table tweets.

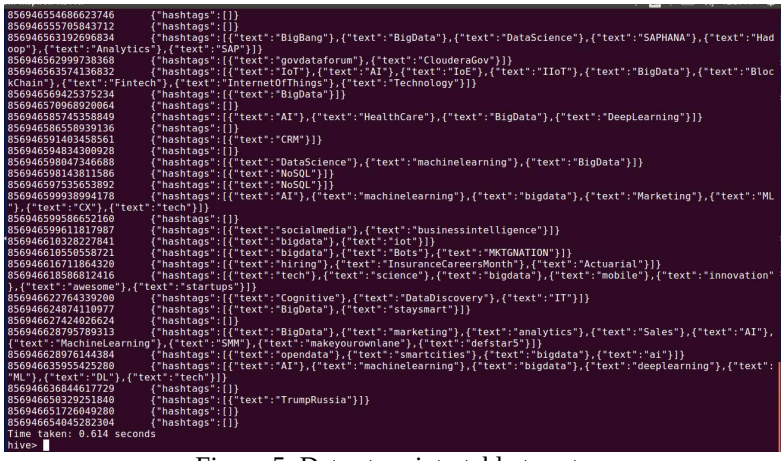

Figure 5: Data store into table tweets

After that from the table tweets we can get the tweets id and the array of hashtag keywords with some unwanted keywords like text , hashtags so we want to prun that keywords by preprocessing the tweets table and collect only the hashtag keywords. After preprocessing the tweets table we can get the hashtag_word table which is shown in figure 6.

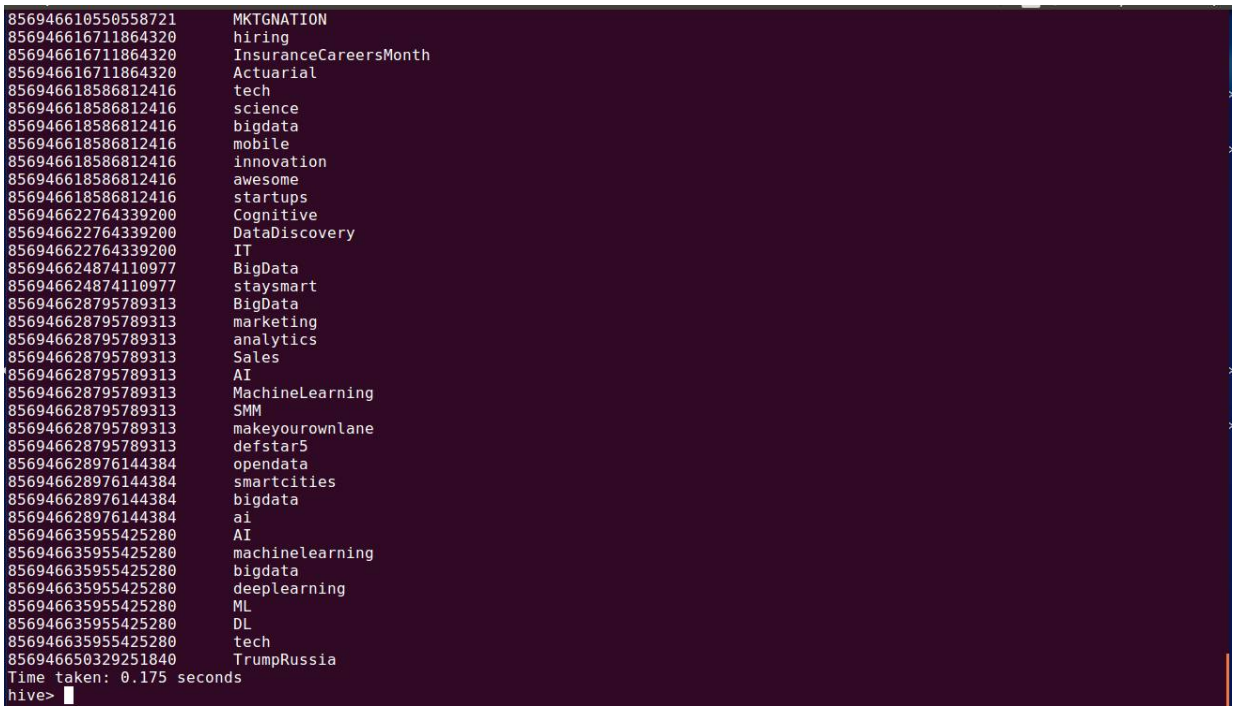

Figure 6: Data store into table hashtag_word table

After getting hashtag_word table we can find the frequency of the hashtag keywords and store the result in to result_top table in the the maximum frequency hashtag keywords are present which is shown in figure 7.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

```
hive> select * from top_result;
OK
BigData 29
AI      16
bigdata 14
analytics       14
Analytics       10
machinelearning 9
Marketing       7
DataScience     7
'tech    7
ML      6
IoT     6
IoE     5
IIoT    5
BlockChain      4
TrumpRussia     4
InternetOfThings        4
fintech 3
NoSQL   3
Fintech 3
Technology      3
Hadoop  2
ai      2
Cognitive       2
Twitter 2
BigBang 2
Tweets  2
socialmedia     2
```

Figure 7. Data store into result_top table

After getting resultant hashtags keywords with their frequency we can easily find the maximum popular keywords and minimum popular keywords whish are shown in table 1 and 2 and their graphical representation are shown in figure 8 and 9.

Table 1: Maximum frequency keywords

HASHTAG KEYWORDS	FREQUNCY OF KEYWORDS
BigData	29
AI	16
Analytics	14
Machinelearning	9
Marketing	7
DataScience	7
Tech	7
IoT	6
IoE	5
IIoT	5

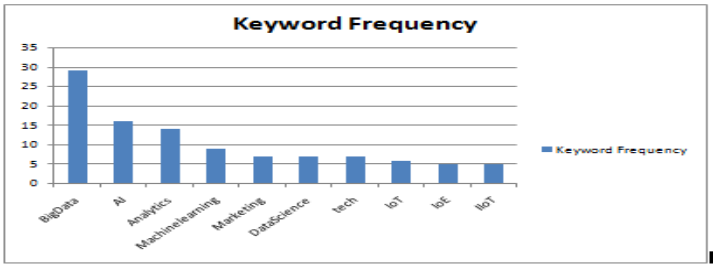

Figure 8: Maximum frequency keywords

Table 2: Minimum frequency keywords

HASHTAG KEYWORDS	KEYWORD FREQUENCY
NoSQL	3
Tweets	2
SocialMedia	2
AWS	1
Analysis	1
BI	1
CRM	1
DataFuture	1
DataScientists	1
DataViz	1

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

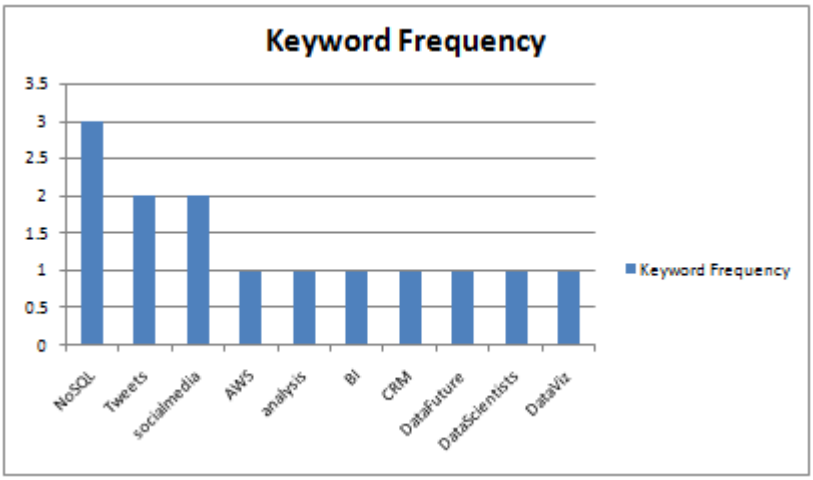

Figure 9: Minimum frequency keywords

VIII. CONCLUSION

In this study, we store original dataset in HDFS file system and analyze the datasets using Hadoop Mapreduce model. This study store analyzes the datasets by using HDFS, Hadoop and Hive respectively, which are more scalable and efficient than traditional RDBMS.The experiment results prove that the present method performs efficiently.
In this, we can fetch real time twitter data and store it into the HDFS and then we develop an trend analysis mechanism which is hive web interface to analyze the twitter hashtag keywords along with its frequency through which we can get the most trendy keywords right now on the twitter through hashtag popularity level.

REFERENCES

[1] Shing H. Doong, “Predicting Twitter Hashtags Popularity Level”, in 2016 49th Hawaii International Conference on System Sciences, IEEE, DOI 10.1109/HICSS.2016.247.
[2] Rahul Kumar Chawda, Dr. Ghanshyam Thakur, “Big Data and Advanced Analytics Tools”, 2016 Symposium on Colossal Data Analysis and Networking (CDAN), IEEE 2016, ISSN: 978-1-5090-0669-4/16.
[3] Skuza, Michal, and Andrzej Romanowski. "Sentiment analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction" In Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on, pp. 1349-1354. IEEE, 2015.
[4] Judith Sherin Tilsha S , Shobha M S, “A Survey on Twitter Data Analysis Techniques to Extract Public Opinion”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 11, November 2015, pp 536-540.
[5] 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.
[6] Praveen Kumar, Dr Vijay Singh Rathore,” Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 6, June 2014, pp 7123-7126.
[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] Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce",6-8 Dec. 2012.
[9] 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/.
[10] Sagiroglu, S., & Sinanc, D, “Big data: A review”, IEEE International Conference on Collaboration Technologies and Systems (CTS), 2013, pp 42-47.
[11] G. Szabo, and B.A. Huberman, “Predicting the Popularity of Online Content”, Communication of the ACM, 2010, 53(8), pp. 80-88.
[12] "Application Programming Interface." *Wikipedia*. Wikimedia Foundation, 23 Oct. 2014. Web. 24 Oct. 2014.
[13] O. Tsur, and A. Rappoport, “What’s in a Hashtag? Content Based Prediction of the Spread of Ideas in Microblogging Communities”. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining, 2012, pp. 643-652.