ANALYSIS OF TWEETS USING HADOOP

[1]Amer Hussain, [2]Dr.Shirisha,

[1]Post Graduate Research Scholar, [2]Professor,

[1,2] Department of Computer Science,

[1,2]Sreenidhi Institute of Science & Technology, Hyderabad, India.

Abstract : Ever since the enhancement of Technology, data has been growing every day. One prominent factor that gave rise to Big data is Social media. Data from these social media sites can be used for many purposes like prediction, analysis, marketing, business etc. Among these social media sites is Twitter, which is prominent and widely used. Traditional systems and tools like R-DBMS etc. do not have the capacity to handle enormous data which is mainly un structured. Hence there is a need to handle Big data and this is solved by Hadoop, a java-based framework for distributed storage & processing of large datasets on commodity hardware. In this paper we propose a method to perform analysis of tweets using Big data tools such as Hadoop, flume, hive Hue etc**.**

Keywords - Bigdata, Social media, Hadoop, Flume, Hive, Hue, twitter, hashtag, tableau**.**

INTRODUCTION

A key factor in the enormous rise in big data is attributed to social media sites. Among them, a widely used social media micro-blogging site is Twitter. Millions of tweets are generated every day. Tweets are short posts (140 char) that contain emoticons, hashtags, slangs etc. These tweets mainly relate to current issues, user-opinions etc. Users tweet on a specific issue with a Hashtag. It's a meta data. User can find all the posts that have been tagged using the hashtag.

ABOUT HADOOP

Hadoop is a java-based programming framework for distributed storage & processing of large datasets on commodity hardware. It is developed by Apache software foundation as open source framework.
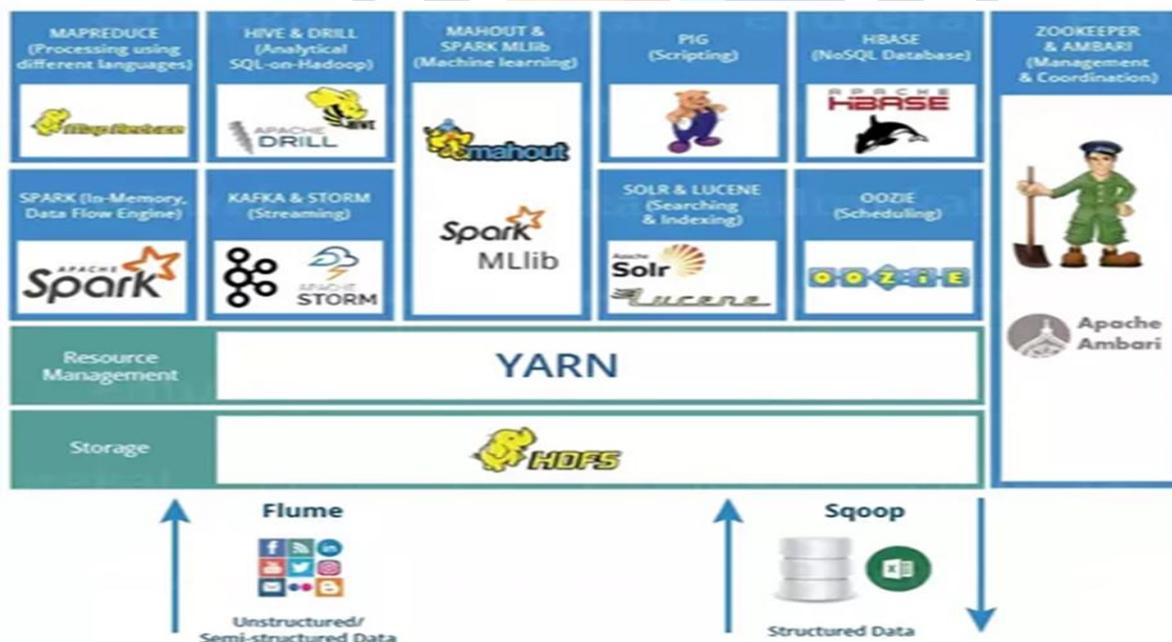

Figure 1 HADOOP ECOSYSTEM

Hadoop mainly consists of 2 components. They are 1) HDFS (Hadoop Distributed File system) for distributed storage 2) MapReduce/Yarn for parallel processing.

HDFS has a master-slave architecture. HDFS mainly contains 1) Name node and 2) Data Node. Name node acts as a master-node. It manages the file system and stores the metadata. Actual data is stored in data node. A file is split into different blocks. These blocks are stored in data nodes.
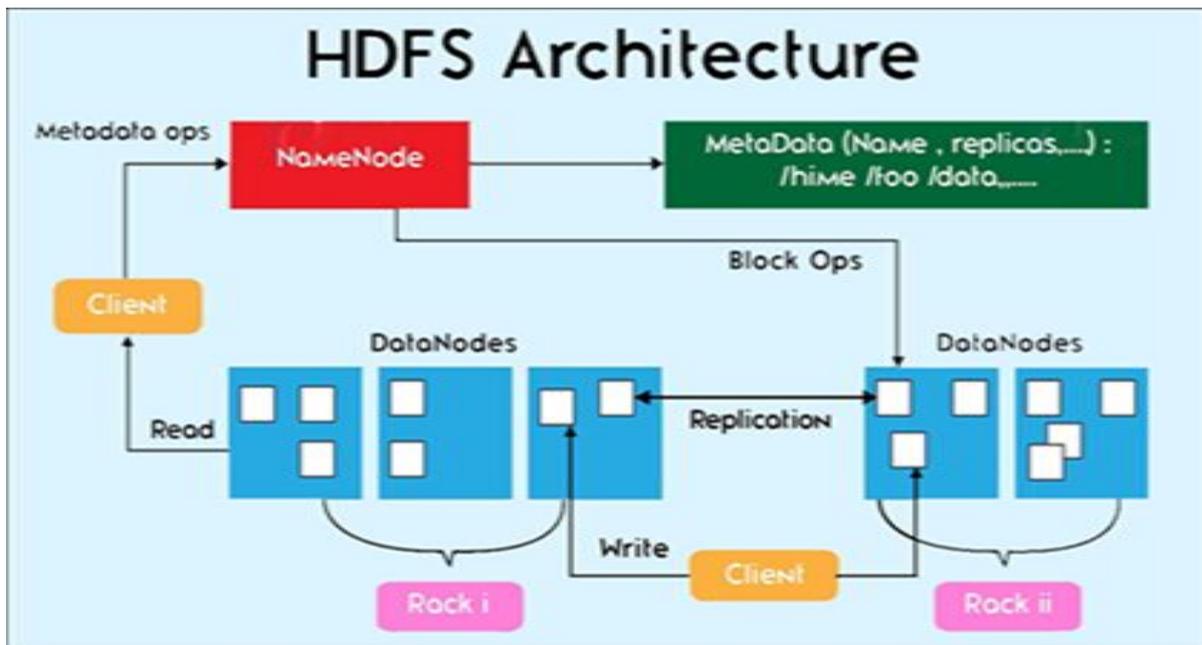
Figure 2 **HDFS ARCHITECTURE**

MapReduce /Yarn is a programming model of Hadoop which is used for parallel and distributed processing on large datasets. MapReduce mainly consists of two components. They are 1) Job Tracker and 2) Task Tracker. A Job tracker gives tasks to specific nodes in a cluster whereas a Task Tracker accepts tasks like map, reduce, shuffle etc. from job tracker.
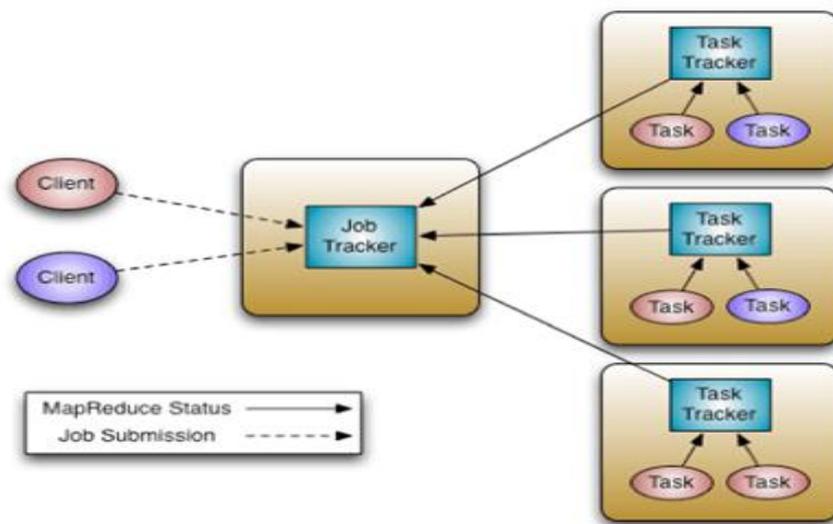

Fig3 **MAPREDUCE ARCHITECTURE**
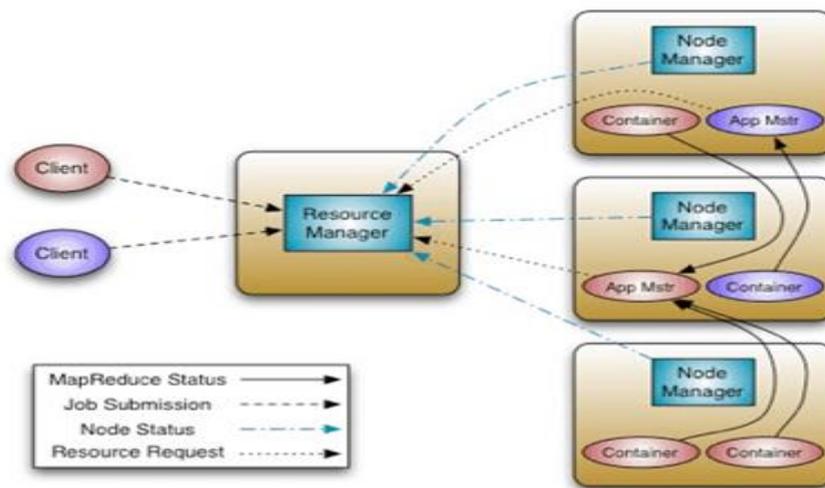
Figure 4 **YARN ARCHITECTURE**

RELATED WORK

As Bigdata and Data analytics are trending now-a-days, a lot of work have been carried in this field
.E.Sivaraman[1] in his paper focused on the importance of Hadoop in solving bigdata use-cases. He explained about the High-performance, Fault-tolerance, Distributed storage and parallel processing of Hadoop. This paper also illustrates about Gartner Hype cycle for emerging technology that shows how bigdata is at its peak of trending technology and will remain for a couple of years.

Can Uzunkaya[2] proposed a method of analyzing tweets using twitter 4J, an un-official java library for twitter API.

Monu Kumar[3] in his paper presented an approach to analyze people's opinions regarding Airtel services. To collect tweets a code was developed in python.

Matthew Edwards[4] in his paper designed a Hadoop-based framework to analyze Synchro phasor data generated in peta-bytes. He explained how Hadoop was capable of computing algorithms on synchro phasor data sets quickly and efficiently.

Anis Zarrad[5] proposed an approach to evaluate public opinion regarding MERS-CoV (Middle East Respiratory Syndrome Corona Virus. He used Apache Hadoop as a big data platform with an opinion mining approach to evaluate the opinions of people.

Eugen Feller [6] in his paper investigated the performance and power implications such as execution time, energy consumption, of running Hadoop in various deployment models like VM's, Physical clusters and in cloud environment.

PROPOSED WORK

 This section presents the proposed approach used for analyzing twitter data using Hadoop. Hadoop can be deployed in a physical cluster, or in a virtual machine or on cloud. For huge data, a physical cluster with 3 or more nodes depending on the size of data is efficient. We have deployed Hadoop in a virtual machine. Instead of installing individual softwares that involve compatibility issues we have used CDH (Cloudera Hadoop Distribution) that comes with latest inbuilt software required for big-data.
CDH is fast and efficient when compared to other traditional Hadoop services.

Cloudera is one of the top vendors of Hadoop. It's products and solutions enable us to deploy and manage Hadoop and related projects. We can manipulate and analyze our data and can keep our data safe, protected and secure.
CDH (Cloudera Hadoop Distribution) is world's most popular Hadoop distribution. It provides security and integration with numerous hardware and software solutions. It consists of many big data tools that include flume, hive, hue, impala, HBase, etc.CDH delivers the core elements of Hadoop i.e. Scalable storage, distributed computing, web-based user interface.

Download Oracle VM VirtualBox and then download cloudera QuickStart Vm (CDH5.13) from cloudera website.
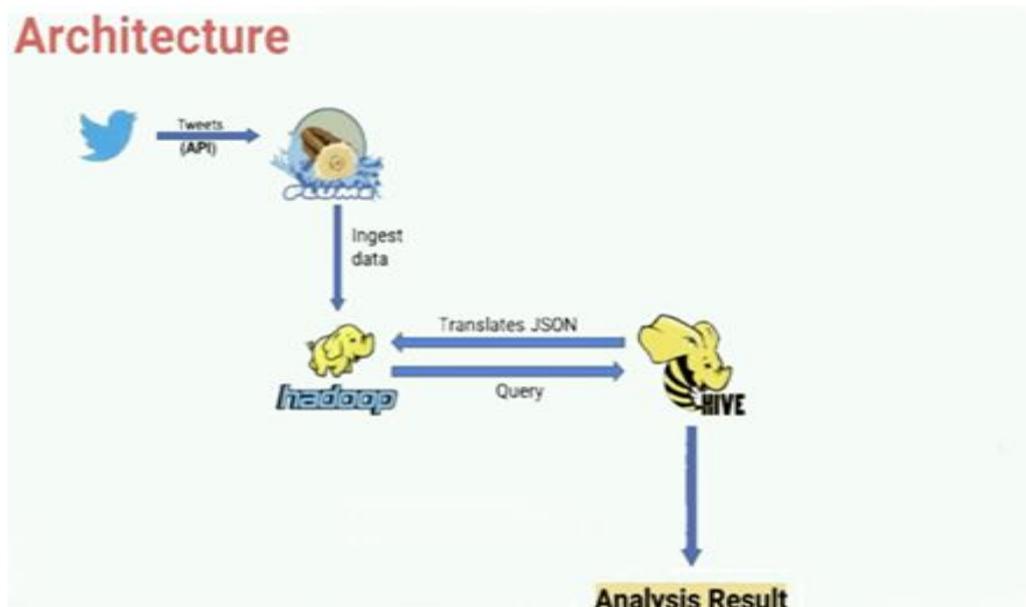
FIG5 PROPOSED ARCHITECTURE OF TWITTER ANALYSIS

Our proposed architecture consists of 3 phases. They are 1) Obtaining twitter-keys.2) Data Ingestion 3) Data Analysis and 4) Data Visualization.

Obtaining Twitter Keys: Twitter gives some access for its development purpose. Users/enterprises can sign into developer account and can make use of some extra features which a normal user does not have access. A user has to request developer account on app.twitter.com and create an app to get twitter API keys and access tokens.
After user authentication following keys will be generated.1) consumer key 2) consumer secret key 3) Access token and 4) Access token secret key.

Data Ingestion: Flume is a real-time big data ingestion tool. It collects and transports large amount of streaming data from various web servers to HDFS. It is also fault-tolerant.
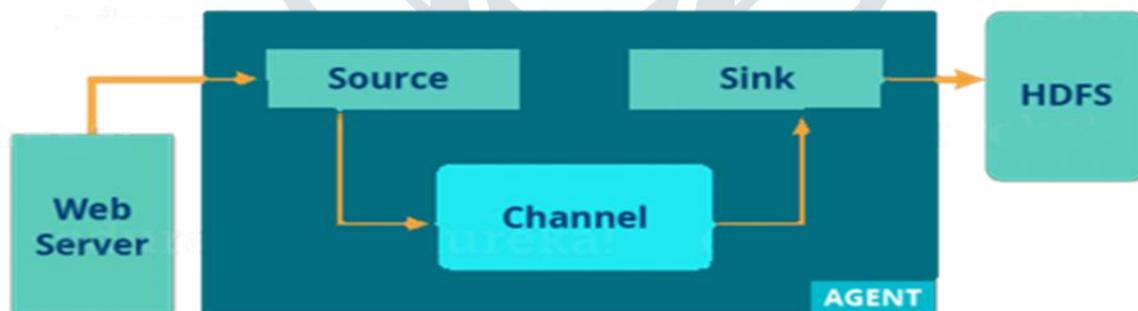
FIGURE 6 FLUME ARCHITECTURE

A flume agent is a (JVM) process that hosts the components through which events flow from an external source to next destination. Flume agent ingests the data from various data sources to HDFS.
 A flume agent has 3 components. They are 1) source 2) channel and 3) sink.
Source accepts the data from the incoming streamline and stores the data in channel. Generally reading speed is faster than writing speed. To match the write and read speed difference there is a buffer, that acts as intermediary storage. It stores the data that is being transferred temporarily thus preventing data loss. Similar to this a channel acts as temporary storage between data source and Hdfs. Sink collects the data from channel and writes it into HDFS permanently.
Open Flume folder using terminal. Create a flume-conf. properties file. Configure this file with the keys that were generated. Create a directory in Hdfs and give its path in flume configuration file. Execute flume with this command.

$flume-ng agent –conf   /usr/lib/flume-ng/conf/-f/usr/lib/flume-ng/conf/flume2conf.properties -D flume. root.logger = DEBUG,console -n Twitter Agent

FLUME-CONFIGURATION FILE

```
TwitterAgent.sources=Twitter
TwitterAgent.channels=MemChannel
TwitterAgent.sinks=HDFS
TwitterAgent.sources.Twitter.type=com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels=MemChannel
TwitterAgent.sources.Twitter.consumerKey=bp3fY697zzelz6rTeHYzNWzOO
TwitterAgent.sources.Twitter.consumerSecret=nIHTO9qrFfRVqjBPlzPGaR06Peprdn6z6Wi
6cxNKtlfGE24FIq
TwitterAgent.sources.Twitter.accessToken=525628994-
OCASvqUb9KVXhF0Rit6w4mvpucYhojE1h0IUcVmc
TwitterAgent.sources.Twitter.accessTokenSecret=vgxEukNA4KmVrhmlRcsvgfOA2j1jdzIN
3Vy0k7dr1YKu2
TwitterAgent.sources.Twitter.keywords=pulwama
TwitterAgent.sinks.HDFS.channel=MemChannel
TwitterAgent.sinks.HDFS.type=hdfs
TwitterAgent.sinks.HDFS.hdfs.path=hdfs://localhost:8020/user/root/flume2
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
TwitterAgent.channels.MemChannel.type=memory
TwitterAgent.channels.MemChannel.capacity=10000
TwitterAgent.channels.MemChannel.transactionCapacity=100
```

Data related to the hashtags that were given as keywords (input) in the flume is generated. The twitter data generated in hdfs is in JSON (Java Script Object Notation) format.

Data Analysis: Hive is a big data tool in Hadoop ecosystem similar to SQL.Hive has the ability to analyze row format data, but data generated in Hdfs is in json format. To overcome this, we use Hive-serde. Serde stands for serializer and deserializer. It converts json data to Querable format. Download Hive-serde-snapshot.jar. We can also create our own serde. Add this jar file to lib folder in flume directory. Open Hive in terminal. Create an external table as follows.

```
CREATE EXTERNAL TABLE tweets_pulwama(
id BIGINT,
created_at STRING,
source STRING,
favorited BOOLEAN,
retweet_count INT,
retweeted_status STRUCT<
text:STRING,
user:STRUCT<screen_name:STRING,name:STRING>>,
entities STRUCT<
urls:ARRAY<STRUCT<expanded_url:STRING>>,
user_mentions:ARRAY<STRUCT<screen_name:STRING,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_nameSTRING
)
```

ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'
LOCATION '/user/root/flume2'

Open Hue editor in cloudera quickstart vm. Hue (Hadoop User Experience) is an open source web interface that supports Apache Hadoop and its ecosystem for browsing, querying and visualizing data.

The tables that were created in hive terminal or hue editor are displayed in hue. We can perform Sql like operations such as
 select * from tweets_pulwama; The results displayed are in row-column format which can be easily understood. Fields such as twitter-userid,friends_count, follower_count,tweets_count,user_location etc. are displayed which can be easily analyzed a better insight the data can be further visualized by using Tableau.

Data Visualization: Tableau is a data visualization tool which generates real time data reports. Download Tableau desktop. Tableau connects to cdh data by using ODBC driver which can be downloaded from Tableau's website. Data can be visualized using charts, maps, graphs etc. in tableau.

RESULT ANALYSIS

A tragedy that occurred on 14th February 2019, in Pulwama district of India killing many jawans was trending topic discussed in Twitter throughout India. We used Pulwama as our hashtag to analyze the public mindset and opinion regarding this incident. Lots of data related to the keyword was generated. Twitter users from different parts of the globe tweeted under this hashtag. User details such as their twitter-id's, friends-count, followers-count, tweets-count, source from which they have posted, location-details etc. were analyzed. We received a data of 600 rows within a few minutes in HDFS. We then established connection to Tableau. For Analysis in Tableau, we took 2 crucial fields tweet-text and location from where the tweet was posted. We have analyzed the opinions of public regarding this incident.
From the results we got we have summarized users into three different categories. They are 1) Users who expressed their concerns and condolences to martyrs' families about the incident. 2) Users who blamed government for their in-ability to curb this incident and failing in taking measures aftermath. 3) Users praising the terrorists. These were mainly people of Pakistan and some Kashmiris.
From the results obtained we have also summarized the people using twitter based on the twitter-source. The decreasing order of users tweeting on different platforms are
Android>iPhone>web>Third-party-service.

FIGURE 7 ANALYSIS REPORT SHOWING TWEET-TEXT IN TABLEAU

View Data: FlumeData.1551190025789   — ☐ X

600 rows ➡ ☑ Show aliases

Text

RT @Manju37172096: On national Level Blood Donation Camp Organised By Dera Sacha Sauda With Teachings Of Saint @Gurmeetramrahim Singh Ji Ir...

RT @amritabhinder: @shahfaesal UN designated, Pak-based terrorist group Jaish-e-Muhammad claimed responsibility for Pulwama attack that kil...

RT @China_SCS_info: #Indian Strikes Thread: Hundreds of jihadis & trainers were moved from Pak Occupied Kashmir to a 5 star, resort style c...

RT @theskindoctor13: If you were criticising central government after Pulwama attack, you must give credit to them after today's retaliatio...

RT @VPSecretariat: Recent diabolical terrorist attack on our CRPF personnel in Pulwama once again underlines the global imperative to isola...

RT @mazdaki: Key points here:- Pak army spokesman concedes Indians came clear into Khyber-Pakhtunkhwa province — not disputed territory

RT @squintneon: UP Police has made 19 arrests while Assam Police has made 16 arrests to those who celebrated/insulted Pulwama Martyrs. apar...

RT @PeacePakistan5: Listen in to what mother of constable Pradeep Kumar, resident of Shamli in UP who was Killed in Pulwama attack, had to...

RT @aajtak: ऑस्ट्रेलिया ने पाकिस्तान से जैश ए मोहम्मद और लश्करे तैयबा जैसे संगठनों पर रोक लगाने की मांग की है.#IndiaStrikesBack

RT @manakgupta: Exactly and same with the 'liberals' here Vir Ji. Many have been calling the nation's Pulwama outrage 'warmongering'. Now q...

This is really how indian crafts came in Pak & ran away by having fear Of #PAF. We knew u were coward & fool so u s... https://t.co/ZwSwIG

RT @cjwerleman: The contrast between the way Pakistan and India's media has responded to the Pulwama Attack couldn't be more stark.

Inspired by Saint Dr @Gurmeetramrahim Singh Ji Insan, millions of @derasachasauda volunteers have pledged to donat... https://t.co/iXjtRYfoIF

#रक्तदान_शिविर_27फरवरी @derasachasauda is world known when it comes to act of human welfare. Continuing with the ch... https://t.co/tdZF45EXUj

RT @DeshGujarat: Fearing surgical strike along LOC/in POK,minimum 325 fidayeen and 25-27 trainers were shifted from POK to resort style cam...

RT @ReenaInsan0987: #रक्तदान_शिविर_27फरवरी@DeraSachaSauda known to stand for every humanitarian cause is conducting blood donation camp c...

RT @muglikar_: Post Pulwama attack, 300 highly trained terrorists moved away from PoK camps to Balakot. They had a 5 star camp with swimmin...

⟨   ▭   ⟩

FIGURE 8 - SHOWING TEXT-DATA OF TWEETS COLLECTED

Sheet 1

Text (Retweeted Status)	Location (User)
Null	Sialkot, Pakistan
1000 kgs bombs dropped ..	Kunda, Pratapgarh, U.P., I..
"Today is a day of pride for the whole of Delhi and the..	chandigarh
	Ulhasnagar, India
@livemint Yes Saw @nare..	Hyderabad, India
@RavinarIN @suhasinih a..	Null
@shahfaesal UN designat..	India, Maharashtra, Mum..
*@derasachasauda..	Null
#India is a liar & Indi..	North of seven billion hu..
#Indian Strikes Thread: Hundreds of jihadis &amp..	Null
	01 Jhandpath
#Indiastrikesback -- After ..	Rajasthan
#रक्तदान_शिविर_27फरवरी..	Dabwali
#रक्तदान_शिविर_27फरवरी..	Null
#रक्तदान_शिविर_27फरवरी Inspired by Saint Dr @Gur..	Null
	Delhi
#रक्तदान_शिविर_27फरवरी..	Null
#रक्तदान_शिविर_27फरवरी..	Panipat, India
#रक्तदान_शिविर_27फरवरी..	Panipat, India
#रक्तदान_शिविर_27फरवरी To Pay the Tribute To Pul..	bhucho mandi bathinda ,p..
	Mansa, India
◆◆ MSG My God ◆◆..	Null
After Pulwama attack,..	Bangalore
easy ho akash, strike rate tera 30 ka tha baten tu 16..	Bahawalpur, Pakistan
	United States of Misery.
Entire Pakistani nation st..	Dadu,Hyderabad Pakistan
Exactly and same with the..	Null
Fearing surgical strike alo..	Null
Going by the Indian claim..	Islamabad, Pakistan
IAF strike destroying terr..	United States
If India's claim of massive..	Srinagar
If you were criticising central government after ..	India
	Republic of UttaraPracha..

Sheet 1 | 📖 Story 1 | ⊞ Dashboard 1 | Sheet 2

Text (Retweeted Status)	Location (User)
Null	Null
	#i supportmsg
	Delhi	Pondicherry
	Delhi, India
	Dil Walo ki New Delhi
	Helabattala
	Hyderabad, India
	Islamabad, Pakistan
	JisDeshMein#SRKRaheta..
	Karachi
	New Delhi, India
	Punjab
	Sialkot, Pakistan
1000 kgs bombs dropped ..	Kunda, Pratapgarh, U.P., I..
"Today is a day of pride for the whole of Delhi and the..	chandigarh
	Ulhasnagar, India
@livemint Yes Saw @nare..	Hyderabad, India
@RavinarIN @suhasinih a..	Null
@shahfaesal UN designat..	India, Maharashtra, Mum..
*@derasachasauda..	Null
#India is a liar & Indi..	North of seven billion hu..
#Indian Strikes Thread: Hundreds of jihadis &amp..	Null
	01 Jhandpath
#Indiastrikesback -- After ..	Rajasthan
#रक्तदान_शिविर_27फरवरी..	Dabwali
#रक्तदान_शिविर_27फरवरी..	Null
#रक्तदान_शिविर_27फरवरी Inspired by Saint Dr @Gur..	Null
	Delhi
#रक्तदान_शिविर_27फरवरी..	Null
#रक्तदान_शिविर_27फरवरी..	Panipat, India
#रक्तदान_शिविर_27फरवरी..	Panipat, India
#रक्तदान_शिविर_27फरवरी To Pay the Tribute To Pul..	bhucho mandi bathinda ,p..
	Mansa, India

FIGURE 9 SHOWING LOCATION OF A USERS AND THEIR RESPECTIVE TWEETS
IN TABLEAU WORKSHEET
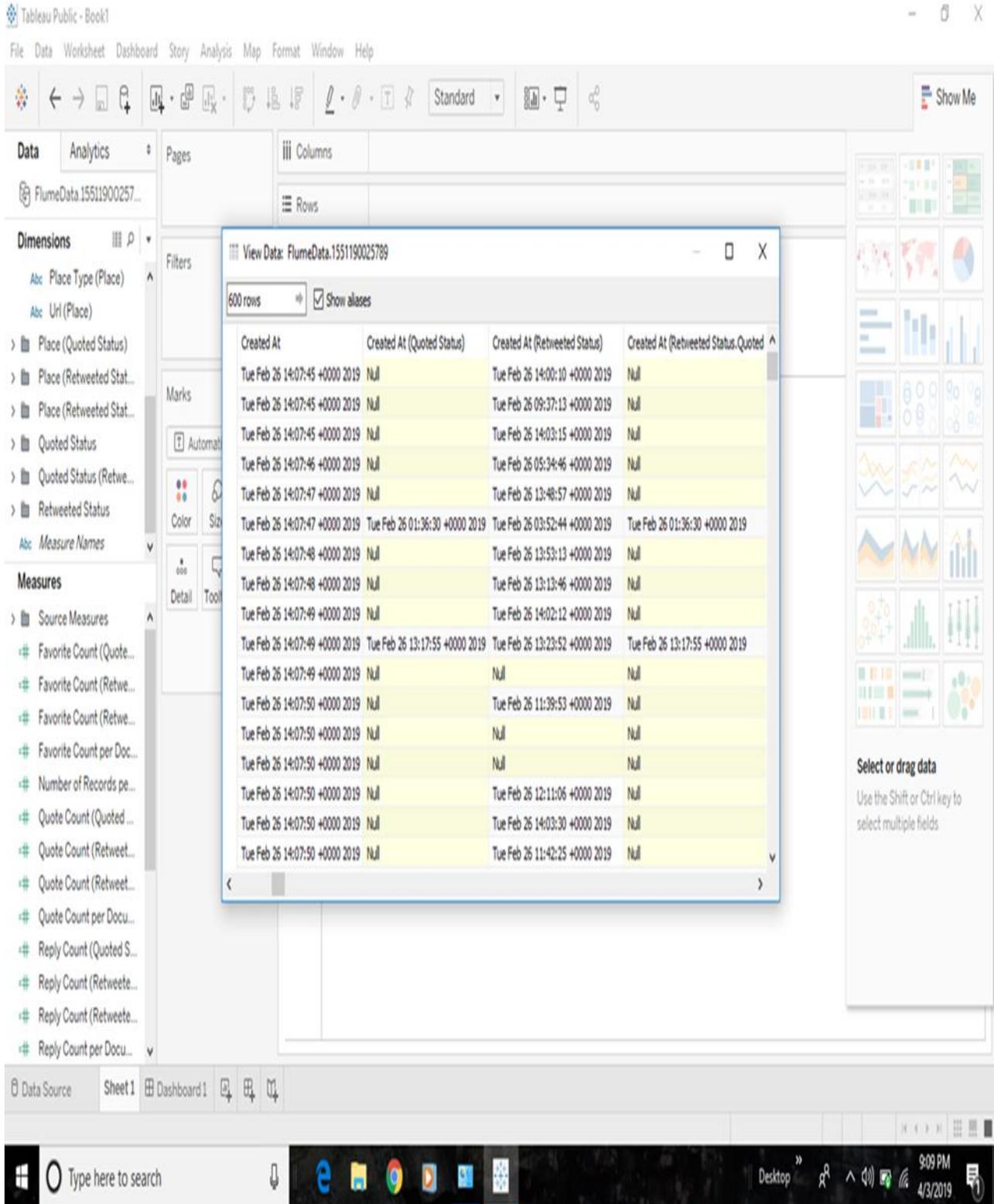
FIGURE 10 SHOWING CREATION DATE AND TIME OF TWEETS AND RETWEETS

FIGURE 11 SHOWING TWEETS SOURCE IN TABLEAU

FIGURE 12 SHOWING TIMESTAMP VALUES OF TWEETS GENERATED IN TABLEAU

EXPERIMENTAL SETUP

Minimum 12Gb ram --   Hardware
Oracle vm VirtualBox –   Software
CDH (version-5.13)   --   Software
Any latest O.S   –   Software
Tableau desktop   --   Software (Min)
10 Mbps Internet   -   Network

CONCLUSION

As twitter is widely used social networking micro-blogging site, tremendous amount of data is generated by users which gives value only if it is subjected to analysis. Public-opinions, product-reviews and various decision making can be done by just giving keywords as hashtags. Hadoop provides a platform for real time distributed storage and parallel processing. In this paper we have used Cloudera Quickstart Vm and collected large amount of tweets. We have used Hive for twitter analysis.
Tweets with the following hashtags as keywords such as "Pulwama"," Budgam", etc. were collected. We have used Hue and Tableau for further analysis. How users throughout the world are reacting to the incident and user-details such as their location, friends and followers count are analyzed. Based on the analysis performed the mindset of a society is judged. Decision making can also be done.

ACKNOWLEDGEMENT

REFERENCES

[1] E. Sivaraman," High Performance and Fault Tolerant Distributed File System for Big Data Storage and Processing using Hadoop".
[2] Can Uzunkaya," Hadoop Ecosystem and Its Analysis on Tweets"
[3] Monu Kumar," Analyzing Twitter Sentiments Through Big Data".
[4] Matthew Edwards "Design of Hadoop-based Framework for Analytics of Large Synchro phasor Datasets".
[5] Anis Zarrad," The Evaluation of The Public Opinion".
[6] Eugen Feller," Performance and Energy Efficiency of Big Data Applications in Cloud Environments: A Hadoop Case Study.
[7] www.cloudera.com
[8] www.wikipedia.com