Location based Analysis of Twitter Data using Apache Hive

Article *in* International Journal of Computer Applications · November 2016

DOI: 10.5120/ijca2016912170

Vijay K Trivedi
LNCT Group of Colleges

Location based Analysis of Twitter Data using Apache Hive

Manish Wankhede
Department of
Computer Science &
Engineering
Lakshmi Narain College of
Technology
Bhopal, India

Vijay Trivedi
Asst. Professor
Department of
Computer Science &
Engineering
Lakshmi Narain College of
Technology
Bhopal, India

Vineet Richhariya, PhD
Prof. & Head
Department of
Computer Science &
Engineering
Lakshmi Narain College of
Technology
Bhopal, India

ABSTRACT

Twitter, one of the largest and famous social media site receives millions of tweets every day on variety of important topic. This large amount of raw data can be used for industrial , Social, Economic, Government policies or business purpose by organizing according to our need and processing. Hadoop is one of the best tool options for twitter data analysis and hadoop works for distributed Big data , Streaming data , Time Stamped data , text data etc. This paper discuss how to use FLUME for extracting twitter data and store it into HDFS for analysis, and after that we are use apache hive for analysing these data. We perform analysis on twitter data to find the number of tweets are posted location wise and also finds the keywords on which maximum and minimum tweets are posted.

Keywords

Hadoop, twitter, Flume, social analysis, hive, HWI, JSON

1. INTRODUCTION

We live in a society and many people used social site where the textual data on the Internet is growing at a rapid pace and many companies are trying to use this flood of data to extract people's views towards their products. Micro blogging today has become a very prevalent communication tool in to Internet users. Twitter, one of the largest social media site and user tweet millions of tweets every day on deferent of important topic. Authors of those messages write about their life, share opinions on variety of issues and discuss current issues. These posts analysis can be used for decision making in different filds like Business, Elections, Product review, government, etc. Also sentiment analysis is one of the most important area of analysis of twitter posts that can be very useful for decision making.

Performing Sentiment Analysis on Twitter is trickier than doing it for large reviews. This is because the tweets are very short (only about 140 characters) and usually contain argot, emoticons, hash tags and other twitter specific jargon. For the development purpose twitter provides streaming API which allows developer an access to one percent (1%) of tweets tweeted at that time bases on the distinctive keyword. The object about which we want to execution sentiment analysis is submitted to the twitter API's which does ahead mining and provides the tweets related to only that keyword. Twitter data is normally unstructured form i.e use of abbreviations is very high. Also it permit the use of emoticons which are direct indicators of the author's view on the topic. Tweet messages also consist of a the user name and timestamp. This timestamp is useful for guessing the future trend application of our project. If User location available we can also help to gauge the trends in different geographical regions.

1.1 Hadoop

The Apache Hadoop 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 a 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 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.

1.2 Flume

Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS). It can be used for dumping twitter data in Hadoop HDFS.

2. LITERATURE REVIEW

Mahalakshmi R, Suseela [2] (2015) Big-SoSA: Social Sentiment Analysis and Data Visualization on Big Data . It proposes a method of sentiment analysis on twitter by using Hadoop and its ecosystems that process the large volume of data on a Hadoop and the MapReduce function performs the sentiment analysis.

Praveen Kumar, Dr Vijay Singh Rathore [3] (2014) Efficient Capabilities of Processing of Big Data using Hadoop Map Reduce Proposes, several solutions to the Big Data problem have emerged which includes the Map Reduce environment championed by Google which is now available open-source in Hadoop. Hadoops distributed processing, Map Reduce algorithms and overall architecture are a major step towards achieving the promised benefits of Big Data.

Sunil B. Mane, Yashwant Sawant, Saif Kazi [1] (2014) Real Time Sentiment Analysis of Twitter Data Using Hadoop. Proposes and provides a way of sentiment analysis using Hadoop which will process the huge amount of data on a Hadoop cluster(faster in real time).

Manoj Kumar Danthala [4] (2015) Tweet Analysis: Twitter Data processing Using Apache Hadoop . 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 pictorial representations of twitter users and their tweets.

Manoj Kumar Danthala [5] (2015) Bigdata Analysis: Streaming Twitter Data with Apache Hadoop and Visualizing using Big Insights. It proposes, twitter data, which is the largest social networking area where data is increasing at high rates every day is considered as big data. This data is processed and analyzed using InfoSphere BigInsights tool which bring the power of Hadoop to the enterprise in real time. This also includes the visualizations of analyzing big data charts using big sheets.

Judith Sherin Tilsha S, Shobha M.S [6] (2015) A Survey on Twitter Data Analysis Techniques to Extract Public Opinion. Using machine learning algorithm ,a feature vector is constructed with the emotion describing words from tweets and are fed to the classifier that classifies the sentiment or opinion. It said that various twitter data analysis techniques that are based on dictionary and that are using the machine learning approaches.

Mr.Sagar Nadagoud [7] (2015), Market Sentiment Analysis for Popularity of Flipkart. It is taking sentiment analysis, for this it is using Hive and its queries to give the sentiment data based up on the groups that have defined in the HQL (Hive Query Language). Here they had categorized this sentiment analysis into 3 groups like tweets that are having positive, neutral and negative comments.

Ramesh R, Divya G, Divya D, Merin K Kurian [8] (2015), Big Data Sentiment Analysis using Hadoop. The main focus of the research was to find such a technique that can efficiently perform Sentiment Analysis on Big Data sets. In this paper Sentiment Analysis was performed on a large data set of tweets using Hadoop and the performance of the technique was measured in form of speed and accuracy. The experimental result shows that the technique exhibits very good efficiency in handling big sentiment data sets.

G.Vinodhini , RM.Chandrasekaran [9] (2012), Sentiment Analysis and Opinion Mining: A Survey. An accurate method for predicting sentiments could enable us, to extract opinions from the internet and predict online customer's preferences, which could prove valuable for economic or marketing research. Till now, there are few different problems predominating in this research community, namely, sentiment classification, feature based classification and handling negations. This paper presents a survey covering the techniques and methods in sentiment analysis and challenges appear in the field.

3. OBSERVATION

Hadoop and its Ecosystems, for getting raw data from the Social Network, we may use Hadoop online streaming tool-using Apache Flume. 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 create the table and filter the information that is needed for us and sort them into the Hive Table. And from this, we are going to perform the Sentiment Analysis by using some UDF's (User Defined Functions) by which we can perform sentiment analysis.

4. PROBLEM DEFINITION

Social media is one of the popular media right now to share opinions or variety of topics and twitter is very popular social site to share every thing related to opinions on variety of topics and discussions on current issues. 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. Twitter sites receives petabytes of data every day and these data is nothing but a collection of tweets so these data is very important in real life to analyse different scenario through which its helps us in decision making. The analysis of twitter data gives real view or different user opinions regarding what they think and to analysis these data provide a better way for making any decision.

5. PROPOSED WORK

For analysing these large and complex data required a power tool, we are using hadoop[10] which is a open source implementation of mapreduce, a powerful tool designed for deep analysis and transformation of very large data.

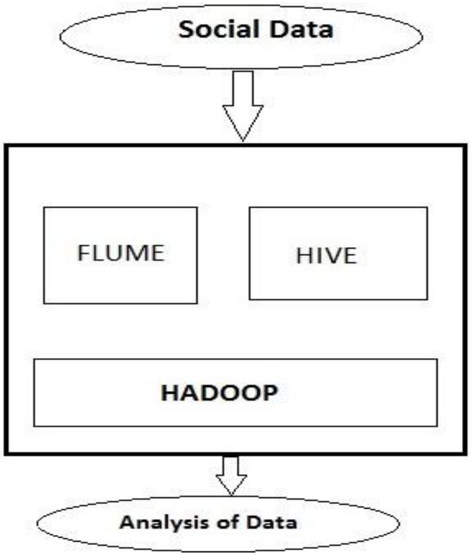


Figure1. Workflow Diagram

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 [11] where the complete data is mapped to frequent datasets and reduced to smaller sizable data to ease of handling ,after this we integrate hadoop ecosystem eg. Flume and Hive on top of the hadoop. The pre-requesite for flume and hive is the hadoop should be pre-install. Flume is used to fetching real time twitter data and stored in HDFS and after the data storage we are performing analysis of these complex data using hive.

6. PROPOSED METHODOLOGY:

Our Steps or Algorithm Steps will follow:

1. In first step We are creating a twitter app using a twitter streaming API for fetching real time twitter data.
2. For doing twitter data analysis first data is uploaded using FLUME in local HDFS. The twitter API used in Flume , through which all the tweets are directly fetch from the twitter site and stored it into the HDFS. Data comes from the twitter site is in un-structure form called JSON data.
3. After storing all twitter data into the HDFS we are performing the analysis part for these we use hive through which we can convert the un-structure complex data in to readable or understandable structure form.
4. Tweets are preprocesses for removing noise and meaningless symbols. And then the data is available in the form of schema oriented , and using hive we are analyze the data by writing a different queries for decision making.

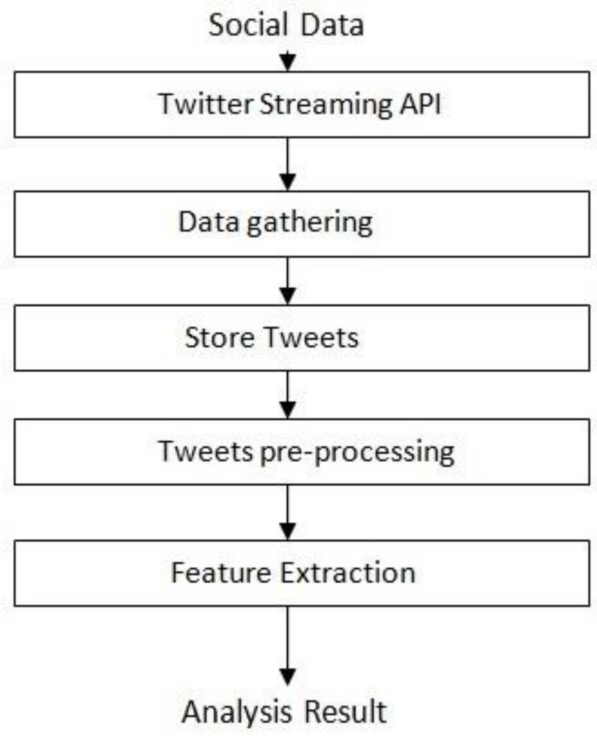


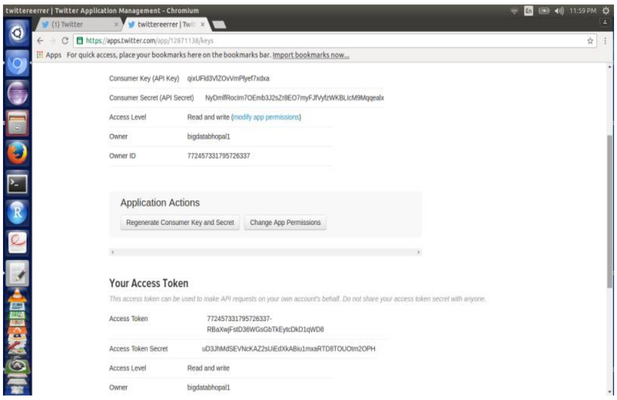
Figure 2. Analysis Step

7. EXPERIMENTAL & RESULT ANALYSIS

All the analysis will perform on intel i3 processor with 4GB of RAM windows machine, in windows machine we can install vmware with configuration of 2 processor and 2GB of RAM and ubuntu is installed on vmware machine. On ubuntu machine we can install java and configure hadoop-1.1.2 version [13] on ubuntu machine.

7.1 Gathering real time twitter data

Before analysis we have to collect twitter data on which we can perform analysis, for collecting twitter data we hav to create a twitter app through which we can generate a consumer and token access keys through which we can fetch real time twiiter data using flume. Apache flume is integrated over hadoop which can getch data from source which is a twitter data and through memory channel workflow stored it into the HDFS which is a sink in our flume configuration. Figure 3 shows that consumer and access token keys are generated through twitter application.


twittereerrer | Twitter Application Management - Chromium
11:59 PM
(1) Twitter
twittereerrer | Twi
https://apps.twitter.com/app/12871138/keys
Apps For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Regenerate Consumer Key and Secret
Change App Permissions

Consumer Key (API Key) qixUFld3VlZOvVmPlyef7xdxa
Consumer Secret (API Secret) NyDmlfRoclm7OEmb3J2sZr8EO7myFJfVyfzWKBLIcM9Mqqealx
Access Level Read and write (modify app permissions)
Owner bigdatabhopal1
Owner ID 772457331795726337

Application Actions

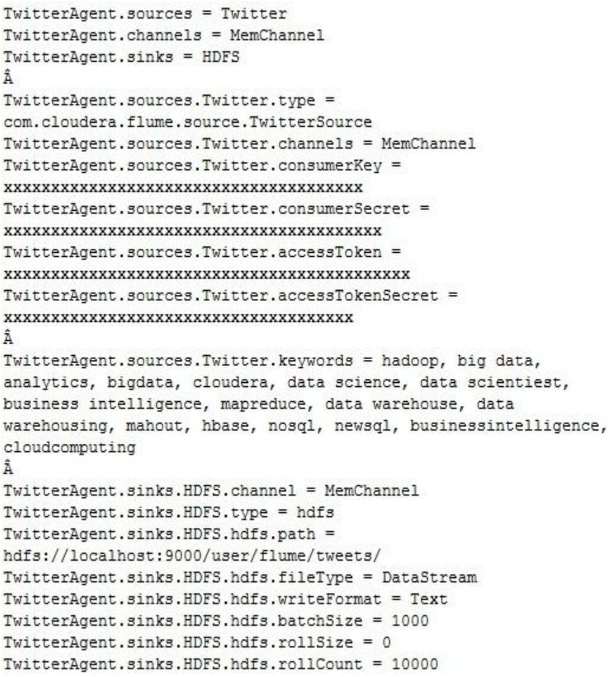
Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token 772457331795726337-RBaXwjFstD36WGsGbTkEytcDkD1qWD8
Access Token Secret uD3JhMdSEVNcKAZ2sUiEdXkABiu1mxaRTD8TOUOtm2OPH
Access Level Read and write
Owner bigdatabhopal1

Figure 3. Generating consumer and access token keys

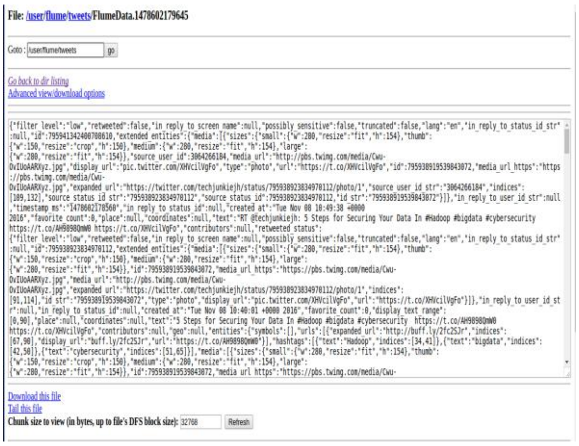
After generating a consumer and access token keys we can configure flume on which we can configure web sources from where the data is coming in our paper it's a twitter and intermediate channels which provide workflow and than configure a sink from which the data is stored and that is HDFS in our paper. And the consumer and access token keys are also written in configuration file shown in figure 4.



```
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 =
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
TwitterAgent.sources.Twitter.consumerSecret =
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
TwitterAgent.sources.Twitter.accessToken =
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
TwitterAgent.sources.Twitter.accessTokenSecret =
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
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
```

Figure 4. Flume-twitter.conf configuration file

After configuring the flume-twiiter.conf file we can start a twiiter agent through terminal which can a twitter API which provide a conncetion establishment between HDFS and twitter and the tweets are start downloading from the twitter and stored it into the the HDFS. Figure 5 shows the twitter JSON data coming from the twitter and stored into HDFS.



File: /user/flume/tweets/FlumeData.1478602179645

Goto : /user/flume/tweets go

Go back to dir listing
Advanced view/download options

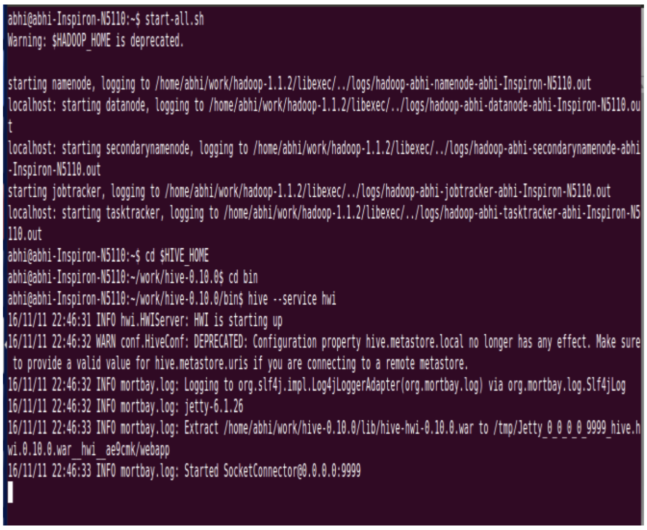
```
{"filter_level":"low","retweeted":false,"in_reply_to_screen_name":null,"possibly_sensitive":false,"truncated":false,"lang":"en","in_reply_to_status_id_str"
:null,"id":795941342400708610,"extended_entities":{"media":[{"sizes":{"small":{"w":280,"resize":"fit","h":154},"thumb":
{"w":150,"resize":"crop","h":150},"medium":{"w":280,"resize":"fit","h":154},"large":
{"w":280,"resize":"fit","h":154}},"source_user_id":3064266184,"media_url":"http://pbs.twimg.com/media/Cwu-
OvIUoAARXyz.jpg","display_url":"pic.twitter.com/XHVcilVgFo","type":"photo","url":"https://t.co/XHVcilVgFo","id":795938919539843072,"media_url_https":"https
://pbs.twimg.com/media/Cwu-
OvIUoAARXyz.jpg","expanded_url":"https://twitter.com/techjunkiejh/status/795938923834970112/photo/1","source_user_id_str":"3064266184","indices":
[109,132],"source_status_id_str":"795938923834970112","source_status_id":795938923834970112,"id_str":"795938919539843072"}]},"in_reply_to_user_id_str":null
,"timestamp_ms":"1478602178560","in_reply_to_status_id":null,"created_at":"Tue Nov 08 10:49:38 +0000
2016","favorite_count":0,"place":null,"coordinates":null,"text":"RT @techjunkiejh: 5 Steps for Securing Your Data In #Hadoop #bigdata #cybersecurity
https://t.co/AH9898QmW0 https://t.co/XHVcilVgFo","contributors":null,"retweeted_status":
{"filter_level":"low","retweeted":false,"in_reply_to_screen_name":null,"possibly_sensitive":false,"truncated":false,"lang":"en","in_reply_to_status_id_str"
:null,"id":795938923834970112,"extended_entities":{"media":[{"sizes":{"small":{"w":280,"resize":"fit","h":154},"thumb":
{"w":150,"resize":"crop","h":150},"medium":{"w":280,"resize":"fit","h":154},"large":
{"w":280,"resize":"fit","h":154}},"id":795938919539843072,"media_url_https":"https://pbs.twimg.com/media/Cwu-
OvIUoAARXyz.jpg","media_url":"http://pbs.twimg.com/media/Cwu-
OvIUoAARXyz.jpg","expanded_url":"https://twitter.com/techjunkiejh/status/795938923834970112/photo/1","indices":
[91,114],"id_str":"795938919539843072","type":"photo","display_url":"pic.twitter.com/XHVcilVgFo","url":"https://t.co/XHVcilVgFo"}]},"in_reply_to_user_id_st
r":null,"in_reply_to_status_id":null,"created_at":"Tue Nov 08 10:40:01 +0000 2016","favorite_count":0,"display_text_range":
[0,90],"place":null,"coordinates":null,"text":"5 Steps for Securing Your Data In #Hadoop #bigdata #cybersecurity  https://t.co/AH9898QmW0
https://t.co/XHVcilVgFo","contributors":null,"geo":null,"entities":{"symbols":[],"urls":[{"expanded_url":"http://buff.ly/2fc2SJr","indices":
[67,90],"display_url":"buff.ly/2fc2SJr","url":"https://t.co/AH9898QmW0"}],"hashtags":[{"text":"Hadoop","indices":[34,41]},{"text":"bigdata","indices":
[42,50]},{"text":"cybersecurity","indices":[51,65]}],"media":[{"sizes":{"small":{"w":280,"resize":"fit","h":154},"thumb":
{"w":150,"resize":"crop","h":150},"medium":{"w":280,"resize":"fit","h":154},"large":
{"w":280,"resize":"fit","h":154}},"id":795938919539843072,"media_url_https":"https://pbs.twimg.com/media/Cwu-
```

Download this file
Tail this file
Chunk size to view (in bytes, up to file's DFS block size): 32768 Refresh

Figure 5. Twitter JSON raw data stored in HDFS

Start analysis of twitter JSON Data

The JSON data comes fro mthe twitter shown into figure 5 are in the unstructured form, so we need a powerful analytical tool to analyse these type of data, so we can use apache hive [12] which is a open source data warehouse analytical tool. So we can integrate hive with hadoop to analyse the data and along with configure a hive web interface shown in figure 6 in which we connect hadoop with hive web interface.

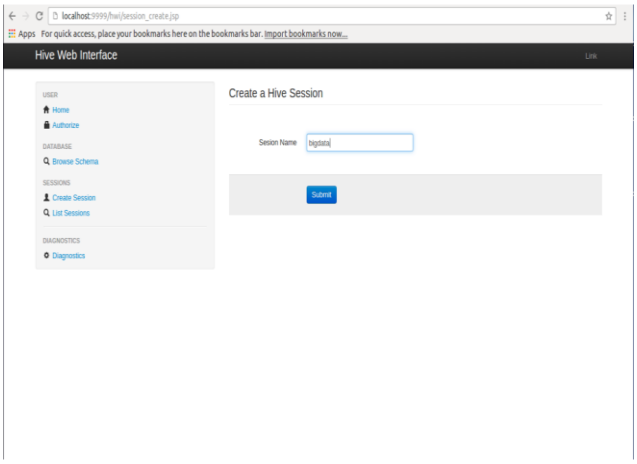


```
abhi@abhi-Inspiron-N5110:~$ start-all.sh
Warning: $HADOOP_HOME is deprecated.

starting namenode, logging to /home/abhi/work/hadoop-1.1.2/libexec/../logs/hadoop-abhi-namenode-abhi-Inspiron-N5110.out
localhost: starting datanode, logging to /home/abhi/work/hadoop-1.1.2/libexec/../logs/hadoop-abhi-datanode-abhi-Inspiron-N5110.ou
t
localhost: starting secondarynamenode, logging to /home/abhi/work/hadoop-1.1.2/libexec/../logs/hadoop-abhi-secondarynamenode-abhi
-Inspiron-N5110.out
starting jobtracker, logging to /home/abhi/work/hadoop-1.1.2/libexec/../logs/hadoop-abhi-jobtracker-abhi-Inspiron-N5110.out
localhost: starting tasktracker, logging to /home/abhi/work/hadoop-1.1.2/libexec/../logs/hadoop-abhi-tasktracker-abhi-Inspiron-N5
110.out
abhi@abhi-Inspiron-N5110:~$ cd $HIVE_HOME
abhi@abhi-Inspiron-N5110:~/work/hive-0.10.0$ cd bin
abhi@abhi-Inspiron-N5110:~/work/hive-0.10.0/bin$ hive --service hwi
16/11/11 22:46:31 INFO hwi.HWIServer: HWI is starting up
16/11/11 22:46:32 WARN conf.HiveConf: DEPRECATED: Configuration property hive.metastore.local no longer has any effect. Make sure
 to provide a valid value for hive.metastore.uris if you are connecting to a remote metastore.
16/11/11 22:46:32 INFO mortbay.log: Logging to org.slf4j.impl.Log4jLoggerAdapter(org.mortbay.log) via org.mortbay.log.Slf4jLog
16/11/11 22:46:32 INFO mortbay.log: jetty-6.1.26
16/11/11 22:46:33 INFO mortbay.log: Extract /home/abhi/work/hive-0.10.0/lib/hive-hwi-0.10.0.war to /tmp/Jetty_0_0_0_0_9999_hive.h
wi.0.10.0.war__hwi__ae9cmk/webapp
16/11/11 22:46:33 INFO mortbay.log: Started SocketConnector@0.0.0.0:9999
```

Figure 6. Integration HWI with hadoop

After integrating hadoop with HWI we can open browser and open hive web interface [13] using url localhost:9999/hwi shown in figure 7.



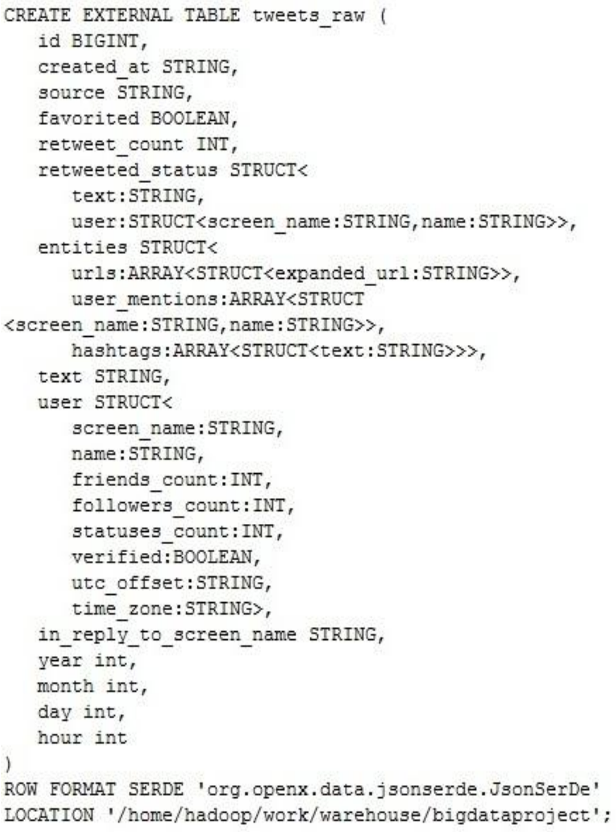
localhost:9999/hwi/session_create.jsp

Create a Hive Session

Sesion Name: bigdata

Figure 7. GUI of Hive with session start

After opening a hive web interface we need to create a session, because analysis cannot be perform on hive web interface without creating a session. After creating a session we can create a hive table to store twitter JSON data in to table means we can store such unstructured data into some structured form. We can user json serde through which we can convert unstructured data into structured data and stored these data into hive table. The hive table named tweets_raw shown in figure 8.



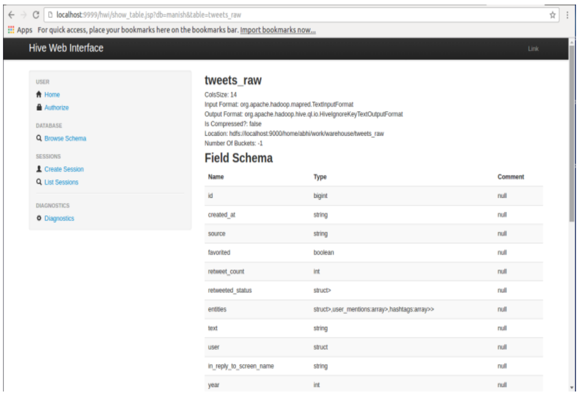
```
CREATE EXTERNAL TABLE tweets_raw (
   id BIGINT,
   created_at STRING,
   source STRING,
   favorited BOOLEAN,
   retweet_count INT,
   retweeted_status STRUCT<
      text:STRING,
      user:STRUCT<screen_name:STRING,name:STRING>>,
   entities STRUCT<
      urls:ARRAY<STRUCT<expanded_url:STRING>>,
      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:STRING,
      time_zone:STRING>,
   in_reply_to_screen_name STRING,
   year int,
   month int,
   day int,
   hour int
)
ROW FORMAT SERDE 'org.openx.data.jsonserde.JsonSerDe'
LOCATION '/home/hadoop/work/warehouse/bigdataproject';
```

Figure 8. hive table which store twitter data

After creating a table we can load the twitter data into hive table with serde properties with command

Load data inpath ‘/user/flume/tweets/’ overwrite into table tweets_raw;

So the data are stored in a structure manner we can also check the structure called schema of the table shown in figure 9.


localhost:9999/hwi/show_table.jsp?db=manish&table=tweets_raw
Apps For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Hive Web Interface
Link
USER
Home
Authorize
DATABASE
Browse Schema
SESSIONS
Create Session
List Sessions
DIAGNOSTICS
Diagnostics

tweets_raw

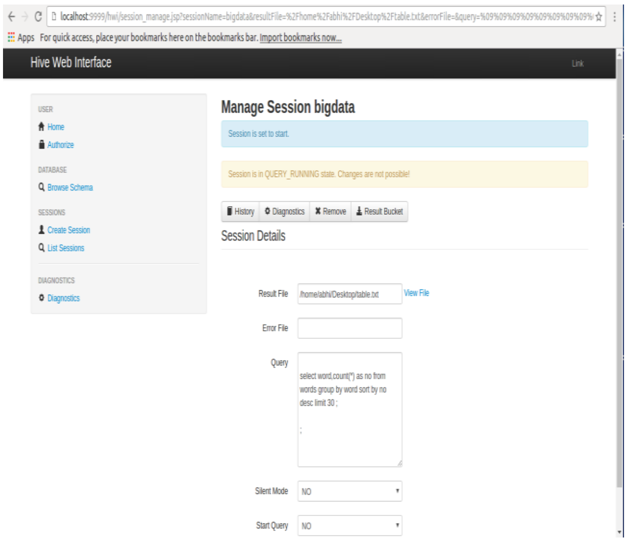
ColsSize: 14
Input Format: org.apache.hadoop.mapred.TextInputFormat
Output Format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Is Compressed?: false
Location: hdfs://localhost:9000/home/abhi/work/warehouse/tweets_raw
Number Of Buckets: -1

Field Schema

Name	Type	Comment
id	bigint	null
created_at	string	null
source	string	null
favorited	boolean	null
retweet_count	int	null
retweeted_status	struct>	null
entities	struct>,user_mentions:array>,hashtags:array>>	null
text	string	null
user	struct	null
in_reply_to_screen_name	string	null
year	int	null

Figure 9. Structure of hive tweets_raw table

After loading JSON data into hive table we can perform analysis , so we can write SQL queries to analyse these data using a session which is created above. Figure 10 shows the session details which contains a result file on which locaton the analysis result will be stored as a textfile and error file also. In query field we can write sql query to analyse data.


localhost:9999/hwi/session_manage.jsp?sessionName=bigdata&resultFile=%2Fhome%2Fabhi%2FDesktop%2Ftable.txt&errorFile=&query=%09%09%09%09%09%09%09%09%
Apps For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Hive Web Interface
Link
USER
Home
Authorize
DATABASE
Browse Schema
SESSIONS
Create Session
List Sessions
DIAGNOSTICS
Diagnostics
History
Diagnostics
Remove
Result Bucket
View File

Manage Session bigdata

Session is set to start.

Session is in QUERY_RUNNING state. Changes are not possible!

Session Details

Result File: /home/abhi/Desktop/table.txt

Error File:

Query: select word,count(*) as no from words group by word sort by no desc limit 30 ;

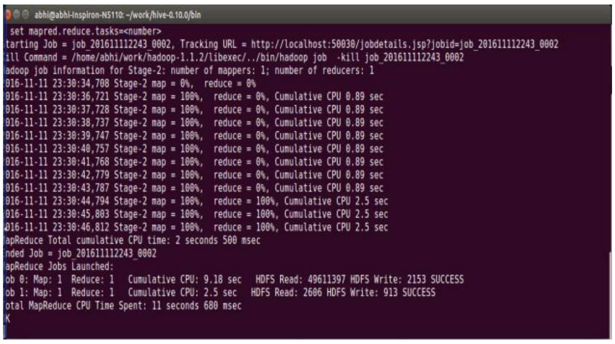
;

Silent Mode: NO

Start Query: NO

Figure 10. Running SQL queries on hive web interface

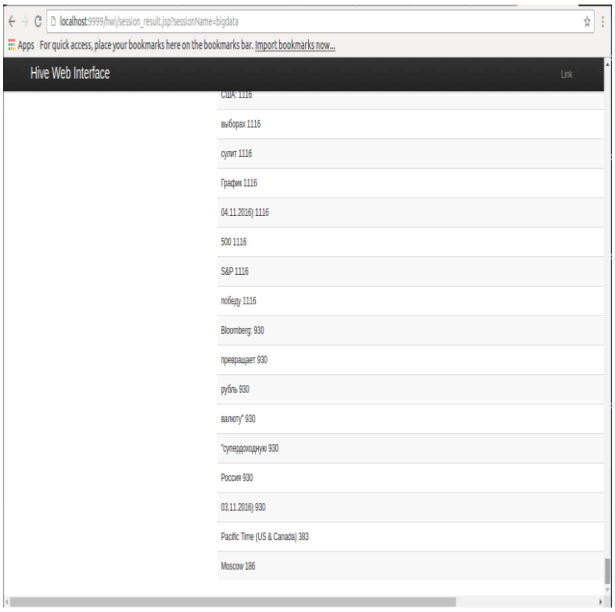
After running a SQL query on hive web interface a mapreduce job is launched for the hive query on the backend of the interface , which is on the terminal shown in figure 11.



```
 set mapred.reduce.tasks=<number>
tarting Job = job_201611112243_0002, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201611112243_0002
ill Command = /home/abhi/work/hadoop-1.1.2/libexec/../bin/hadoop job  -kill job_201611112243_0002
adoop job information for Stage-2: number of mappers: 1; number of reducers: 1
016-11-11 23:30:34,708 Stage-2 map = 0%,  reduce = 0%
016-11-11 23:30:36,721 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.89 sec
016-11-11 23:30:37,728 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.89 sec
016-11-11 23:30:38,737 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.89 sec
016-11-11 23:30:39,747 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.89 sec
016-11-11 23:30:40,757 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.89 sec
016-11-11 23:30:41,768 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.89 sec
016-11-11 23:30:42,779 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.89 sec
016-11-11 23:30:43,787 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 0.89 sec
016-11-11 23:30:44,794 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.5 sec
016-11-11 23:30:45,803 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.5 sec
016-11-11 23:30:46,812 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 2.5 sec
apReduce Total cumulative CPU time: 2 seconds 500 msec
nded Job = job_201611112243_0002
apReduce Jobs Launched:
ob 0: Map: 1  Reduce: 1   Cumulative CPU: 9.18 sec   HDFS Read: 49611397 HDFS Write: 2153 SUCCESS
ob 1: Map: 1  Reduce: 1   Cumulative CPU: 2.5 sec   HDFS Read: 2606 HDFS Write: 913 SUCCESS
otal MapReduce CPU Time Spent: 11 seconds 680 msec
K
```

Figure 11. launching MapReduce Job for each query

The result of the query are stored in text file which is mention on the session details and also the result are display on hive web interface shown in figure 12.


localhost:9999/hwi/session_result.jsp?sessionName=bigdata
Apps For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Hive Web Interface
Link

```
США: 1116
выборах 1116
сулит 1116
График 1116
04.11.2016) 1116
500 1116
S&P 1116
победу 1116
Bloomberg: 930
превращает 930
рубль 930
валюту" 930
"супердоходную 930
Россия 930
03.11.2016) 930
Pacific Time (US & Canada) 383
Moscow 186
```

Figure 12. Result of the query analysis

We can analyse the data and file the number of tweets are posted on the selected fields location wise.thw query will be

select count(*) as no,user.time_zone from tweets_raw group by user.time_zone sort by no desc; which return a total number tweets are tweeted location wise and also we can find the keywords on which maximum and minimum tweets are posted shown in figure 12.

8. CONCLUSION

On analysing complete scenario regarding the analysis of social data we say that using traditional analytical tool we can not perform analysis on such huge and complex data , so we uses a new powerful tool which is designed for deep analysis called hadoop and also integrate with its ecosystem FLUME, HIVE . both the ecosystem runs on top of the hadoop and flume is uses for fetching data and stored it in HDFS and than we uses hive for analysing these huge and complex data. We perform analysis on twitter data to find the number of tweets are posted location wise and also finds the keywords on which maximum and minimum tweets are posted. In future enhancement of this paper is to analyse this data for some other prarmeter to helps us in decision making.

9. REFERENCES

[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] 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] 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.

[4] Manoj Kumar Danthala, "Tweet Analysis: Twitter Data processing Using Apache Hadoop", International Journal Of Core Engineering & Management (IJCEM) Volume 1, Issue 11, February 2015, pp 94-102.

[5] Manoj Kumar Danthala, "Bigdata Analysis: Streaming Twitter Data with Apache Hadoop and Visualizing using BigInsights", International Journal of Engineering Research & Technology, Volume. 4 - Issue. 05 , May – 2015.

[6] 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.

[7] Mr. Sagar Nadagoud, Mr. Kotresh Naik.D, "Market Sentiment Analysis for Popularity of Flipkart ", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume4Issue5,May2015,pp 2117-2123.

[8] 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.

[9] 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.

[10] 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/

[11] Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce",6-8 Dec. 2012.

[12] Hive kiwi at http://www.apache.org/hadoop/hive.

[13] Hadoop Map-Reduce Tutorial at http://hadoop.apache.org/common/docs/current/hdfs_user_guide.html.

[14] Hive Performance Benchmark. Available at http://issues.apache.org/jira/browse/HIVE-396.