Opinion Mining of Twitter Data using Hadoop and Apache Pig

Anjali Barskar
Department of
Computer Science &
Engineering
Shri Balaji Institute of
Technology & Management
Betul, India

Ajay Phulre
HOD(CSE)
Department of
Computer Science &
Engineering
Shri Balaji Institute o
Technology & Management
Betul, 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 opinion mining because twitter contains variety of opinions on various topics so we have to analyse these opinions using hadoop and its ecosystems to check every tweets polarity either tweets contains positive ,negative or neutral opinions on particular topic. This paper provides an efficient mechanism to perform opinion mining by coming up with a finish to finish pipeline with the assistance of Apache Flume ,Apache HDFS, and Apache Pig.

Here we have used dictionary based approach for analysis for which we have implemented pig statements through which we can analysis these complex twitter data to check polarity of the tweets based on the polarity dictionary through which we can say that which tweets have negative opinion or positive opinion.

Keywords
Hadoop, twitter, Flume, opinion mining, social analysis, apache pig.

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 fields 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 [1] 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.

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 [2] 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.

APACHE 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.

APACHE PIG
Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

2. LITERATURE REVIEW

Mahalakshmi R, Suseela [3] (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 [10] (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 [9] (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 [11] (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 by using apache pig. 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[13] which is a open source implementation of mapreduce, a powerful tool designed for deep analysis and transformation of very large data.

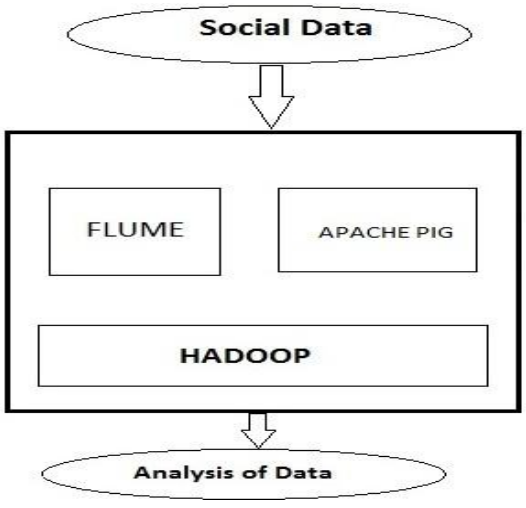


Figure1. Workflow Diagram

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 [12] 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 and Pig on top of the hadoop. The pre-requesite for flume,hive and pig 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 pig.

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 structure data and than we can analyse this data by using apache pig which is a very powerful tool for analysis.

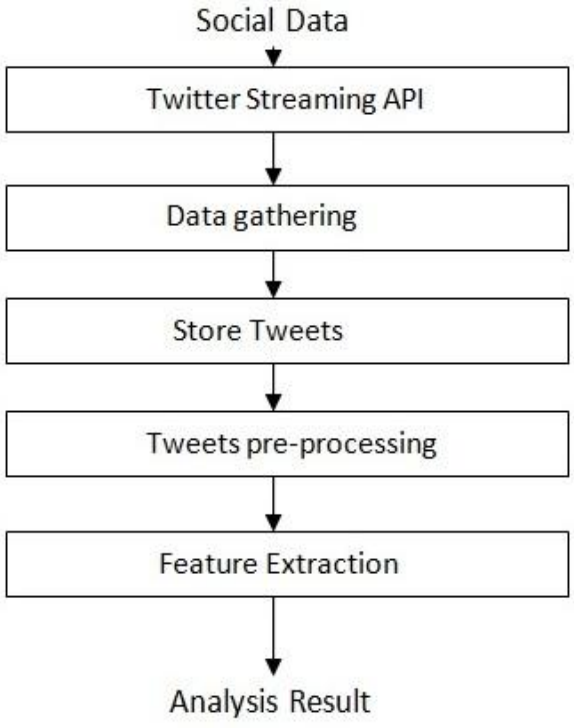


Figure 2. Analysis Step

7. EXPERIMENTAL AND RESULT ANALYSIS

All the experiments were performed using an i5-2410M CPU @ 2.30 GHz processor and 4 GB of RAM running ubuntu 14 . 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.
- Analyze using apache pig.

Creating Twitter Application

First of all if we want to do opinion 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. 3 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.

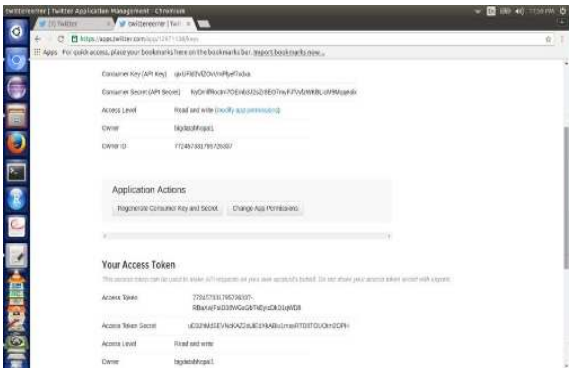
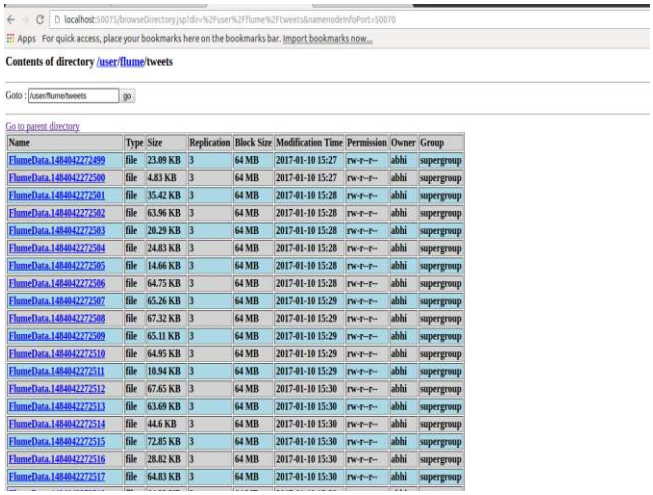


Figure 3. Creating twitter application

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.

```
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 = xxxxxxxxxxxxxxxxxx
TwitterAgent.sources.Twitter.consumerSecret = xxxxxxxxxxxxxxxxx
TwitterAgent.sources.Twitter.accessToken = xxxxxxxxxxxxxxxxxxxxx
TwitterAgent.sources.Twitter.accessTokenSecret = xxxxxxxxxxxxxxx
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
TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
```



Contents of directory /user/flume/tweets

Goto : /user/flume/tweets go

Go to parent directory

Name	Type	Size	Replication	Block Size	Modification Time	Permission	Owner	Group
FlumeData.1484042272499	file	23.09 KB	3	64 MB	2017-01-10 15:27	rw-r--r--	abhi	supergroup
FlumeData.1484042272500	file	4.83 KB	3	64 MB	2017-01-10 15:27	rw-r--r--	abhi	supergroup
FlumeData.1484042272501	file	35.42 KB	3	64 MB	2017-01-10 15:28	rw-r--r--	abhi	supergroup
FlumeData.1484042272502	file	63.96 KB	3	64 MB	2017-01-10 15:28	rw-r--r--	abhi	supergroup
FlumeData.1484042272503	file	20.29 KB	3	64 MB	2017-01-10 15:28	rw-r--r--	abhi	supergroup
FlumeData.1484042272504	file	24.83 KB	3	64 MB	2017-01-10 15:28	rw-r--r--	abhi	supergroup
FlumeData.1484042272505	file	14.66 KB	3	64 MB	2017-01-10 15:28	rw-r--r--	abhi	supergroup
FlumeData.1484042272506	file	64.75 KB	3	64 MB	2017-01-10 15:28	rw-r--r--	abhi	supergroup
FlumeData.1484042272507	file	65.26 KB	3	64 MB	2017-01-10 15:29	rw-r--r--	abhi	supergroup
FlumeData.1484042272508	file	67.32 KB	3	64 MB	2017-01-10 15:29	rw-r--r--	abhi	supergroup
FlumeData.1484042272509	file	65.11 KB	3	64 MB	2017-01-10 15:29	rw-r--r--	abhi	supergroup
FlumeData.1484042272510	file	64.95 KB	3	64 MB	2017-01-10 15:29	rw-r--r--	abhi	supergroup
FlumeData.1484042272511	file	10.94 KB	3	64 MB	2017-01-10 15:29	rw-r--r--	abhi	supergroup
FlumeData.1484042272512	file	67.65 KB	3	64 MB	2017-01-10 15:30	rw-r--r--	abhi	supergroup
FlumeData.1484042272513	file	63.69 KB	3	64 MB	2017-01-10 15:30	rw-r--r--	abhi	supergroup
FlumeData.1484042272514	file	44.6 KB	3	64 MB	2017-01-10 15:30	rw-r--r--	abhi	supergroup
FlumeData.1484042272515	file	72.85 KB	3	64 MB	2017-01-10 15:30	rw-r--r--	abhi	supergroup
FlumeData.1484042272516	file	28.82 KB	3	64 MB	2017-01-10 15:30	rw-r--r--	abhi	supergroup
FlumeData.1484042272517	file	64.83 KB	3	64 MB	2017-01-10 15:30	rw-r--r--	abhi	supergroup

Figure 4. Tweets Flumedata stored in HDFS

Analyze using apache pig:-

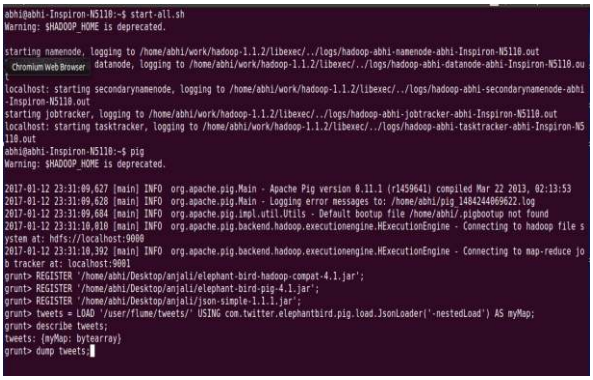
For analyzing JSON data which are coming from flume we are configuring apache pig on top of the hadoop. After configuring pig on top of the hadoop we have to load list of jar files which is registred in pig grunt shell. The list of jar files are listed below:-

REGISTER '/home/user/Desktop/elephant-bird-hadoop-compat-4.1.jar';

REGISTER '/home/user/Desktop/elephant-bird-pig-4.1.jar';

REGISTER '/home/user/Desktop/json-simple-1.1.1.jar';

After registering jar files we can load the tweets from hdfs into pig using elephant bird jsonLoader , with the help of these jar file we can load the json twitter data into pig, the command or query are shown in figure 5.



```
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
 Chromium Web Browser 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:~$ pig
Warning: $HADOOP_HOME is deprecated.

2017-01-12 23:31:09,627 [main] INFO  org.apache.pig.Main - Apache Pig version 0.11.1 (r1459641) compiled Mar 22 2013, 02:13:53
2017-01-12 23:31:09,628 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/abhi/pig_1484240866622.log
2017-01-12 23:31:09,684 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/abhi/.pigbootup not found
2017-01-12 23:31:10,010 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file s
ystem at: hdfs://localhost:9000
2017-01-12 23:31:10,392 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce jo
b tracker at: localhost:9001
grunt> REGISTER '/home/abhi/Desktop/anjali/elephant-bird-hadoop-compat-4.1.jar';
grunt> REGISTER '/home/abhi/Desktop/anjali/elephant-bird-pig-4.1.jar';
grunt> REGISTER '/home/abhi/Desktop/anjali/json-simple-1.1.1.jar';
grunt> tweets = LOAD '/user/flume/tweets/' USING com.twitter.elephantbird.pig.load.JsonLoader('-nestedLoad') AS myMap;
grunt> describe tweets;
tweets: {myMap: bytearray}
grunt> dump tweets;
```

Figure 5. loading twitter data into pig

After loading json twitter data into pig using elephant bird JsonLoader , the data is loaded is shown in figure 6. Dump keyword is used in pig to display a result on terminal, when we load the data into pig we can display it using dump keyword.

```
dish, id_str=4351711216, profile_image_url=http://pbs.twimg.com/profile_images/672070143892185088/grY8I_4a_normal.jpg, is_transla
tor=false, listed_count=445}})
({filter_level=low,retweeted=false,in_reply_to_screen_name=,possibly_sensitive=false,truncated=false,lang=en,in_reply_to_status_i
d_str=,id=818758773170913280,in_reply_to_user_id_str=,timestamp_ms=1484042277983,in_reply_to_status_id=,created_at=Tue Jan 10 09:
57:57 +0000 2017,favorite_count=0,place=,coordinates=,text=RT @jose_garde: Business Intelligence (BI) &amp; Advanced #Analytics |
 Shantanu Bhanare, https://t.co/MKPKlEhYp4 #BusinessIntelligence #BI,contributors=,retweeted_status=,filter_level=low, text=Busin
ess Intelligence (BI) &amp; Advanced #Analytics | Shantanu Bhanare, https://t.co/MKPKlEhYp4 #BusinessIntelligence #BI, contributo
rs=null, geo=null, retweeted=false, in_reply_to_screen_name=null, possibly_sensitive=false, truncated=false, lang=en, entities={h
ashtags={[{text=Analytics,indices=[42,52]}],[{text=BusinessIntelligence,indices=[97),(118)]}],{[text=BI,indices=[(119),(122)
]}]}, symbols={}, urls={{display_url=klou.tt/13l1awx2qogih,expanded_url=http://klou.tt/13l1awx2qogih,indices=[73),(96)],url=htt
ps://t.co/MKPKlEhYp4]}}, user_mentions={}}, in_reply_to_status_id_str=null, is_quote_status=false, id=818649320922599425, source=
<a href="http://klout.com" rel="nofollow">Post with Klout</a>, in_reply_to_user_id_str=null, favorited=false, in_reply_to_status
id=null, retweet_count=28, in_reply_to_user_id=null, created_at=Tue Dec 27 07:34:49 +0000 2016, favorite_count=22, id_str=8136493
20922599425, place=null, user={location=Navarra, default_profile=false, statuses_count=173966, profile_background_tile=true, lang
uage=en, profile_link_color=0C0E7A, profile_banner_url=https://pbs.twimg.com/profile_banners/2284237477/1401356354, id=2284237477, fo
llowing=null, favourites_count=238709, protected=false, profile_text_color=333333, contributors_enabled=false, description=#Digit
alMarketing #Marketing #SocialMedia #Bigdata #Analytics #innovation #BusinessIntelligence #SalesManagement #B2B #Strategy #IoT #A
I #VR #AR- keys to succes, verified=false, name=Jose Javier Garde, profile_sidebar_border_color=EEEEEE, profile_background_color=
131516, created_at=Thu Jan 09 23:06:34 +0000 2014, default_profile_image=false, followers_count=45071, geo_enabled=false, profile
_image_url_https=https://pbs.twimg.com/profile_images/471395209494802432/vec86zE2_normal.png, profile_background_image_url=http:/
/abs.twimg.com/images/themes/theme14/bg.gif, profile_background_image_url_https=https://abs.twimg.com/images/themes/theme14/bg.gi
f, follow_request_sent=null, url=https://www.linkedin.com/in/josejaviergardelecumberri, utc_offset=3600, time_zone=Madrid, notifi
cations=null, friends_count=39004, profile_use_background_image=true, profile_sidebar_fill_color=EFEFEF, screen_name=jose_garde,
id_str=2284237477, profile_image_url=http://pbs.twimg.com/profile_images/471395209494802432/vec86zE2_normal.png, is_translator=fa
lse, listed_count=7001, coordinates=null},geo#,entities#{hashtags=[{[text#Analytics,indices#{(50),(60)]}],{[text#BusinessIntelli
gence,indices#{(113),(134)]}},{[text#BI,indices#{(135),(138)]}}, symbols={}, urls={{display_url#klou.tt/13l1awx2qogih,expanded
_url#http://klou.tt/13l1awx2qogih,indices#{(89),(112)},url#https://t.co/MKPKlEhYp4]}}, user_mentions={{{id#2284237477,indices#{(3)
,(14)},screen_name#jose_garde,id_str#2284237477,name#Jose Javier Garde}}},is_quote_status#false,source#<a href="http://twitter.c
om" rel="nofollow">Twitter Web Client</a>,favorited#false,retweet_count#0,in_reply_to_user_id#,id_str#818758773170913280,user#{lo
cation=null, default_profile=true, statuses_count=72966, profile_background_tile=false, lang=en, profile_link_color=1DA1F2, id=43
51711216, following=null, favourites_count=66340, protected=false, profile_text_color=333333, contributors_enabled=false, descrip
tion=null, verified=false, name=Martin Geldish, profile_sidebar_border_color=C0DEED, profile_background_color=C0DEED, created_at=
Wed Dec 02 15:04:24 +0000 2015, default_profile_image=false, followers_count=597, geo_enabled=false, profile_image_url_https=http
s://pbs.twimg.com/profile_images/672070143892185088/grY8I_4a_normal.jpg, profile_background_image_url=http://abs.twimg.com/images
/themes/theme1/bg.png, profile_background_image_url_https=https://abs.twimg.com/images/themes/theme1/bg.png, follow_request_sent=
null, url=null, utc_offset=-28800, time_zone=Pacific Time (US & Canada), notifications=null, friends_count=156, profile_use_backg
round_image=true, profile_sidebar_fill_color=DDEEF6, screen_name=martingeldish, id_str=4351711216, profile_image_url=http://pbs.t
wimg.com/profile_images/672070143892185088/grY8I_4a_normal.jpg, is_translator=false, listed_count=445}})
grunt>
```

Figure 6. Twitter data is loaded into pig

After loading data into pig we can preprocessed the data with pig function through which we can collect only twitter id and text from all the data, And after that we can divide our text string into multiple replicated words using flatten function in pig, means we have data in the form of id and collection of words.

For checking polarity we also need a sentiment dictionary on which basis we can chack the polarity of the tweets. So we can create a dictionary table which is shown in figure 7.

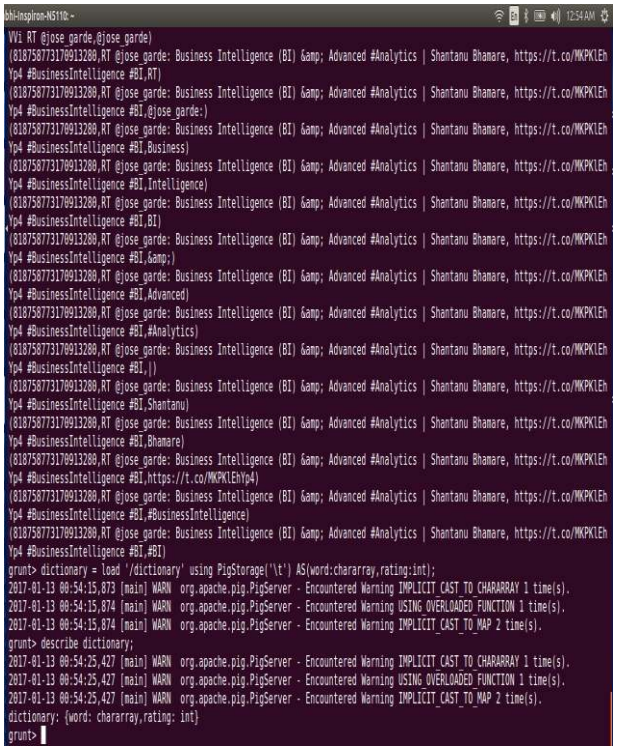


Figure 7. creating dictionary table

After creating dictionary table , we can load the dictionary words and their rating on dictionary table which is shown in figure 8.

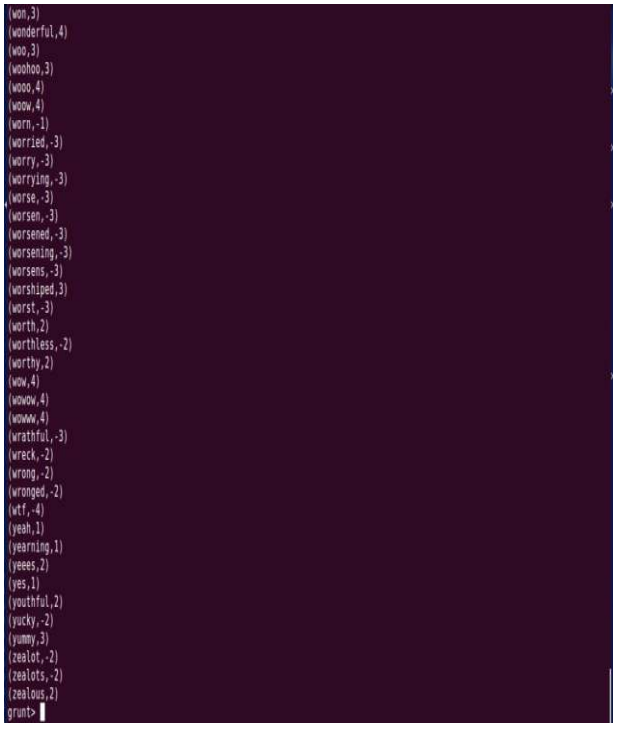


Figure 8. Dictionary words are loaded into pig

Now we have two table first we contains twitter id and collection of words and another table is dictionary which contains words along with their rating. Now we can join both the table using left outer join so we can get resultant table which contains twitter id , text and the overall rating of the text based on the sentiment dictionary. The resultant data is shown in figure 9.

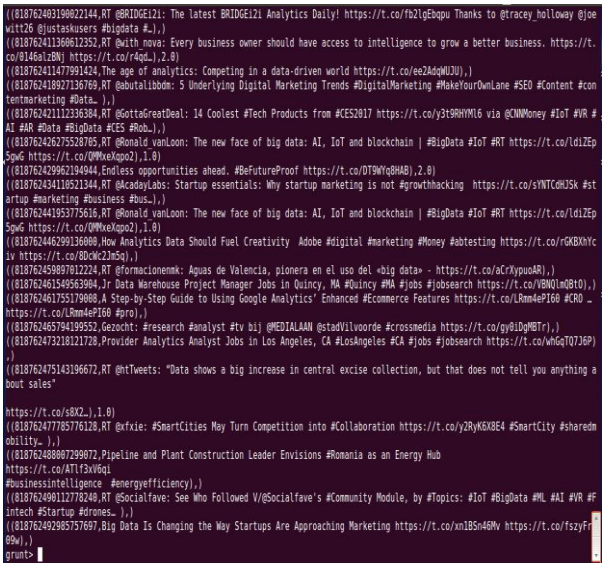


Figure 9. Polarity result of all tweets

After that we can just filter the resultant data into positive, negative or neutral polarity basis. All the positive opinions are filter and which is shown in figure 10. In this we can display only positive tweets.

Figure 10. Positive tweets with polarity rating

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, PIG . 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 pig for analysing these huge and complex data. In this we can also identify the polarity of the tweet by which we can say that which tweet have a positive meaning or a negative

meaning.Taking the system even further, we can also analyse the twitter data using some other viualizing techniques through which we can easily understand the analysis results.

9. REFERENCES

[1] Marco Furini, Manuela Montangero, "TSentiment: On Gamifying Twitter Sentiment Analysis", IEEE ISCC 2016 Workshop: DENVECT, IEEE 2016, ISSN: 978-1-5090-0679-3/16.

[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] 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.

[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] 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.

[10] 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.

[11] 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.

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

[13] 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/