Opinion Mining of Twitter Data using Hive

Pratyancha Kirar
Department of Information Technology
Samrat Ashok Technological Institute
Vidisha, India

Deepak Sain
Asst. Professor
Department of Information Technology
Samrat Ashok Technological Institute
Vidisha, India

S. K. Shrivastava
Prof. & Head
Department of Information Technology
Samrat Ashok Technological Institute
Vidisha, India

ABSTRACT
In todays extremely developed world, each minute, individuals round the globe specific themselves via numerous platforms on the net. And in every minute, an enormous quantity of unstructured information is generated. This information is within the style of text that is gathered from forums and social media websites. Such information is termed as massive information. User opinions square measure associated with a good vary of topics like politics, latest gadgets and merchandise. Social Networking sites provides tremendous impetus for large information in mining people's opinion. Public API's catered by sites like Twitter provides North American nation with helpful information for studying writer's perspective in terms of of a specific topic, product etc. To distinguish people's opinion, tweets square measure labeled into positive, negative or neutral indicators. 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 Hive. Here we proposed to develop a opinion Analysis mechanism to analyze the various polarity of opinions of Twitter users through their tweets in order to extract what they think.

Here we have used dictionary based approach for analysis for which we have implemented hive queries 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
Opinion mining, hadoop, apache flume, hive, Dictionary based approach, bigdata.

1. INTRODUCTION
Opinions are subjective expressions that outline people's, appraisals, feelings or sentiments toward entities, events and their properties. Recently there has been a massive escalation in use of Social Networking sites such as Twitter to express people's opinions. Impelled by this growth, companies, media, review groups are progressively seeking ways to mine Twitter for information about what people think and feel about a particular product or service. Twitter data [1] is a valuable source of information for marketing intelligence and trend analysis in all industries. Twitter generates gigantic data that cannot be handled manually hence the requirement of automatic categorization. Tweets are unambiguous short texts messages that are up to a maximum of 140 characters. These texts are polarized based on the nature of the comment. Focus of this paper is to provide an automated mechanism for collecting, aggregating, streaming and analyzing tweets in near real time environment and a glimpse of two of its use case scenarios.

Data from Twitter
Twitter provides us with a Streaming API which will be employed to obtain a constant stream of tweets enabling us to collect and analyze user opinion. The Streaming API works by making a request for a specific type of data which is filtered by keyword, a user, geographic area etc.[2] Once connection to the Twitter API is established via the Streaming API, data collection takes place. The tweets collected will be encoded in JavaScript Object Notation (JSON). JSON provides us with a way to encode this data. The whole tweet is regarded as a dictionary consisting of various fields. The fields may be contributors (indicates users who have authored the tweet), coordinates (Represents the geographic location of the Tweet as reported by client application), favorite_count (No. of times the tweet has been "favorited"), text (actual text of the tweet) and several other fields.

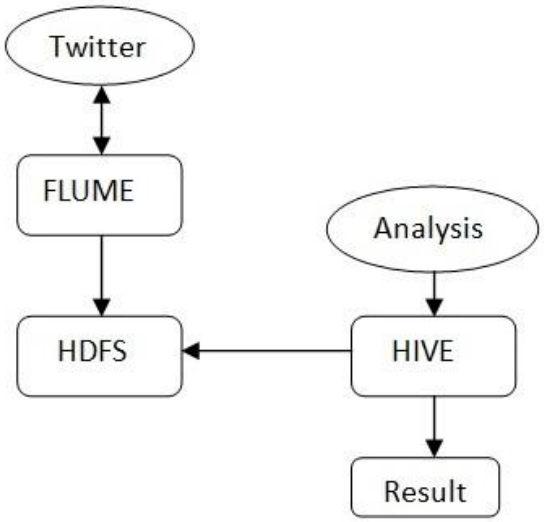

Fig 1. Workflow

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.

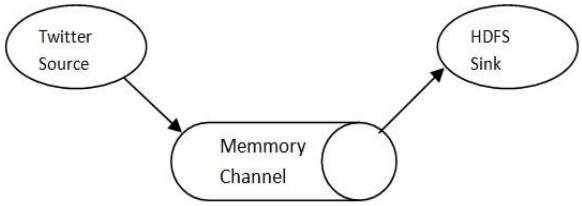

Fig 2. Flume Components

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

Hadoop is a open source framework for processing [11] and storing large datasets over a cluster.It is used in handling large and complex data which may be structured,unstructured or semi-structured that does not fit into tables. Twitter data falls into the category of "semistructured" data which can be best stored and analyzed using Hadoop and it's underlying file system.

Hadoop Distributed File System

Hadoop Distributed File System (HDFS) is a distributed file system which rests on top of the native file system and is written in java. It is highly fault tolerant and is designed for commodity hardware. HDFS has a high throughput access to application and is suitable for applications with large amount of data. The master-server architecture of HDFS having single name node helps in regulating the file system access. Requests from file system clients are handled by the data nodes. Data is stored as Input splits (blocks) on the underlying file system. The replication factor is set as 3 by default in order to maintain redundancy of data. In this case, huge amounts of tweets are collected, stored and analyzed.

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 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.

2. LITERATURE REVIEW

Mahalakshmi R, Suseela [4] (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 [5] (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 [3] (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 [6] (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 [7] (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 [8] (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 [9] (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.

3. RELATED WORK

Opinion mining is one of the most popular trends in today's world. Lot of research and Literature surveys are being done in this sector. Bo Pang and Lillian Lee are pioneers in this field. Current works in this field which uses a mathematical

approach using algorithms for opinion polarity are based on a classifier trained using a collection of annotated text data. Before training, data is preprocessed so as to extract only the main content .Some of the classification methods have been proposed are Naïve Bayes, Support Vector Machines, K-Nearest Neighbors etc. Continuous research is being done to determine most efficient method for opinion mining.

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.

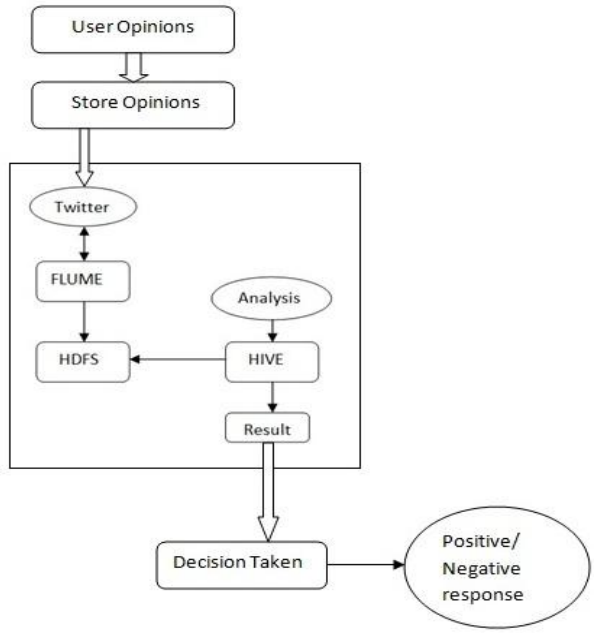

Fig 3. Proposed system workflow

Our Steps or Algorithm Steps will follow:

Step 1: users can share their opinions by posting a variety of tweets on twitter.

Step 2: all these tweets are stored in twitter database centre , their are millions of tweets are posted everyday on twitter which can generates petabytes of data which is stored on twitter data centre.

Step 3: for analysis we need these large and complex twitter data which contains variety of opinions posted by different users, we uses flume to fetch these twitter data and stored it into HDFS, we can create a twitter API through which we can fetch real time twitter data from web and stored it into HDFS.

Step 4: After storing these large and complex twitter json data we need a analysing tool to analyse these complex data, for these we uses hive which runs on top of the hadoop and takes input from HDFS and its support SQL queries through which we can analysis these data.

Step 5: Based on the analysis result from hive , we can check polarity of the tweets with the help of polarity dictionary which contains a number of English words with their polarity from -5 to +5 which indicates negative to positive and by joining these words polarity we can take a decision that which tweets are positive meaning and a negative meaning .

6. 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 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. 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.

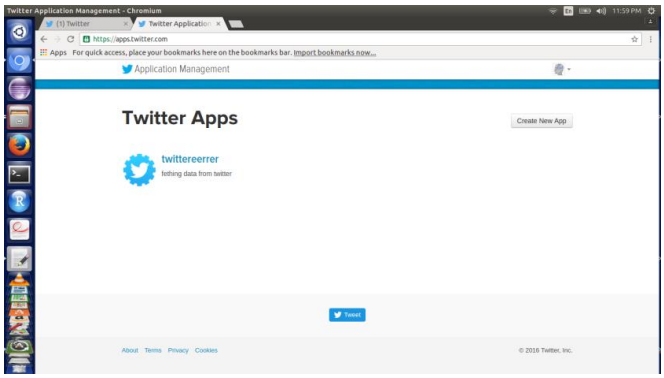

Figure 4. Creating twitter application

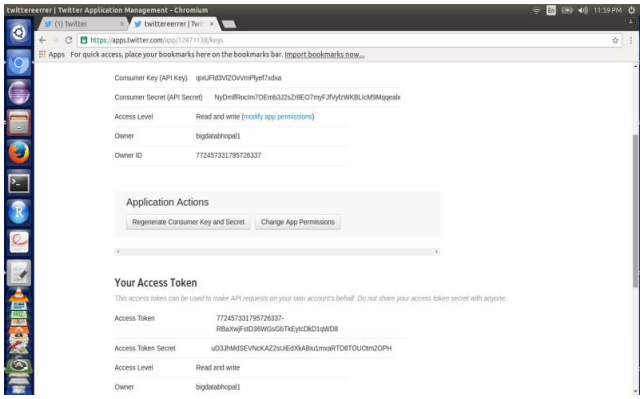

Figure 5. Generation access token keys

The figure 5 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.

```
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.

```
create external table raw(
created_at string,
id bigint,
text string,
user STRUCT<
screen_name:string,
name:string,
locations:string,
description:string,
created_at:string,
followers_count:int,
url:string>)
ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'
location '/home/abhi/work/warehouse/new_bigdata';
```

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

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

Using these jar file and custom serde we can store the unstructured data into hive table raw which has been created above in a structure manner. The figure 6 shows the structure data stored in table raw.

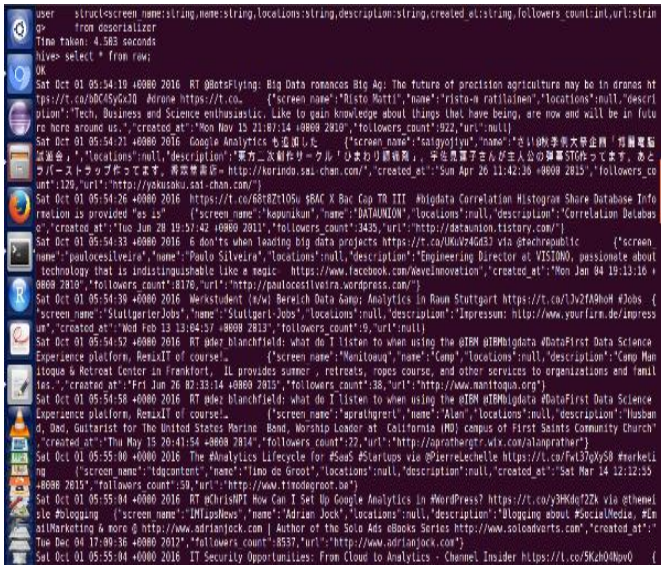

Figure 6. Data store into table raw

After that from the table raw, there are many fields and many information but we need only the twitter id and text because every text or tweet posted by user have a unique tweeter id by which we can easily identify the user which post the tweet. So we can creat another table to store only two information first is tweet id and another is text. And than we create another table as split_words which can store tweet id and text as a array of string or word, In these table we can split the text into words based on the space separator using split function. The split_words table are shown below in fig 7.

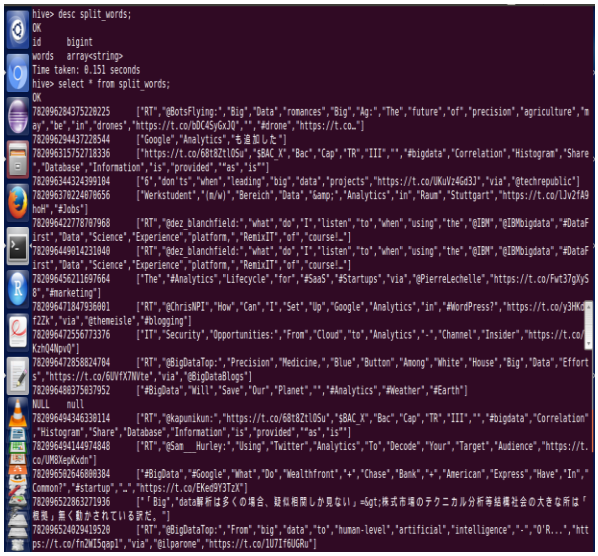

Figure 7. Data store into table splits_word table

After that we can store these splited words into another table as a new row, which means many rows are created for a single tweet id based on the number of words in that tweet text. We can use explode function in hive which can take a array of strings and generate a new row for each words so we can found a table which consist a tweet id and a word related to that tweet.

After that we can create a another table called dictionary table to store a dictionary which consist a two information first stored a English word and another field store a polarity of that tweet which is from -5 to +5. And the dictionary table are shown in fig 8.

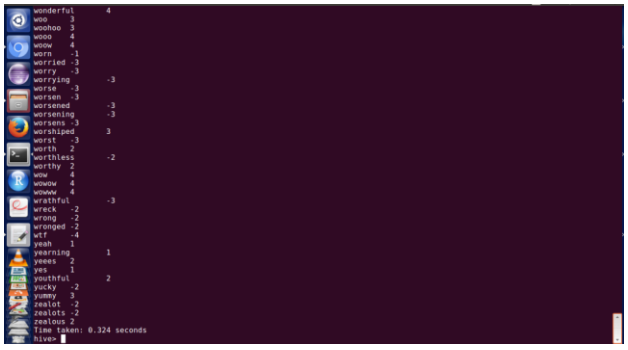

Figure 8. Data store into dictionary table

After that we have a two table , first table contains a two field which store a tweet id and the word related to that tweet and second table consist a two fields a English word and a polarity of that English word. Than we can join this two table based on the words which is common from both the table , we can perform a left outer join and the resultant table are shown in fig 9.

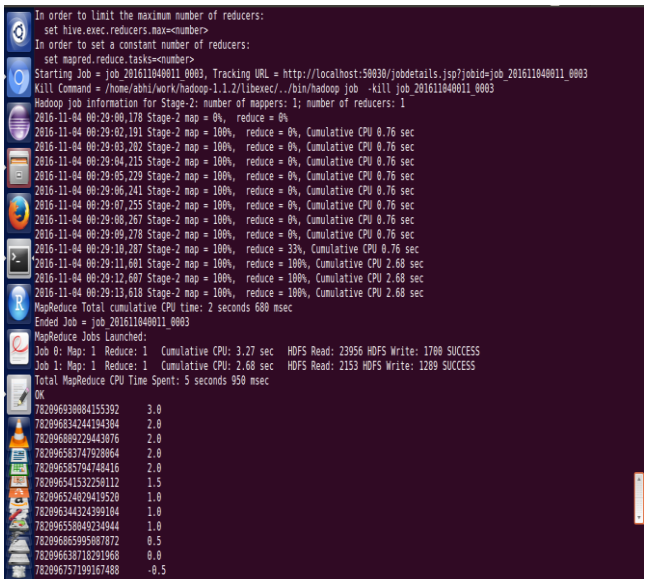

Figure 9. Data store in resultant table

The resultant table consist of a two fields which store a unique tweet id and a polarity of that tweets which comes from the average rating from the joining of split_words and dictionary tables. The resultant table shown in fig. show that unique tweet id and the polarity of that tweet on the polarity basis we can say that the tweet indicates positive meaning or a negative meaning, and with the help of unique tweet id we can easily identify the user name and other information on the basis of unique tweet id.

7. CONCLUSION

Opinion Mining may be a terribly wide branch for analysis. we've lined a number of its necessary aspects. an equivalent design may well be used for a spread of applications designed to seem at Twitter knowledge, like distinguishing spam accounts, or distinguishing clusters of keywords. 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 visualize these analysie by some other language such as R, with the help of wordcloud in R we can also visualize these analysis.

8. REFERENCES

[1] Aditya Bhardwaj, Vineet Kumar Singh,Vanraj, Yogendra Narayan, “Analyzing BigData with Hadoop Cluster in HDInsight Azure Cloud”,IEEE 2015, **978-1-4673-6540-6/15.**

[2] Aditya Bhardwaj, Vanraj, Ankit Kumar, Yogendra Narayan , Pawan Kumar, “Big Data Emerging Technologies: A CaseStudy with Analyzing Twitter Data using Apache Hive”, in IEEE 2015, **978-1-4673-8253-3/15.**

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[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.