**Sentiment Analysis on Twitter Data using Hadoop-Ecosystem and AFINN Dictionary**

A

**Major Project**

*Submitted*

*In partial fulfillment*

*For the award of the Degree of*

***Bachelor of Technology***

***In Department of******Computer Science &******Engineering***



Submitted to:                          Submitted By:

Vijay Parkash Sharma Jitender Singh Virk

Guided by: 1302041054

Chetanya Sharma

**Department of** ***Computer Science &* Engineering**

**JECRC UNIVERSITY**

Ramchandrapura, Jaipur

**CONTENTS**

**TITLE Page No.**

1. Company Profile 3
2. Abstract 4
3. Introduction 5
4. Problem Statement 6
5. Methodology 7
6. Conclusion 14
7. References 15

**Company Profile**

IBMCE-Headstart is Delhi Based software development and IT Consultancy Company, incorporated in the year 2000 and involved in Training of students & Working Professionals in the field of Information Technology, Banking & Finance, and Management. With Multiple Centres in Delhi/NCR and strong Association with NIIT since 2001, we have trained over 80000 students from last 15 years. We have won various awards for excellence over last 10 years. We have a strong and committed team of faculties on Latest Technologies.

IBM is more than 100 year old company and pioneer in computation tech domain. We feel that in a world, where there is growing demand for skilled IT professionals, the key to transforming today's students into tomorrow's working professionals is to develop industry capabilities right from foundation level. To this effect, the Career Education of IBM (under Software Group) has caused revolutionary changes in academia, rapidly building industry-relevant software capabilities that organizations need today and tomorrow.

**Abstract**

In today‘s highly developed world, every minute, people around the globe express themselves via various platforms on the Web. And in each minute, a huge amount of unstructured data is generated. This data is in the form of text which is gathered from forums and social media websites. Such data is termed as big data. User opinions are related to a wide range of topics like politics, latest gadgets and products. These opinions can be mined using various technologies and are of utmost importance to make predictions or for one-to-one consumer marketing since they directly convey the viewpoint of the masses. Here we propose to analyse the sentiments of Twitter users through their tweets in order to extract what they think. Hence we are using hadoop for sentiment analysis which will process the huge amount of data on a hadoop cluster faster.

Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. Social media monitoring tools like Brand-watch Analytics make that process quicker and easier than ever before. The applications of sentiment analysis are broad and powerful. The ability to extract insights from social data is a practice that is being widely adopted by organizations across the world. Shifts in sentiment on social media have been shown to correlate with shifts in the stock market. The Obama administration used sentiment analysis to gauge public opinion to policy announcements and campaign messages ahead of 2012 presidential election.

The ability to quickly understand consumer attitudes and react accordingly is something that Expedia Canada took advantage of when they noticed that there was a steady increase in negative feedback to the music used in one of their television adverts.

**INTRODUCTION**

From 20th century onwards this WWW has completely changed the way of expressing their views. Present situation is completely they are expressing their thoughts through online blogs, discussion forms and also some online applications like Facebook, Twitter, etc. If we take

Twitter as our example nearly 1TB of text data is generating within a week in the form of tweets. So, by this it is understand clearly how this Internet is changing the way of living and style of people. Among these tweets can be categorized by the hash value tags for which they are commenting and posting their tweets. So, now many companies and also the survey companies are using this for doing some analytics such that they can predict the success rate of their product or also they can show the different view from the data that they have collected for analysis. But, to calculate their views is very difficult in a normal way by taking these heavy data that are going to generate day by day.



Fig. 1: Describes clearly Apache Hadoop Ecosystem.

The above figure shows clearly the different types of ecosystems that are available on Hadoop so, this problem is taking now and can be solved by using BIGDATA Problem as a solution. And if we consider getting the data from Twitter one should use any one programming language to crawl the data from their database or from their web pages. Coming to this problem here we are collecting this data by using BIGDATA online streaming Eco System Tool known as Flume and also the shuffling of data and generating them into structured data in the form of tables can be done by using Apache Hive.

**PROBLEM STATEMENT**

**2.1 Existing System**

As we have already discussed about the older way of getting data and also performing the sentiment analysis on those data. Here they are going to use some coding techniques for crawling the data from the twitter where they can extract the data from the Twitter web pages by using some code that may be written either in JAVA,

Python etc. For those they are going to download the libraries that are provided by the twitter guys by using thisthey are crawling the data that we want particularly.

After getting raw data they will filter by using some old techniques and also they will find out the positive, negative and moderate words from the list of collected words in a text file. All these words should be collected by us to filter out or do some sentiment analysis on the filtered data. These words can be called as a dictionary set by which they will perform sentiment analysis. Also, after performing all these things and they want to store these in a database and coming to here they can use RDBMS[14] where they are having limitations in creating tables and also accessing the tables effectively.

**2.2Proposed System**

As it can have seen existing system drawbacks, here we are going to overcome them by solving this issue using Big Data problem statement. So here we are going to use Hadoop and its Ecosystems, for getting raw data from the Twitter we are using Hadoop online streaming tool using Apache Flume. In this tool only we are going to configure everything that we want to get data from the Twitter. For this we want to set the configuration and also want to define what information that we want to get form Twitter. All these will be saved into our HDFS (Hadoop Distributed File System) in our prescribed format. From this raw 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 form that we are going to perform the Sentiment Analysis by using some UDF’s (User Defined Functions) by which we can perform sentiment analysis by taking Stanford Core NLP[11] as the data dictionary so that by using that we can decide the list of words that coming under positive, moderate and negative.

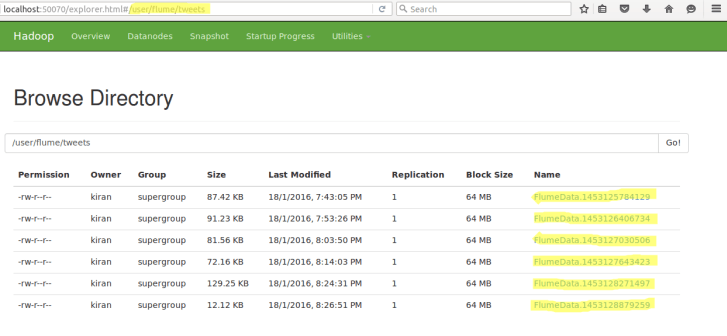
The following figure shows clearly the architecture view for the proposed system by this we can understand how our project is effective using the Hadoop ecosystems and how the data is going to store form the Flume, also how it is going to create tables using Hive also how the sentiment analysis is going to perform.



Fig. 2: Architecture diagram for proposed system.

**METHODOLOGY**

All the real-time tweets is kept it the location ‘**/user/flume/tweets**‘ HDFS. You can refer to the below screen shot for the same.



The data from Twitter is in ‘Json’ format, so a Pig JsonLoader is required to load the data into Pig.

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

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

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

After registering the required jars, we can now write a Pig script to perform Sentiment Analysis.

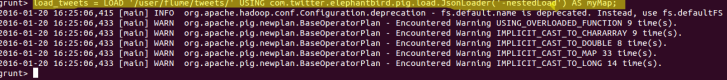
Below is a sample tweets collected for this purpose:

|  |  |
| --- | --- |
| 1 | {"filter\_level":"low","retweeted":false,"in\_reply\_to\_screen\_name":"FilmFan","truncated":false,"lang":"en","in\_reply\_to\_status\_id\_str":null,"id":689085590822891521,"in\_reply\_to\_user\_id\_str":"6048122","timestamp\_ms":"1453125782100","in\_reply\_to\_status\_id":null,"created\_at":"Mon Jan 18 14:03:02 +0000 2016","favorite\_count":0,"place":null,"coordinates":null,"text":"@filmfan hey its time for you guys follow @acadgild To #AchieveMore and participate in contest Win Rs.500 worth vouchers","contributors":null,"geo":null,"entities":{"symbols":[],"urls":[],"hashtags":[{"text":"AchieveMore","indices":[56,68]}],"user\_mentions":[{"id":6048122,"name":"Tanya","indices":[0,8],"screen\_name":"FilmFan","id\_str":"6048122"},{"id":2649945906,"name":"ACADGILD","indices":[42,51],"screen\_name":"acadgild","id\_str":"2649945906"}]},"is\_quote\_status":false,"source":"<a href=\"https://about.twitter.com/products/tweetdeck\" rel=\"nofollow\">TweetDeck<\/a>","favorited":false,"in\_reply\_to\_user\_id":6048122,"retweet\_count":0,"id\_str":"689085590822891521","user":{"location":"India ","default\_profile":false,"profile\_background\_tile":false,"statuses\_count":86548,"lang":"en","profile\_link\_color":"94D487","profile\_banner\_url":"https://pbs.twimg.com/profile\_banners/197865769/1436198000","id":197865769,"following":null,"protected":false,"favourites\_count":1002,"profile\_text\_color":"000000","verified":false,"description":"Proud Indian, Digital Marketing Consultant,Traveler, Foodie, Adventurer, Data Architect, Movie Lover, Namo Fan","contributors\_enabled":false,"profile\_sidebar\_border\_color":"000000","name":"Bahubali","profile\_background\_color":"000000","created\_at":"Sat Oct 02 17:41:02 +0000 2010","default\_profile\_image":false,"followers\_count":4467,"profile\_image\_url\_https":"https://pbs.twimg.com/profile\_images/664486535040000000/GOjDUiuK\_normal.jpg","geo\_enabled":true,"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":19800,"time\_zone":"Chennai","notifications":null,"profile\_use\_background\_image":false,"friends\_count":810,"profile\_sidebar\_fill\_color":"000000","screen\_name":"Ashok\_Uppuluri","id\_str":"197865769","profile\_image\_url":"http://pbs.twimg.com/profile\_images/664486535040000000/GOjDUiuK\_normal.jpg","listed\_count":50,"is\_translator":false}} |

The tweets are in nested Json format and consists of map data types. We need to load the tweets using JsonLoader which supports maps, so we are using **elephant bird JsonLoader**to load the tweets.

Below is the first Pig statement required to load the tweets into Pig:

|  |  |
| --- | --- |
| 1 | load\_tweets = LOAD '/user/flume/tweets/' USING com.twitter.elephantbird.pig.load.JsonLoader('-nestedLoad') AS myMap; |

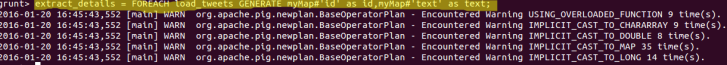


When we dump the above relation, we can see that all the tweets got loaded successfully.

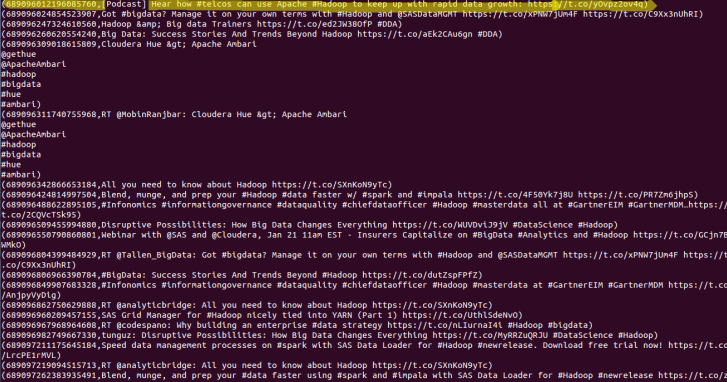


Now, we shall extract the **id**and the **tweet text** from the above tweets. The Pig statement necessary to perform this is as shown below:

|  |  |
| --- | --- |
| 1 | extract\_details = FOREACH load\_tweets GENERATE myMap#'id' as id,myMap#'text' as text; |

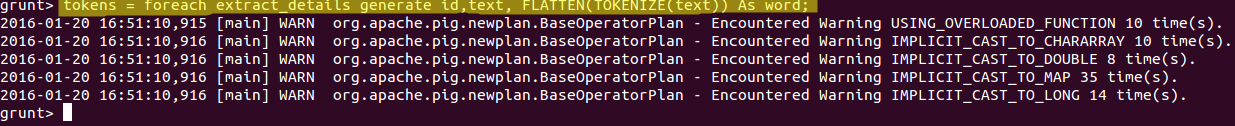


We can see the extracted **id** and **tweet text**from the tweets in the below screen shot.

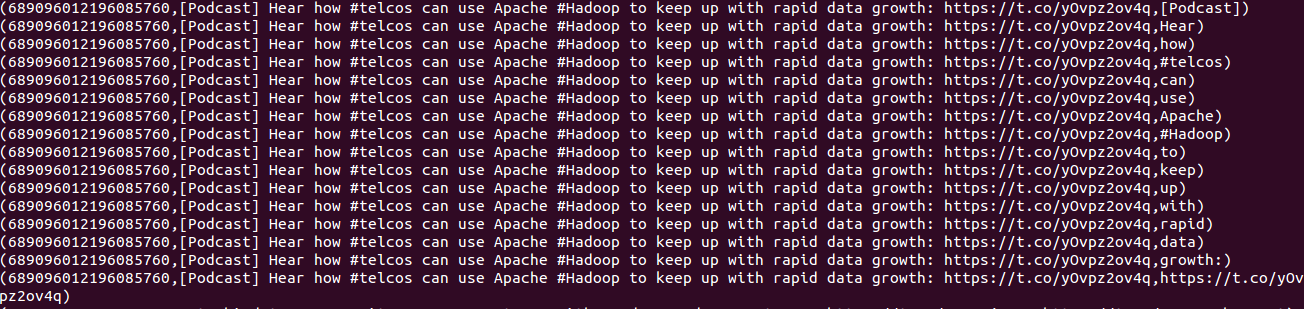


We have the tweet id and the tweet text in the relation named as **extract\_details**. Now, we shall extract the words from the text using the TOKENIZE key word in Pig.

|  |  |
| --- | --- |
| 1 | tokens = foreach extract\_details generate id,text, FLATTEN(TOKENIZE(text)) As word; |



From the below screen shot, we can see that the text got divided into words.



Now, we have to analyse the Sentiment for the tweet by using the words in the text. We will rate the word as per its meaning from +5 to -5 using the dictionary AFINN. The AFINN is a dictionary which consists of 2500 words which are rated from +5 to -5 depending on their meaning. You can download the dictionary from the following link:

[AFINN dictionary](https://drive.google.com/open?id=0ByJLBTmJojjzZ0d1RVdBTDVjT28)

We will load the dictionary into pig by using the below statement:

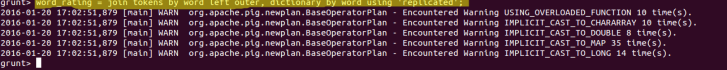
|  |  |
| --- | --- |
| 1 | dictionary = load '/AFINN.txt' using PigStorage('\t') AS(word:chararray,rating:int); |

We can see the contents of the AFINN dictionary in the below screen shot.



Now, let’s perform a map side join by joining the **tokens**statement and the dictionary contents using this command:

|  |  |
| --- | --- |
| 1 | word\_rating = join tokens by word left outer, dictionary by word using 'replicated'; |



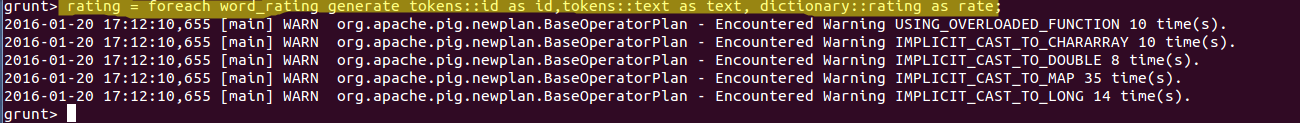
We can see the schema of the statement after performing join operation by using the below command:

|  |  |
| --- | --- |
| 1 | describe word\_rating; |
|  |  |

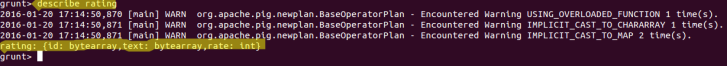
https://i0.wp.com/s3.amazonaws.com/acadgildsite/wordpress_images/bigdatadeveloper/Sentiment+Analysis+on+Twitter+Data+using+Pig/describe_word_rating.png?zoom=0.800000011920929&resize=727%2C61&ssl=1

In the above screenshot, we can see that the word\_rating has joined the **tokens(**consists of id, tweet text, word**)**statement and the **dictionary**(consists of word, rating). Now we will extract the **id,tweet text**and **word rating(**from the dictionary**)**by using the below relation:

|  |  |
| --- | --- |
| 1 | rating = foreach word\_rating generate tokens::id as id,tokens::text as text, dictionary::rating as rate; |

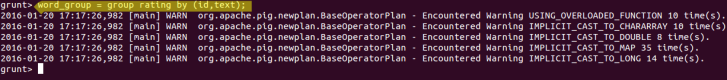


We can now see the schema of the relation **rating**by using the command describe rating.



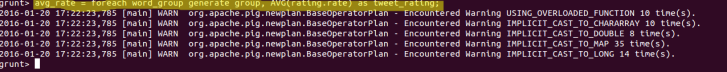
In the above screen shot we can see that our relation now consists of **id,tweet text** and **rate(**for each word**).** Now, we will group the **rating of all the words in a tweet**by using the below relation:

|  |  |
| --- | --- |
| 1 | word\_group = group rating by (id,text); |

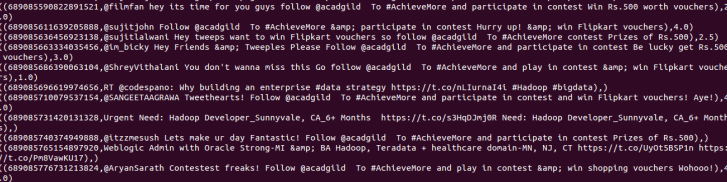


Here we have grouped by two constraints, **id** and **tweet text**. Now, let’s perform the **Average** operation on the **rating of the words per each tweet**.

|  |  |  |
| --- | --- | --- |
|  | avg\_rate = foreach word\_group generate group, AVG(rating.rate) as tweet\_rating; |  |



**Now we have calculated the Average rating of the tweet using the rating of the each word.**You can refer to the below image for the same.



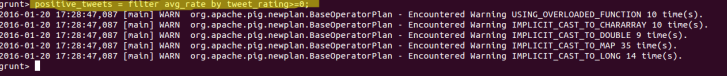
From the above relation, we will get all the tweets i.e., both positive and negative.

Here, we can classify the positive tweets by taking the rating of the tweet which can be from **0-5.**We can classify the negative tweets by taking the rating of the tweet from**-5 to -1.**

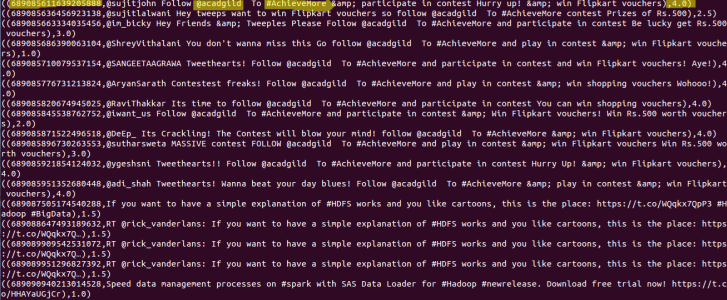
We have now successfully performed the Sentiment Analysis on Twitter data using Pig. We now have the tweets and its rating, so let’s perform an operation to filter out the positive tweets.

Now we will filter the positive tweets using the below statement:

|  |  |
| --- | --- |
| 1 | positive\_tweets = filter avg\_rate by tweet\_rating>=0; |



We can see the positive tweets and its rating in the below screen shot.



In the above screen shot we can see the tweet\_id,tweet\_text and its rating.

**CONCLUSION**

There are different ways to get Twitter data or any other online streaming data where they want to code lines of coding to achieve this. And, also they want to perform the sentiment analysis on the stored data where it makes some complex to perform those operations. Coming to this paper we have achieved by this problem statement and solving it in BIGDATA by using Hadoop and its EcoSystems. And finally we have done sentiment analysis on the Twitter data that is stored in HDFS. So, here the processing time taken is also very less compared to the previous methods because Hadoop MapReduce and Hive are the best methods to process large amount of data in a small time.

**REFERENCES**

[1] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1-12.

[2] Tang, H., Tan, S., Cheng, X., A survey on sentiment detection of reviews, Expert Systems with Applications: An International Journal, v.36 n.7, p.10760-10773, September, 2009.

[3] A. Pak and P. Parouek, “Twitter as a corpus for sentiment analysis and opinion mining,” in Proceedings of LREC, vol. 2010.

[4] J. Dean and S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters,” Communications of the ACM, Vol. 51, Iss. 1, pp. 107-113, January 2008.

[5] S. Ghemawat, H. Gobioff and S-T. Leung, “The Google File

System,” ACM SIGOPS Operating System Review, Vol. 37, Iss. 5, pp. 29-43, December 2003.

[6] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, “The Hadoop Distributed File System,” in the 26th IEEE Symposium on Mass Storage Systems and Technologies, pp. 1-10, May 2010.

[7] Bahrainian, S.A., Dengel, A., Sentiment Analysis using Sentiment Features, In the proceedings of WPRSM Workshop and the Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, Atlanta, USA, 2013.

[8] “Sentimental Analysis”, Inc. [Online]. Available: <http://www.cs.uic.edu/~liub/FBS/sentiment-analysis> [Accessed 23 March 2013].

[9] (Online Resource) Hive (Available on:http://hive.apache.org/).

[10] (Online Resource)http://jsonlint.com/

[11] (Online Resource)http://nlp.stanford.edu/software/corenlp.shtml.

[12] T. White, "The Hadoop Distributed Filesystem," Hadoop: The Definitive Guide, pp. 41 73, GravensteinHighwaNorth, Sebastopol: O’Reilly Media, Inc., 2010.

[13] (Online Resource) http://flume.apache.org/

[14] S. W. Ambler. Relational databases 101: Looking at the whole picture.www.AgileData.org, 2009.