

# **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***



**JECRC**<sup>TM</sup>  
**UNIVERSITY**  
BUILD YOUR WORLD

Submitted to:

Vijay Parkash Sharma

Guided by:

Chetanya Sharma(IBMCE-Headstar)

Submitted By:

Jitender Singh Virk

1302041054

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

**JECRC UNIVERSITY**

Ramchandrapura, Jaipur

## CONTENTS

<b>TITLE</b>	<b>Page No.</b>
1. Company Profile	3
2. Abstract	4
3. Introduction	5
4. Problem Statement	6
5. Methodology	7
6. Conclusion	14
7. Learnings	15
8. References	16

## **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.

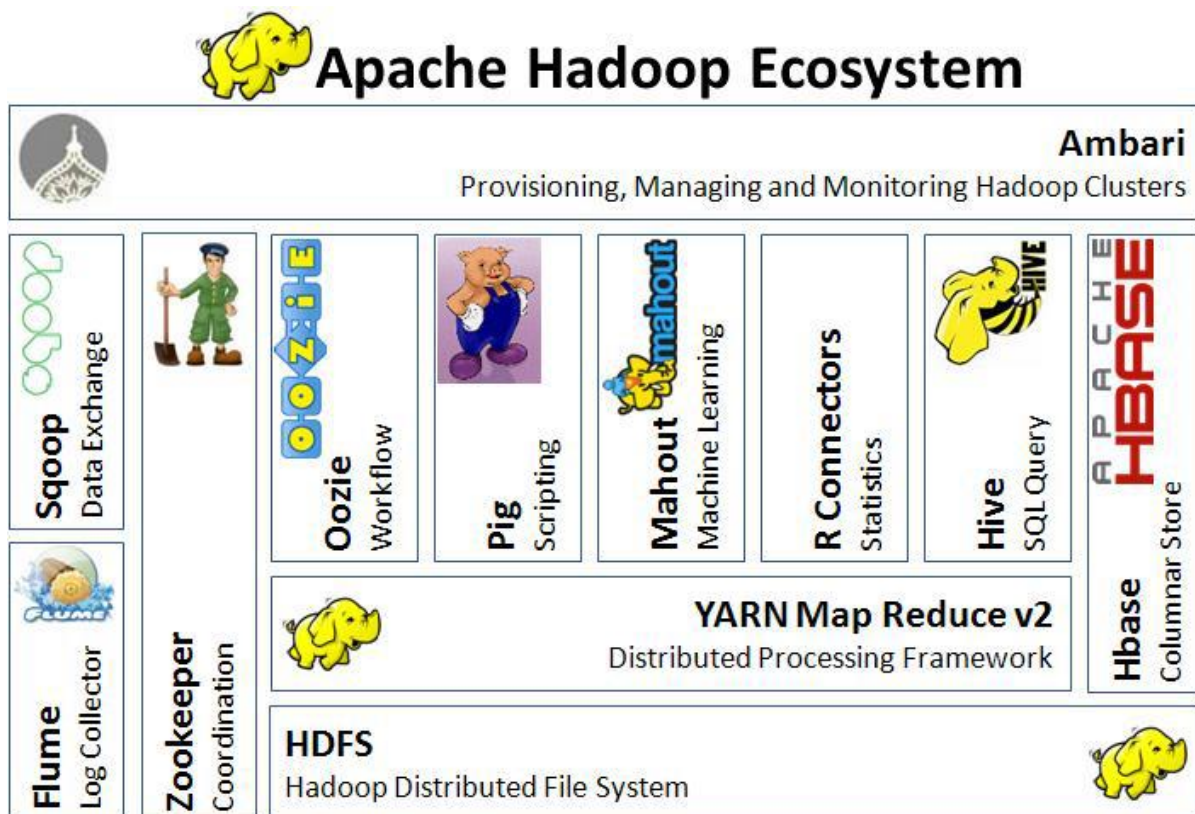


Fig. 1: Describes clearly Apache Hadoop Ecosystem.

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 this they 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.

### 2.2 Proposed 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.

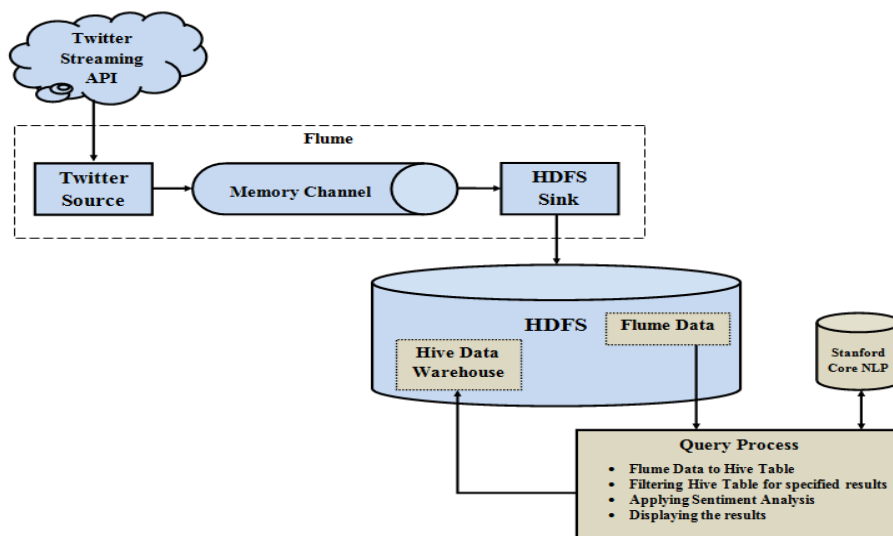
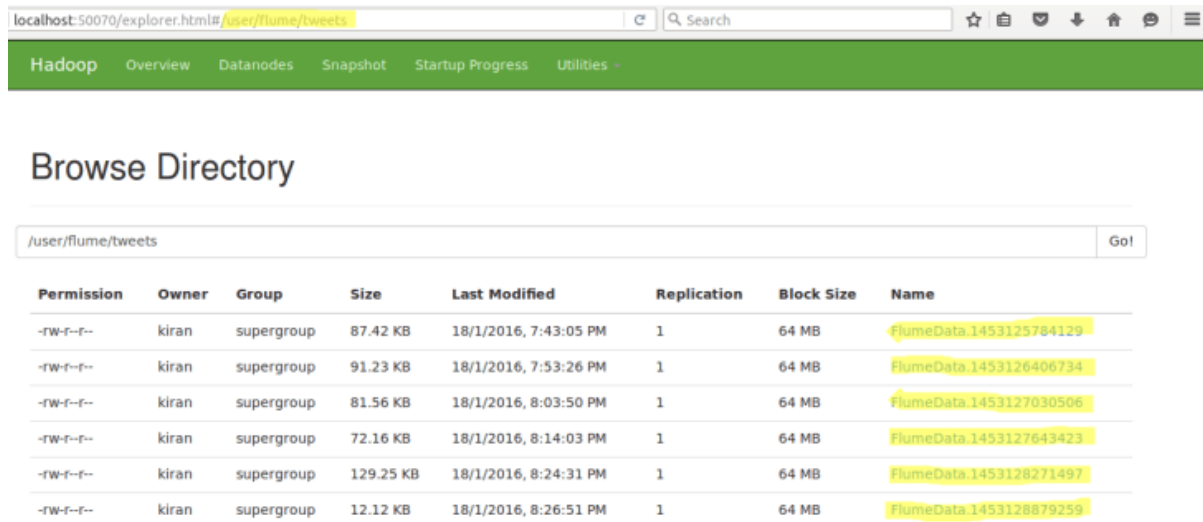


Fig. 2: Architecture diagram for proposed system.

## METHODOLOGY

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



Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	kiran	supergroup	87.42 KB	18/1/2016, 7:43:05 PM	1	64 MB	flumeData.1453125784129
-rw-r--r--	kiran	supergroup	91.23 KB	18/1/2016, 7:53:26 PM	1	64 MB	flumeData.1453126406734
-rw-r--r--	kiran	supergroup	81.56 KB	18/1/2016, 8:03:50 PM	1	64 MB	flumeData.1453127030506
-rw-r--r--	kiran	supergroup	72.16 KB	18/1/2016, 8:14:03 PM	1	64 MB	flumeData.1453127643423
-rw-r--r--	kiran	supergroup	129.25 KB	18/1/2016, 8:24:31 PM	1	64 MB	flumeData.1453128271497
-rw-r--r--	kiran	supergroup	12.12 KB	18/1/2016, 8:26:51 PM	1	64 MB	flumeData.1453128879259

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:

```
{ "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, "p
```



```

"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:

```

load_tweets = LOAD '/user/flume/tweets/' USING com.twitter.elephantbird.pig.load.JsonLoader('nestedLoad') AS myMap;

```

```

prunt> load_tweets = LOAD '/user/flume/tweets/' USING com.twitter.elephantbird.pig.load.JsonLoader('nestedLoad') AS myMap;
2016-01-20 16:25:06,415 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2016-01-20 16:25:06,433 [main] WARN org.apache.plg.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 9 time(s).
2016-01-20 16:25:06,433 [main] WARN org.apache.plg.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 9 time(s).
2016-01-20 16:25:06,433 [main] WARN org.apache.plg.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 16:25:06,433 [main] WARN org.apache.plg.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 33 time(s).
2016-01-20 16:25:06,433 [main] WARN org.apache.plg.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
prunt>

```

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

```

{[filter_level#low,text#Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/yOvpz2ov4q,contributors
#,geo#,retweeted#false,in_reply_to_screen_name#,possibly_sensitive#false,truncated#false,lang#en,entities#{hashtags:[{text#telcos,indices#(19),
(26)}]},([text#Hadoop,indices#(42),(49)])),symbols={},urls=[{display_url#tbn.co/1KhvqXJ,expanded_url#http://tbn.co/1KhvqXJ,indices#(85),(16
3)],url_https#https://t.co/yOvpz2ov4q}},user_mentions={},in_reply_to_status_id_str#,is_quote_status#false,id#689096012196085760,source#a href="http
://twitter.com/download/android" rel="nofollow">Twitter for Android/a>,in_reply_to_user_id_str#,favorited#false,timestamp_ms#1453128266749,in_re
ply_to_status_id#,retweet_count#,in_reply_to_user_id#,created_at#Mon Jan 18 14:44:26 +0000 2016,favorite_count#,id_str#689096012196085760,place
#,user#(location=Commonwealth of Massachusetts, default_profile=false, statuses_count=22991, profile_background_image=true, lang=en, profile_link
_color=190069, profile_banner_url=https://pbs.twimg.com/profile_banners/237413764/1449519792, id=237413764, following=null, favourites_count=7992,
protected=false, profile_text_color=000000, contributors_enabled=false, descriptions=Content Marketing @IBAnalytics. Father to the #AdventureMen
Cheshire YMCA Board Member. Volunteer @CampTakodah. Chaplain of Trinity Lodge AF&AM. Loud speaker., verified=false, name=J. Graeme Noseworthy, fol
lowers_count=3153, geo_enabled=true, profile_image_url_https=https://pbs.twimg.com/profile_images/686880402871562242/Lxn73Ql1_normal.jpg, profile
_image_url_https=https://pbs.twimg.com/profile_images/686880402871562242/Lxn73Ql1_normal.jpg, profile_background_image_url_https=https://pbs.
twimg.com/profile_images/674975457109008384/EXfxcv38.jpg, profile_background_image_url_https=https://pbs.twimg.com/profile_images/674975457109008384/EXfxcv38.jpg, follow_request_sent=null, url=http://linkd.in/bDH7p1, utc_offset=-18000,
time_zone=Eastern Time (US & Canada), notifications=null, friends_count=2542, profile_use_background_image=true, profile_sidebar_fill_color=000000,
screen_name=graemeknows, id_str=237413764, profile_image_url=http://pbs.twimg.com/profile_images/686880402871562242/Lxn73Ql1_normal.jpg, is_t
ranslator=false, listed_count=404,coordinates#)}
{[filter_level#low,retweeted#false,in_reply_to_screen_name#,possibly_sensitive#false,truncated#false,lang#en,in_reply_to_status_id_str#,id#689096
024854523907,extended_entities#{media:[{id#689096024632225792,sizes#{small:{w=340,h=167,resize=fit},thumb#{w=150,h=150,resize=crop},medium
#{w=600,h=295,resize=fit},large#{w=640,h=315,resize=fit}},media_url_https#https://pbs.twimg.com/media/CZApLVrWCAAIexa.png,media_url#http://p
bs.twimg.com/media/CZApLVrWCAAIexa.png,expanded_url#http://twitter.com/Tallen_BigData/status/689096024854523907/photo/1,indices#(96),(119)},id_s
tr#689096024632225792,display_url#pic.twitter.com/C9Xx3nUHR1,type#photo,url_https#https://t.co/C9Xx3nUHR1}},in_reply_to_user_id_str#,timestamp_ms#145
3128269767,in_reply_to_status_id#,created_at#Mon Jan 18 14:44:29 +0000 2016,favorite_count#,place#,coordinates#,text#Got #bigdata! Manage it on
your own terms with #Hadoop and @SASDataMGMT https://t.co/xPNW7jUn4F,contributors#,geo#,entities#{hashtags:[{text#bigdat
a,indices#(4),(12)}]},([text#Hadoop,indices#(47),(54)])),symbols={},media:[{id#689096024632225792,sizes#{small:{w=340,h=167,resize=fit},
thumb#{w=150,h=150,resize=crop},medium#{w=600,h=295,resize=fit},large#{w=640,h=315,resize=fit}},media_url_https#https://pbs.twimg.com/me
dia/CZApLVrWCAAIexa.png,media_url#http://pbs.twimg.com/media/CZApLVrWCAAIexa.png,expanded_url#http://twitter.com/Tallen_BigData/status/68909602485
4523907/photo/1,indices#(96),(119)},id_str#689096024632225792,display_url#pic.twitter.com/C9Xx3nUHR1,type#photo,url_https#https://t.co/C9Xx3nUHR1}},
urls=[{display_url#bit.ly/1nekW08,expanded_url#http://bit.ly/1nekW08,indices#(72),(95)},url_https#https://t.co/xPNW7jUn4F}},user_mentions=[{id#260
93930,indices#(59),(71)},screen_name#SASDataMGMT,id_str#26093930,name#SAS Data Management}}],is_quote_status#false,source#a href="http://logit
.vocestorm.com" rel="nofollow">VoiceStorm/a>,favorited#false,retweet_count#,in_reply_to_user_id_str#689096024854523907,user#(location=null,
default_profile=true, statuses_count=407, profile_background_image=false, lang=en, profile_link_color=008484, profile_banner_url=https://pbs.twi
ng.com/profile_banners/3875179092/1425667328, id=3875179092, following=null, favourites_count=1, protected=false, profile_text_color=333333, cont
ributors_enabled=false, description=null, verified=false, name=Taylor Allen, profile_sidebar_border_color=C0E0E0, profile_background_color=C0E0E0,
created_at#Fri Mar 06 16:11:55 +0000 2015, default_profile_image=false, followers_count=19, geo_enabled=false, profile_image_url_https=https://p
bs.twimg.com/profile_images/573879073904029696/h8c_3mg2_normal.jpeg, 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=-18000,

```



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;
```

```
grunt> extract_details = FOREACH load_tweets GENERATE myMap#'id' as id,myMap#'text' as text;
2016-01-20 16:45:43,552 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 9 time(s).
2016-01-20 16:45:43,552 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 9 time(s).
2016-01-20 16:45:43,552 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 16:45:43,552 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 16:45:43,552 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
```

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

```
(689096012196085760, [Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vzp20v4g)
(689096024854523907, Got #bigdata? Manage it on your own terms with #Hadoop and @SASDataMGMT https://t.co/xPNW7jUm4F https://t.co/C9Xx3nUhrI)
(689096247324610560, Hadoop & Big data Trainers https://t.co/ed2JN380YP #DDA)
(689096260620554240, Big Data: Success Stories And Trends Beyond Hadoop https://t.co/aEk2CAu6gn #DDA)
(689096309018615809, Cloudera Hue & Apache Ambari)
@gethue
@ApacheAmbari
#hadoop
#bigdata
#hue
#ambari)
(689096311740755968, RT @MobinRanjbar: Cloudera Hue & Apache Ambari)
@gethue
@ApacheAmbari
#hadoop
#bigdata
#hue
#ambari)
(68909634286653184, All you need to know about Hadoop https://t.co/5XnKoN9yTc)
(689096424814997504, Blend, nunge, and prep your #Hadoop #data faster w/ #spark and #Impala https://t.co/4F50Yk7j8U https://t.co/PR7Zm6jhp5)
(689096488622895105, Infonomics #informationgovernance #dataquality #chiefdataofficer #Hadoop #masterdata all at #GartnerEIM #GartnerMDM https://t.co/2CQVcT5k95)
(689096509455994880, Disruptive Possibilities: How Big Data Changes Everything https://t.co/WUVDviJ9jV #DataScience #Hadoop)
(689096550790860801, Webinar with @SAS and @Cloudera, Jan 21 11am EST - Insurers Capitalize on #BigData #Analytics and #Hadoop https://t.co/GCjn7EwMko)
(689096804399484929, RT @Tallen_BigData: Got #bigdata? Manage it on your own terms with #Hadoop and @SASDataMGMT https://t.co/xPNW7jUm4F https://t.co/C9Xx3nUhrI)
(689096806966390784, #BigData: Success Stories And Trends Beyond #Hadoop https://t.co/dutZspFPfZ)
(689096849907683328, #Infonomics #informationgovernance #dataquality #chiefdataofficer #Hadoop #masterdata at #GartnerEIM #GartnerMDM https://t.co/AnjpyVvDlq)
(689096862750629888, RT @analyticbridge: All you need to know about Hadoop https://t.co/5XnKoN9yTc)
(689096862809457155, SAS Grid Manager for #Hadoop nicely tied into YARN (Part 1) https://t.co/UthLSdeNv0)
(689096867968064608, RT @codespano: Why building an enterprise #data strategy https://t.co/nLiurnaI4I #Hadoop #bigdata)
(689096882749667330, tunguz: Disruptive Possibilities: How Big Data Changes Everything https://t.co/MyRRZuQRJU #DataScience #Hadoop)
(689097211175645184, Speed data management processes on #spark with SAS Data Loader for #Hadoop #newrelease. Download free trial now! https://t.co/LrcPEiRMVL)
(689097219094515713, RT @analyticbridge: All you need to know about Hadoop https://t.co/5XnKoN9yTc)
(689097262383935491, Blend, nunge, and prep your #data faster using #spark and #Impala with SAS Data Loader for #Hadoop #newrelease https://t.co/5XnKoN9yTc)
```

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;
```

```
grunt> tokens = foreach extract_details generate id,text, FLATTEN(TOKENIZE(text)) As word;
2016-01-20 16:51:10,915 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 10 time(s).
2016-01-20 16:51:10,916 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 10 time(s).
2016-01-20 16:51:10,916 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 16:51:10,916 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 16:51:10,916 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
grunt>
```

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

```
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,[Podcast])
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,Hear)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,how)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,#telcos)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,can)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,use)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,Apache)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,#Hadoop)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,to)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,keep)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,up)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,with)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,rapid)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,data)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,growth:)
(689096012196085760,[Podcast] Hear how #telcos can use Apache #Hadoop to keep up with rapid data growth: https://t.co/y0vpz2ov4q,https://t.co/y0vpz2ov4q)
```

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](#)

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.

```
(tricked,-2)
(trickery,-2)
(triumph,4)
(triumphant,4)
(trouble,-2)
(troubled,-2)
(troubles,-2)
(true,2)
(trust,1)
(trusted,2)
(tumor,-2)
(twat,-5)
(ugly,-3)
(unacceptable,-2)
(unappreciated,-2)
(unapproved,-2)
(unaware,-2)
(unbelievable,-1)
(unbelieving,-1)
(unbiased,2)
(uncertain,-1)
(unclear,-1)
(uncomfortable,-2)
(unconcerned,-2)
(unconfirmed,-1)
(unconvinced,-1)
(uncredited,-1)
(undecided,-1)
(underestimate,-1)
(underestimated,-1)
(underestimates,-1)
(underestimating,-1)
(undermine,-2)
(undermined,-2)
(undermines,-2)
(undermining,-2)
(undeserving,-2)
(undesirable,-2)
(uneasy,-2)
(unemployment,-2)
(unequal,-1)
```

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';
```

```

grunt> word_rating = join tokens by word left outer, dictionary by word using replicated ;
2016-01-20 17:02:51,879 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 10 time(s).
2016-01-20 17:02:51,879 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 10 time(s).
2016-01-20 17:02:51,879 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 17:02:51,879 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 17:02:51,879 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
grunt>

```

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

1 describe word\_rating;

```

grunt> describe word_rating;
2016-01-20 17:06:11,823 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 1 time(s).
2016-01-20 17:06:11,823 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2016-01-20 17:06:11,823 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 2 time(s).
word_rating: {tokens::id: bytearray,tokens::text: bytearray,tokens::word: chararray,dictionary::word: chararray,dictionary::rating: int}
grunt>

```

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:

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

```

grunt> rating = foreach word_rating generate tokens::id as id,tokens::text as text, dictionary::rating as rate;
2016-01-20 17:12:10,655 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 10 time(s).
2016-01-20 17:12:10,655 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 10 time(s).
2016-01-20 17:12:10,655 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 17:12:10,655 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 17:12:10,655 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
grunt>

```

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

```

grunt> describe rating
2016-01-20 17:14:50,870 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 1 time(s).
2016-01-20 17:14:50,871 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2016-01-20 17:14:50,871 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 2 time(s).
rating: {id: bytearray,text: bytearray,rate: int}
grunt>

```

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);

```

grunt> word_group = group rating by (id,text);
2016-01-20 17:17:26,982 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 10 time(s).
2016-01-20 17:17:26,982 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 10 time(s).
2016-01-20 17:17:26,982 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 17:17:26,982 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 17:17:26,982 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
grunt>

```

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;  
up, AVG(rating.rate) as tweet\_rating;

```
grunt> avg_rate = foreach word_group generate group, AVG(rating.rate) as tweet_rating;
2016-01-20 17:22:23,785 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 10 time(s).
2016-01-20 17:22:23,785 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 10 time(s).
2016-01-20 17:22:23,785 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 8 time(s).
2016-01-20 17:22:23,785 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 17:22:23,785 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
grunt>
```

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.

```
((689085590822891521,@filmfan hey its time for you guys follow @acagdild To #AchieveMore and participate in contest Win Rs.500 worth vouchers),4.0)
((689085611639205888,@sujitjohn Follow @acagdild To #AchieveMore & participate in contest Hurry up! & win Flipkart vouchers),4.0)
((689085636456923138,@sujitlalwani Hey tweeps want to win Flipkart vouchers so follow @acagdild To #AchieveMore contest Prizes of Rs.500),2.5)
((689085663334035456,@in_bicky Hey Friends & Tweeps Please Follow @acagdild To #AchieveMore and participate in contest Be lucky get Rs.500 vouchers),3.0)
((689085686390863104,@ShreyVithalan! You don't wanna miss this Go follow @acagdild To #AchieveMore and play in contest & win Flipkart vouchers),1.0)
((689085696619974656,RT @codespano: Why building an enterprise #data strategy https://t.co/nLiurnaI4l #Hadoop #bigdata),)
((689085710079537154,@SANGEETAAGRAWA Tweethearts! Follow @acagdild To #AchieveMore and participate in contest and win Flipkart vouchers! Aye!),4.0)
((689085731420131328,Urgent Need: Hadoop Developer_Sunnyvale, CA_6+ Months https://t.co/s3Hq0Jmj0R Need: Hadoop Developer_Sunnyvale, CA_6+ Months),)
((689085740374949888,@itzzmesush Lets make ur day Fantastic! Follow @acagdild To #AchieveMore and participate in contest Prizes of Rs.500),)
((689085765154897920,Weblogic Admin with Oracle Strong-MI & BA Hadoop, Teradata + healthcare domain-MN, NJ, CT https://t.co/Uy0t5B5P in https://t.co/Pm8VawKU17),)
((689085776731213824,@AryanSarath Contest freaks! Follow @acagdild To #AchieveMore and play in contest & win shopping vouchers Wohooo!),4.0)
```

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;

```
grunt> positive_tweets = filter avg_rate by tweet_rating>=0;
2016-01-20 17:28:47,087 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning USING_OVERLOADED_FUNCTION 10 time(s).
2016-01-20 17:28:47,087 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 10 time(s).
2016-01-20 17:28:47,087 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 9 time(s).
2016-01-20 17:28:47,087 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_MAP 35 time(s).
2016-01-20 17:28:47,087 [main] WARN org.apache.pig.newplan.BaseOperatorPlan - Encountered Warning IMPLICIT_CAST_TO_LONG 14 time(s).
grunt>
```

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

```

((689085611639205880,@sujitjohn Follow @acadgild To #AchieveMore & participate in contest Hurry up! & win Flipkart vouchers),4.0)
((689085636456923138,@sujitlalwani Hey tweeps want to win Flipkart vouchers so follow @acadgild To #AchieveMore contest Prizes of Rs.500),2.5)
((689085663334035456,@ln_bicky Hey Friends & Tweeples Please Follow @acadgild To #AchieveMore and participate in contest Be lucky get Rs.500 vouchers),3.0)
((689085686390063104,@ShreyVithalani You don't wanna miss this Go follow @acadgild To #AchieveMore and play in contest & win Flipkart vouchers),1.0)
((689085710079537154,@SANGEETAAGRAWA Tweethearts! Follow @acadgild To #AchieveMore and participate in contest and win Flipkart vouchers! Aye!),4.0)
((689085776731213824,@AryanSarath Contestest freaks! Follow @acadgild To #AchieveMore and play in contest & win shopping vouchers Wohooo!),4.0)
((689085820674945025,@RaviThakkar Its time to follow @acadgild To #AchieveMore and participate in contest You can win shopping vouchers),4.0)
((689085845538762752,@Iwant_us Follow @acadgild To #AchieveMore and participate in contest & Win Flipkart vouchers! Win Rs.500 worth vouchers),2.0)
((689085871522496518,@DeEp_ Its Crackling! The Contest will blow your mind! follow @acadgild To #AchieveMore & win Flipkart vouchers),4.0)
((689085896730263553,@sutharsweta MASSIVE contest FOLLOW @acadgild To #AchieveMore and play in contest & win Flipkart vouchers Win Rs.500 worth vouchers),3.0)
((689085921854124032,@ygeshsni Tweethearts!! Follow @acadgild To #AchieveMore and participate in contest Hurry Up! & win Flipkart vouchers),4.0)
((689085951352680448,@adi_shah Tweethearts! Wanna beat your day blues! Follow @acadgild To #AchieveMore & play in contest & win Flipkart vouchers),4.0)
((689087505174540288,If you want to have a simple explanation of #HDFS works and you like cartoons, this is the place: https://t.co/WQqkx7QpP3 #Hadoop #BigData),1.5)
((689088647493189632,RT @rick_vanderlans: If you want to have a simple explanation of #HDFS works and you like cartoons, this is the place: https://t.co/WQqkx7Q...),1.5)
((689089909542531072,RT @rick_vanderlans: If you want to have a simple explanation of #HDFS works and you like cartoons, this is the place: https://t.co/WQqkx7Q...),1.5)
((689089951290827392,RT @rick_vanderlans: If you want to have a simple explanation of #HDFS works and you like cartoons, this is the place: https://t.co/WQqkx7Q...),1.5)
((689090940213014528,Speed data management processes on #spark with SAS Data Loader for #Hadoop #newrelease. Download free trial now! https://t.co/HHAYaUGjCr),1.0)

```

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.



## **LEARNINGS**

1. Big Data and its evolution
2. 5 important V's of Big Data
3. Hadoop Distributed File System
4. Hadoop Clusters
5. Hadoop MapReduce
6. DataNode
7. NameNode
8. Hive
9. Flume
10. Pig
11. Sqoop
12. Linux

## 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](http://www.AgileData.org), 2009.