ISSN 2321-807X
Volume 12 Number20
Journal of Advances in chemistry

Prediction of sales using Big data analytics

I.Karthika, P.Gokulraj, S.Saravanan
Assistant Professor, Department of Computer Science and Engineering. M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India
(e-mail:karthikai.cse@mkce.ac.in).
Assistant Professor, Department of Computer Science and Engineering.M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India
(e-mail:gokulrajp.cse@mkce.ac.in).
Assistant Professor, Department of Computer Science and Engineering.M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India
(e-mail: saravanans.cse@mkce.ac.in).

ABSTRACT:

Social media is a main source of collecting big-data. Data analysis converting their bigger data to smart data. Smart data is acquired with the help of Apache Flume, Apache hive and Apache HDFS, smart data increase the sales of Marketing industry. It helps product owner to analyze people's opinion about their product and consumer can analyze the reviews of product before purchase. If tweets came along with Location, data analyzed based on the location.

Keywords : FLUME , HIVE , HDFS, Smart data

I.INTRODUCTION

Marketing industry increase their sales turnover with the help big data analysis. They concentrating on the source of social media to posts to know about product review to increase the sales. Among all social media, twitter is one of the most famous one. Raw data is extracted from twitter using an Apache FLUME.

once the raw data is extracted it store in the Apache HDFS. On data analysis raw data is converted into smart data. Data analysis can be done using Apache hive. Text tweets are filtered with the product name. Location also analyzed, if tweet is given with location.

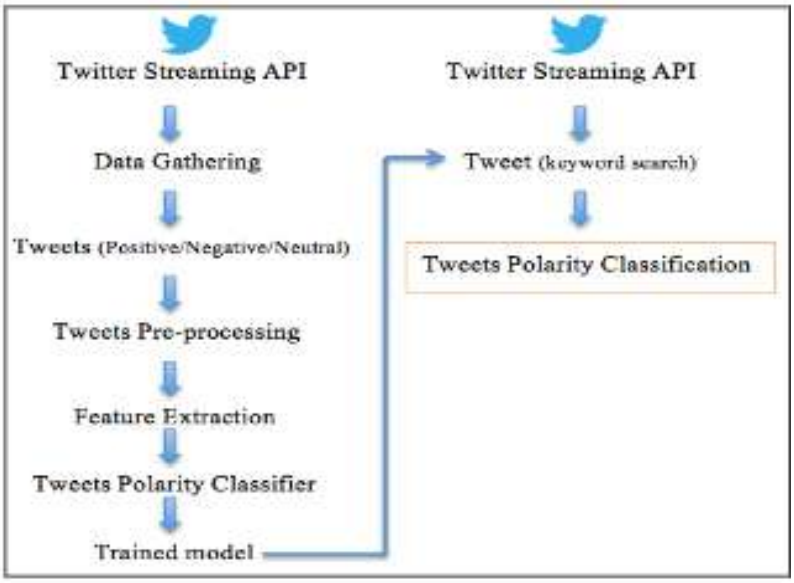

Figure 1 : Twitter Analysis Steps

II.DATA EXTRACTION

Raw data is extracted with help of twitter streaming API. Configuring an flume agent which has a source and sink. Twitter posts will acts as an source, which is an twitter source, Apache HDFS will acts an sink. Apache Flume extracted raw data from twitter via memory channel put the tweets in HDFS.

Apache Flume:

Apache Flume configuring according to the product search. Twitter source which is configured to an twitter. Twitter data is carried over the channel called memchannel. Storage area of an Twitter is sink ,configured to an HDFS. Along with configuration can made to an keyword to be searched ,HDFS path, Write format, memory capacity and Transaction capacity.

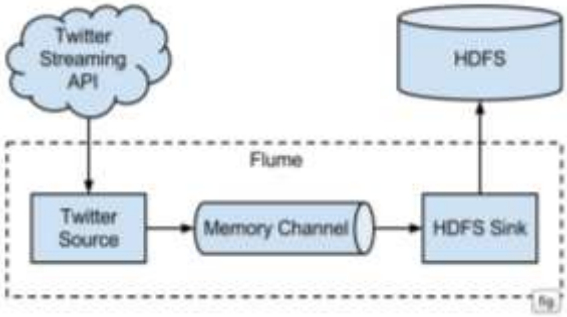

Figure 2 : Workflow of Flume

```
agent1.sources.source_read.type =
    com.cloudera.flume.source.TwitterSource
agent1.sources.source_read.channels = MemChannel
agent1.sources.source_read.consumerKey =
agent1.sources.source_read.consumerSecret =
agent1.sources.source_read.accessToken =
agent1.sources.source_read.accessTokenSecret =
agent1.sources.source_read.keywords = hadoop
```

Figure 3 : Flume configuration

HDFS:

Hadoop distributed file system is used for storing huge volume of data. Instead of storing in an local file system, put up the data in hadoop cluster Raw Data is segregated into 64MB input splits. Input splits were stored in the data node. To overcome the fault tolerance ,by default 3 replication resides in the data node.

Data from the source

Aug 25 18:26:22 +0000
2015","favorite_count":0,"place":null,"coordinates":null,"text":
"RT @AAsoftCo: Hortonworks buys better Hadoop data flow
management http://t.co
/w4lvWXOvWM","contributors":null,"retweeted_status":
{"filter_level":"low","contributors":null,"text":"Hortonworks
buys better Hadoop data flow management http://t.co
/w4lvWXOvWM","geo":null,"retweeted":false,"in_reply_to_screen_na
me":null,"possibly_sensitive":false,"truncated":false,"lang":"en
","entities":{"trends":[],"symbols":[],"urls":
[{"expanded_url":"http://arsoft-company.biz/hortonworks-
buys-better-hadoop-data-flow-management/","indices":
[52,74],"display_url":"arsoft-company.biz/hortonworks-
bu\u2026","url":"http://t.co/w4lvWXOvWM"}],"hashtags":
[],"user_mentions":
[]},"in_reply_to_status_id_str":null,"id":656233138667130880,"so
urce":"WordPress.com\/a>","in_reply_to_user_id_str":null,"favor
ited":false,"in_reply_to_status_id":null,"retweet_count":1

III.DATA PROCESSING

Data processing can be done with the help of Apache hive. It is an analysis tool present in Hadoop ecosystem. Hive configured an top of the hadoop. It's query language HiveQL, it helps in querying and managing large data sets .tweets are in the form of JSON format.

By default hive is processing with row format. Hive has an Hive SerDe is serializer and Deserializer, It is an interface to convert the data than hive can process. Hive also has an feature called partitions helps to make the product search. Segregating words into partitions, those partitions a can further divided into partitions. Before querying with

the partitions, make sure that

```
set hive.exec.dynamic.partition = true;

set hive.exec.dynamic.partition.mode = nonstrict;
```

create a hive table for raw data set

```
CREATE EXTERNALTABLE rawdata(

tweetId BIGINT, username STRING,

word STRING, CreatedAt STRING,

profileLocation STRING COMMENT 'Location of user',

favc BIGINT,retweet STRING,retcount BIGINT,followerscount BIGINT)

PARTITIONED BY (datehour INT)

COMMENT 'This is the Twitter streaming data'

ROW FORMAT DELIMITED

FIELDS TERMINATED BY 't';
```

Loading data into hive

```
LOAD DATA INPATH '/user/flume/tweets/2013/02/25/16' INTO TABLE rawdataPARTITION (datehour='2013022516');
```

Apache Oozie:

It is coordinator application, putting raw data to hive table for every hour. It will add new partition for every hour.

Data processing is done with the help of hive queries. Data is partitioned according to hour of arrival. Tweets are filtered according to the product name.

IV. SCORE ANALYSIS

Score can be analyzed according to positive ,neutral, negative. Tweets were analyzed according to filtered words which is compared to data set which is already available. on that comparison tweets are rated and categorized into positive, negative and neutral. This helps the marketing industry to know feedback about product. if tweet came along with a location, sales of the product can be reviewed according to the location.

V. CONCLUSION

In this decade Marketing industry uses the social media to review the feedback. By using an hadoop ecosystem tools, extracted raw data from the twitter source and hiveQl query language is used for analyzing this data. Based on the analysis the data is segregated into positive ,negative and neutral. This process helps in analyzing best opinion mining.

VI. REFERENCES

[1] Efthymios Kouloumpis, TheresaWilson and Johanna Moore, “Twitter Sentiment Analysis: The Good the Bad and the OMG!”, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media(2011).

[2] Mohammad Samiul Islam, Faysal Ahmed, Rashedur M. Rahman, Syed Akib Anwar Hridoy and M. Tahmid Ekram, ”Localized twitter opinion mining using sentiment analysis”,Springer (2015).

[3] Carolina Bigonha, Mirella M. Moro, Marcos A. Gonçalves, Virgílio A. F. Almeida, Thiago N. C. Cardoso, “Sentiment-based influence detection on Twitter”,The Brazilian Computer Society (2011).

[4] Justin Zhan and Xing Fang,”Sentiment analysis using product review data”,Springer (2015).

[5] Borut Sluban, Igor Mozetič, Jasmina Smailović, Petra Kralj Novak, “Sentiment of Emojis”,Plos one (2015).

[6] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau, “Sentiment Analysis of Twitter Data”, NextGen Invent (NGI) Corporation(2012).

[7] Hailin Jin and Jianchao Yang, Quanzeng You and Jiebo Luo, “Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks”,Association for the Advancement of Artificial Intelligence (2015).

[8]Zikmund,Babin, “Essentials of Marketing Research (Book Only),” South-Western Pubublication,2009.

[9]Twitter,“Twitter is an online social networking service websites,” https://twitter.com.

[10]Twitter development, “Twitter Application Management: Create and maintain,” https://apps.twitter.com.

[11]Hair, J. F., Anderson, R. E., Tatham, R. L.,Black,“Multivariate analysis,” Englewood: Prentice Hall International,1998.

[12]Hernandez, Stolfo, “Real-world data is dirty: Data cleansing and the merge/purge problem. Communication of the ACM, vol.2,pp.23-34,1998.

[13]Herr,Kardes, F.Kim,"Effects of Word-of-Mouth and Product-Attribute Information on Persuasi

on: An Accessibility-Diagnosticity

[14]. S Saravanan, V Venkatachalam "Optimization of SLA violation in cloud computing using artificial bee colony" Int. J. Adv. Eng Int. J. Adv. En ,Vol.1,pp410-414,2015.

[15] S Saravanan, V Venkatachalam ," Improving map reduce task scheduling and micro-partitioning mechanism for mobile cloud multimedia services" International Journal of Advanced Intelligence Paradigms ,Vol 8(2),pp157-167,2016.

[16] S Saravanan, V Venkatachalam ," Advance Map Reduce Task Scheduling algorithm using mobile cloud multimedia services architecture" IEEE Digital Explore,pp21-25,2014.

[17]] S Saravanan, V Venkatachalam ," Enhanced bosa for implementing map reduce task scheduling algorithm" International Journal of Applied Engineering Research,Vol 10(85),pp60-65,2015.