Investigation of Opinion Mining on Derived Twitter Data using Big Data Tools

[1]MD Waliur Rahman, [2]Prof. Sarvesh Site

M. Tech. Scholar, Department of Computer Science and Engineering, All Saints' College of Technology, Bhopal[1]

Guide, Department of Computer Science and Engineering, All Saints' College of Technology, Bhopal[2]

Abstract: With speedy innovations and growing web population, petabytes of data area unit being generated each second. Process this monumental knowledge and analyzing may be a tedious method now-a-days. The quantity of information in period of time is growing rapidly. Nearly 80% of the info is in unstructured format. Analysis of unstructured knowledge in period of time may be a terribly difficult task. Existing traditional business intelligence (BI) tools perform best only in a pre-defined schema. In this paper, a solution has been proposed that fetches real time twitter data and stored into hadoop components. After storing, sentiment analysis has been performed on these data using big-data analytical tools like: Apache Flume, Apache hive and Apache pig. Finally, their performance comparison has been presented. The results and analysis done on the twitter data, which is shown with the help of tables, diagrams and snapshots, later the comparison is done between the tools on which the sentiment analysis has been done. And after that, this idea and conclusion gotten that pig runs faster and works in fewer map-reduce works compare to hive.

Keywords: - Opinion Mining, Twitter Data, Sentiment Analysis, Hadoop Component

I. INTRODUCTION

Micro blogging is a very famous and popular communication tool used among the Internet users [1]. Twitter is one of the big and largest social media sites which receive millions of tweets every day on different and variety of important and trending issues. Users who post their tweets write about their condition, life, share opinions on variety of topics and discuss the hot and current issues. These posts are then analyzed by Government, Elections, Business, Product review etc. for decision making. Sentiment analysis is therefore, one of the important areas of analysis of twitter posts that can be very helpful in decision making.

Social media has gained enormous popularity within marketing teams [2], and Twitter is an effective tool for a corporation to get people excited about its new products launched. Twitter makes this easy to engage users and communicate straightly with them, and in turn, users will be able to provide word-of-mouth marketing for companies by discussing the products [3]. Given limited resources, and understanding it may not be able to speak with everyone that is the target straightly, marketing departments can be more effective by being selective about whom you reach out to rather than carrying out field surveys for acquiring feedback.

Performing and doing Sentiment Analysis on Twitter is more difficult than performing it for huge reviews [4]. This is because the tweets are very small and short (only about 140 characters) and usually contain emotions, slangs, hash tags and other twitter exact jargon. For the improvement of purpose twitter provides streaming API [5] which permits the developer an access to 1% of tweets tweeted at that time bases on the specific keyword. The object about that the sentiment analysis is done and performed on, is submitted to the twitter API‟s which does additional mining and provides the tweets related to only those objects. Twitter data is commonly unstructured example: using of abbreviations is very high. Also it sanctions the use of emoticons which are direct pointers of the authors view on the subject. Tweet messages as well as consist of a timestamp and the user name. This timestamp is useful for guessing the future trend application [6] of this project. User location if available can also help to gauge the trends in different geographical regions.

II. METHODOLOGY

2.1 Collecting of Twitter Data

Human analysts with no special tools will now not be of huge volumes of information. Data processing will alter the method of finding patterns in data. The results are often either utilized by machine-driven call support systems or bimanual human testing. Data processing is AN integral a part of science and business areas to investigate huge amounts of information to find trends. Twitter may be a great tool for social internet mining as a result of it's an expensive supply of social information attributable to its inherent openness for public consumption. It's a clean and well-documented API [25], wealthy developer tool, and encompasses a broad attractiveness tousers. Data processing in Twitter is easy and might bring vital worth.

Data Mining is extraction of data hidden from giant volumes of data. Information discovery varies from ancient info retrieval from databases. In a very ancient DBMS, information records area unit came in response to

a query; whereas in information discovery, what's retrieved is implicit patterns. The method of discovering such patterns is termed data processing.

2.2 Twitter API

In computer programming, associate degree Application Programming Interface (API) needs a code part in terms of its operations, their inputs and outputs and underlying sorts. Its main purpose is to outline a group of functionalities that area unit freelance of their several implementation, permitting each definition and implementation to vary while not compromising one another. Associate degree application--programming interface (API) could be a set of programming directions and standards for accessing a net--based code application or Web tool. A code company releases its API to the general public so alternative code developers will style merchandise that area unit power-driven by its service.
In addition to accessing databases or constituent, like onerous disk drives or video cards, associate degree API will be accustomed ease the work of programming graphical program elements, to permit integration of latest options into existing applications (a so--called "plug--in API"), or to share knowledge between otherwise distinct applications. In observe, many times an API come within the kind of a library that has specifications for routines, knowledge structures, object categories, and variables. In other cases, notably for SOAP and REST services, associate degree API comes as simply specification of remote calls exposed to the API customers.

2.3 Data Pre-processing

Data preprocessing is a data mining and extracting technique that contains transforming raw data into an understandable format. Real-world data is often inconsistent, incomplete and/or lacking in certain trends or behaviors, and is likely to contain many errors. Data preprocessing is a confirmed and proven method of resolving such issues. Data preprocessing makes ready raw data for further processing.
Preprocessing steps are performed prior to classification. They are as follows:

1. Fetch data from Twitter Streaming API.
2. Convert unstructured JSON data into structured data.
3. Apply stemming to remove stop words.
4. Store the preprocessed data for analysis.

III. PROPSOED WORK

For analyzing these huge and complex data requires a powerful tool, hadoop is used that is an open source implementation of map-reduce, a powerful tool designed for deep analysis and transformation of very large data

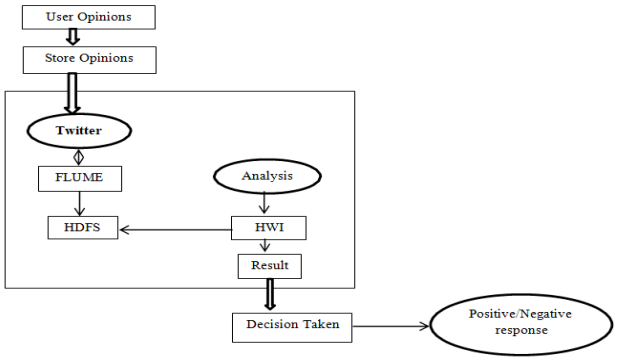

Fig: 1. Proposed system workflow

Algorithm Steps are as 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 center , there are millions of tweets are posted everyday on twitter which can generates petabytes of data which is stored on twitter data center.

Step 3: for analysis these huge and complex twitter data are needed which contains variety of opinions posted by different users, flume is used to fetch these twitter data and store them into HDFS, a twitter API is generated through which the real time twitter data is fetched from web and store them into HDFS.
Step 4: After storing these huge and complex twitter json data, an analyzing tool is needed to analyze these complex data, for these hive is used which runs on top of the hadoop and takes input from HDFS and its support SQL queries through which the data can be analyzed.

Step 5: Based on the analysis result from hive, the polarity of the tweets can be checked with the aid 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.

IV. SIMULATION ENVIRONMENT

For collecting the twitter data apache flume is configured on the top of Hadoop to collect and store tweets in HDFS and also config apache hive for analyse the twitter data comes from web.
The following tools are required for extracting and analyses of data:

1. ApacheHadoop
2. ApacheFlume
3. ApacheHive
4. ApachePig

Hadoop: Hadoop is associate degree Apache open supply framework that has been written in java that enables distributed process of large datasets across clusters of computers mistreatment straightforward programming models. The Hadoop framework application works in associate degree setting that gives distributed storage and computation across clusters of computers. Hadoop is meant to rescale from single server up to thousands of machines, every providing native computation and storage.

There are lots of work have been done by the researchers which is based on big data. So, Hadoop gives a very good platform to work on them. It has 2 parts: Hadoop map reduce and other is distributed file system also known as Hadoop Distributed File System. Map reduce was utilizing its file to work on in HDFS. The various input and output stores by HDFS and it easy to monitoring, managing communication and controlled system component by frame work.

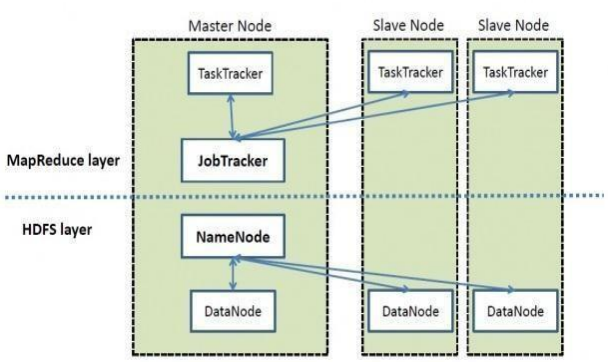

Fig. 2: Hadoop Architecture [5]

Hadoop Architecture

At its core, Hadoop has two major layers namely:

1. Processing/Computation layer (Map Reduce).
2. Storage layer (Hadoop Distributed File System).

V. EXPERIMENTL RESULTS

(HQL)After running the Flume by setting the above configuration then the Twitter data will automatically will be saved into HDFS where 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 has been got form the Twitter is also in the JSON format.

Create external table load _tweets
{
Id BIGINT, Text STRING
ROW
FORMAT
SERDE
„com.cloudera.h
ive.serde.JSONS
erDe"
LOCATION
„/home/hadoop/
hivewarehouse"
}

Figure 2: Data Stored In Load_Tweets

After finishing mapreduce job execution a new table is obtain called words table which consist of two fields first is which has bigint datatype and another fields consist splited words which has a arrasy of stringtypes. After creating a words table from load_tweets table the data which is stored into words table, in which one fields consists twitter id and another fields consist array of splited words.

Figure 3: Data stored in words_table

Analysis Performance Comparison

After analyzing the twitter data the polarity of tweets is gotten, in this thesis the compression between performance of Apache pig and Apache hive is don for analyzing JSON data. For this different size of dataset is gotten on which the analysis can be performed using hive and pig , The execution time taken by both the analytical tools on different size datasets are shown in figure 4.

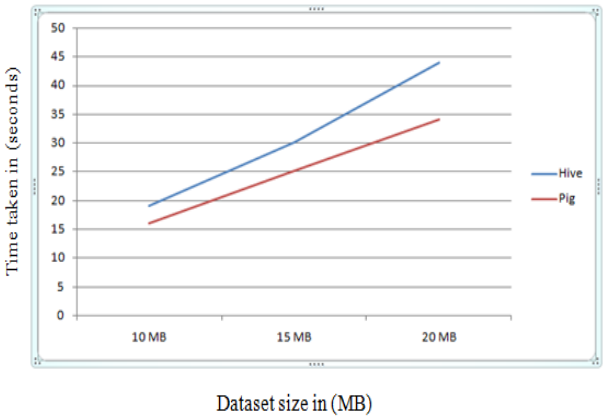

Figure 4: Execution time taken by hive & pig

Table 1: Classification result of reviews

Description	Bigdata	Hadoop	Cloud	Analytics
Positive	25	49	25	1
Neutral	20	2	9	0
Negative	0	0	0	0
Result	Positive	Positive	Positive	Positive

VI. CONCLUSION

In this the popularity of tweets can also be identified by which it can be said that which tweet have a positive meaning or a negative meaning. In this paper the twitter data is fetched by using flume and store them into the HDFS and then these data are analyzed by using hive and pig, the results and analysis done on the twitter data, which is shown with the help of tables, diagrams and snapshots, later the comparison is done between the tools on which the sentiment analysis has been done.

REFERENCES

[1] Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub,and More--- Matthew A.Russell.

[2] G. Szabo, and B.A. Huberman, “Predicting the Popularity of OnlineContent”, Communication of the ACM, 2010, 53(8), pp.80-88.

[3] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, “Improving LDA topic models for microblogs via tweet pooling and automatic labeling,” in Proceedings of the 36th International ACM SIGIR Conference on Research and Development inInformation Retrieval, ser. SIGIR “13. New York, NY, USA: ACM, 2013, pp.889–892.

[4] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. A. Goncalves, and F. Benevenuto, “Analyzing the dynamic evolution of hashtags on twitter: a language-based approach,” in Proceedings of the Workshop on Language in Social Media (LSM 2011). Portland, Oregon: Association for Computational Linguistics, 2011, pp. 58–65.

[5] "The Streaming APIs." *Twitter Developers*. N.p., n.d. Web. 23 Oct.2014.

[6] Y. Wang, J. Liu, J. Qu, Y. Huang, J. Chen, and X. Feng, “Hashtag graph based topicmodel for tweet mining,” in Data Mining (ICDM), 2014 IEEE International Conference on, Dec 2014, pp.1025–1030.

[7] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a News Media?", In: Proceedings of the 19th International Conference on World Wide Web, 2010, pp.591-600.

[8] McKinsey, Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey & Company, 2011, http: //www.mckinsey.com/.

[9] Sagiroglu, S., &Sinanc, D, “Big data: A review”, IEEE International Conferenceon Collaboration Technologies and Systems (CTS), 2013, pp42-47.

[10] K. W. Lim and W. Buntine, “Twitter opinion topic model: Extracting productopinions from tweets by leveraging hashtags and sentiment lexicon,” in Proceedings of the23rd ACM International Conference on Conference on Information and Knowledge Management, ser. CIKM “14. New York, NY, USA: ACM, 2014, pp.1319–1328.