IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Survey on Buddy Analytics Based on Social Media

Mangesh U. Sanap[1], Prof.V.S.Phad[2]

Dept. of Computer Engineering, SKNCOE, Pune, India[1, 2]

Abstract: 'BIG-DATA' used in different industries over the last few years, on a scaling that generated lots of data every day. Big Data is a term applied to data sets of very large size such that the traditional databases are unable to process their operations in a significant amount of time. Big Data is a collection of data that is large and/or complex to process using data processing applications, Hadoop is a distributed paradigm used to manipulate the large amount of data. It actually holds the huge amount of data & perform the operations like data analysis, result analysis, data analytics etc. It is highly scalable computing platform. Productive E-commerce sites, Facebook, Twitter one of the largest social media site receives comments, tweets or customer reviews in millions every day in the range of terabyte or petabytes per day. Ideas and opinions of people are influenced by the opinions of other people. Lot of research is going on analysis of reviews given by people. We can collect the data from the social media site by using BIGDATA eco-system using online streaming tool Flume. We are using Hive and its queries to give the sentiment data based up on the groups that we have defined in the HQL (Hive Query Language). Here we have categorized this sentiment analysis into 3 groups like comments that are having positive, moderate and negative comments. This Analytics paper provides a way of analyzing of big data such as Facebook data using Apache Hadoop which will process and analyze the comments on a Hadoop clusters.

Keywords: Hadoop, Big Data, Map Reduce, Facebook, HDFS, Sentimental Analysis, Flume.

I. INTRODUCTION

Before Last few years, industries and organizations didn't need to store and perform much operations and analytics on data of the customers. However since 2005, the need to transform everything into data is much entertained to satisfy the requirements of the people. So Big data came into picture in the real time business analysis of processing data. Some well-known internet companies like Google, Amazon, LinkedIn, Yahoo! etc. have generated a huge amount of structured and unstructured data every day. This exponential growth of data leads to some challenges like processing of large data sets, extraction of useful information from online generated data sets etc. Hadoop is one of the processing tools that is used to analyze and process large data sets. It has two main components: HDFS, Hadoop Distributed File System used for reliable storage of data and MapReduce, which is used to process the data.

The term big data refers to the data that is generating around us everyday life. It is generally exceeds the capacity of normal conventional traditional databases. For example by combining a large number of signals from the user's actions and those of their friends, Facebook developed the large network area to the users to share their views, ideas and lot many things. 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 Facebook as our example nearly 1TB of text data is generating within a week in the form of comments. So, by this it is understand clearly how this Internet is changing the way of living and style of people. Among these comments can be categorized by the hash value tags for which they are commenting and posting their comments. 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.

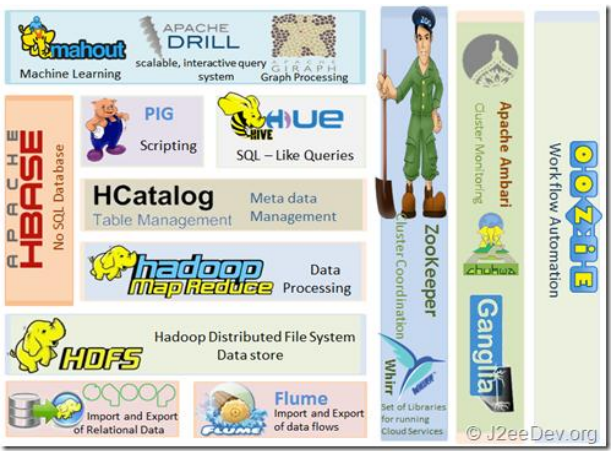

Fig1: Apache Hadoop Eco-System

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 BIG-DATA [2] Problem as a solution. And if we consider getting the data from Facebook [5] 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

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

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 [3].

Today, the textual data on the internet is growing at a rapid pace. Different industries are trying to use this huge textual data for extracting the people's views towards their products. Social media is a vital source of information in this case. It is impossible to manually analyze the large amount of data. This is where the need of automatic categorization becomes apparent.[4] Subjective data is analyzed generally in this case. There are a large number of social media websites that enable users to contribute, modify and grade the content. Users have an opportunity to express their personal opinions about specific topics. The example of such websites include blogs, forums, product reviews sites, and social networks. In this case, Facebook data is used. Sites like Facebook contain prevalently short comments, like status messages on social networks like Facebook. Additionally many web sites allow rating the popularity of the messages which can be related to the opinion expressed by the author. The focus of our project is to assign the polarity to each comment i.e. whether the author express positive or negative opinion.

II. RELATED WORK

Zhibo Wang, Jilong Liao, Qing Cao, Hairong Qi, and Zhi Wang[1] proposed Friendbook paper implement existing social networking services is how to recommend a good friend to a user. Most of them rely on pre-existing user relationships to pick friend candidates. For example, Facebook relies on a social link analysis among those who already share common friends and recommends symmetrical users as potential friends.

In previous Friendbook paper, author presented the design and implementation of Friendbook, which is semantic-based friend recommendation system for social networks. Different from the friend recommendation mechanisms relying on social graphs in existing social networking services, Friendbook extracted life styles from user-centric data collected from sensors on the smartphone and recommended potential friends to users if they share similar life styles. Existing system implemented Friendbook on the Android-based smartphones, and evaluated its performance on both small scale experiments and large-scale simulations. The results showed that the recommendations accurately reflect the preferences of users in choosing friends.

Minqing Hu and Bing Liu [10] analyzed set of techniques for mining and summarizing product reviews based on data mining and natural language processing methods. The objective is to provide a feature-based summary of a large number of customer reviews of a product sold online.

Penchalaiah.C1, Murali.G2Suresh Babu [6] proposed Sentiment Analysis on Twitter Data using: Apache Flume and Hive and solving it in BIGDATA by using Hadoop and its Eco Systems. And finally we have done sentiment analysis on the Twitter data that is stored in HDFS.

Piyush Gupta, Pardeep Kumar, Girdhar Gopal [12]proposed A new approach about doing the sentiment Analysis with theuse of MapReduce to run it faster.

III. PROPOSED WORK

In our Analytics paper evaluate existing system extends on large-scale Experiments. We proposed evaluate Friendbook system on large scale experiments using social media data.

Recently analysis is worked for few Twitter tweets analysis but here we are to doing work for user behavior analysis using Facebook Comments or any social media data reviews, user likes, interest's So here we are going to use Hadoop and its Ecosystems, for getting raw data from the Facebook or any social media sites we are using Hadoop online streaming tool using Apache Flume [6]. Using Flume tool only we configure everything that we want to get data from the Facebook.[7] For analysis we want to set the configuration and also want to define what information that we want to get form Facebook All these will be saved into our HDFS (Hadoop Distributed File System)[8] in our prescribed format.

In this proposed system, a method to calculate Analysis of reviews or comments given by the customers or user is proposed and implemented in Java on Hadoop. The method works in two phases: Mapper phase and Reducer phase. We are use a positive and negative word dictionary to identify positive and negative words [9] [10]. Stop word dictionary is used to identify and remove stop words from the reviewed product [11]. The focus of our project is to assign the polarity to each comment i.e. whether the author express positive or negative opinion.[12]

IV. ARCHITECTURAL VIEW

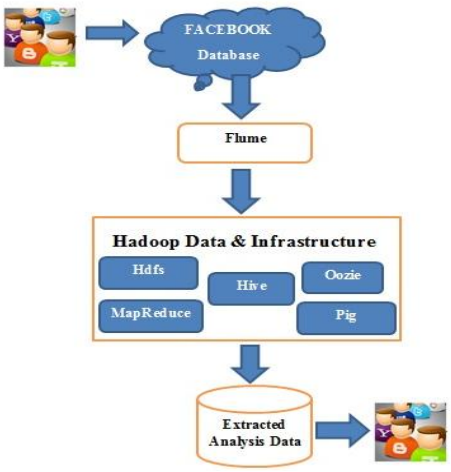

Fig2: Proposed Architecture for Buddy Analytics

The above figure2 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 analysis is going to perform.

DOI 10.17148/IJARCCE.2015.412124

Table1: Survey Table

Sr. no	Paper	Technique	Advantage	Disadvantage	Result
1	Friendbook: A Semantic-based Friend Recommendation System for Social Networks[1]	Friend Recommendation System for Social Networks	Recommended potential friends to users if they share similar life styles	Recommend friends to users based on their social graphs, which may not be the most Appropriate to reflect a user's preferences on friend selection in real life.	Friendbook extracted life styles from user-centric data collected from sensors on the smartphone and recommended potential friends to users if they share similar life styles
2	Effective Sentiment Analysis on Twitter Data using: Apache Flume and Hive [6]	HQL (Hive Query Language) & Flume.	Easily extracted data from huge database	Need Oozie by creating a work flow so that can give a time slang such that it will work based upon that time we allocated for performing a particular work	The best methods to process large amount of data in a small time.
3	Sentiment Analysis on Hadoop with Hadoop Streaming. [12]	Sentiment analysis computational technique	method that assigns scores indicating positive and negative opinion about the product is proposed	This not work on online data	Sentiment Analysis is one of the analytics which will tell the writers/ producers sentiment about their thoughts
4.	Mining and Summarizing Customer Reviews. [10]	Feature-based opinion summarization	are use a positive and negative word dictionary to identify positive and negative words	Set of techniques for mining and summarizing product reviews based on data mining and natural language processing methods.	It provides a feature based summary of a large number of customer reviews of a product sold online.

V. CONCLUSION

Data volume increasing rapidly nowadays, it is required to process on data speedily. Analytics is used to tell the writers/producers sentiment about their thoughts. There are several ways to define and analyze the social media data such as facebook, Twitter etc. Here anyone can perform different operations queries in these type of data. But the problem arises when dealing with BIGDATA. In This analytics paper we have try to execute problem statement and solving it in BIGDATA by using Hadoop and its Eco Systems. And finally we will try to done Buddy analytics based on user Facebook comments or reviews, likes, interests. Here it is solving by using Hadoop and its packages. And we have trying to done some Buddy analysis on their comments and the most number of comment ids.

REFERENCES

[1] Zhibo Wang, Jilong Liao, Qing Cao, Hairong Qi, and Zhi Wang, Member, IEEE," Friendbook: A Semantic-based Friend Recommendation System for Social Networks" IEEE TRANSACTIONS ON MOBILE COMPUTING (Volume:14,Issue:3)

[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big Data: The Next Frontier For Innovation, Competition, And Productivity", May 2011.

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

[4] Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis".

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

[6] Penchalaiah.C1, Murali.G2Suresh Babu.A3" Effective Sentiment Analysis on Twitter Data using: Apache Flume and Hive "

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

[8] T. White, "The Hadoop Distributed File system," Hadoop: The Definitive Guide, pp. 41-73, Gravenstein Highway North, Sebastopol: O'Reilly Media, Inc., 2010.

[9] "Opinion Mining, Sentiment Analysis, and Opinion Spam Detection," (Last visited in june 2015) [online]. Availaible: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[10] Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA.

[11] "stop-words," (Last visited in june 2015) [online]. Available: https://code.google.com/p/stop-words/

[12] Piyush Gupta, Pardeep Kumar, Girdhar Gopal "Sentiment Analysis on Hadoop with Hadoop Streaming" International Journal of Computer Applications (0975 – 8887) Volume 121 – No.11, July 2015