ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

User Analytics on Twitter Stream Data

Sagar Rane, Manik Hendre, Sharayu Lokhande
Assistant Professor, Dept. of CSE, Army Institute of Technology, Savitribai Phule Pune University, Pune, India.

Manuscript Info

Manuscript History:

Received: 14 December 2015
Final Accepted: 19 January 2016
Published Online: February 2016

Key words:
Hadoop, Big Data, Map Reduce, Twitter, HDFS, Tweets, Sentimental Analysis, Flume.

Corresponding Author

Sagar Rane.

Abstract

Twitter, is one of the largest social media site that receives tweets in millions of data in every day in range of Petabytes per year. Big Data is a pool of information that is outsized and difficult to progression by data processing applications, Hadoop is a disseminated archetype used to handle the huge quantity of documents. It grasps the vast quantity of documents and carry out the procedures like documents analysis, outcome analysis, and records analytics. 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. This huge amount of raw data can be used for industrial or business. This Analytics paper provides a way of analyzing of big data such as Twitter data using Apache Hadoop which will process and analyze the tweets on a Hadoop clusters. In this paper, we are going to talk how effectively sentiment analysis is done on the data which is collected from the Twitter using Flume. Twitter is an online web application which contains huge amount of data that can be a structured, semi-structured and un-structured data. Twitter is also difficult due to language that is used for comments. So here we are taking sentiment analysis, for this 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.

Copy Right, IJAR, 2016,. All rights reserved.

Introduction:-

Big data is recycled ubiquitously at the present in disseminated archetype on web. BIG data is the group of collections of massive volume of data. 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. Twitter.com is a popular microblogging website. Each tweets is 140 characters in length. Tweets are frequently used to express a twitter's emotion on particular subject. The upcoming of operational societal mass media and communication machineries has activated a quick rise in the stream of user produced content of several forms. Persons are precise their responses, desires and preferences through societal mass media via means of word-based piece of short nature relatively scripting extensive writing. We appeal groups complete through this old-fashioned way as G-friends, which stands for topographical geo location based groups. 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 gears: HDFS used for reliable storing of files & Map Reduce, which is used to process the documents. In social media most of people adds their habits, daily activities, etc. on that data (habits, activities, comments) I found who is most matched to another person... I analyze that person to user by calculating polarity.

One difficult task with current social interacting facility area is that large data is stored in HDFS means it is more scalable... Masses of tweets are produced every period of time on diverse issues. I plan an unsupervised and domain-independent tactic by means of the polarization scores using three lexical sources SentislangNet, SenticNet2, SentiWordNet 3.0.SentiWordNetcomprisespolarizationmarks of small units for that I express positive or negative opinion.

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, Twitter 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 Twitter 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.

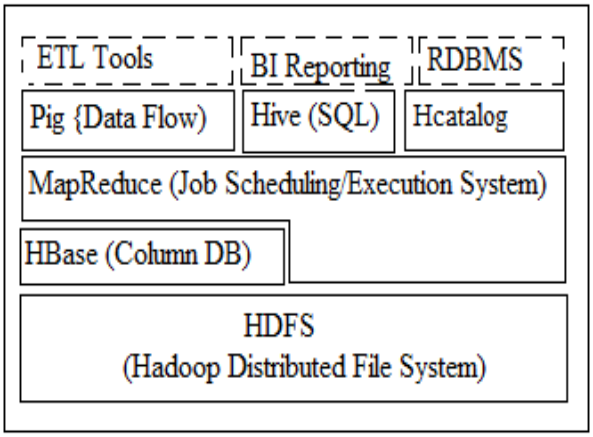

fig. Ecosystem

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 BIG-DATA [8]. Problem as a solution. And if we consider getting the data from Twitter [11] 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 showing of data and generating them into structured data in the form of tables can be done by using Apache Hive [9].

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 [10] Subjective data is analyzed generally in this case. There are an enormous sum of societal mass media webs that permit persons to donate, revise and mark the data. Consumers have a chance to direct their individual feelings about particular subjects. The cases of such webs include blogging sites, media sites, artifact examinations sites, and social nets. In this case, Facebook data is used. Sites like Facebook cover generally

small explanations, like status msgs on social systems like FB. Further more numerous websites permit marking the popularity of the mgs which can be associated to the view articulated by the writer. The focus of our project is to assign the polarity to each comment i.e. whether the author express positive or negative opinion.

Related work:-

In this paper, L. Page, S. Brin, R. Motwani, and T. Winograd [1] have got on the bold job of summarizing each page on the WWW into a sole figure, its PageRank. PageRank is a universal standing of all web sheets, irrespective of their data, based only on their place in the website graph assembly. It establish a numerals of requests for PageRank in accumulation to search which include stream of traffic estimation, and user map reading. Also, it can produce modified PageRank's which can make an outlook of web from an individual standpoint. Generally, our trials with PageRank put forward that the assembly of the web map is precise and valuable for a diversity of info retrieval jobs.

In this paper, P. Desikan, N. Pathak, J. Srivastava, and V. Kumar. [2] Suggest a technique to incrementally calculate PageRank for a big graph that is developing. Here they have given method to calculate PageRank incrementally for developing graphs. The key opinion is that evolution of the web graph is sluggish, with big portions of it remaining unaffected. By sensibly describing the altered and unaffected slices and the dependency across them, it is thinkable to progress effective procedures for calculating the PageRank metric incrementally.

In this paper, W. H. Hsu, A. King, M. Paradesi, T. Pydimarri, and T. Weninger[3] spoken the challenge of link commendation in web blogging and similar social nets. It discussed the crushed Structures accessible in *Live Journa*l's community user info pages &define graph procedures which are important for analysis of the social network.

In this paper E. Miluzzo, N. D. Lane, S. B. Eisenman, and A. T. Campbell [4], addressed a full picture of the cenceme design. Over our archetype implementation they have shown an effective mixing with an amount of widespread standard customer computer connected nodes and social interacting applications.

In this paper, T. Huynh, M. Fritz, and B. Schiel [5] given an innovative technique to identify day-to-day habits as a probabilistic mixture of doings patterns. The usage of theme mockups enables the programmed detection of such designs in a consumer's everyday routine.

In this paper, Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma [6] set goal in the direction of thoughtful user mobility based on Global Positioning System content. An effort pointing to conclude convey ancestyles from Global Positioning System logs created by supervised knowledge is described.

Zhibo Wang, Jilong Liao, Qing Cao, Hairong Qi, and Zhi Wang[7] given Friend book paper instrument surviving social interacting facilities is how to endorse a perfect colleague or associate to a consumer. Maximum of them depend on pre-existing consumer associations to pick colleague or associate applicants. For e.g., FB relies on a social relation investigation among those who already part of common contacts and endorses symmetrical consumers as possible friends.

In previous Friend book paper, author given the strategy and execution of Friend book, which is based on meaning friend endorsement system for social nets. Dissimilar from the freeing or colleague or associate endorsement appliances relying on social graphs in current social interacting services, Friend book take out natural life styles from consumer-centric content composed from smartphone sensors and suggested possible families to consumers if they share similar natural life styles. Current system implemented Friendbook on the Android OS phones, and assessed its performance on both minor scale trials and extensive simulations. The outcomes presented that the endorsement precisely replicate the favorites of consumers in choosing families of colleagues.

This paper contains of three stages: alteration in point-based segmentation or Subdivision technique, an implication model and a graph-based post-processing algo. It takes a transformation point-based segmentation or subdivision method to panel each Global Positioning System route into separate divisions of different transport modes. Other, from each segment or subdivision, It classify a set of cultured features, are not affected by different traffic situations. After these structures of features are nourished to a procreative suggestion model to organize the segments or

subdivisions of diverse modes. Lastly, it takes graph-based post treating to further progress the suggestion performance.

System overview:-

As it can have seen existing system drawbacks, here we are going to overcome them by using Big Data problem statement. In our Analytics paper evaluate existing system extends on large-scale Experiments. So here we are going to use Hadoop, for getting raw data from the Twitter or any social media sites we are using Hadoop online streaming tool using Apache Flume [12].

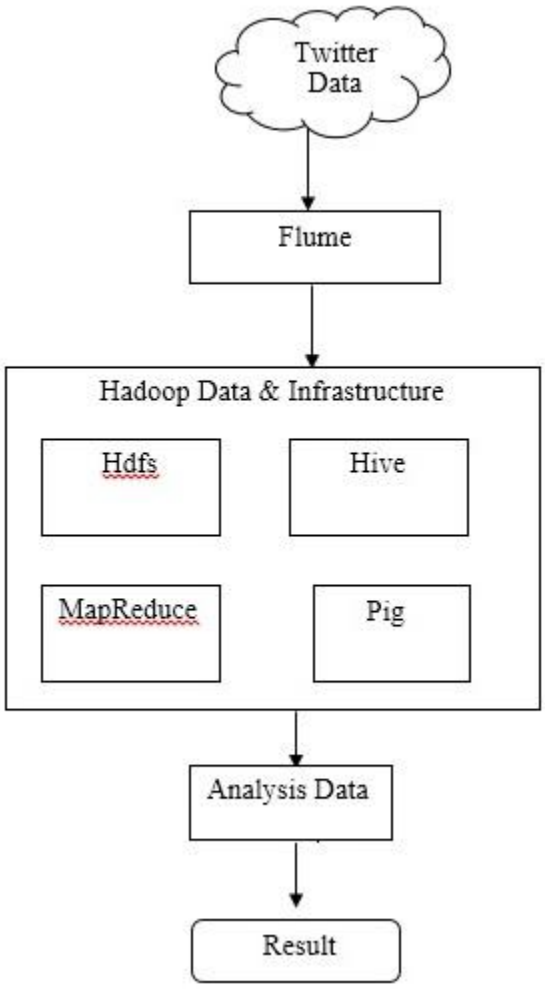

fig. Architecture of Twitter Analytics

First the raw data of Twitter accessed by using flume. All these will be saved into our HDFS (Hadoop Distributed File System) [13] in our prescribed format. Secondly, HDFS can give the data that is stored in it to Hadoop Framework.

Hadoop Framework are nothing but map reduce, hive, pig. In this 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 [14] [15]. Stop word dictionary is used to identify and remove stop words from the

reviewed product [16]. The focus of our project is to assign the polarity to each comment i.e.it gives valuable opinion or suggestion which may be positive or negative. [17]. 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.

Conclusion:-

In this paper, we have taken only the polarization marks or scores from above mentioned source i.e. SenticNet to analyze twits. We have to assess this model on extensive field experiments. We have done some analysis on the tweets and the most number of tweet ids. There are several ways to define and analyze the social media data such as Twitter, Facebook. 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 User analytics based on user Twitter comments or Tweets, reviews, likes, interests.

References:-

1. L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab, 1999.

2. P. Desikan, N. Pathak, J. Srivastava, and V. Kumar. Incremental page rank computation on evolving graphs. Proc. of WWW, pages 1094-1095, 2005.

3. W. H. Hsu, A. King, M. Paradesi, T. Pydimarri, and T. Weninger. Collaborative and structural recommendation of friends using weblog-based social network analysis.

4. E. Miluzzo, N. D. Lane, S. B. Eisenman, and A. T. Campbell. Cenceme-Injecting Sensing Presence into Social Networking Applications. In Proc. of EuroSSC, pages 1–28, October 2007.

5. T. Huynh, M. Fritz, and B. Schiel. Discovery of Activity Patterns using Topic Models. In Proc. of UbiComp, 2008.

6. Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma. Understanding Transportation Modes Based on GPS Data for Web Applications. ACM Transactions on the Web (TWEB), 4(1):1–36, 2010.

7. 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)

8. 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.

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

10. Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis".

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

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

13. T. White, "The Hadoop Distributed File system," Hadoop: The Definitive Guide, pp. 41-73, Gravenstein Highway North, Sebastopol: O"Reilly Media, Inc., 2010.

14. "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

15. 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.

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

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