JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Survey on Twitter Data Analysis Using Kafka

[1]Mrs.Sophia G,
Assistant Professor, Department of Computer Science and Engineering, MVJ College of Engineering, India

Ranjith V, Roshan A, Tharun S [234],
Student, Department of Computer Science and Engineering, MVJ College of Engineering, India

Abstract: Twitter is an online social networking service with more than 300 million users, generating a huge amount of information every day. Twitter's most important characteristic is its ability for users to tweet about events, situations, feelings, opinions, or even something totally new, in real time. Currently there are different workflows offering real-time data analysis for Twitter, presenting general processing over streaming data. This study will attempt to develop an analytical framework with the ability of in-memory processing to extract and analyze structured and unstructured Twitter data. The proposed framework includes data ingestion, stream processing, and data visualization components with the Apache Kafka messaging system that is used to perform data ingestion task. Furthermore, Spark makes it possible to perform sophisticated data processing and machine learning algorithms in real time. We have conducted a case study on tweets about the earthquake in Japan and the reactions of people around the world with analysis on the time and origin of the tweets.
Key words: Streaming processing, Big Data, Kafka, Spark Twitter, Real time analysis, Data analytics.

1.INTRODUCTION

The most Today, as of now dealing with big data is a big deal. If we see Twitter, the data is continuous flow which is really huge. My Paper mainly deals with gathering real time twitter tweets, either those are of reply/retweeted/normal tweets from TwitterAPI and calculating the sentiment analysis on that particular tweet using Kafka Streaming and finally sending all those sentiment analytical data into the kafka topic, hence by making use of this data with the help of KSQL, we could able to calculate the overall sentiment analysis on a particular keyword, like how people reacted (positive/negative/neutral) about particular keyword. And here the keyword might be a person or an organization or anything else, where we wanted to find how people are speaking about that keyword.

Twitter is a popular social networking site where millions of people tweet every second about various topics related to society, politics, sports, entertainment, and many more. The standard syntax followed by Twitter users while tweeting involves hashtags, retweets, and user mentions. Hashtags are words or phrases which are prefixed with "#," and user mention means mentioning other people, companies, brands, or precisely other Twitter users in the tweet by using the "@" symbol at the beginning of their user name. There is a restriction of 140 characters on the length of any tweet which allows users to post tweets quickly. At the same time, users all across the globe can tweet about anything happening or their thoughts at any given time of the day. Tweets thus help people to understand how others feel about different ongoing events, government policies, sports tournaments, etc. Brands can analyze tweets to know people's sentiments towards their products. Government and politicians get an idea of how people are responding to the different policies, acts, and amendments. During elections, Twitter plays a vital role in campaigning too. For a given day or a span of days, any topic can be made trending by the repeated use of the same hashtag. Thus, Twitter trends play an important role in the process of decision-making by different organizations and companies. The main motivation for the Twitter trend analysis is to identify the recent trends happening across the world using big data machine learning techniques. This will help to analyze what has happened in the past and what may happen in the future. It helps to track customer trends and interests especially what customers like, what their behaviors are, and how these changes over the time. Big data analytics applications enable data scientists, predictive modelers, statisticians and other analytics professionals to analyse growing volumes of structured transaction data, plus other forms of data that are often left untapped by conventional business intelligence (BI) and analytics programs. That contains semi-structured and unstructured data -- for example,

internet clickstream data, web server logs, social media content, text from customer emails and survey responses, mobile-phone call-detail records and machine data captured by sensors connected to the internet of things.

On a broad scale, data analytics technologies and techniques provide means of analysing data and drawing conclusions about them to help companies make informed business decisions. Big data analytics is a form of advanced analytics, which involves complex applications with elements such as models, statistical algorithms and what-if analyses powered by high-performance analytics.

Sentiment Analysis is finding out the type of emotion from the textual information using several text evaluation techniques. Sentiment evaluation allows businesses to identify user's sentiment closer to products, brands or services in online conversations and feedback. Sentiment Analysis has grown to become a motivating field for domains which deals with user experience. The expression resultant from sentiment analysis refers to the experience of the user on some issues. Now-a-days social networking sites are playing vital role since majority of the population uses several social media platforms to show their views and experience. Because of most of the population being on the social media platform, I generate a huge amount of data in the form of comments, articles etc. about certain topic. So by implementing an automated process which investigates and classifies the user experience would be quite essential.

Figure 1: The diagram of 3Vs of big data

2. METHODS OF DETECTION

Big Data

Big data primarily refers to data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many entries (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher rate. Though used sometimes loosely partly because of a lack of formal definition, the interpretation that seems to best describe big data is the one associated with large body of information that we could not comprehend when used only in smaller amounts.

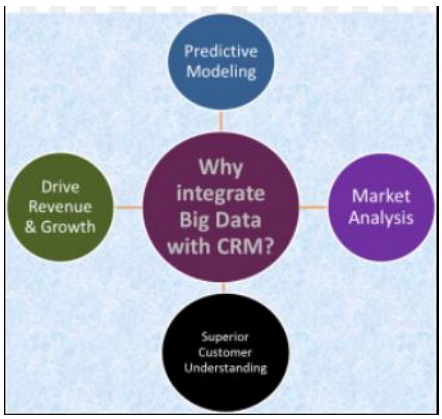

Figure 2. Why Big Data ?

Big data analysis challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and data source. Big data was originally associated with three key concepts: volume, variety, and velocity. The analysis of big data presents challenges in sampling, and thus previously allowing for only observations and sampling. Thus, a fourth concept, veracity, refers to the quality or insightfulness of the data. Without sufficient investment in expertise for big data veracity, then the volume and variety of data can produce costs and risks that exceed an organization's capacity to create and capture value from big data.

Spark

Spark has its architectural foundation in the resilient distributed dataset (RDD), a read-only multiset of data items distributed over a cluster of machines, that is maintained in a fault-tolerant way. The Data frame API was released as an abstraction on top of the RDD, followed by the Dataset API. In Spark 1.x, the RDD was the primary application programming interface (API), but as of Spark 2.x use of the Dataset API is encouraged even though the RDD API is not deprecated. The RDD technology still underlies the Dataset API.

Spark and its RDDs were developed in response to limitations in the MapReduce cluster computing paradigm, which forces a particular linear dataflow structure on distributed programs: MapReduce programs read input data from disk, map a function across the data, reduce the results of the map, and store reduction results on disk. Spark's RDDs function as a working set for distributed programs that offers a (deliberately) restricted form of distributed shared memory.

Kafka

Kafka stores key-value messages that come from arbitrarily many processes called producers. The data can be partitioned into different "partitions" within different "topics". Within a partition, messages are strictly ordered by their offsets (the position of a message within a partition), and indexed and stored together with a timestamp. Other processes called "consumers" can read messages from partitions. For stream processing, Kafka offers the Streams API that allows writing Java applications that consume data from Kafka and write results back to Kafka. Apache Kafka also works with external stream processing systems such as Apache Apex, Apache Beam, Apache Flink, Apache Spark, Apache Storm, and Apache NiFi.

Kafka runs on a cluster of one or more servers (called brokers), and the partitions of all topics are distributed across the cluster nodes. Additionally, partitions are replicated to multiple brokers.

Natural language processing

Natural language processing (NLP) is an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

NLP enables computers to understand natural language as humans do. Whether the language is spoken or written, natural language processing uses artificial intelligence to take real-world input, process it, and make sense of it in a way a computer can understand. Just as humans have different sensors -- such as ears to hear and eyes to see -- computers have programs to read and microphones to collect audio. And just as humans have a brain to process that input, computers have a program to process their respective inputs. At some point in processing, the input is converted to code that the computer can understand.

Integrated Dashboard

Creating an integration between your analytic software and a dashboard is a great way of understanding customer experience. For instance, when new feedback has been received the data can be easily input into the corresponding fields. These fields could indicate customer satisfaction, engagement and so on.

The dashboard is designed to visualize key metrics for each persona. From the summary level dashboards (top level), you can drill down to the detailed level dashboards or to go into next level details for easier data analysis. The out-of-the-box feature integrates ICABI dashboards engine with the Infosphere Master Data Management Collaboration Server - Collaborative Edition. By default, the feature is enabled, and no further configuration is required.

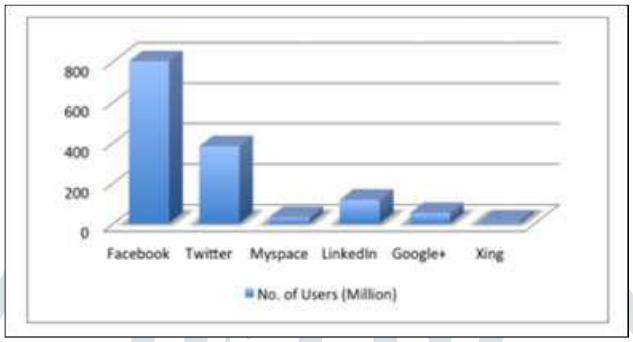

Figure 3: Example dashboard showing number of users

3. Survey Papers

In the previous years, studies of Sentiment Analysis and emotional models had a wide attention. The reason for that is basically due to the recent enlargement of data which exists on the social networks, particularly of those that describe people's point of view, thoughts and comments [5]. Walaa Medhat et al. [6] presented and discussed in brief details different types of sentiment analysis and its applications. Algorithms and their originating references of various SA techniques are categorized and shortly explained. Yang et al. [7] introduced the common sentiment analysis methods from the perspective of machine learning technologies, which encompass Naive Bayes technique, Maximum Entropy method, Support Vector Machine technique, and Artificial Neural Network method and performance assessment and difficulties. Pang et al. [8] were the first to apply Machine Learning for sentiment mining on movie reviews corpus, many classification algorithms were used, whereas unigram and bag of words are utilized for obtaining features. The ratio of accuracy differs according to what they applied for example it was 82.9% by applying Support Vector Machines, while it was78.7% by applying Naive Bayes classifier. Wang et al. [9] used training dataset which contains 17000 Tweets to come up with a real time Twitter Sentiment Analysis System regarding to U.S. voting Presidential Election Cycle in 2012. In [1], authors introduced a new method combining the help of SentiWordNet alongside with an implementation of Naive Bayes; therefore more accuracy can be achieved.

One of the possible techniques to get more accuracy of classification of tweets is applying SentiWordNet and Naïve Bayes that give positive, negative and objective degree of the words exist in tweets. Bindalet al. [10] proposed a twostep system can be applied for sentiment classification of the tweet. During the initial step, sentiment lexicons are used to classify tweets, while the polarity of each tweet is also assigned by aggregating the scores of each token. During the next step, the SVM classifier receives all the tweets with low absolute scores to strengthen the whole accuracy. In [11], authors introduced a novel manner for Sentiment Learning depend on Spark platform; the hashtags and emotions within a tweet are exploited by the suggested algorithm, as sentiment labels, and continue to a classification step of various sentiment types using parallel processing methods. In [12], authors suggested a real-time solution using spark framework, for processing sentiment analysis Saudi dialect in twitter based on lexicon-based algorith. In [13], authors recommended an efficient sentiment prediction technique in Big Data, using Spark. The outcomes got from the suggested work were subject to analysis to demonstrate high levels of scalability in relation to accuracy and time. It was noted that even with the growth of data volume, the processing time indicated very less variance. Authors in [14] suggested a system based on an SVM alongside with a rule-based classifier in order to enhance system accuracy. Authors in [2] suggested preprocessing of data to ignore noise and implemented sentiment analysis for movie dataset, on Hadoop framework and analyzed with agreat number of tweets. Yan et al. in [4]

proposed a microblog sentiment classification approach with parallel support vector machine (SVM) technique and Spark is utilized to improve the performance.

4. CONCLUSION

Data analytics (Sentiment Analysis) became a major part while taking major decisions in any organization, hence in this paper we discussed how to calculate sentiment by using advanced technologies such as kafka streaming and ksql with confluent platform. On top of that, we are not using extra service called storm, that increases the speed of overall analysis. Twitter Data in the form of reviews, thoughts, opinion,comments, feedback, and grievance are treated as big data and it cannot be interpreted directly; it should be preprocessed in order to be suitable for mining tasks. In this research, we propose an efficient sentiment prediction technique, utilizing the Apache Spark's Machine Learning library to execute different classification algorithms. The results indicate a significant enhancement in the accuracy of Naive Bayes and Logistic Regression with respect to increasing the volume of dataset, while the improvement is not strong in Decision Trees, also, experiment's results conclude that there is an inverse proportional relation between running time and the number of machines in the Spark Cluster, So in case of adding extra nodes in the cluster, higher performance capability will be obtained. From the former outcomes, our system can be described aseffective and scalable.

REFERENCES

[1] Parveen, Huma, and Shikha Pandey. "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm." Applied and Theoretical Computing and Communication Technology (iCATccT), 2016 2nd International Conference on. IEEE, 2016.

[2] Baltas, Alexandros, Andreas Kanavos, and Athanasios K. Tsakalidis."An apache spark implementation for sentiment analysis on twitter data." International Workshop of Algorithmic Aspects of Cloud Computing. Springer, Cham, 2016.

[3] Yan, Bo, et al. "Microblog Sentiment Classification Using Parallel SVM in Apache Spark." Big Data (BigData Congress), 2017 IEEE International Congress on. IEEE, 2017.

[4] Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf.Retrieval 2(1–2), 1–135 (2008) .

[5] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams Engineering Journal 5.4 (2014): 1093-1113.

[6] Yang, Peng, and Yunfang Chen. "A survey on sentiment analysis by using machine learning methods." Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2017 IEEE 2nd Information. IEEE, 2017.

[7] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?:sentiment classification using machine learning techniques. "Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[8] Wang, Hao, et al. "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle." Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012.

[9] Geetika Gautam, Divakar Yadav. (2014). Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis. IEEE 2014.

[10] Neethu M S, Rajasree R. Sentiment Analysis in Twitter using Machine Learning Techniques. IEEE 2013.

[11] W. Yang, X. Liu, L. Zhang, and L. T. Yang. Big Data realtime processing based on Storm. In 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, pages 1784–1787. IEEE, 2013.

[12] Datadog Engineering Blog. Monitoring Kafka performance metrics. 23 May 2016.

[13] A. R. Baig and H. Jabeen. Big Data analytics for behavior monitoring of students. Procedia Computer Science, 82:43– 48, 2016.

[14] M. T. Jones. Process real-time Big Data with twitter Storm. IBM Technical Library, 2013.

[15] V. Ta, C. Liu, G. Wandile. Big Data Stream Computing in Healthcare Real-Time Analytics. IEEE International Shahrivari. Beyond batch processing: towards real-time and streaming Big Data. Computers, 3(4):117–129, 2014.