RESEARCH ARTICLE OPEN ACCESS

A Review paper on Big Data Analytics of Geo-Social Media. #Earthquake

Yogesh Sharma [1], Aastha Jaie [2], Heena Garg [3], Sagar Kumar [4]
Assistant Professor, CSE Btech(CSE), Btech(CSE), Btech(CSE)
Department of Computer Science and Engineering
Maharaja Agrasen Institute of Technology, Rohini, New Delhi
GGSIPU, New Delhi, India

ABSTRACT
Geosocial Network data can be a great asset for making real-time decisions and future planning by analyzing geosocial media posts. There are millions of Geosocial Network users who are producing overwhelming data, called " Big Data" which is challenging to be analyzed and make real-time decisions. Hence, here, we put forward an efficient system for traversing Geosocial Networks while harvesting data and the user's location information.
A system architecture is proposed that processes an abundant amount of various social networks' data to check various Earth events, diseases, user trends, and views to make future real-time decisions and facilitate future planning.
The proposed system consists of five layers: Data collection, Data processing, Application, Communication, Data Storage.
Twitter is analyzed using the proposed architecture in order to identify current events or disasters, such as earthquakes, fires, Ebola virus, and snow. The system is calculated with respect to efficiency (correctness) while considering system throughput. We showed that the system has higher throughput and is capable of analyzing massive Geosocial Network data at real-time.
Keywords: - Geosocial Network; Big Data; Hadoop; Twitter Analysis; Analytics

I. INTRODUCTION

Social media is drastically advancing their feature day by day while making themselves from social networks to Geosocial networks. It gives a user the power to make their content public along with their geographical information, which has resulted in an increase in the use of Geosocial Networks. The information shared on any media is geosocial because 1) the posts have extensive content that represents geographical information with specific locations that are either entered explicitly or added implicitly (by Earth coordinates, such as latitude or altitude), and 2) the views shared on social media reveal social knowledge and strengthen relationship and communication.

Microblogging websites have evolved to become a source of varied kind of information. This is due to nature of microblogs on which people post real-time messages about their opinions on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in daily life. In fact, companies manufacturing such products have started to poll these microblogs to get a sense of general sentiment for their product. Many times these companies study user reactions and reply to users on microblogs. One challenge is to build technology to detect and summarize an overall sentiment.

Big Data: Definition

Big data is a term that refers to data sets or combinations of data sets whose size (volume), complexity (variability), and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases and desktop statistics or visualization packages, within the time necessary to make them useful. Figure No. 1 gives Layered Architecture of Big Data System. It can be decomposed into three layers, including Infrastructure Layer, Computing Layer, and Application Layer from top to

bottom.

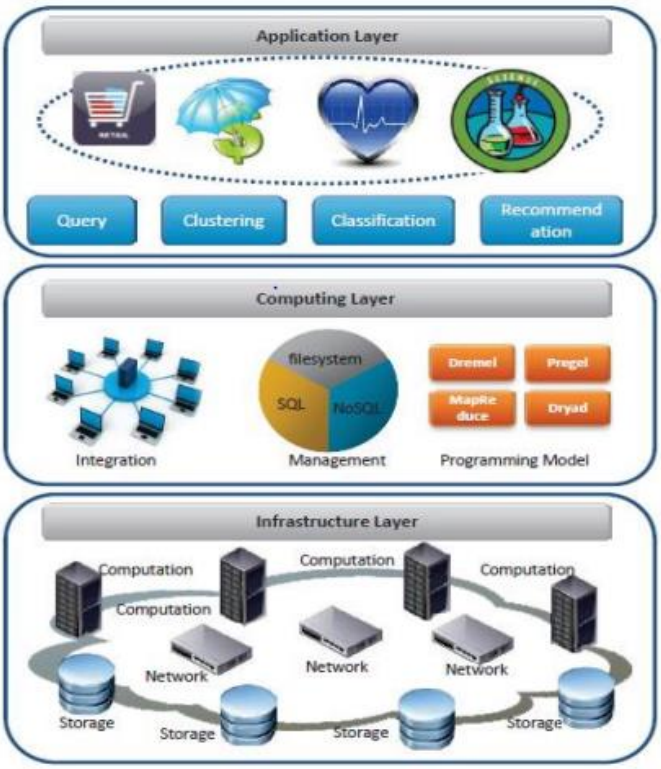


Figure 1: Layered Architecture of Big Data System

3 Vs of Big Data:

Volume of data: Volume refers to the amount of data. The volume of data stored in enterprise repositories has grown from megabytes and gigabytes to petabytes.

Variety of data: Different types of data and sources of data. Data variety exploded from structured and legacy data stored in enterprise repositories to unstructured, semi-structured, audio, video, XML etc.

Velocity of data: Velocity refers to the speed of data processing. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.

Problem with Big Data Processing

i. Heterogeneity and Incompleteness

When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data require further work.

ii. Scale

Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, following Moore's law, to provide us with the resources needed to cope with increasing volumes of data. But, there is a fundamental shift underway now: data volume is scaling faster than compute resources, and CPU speeds are static.

iii. Timeliness

The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when one speaks of Velocity in the context of Big Data. Rather, there is an acquisition rate challenge

iv. Privacy

The privacy of data is another huge concern and one that increases in the context of Big Data. There is great public fear regarding the inappropriate use of personal data, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

v. Human Collaboration

In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Ideally, analytics for Big Data will not be all computational rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modeling and analysis phase in the pipeline. In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from

multiple human experts, and shared the exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input and support their collaboration

2. Hadoop: Solution for Big Data Processing

Hadoop is a Programming framework used to support the processing of large data sets in a distributed computing environment. Hadoop was developed by Google's MapReduce that is a software framework where an application breaks down into various parts. The Current Apache Hadoop ecosystem consists of the Hadoop Kernel, MapReduce, HDFS and numbers of various components like Apache Hive, Base and Zookeeper. HDFS and MapReduce are explained in the following points.

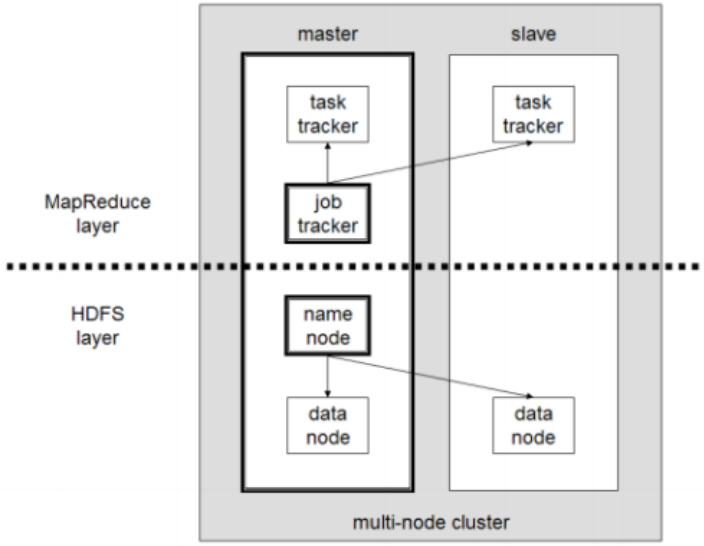


Figure 2: Hadoop Architecture

HDFS Architecture

Hadoop includes a fault-tolerant storage system called the Hadoop Distributed File System, or HDFS. HDFS is able to store huge amounts of information, scale up incrementally and survive the failure of significant parts of the storage infrastructure without losing data. Hadoop creates clusters of machines and coordinates work among them. Clusters can be built with inexpensive computers. If one fails, Hadoop continues to operate the cluster without losing data or interrupting work, by shifting work to the remaining machines in the cluster. HDFS manages storage on the cluster by breaking incoming files into pieces, called "blocks," and storing each of the blocks redundantly across the pool of servers. In the common case, HDFS stores three complete copies of each file by copying each piece to three different servers.

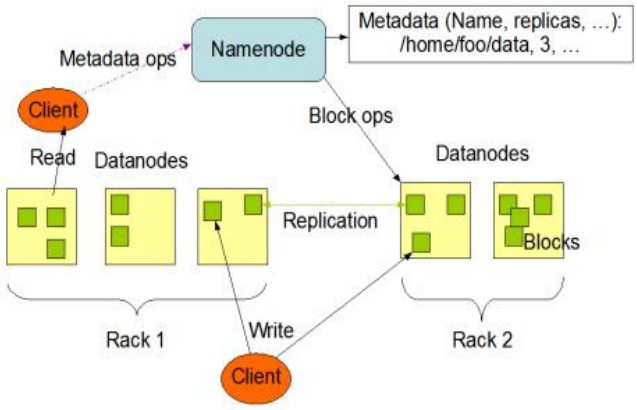


Figure 3: HDFS Architecture

MapReduce Architecture

The processing pillar in the Hadoop ecosystem is the MapReduce framework. The framework allows the specification of an operation to be applied to a huge data set, divide the problem and data, and run it in parallel. From an analyst's point of view, this can occur in multiple dimensions. For example, a very large dataset can be reduced into a smaller subset where analytics can be applied. In a traditional data warehousing scenario, this might entail applying an ETL operation on the data to produce something usable by the analyst. In Hadoop, these kinds of operations are written as MapReduce jobs in Java. There are a number of higher-level languages like Hive and Pig that make writing these programs easier.

The outputs of these jobs can be written back to either HDFS or placed in a traditional data warehouse. There are two functions in MapReduce as follows:

map – the function takes key/value pairs as input and generates an intermediate set of key/value pairs
reduce – the function which merges all the intermediate values associated with the same intermediate key

II. PROPOSED COMPUTING MODEL

This section describes the proposed model, including system overview, proposed architecture, application, and limitations.

System Overview

The proposed system analyzes all the tweets generated by Twitter users. All such information corresponds to user locations and can provide real-time monitoring of disasters, fatal diseases, or accidents. Moreover, user data and their location information can be used to recommend various systems based on the user's current location, such as useful products, restaurants, hotels, and transportation.

Our system harvests user data from Geosocial Networks as a tuple (l, t, u) of location l, time t, and user u. The tuple describes, what user u posted on a Geosocial Network at time t from location l. Obtaining user location is also a challenging task that has become easier because of advanced remote smart devices and GPS systems. Smart devices send location coordinates by default to the network while the user posts any activity or event. The location indicates the particular area in order to determine the events or activities that are occurring at different time variants based on the Geosocial Network being used. Such location is attached with the post, and it is used to find the authenticity and relevance of the activity or event that corresponds to the location.

The major challenge is how to provide analysis for such a huge amount of data generated by many social networks. The proposed system uses advanced computing technologies to meet the challenge.

A detailed description of the proposed system, including the proposed architecture, implementation model, application, issues, and challenges are given in the following sections.

Proposed System Architecture

The system has three basic top to bottom layers, i.e., Data Collection, Data Processing, and Services and Application. Two additional layers work side by side with the basic layers. Such additional layers provide communication and storage for raw and structured data. The communication layer provides internal communication between various servers through different communication technologies, such as Wi-Fi and Ethernet, as well as external communication to the Geosocial Network servers for data harvesting using any fast Internet technology, such as WiMAX, 3G, and LTE. The storage layer manages services for storing data that can be structured, unstructured, or raw for future use or planning. Instead of working with all three basic layers, as is done by the communication layer, the storage layer only works side by side with the last two basic layers, i.e., data processing and services and application.

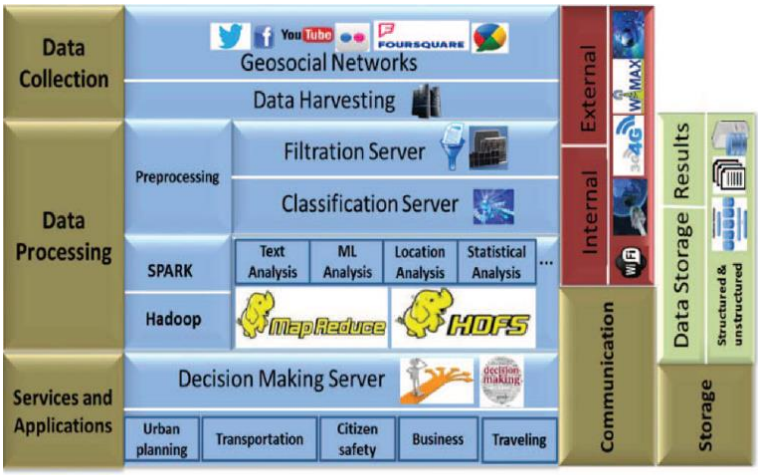


Fig. 2. Proposed system architecture

The data are stored in the database after classification, and the results are stored after analysis and decision-making. Because the data sources is Geosocial Network, such as Twitter, the data collection layer is responsible for the data harvesting server that obtains data or content from Geosocial Network servers.

Data harvesting is one of the main and challenging tasks of the system because data are captured and harvested at high speed in real time. We can harvest data that is publically available to anyone by the social networks. However, this method cannot guarantee to harvest the real-time data for all the social networks. We have extracted Twitter Tweets using Flume tool which is used for capturing streaming data. The query response is JSON format. In the case of Twitter, two APIs are provided to obtain tweeter data, i.e., 1) Streaming, used to continuously obtain real- time data without delay by giving the topic/content, location, and maximum size limit and 2) RESTful, which can work similarly to the state transfer API, and is used to obtain data in discrete delays, such as after 5 minutes, 1 hour, or 1 data [28]. In both cases, the received data contain the actual data and metadata.

The data processing layer has two subparts:

1) preprocessing and computations and 2) analysis. This layer is responsible for all types of complex processing, analysis, and results generation. Because the system manages a significant amount of social network data, it is not possible to process each and every byte, regardless of its usefulness. Thus, filtration is performed in order to reduce

data size by discarding unnecessary data and decrease computational overhead. The unnecessary metadata, irrelevant posts with respect to content, area, and time, are discarded. At the preprocessing level, classification is performed in order to further reduce computational overhead by forming an organized structure for data, which makes data access very fast. The data are classified based on location, time, and content. Various content types are already notified, such as earthquakes, fires, and Ebola virus, based on the requirements for classifying the data. The next phase of data processing is the analysis of the classified data.

Given that such analysis requires significant data computation, we need a powerful hardware and software system with the ability to manage such large amount of social network data. Thus, we deployed a Hadoop ecosystem with a strong distributed file system (HDFS) that can store data across multiple nodes with higher security, reliability, and fast access. The Hadoop ecosystem can perform parallel processing on the same data stored on HDFS nodes using its parallel programming paradigm, i.e., Map Reduce. Hadoop was initially developed for batch processing, but here we need real-time processing on continuously incoming geosocial data. Various machine learning algorithms, statistical analysis, text analytics, and content-based analysis are performed on the processing layer to generate results for decision-making. In our case, we simply provide basic analysis to determine various disaster events, such as earthquakes, fires, and diseases.

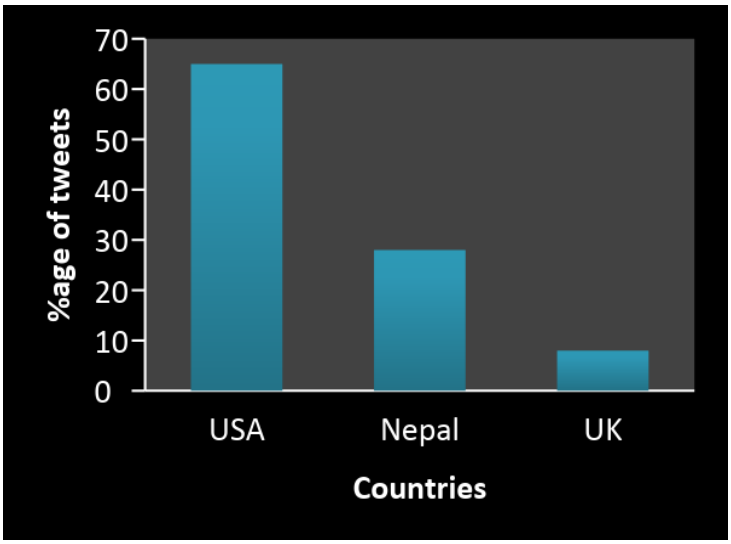
Finally, we have the service and application layer that is responsible for decision-making and using the results for various applications based on needs. For example, in the case of earthquake detection in various areas, Hadoop provides all the IEEE ICC 2017 SAC Symposium Big Data Networking Track computation and statistical analyses of the tweets that contain earthquake information at time t. Hadoop can only provide the results, which can be manipulated by the decision server. The decision server then identifies the location based on the results, such as time, how many tweets are received, and authenticity of the results. The decision server determines whether there is an earthquake or any other event at a particular location. Later, these results can be used for many applications, such as urban planning, transportation, citizen safety, business, and traveling.

Analysis and discussion

For Analysis of a really big data, it takes high processor, Ubuntu 14.04 LTS core TMI 5 machine with 3.2 GHz,* four processor of 4GB RAM. The offline data can be available from various APIs provided. But we have used Hadoop ecosystem which has HDFS(Hadoop Distributed File System) making it more powerful and effective. In this paper, we have taken streams of data from twitter handle where humans are treated as a sensor system. Humans use twitter In most of the cases to express various events, news, sentiments etc. It has been observed that social media covers an event even more than a news channel. In less than a minute of a disaster or anything else, tweets regarding this can be seen.

In this paper, we have used the hashtag convention. The hashtag is used generally in tweets to easily address the issue to someone who needs it or it also helps to spread awareness among the people near the location.

So we have taken the #Earthquake or #EARTHQUAKE. In this paper, we have analyzed two things-Time between Earthquake occurrence and Tweets, Second is Originating locations of Earthquake-reporting tweets at distinct time intervals. For this, we need to know the geolocations of the tweets.Through the graph, we can see the %age of tweets in various countries.



The below graph shows the adjusted times between event occurrence and tweets. It gives the analysis of adjusted times between event occurrence and tweets. The Data is taken from the Earthquake in the US on 23 March 2011. It was a dangerous earthquake covered almost half of the east part of the United States. Thus it got a lot number of tweets.

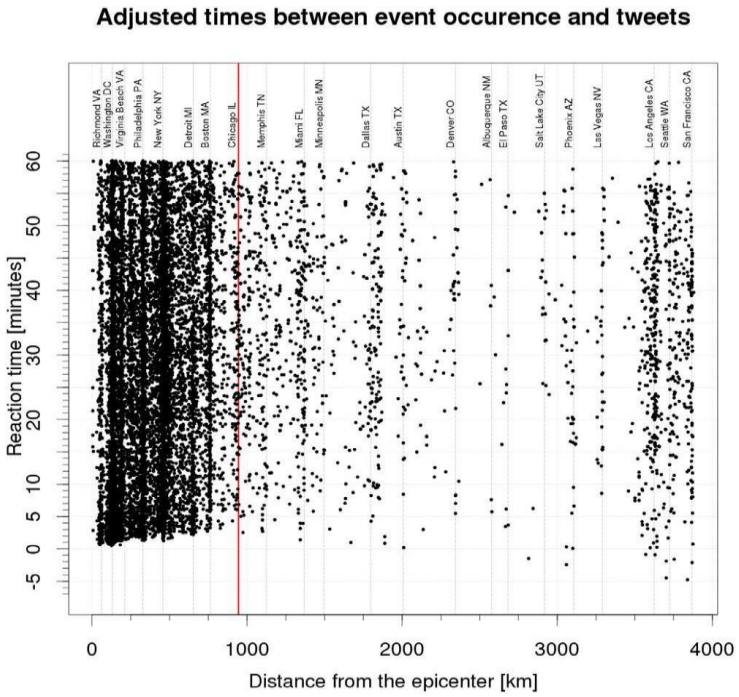


Fig.2 Adjusted times between event occurrence and tweets

When we talk about Hadoop over other systems, we focus more on Efficiency. Therefore, a system is evaluated w.r.t throughput given due to dataset size. The good thing about our system is its throughput increases as data size increases. When Data size increases, the major number of switches between map and reduce function reduces and thereby giving us a good throughput.

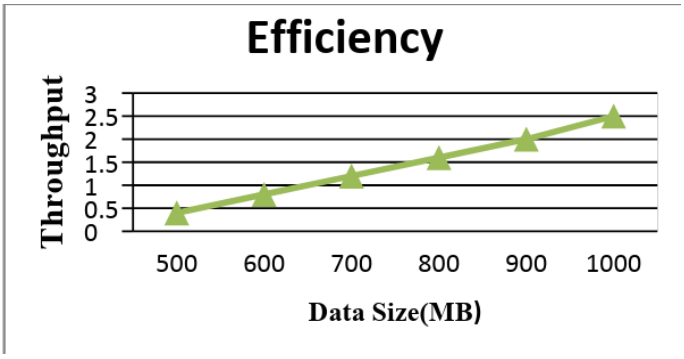


Fig. The throughput of the dataset based on the data set.

III. CONCLUSION

Geosocial network has proved itself as an asset to society. Social Media feeds are rapidly emerging as a social avenue for dissemination of geographical information as the social media content generally includes the location and the various keywords. The objective of this project was to paper is to assess the quality of the tweets from this really huge amount of data. This further helps to increase advanced situational awareness. We were mainly interested in the analysis the twitter feeds and treating human as a sensor to get the information as early as possible. And the results through this we get is really interesting. In this paper, we proposed a system in which data from Twitter is collected through Hadoop framework, which cleans or filters the data provided and then analyzing those tweets which prove us the power of the geosocial network.

REFERENCES

[1] M.Mazhar Rathor, Anand paule, Awais Ahmed, “Big data Analytics of geosocial data” The school of CSE, Korea pp 3-5,2017.

[2] Andrew crooks, '#Earthquake: Twitter as a Distributed Sensor System' pp 129-140, 2014-15.

[3] Jeiming Shi, Nikos Mamoukis "Density-Based Clustering in Geo-Social networks" , The University of hong-kong, pp 2-9, 2000

[4] Zhao S, Zhong L, Wickramasuriya J, and Vasudevan V 2011 Human as real-time sensors of social and physical events.

[5] Hern, “Online volunteers map Philippines after typhoon Haiyan,” The Guardian, November 15, 2013.

[6] Andrew crooks ” Data of Earthquake in US on 23rd March, 2011” pp 7-20

[7] Yyes van Gennip, Huiy hu, Blake hunter“Geosocial Graph Based Community Detection”, Department of Mathematics, University of California, Los Angles, USA.