BIG DATA ANALYSIS ON TWITTER USING HADOOP AND MAPREDUCE

[1]Navyashree S, [2]A .V. Krishna Mohan
[1]PG Student, Dept of CSE, SIT, Tumakuru, Karnataka, India
[2] Asst. Professor, Dept of CSE, SIT, Tumakuru, Karnataka, India

Abstract : Geosocial network information would be aided in the form of boon for the government to give decisions in real-time and future outlining for the event happens of the real world like, the user can able to take decisions for the real-time by subsequently analyzing the information straightway for the future which are the areas that will be concerned by the particular events. Thereby, in this paper, a system framework is proffered that processes a rich quantity of diversified social network data to supervise phenomenon, earth events, medical morbidities, etc., by reaping details as well as user's site information in a secure manner. Twitter is examined using the intended framework to identify events like Earthquake, Bomb blast, Fire, etc. Hadoop ecosystem is used to analyze big data propagating from the various sensors. ECC algorithm is applied to secure the sensor data.

IndexTerms **- Geosocial Network, Big Data, Hadoop, MapReduce, ECC.**

I. INTRODUCTION

Each day, many people communicate with each other through social media i.e. people live through social media where they send messages on websites like Instagram, Facebook, Twitter, and etc. The public network has drastically substituted the habit of people convey and promoting their quality on a daily basis building itself from social media to geosocial networks. As the technology is advancing daily that allows the smart phones to use GPS systems, that made location data more effective and it allows the user to make their text general along with the geodetic report.

The data delivered in any media is geosocial owing to the fact that,

1. The messages hold immense substance which depicted spatial data beside particular areas which are neither recorded ambiguously nor unambiguously.
2. Social information is made known by the perspectives shared via social media which strengthen relationship and communication.

Geosocial network information would be used as strength in favour of the government to take decisions in real time and planning in the future by examining geosocial broadcasting posts. Making use of geosocial network information is not only favourable to authorities, more it can too affect individual existence. Geosocial network facts are able to come up with assists to standard nationals to businessmen. Despite that when gathering geosocial information via social media forums such as Facebook, Skype, Whatsapp, Twitter, etc., it must be emphasized that these systems hold plenty of end users who put lots of messages. Twitter, for instance, is one of the leading provenances of real-time evidence, it's been utilized by many of internet users nowhere in the world where people share their views, feelings in brief messages called tweets. Any case something would happen over the globe, netizens begin to speak about that over twitter. People who want to find for specific places, incidents, or issues become lenient because of the presence of the hashtags in tweets. A number of twitter end-users commonly involve in exchange information by putting tweets and retweeting others tweets successively recognizes the top vogue in the twitter world.

Thence, it could be clearly considered that all the members of different popular networks spawning formidable quantity of data called "BIG DATA"; that kind of data might reach in terabytes. Therefore, collecting and analyzing that kind of geo-social information which is on the air is an extremely conflicting job. Hence, it requires a peculiar data-processing setting and progressive computational mechanisms along with sapient discipline so as to issue in-time or real-time analysis. Thereby, with an eye to addressing these data-processing challenges, in this paper innovative geosocial evidence assayed is proposed that not merely makes fragmentary data adroitly inside a duration limit besides providing persistent information probe to various mutual enterprise, deeming Twitter, Flicker, Facebook, YouTube, and so on.

A. OBJECTIVE

The main motivation of this paper is to analyze huge quantities of data that are generated from geosocial network. For that, a secure sensor data is to be generated for generating of the specific events like Fire, Bomb blast, Earthquake so that the user can quest for peculiar events by fetching the secure data to get secure information.

B. PROBLEM STATEMENT

There are number of Geosocial Network members be producing staggering of information, called "Big Data" which is difficult to be store, process and analyze to take decisions in real-time using traditional approach and that data is not stored in a secure way, any unauthorized user can access these data. To solve storage and processing problem Hdfs and MapReduce are used and to provide security for the system ECC algorithm is used to secure the data.

II. BIG DATA

A. CHARACTERISTICS

- **Velocity**: It indicates the speed of fresh information developed and deployed.
- **Volume**: It indicates the quantity of information that is generated and stored.
- **Variety:** It indicates the various sources and kinds of information i.e. Structured and unstructured data.
- **Veracity:** It refers to the noise, biases, and deformity in the information.
- **Volatility:** It refers to how long is information valid and how long does that information need to be stored.

B. TYPES

Structured Data

- Commonly resides in relational databases (RDBMS).
- Data has structure and is highly organized in the form of tables.
- It is easier to upload, query, search, analyze.
- **E.g.:** Spreadsheet.

Semi-Structured Data

- It is primarily structured data that is unorganized.
- The data doesn't abide in fixed fields or records.
- Difficult to store, analyze and retrieve.
- **E.g.: JSON, XML, CSV.**

Unstructured Data

- Data cannot be stored in rows and columns.
- No-specific pattern and not organized.
- Difficult to query and retrieve by applications.
- **E.g.:** Audio, video, WebPages.

III. HADOOP MAPREDUCE IMPLEMENTATION

It is an open-source distributed processing framework that allows storing Big Data in a distributed setting and process utilizing the MapReduce programming model. The core components of Hadoop are HDFS and MapReduce.

A. HDFS (Hadoop Distributed File System)

HDFS is a distributed file system schemed to stock and manages huge files which are of GB/TB in size in an efficient manner. HDFS is highly faulting tolerant and designed using low-cost hardware. Each file is stored on HDFS as blocks. HDFS follows the master-slave framework and it has the following components.

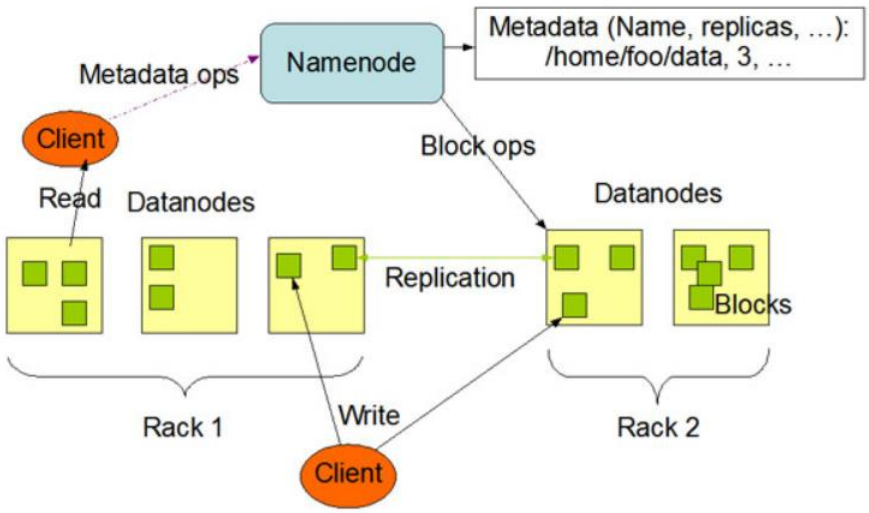

Fig.1. HDFS Architecture

Namenode

- It contains the information that data is stored on which data node.
- Maintains and Manages data node.
- Records metadata.

Datanode

- It stores the data that are sent to the Hadoop cluster.
- Stores actual data.
- Serves read and write calls from the client.

B. MAPREDUCE

Map-Reduce introduced by Google for the purpose of process and store huge datasets in clusters and it is the core element of the Apache Hadoop software framework. MapReduce is a processing procedure and a programming model for distributed computing based on java. MapReduce framework mainly consists of two processing phases-Map and Reduce. In the Map job block of data is read and processed to produce key-value pairs as intermediate outputs and it is stored in the local file system. The output of the map job serves as input to the reducer and it receives the key-value pair from multiple map phases. Then, those intermediate key-value pair is aggregated in reduce phase into a smaller key-value-pairs which is the final output and it is stored in Hdfs.

Map Reduce Components are:

1. **Name Node:** Executes large databases of HDFS, doesn't tackle with records explicitly.
2. **Data Node:** It caches blocks of HDFS.
3. **Job Tracker:** It records, splits and tracks job slaying on slaves i.e. Task Trackers.
4. **Task Tracker:** It organizes Map-Reduce functions.

IV. LITERATURE SURVEY

Carlo Ratti et al [1] as the technology for formative the geographic areas of cellular phones and other hand-held gadgets is seemly even more available and it is paving the way to a broad range of applications which are collectively known as Location Based Services (LBS). This paper aims to study and present the possibility of the LBS technology to the metropolitan planning communal. In addition, it unveils the 'Mobile Landscapes' design: an application in the urban area of Milan, Italy, predicated on the geographical mapping of cellular phone practice at various periods of the day. The outcomes enable a pictorial representation of the severity of urban actions and their development through time and place. In the end, a number of future applications are addressed and their ability for urban surveys and planning is evaluated.

Anand Paul et al [2] designed the essential role of incorporating IoT with social networking in fulfilling human dynamics on the basis of big data analytics. Through careful investigation of incorporation and interaction of people in the emerging smart cities, they have introduced the idea of Smartbuddy and proved its range of application through a Hadoop ecosystem. Smart buddy thoroughly overthrows the difficulty of understanding by offering a response to users that tender them the opportunity to enhance their behavior using taxonomy of the attentive note.

Crooks et al [3] access a procedure for real-time networking sustains are speedily evolving as an emerging road for the promise and scattering of data that is regularly territorial. Their contentment frequently incorporates notes to cases taking place at, or influencing particular areas. Within this article, they examine the extensional and transient features of the twitter feed movements answering to the earthquake of 5.8 magnitudes which happened in the United States (US) on August 23, 2011. They dispute that these feeds portray a hybrid format of a sensor device which permits in order to classify and tracking of the affected location of the incident. By conflicting this with similar content assembled through the consecrated collaboration 'Did You Feel It?' (DYFI) website of the U.S. Geological survey this paper assesses the feasible offed employ of reaped social networks contentment for incident auditing. The trials assist the opinion that the user serves as sensors to provide us equivalent outcomes within a prompt method and be able to supplement further origins of information in order to improve our circumstantial perception and improve our awareness and reaction to such events.

V. PROPOSED SYSTEM

The proposed system consists of a sensor that generates a sensor data of a particular event related to disasters and it senses the data in which particular time and area the event occurred, the information can be encrypted by using a particular sensor public key. Only the authorized user can search in the tweets where the specific event happened. By using the specific private key the user can allow decrypting information which can be used to make future planning and real-time decisions. ECC is public key cryptography which is utilized for encrypting and decrypting the information presented in the geo-social network. The information presented in the geo-social network is processed using the MapReduce algorithm.

A. ARCHITECTURE DIAGRAM

- **Data collection:** In data collecting time, to create the fussy event related to the disasters akin Earthquake, Bomb blast, Fire the sensor ought to sense the information in which distinct locations, time, and year the event occurred.
- **Upload data**: The data generated by the sensor are sent to the geosocial network in the encrypted format where it will be stored. The ECC algorithm is used for encryption.
- **Data storage:** the sensor data will be stored in the geo-social network.
- **User:** The user must log in to the framework thereby user could probe for the specific information for which the event has happened. After login, the user could probe for the data by hash keyword and in year range. In which uncommon year the event has happened and in which location-specific event has occurred the user could decrypt the information.
- **Map Reduce:** After decrypting the data, the MapReduce algorithm is applied to obtain the count of every certain event in which countries intervened and the result will be displayed in the sort of graph.

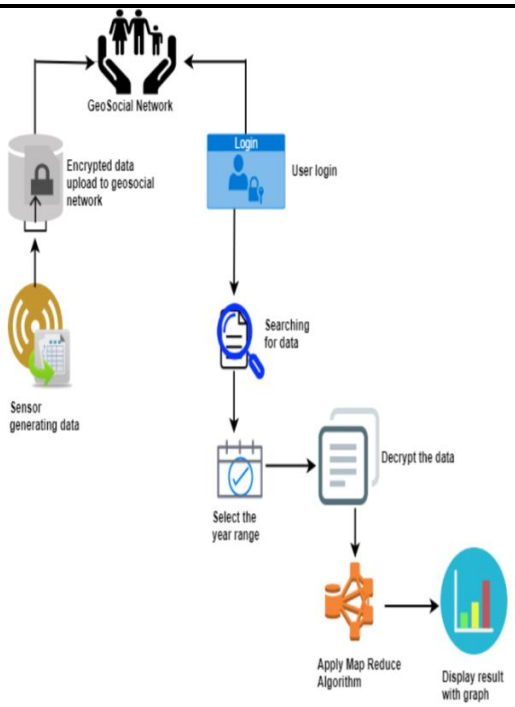

Fig.2. shows the overall architecture of the proposed system

VI. ALGORITHM

A. WORD COUNT MAPREDUCE ALGORITHM

Word count is one of the critical functions in big data world findings its needs in search technologies (TF/IDF), recommendation engines, etc. This algorithm counts the variety of times every word occurs in a file.

Hadoop Word Count Map Reduce operation job occurs in 3 stages –

- **Mapper Phase**

In the mapper phase <Key, Value> pair will be generated from the input text file by tokenizing the text into words. For instance, in this paper <Fire, 1> pair is generated where the key is a unique word and value is one.

- **Shuffle and Sort Phase**

It is carried out within by the Hadoop framework. Here, the output of the mapper phase i.e. <Key, Value> pair is taken as input and then sorted in order i.e. is categorized by the key and sent to the reducer class. For instance, the <Key, Value > pairs <Fire, [1, 1]> to the reducer class.

- **Reducer Phase**

All the values are aggregated in the reducer class which is sent from the shuffle and sort phase. All the keys are classified with each other and the identical key values are summed up to find the appearances for a certain word and generate another <Key, Value> pair. For instance <Fire, 2>.

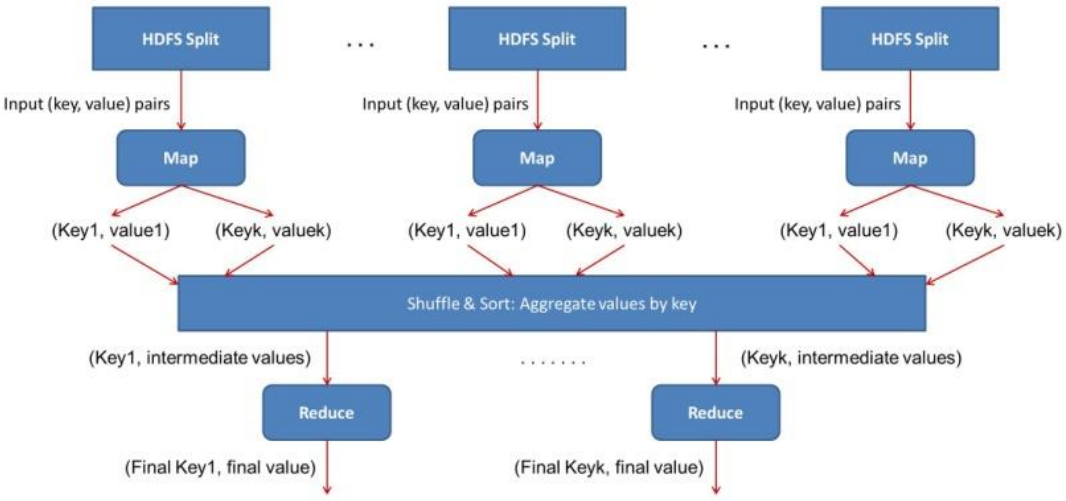

Fig.3. MapReduce algorithm

B. ECC (ELLIPTIC CURVE CRYPTOGRAPHY)

ECC (elliptic curve cryptography technique will be used for encryption and decryption of data.
The following symbols are utilized,

- E- Elliptic curve
- P- Point on the curve
- n- Maximum limit (prime number)

➢ **Generation of keys**

Keys will be utilized for encryption and decryption. Here the public key is utilized for encryption and private key for decryption. A number‘s’ has to choose within a range of ‘n’. The public key will be generated using the following formula

- W=s*p
- Where s= random number selected within the range(1 to n-1)
- P is a point on the curve.
- ‘W’ is public key and‘s' is a private key.

➢ **Encryption**

Assume ‘x' is the data sensed by the sensor sent to the Geo-social network. Represent this data on a curve. Consider’x’ as point ‘M’ on the curve ‘E’. Randomly select ‘k’ from [1-(n-1)].Two cipher texts will be generated let be m1 and m2

- M1=k*p
- M2=M+k*W

➢ **Decryption**

The data have to decrypt which are sent by the sensor

X=M2-s*M1;

Where x is the original message.

1) Proof

X=M2-s*M1

‘x’ can represent as ‘M2-s*M1’

M2-s*M1=(x+k*Q-s*(K*P) M2=x+K*Q & M1=K*p)

=X+k*s*P-S*K*p (cancel k*s*p)

=x (original message)

VII. RESULT ANALYSIS

The Twitter dataset is used to analyze the proposed system which contains the distinct disaster data like Earthquakes, Bomb blast, and Fire. The system analyzed all sensor data. The analysis is carried out relying on the user searching for which particular event.

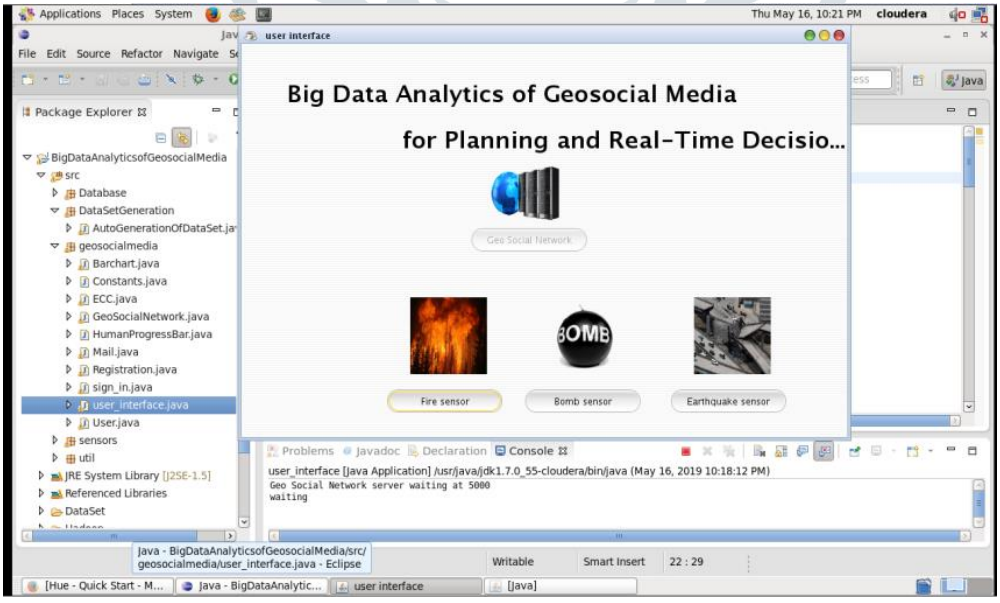

Fig.4. the snapshot shows the user interface and clicks on Geo Social Network button

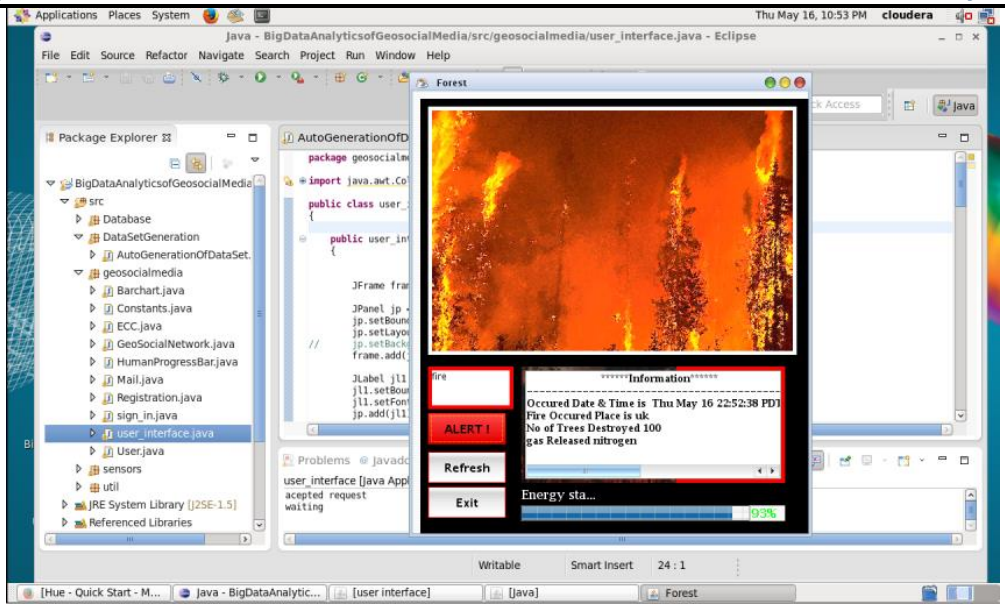

Fig.5. the snapshot shows the fire sensor sensing data.

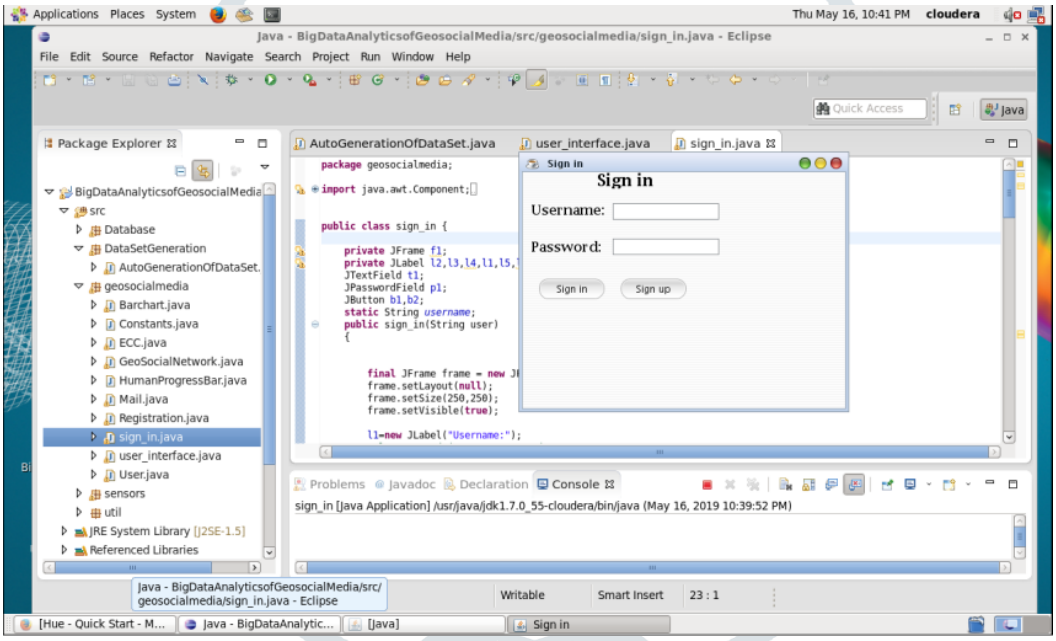

Fig.6. the snapshot shows the sign in.

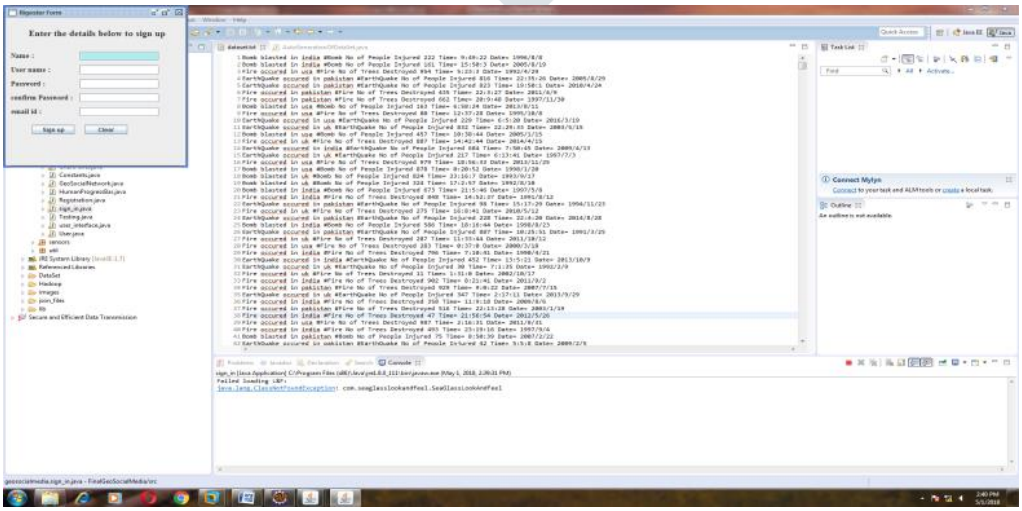

Fig.7. the snapshot shows user entering details to register

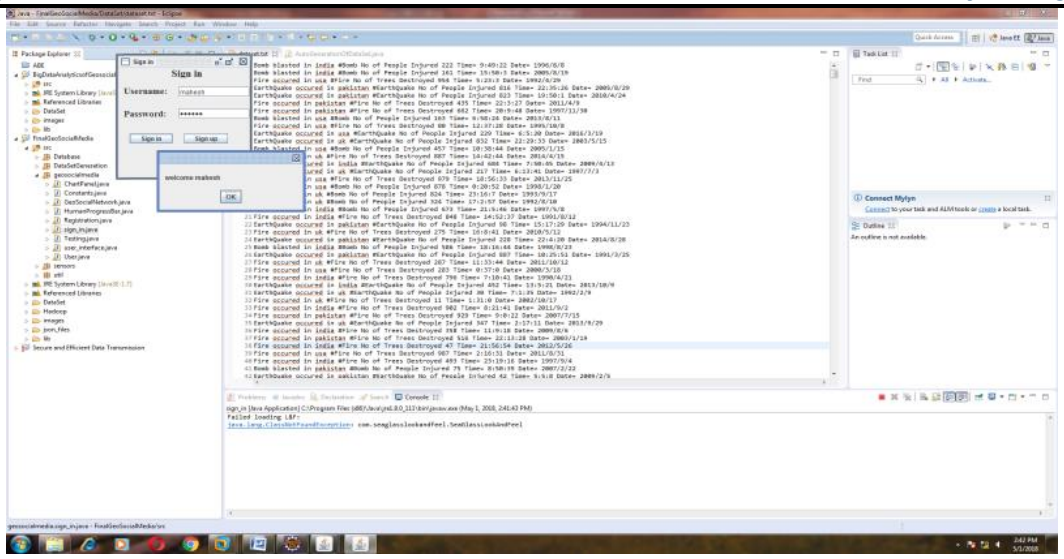

Fig.8. the snapshot shows login.

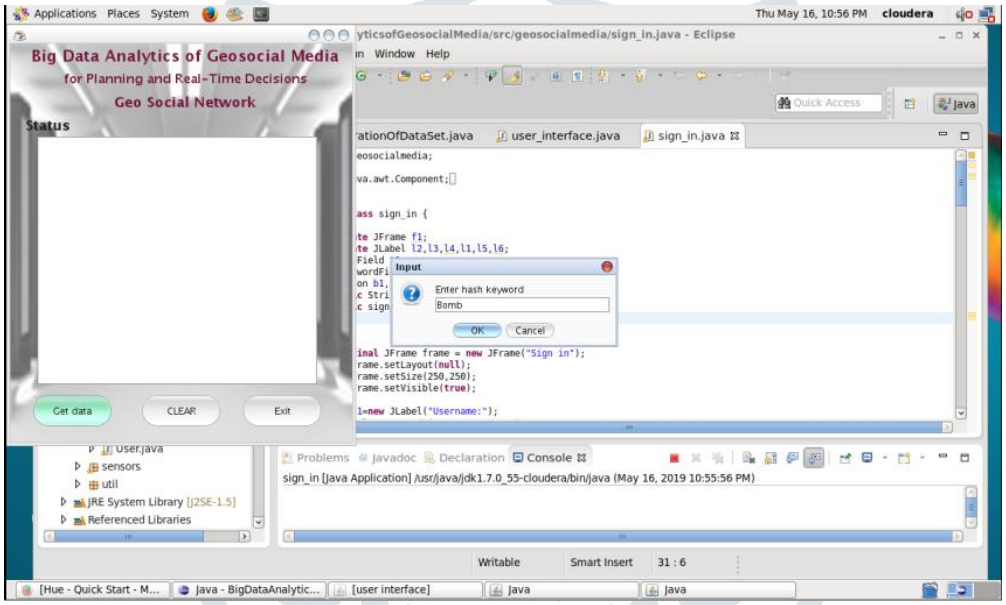

Fig.9. the snapshot shows the user has to enter the hash keyword

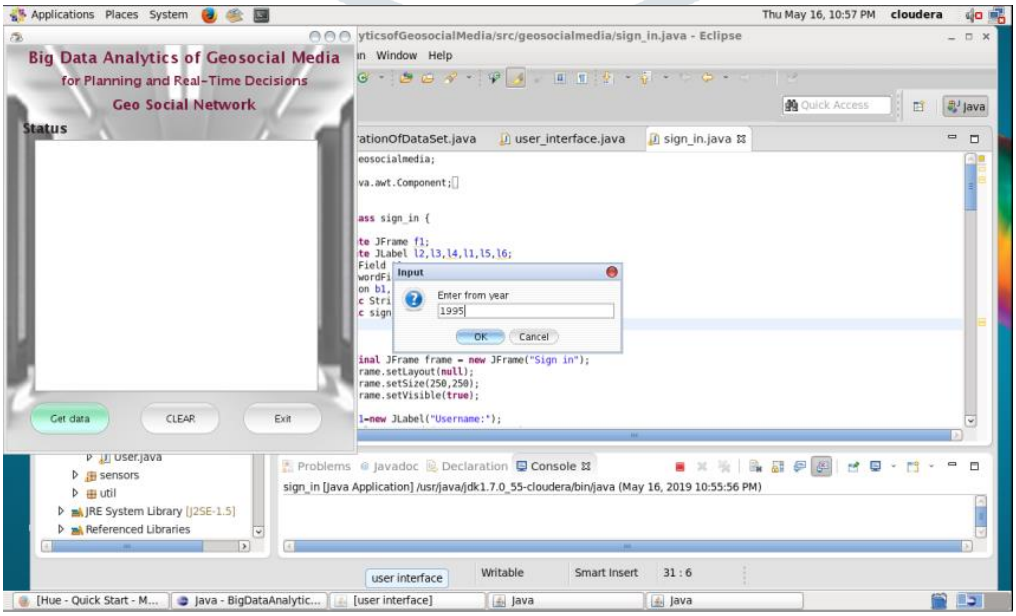

Fig.10. the snapshot shows the user has to enter from the year.

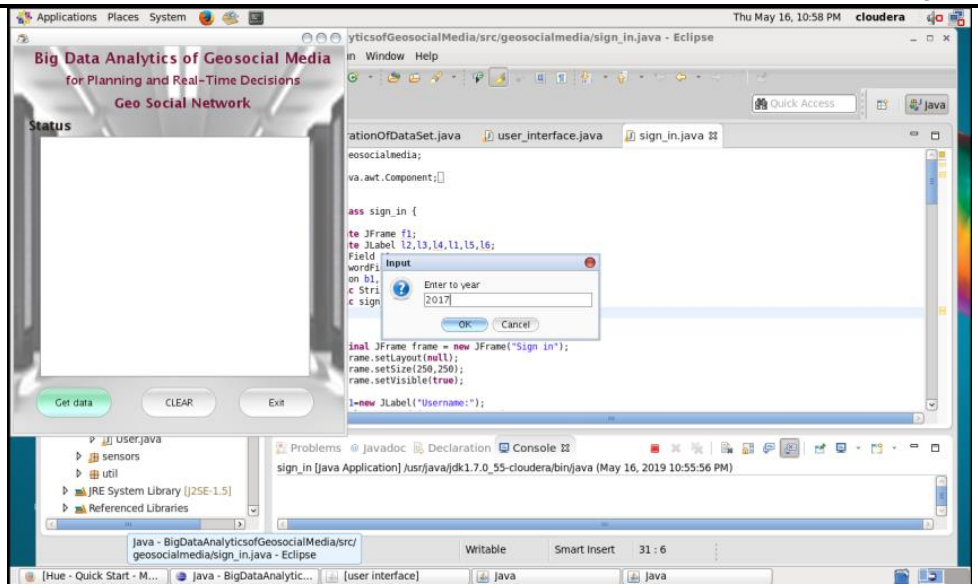

Fig.11. the snapshot shows the user has to enter to year.

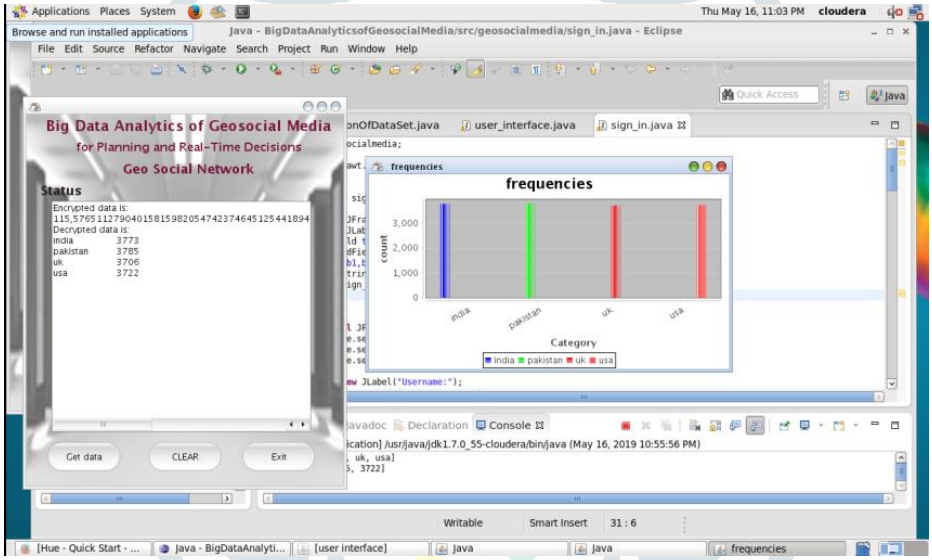

Fig.12. the status of the generated output.

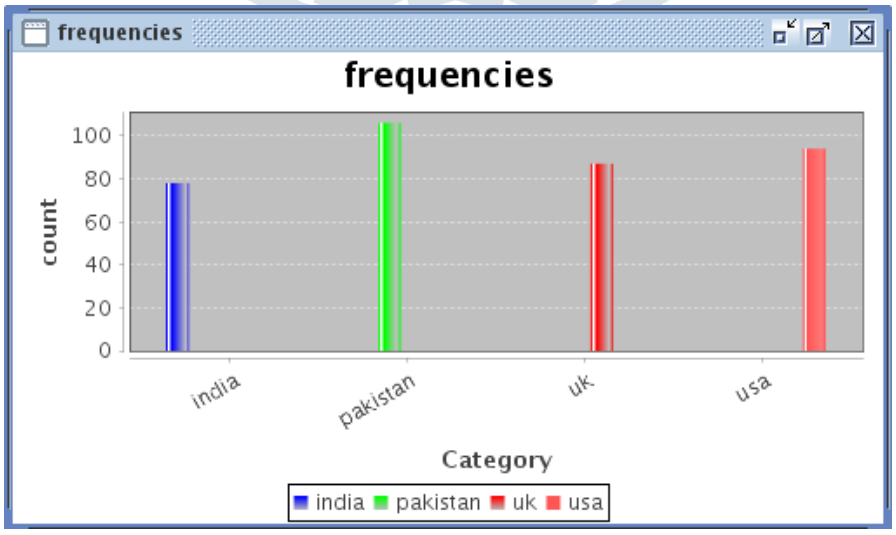

Fig.13. the snapshot shows the result with graph

VIII. CONCLUSION

Geosocial Networks could be an advantage in favour of authorities in the matter of offering proficiencies and offering from miseries using appropriate regulation and diminution of the frightened offing disperse of some diseases. Likewise, that kind of network could be profitable to normal people through affording commended order, transportation security, medical services, etc. and to catalysts because of the beginning latest commodity in different locations by controlling the geosocial information of a certain location. Despite that, such advantages could merely be originated from further analytics which engages a meaningful quantity of information produced from heterogeneous Geosocial Networks.

The issue is feasible along with progressed techniques and improved analysis, and a structure among major informatics masteries. Wherefore, in this paper, the suggested method utilizes geosocial information for improved preparation, protection out of catastrophes, and appropriate governance, perception, etc., on the basis of different geolocations. The architecture does not merely gather a huge quantity of information at great- velocity from Geosocial networks, but it can be able to handle, investigate, and take real-time decisions. This paper studied Twitter information about different events using the proposed technique. The technique was established utilizing a Hadoop framework and ECC algorithm is used for data security. The technique was more capable when handling a lot of datasets.

REFERENCES

[1] Ratti, S. Williams, D. Frenchman, R. Pulselli, "Mobile landscapes: using location data from cell phones for urban analysis", *Environment and Planning B Planning and Design*, vol. 33, no. 5, pp. 727, 2006.

[2] Paul Anand et al., "Smartbuddy: defining human behaviors using big data analytics in social internet of things", *IEEE Wireless Communications 23.5*, pp. 68-74, 2016.

[3] Crooks, A. Croitoru, A. Stefanidis, J. Radzikowski, "#Earthquake: Twitter as a distributed sensor system", *Transactions in GIS*, vol. 17, no. 1, pp. 124-147, 2012.

[4] M. Zook, M. Graham, T. Shelton, S. Gorman, "Volunteered geographic information and crowdsourcing disaster relief: A case study of the Haitian earthquake", *World Medical & Health Policy*, vol. 2, no. 2, pp. 7-33, 2010.

[5] Vinutha P "Dynamic Decision Making Framework for Geo-Social Media for Making Smart Plans on Big-Data", www.ijert.org Volume 6, no. 13, 2018

[6] M. Mazhar Rathore , Anand Paul , Awais Ahmad , Muhammad Imran and Mohsen Guizani "Big Data Analytics of Geosocial Media for Planning and Real-Time Decisions " 2017 IEEE International Conference on Communications (ICC) 21-25 May 2017

[7] **RAJDEEP PAUL "BIG DATA ANALYSIS OF INDIAN PREMIER LEAGUE USING HADOOP AND MAPREDUCE"**, 2017 INTERNATIONAL CONFERENCE ON COMPUTATIONAL INTELLIGENCE IN DATA SCIENCE(ICCIDS) **2-3 JUNE 2017**

[8] M. Zook, M. Graham, T. Shelton, S. Gorman, "Volunteered geographic information and crowdsourcing disaster relief: A case study of the Haitian earthquake", *World Medical & Health Policy*, vol. 2, no. 2, pp. 7-33, 2010.

[9] www.dezyre.com

[10] Stefanidis, A. Crooks, J. Radzikowski, "Harvesting ambient geospatial information from social media feeds", *GeoJournal*, pp. 1-20, 2012.