Big Data Prediction on Crime Detection

[1]Dev Naomi.G [2]Karthigaa.M [3]Keerthana.B [4]Janani A
[1,2,3]Student [4]Assistant Professor
[1,2,3,4]Department of Information Technology
[1,2,3,4]Loyola-ICAM College of Engineering and Technology
Chennai, India

Abstract

Big data is the collection of large amount of data which is generated from various application such as social media, e-commerce etc. Those large amount of data were found to be tedious for storage and analysis. Now a day's various tools and techniques are coming to an existence to solve this problem. One of the application where huge amount of data is massively increasing is Crime which makes a huge issue for government to provide deliberate decision by following the laws and order. These large data were kept with the solution called Big Data Analytics. The processing of this huge data generated can be analyzed with high performance when compared to traditional way of processing. By performing these analytics, the government and the people will be provided with security and a healthy nation.

Keyword- Big Data, Hadoop, HDFS, Sqoop, Pig, Hive

I. INTRODUCTION

Nowadays large amount of data has been generated day by day. These large amount of data cannot be processed in a traditional way and hence big data analytics came into existence. This technique provides to overcome as the challenges such as storage, processing, visualization and privacy. There are three V's which plays a major role in this technique. They are *Volume: This signifies huge voluminous data; it is in order of terabytes and even petabytes. *Velocity: This signifies the high velocity with which the data is generated. *Variety: This refers to the huge variety of data generated in various fields.

Hadoop is an open source and durable platform for Big Data which can process huge data with high speed, scalability and reliability. Hadoop framework is capable to develop applications that run on clusters of computers and they could also perform complete statistical analysis for a huge amounts of data. The two major components of Hadoop are Map Reduce and Hadoop Distributed File System (HDFS). Hadoop consists of number of components which can be used for storing, processing and analysing data efficiently. Pig, Hive, Sqoop tools works on the top of the Hadoop by performing the above actions.

II. PROBLEM CREATION

In today's world, every establishment is facing growing challenges which need to be coped up quickly and efficiently. With continually increasing population and crime rate, analysing the data is a huge issue for governments to make decisions. This is really necessary to keep the citizens of the country secure from crimes. The best place to find way for improvement of voluminous raw data that is generated on a regular basis from various sources is by applying Big Data Analytics. Big Data Analytics refers to the tools and practices that can be used for transforming the raw data into meaningful content which helps in forming a decision support system to take steps towards keeping crimes in control. With the frequently increasing population and crime rates, certain trends must be discovered, studied and discussed to take well analysed decisions so that law and order can be maintained properly through by providing sense of safety and prosperity among the citizens of the country.

III. EXISTING METHOD

In the existing system, we can view only the details of particular information about the crime in our city by. Thus it produces more workload for the authorized person. The size of database is increasing day-to-day, so the load on database is consequently increasing. All the crime records are stored in a file. When other police station requires any criminal information, at that time they need to call that police station.

This leads to disadvantage by consuming time Drawbacks of Existing System

1) More man power.
2) Time consuming.
3) Consumes large volume of pare work.
4) Need manual calculation.
5) No direct role for the higher officials.

6) Damage of machines due to lack of attention.
7) To avoid all these limitations, the system needs to be computerized.

A. Literature Survey

Lenin Mookiah,William Eberle and Ambareen Siraj 2014, suggested that crime analytics and prediction have been studied among many research communities. In recent years, crime data from different heterogeneous fields have given several opportunities to the research community for effectively studying crime pattern and predicting the tasks in actual real data. Here they also discussed that the research takes into account a variety of crime related variables, and shows that in some cases information has been widely accepted as influencing the crime rate.

Tahani Almanie, Rsha Mirza and Elizabeth Lor 2015, focused on finding spatial and temporal criminal hotspots. It analyses two different real-world crimes datasets and provides a comparison between them through a statistical analysis supported by several graphs. Then, it clarifies how we conducted Apriori Algorithm to produce interesting frequent patterns for criminal hotspots. In addition, they showed how to use Decision Tree classifier and Naïve Bayesian classifier in order to predict potential crime types. To further analyse crimes' datasets, they introduced an analysis study by combining dataset with its demographics information in order to capture the factors that might affect the safety of neighbourhoods. The results of this solution could be used to raise people's awareness regarding the dangerous locations in that area and to help agencies to predict future crimes in a specific location within a particular time.

Anisha agarwal and Dhanashree chougule 2016, suggested that a threat can be reduced if a prediction analysis is done on the concerned person to determine if he is about to do the crime or not. This aspect can be beneficial both for law enforcement and the safety of our country. Today, time is a concerning factor for sentencing criminals. Many a time a criminal released on bail may yet become potential threat to the society, even after they have served their sentence. Data mining is an approach that can handle large voluminous datasets and can be used to predict desired patterns. Our users will be the police officers who is viewing from time to time shall be able to predict the possibility of the crime and the criminal is probable to commence in the nearest future as well as which particular crime he will be committing. They used frequent pattern mining with association rule mining to analyse the various crimes done by a criminal and predict the chance of each crime that can again be performed by that criminal. So that the threats on crime areas can be reduced. This analysis may help the law enforcement of the country to take a more accurate decision or may help in safeguarding that particular area. They also concentrated on Apriori algorithm with association rule mining technique to achieve the result by predicting the crimes. By using this Apriori algorithm, the crimes and criminal's details have been classified and can be easily viewed by the police officers and people.

IV. PROPOSED METHOD

Citizens need not go to the police station to complain about crime. They can directly view information on site about the crimes and criminal. Thus it reduces the manpower and also reduces the time. The user can also view the current status of their particular case updated from police station. Due to generation of huge amount of data, the space for storage become restricted. This produces storage deficiency which is overcome by using HDFS storage.

We mainly focusing on the following two specific topics:

1) Performance trade-offs of Hadoop deployment models on both physical and virtual clusters for analysing the crime data
2) Investigation of Hadoop applications provides power consumption.

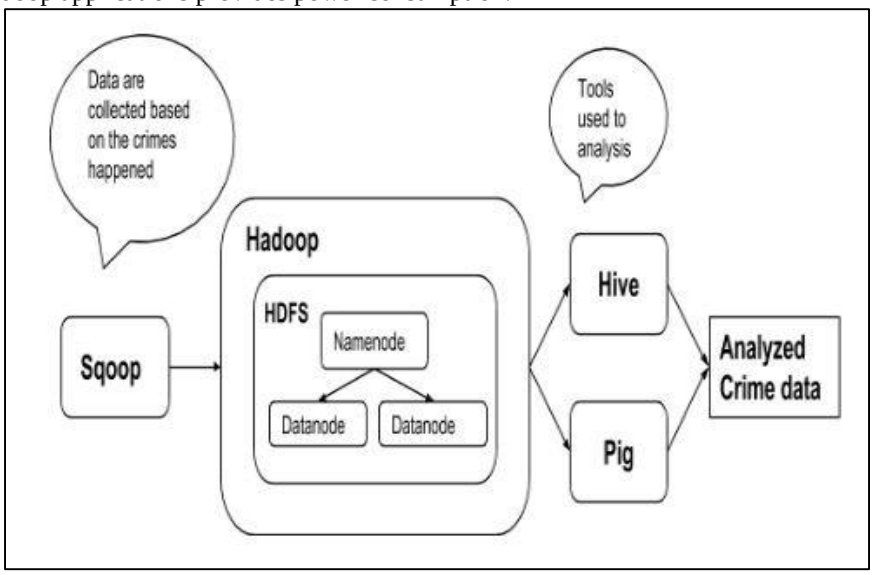

Fig. 1: Architecture for the Analysis of Crime Data

A. Loading the Data

Sqoop has been designed to move complex amount of structured data into HDFS. It is connected with RDBMS and moves the data into the HDFS and vice versa. Sqoop is used for both importing and exporting the data. It has a generic JDBC connector which helps to move data from all kinds of RDBMS's along with SQL services such as MySQL and oracle's SQL+. It is a distributed and reliable service for aggregating, moving and storing data.

B. Storing and Processing the Data

The Hadoop file system is designed to use in commodity hardware. The Hadoop distributed file system (HDFS) is highly scalable in nature to store data which is big in size across the multiple nodes of the clusters. It is reliable because it replicates the data across the nodes, and hence it's easy to recover the data from failure. HDFS has a master/slave architecture in which master plays as a single NameNode that manage the file system whereas slave plays as a multiple DataNode by serving those distributed files from the Name Node and regulates access to files by users. HDFS exposes a file system namespace and allows data to be stored in files. The NameNode executes file system operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. DataNodes are created for serving read and write requests from the file system's clients and also perform block creation, deletion, and replication upon instruction from the NameNode.

There are two different trackers in HDFS. One is JobTracker and another one is TaskTracker. JobTracker runs on separate node by receiving the request for the MapReduce and execute for the clients. JobTracker uses NameNode to find the location of the data then search for the best TaskTracker nodes to execute the data. Its job is to monitor the TaskTrackers and to submit the status of work back to client. If the JobTracker is slow, HDFS will run but the MapReduce execution cannot be started and existing MapReduce jobs will be stopped.

TaskTracker runs on all DataNodes. Both Mapper and Reducer tasks of mapreduce is executed on Data Nodes by assigning to TaskTracker by JobTracker by updating the status of processing of data. If the TaskTracker fails, the tasks are assigned to another node to execute.

MapReduce is achieved by programming paradigm which can do parallel processing on cluster nodes. It takes the input and gives the output in form of key-value pairs. MapReduce contain three stages/phases by processing each data sequentially and independently on every cluster node and generates intermediate key-value pairs.

The Mapper Phase is the first stage in the process which it splits out each word into a separate string and for each string, it will provide the output as 1.

Map (k1, v1) A list (k2, v2)

The Shuffle / Combiner phase will use the word as the key, by hashing the records to the specific reducers.

The output of the Mapper phase will be the input of Reducer phase. The Reducer phase will then sum the number of times that how many times each word was repeated and write that sum count together with the word as its output. It processes and merges the information values to give the final output, will be in the form of key-value pairs.

Reduce (k2, list (v2)) A list (k3, v3)

Thus output will gets sorted after each phase, thus providing the user with the aggregated output from all nodes in an orderly fashion.

C. Analysis of Data

After collecting the data, it has to be processed so that meaningful information can be extracted which can serve as a decision support system. However, Writing MapReduce requires basic knowledge of Java along with sound programming skills. Even after writing the code, which is in itself a labour intensive task, consumes additional time is required for the overview of code and its quality check. But now, analysts have additional options of using the Pig Latin which is the scripting language used to construct MapReduce programs for an Apache project which runs on Hadoop. The benefit of using this is that fewer lines of code to be written which reduces the overall development, testing cost and time. These scripts take just about 5% time compared to writing MR programs but 50% more time in execution. Although Pig scripts are 50% slower in execution compared to MR programs, they are still very effective in increasing productivity of data engineers and analysts by saving lots of time in writing phase. The scripting language consists of flow instructions which are first converted into logical plan in which it allows to maximize the processing and guarantees the users' needs and then the physical plan of the pig allows for optimization of joining, grouping and parallelism of tasks. Also, different operators can be having different factor of parallelization to cope with the varying volumes of data flow. The script is then converted into MapReduce instructions through its framework and used to process data. The steps to be followed in executing the Pig Latin scripts are

1) The first step in a Pig program is to LOAD the data you want to manipulate from HDFS.
2) Then, run the data through a set of transformations (which, under the covers, are translated into a set of mapper and reducer tasks).
3) Finally, DUMP the data to the screen or STORE the results in a file somewhere.

By using Hive, it acts as a data warehouse facilitating reading and writing operation using SQL queries called Hive Query Language (HQL).HQL statements are broken down by the Hive service into MapReduce jobs and executed across a Hadoop clustering.

V. IMPLEMENTATION

A. Experimental Finding

1) Total Number of Crime in Chennai City

Algorithm for finding the Crime using Grunt Shell
Input: Crime data
Output: Crime rate in 2015 Chennai

1) Enter into the Grunt shell using command: Pig
2) X = Load the data set using PIG;
3) Y = for each X generate the crime year by area;
4) Z = Group by area;
5) Data = for each Z generate group, SUM (X.year);
6) Store output;

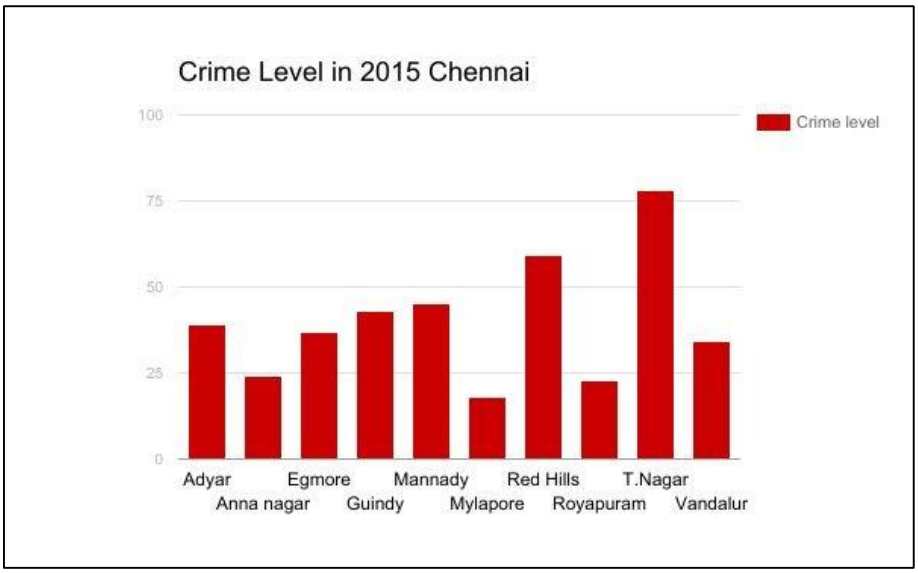

Fig. 2: Total Number of Crime in Chennai City

From the above graph it is clear that if the number of complaints from a particular city is found to be very high, extra security must be provided to that area thereby increasing police and strict vigilance. So that the crime cities can be avoided from further attacks.

2) Total Number of Crime by their Types

Algorithm for finding the Crime data using Grunt Shell
Input: Crime data
Output: Types of Crime in Chennai:

1) Enter into the Grunt shell using command: Pig
2) X = Load the data set using PIG;
3) Y = For each X generate the crime year by crime;
4) Z = Group by crime;
5) Data = For each Z generate group, SUM(X.year);
6) Store output;

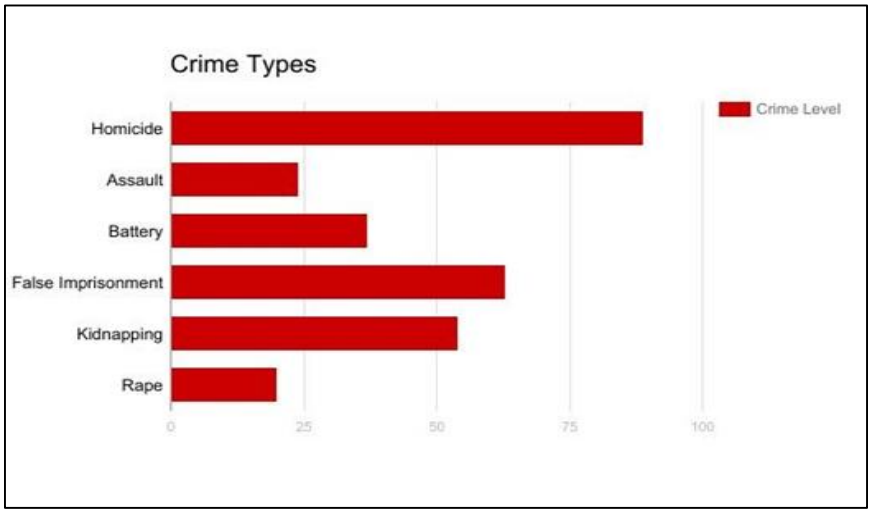

Fig. 3: Total number of Crime by their types

From the above graph it indicates that the crime in which the objective is violent with the increasing crime, certain trends must be discovered by spreading awareness among the citizens of that particular area. This analysis describes the rate of crime occurred in that particular city.

3) Total number of crime happened based on age

Algorithm for finding the Crime using Grunt Shell
Input: Crime data
Output: Grouping crime based on age

1) Enter into the Grunt shell using command: Pig
2) X = Load the data set using PIG;
3) Y = For each X generate the crime by age;
4) Z = Group by age;
5) Data = For each Z generate group, SUM(X.year);
6) Store output;

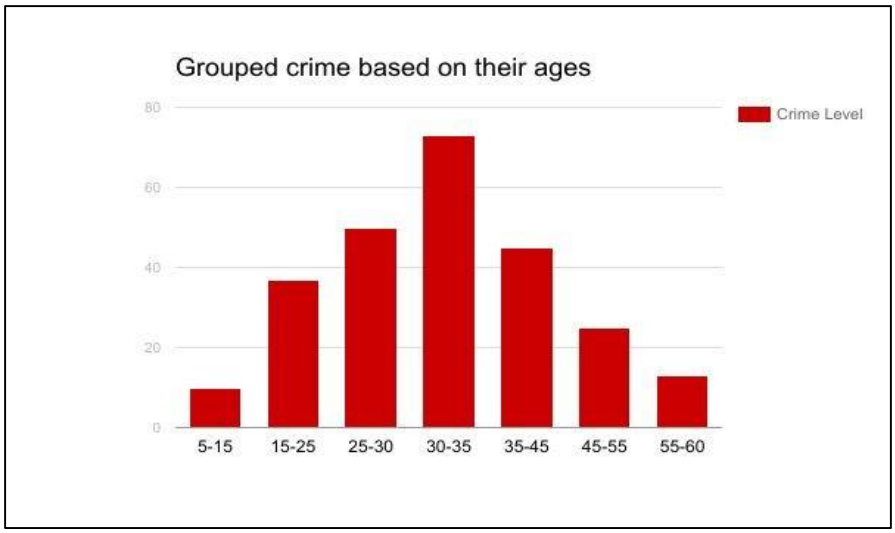

Fig. 4: Total Number of Crime based on the age

From the above graph it indicates the crimes that are happening around the city based on the age of the persons. The increasing population and crime rates, certain trends must be discovered, so that law and order can be maintained properly among those ages of people of the city.

If the number of complaints from a particular area is found to be very high, extra security must be provided to those residents, thereby increasing police presence, quick redressed of complaints and strict vigilance.

By analyzing through Grunt shell algorithm, the performance speed is increased up to 20% so that the time consumption plays a major role in analyzing those huge amount of data. Thus this algorithm defines about the crime taking place, the culprit involved and the type of person affected were easily recognized. The limitation of year, describes the total number of crime detected

in that particular area were predicted by providing awareness and security to that area. Therefore, the tedious job for government by analyzing such a huge number of data have been easily analyzed through this grunt shell algorithm. Distributed and parallel analysis on the huge number of crime data makes the processing and analyzing speed to be completed with high speed rather than the traditional way of processing the data which is not possible for huge amount of data. Overcoming the problems in traditional processing through by completing the appropriate work in Big Data Analytics through Grunt Shell Algorithm.

VI. CONCLUSION

Big data analytics refers to the process that can be used for transforming raw data into meaningful information. This also helps in forming a decision support system for the legislature towards crime detection. Since the population and the crime rates are increasing drastically, certain steps should be taken in order to maintain law and order. If the number of crimes are found to be increasing in a particular area extra security should be provided. These measures should be carried out so that a sense of safety and awareness can be maintained among the citizens of the country.

VII. FUTURE WORK

This analysis can be further carried out by using Apache spark which provides support for structured and semi-structured data. Its processing engine's in-memory computing layer is supposed to run batch-processing programs 100 times faster than MapReduce.

REFERENCES

[1] Tahani Almanie, Rsha Mirza and Elizabeth Lor "Crime Prediction Based on Crime Types and Using Spatial and Temporal criminal Hotspots" International Journal of Data Mining & Knowledge Management Process Vol.5, No.4, July 2015
[2] S. Sathyadevan, M. Devan, and S. Surya Gangadharan, "Crime analysis and prediction using data mining," in Networks Soft Computing(ICNSC), 2014 First International Conference on, Aug 2014, pp. 406–412.
[3] J.Archenaa1 and E.A.Mary Anita2 "A Survey of Big Data Analytics in Healthcare and Government" Procedia Computer Science 50 on, 2015, pp.408 – 413.
[4] Liang Zhao, Zhikui Chen, Senior Member, IEEE, Yueming Hu, Geyong Min, Senior Member, IEEE,and Zhaohua Jiang "Distributed Feature Selection for Efficient Economic Big Data Analysis" Journal Of Latex Class Files, Vol. 13, No. 9, Sep 2014.
[5] Arushi Jaina, and Vishal Bhatnagara "Crime Data Analysis Using Pig with Hadoop" International Conference on Information Security & Privacy (ICISP2015), 11-12 Dec2015, pp.571 – 578
[6] Janez Kranjc,Roman Orac,Vid Podpecana,Nada Lavrac, Marko Robnik-Sikonja c "ClowdFlows:Online workflow for distributed ig data mining" Future Generation Computer Systems on 2017,pp.38-58.
[7] Tomasz Jacha, Ewa Magieraa, Wojciech Froelicha "Application of HADOOP to Store and Process Big Data Gathered from an Urban Water Distribution System" Procedia Engineering 119 on 2015,pp.1375-1380
[8] Y. Chen, S. Alspaugh, D. Borthakur, R. Katz, "Energy efficiency for large-scale mapreduce workloads with significant interactive analysis", Proceedings of the 7th ACM European conference on Computer Systems, EuroSys '12, 2012,pp. 43–56.
[9] R.Saranya, V.P.MuthuKumar, "Security issues associated with big data in cloud computing" International Journal of Multidisciplinary Research and Development Vol.2,No.4,April 2015

