International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307

Crime Analysis and Prediction Using Big Data

Pournima B. Minajagi*, Prof. Ranjana Nadagoudar

Department of Computer Science and Engineering, Visvesvaraya Technological University, Belgaum, Karnataka, India

ABSTRACT

Big data involves large-scale storage and processing of large data sets. Crime analysis and prevention is a systematic approach for identifying and analyzing patterns and trends in crime. It can predict regions which have high probability for crime occurrence and can visualize crime prone areas. The use of frequent pattern mining with association rule mining to analyze the various crimes done by a criminal and predict the chance of each crime that can again be performed by that criminal. This analysis may help the law enforcement of the country to take a more accurate decision or may help in safeguarding an area if a criminal released on bail is very much likely to perform crime. Apriori algorithm with association rule mining technique to achieve the result.

Keywords: Apriori algorithm, association rule mining, crime analysis, prediction.

I. INTRODUCTION

Sets of big data which is used for software tools to manage, process and capture the data within reasonable elapsed time. Big Data sizes are currently ranging from a few dozen terabytes to many petabytes of data in a single data set. Big Data have set of technologies and techniques among new type of integration to expose deep assessment from datasets which are complex, diverse and of enormous scale. It is a wide set of data therefore complex or large, that conventional processing of data applications are insufficient. The term frequently refers to remove value from data, and barely to a data set of particular size. Accurateness in Big Data could guide to additional certain decision making and improved decisions are capable of outcome in better reduced risk, cost reduction and operational efficiency.

Data mining means retrieval of essential data which is hidden in huge set of data. It is a great technology with high potential to help organizations focus on the most important information in their data warehouses. It is non-trivial removal of implicit, potentially helpful information and subsequently unknown data from data base. In an uncertain and highly competitive business environment, the value of strategic information system is easily recognized. In today's business environment, efficiency or speed is not only the key for competitiveness. The tremendous amount of data, in the order of tera- to peta-bytes, has fundamentally changed science and engineering, transforming many disciplines from data-poor to increasingly data-rich, and calling for new data-intensive methods to conduct researches. To make, manage and analyze a conclusion of huge quantity of data we require methods known as data mining which will be transform in numerous fields. Data mining tools predict future trends and behaviours. It is popularly known as knowledge discovery in databases. Using traditional tools such as RDBMS, NoSQL, HPC it is difficult to store, manage and process the unstructured large amount of data. In order to overcome all the problems faced by traditional tools, now we are using Hadoop as main platform to store big data through its Hadoop Distributed File System.

II. LITERATURE SURVEY

Literature survey is an important activity, which we have to do while gathering information for the project. It will help us to get required information or ideas to do the project. The following paragraphs discuss the related work and issues in the area of crime analysis in Big Data.

Dr. V. M. Thakare and Mr. S. P. Deshpande[1] represents the removal of unseen predictive data from huge databases; it is a dominant technology with enormous potential to facilitate organization center on the mainly essential information in their information warehouses. Tools of mining predict future behaviours and trends, helps organizations to create practical knowledge-driven decision. The prospective, automated analysis presented by means of data mining move about outside the analysis of ancient times procedures given by retrospective tools distinctive of decision supporting system. Tools of mining can respond the questions that conventionally were also consuming of time to decide. They organize databases for predictive data and looking for discovery of unknown patterns where experts may overlook since it lies outer surface of their potential.

Chin-Feng Lai, Athanasios V. Vasilakos, Han-Chieh Chao and Chun-Wei Tsai [2] describes that the problems of analyzing the large scale data were not suddenly occurred but have been there for several years because the creation of data is usually much easier than finding useful things from the data. Even though computer systems today are much faster than those in the 1930s, the large scale data is a strain to analyze by the computers we have today. The troubles of managing the large-scale data still exist once we are inserting the era of big data; means that the data is unable to be handled and processed by most current information systems or methods because data in the big data era will not only become too big to be loaded into a single machine, it also implies that most traditional data mining methods or data analytics developed for a centralized data analysis process may not be able to be applied directly to Big Data.

D. Carstoiu, A. Cernian, A. Olteanu[3] provides HDFS is a dispersed file system structure for working on general hardware structures (commodity computers) characterized by low cost implementation. Through HDFS, applications can rapidly access data in the context of applications that handle large volumes of data. An HDFS instance may consist of hundreds or thousands of machines, each keeping parts of data files. In case of failure, it can be restored automatically. HDFS supports even millions of files in one instance, aggregating a scalable multitude of nodes in the same cluster. The simple consistency model implemented is write-once-read-many. Processing in an application with large amounts of data is more efficient if executed near where data are stored. This minimizes network congestion and increases the system performance.

Manashvi Birla, Aditya B. Patel, Ushma Nair[4] says that: "Addressing Big Data Problem Using Hadoop and MapReduce" reports the experimental work on the Big data problems. It shows the most favorable solution by means of Map Reduce programming framework, Hadoop Distributed File System (HDFS) for storage and cluster of Hadoop for parallel processing to process huge set of data crimes, therefore we utilize data mining method for analysis and prediction of it.

Sanjay Ghemawat and Jeffrey Dean [5] describes that programming model is MapReduce and related performance for generating and processing huge set of data. Users identify function of map that processes a input/value pairs to produce a set of transitional pairs of input/value, and a decrease function that merge all intermediary standards coupled with the similar intermediate key. Written programs in this practical manner are routinely executed and parallelized on a huge cluster of commodity equipment. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This provides programmers with no any knowledge with distributed and parallel systems to simply operate the huge system of distributed resources.

T.Kalaikumaran, Sukanya.M and Dr.S.Karthik [6] represents analysis of crime is a set of analytical and systematic procedure for providing the data concerning with crime patterns at the exacting time. Investigation of crime is an significant action for identify the crime hotspot. It maintains the amount of department function that contains special operations, patrol deployment, investigations, tactical units, crime prevention and administrative, planning and research services.

III. METHODOLOGY

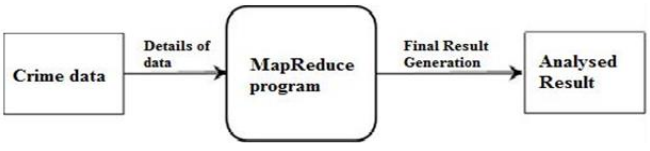

Figure 1: Zero Level DFD Diagram

Figure 3.1 shows the zero-level DFD for the Invoice Analysis Process. The data which has to be analysed is the input for the project. The analysing process will be carried out using Hadoop and the final result will be the analysed data which will be stored in a file.

3.1 First Level DFD

Figure 3.2 shows the first-level DFD for the Invoice Analysis Process. Mapper, Reducer and Storing the analysed data are the main process involved in the Analysis process. Analyst will provide the invoice data, Run mapper on the data and the data is passed on to run reducer. The outcome of data from reducer will be the analysed data which will be stored and returned back to the Analyst.

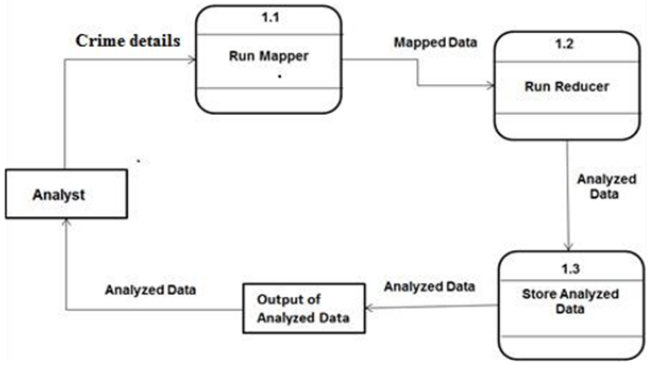

Figure 2: First Level DFD Diagram

3.2 Activity Diagram

The representations in terms of graphical where workflows of stepwise activities and actions with maintenance for iteration, choice and concurrency.
Figure 3.3 shows the Activity Diagram for invoice Analysis using Hadoop Initially the Hadoop Cluster is set up and single node cluster is established. After that the invoice data is taken as input of analysis, Mapper and Reducer are run on the data using the key value. The final analysed data is stored in a file as output of Analysis process.

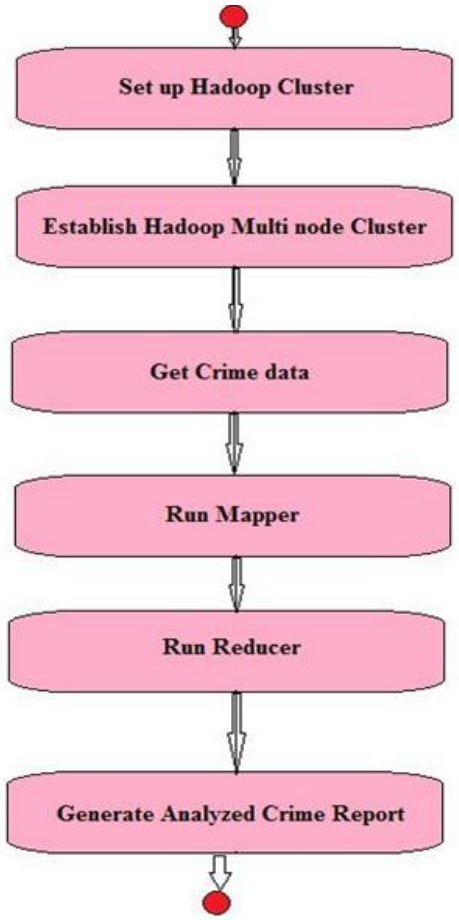

Figure 3: Activity Diagram to represent Crime data analysis

IV. IMPLEMENTATION

4.1 Hadoop Cluster Architecture

HDFS intended to run on commodity hardware. It has lots of similarity among accessible distributed file system and is considered to be deployed on low-priced hardware and also extremely fault-tolerant.

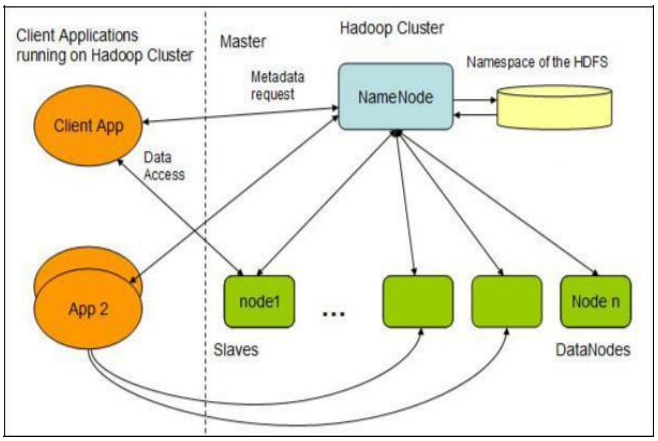

Figure 4 : HDFS Architecture

Figure 4.1 shows the master/slave architecture of Hadoop with client applications accessing the cluster. For data access, a client application communicates with

the NameNode server to obtain the file system metadata. Then transfer of information takes place among the DataNodes and client application. The structural design therefore establishes least amount contribution of the single NameNode to decrease the workload scheduled on the cluster. MapReduce jobs executing on the cluster, the tasks are submitted from the NameNode which are additionally dispersed among the necessary DataNodes. The data is processed on the DataNodes as Map and Reduce tasks and the output is written back to the file system for client access.

4.2 MapReduce

It is a framework of software developed in 2004 by Google to maintain distributed computing on clusters of computers on huge sets data. It is a model of programs for generating and processing huge sets of data. Users identify a function of map that processes pair of key/value to produce a reduce function and a set of intermediary pairs of key/value that merges the entire transitional standards connected among the similar intermediate key.

Step for Map: The node of master receives the input, distributes it up into minor sub-difficulties, and spreads them to nodes of worker. A node of worker could perform this another time, in turn foremost to structure of multi-level tree. The worker node processes the lesser difficulty and passes the response back to its node of master. Map takes data of one pair with a category in one domain of data, and proceeds pairs list in several domain:

Map (k1, v1) → list (K2, v2)

Step for Reduce: The node of master then contributes the solutions to every the sub-difficulties and collects them in several way to outline the output – the response to the problem it was initially annoying to resolve. The function of Reduce is applied in parallel to every cluster, which in turn provides a gathering of standards in the identical domain:

Reduce (K2, list (v2)) → list (v3)

When MapReduce jobs are submitted to the cluster, the NameNode forwards them to appropriate DataNodes where the data resides. Ahead getting the tasks, the DataNode generates the task and returns the result on its local system. The implementation of task is handled by framework of MapReduce. In the MapReduce programming model, the computations are divided into map and reduce tasks. The tasks are simultaneously performed on the DataNodes. In the mapping task, the data is processed into <key, value> pairs with a minimal coordination of DataNodes. In the reducing task, each output from DataNodes is combined to produce single output for the application. Figure 4.2 illustrates how the programming model works.

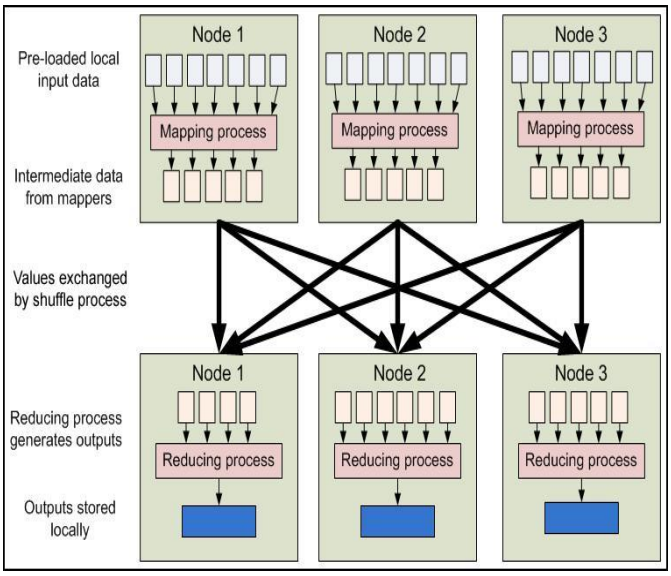

Figure 5 : MapReduce programming mode

4.3 Algorithm-Apriori algorithm

The Apriori-Algorithm is influential algorithm for mining frequent item sets for boolean association rules. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation, and groups of candidates are tested against the data.

```
Apriori (T, minSupport) {
//T is the database and minSupport is the minimum support
L1= {frequent items};
for (k= 2; Lk-1 !=∅; k++) {
Ck= candidates generated from Lk-1
//that is Cartesian product Lk-1 x Lk-1 and eliminating any k-1 size itemset that is not //frequent
for each transaction t in database do{
#increment the count of all candidates in Ck that are contained in t
Lk = candidates in Ck with minSupport
}//end for each
}//end for return
return Uk, lk; }
```

V. RESULTS ANALYSIS

Analysis of different attributes

Here we considered Category, Day, PDDistrict attribute. The outcome of this is the number of crimes that had happened in each attribute based on the year.

Category

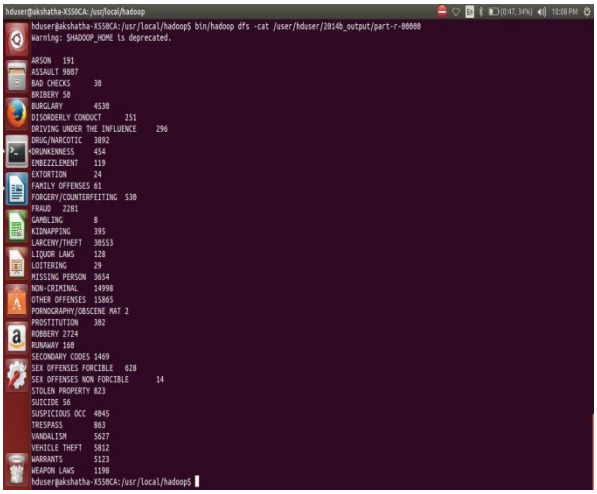

Figure 6 : Shows the Category wise analysis for the year 2016.

Day

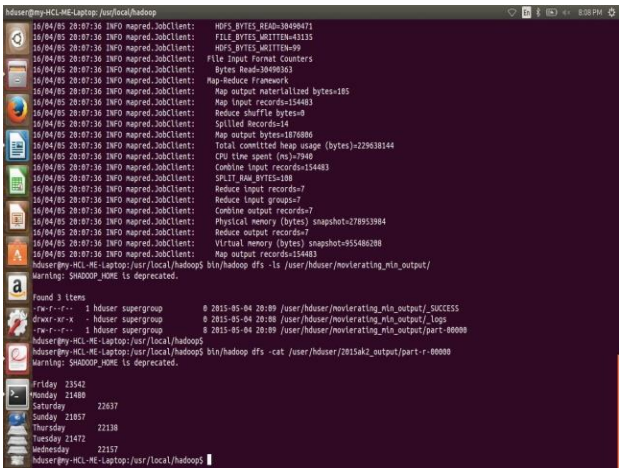

Figure 7 : Shows the Day wise analysis for the year 2016.

PDDistrict

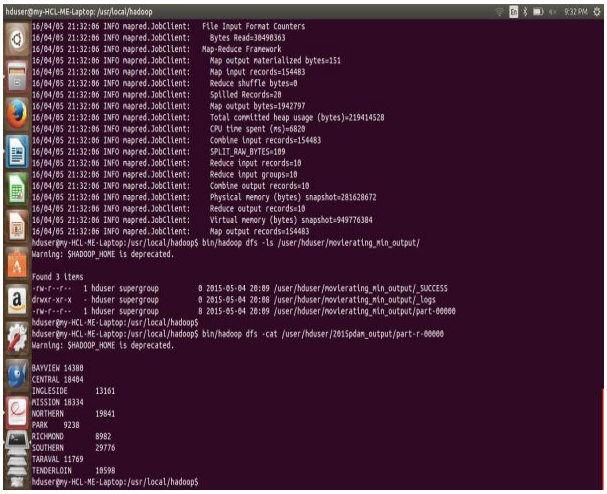

Figure 8 : Shows the PDDistrict wise analysis for the year 2016.

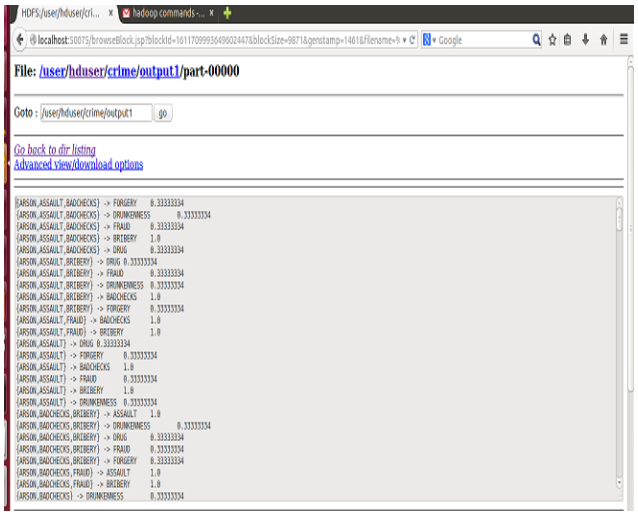

Figure 9 : Shows the result of accuracy against different item sets.

VI.CONCLUSION

The result of Apriori algorithm gives the maximum accuracy against the different datasets. So the system will automatically learn the changing patterns in crime by examining the crime patterns. The result of this project will facilitate to predict the crime against the individual criminal who is going to do the crime. As a part of prediction, we can find the location of the crime. Against several datasets, we will predict the frequent item sets and the system will understand the patterns by verifying the crime patterns. Bringing the change in a crime data we have to identify the new changes which leads the crime.

For achieving the better prediction, we have to find more attributes. As the software will perform to predict the criminal on individual crimes using "Apriori algorithm" with association rule and this will facilitate to find the criminal who is about to commit the crime.

VII. REFERENCES

[1]. Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos, Big data analytics: a survey, Department of Computer Science and Information Engineering, National Ilan University , 2015.

[2]. Jeffrey Dean, Sanjay Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, Google, 2014.

[3]. Dr. V. M. Thakare and Mr. S. P. Deshpande, A data Mining Systems and application: A Review, Amravati University, Volume 1, Issue 6, May 2016.

[4]. Manashvi Birla, Aditya B. Patel, Ushma Nair, The Big Data Analytics with Hadoop: Review, Volume 6 issue 8, December 2012.

[5]. Dr. A.Bharathi, R. Shilpa, A Survey on Crime Data Analysis of Data Mining Using Clustering Techniques, Volume 2, Issue 8, August 2014.