A Survey of Cloud era Hadoop Analytics in London Crime Data

ABSTRACT

The present research aims at the creation of a crime data dashboard using Hadoop Big data technology which could perform analysis of the data collected from the London City police. The study of crime and the development of science and technology has been in a long lasting interaction. The present research focuses on the small area of the London city where the population is not more than 10,000 which is protected by the London city police. In this study, semi structured crime data from 2013-2017 was collected for analysis especially in London city. MapReduce algorithm is used for analysing the London criminal data and creates a dashboard for displaying the classified criminal activities. Crime table generated in the research will helpful to the citizens and police organisations to understand which crime type are standing at top of list, and then people will understand the crime possibility at their location. This present study based on cloud using Hadoop was exercised in the small scale which seems to reduce the crime in the London city in a considerable manner. Researches should be done further in order to utilize the methodology in a global manner for the crime investigation and control.

Table of Contents

CHAPTER I: INTRODUCTION

1.1 Background of the research

The study of crime and the development of science and technology has been in a long lasting interaction. Crime being a social threat needs to be controlled, and it becomes fundamental in the development of societies all around the world and also for the welfare of the people. This crime control also brings in the stability of countries. It is observed that in today's' world there is a large increase in the volume of crimes that occur in the society. As this increased, the need arises to focus on the prevention of crime with the strengthening of public security. The quality of life and the fundamental freedom is highly influenced negatively by the crime in both developed and developing countries. This challenges the human rights of the people of any nation. The crime occurrence did not spare any country, but the seriousness and intensity of the problem vary from country to country (Li, 2014). In the UK, the studies were made on the criminal activities by Budd (2001) and found that 1 % of UK households experienced 42 % of residential burglaries, and (Bowers, 2014) found that 80 % of thefts in bars in London (UK) occurred in 20 % of facilities (Rosser *et al*., 2017). From the ONS report of Crime Survey on England and Wales (CSEW) showed there were 5.9 million incidents of crime and 11 million incidents of crime were filed related to the criminal activities. In March 2017, the police have recorded nearly 5 million offenses, and it confirms the gradual increase in the rate of some crime types (ONS, 2017).

The scientific study of crime and criminal behavior is the key focus in the study of criminology. This process also focuses on law enforcement and aims to identify the crime characteristics. Crime analysis includes a relationship with criminals and also in exploring and detecting crimes. The increase in the crime rate and the high volume of crime data sets with the complex relationship between them has proved the need for data mining techniques in crime analysis. The data mining tools and their results help and support police forces (Keyvanpour *et al*., 2011). According to Nath (2007), Crime detection problems are assisted with data mining technique and solving the crime involves human intelligence and experience which is a complex task. In this method, the data mining is employed to develop computer models with the human experience in crimes over the years. The crime is viewed from the bigger picture perspective with big data analytics where the focus is moved from the particular data points and the possible inaccuracies. The new inferences are made with the

correlations and the patterns that are included in the data. The theory has been already proved the practical implementation of the Big Data process makes it more significantly feasible for practical applications. The application of big data in crime analysis is thus significant(Ravna, 2015).

1.2 Problem statement

The growing challenges must be met and handled efficiently as it is inevitable in the every establishment. As the population is increasing it is also seen that the crimes and crime rates are also increasing and hence the analysis of crime-related data becomes important. This analysis becomes a concern for government to make strategic decisions thus maintaining law and order. This analysis is essential to keep the country and its citizens safe from crimes. Big data analytics thus enters the scenario as the voluminous data that are being generated by various sources on a regular basis must be handled. BDA is a tool and also a practice that is used for transforming the raw data into meaningful information that help the decision of the legislative and judiciary thus helps in keeping the crime in check(Jain & Bhatnagar, 2016).

As the problem is identified, present research focuses on the development of the crime analysis dashboard using Hadoop Big data technology. Our research focuses on the small area of the London city where the population is not more than 10,000 which is protected by the London city police. The metropolitan police force is the largest police force which handles the entire United Kingdom. Thus they maintain the dashboard of crimes of UK. But the city of London police do not have any dashboard or crime analysis report that can control the crime. Our present research aims to create a crime analysis dashboard for them to improve security aspects in the nation.

1.3 Research aims and objectives

The present research aims at the creation of a crime data dashboard using Hadoop Big data technology which could perform analysis of the data collected from the London City police. In this regard, the following objectives are framed:

- To design and develop a Hadoop based Crime data analysis dashboard for the analysis of crime in the London city.
- To utilize the developed system to examine crime-prone zones, and total number of the crimes per year, month, quarter in the select region.

1.4 Scope and significance of the research

The present research limits its scope towards the development of a crime analysis dashboard for the region of London. However, the system designed can act as a framework for any region and can be utilized to analyze crime situations in any region.

1.5 Summary

The utilization of enormous information in Crime investigation is therefore critical. The hypothesis has been as of now demonstrated the reasonable usage of the Big information Data process makes it all the more fundamentally attainable for pragmatic applications. The investigation of Crime and the advancement of science and innovation have been in an enduring collaboration. Crime being a social danger should be controlled, and it ends up plainly central in the advancement of social orders all around the globe and furthermore for the welfare of the general population.

The present research means to make a Crime investigation dashboard for London police to enhance security viewpoints in the country. As the issue is distinguished, show examines concentrates on the improvement of the Crime examination dashboard utilizing Hadoop Big Data innovation. However, the city of London police don't have any dashboard or wrongdoing investigation report that can control the wrongdoing. The examination concentrates on the little zone little part in of the London city where the populace is not more than 10,000 which is secured by the London city police. The developing difficulties must be met and taken care of productively as it is inescapable in the each foundation. As the populace is expanding it is likewise observed that the violations and wrongdoing rates are additionally expanding and thus the investigation of Crime related information ends up noticeably imperative.

CHAPTER II: LITERATURE REVIEW

2.1 Introduction

"Big Data" shall be explained as the quantity of data produced off late. However there is much more than this. This concept evaluates the world in a new method, and this requires a unique platform which facilitates the near-constant exponential growth of data. As well, this requires high speeds for data collection and data processing. Several projects are available such as ClearMap, Citizens Data Project and lot more are trying to enter into the data pools based on the City of Chicago. As this is the beginning phase of these projects, they are on the lookout for the methods as for how to execute the full power of data. The main aim of this assessment will focus on the core of this data which is quite complex in specifically explaining about the city, however of the prognostic, action-based applications shall be generated in case if all the data sources are equally assessed in a data pool within a Big Data Ecosystem (Christopher, 2016). At times, Big data comprises of data sets with distant sizes from the power of generally utilized software tools for capturing, curating, managing and processing data within a manageable duration. Big data "size" is an unendingly moving goal, because it commences from several dozens of terabytes and passes through various petabytes of data. Big data shall be defined as a group of strategies and technologies which require novel methods of integration to open huge veiled values from huge datasets which are several, tough and of highly huge. Big data environment engages a mass, systematize and valuate the several data types. It is noted that the respective Map Reduce framework that structure creates a huge deal of transitional data (Yadav & Chaudhary, 2016).

2.2 Related studies

Wilson (1963) identified crime analysis as an essential police function and recommended that a crime analysis, based on the Wilson techniques, in 1982 the final report of the integrated criminal apprehension program (ICAP), the definition of the crime analysis was mentioned as *"*crime analysis Function is defined a set of systematic analytical processes directed at providing timely and pertinent information relative to crime patterns and trend correlations to assist- operational and administrative personnel in planning the deployment of resources for prevention and suppression of criminal activities, aiding the investigative

process, and increasing apprehensions and clearance of cases especially against career criminals" (Fyfe & Wilson, 1997).

Crime mapping and spatial analysis by Ahmadi (2003) proposed that they are the key tools in crime and justice. The digital mapping in crime control and prevention programs using GIS has shown increased advantages in the recent days. This is the first official definition of the crime analysis, now called as tactical crime analysis (Gwinn *et al*., 2008), but this definition did not satisfy the different demands that would be placed in the tiny police agencies. The crime analysis aided by the GIS system in a city identifies and highlights suspicious incidents and events in a city. The identified events and the implementation of policing methodologies may reduce the overall crime rate and disorder (ESRI, 2008).

Malathi and Baboo (2011) studied the enhanced algorithm for crime analysis using data mining .As large datasets are available efficient algorithm is needed for the analysis of such crime data. The spatio temporal modelling of criminal incidents are analysed by Wang and Brown (2012). The future criminal patterns are identified and the necessary measures are taken by the law enforcing authority to predict and minimize the occurrence of such crimes in the near future. The fear of crime in gated and non-gated areas studied by Abdullah *et al*. (2012) exhibit the crime analysis in different residential environments and the results helped in predicting the choice of residents in choosing their inhabitation.

Shibata *et al*. (2012) worked on the Japan crime data to identify the attitudes of crime prevention and perception of the public safety. In this project, he worked on principal component analysis (PCA) method to find out the public behaviour towards the crime. "The result showed that elder people have a much stronger attitude toward community crime prevention compared to younger peopl" (Shibata *et al*., 2012).

Soh revised above model in the 2012 itself; he worked on the crime and urbanisation relation of Malaysia country. In his journal, he cited that urbanisation is directly proportional to crime rate. In his report, he has concluded with "The involvement and responsibility to reduce the crime rate not only the government's responsibility but also must involve others especially public, private sector's as well" (Soh, 2012).

Rasmi and Jantan (2013) worked on the cybercrime analysis, in their work they delivered that the investigation techniques take much time and money consuming to understand and analyse the cybercrime data such as the intention of crime. so, Rasmi and

Jantan (2013) they found a new algorithm called Similarity of attack intension (SAI), this algorithm works on the cosine similarity, which follows the distance-based similarity measures the similar crime intensions, this algorithm is a revised version of the Attack Intension Analysis(AIA) algorithm, these algorithms predict the new cybercrime intensions and allocates the probability value for these intensions. In the result of their study they proposed that SAI algorithm finds the similarity between the cybercrimes called cybercrime intensions. Based on AIA algorithm they identified the probability values of the accuracy detection for the cybercrimes intentions. The final analysed result explains the relation between the new cybercrime and pre-defined cybercrimes, it is also proved that the similarity measurements of the intensions help the crime analysts to identify the similar cases of the new cybercrimes with old crimes, which reduces the processing time and cost.

Nath (2006) worked on the crime analysis using data mining, in this research the researcher defined that out of 100 % crimes 50 % of the crimes are committed by the 10 % of the criminals, so to solve the crimes, Shyam Varun Nath has used the clustering algorithm in the data mining field to reduce the crimes, He has applied the clustering algorithm technique to real time crime data which he has collected from the sheriff's office. Alongside, researcher applied the semi-supervised learning techniques to increase the predictive accuracy. The result has projected using geo-spatial plot of crime which helps to develop the productivity of the detectives and police agencies.

Sathyadevan *et al*. (2014) studied the crime analysis and prediction using data mining which predicts the frequently occurring crimes and with the advanced technology the crime prone areas are identified and the law enforcements can pay more attention in such areas thereby minimizing the occurrence of crimes. Ansari and Kale (2014) proposed a study where the crime mapping and analysis is done in Aurangabad city which is also aided by GIS. In the study the hotspot detection such as spatial analysis, interpolation and spatial auto correlation is used for finding the crime hotspot. Woodworth *et al*. (2014) studies the non-local crime density estimation incorporating housing information. With the discrete sample of event locations the probability of events occurring in spatial domain is analysed, the density of the crime can be predicted in residential burglaries happening in areas where there are no or minimal residences. The crime analysis and mapping used by the law enforcement organization initially used GIS system with which the crime prediction and accordingly

minimizing the crime in the hotspots detected. As there is increase in data further research is carried out with Hadoop environment which are discussed as follows.

Saoumya and Baghel (2015) discussed the predictive model for crime mapping using big data analytics. The crime event forecasting is done using hadoop in the areas where there are greater risks and these places are termed as hotspots. The artificial neural network specification and estimation approach is enhanced by the processing capability of hadoop platform. The crime mapping and analysis using GIS by Kedia (2016) is utilized for law enforcement and crime management. With the crime hotspots identified crime management is done effectively. Crime analysis of Chicago (Wang & Li, 2016) using the modern technology such as Hadoop that process the large amount of dataset and it analyses the relationship between the crime rate and the factors related to it. With this the key factors for crime analysis can be done and the crime management is the result of this analysis.

Minocha *et al*. (2016) discussed the map reduce frame work for big data analytics. The map reduce algorithms for Chicago crime data sets using hadoop web is used for analysing the crime data sets and the interpretations help in identifying the crimes in a particular locality.

Jain and Bhatnagar (2016) studied the crime data analysis with Pig with Hadoop. With increasing population and crime rate it becomes voluminous to analyse the crime data set when the data are analysed efficiently the citizen rest assured as the crimes analysed can be prevented in future.

Ghosh *et al*. (2016) initiated a Machine Learning technique to mechanize and aid illegal analysts spotting and work in situations which could bring in a big change. Effectively, it facilitates them to recognize dead end leads and shun having the researchers provide duration on certain factors such as:

- The system backs to rapidly dig all the way through several data sources and choose if the executioners or losses comprised of associated and spot and previous accounts
- This essential technique aids analysts in the data fusion centers expand helpful ‘tactical information profiles’ of the supposes promptly and give it to the researchers in the field

- Synchronized back-end functions and ground work will help get good output

Pushpa *et al*. (2016) it is gradually getting tougher to manage high quantity of data that creates about the business projection, range of fields from social networking to money, flight data, surroundings and fitness. Huge Data has high danger in the indemnity industry and to follow the practical responses to products. Big Data supervises qualities such as several wave movements, flight data, traffic data, financial transactions, fitness and offence. Huge Data confront forms something that is value to the consumer. There is a method in which it could be formed, stocked up, processed and evaluated it to twist the raw data information to hold the ultimate supervisory activity. In this research analysis, Big Data is portrayed as a case study for Airline data based on hive devices.

Williams *et al*. (2016) critically analyses the application and disadvantages of big data for the understanding of crime and uncertainty. This study hypothesize that disorder-related posts on Twitter are related with rate of police crime. The results show indication that naturally occurring social media data possibly give a different data source on the crime issue.

Joseph *et al*. (2017) discussed the data mining techniques for analyzing the crime patterns. The previously stored data from various sources are used for identifying the trends in crime patterns and the methods used in crime data mining are discussed.

Malarvizhi and Ibrahim (2017) an efficient clustering for crime analysis is studied. Crime incidents are increasing day by day and hence analyzing it with efficient means becomes important based on the occurrence of crimes in particular regions.

Pramanik *et al*. (2017) recognize five major technologies such as link analysis, intelligent agents, text mining, ANNs, and ML that was operated in several areas for developing the basic techniques of an automated security and illegal examination method. Based on the big data environment, there are several chances for beating into big data to develop security and illegal probing.

Osman *et al*. (2017) depending on the skill researched in the previous topic, design strategies of integrated scalable BDA frameworks for SCs are summed up in the next points:

- Skill to evaluate stream data for online and concurrent applications (D. (Singh *et al*., 2016; Ghosh *et al*., 2016b).

- Skill to evaluate batch data for applications that give suspension (Barbieru & Pop, 2016).
- Skill to hold extorted models for further processing,
- Skill to combine one or more extorted models for higher level of analytics

The evaluation method specified that existing proposed structures lack two main qualities, such as model persistence and ensemble. For coping up with lack, a new theoretical structure is presented to give the skill to persevere and assemble extracted models. The projected structure facilitates the skill to withhold the extracted models, facilitating further analysis on these models devoid of the requirement to re-evaluate the data. This quality protects the substantial duration for data re-evaluation. As well, extracted model assemble will facilitate the skill for more complex cross-domain analytics.

According to Josep *et al*. (2017) the data mining is a data analyzing techniques that applied to investigate crime data previously stored from different sources to identify methods and trends in crimes recently. Moreover, it can be used to develop capability in solving the crimes faster and further can be used to indicate the crimes automatically. However, there are numerous data mining techniques are available. Select ion of the appropriate data mining techniques is important in order to increase potential of crime detection. This paper reviews the literatures on different data mining uses, specifically applications that used to solve the crimes. Survey also throws light on research gaps and challenges of crime data mining. Furthermore, this paper provides in-depth information about the data mining for identifying the patterns and trends in crime to be applied suitably and to be an aid for researchers in the field of crime data mining. Big Data Analytics considered being the tools and practices that can be applied for changing this raw data into crucial and meaningful information which aids in producing a decision help system for the legislature and judiciary to the consideration of continuation of crimes investigation in. Certain trends and technologies must be discovered, studied and discussed to take well informed decisions with the elevating population and crime rates, this can be aid to proper maintenance of law and order properly. Additional security must be given to the residents there by increasing number of police, immediate redressed of complaints and strict investigation when the number of complaints from a specific state is identified to be very high. Crimes against women are becoming an increasingly worrying and serious issue for current government. In order to reduce these issues data mining application can be involved.

Sanders and Sheptycki (2017) defines moral economy is understood to be the distribution, circulation, production, and application of moral sentiments emotions and values, norms and obligations in social environment by which it enhances a contribution to the critique of stochastic governance. In reality, the argument is that some technological growths in connection to policing, conditions, law and governance are occurring in the context of a neo-liberal moral economy that is structuring the social results of stochastic governance. Considering about policing in both the precise of crime conflicting and extensively in its Foucaldian sense as governance, empirical indication of 'policing with Big Data' exhibit the hallmarks of the moral economy of neo-liberalism. This recommends that a toughening of the socio-legal and technical models of stochastic governance has already formed largely.

Strang and Sun (2017) applied big data software Hadoop in Google News to get complicated terrorism data. Big text search is used to code the aspects of interest into nominal areas. New areas and information into an existing database obtained from other researchers are integrated. Noticeable connection between terrorist group ideology and terrorist attack type is assumed in this study. This assumption is examined with correspondence analysis in SPSS. The analysis showed positive results; hence a symmetric model to find out the hidden relationships between terrorist ideology and attack type is developed. The main aim of the study was to demonstrate how statistical software methods possibly applied in big data analytics. These methods will aid to other researchers and practitioners. The results of a noticeable relationship between terrorist ideology and attack type perhaps support to national security planning and supply chain processes.

Tutun *et al*. (2017) defines terrorism as a complicated phenomenon with large disorders in user strategy. A new network model is formed in the framework, and the shape of the connections is examined to deduce knowledge about terrorist assaults. Particularly, an Evolutionary Simulating Annealing Lasso Logistic Regression (ESALLOR) model is presented to choose main characters for similarity performance. Followed by, a heterogeneous similarity function is introduced to evaluate the connections among assaults. In addition, graph based outbreak detection is presented to define dangerous areas for the rash of terrorism. Results of experiment indicates the effectiveness of proposed framework with more than 90% accuracy for identifying patterns when related with that of real terrorism assaults in 2014 and 2015. As a result, governments can understand automatically future

impact events of terrorism, and governments can identify terrorists' activities and strategy to reduce the risk of future attacks by applying this insightful framework.

Clustering methods:

Clustering of statistically important crimes in a particular area and detecting the problematic areas by clustering methods were discussed by (Vural *et al*., 2014; Andresen, 2011). Association of some areas with the major crime features were described in Vural *et al*. (2013, 2014) and Sun *et al*. (2014) for analyzing the crime. The research work discussed above dealt with the associations of the crime incidents within a certain geographical area. The results provided by these studies concluded that the analysis of the criminal characteristics is less studied on the perspective of spatial and temporal basis. They Ewart and Oatley (2003), Holst and Bjurling (2013) have made attempt to analyse the behavior of criminals and classified the crimes committed by the criminals over time. The criminal behavior analysis was lead to predict the results of various crime problems by evaluating the behavior of criminals over time (Ewart & Oatley, 2003). Criminal behaviors can reflect the characteristics of the criminals to a great extent. To predict the crime types according to characteristics of vast amounts of criminals is an important part of criminal behavior analysis. In Upadhyaya and Jain, (2013), Wanawe *et al*. (2014), three typical classification algorithms, including C4.5 algorithm, Naive Bayesian algorithm and K-nearest neighbor (KNN) algorithm were used for criminal behavior analysis with the solutions for finding the missing data.

Classification methods:

Decision tree and support vector classification were used in the higher rate among the other classification algorithms. Wang *et al*. (2014) used a decision tree to identify what attributes highly affect a specific type of offense and the criminal behavior and specifying offense types related to criminal groups (Al-Janabi, 2011). The similar problem has been studied by Sharma (2014) and is about the importance of data mining technology to design proactive application to detect the suspicious criminal activities. Among other classification, the decision tree represented by ID3 algorithm is widely applied. For prototype system of computer forensics, ID3 is used generally. This research analyzes advantages and disadvantages of algorithm ID3 in detail (Chen Jin *et al*., 2009). Because performance was not good, Wang *et al*. (2014) proposed an approach for constructing a decision tree-based classification model for crime prediction. Nonetheless, this method has never been used to predict criminal. Recently, following the rising interest toward the support vector machine (SVM), various studies showed that SVM outperforms other classification algorithms.

Kianmehr and Alhajj (2008) worked on the performance of SVMs for predicting the hot spot crime location when a predefined level of crime rate and a percentage for selecting a portion of that are given. The main drawback with the SVM is its high computational complexity. In this study, the Naı¨ve Bayes classifier is proposed to predict criminal of particular crime incident by using incident-level crime data. In the literature, another study about criminal is the criminal profiling (Baumgartner & Ferrari Palermo, 2008). Criminal profiling aims to estimate the criminal characteristics using the data collected from the field and narrow the list of possible suspects. These works can particularly make influence on the criminal profiling over the crime series (Pflueger et al., 2015; Rossmo et al., 2005; Ewart & Oatley, 2003), committed by the same criminal. Research suggests that the majority of crime is committed by a minority of offenders (Liu & Brown, 2003). The task of determining whether multiple crimes were committed by the same offender is challenging (Pflueger *et al*., 2015; Ewart & Oatley, 2003). The behaviors exhibited by offenders across the crimes are often used to establish crime linkages (Bennell et al., 2014; Rossmo et al., 2005). In Tollenaar and van der Heijden (2013), Toole *et al*. (2010) researchers use statistical modeling to detect changes in criminal activities. A statistically significant change in the distribution for a region may be indicative of criminal activities that need to be investigated.

The works related to the criminal prediction in the literature are very limited due to limited access to incident-level data. In order to find the criminals, the criminal data including their identities, characteristics and even their social network should be analyzed together with other available information such as GIS data. Therefore, the approaches used in the criminal prediction vary according to the description of dataset features. For example, while some works (Baumgartner & Ferrari Palermo, 2008) make association analysis based on time, crime and criminal information, some others do crime analysis or criminal profiling on the basis of crime characteristics (Canter, 2009; Baumgartner & Ferrari Palermo, 2008) , or criminal behaviors (Ewart & Oatley, 2003). In Brunsdon *et al*. (2007) , Dombroski and Carley (2002) a quick response system was proposed which can identify the most probable local suspects involved in a crime case. It maps the current location of the probable suspects using mobile GIS (Saravanan *et al*., 2013).

2.3 Prospective Mapping

Traditional methods of crime mapping, such as kernel density estimation (KDE), are used to generate risk surfaces that indicate where crime has previously clustered. As such,

they consider the location of crime events but ignore their timing. Inspired by the above findings, Bowers *et al*., (2004) proposed a method of predictive crime mapping, named ProMap, that models the way in which crime clusters (or appears to spread) in space and time. To do this, the expected risk at a location for a particular period (usually the next day, few days or the next week) is estimated as a function of the density of crime that has occurred at or near to that location. However, events are also inversely weighted according to when they occurred, so that more recent crimes receive a greater weighting.

The simplest form of the function sums the product of inverse time and distance weights given to each crime in the data set for the locations of interest (e.g. a series of grid cells). In a series of studies Bowers *et al* (2004), Johnson *et al*. (2007) have examined residential burglary, prospective mapping ('ProMap') has been shown to offer a modest but reliable predictive gain relative to KDE maps, particularly where the ProMap models used have incorporated data on housing density and the location of major roads. In an attempt to establish the usefulness of this technique in practice, a field trial of ProMap was undertaken in collaboration with the East Midlands Police (UK) and the findings disseminated via a Home Office research report (Johnson *et al*., 2007a). The application developed as part of that research identified and displayed high-risk grid squares (100 m by 100 m) against a backdrop of the street network to clearly delineate the areas of suggested intervention. Consultation with police practitioners suggested that they thought the maps were useful in an operational context. However, as is true with so many crime reduction interventions VVB (Knutsson & Clarke, 2006), problems were experienced with implementation on the ground. Due to organisational changes, resources were not devoted to using the maps in practice as much as had been anticipated prior to the study, so unfortunately it was difficult to evaluate the potential of the system in terms of real-world crime reduction.

Subsequent to the East Midlands trial, approaches based on the ProMap approach have been implemented in other areas of the UK. In Greater Manchester (UK), for example, Fielding and Jones (2012) developed their own system which followed the same principles as those described above. Over an initial 12-month implementation interval, intervention included increased guardianship provided by police patrols and other emergency service staff, and burglary was found to decline by 27 % in the treatment area. This occurred in the context of a force-wide increase of 7 %, and a reduction in burglary of 10 % in the next most similar area. Other trials have been implemented and a common feature of them is that the unit of

analysis—in terms of the mapped regions of risk—is grid-based, often with the areas of risk displayed being 50 m by 50 m or larger.

The related studies show that the crime analysis can be done efficiently with Hadoop as there is an increase in the amount of data that needs to be analysed and with this analysis the crime prediction thereby minimizing the crime rate.

2.4 Conclusion

Consequently, the involvement of very large data for analysis impel to the application of Hadoop model for data analysis. Although traditional RDBMS system can also be as used for the data analysis, however since the data is very large, the time needed for analysis would be quite long. The uses of Hadoop framework for the data analysis sustain many applications: firstly, flexibility of Hadoop model is quite appropriate for big data analysis since it can compute any format of data including structured, un-structured, multi-structured data. Second application is scalability of Linear processing of the big data. Hadoop's application for data analysis would offer great scalability for the presented system as the system consists of petabytes of data. Third is reliability of the data availability which increases the Hadoop's file system HDFS involves redundant storage of data. 3 copies of the data to increase the system reliability can be stored by the HDFS. Finally it provides economical application; Since the Hadoop framework is completely free. Hence, the adoption of Hadoop for the presented system increases the entire potential of the system.

CHAPTER III- RESEARCH METHODOLOGY

3.1 Introduction

The present chapter presents the research methodology to establish a framework for data collection and to evaluate whether the existing information is carried out in a correct manner to arriving finally at the purpose of validating new information. The main objective of this research methodology is to analyze the crime data dashboard using Hadoop Big data analytics.

The primary significance of this research methodology is to,

- discuss the research design and method,
- describe the sample data collection and storage,

- ❖ designate the process used in collecting and designing the sample data,
- ❖ gives a description of the theoretical and statistical methods used to analyze the data.
- ❖ Discuss the reliability and validity of the research tool

Proposed objective of the study:

- To design and develop a Hadoop based Crime data analysis dashboard for the analysis of crime in the London city.
- To utilize the developed system to examine crime prone zones, and total number of the crimes per year, month, quarter in the select region.

3.2 Methodology

The research methodology is composed of four stages. These are discussed as follows:

1. Loading the data
2. Sorting and searching the data
3. Processing the data
4. Analyzing the data.

The first phase, collect the semi-structured data and stored into Hadoop hdfs file system. Here the command-line interface application is used for transferring data between relational databases and Hadoop. Once the the data are in unstructured form, need to use some other components (i.e., flume, hive, pig) to load the data into hdfs.

In the second phase, will use map-reduce algorithm paradigm to sort and search the data from the collective resources. Here mainly data will be divided according to type and attributes. The Map Reduce algorithm contains two important tasks, namely Map and Reduce. In this module also used for analyzing the data set using MAP REDUCE. Map Reduce Run by Java Program.

In the third and fourth phases, will use some Hadoop components, i.e., hive, pig and impala to process the data and produce an effective result. Here, need to do some research on "how to process the data especially, if the data stored in hdfs/ Hive warehouse, or else if the data collected from the social networking websites, then how to store the data in Hadoop and

how to process the data using Hive/Impala components, finally how to connect the Hadoop with a visualisation tool''.

3.4 Research design and approach

The term research is an original contribution towards the existing stock of knowledge, thus allowing its advancement (Kothari & Garg, 2011). On the other hand, its defined as "systematic method consisting of enunciating the problem, collecting the facts or data, formulating a hypothesis, analysing the facts and reaching certain conclusions either in the form of solutions towards the concerned problem or in certain generalizations for some theoretical formulation" (Kothari & Garg, 2011). Research approaches can be broadly divided into quantitative, qualitative, and mixed methods. Quantitative research is based on the measurement of quantity or amount; qualitative research is based on non-numerical data; mixed methods fall somewhere between the other two. A detailed explanation of these approaches appears in (Creswell, 2013). Further, depending on the research purpose, research can be subdivided into exploratory research, descriptive research and explanatory research (Zhang, 2016).

The predominant purpose of this research is to develop an informative crime analysis dashboard which can able to show the crime locations, the crime rate per area and as well as month, maximum crimes per locations, identify the crime locations according to higher crime types in the respective locations, total count of the crime types according to locations and months and identify the maximum crime happening areas though geographic graph.

The major objective is to reduce the crimes and control the crime rate in the city of London police region, The research provides the dashboard based analysis which can give proper information to the police organisations to increase their focus on locations and protections near higher crime zones where maximum crimes are happening, and the proposed dashboard have to suggest the police that how the crime rate is aggrandizing and plummeting by month to month, from area to area and year to year. The current dashboard of the city of London police not up to the mark and it is not so informative to one can understand the significance of crime information. This proposed crime analysis is not only useful to police organizations, but also for citizens of the city of London region, they can protect themselves once they were aware of the crimes that which is happening mostly in their areas.

The current research work focused on the parameters that were absent in the dashboard of the London police. The existing London police dashboard can show information merely for one year; it is the main drawback to understand the flow of crime rate from the past few years. The dashboard is providing the single year count of the criminal data and percentage of criminal activities; this is also a vital flaw and it does not provide the information about the location of the crime.

The proposed work contributed its effort for making the effective and dashboard which can be easy to understood and interpreted very easily. The crime analysis dashboard will provide in detail information for the past 4 years, so one can easily understand the flow of crimes. To do this 4 different year's crime information were combined at single place. The crime analysis dashboard will make it more informative and provide the count rate of crimes for the last four years. The dashboard will give clear information regarding the crime locations, the maximum crime rate per crime location, for that location attribute was used in map reduce algorithm, which hasn't used in the old dashboard of the city of London police. The dashboard will show the geographic information on the prime locations where are more probable to happen the crimes. For that, longitude, latitude and location attributes were used. The old dashboard does not have location and count of crime information in a single table, in the proposed analysis, the dashboard will provide the count of the crimes according to locations.

It is arguable that the proposed dashboard will help the government organization as well as citizens with the clear out from the threatening of the crime frightens and aware of the criminal activities. The flowchart showed in figure 2 explains how the proposed method is working and the thematic overview of the proposed scheme. The procedure starts with loading the unstructured criminal data into the cloud environment and process the database using the Hadoop framework. Following the previous section, the contents in the database are checked for its relevant with the criminal data.

3.5 Data collection and storing

In this study, have collected the semi structure crime data from 2013-2017 for the analysis especially in London city. Moreover, uses the data that has happened before a particular date, to create cluster maps, and test its robustness for forecasting when as well as where the crime is most likely to happen. Also, will analyze which type of the crimes is happening in the city. In this database, the crime incidents are variables such as day, month,time, weather, and

location which are mapped as geographical coordinates (Big Data Startup, 2013). After collecting the input data, stored into Hadoop hdfs file system. The Hadoop distributed file system (HDFS) is a distributed file system which designed to run on the commodity hardware. HDFS is highly fault-tolerant and deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for application that has large data sets. The database for each year crime data contains the attributes respectively as follows. They are Crime ID, month, reported by, falls within, longitude, latitude, location, lsoa code, lsoa name, crime type, last outcome and context.

3.6 Data processing and analysing

After collecting the input data, it has to be processed so that meaningful information can be extracted out of it which can be served as the decision support system. In the data processing procedure will use MapReduce algorithm for parallelization of the huge data. Generally, the MapReduce functions operates on set of key/value pairs. First, the Mapper takes the input and generates a set of intermediate key/value pairs. Seond, the Reducer merges all intermediate values associated with the same intermediate key (Sharma et al., 2017). By the use of hive and pig/ impala languages the data will be processed based on MapReduce algorithm paradigm giving by tasks and jobs to the data. Here, the hive is SQL based language, research should be done on how a SQL language can use the MapReduce algorithm to process the semi-structure or unstructured data to in the processing stage. Pig is a Latin language is developed by Apache, it is a scripting language. Additionally, it does not have an inbuilt data warehouse, pig language results will be stored in hdfs only, so need to do a lot of research in these areas how can two languages handle the large data sets, how to process them. After that the result of the structured data sets will connect with visualization tool to produce in detail analytics result. If the result comes from the unstructured data sets then will be shown in a meaningful structure, i.e., tables to understand.

3.3 Framework for the proposed work

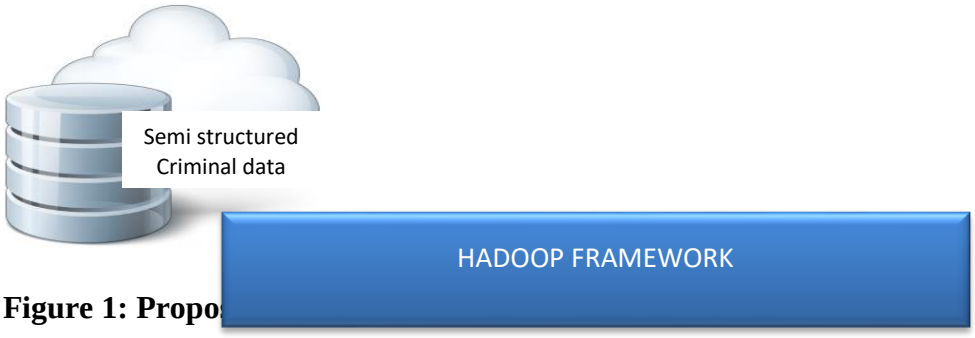


Figure 1: Propo

The gene the rit gure1. It describes about the two main structural frameworks such as Hadoop and visualisation framework. These structures t ess ses ding, Sorting and searching, Proce the obt cloud server. The Hadoop framew ces

The first S, PIG and HIVE. Hadoop distribut System (HDFS) is designed to store very l a sets at high bandwidth to user applications. In directly attached storage and execute user applic

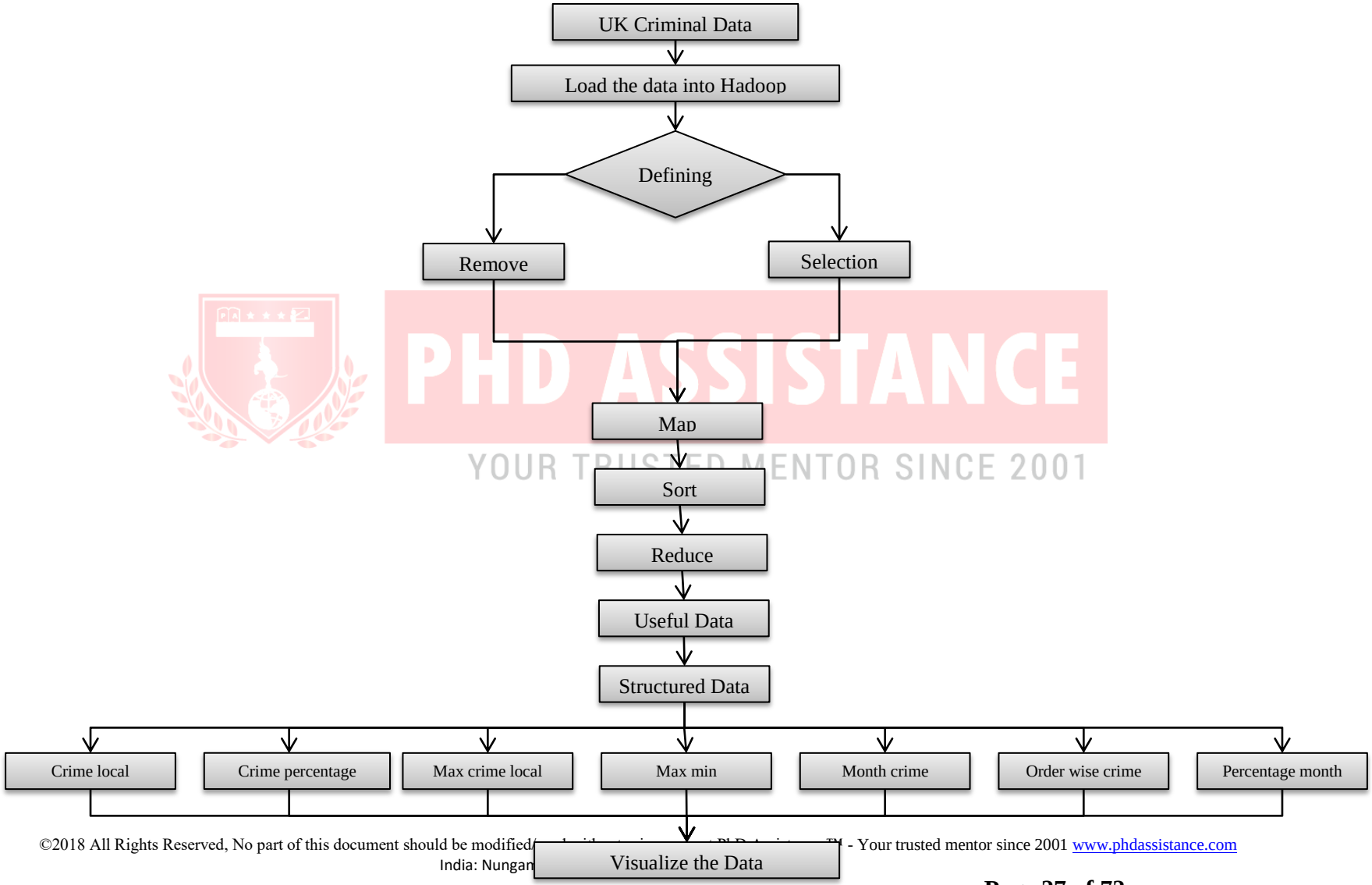
By default, Hadoop uses the Java-based system called as MapReduce. This tool is served as the main component for processing the data. Processing can occur on data stored either in an unstructured file system or in a database. MapReduce can take advantage of the locality of data; processing it near the place it is stored in order to minimize communication overhead. In the procedure of Mapping, each worker node applies the "map()" function to the local data, and writes the output to a temporary storage. A master node ensures that only one copy of redundant input data is processed. Shuffle and Reduce allows the worker nodes to redistribute the data based on the output keys and process each group of output data for each output key generated in parallel.

MapReduce allows for distributed processing of the map and reduction operations. Provided that each mapping operation is independent of the others, all maps can be performed in parallel though in practice this is limited by the number of independent data sources and/or the number of CPUs near each source. Similarly, a set of reducers can perform the reduction phase, provided that all outputs of the map operation that share the same key are presented to the same reducer at the same time, or that the reduction function

is associative. While this process can often appear inefficient compared to algorithms that are more sequential, MapReduce can be applied to significantly larger datasets than "commodity" servers can handle – a large server farm can use MapReduce to sort a petabyte of data in only a few hours. The parallelism also offers some possibility of recovering from partial failure of servers or storage during the operation, if one mapper or reducer fails, the work can be rescheduled.

Visualization framework has proved to be a great ally for finding useful patterns and information hidden in the data. It is not a new thing, since it has been used from decades to communicate information in a better and easier way. A vast amount of data is generated every day, only 20% of this is structured rest is all unstructured data, visualizing data helps in exploring and analysing data in easier way as it helps in understanding data, finding correlations, finding general trends in data etc. There exist some special visualization tools that are designed especially for handling big data like Tableau, Microsoft PowerBI, and Gephie etc. The proposed work uses the Tableau for visualization to do in-depth analysis.

Figure 2: Flow Chart for the proposed work



This decider step will help the user by defining the parameters in the criminal data whether it is relevant or not; if it is found to be not matching with the relevant data, the data will be removed from the database and the selected contents are fed into the HDFS system. The integrated file system of the HDFS will correlate the similar events, arrange and shrink the size of the data required for processing using the MapReduce algorithm. These steps will be covered in the consecutive functions such as Map, sort and reduce in the flow chart. The useful data is collected from the previous step and it's designed to get a structured data for the purpose of classification. As required the data is structured as per the commands and resulted in a database with seven different classifications named as crime local, crime percentage, max crime local, Max min, Month crime, order wise crime and Percentage month. The classified data are in the table form and it is visualised by adding the Tableau framework into the Hadoop framework. It enables the local people and the police officers to interpret the details listed in the table and provide the clear picture to the police about the criminal activities for taking the necessary actions.

3.7 Evaluation procedure

To evaluate the performance of proposed approach, will measure the processing time, precision and reliability. Techniques and technologies used effect the processing time. The challenge is to develop algorithms and technologies to increase the accuracy and precision.

3.8 Summary

This chapter discussed the methodology adopted during the study as well as the aspects of reliability and validity of data. In order for the study to be reliable and valid, information gathered must be related to the objectives and sub-objectives of the study. All attempts were made to ensure that errors were minimized and that the appropriate sample was obtained by the quantitative analysis and thus the findings of the study can be accepted with a reasonable degree of confidence. Once the data had been collected, analyses were conducted on the data. These will be covered in the following chapter.

CHAPTER IV: IMPLEMENTATION PROCEDURE

4.1 Introduction

The purpose of this study is to apply the MapReduce algorithm for analysing the London criminal data and create a dashboard for displaying the classified criminal activities. This study represented the efforts to innovatively apply the MapReduce to the study of crime, which was failed to address by many other studies. Almost all of the previous studies were microscopic, while this study was macroscopically overviewed. It did not look into individual offences in reality taking place; rather it revolved around a broad range of criminal activities and classifying it on the basis of the type of the criminal activities.

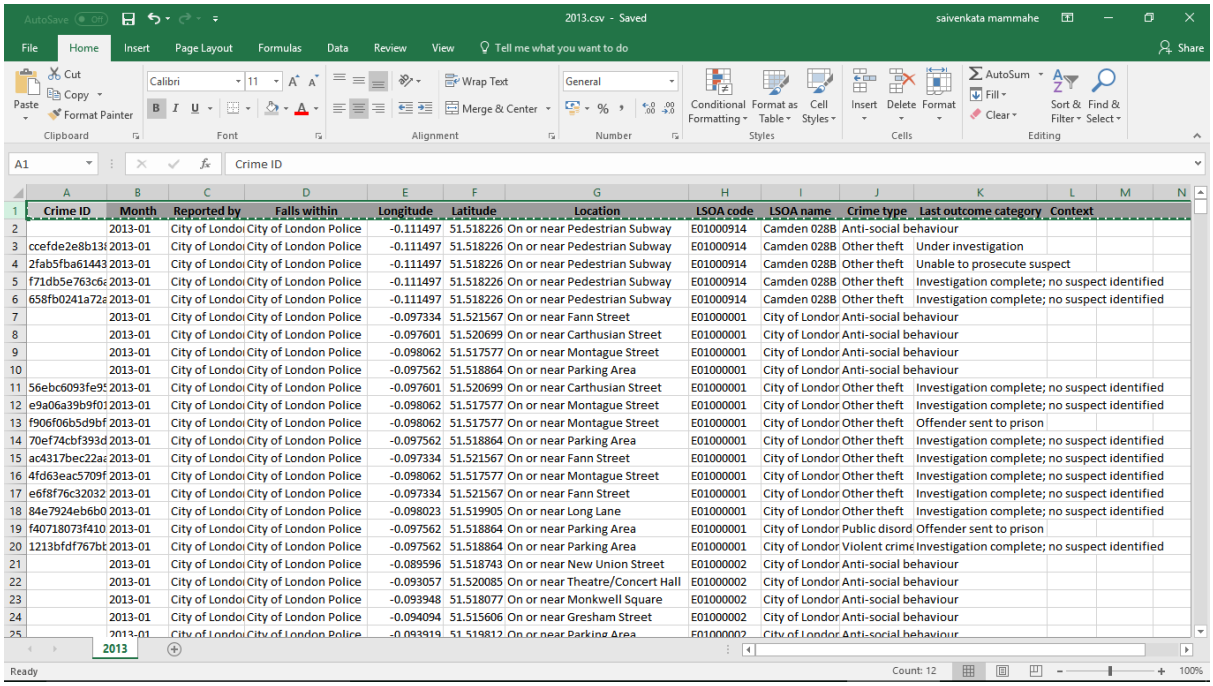
The methodology which the proposed work follows for the crime analysis of city of London police crime data is based on four basic procedures such as Data collection, Data storing and storing, Data processing and Data analysing.

4.2 Data collection

The BigData database was collected from the website of the city of London police crime data for the years ranging from 2013 to 2017 June. After that the data in the month wise were merged on the basis of the respective year as shown in figure 3. The crime data contains the attributes such as Crime ID, month, reported by, falls within, longitude, latitude, location, lsoa code, lsoa name, crime type, last outcome and context.

The semi structured data in the form of .csv format cannot be processed using the traditional database analyser. Hence the system called Hadoop is introduced and installed with the supporting software known as VMware which works on cent OS. Using this procedure, the database for the years 2013-2017 were stored on the cloud.

Table 1: Crime database



4.3 Data Storing

In data storing stage, all the data were stored into HDFS using the command given below stored the 2013 crime data into the directory using put command.

*-rw-r--r-- 3 cloudera cloudera 1474278 2017-09-17 12:55 ****/2013.csv*

The same procedure is repeated for the remaining 5 data and imported the entire file into hive shell for creating the 6 data bases called d2013, d2014, d2015, d2016, d2017 and final.

Now data base creation has completed, after this step, a year wise crime data was stored into respective year database. Below shown is an example of table creation for the crime data which is stored on cloud era desktop.

4.3.1 Table creation for crime data in database

The below created tables are root tables for the proposed result tables with the attributes such as the count of crimes according to month, year and location, percentage of crimes per month, percentage of crimes per location, max crimes, order wise crime rate per location and most of the crimes according to location.

4.4 Data Sorting

In the Hadoop, HDFS is highly scalable and reliable in the performance, HDFS replicates the data across the data nodes, these data nodes work under the name node, name node is the master node, and it distributes the jobs among the data nodes by key and pair values. Hence, there is no need of the RAID (Redundant Array of Integrated Devices) to store the data into different clusters. As like as name node and data node in the HDFS, there are two main tasks in the MapReduce as well, those are Job tracker and task tracker. Here, Job tracker is the master and Task tracker is the slave. The Job tracker useful to schedule the jobs to the task trackers, task trackers know the data location, the tasks and nodes communicate via periodic messages which are called as heartbeats. The main paradigm of the MapReduce process is the parallel processing on the nodes in the cluster. These task trackers take the input and produce the output in the form of the key and pair values. In the map phase, data nodes process each record sequentially and independently, and nodes generate the intermediate key-value pairs.

$$Map\ (k1, v1) \rightarrow List\ (k2, v2)$$

In the reduce phase, the task trackers take the input from the data nodes and it processes the merges of the intermediate values to produce the final output, in the form of the key and pair values.

$$Reduce\ (k2, List\ (v2)) \rightarrow List\ (k3, v3)$$

The output gets sorted after completion of each phase of the reduce function, and which provides the user with the aggregated output from the all the nodes. MapReduce framework automatically sorts the data into keys by using a mapper, after the mapper job, all the data will be grouped according to the keys, intermediate key-pair values generated by mapper task, In the next phase, the data will be sorted by key (not by value) in the reducer task. The data will be shuffled, sorted and organized according to values and can be grouped by the same key and pair values by the reducer, This MapReduce paradigm works behind of the hive and pig languages.

After creation of the root tables in every database, to address the proposed problems according to year wise, MapReduce paradigm is used to sort the table according to achieve the meaningful and understandable format. MapReduce works behind of the Hadoop console works according to the commands which is given in HQL language and produce the result with an infraction of the seconds.

4.4.1 Crime per location (crimeloc) table

After creation of the root table of the crime data in every database, to address the first problem that count of crimes per location, the data are sorted per location and number of the crimes as per the location using MapReduce paradigm. In the MapReduce process, mapper will take the location, count(crimetyp) data from the root table and stores the location, count(crimetyp) at a temporary location based on the same key value, meanwhile it counts the total crimes in the year, in the next phase each location will be sorted based on the same keys, in the reduce phase reduce will apply the count method over the count(crimetyp) and the combine the result of count(crimetyp) per location based on the same key and pair values.

1. **maxmin table**

Following the creation of the crimeloc table from the root table in every database, maxmin table was created. This table satisfies the proposed condition that shows the maximum and minimum crimes per location.

2. **Month crime table**

The next problem is the counts of the crime per month, to address this problem table were created in every database. This table satisfies the above-proposed condition that count of the crimes per month. Firstly the monthcrime table was created with the month and count13 attributes, which row format delimited, it means records stored in the row format, and the fields terminated by ‘,’, it means the final records which are going to store in hdfs with ‘,’ delimiter.

3. **Total percentage crime table**

After the creation of the orderwisecrime table from root data in every database, table for totalpercenrtagecrime was created, which satisfies the proposed condition that the percentage of the crimes are classified according to crime types.

4.5 Data Processing

After creation of individual tables in every database, started creating the final table using full outer joins. As like the crime loc final, the flowing final table such as *orderwisecrimefinal*, *monthcrimefinal*, *totalpercentagecrimefinal*, *maxminfinal* were created.

In the data processing, the tables which were created are gathered to get the appropriate informative result by using the MapReduce algorithm. Under the outline of the MapReduce algorithm, data tables were created in every database; when the data tables are merged with one another in year wise manner then it would be helpful to analyze the data easily with the visualization tool.

- In First phase, duplicate tables were created in default database as crimeloc13, crimeloc14, crimeloc15, crimeloc16, crimeloc17.
- In Second phase, first of all the tables were joined (crimeloc13 and crimeloc14) using full outer join and created a crimeloc1 table, the select command was inserted into crimeloc1 table.
- In third phase, crimeloc1 and crimeloc2 were joined with full outer join and created crimeloc3, and inserted the select command into crimeloc3 table.
- In the final phase, crimeloc3 was joined with crimeloc17 and created crimelocfinal table from this step. Crimeloc3 attributes are loc, cnt13, cnt14, cnt15, cnt16.

Like this all the final result tables were created using the same procedure and created below tables with the new database called final and stored all the data into final database.

In the data sorting process, 6 tables were formed for every database; so that every individual database contains same schema tables. In the first mapping process, mapped the 2013, 2014 tables with the select command, have combined both tables with full outer join connection in the combined phase. In the reduction process and stored the result of 2013 and 2014 data tables as a t1 table.

In the next mapping phase, 2015 and 2016 data tables are mapped using the select command, using full outer join connection and combined the both tables in the combiner

phase, In the reduce phase the data are stored the result of the 2015, 2016 data tables in combiner phase as a t2 table.

In the 3rd stage, mapped the both t1 and t2 tables with the select command, and the combined the both tables using full outer join and stored this combined table as t3 in the reduce phase.

In the final stage, both t3 and 2017 data tables are mapped using the select command. The tables were joined using full outer join connection in the combiner phase and stored the combined the result table as the final table in the final database.

Data Analysis and Visulaisation

After creation of the final tables, windows 7 OS was installed on VMware workstation using windows ISO file, by making this the data can be transferred lively from one VM to multiple virtual machines.

Following that tableau 14 days trail version on windows7 OS was installed. Next to that, instal the cloud era hadoop ODBC driver, this driver is useful to share the live data to one virtual machine to another virtual machine; after installation, when the cloud era hadoop databases opens in the tableau, there the server name must be given, in the proposed method where the ip address of the cloud era hadoop along with ip given at the port number as 21000.

After that automatically impala option will be activated, the connect button should be pressed, followed in the next step the final database need to be selected, in the very next step an option named import multiple tables has to be chosed; after completion of this step two options will be displayed, “connect live”, “import all and import some data”. For making any changes in the cloud era hadoop data, the ‘connect live’ option must be chose so then that would be reflected in the tableau interface as well.

PROCESSING STEPS:

After creation of individual tables in every database, the final table using full outer joins was created. Below I am showing a schema of the creation of the crimelocfinal table, as like as crime loc final, I have created the final tables for all table **(orderwisecrimefinal, monthcrimefinal, totalpercentagecrimefinal, maxminfinal)** by following the same schema and algorithm**.**

- First, duplicate tables in default database as crimeloc13, crimeloc14, crimeloc15, crimeloc16, crimeloc17 has to be created as like as follows.

MapReduce 1:

In Second phase, crimeloc13 and crimeloc14 should be merged together using full outer join and crimeloc1 table needs to be created by using the select command.

MapReduce 2:

In third phase, both crimeloc15 and crimeloc16 has to be merged using full outer join and crimeloc2 table needs to be created by using the select command.

MapReduce 3:

In the third phase, both crimeloc1 and crimeloc2 has to be merged with full outer join and and crimeloc3 table needs to be created by using the select command.

MapReduce 4:

In the final phase, the above crimeloc3 with crimeloc17 has to be joined and and crimelocfinal table needs to be created from this step. crimeloc3 table attributes are loc, cnt13,cnt14,cnt15,cnt16.

Create a replica of the table in default database	Create table crimeloc13 as
Select the table from an outside database	Select * from d2013.crimloc13
First map reducer hql script of crimeloc2013 and crimeloc2014 data combined by using full outer join on loc attribute.	SELECT crimeloc13.loc, crimeloc13.cnt13,crimeloc14.cnt14 from crimeloc13 FULL OUTER JOIN crimeloc14 ON crimloc13.loc = crimeloc14.loc;
Creation of the table for crimeloc2013&crimeloc2014 combined data	Create table crimeloc1(loc string, cnt13 int,cnt14 int)
Single row per line	Row format delimited
Columns divide with the comma delimiter	Fields terminated by ‘,’
Store the select command of the MapR 1 into table crimeloc1	Insert into table crimeloc1
Second map reducer hql script of 2015 and 2016 data combined by using full outer join on loc attribute	SELECT crimeloc16.loc, crimeloc15.cnt15,crimeloc16.cnt16 from crimeloc15 FULL OUTER JOIN crimeloc16 ON crimloc15.loc = crimeloc16.loc;
Creation of the table for 2015&2016 combined data	Create table crimeloc2(loc string, cnt15 int,cnt16 int)
Single row per line	Row format delimited
Columns divide with the comma delimiter	Fields terminated by ‘,’
Store the select command of the MapR 2 into table crimeloc2	Insert into table crimeloc2

third map reducer hql script of crimeloc1 and crimeloc2 data combined by using full outer join on loc attribute	SELECT crimeloc1.loc, crimeloc1.cnt13,crimeloc1.cnt14, crimeloc2.cnt15,crimeloc2.cnt16 from crimeloc1 FULL OUTER JOIN crimeloc2 ON crimloc1.loc = crimeloc2.loc;
Creation of the table for crimeloc1&crimeloc2 combined data	Create table crimeloc3(loc string,cnt13 int,cnt14 int, cnt15 int,cnt16 int)
Single row per line	Row format delimited
Columns divide with the comma delimiter	Fields terminated by ‘,’
Store the select command of the MapR 3 into table crime loc3	Insert into table crimeloc2
final map reducer hql script of crimeloc3 and cimeloc2017 data combined by using full outer join on loc attribute	SELECT crimeloc3.loc, crimeloc3.cnt13,crimeloc3.cnt14, Crimeloc3.cnt15,crimeloc3.cnt16,crimeloc17.cnt17 from crimeloc3 FULL OUTER JOIN crimelc17 ON crimloc3.loc = crimeloc17.loc;
Creation of the table for crimeloc1&crimeloc2 combined data	Create table crimeloc3(loc string,cnt13 int,cnt14 int, cnt15 int,cnt16 int, cnt17 int)
Single row per line	Row format delimited
Columns divide with the comma delimiter	Fields terminated by ‘,’
Store the select command of the MapR 3 into table crimelocfinal	Insert into table crimelocfinal

All the result tables followed the same schema which was used in the above procedure, The following tables was created and it was named as “final” to store all the data into the final database.

Crime percentage final table MapR hql command	Select crimepercentage3.crimetyp, crimepercentage3.cnt13, crimepercentage3.cnt14, crimepercentage3.cnt15, crimepercentage3.cnt16, crimepercentage17.cnt17 from crimepercentage3 FULL OUTER JOIN crimepercentage17 On crimepercentage3.crimetyp = crimepercentage17.crimetyp;
Maxmin final table MapR HQL script	Select maxmin3.loc, maxmin3.max13,maxmin3.max14, maxmin3.max15,maxmin3.max16 maxmin3.min13,maxmin3.min14 maxmin3.min15,maxmin3.min16 maxmin17.max17,maxmin3.min17 FROM maxmin3 FULL OUTER JOIN maxmin17 ON maxmin3.loc = maxmin17.loc;
Month crime final table MapR HQL script	Select montcrime3.month,monthcrime3.cnt13, monthcrime3.cnt14, monthcrime3.cnt15, monthcrime3.cnt16, monthcrime17.cnt17 FROM monthcrime3 FULL OUTER JOIN monthcrime17 ON monthcrime3.month = monthcrime17.month;
Orderwise crime final table MapR HQL script	Select orderwisecrime3.crimetyp, orderwisecrime3.cnt13, orderwisecrime3.cnt14, orderwisecrime3.cnt15, orderwisecrime3.cnt16, orderwisecrime17.cnt17 from orderwisecrime3 FULL OUTER JOIN orderwisecrime17 ON orderwisecrime3.crimetyp = orderwisecrime17.crimetyp;
Percentagemonth final table MapR HQL script	Select percentagemonth3.month,percentagemonth3.cnt13,

	percentagemonth3.cnt14, percentagemonth3.cnt15, percentagemonth3.cnt16, percentagemonth17.cnt17 from percentagemonth3 FULL OUTER JOIN percentagemonth17 ON percentagemonth3.month = percentagemonth17.month;

Summary:

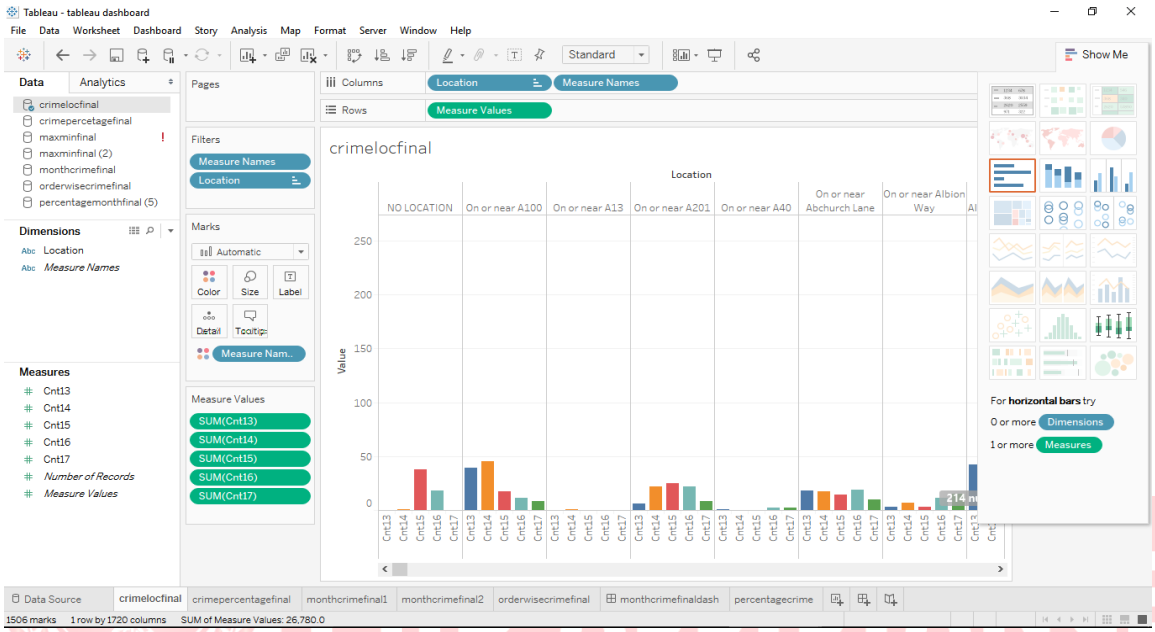
In the implementation process, semi-structured crime data has converted into the meaningful data tables through the data collection, data storing and sorting and data processing process. In data analysis phase the meaningful data tables converted into visualization result, through these worksheets one can easily understand the crime-prone zones and time period, which is the most crime happening month, which are the crimes rate are aggrandizing from the last 4 years. The main difference between my crime analysis dashboard and current city of london police dash board are, city of london police dash board provides very limited information, it shows the just one year crime count and percentage information, meanwhile, It does not show location crime data information, When it is come to my crime analysis dashboard, it gives the clear information of the crime rate from the past 4 years, and it provides the crime information by location in my dashboard, Moreovere the dashboard will show the geo graphic locations where the crime rates are high through visual map, thesea are the main advantages with the dashboard which will provide the more information regarding the crimes, so this will help to the citizens and security organisations to provide security in those locations and the particular months. the next chapter will discuss about the results, and will provide some suggestions regarding the crime preventions based on the tableu worksheets.

CHAPTER V: RESULTS AND DISCUSSION

5.1 Introduction

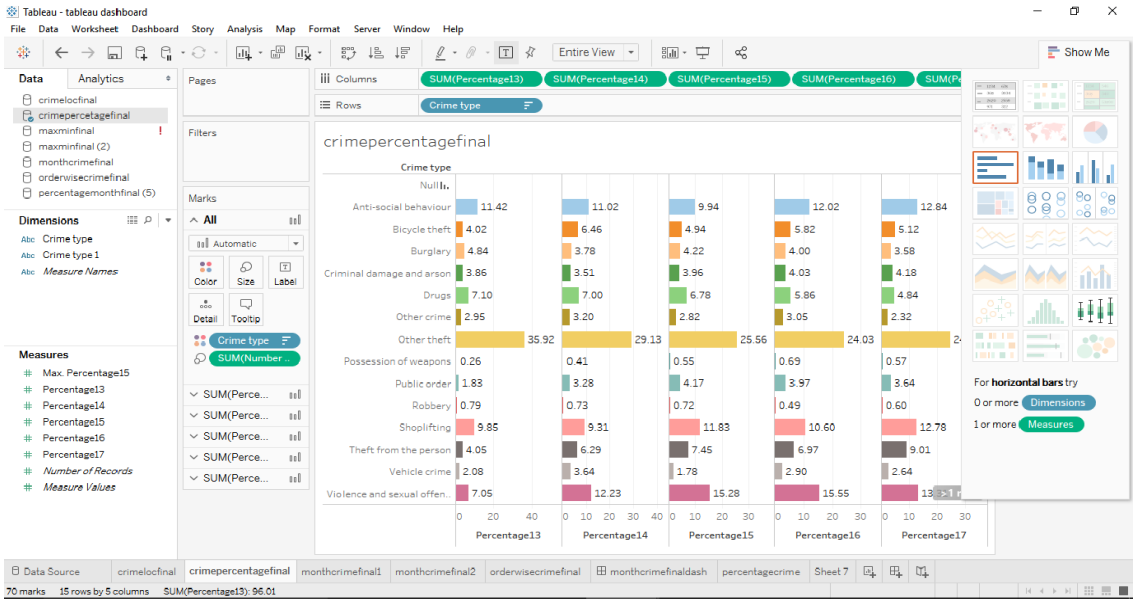
5.2.1 Prevalence of Crime rate based on Location

Crimelocfinal



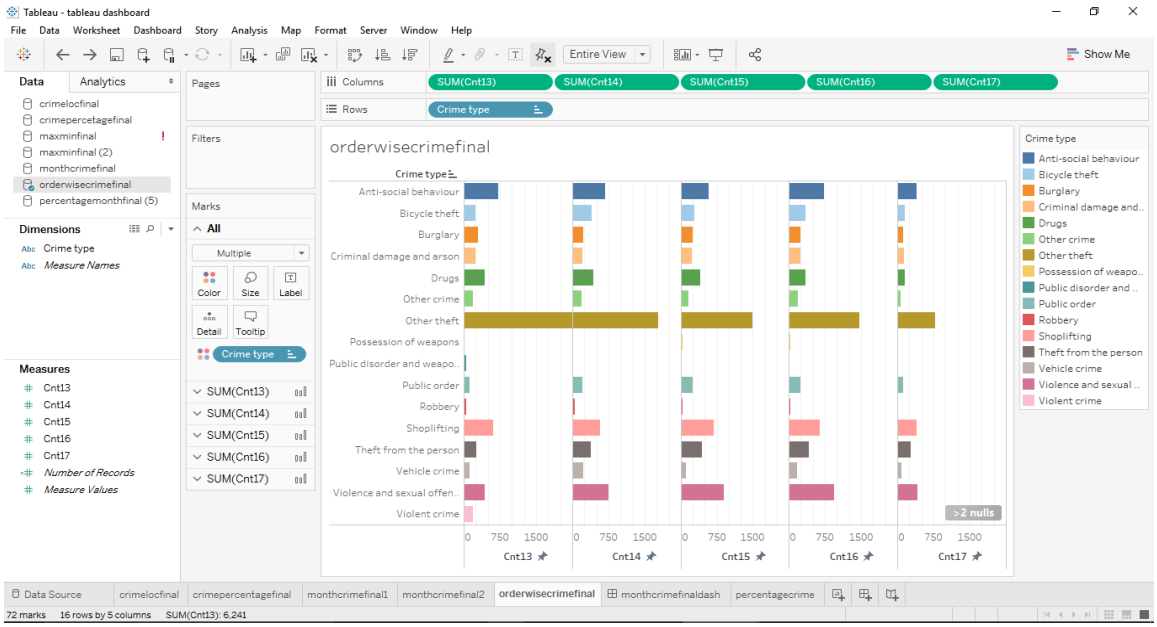
On the bar graph worksheet, how was the total number of crime rate between 2013 and 2017 (June) in every location could be easily observed. So, from this graph one can identify the locations where the crime rate has been increasing from the last 4 years, this graph will help the police organisations to identify the most crimes happening areas to protect citizens from crimes.

5.2.2 Crime Percentage Final



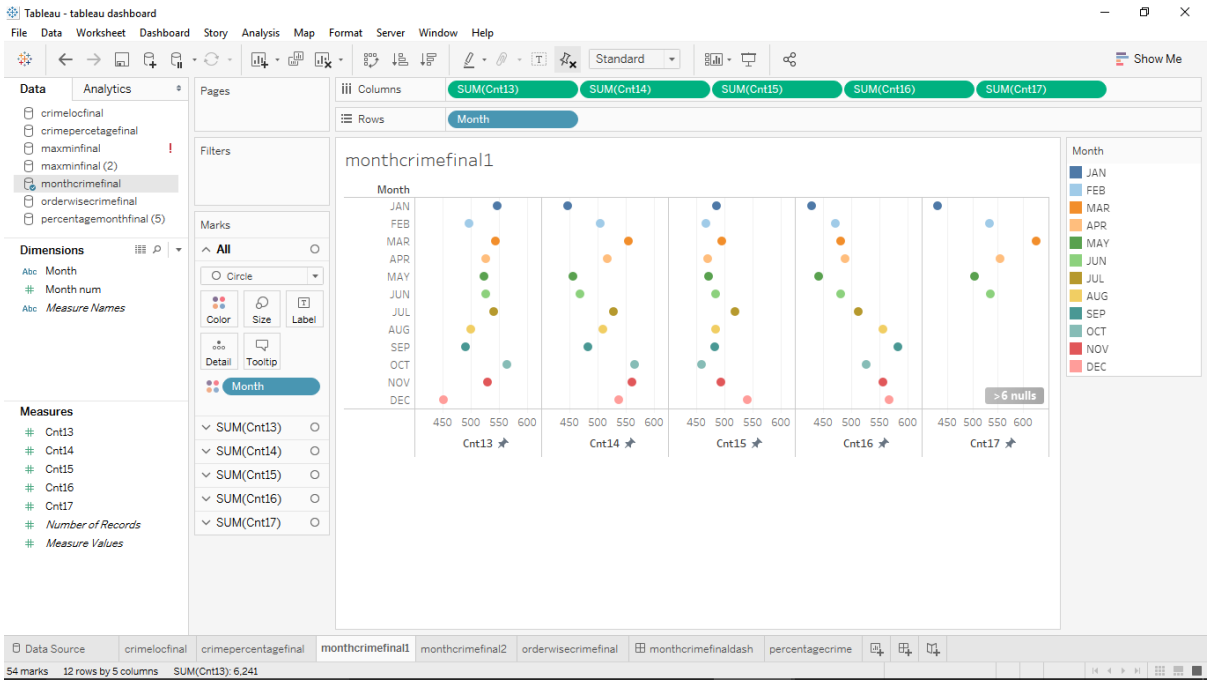
The above worksheet explains the crime types and the percentage of the crime types from 2013 to 2017. This graph is a predictive graph about the crime rate of the 2017. Might be there are some chances to increase the crime rate as mentioned in the graph. This explains which crimes are happening most and how much percentage those are contributing in whole crime percentage of the year. So, from above graph one can understand the crime rate in the percentages, so it would be helpful to police officers to understand the crime increasing and decreasing rate according to crime type, so they can focus on crimes.

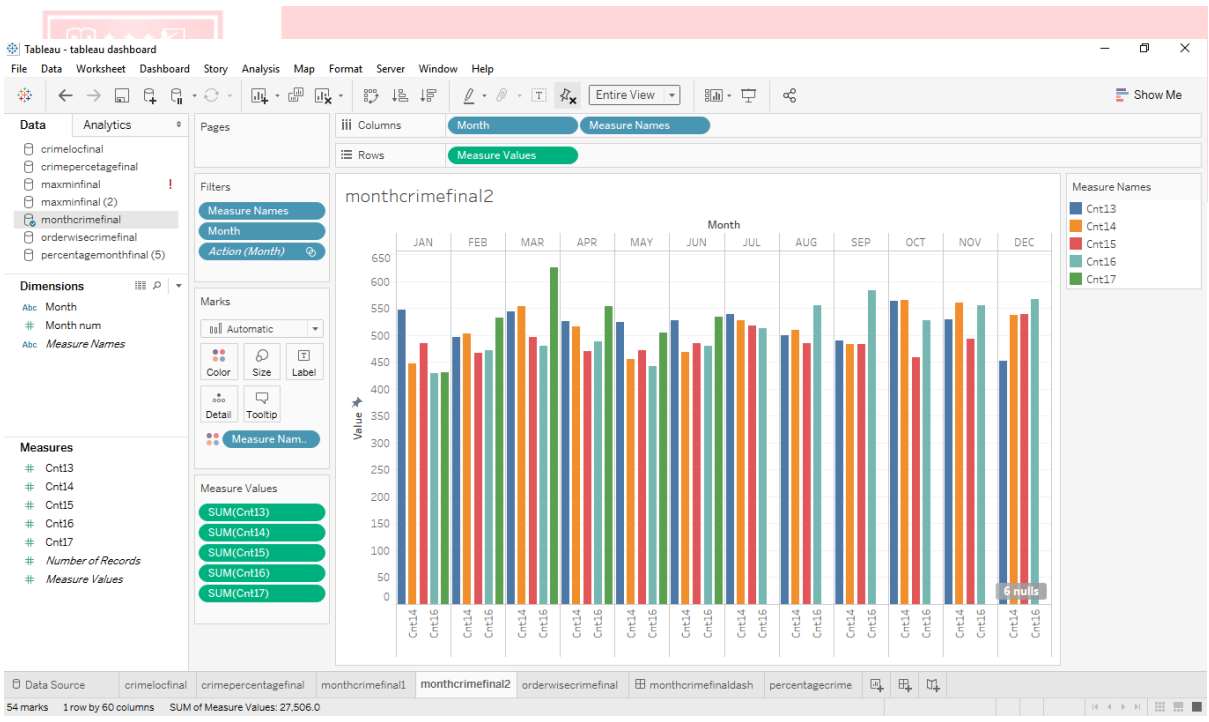
5.2.3 Order wise crime final



The above worksheet explains the crime rate and the number of crimes from 2013 to 2017 June. From above graph, one can understand the increasing and decreasing of the crimes in number value. This graph will help the security agencies to understand which crime is increasing from the past 4 years in numbers and which is the decreasing in numbers, the 2017 data included from only jan to till june 2017)

5.2.4 Month crime final

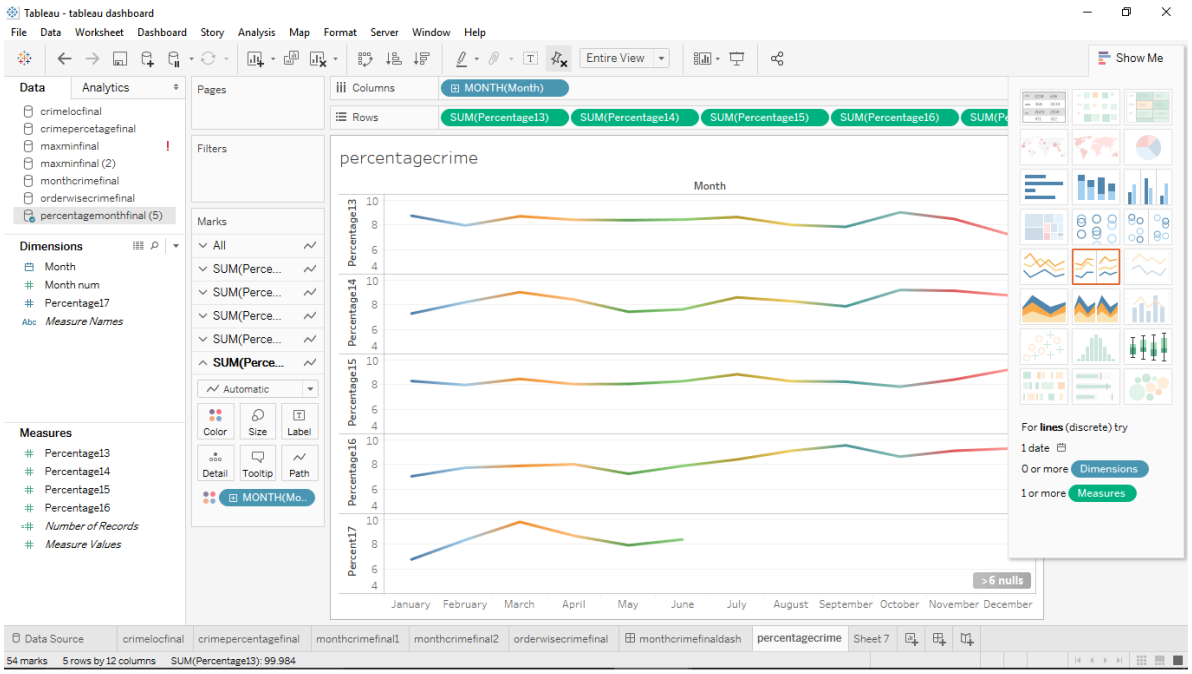




Above two graphs will explain the crime rate according to months from 2013 to 2017. it shows the how the crime rate is fluctuating in every month of year. Here one can observe that how many number of crimes have been happening in every month from 2013 to 2017

June. These graphs will help the police organisations to identify the most crimes happening months, and then can increase the security aspects more months.

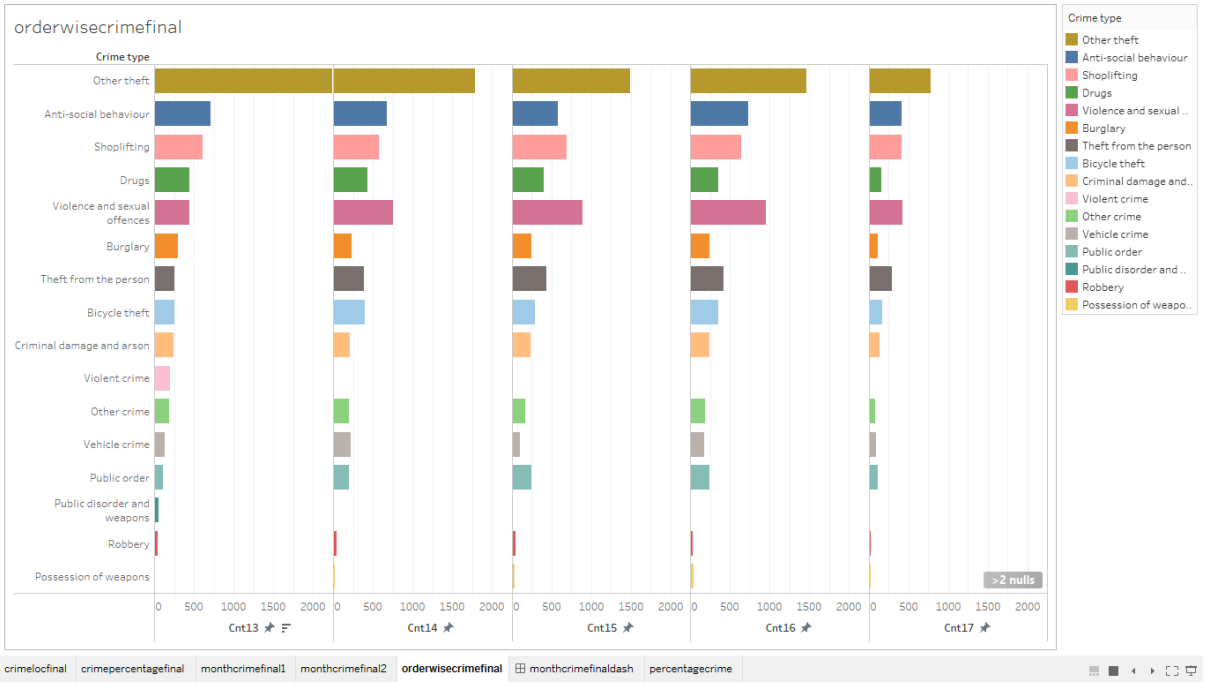
5.2.5 Percentage Month Final



The above graph is percentage month final, this graph explains the percentage of the crime rate according to month from 2013 to 2017 june. Here it is observed the fluctuating in the percentage of the crime rate from one month to other month; this graph will help to understand which months are well known for crimes, so this will help to security agencies to increase the security in the particular months.

From the above avg crimes table, police organisations must take more actions In last quarter of the year, these months are having more crime rate than other months, after these July and august months are also well known for the crime rate, so police people must put more concentration in these two months. Even though first quarter is having the less crime rate rather than other two quarters, the march month stands at first place in the avg crime rate, this month is having more crime rate than any other months, so police and citizens of the city of London police region must be more careful in the march month.

5.2.6 Orderwise Crime Table



From the order wise crime table, it is understood that the crime rate order according to crime headers, as seen in above table, in every year from 2013 to 2017 the other theft crime always stands at first place in the crime rate, after other theft, violence and sexual assault stands in the second position in the more number of the crimes. In the third position, antisocial behaviour crime will come in the number of the crimes, but from the last few years the number of anti-social behaviour crimes is decreasing. The very next to antisocial behaviour, shoplifting will be the crime type which is having more crime rate. After these all drugs and bicycle theft crimes are having almost same number of the crime rate. Even other crime, vehicle crime and public order crimes are having the almost same crime rate, but when it is come to possession weapons, burglary and violent crimes are standing just above to the negligible values in the crime rate.

No.	Order of the crimes according to crime rate.
1	Other theft
2	Violence and sexual assaults
3	Anti-social behaviour

4	shoplifting
5	Drugs and bicycle theft
6	Other crime, vehicle crime and public order
7	Possession weapons and burglary
8	Violent crime

Above order wise crime table will helpful to the citizens and police organisations to understand which crime type are standing at top of list, and then people will understand the crime possibility at their location. Other side, government should concentrate about the top of the crimes which are in the list, to provide more security to the citizens of the city of London police region.

Analysis:

The crimelocfinal table proposed in the study is given in the table 5.1, this table indicates the areas which are most crime prone zones. From the obtained result the database was divided the crime probability areas into 3 sections. They are

- Extreme probability areas (more than 100 + avg crimes per year)
- High probability areas (50 -100 avg crimes per year)
- Medium probability areas(about to 50 avg crimes per year)

Table 2: Crime prone zones

Extreme probability areas	**High probability areas**	**Medium probability areas**
On or near Bloomfield street	On or near Artillery lane	On or near A100
On or near conference/Exhibition	On or near Bear Alley	On or near Alderman's walk
On or near fish street	On or near Bell inn yard	On or near Aldgate
On or near Gravel Lane	On or near Bride lane	On or near Austin friars

On or near new change	On or near Clément's Lane	On or near bow lane
On or near night club	On or near fleet street	On or camomile street
On or near parking area	On or near grace church street	On or near cannon street
On or near pedestrian subway	On or near Leadenhall street	On or near Carthusian street
On or near police station	On or near primrose street	On or near Cheapside
On or near st Martin's Le Gra	On or near queen Victoria street	On or near creed lane
On or near Super market	On or near rood lane	On or near distaff lane
	On or near shopping area	On or near east cheap
		On or near Eldon street
		On or near fann street
		On or near fetter lane
		On or near finch lane
		On or near Finsbury avenue
		On or near further/higher street
		On or near great tower street
		On or near Gresham street
		On or near hospital
		On or near lime street
		On or near little Somerset

		On or near mark lane
		On or near moor lane
		On or near Moorgate
		On or near new broad street
		On or near park open space
		On or near Philpot lane
		On or near primrose lane
		On or near pudding street
		On or near Southampton bay
		On or near swithin's lane
		On or near Watling court
		On or near wormwood street

Table 5.1 gives information regarding the most crime prone zones, extreme probability zones are nothing but these areas having 100+ avg crimes per year, so police organisations should provide more security at these areas. High probability areas are also having the more crimes from the last 5 years, in these areas avg crime rate is between 50 -100 crimes per year. Medium probability zones are having 25 – 50 avg crimes per year, so the citizens of the London should be careful in above mentioned areas.

5.2 Crime percentage result and discussion

Table 3: Crime percentage dashboard

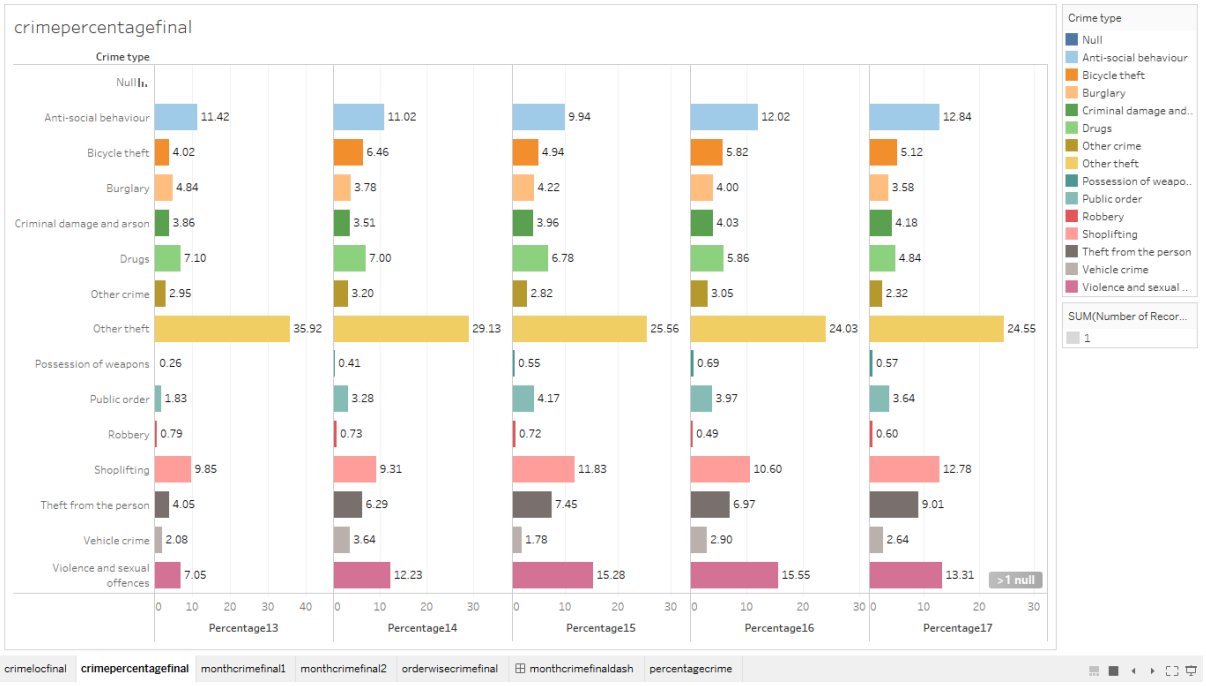


Figure 2 shows that even the ordinary citizen can able to understand one thing that the top most crimes of city of London police are other theft, anti-social behaviour, shoplifting violence and, violence and sexual assault. So, police agencies should put more concentration among these crimes. Now we will discuss deeply into it, the first crime is anti-social behaviour has increased consistently from the 2013 to 2017, So this crime rate is increasing from the last 3 years, the next crime is bicycle theft is the one of the small crimes then others, the bicycle theft cases are slightly decreasing from the past two years, this is a good sign to the police and citizens, the next crime is burglary crime rate is maintain the steady flow from the past 3 years, but when compare with the last two years this year this crime rate is sank gradually. The criminal damage and arson also smaller crime then other crimes, though this crime rate has risen significantly from the year of 2013 to 2017. Possession of drugs is the dangerous crime in this modern world, but shockingly this crime rate slipped back from last 4 years. Other crime is one the crimes which are having less contribution among the crime rate, the other crime rate also decreasing marginally from the last 2 years. Other theft is the largest crime in city of the London police areas, even though the crime rate of the other theft has been plummeting from the 2013 to 2017, shoplifting one the major crimes in this area, this crime rate has surged significantly from the last 3 years. Along with shoplifting, theft from person crime also increased significantly from the last two years. In the contrast to above crimes, violence and sexual assault also one of the largest crimes in

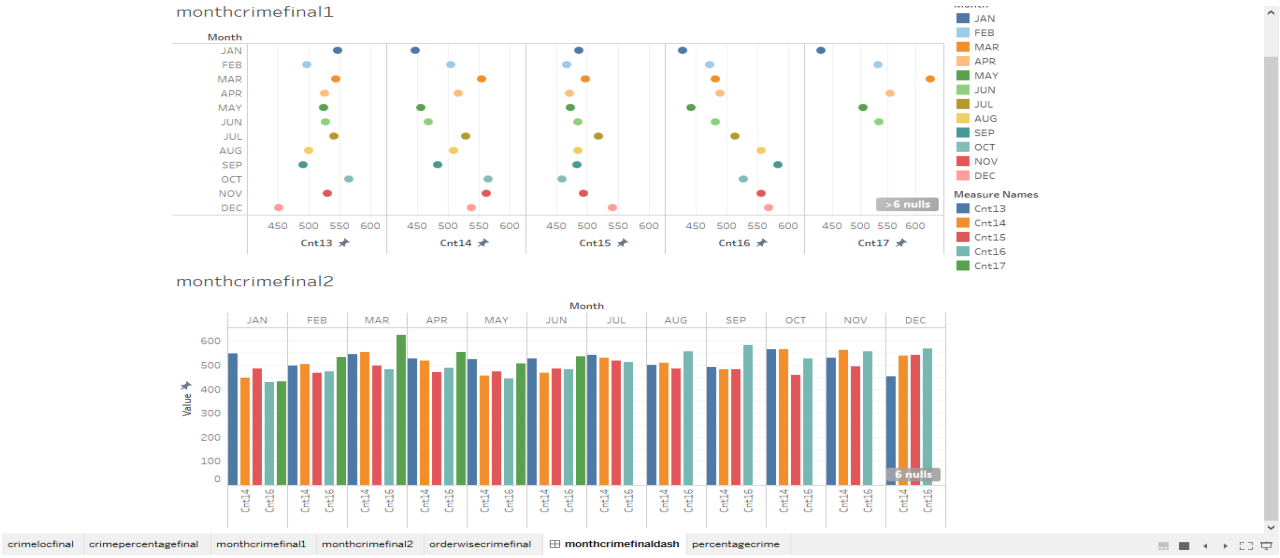
these areas, but this crime rate has dipped significantly in current when compare with the last two years,

Table 4: Statistics on Criminal activities

Anti-social behaviour	Increasing slightly
Bicycle theft	Maintain constant flow
Burglary	Decreased marginally
Criminal damage and arson	Risen gradually
Drugs	Dropped slightly
Other crime	Decreasing significantly
Other theft	Plummeting slightly but the largest crime
Possession weapons	A significant rise happened from last 2yr's
Public order	Increased moderately
Robbery	Decreased marginally
Shoplifting	Climbed slightly
Theft from the person	Aggrandised gradually
Vehicle crime	Increasing slightly
Violence and sexual assaults	Decreasing gradually from the last 1 year.

Table 4 gives clear information about the Police organisations need to concentrate more on the anti-social behaviour, criminal damage and arson, possession weapons, public order, shoplifting and Theft from person crimes. These crimes rate increasing significantly from the last few years, even though, other theft crime is decreasing, but this is the largest crime among the all, so police people need to concentrate on this crime more than other crimes.

5.3 Monthcrime



From the monthcrime table we can understand that, the crime rate is more in the last quarter of the years, First, in the first quarter, crime rate in the January month is decreasing significantly from 2013 to 2017, it was about to 550 in 2013 and it has dropped to just below of 450 crimes in the 2017.When it is come to February month, the crime rate was below 500 from 2013 to till 2016 and it has increased significantly in the 2017 year with more than 525 crimes. The crime rate in the march never been constant from the last 4 years, It is always more crimes happening month among all from the last 5 years except in 2015 and 2016 years, but in 2017 this month got the largest crime number with the 625+ crimes. along with march, April also stands in the largest crime happening months in the first quarter of every year, the crime rate always more than 550 in most of the years, in the current running year the nearly 560 crimes happened in the April month.

In the second quarter of the year, this is the one of the quarters of the year which has more crime rate than the first quarter, following to it, the crime rate in the July and august months above than the 500+ crimes, but when it comes to may an July month little back than other two months, in the may month crime rate is decreasing significantly from 2013 to till 2017 and it has dropped down and reached just above of the 500. but the coming three months such as June, July and august having 520+ crime rate in 2017.

In the last quarter of the year, the crime rate is higher than other two quarters in every year, September Is the only month which has less crime rate than other three months in this quarter.

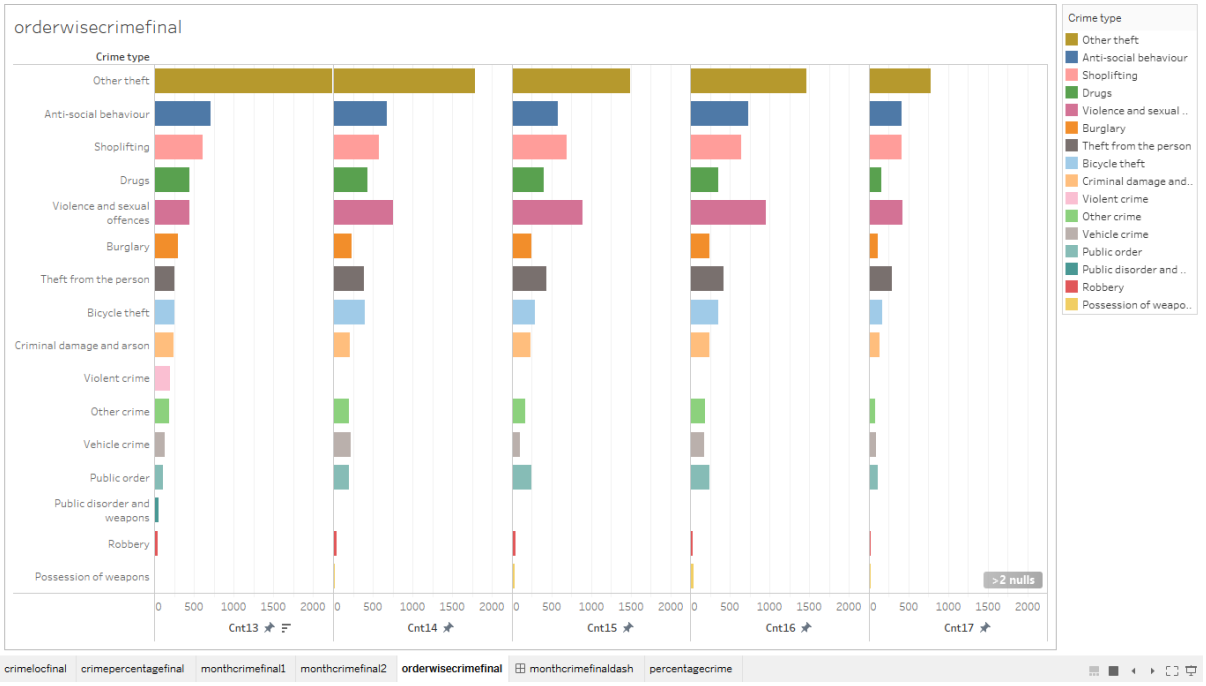
The avg crime rate of the October month is 525+ this is the one the crime happening month in every year. The avg crime rate is slight more in November month when compare with October, the avg crime rate of the November is 535, which is just above than October month. Even December month also stands behind of November month in avg crime rate December and October month's stands at same place with 525+ avg crimes.

Month	**AVG crime rate Descending order**
March	540
November	535
October	528
December	525
July	524
August	512
April	511
September	509
June	499
February	494
May	479
January	468

From the above avg crimes table, police organisations must take more actions In last quarter of the year, these months are having more crime rate than other months, after these July and august months are also well known for the crime rate, so police people must put more concentration in these two months. Even though first quarter is having the less crime

rate rather than other two quarters, the march month stands at first place in the avg crime rate, this month is having more crime rate than any other months, so police and citizens of the city of London police region must be more careful in the march month.

5.4 Orderwise Crime Table



From the order wise crime table, it is understoodd the crime rate order according to crime headers, as can see above table, in every year from 2013 to 2017 the other theft crime always stands at first place in the crime rate, after other theft, violence and sexual assault stands in the second position in the more number of the crimes. In the third position, antisocial behaviour crime will come in the number of the crimes, but from the last few years the numbers of anti-social behaviour crimes are decreasing as it can be seen from the above worksheet. The very next to antisocial behaviour, shoplifting will be the crime type which is having more crime rate. After these all drugs and bicycle theft crimes are having almost same number of the crime rate. Even other crime, vehicle crime and public order crimes are having the almost same crime rate, but when it is come to possession weapons, burglary and violent crimes are standing just above to the negligible values in the crime rate.

Table 5: Crime classification

No.	Order of the crimes according to crime rate.
1	Other theft
2	Violence and sexual assaults

3	Anti-social behaviour
4	shoplifting
5	Drugs and bicycle theft
6	Other crime, vehicle crime and public order
7	Possession weapons and burglary
8	Violent crime

Above order wise crime table will helpful to the citizens and police organisations to understand which crime type are standing at top of list, and then people will understand the crime possibility at their location. Other side, government should concentrate about the top of the crimes which are in the list, to provide more security to the citizens of the city of London police region.

CHAPTER VI: CONCLUSION

Crime being a social menace that should be controlled, and it ends up plainly dominant in the advancement of social orders all around the globe and furthermore for the welfare of the general populace. The present research means to make a Crime investigation dashboard for London city police to boost security perspectives in the country. The scrutiny concentrates on the pint-sized region little part in of the London city where the populace is not more than 10,000 which is secured by the London city police. The developing difficulties must be met and taken care of effectively as it is inescapable in the each foundation. The primary significance of this research methodology was to discuss the research design and method in sample data collection and storage followed by the process used in collecting and designing the sample data, which gives a description of the theoretical and statistical methods used to analyze the data. Acceptable for the study to be reliable and valid, sample that was gathered must be related to the intentions of the study and also ensure that errors were minimized by the quantitative analysis and thus the findings of the study can be accepted with a reasonable scale of certainty. This present study based on cloud using Hadoop was exercised in the small scale which seems to reduce the crime in the London city in a considerable manner. Researches should be done further in order to utilize the methodology in a global manner for the crime investigation and control.

References

Abdullah, A., Salleh, M.N.M. & Sakip, S.R.M. (2012). Fear of Crime in Gated and Non-gated Residential Areas. *Procedia - Social and Behavioral Sciences*. [Online]. 35. p.pp. 63–69. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1877042812003758.

Ahmadi, M. (2003). *Crime Mapping and Spatial Analysis*. International Institute for Geo-Information Science and Earth Observation Enschede, The Netherlands.

Al-Janabi, K. (2011). A proposed framework for analyzing crime data set using decision tree and simple k-means mining algorithms. *Journal of Kufa for Mathematics and Computer*. 1. p.pp. 8–24.

Andresen, M.A. (2011). Estimating the probability of local crime clusters: The impact of immediate spatial neighbors. *Journal of Criminal Justice*. 39 (5). p.pp. 394–404.

Ansari, S.M. & Kale, K. V. (2014). Mapping and Analysis of Crime in Aurangabad City using GIS. *IOSR Journal of Computer Engineering (IOSR-JCE)*. [Online]. 16 (4). p.pp. 67–76. Available from: http://www.iosrjournals.org/iosr-jce/papers/Vol16-issue4/Version-7/L016476776.pdf.

Barbieru, C. & Pop, F. (2016). Soft Real-Time Hadoop Scheduler for Big Data Processing in Smart City. In: *IEEE 30th International Conference on Advanced Information Networking and Applications*. [Online]. 2016, Crans-Montana, Switzerland: IEEE, pp. 863–870. Available from: http://www.diva-portal.org/smash/get/diva2:1105669/FULLTEXT01.pdf.

Baumgartner, K. & Ferrari Palermo, G. (2008). Constructing Bayesian networks for criminal profiling from limited data. *Knowledge-Based Systems*. 21. p.pp. 563–572.

Bennell, C., Mugford, R., Ellingwood, H. & Woodhams, J. (2014). Linking Crimes Using Behavioural Clues: Current Levels of Linking Accuracy and Strategies for Moving

Forward. *Journal of Investigative Psychology and Offender Profiling*. 11 (1). p.pp. 29–56.

Big Data Startup (2013). *The Los Angeles Police Department Is Predicting and Fighting Crime with Big Data*. [Online]. 2013. Big Data Startup. Available from: http://esatjournals.net/ijret/2015v04/i04/IJRET20150404061.pdf. [Accessed: 12 September 2017].

Bowers, K. (2014). Risky Facilities: Crime Radiators or Crime Absorbers? A Comparison of Internal and External Levels of Theft. *Journal of Quantitative Criminology*. [Online]. 30 (3). p.pp. 389–414. Available from: http://link.springer.com/10.1007/s10940-013-9208-z.

Bowers, K., Johnson, S. & Pease, K. (2004). Prospective Hot-Spotting: The Future of Crime Mapping. *British Journal of Criminology*. 44 (5). p.pp. 641–658.

Brunsdon, C., Corcoran, J. & Higgs, G. (2007). Visualising space and time in crime patterns: a comparison of methods. *Comput Environ Urban Syst*. 31. p.pp. 52–75.

Budd, T. (2001). *Practice Messages from the Britsh Crime Survey*. London.

Canter, D. (2009). Developments in geographical offender profiling: commentary on Bayesian journey-to-crime modelling. *Journal of Investigative Psychology and Offender Profiling*. 6 (3). p.pp. 161–166.

Chen Jin, Luo De-lin & Mu Fen-xiang (2009). An improved ID3 decision tree algorithm. In: *2009 4th International Conference on Computer Science & Education*. July 2009, IEEE, pp. 127–130.

Christopher, M. (2016). *Using Big Data and Hadoop Ecosystems to Fight Crime and Inequality*. [Online]. Chicago. Available from: http://www2.rmcil.edu/dataanalytics/v2016/papers/Using_Big_Data_and_Hadoop_Ecos

ystems_to_Fight_Crime_and_Inequality.pdf.

Creswell, J.W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*. New York & London: Sage publications.

Dombroski, M.J. & Carley, K.M. (2002). Netest: estimating a terrorist network's structure. *Computational and Mathematical Organization Theory*. 8 (3). p.pp. 235–241.

ESRI (2008). *Crime Analysis: GIS Solutions for Intelligence-Led Policing*. [Online]. California: ESRI. Available from: http://www.esri.com/library/brochures/pdfs/crime-analysis.pdf.

Ewart, B.W. & Oatley, G.C. (2003). Applying the Concept of Revictimization: Using Burglars' Behaviour to Predict Houses at Risk of Future Victimization. *International Journal of Police Science and Management*. 5 (2). p.pp. 69–84.

Fielding, M. & Jones, V. (2012). 'Disrupting the Optimal Forager': Predictive Risk Mapping and Domestic Burglary Reduction in Trafford, Greater Manchester. *International Journal of Police Science and Management*. 14 (1). p.pp. 30–41.

Fyfe, J. & Wilson, O. (1997). *Police administration*. New York: McGraw-Hill.

Ghosh, D., Ae Chun, S., Shafiq, B. & Adam, N.R. (2016a). Big Data-based Smart City Platform. In: *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research - dg.o '16*. [Online]. 2016, New York, New York, USA: ACM Press, pp. 58–66. Available from: http://dl.acm.org/citation.cfm?doid=2912160.2912205.

Ghosh, D., Chun, S.. & Shafiq, B. (2016b). Big Data-based Smart City Platform: Real-Time Crime Analysis. In: *17th International Digital Government Research Conference on Digital Government Research.* [Online]. 2016, Shanghai, China: Digital Government Research, pp. 58–66. Available from: http://www.diva-

portal.org/smash/get/diva2:1105669/FULLTEXT01.pdf.

Gwinn, S., Bruce, C., Hick, S. & Cooper, J. (2008). *Exploring crime analysis*. Overland Park, Kan: International Association of Crime Analysts.

Holst, A. & Bjurling, B. (2013). A Bayesian Parametric Statistical Anomaly Detection Method for Finding Trends and Patterns in Criminal Behavior. In: *2013 European Intelligence and Security Informatics Conference*. August 2013, IEEE, pp. 83–88.

Jain, A. & Bhatnagar, V. (2016). Crime Data Analysis Using Pig with Hadoop. *Procedia Computer Science*. [Online]. 78. p.pp. 571–578. Available from: http://linkinghub.elsevier.com/retrieve/pii/S187705091600106X.

Johnson, S., Birks, D., McLaughlin, L., Bowers, K. & Pease, K. (2007a). *Prospective mapping in operational context: Home Office Online Report*. London: Home Office.

Johnson, S.D., Bernasco, W., Bowers, K.J., Elffers, H., Ratcliffe, J., Rengert, G. & Townsley, M. (2007b). Space–Time Patterns of Risk: A Cross National Assessment of Residential Burglary Victimization. *Journal of Quantitative Criminology*. 23 (3). p.pp. 201–219.

Josep, I.S., Subramani, B.., Harish, T. & Raj, M. (2017). A Survey of Data Mining Techniques for Analyzing Crime Patterns. *International Journal of Engineering Science and Computing*. [Online]. 7 (3). p.pp. 5721–5723. Available from: http://ijesc.org/upload/78a7afea0768dcab05be2b37731764c5.A Survey of Data Mining Techniques for Analyzing Crime Patterns.pdf.

Joseph, S.I., Bala Subramani, C., Harish, T. & Raj, M. (2017). A Survey of Data Mining Techniques for Analyzing Crime Patterns. *International Journal of Engineering Science and Computing*. [Online]. 7 (3). p.pp. 5721–5723. Available from: http://ijesc.org/upload/78a7afea0768dcab05be2b37731764c5.A Survey of Data Mining Techniques for Analyzing Crime Patterns.pdf.

Kedia, P. (2016). *Crime Mapping and Analysis using GIS*. [Online]. 2016. Geomatics 2000: Conference on Geomatics in Electronic Governance. Available from: https://www.researchgate.net/publication/309125859_Crime_Mapping_and_Analysis_using_GIS?channel=doi&linkId=57ffe91608ae32ca2f5db4da&showFulltext=true. [Accessed: 16 August 2017].

Keyvanpour, M.R., Javideh, M. & Ebrahimi, M.R. (2011). Detecting and investigating crime by means of data mining: a general crime matching framework. *Procedia Computer Science*. [Online]. 3. p.pp. 872–880. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1877050910005181.

Kianmehr, K. & Alhajj, R. (2008). Effectiveness of Support Vector Machine for Crime Hot-Spots Prediction. *Applied Artificial Intelligence*. 22 (5). p.pp. 433–458.

Knutsson, J. & Clarke, R. (2006). *Putting theory to work: implementing situational prevention and problem-oriented policing, vol. 20 of crime prevention studies*. Monsey: Criminal Justice Press.

Kothari, C.. & Garg, G. (2011). *Research methodology: methods and techniques*. [Online]. 2011. New Age International. Available from: https://www.diva-portal.org/smash/get/diva2:1046794/FULLTEXT01.pdf. [Accessed: 12 September 2017].

Li, X. (2014). *Application of Data Mining Methods in the Study of Crime Based on International Data Sources*. [Online]. Tampere University Press. Available from: http://uta32-kk.lib.helsinki.fi/bitstream/handle/10024/95108/978-951-44-9419-2.pdf?sequence=1.

Liu, H. & Brown, D. (2003). Criminal incident prediction using a point-pattern-based density model. *International Journal of Forecasting*. 19. p.pp. 603–622.

Malarvizhi, S. & Ibrahim, S. (2017). An Efficient Clustering for Crime Analysis. *International Journal of Innovative Research in Computer and Communication Engineering*. 5 (2). p.pp. 1302–1309.

Malathi, A. & Santhosh Baboo, S. (2011). An Enhanced Algorithm to Predict a Future Crime using Data Mining. *International Journal of Computer Applications*. 21 (1). p.pp. 1–6.

Minocha, S., Kumar, J., Singh, H., Bawa, S. & Engineering, N.C.C. (2016). HadoopWeb : MapReduce Platform for Big Data Analysis. *International Research Journal of Engineering and Technology*. [Online]. 3 (7). p.pp. 1356–1361. Available from: https://scholarworks.bridgeport.edu/xmlui/bitstream/handle/123456789/1654/265-crime_analysis_of_chicagoposter.pdf?sequence=1.

Nath, S. (2006). Crime Pattern Detection Using Data Mining. In: *2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*. [Online]. December 2006, IEEE, pp. 41–44. Available from: http://ieeexplore.ieee.org/document/4053200/.

Nath, S.V. (2007). Crime Data Mining. In: *Advances and Innovations in Systems, Computing Sciences and Software Engineering*. [Online]. Dordrecht: Springer Netherlands, pp. 405–409. Available from: http://link.springer.com/10.1007/978-1-4020-6264-3_70.

ONS (2017). *Crime in England and Wales: year ending Mar 2017*. [Online]. 2017. Office for National Statistics. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/crimeinenglandandwales/yearendingmar2017. [Accessed: 31 August 2017].

Osman, A.M.S., Elragal, A. & Bergvall-Kareborn, B. (2017). *Big Data Analytics and Smart Cities: A Loose or Tight Couple*. [Online]. Sweden. Available from: http://www.diva-portal.org/smash/get/diva2:1105669/FULLTEXT01.pdf.

Pflueger, M.O., Franke, I., Graf, M. & Hachtel, H. (2015). Predicting general criminal recidivism in mentally disordered offenders using a random forest approach. *BMC Psychiatry*. 15 (1). p.p. 62.

Pramanik, M.I., Lau, R.Y.K., Yue, W.T., Ye, Y. & Li, C. (2017). Big data analytics for security and criminal investigations. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. [Online]. 7 (4). p.p. e1208. Available from: http://doi.wiley.com/10.1002/widm.1208.

Pushpa, S.K., Manjunath, T.N. & Srividhya (2016). Analysis of Airport Data using Hadoop-Hive: A Case Study. *International Journal of Computer Applications*. [Online]. p.pp. 1–6. Available from: http://research.ijcaonline.org/ncrtit2016/number2/ncrtit201634.pdf.

Rasmi, M. & Jantan, A. (2013). A New Algorithm to Estimate the Similarity between the Intentions of the Cyber Crimes for Network Forensics. *Procedia Technology*. [Online]. 11. p.pp. 540–547. Available from: http://linkinghub.elsevier.com/retrieve/pii/S2212017313003800.

Ravna, A.K. (2015). *Electronic Fortune-Tellers Predictive policing as a sociotechnical phenomenon*. [Online]. UNIVERSITY OF OSLO. Available from: https://www.duo.uio.no/bitstream/handle/10852/45335/Master-thesis---Ailo-Ravna.pdf?sequence=1.

Rosser, G., Davies, T., Bowers, K.J., Johnson, S.D. & Cheng, T. (2017). Predictive Crime Mapping: Arbitrary Grids or Street Networks? *Journal of Quantitative Criminology*. [Online]. 33 (3). p.pp. 569–594. Available from: http://link.springer.com/10.1007/s10940-016-9321-x.

Rossmo, D., Laverty, I. & Moore, B. (2005). Geographic profiling for serial crime investigation. In: *Geographic information systems and crime analysis*. Hershey: Idea

Group Publishing, pp. 102–104.

Sanders, C.B. & Sheptycki, J. (2017). Policing, crime and 'big data'; towards a critique of the moral economy of stochastic governance. *Crime, Law and Social Change*. [Online]. p.pp. 1–15. Available from: http://link.springer.com/10.1007/s10611-016-9678-7.

Saoumya & Baghel, A.S. (2015). A Predictive Model for Mapping Crime Using Big Data Analytics. *International Journal of Research in Engineering and Technology*. 4 (4). p.pp. 344–348.

Saravanan, M., Thayyil, R. & Narayanan, S. (2013). Enabling Real Time Crime Intelligence Using Mobile GIS and Prediction Methods. In: *2013 European Intelligence and Security Informatics Conference*. August 2013, IEEE, pp. 125–128.

Sathyadevan, S., S, D.M. & S., S.G. (2014). Crime analysis and prediction using data mining. In: *2014 First International Conference on Networks & Soft Computing (ICNSC2014)*. [Online]. August 2014, IEEE, pp. 406–412. Available from: http://ieeexplore.ieee.org/document/6906719/.

Sharma, D., Asghar, Z. & Aalsaunet, A. (2017). *An introduction to Big Data*. [Online]. 2017. IFI. Available from: http://heim.ifi.uio.no/~infpri/Presentasjoner/INF5050BigData.pdf. [Accessed: 12 September 2017].

Sharma, M. (2014). Z - CRIME: A data mining tool for the detection of suspicious criminal activities based on decision tree. In: *2014 International Conference on Data Mining and Intelligent Computing (ICDMIC)*. September 2014, IEEE, pp. 1–6.

Shibata, S., Hanyu, K., Asakawa, T., Shimada, T. & Omata, K. (2012). Community Activities to Protect Children from Crime, People's General Trust, and Attitude toward the Activities in Japan. *Procedia - Social and Behavioural Sciences*. [Online]. p.pp. 40–50. Available from: http://www.sciencedirect.com/science/article/pii/S1877042812008014.

Singh, D., Vishnu, C. & Mohan, C.K. (2016). Visual Big Data Analytics for Traffic Monitoring in Smart City. In: *15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. [Online]. 2016, Los Angeles, CA: IEEE, pp. 886–891. Available from: http://www.diva-portal.org/smash/get/diva2:1105669/FULLTEXT01.pdf.

Soh, M.B.C. (2012). Crime and Urbanization: Revisited Malaysian Case. *Procedia - Social and Behavioral Sciences*. [Online]. 42. p.pp. 291–299. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1877042812010750.

Strang, K.D. & Sun, Z. (2017). Analyzing Relationships in Terrorism Big Data Using Hadoop and Statistics. *Journal of Computer Information Systems*. [Online]. 57 (1). p.pp. 67–75. Available from: https://www.tandfonline.com/doi/full/10.1080/08874417.2016.1181497.

Sun, C., Yao, C.-L., Li, X. & Lee, K. (2014). Detecting crime types using classification algorithms. *Journal of Digital Information Management (JDIM)*. 12. p.pp. 321–327.

Tollenaar, N. & Van-der-Heijden, P.G.M. (2013). Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 176 (2). p.pp. 565–584.

Toole, J., Eagle, N. & Plotkin, J. (2010). *Quantifying behavioral data sets of criminal activity: In: 2010 AAAI spring symposium series*. California.

Tutun, S., Khasawneh, M.T. & Zhuang, J. (2017). New Framework That Uses Patterns and Relations to Understand Terrorist Behaviors. *Expert Systems with Applications*. [Online]. Available from: http://www.start.umd.edu/publication/new-framework-uses-patterns-and-relations-understand-terrorist-behaviors.

Upadhyaya, D. & Jain, S. (2013). Hybrid approach for network intrusion detection system

using k-medoid clustering and Naı¨ve Bayes classification. *International Journal of Computing Science*. (10). p.pp. 231–236.

Vural, ehmet S., Gok, M. & Yetgin, Z. (2014). Analysis of incident-level crime data using clustering with hybrid metrics. *GAU Journal of Social and Applied Science*. 8. p.p. :8–20.

Vural, M.S., Gok, M. & Yetgin, Z. (2013). Generating incident-level artificial data using GIS based crime simulation. In: *2013 International Conference on Electronics, Computer and Computation (ICECCO)*. November 2013, IEEE, pp. 239–242.

Wanawe, M., Awasare, S. & Puri, N. (2014). Efficient approach to detecting phishing a web using k-means and Naı¨ve-Bayes algorithms. *International Journal of Research in Advent Technology*. 2. p.p. 1 06–111.

Wang, S. & Li, W. (2016). *Crime Analysis of Chicago*. [Online]. University of Bridgeport, Bridgeport, CT. Available from: https://scholarworks.bridgeport.edu/xmlui/bitstream/handle/123456789/1654/265-crime_analysis_of_chicagoposter.pdf?sequence=1.

Wang, X. & Brown, D.E. (2012). The spatio-temporal modeling for criminal incidents. *Security Informatics*. [Online]. 1 (1). p.p. 2. Available from: http://security-informatics.springeropen.com/articles/10.1186/2190-8532-1-2.

Wang, Y., Peng, X. & Bian, J. (2014). Computer Crime Forensics Based on Improved Decision Tree Algorithm. *Journal of Networks*. 9 (4). p.pp. 1005–1011.

Williams, M.L., Burnap, P. & Sloan, L. (2016). Crime Sensing with Big Data: The Affordances and Limitations of using Open Source Communications to Estimate Crime Patterns. *British Journal of Criminology*. [Online]. 57 (2). Available from: https://academic.oup.com/bjc/article-lookup/doi/10.1093/bjc/azw031.

Woodworth, J.T., Mohler, G.O., Bertozzi, A.L. & Brantingham, P.J. (2014). Non-local crime density estimation incorporating housing information. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. [Online]. 372 (2028). p.pp. 20130403–20130403. Available from: http://rsta.royalsocietypublishing.org/cgi/doi/10.1098/rsta.2013.0403.

Yadav, M. & Chaudhary, S. (2016). A Survey on Big Data Analytics Using Hadoop Ecosystem Tools. *International Journal of Computer Science and Information Technologies*. [Online]. 7 (4). p.pp. 2100–2102. Available from: http://ijcsit.com/docs/Volume 7/vol7issue4/ijcsit2016070492.pdf.

Zhang, L. (2016). *Big Data Analytics for Fault Detection and its Application in Maintenance*. [Online]. Sweden. Available from: https://www.diva-portal.org/smash/get/diva2:1046794/FULLTEXT01.pdf.

APPENDIX

A. Crimeloc Table

Algorithm:

- Enter the Hadoop console using command: hive
- Select the database using use command;
- x = map the location, crimetyp dataset using select command;
- y = For each x generate the number of crimes by location respectively;
- z = Group by location;
- Data = For each Z generate group, count(crimetyp);
- Result = for whole data follow the order by locations;
- Create table crimeloc using create table
- insert the output of result into crimeloc table;
- Store output;

B. Maxmin table

Algorithm:

1. Enter the Grunt shell using command: hive
2. Select the database using the use command;
3. x = map the location, crimetyp data set using select command;
4. y = For each x generate the crime types by location respectively;
5. z = Group by location, count(crimetyp);
6. Data = For each Z generate group, max(count(crimetyp)), min(count(crimetyp));
7. Result = for whole data follow the order by locations;
8. create table maxmin
9. insert the result into the maxmin table
10. Store output;

C. month crime table

Algorithm:

1. Enter the Grunt shell using command: hive
2. Select the database with use command;
3. X = map the month, crimetyp data set using select command;
4. Y = For each X generates the each crimetyp by month;
5. Z = Group by month;
6. Data = For each Z generate group, count (y. crimetyp), order by count;
7. create the table monthcrime;
8. insert the result into table month crime;
9. Store output

D. orderwisecrime table

Algorithm:

1. Enter the Grunt shell using command: Hive
2. Select the database with use command;
3. w = Load the crime type and count(crimetyp) data set using select command;
4. x = For each w generate crimetyp by month

5. Y = For each X sort the crimetyp by crime head;
6. Z = Group by crime head;
7. Data = For each Z generate group, count (y. crimehead);
8. Result = for each data place in order by count;
9. create the table orderwisecrime;
10. insert the Data into orderwisecrime table;
11. Store output;

E. percentage-month crime:

Algorithm:

1. Enter the Grunt shell using command: Hive
2. select the database with use command;
3. w = map the month, crimetyp data set using show command;
4. x = For each w generate crimetyp by month
5. Y = For each X combine the crimetyp by crime head;
6. Z = count by crime head;
7. A = for each z multiply with 100 and divide by count(crimetyp);
8. Data = For each A generate group (x. month), percentage (A. (count (crime head) * 100 / count(crimetyp)));
9. Result = for each data place in order by month;
10. create table percentagemonthcrime;
11. insert the selection command into table percentagemonthcrime;
12. Store output;

F. totalpercentagecrime table

Algorithm:

1. Enter the Grunt shell using command: Hive
2. select the database with the use command;
3. w = map the crimetyp data set using select command;
4. x = For each w generate crimetyp by month
5. Y = For each X combine the crimetyp by crime head;

6. Z = count by crime head;
7. A = for each z multiply with 100 and divide by count(crimetyp)
8. Data = For each A generate group (x. crimehead), percentage(A.(count(crime head) * 100 / count(crimetyp)));
9. Result = for each data place in order by crime head.
10. create table totalpercentagecrime;
11. insert the select command into the totalpercentagecrime table;
12. Store output;

G. Data Processing

Algorithm:

1. Enter the Grunt shell using command: hive
2. x = map the 2013 and 2014 data sets using select command;
3. y = combine the result of 2013 & 2014 data using full outer join;
4. z = reduce the y result as t1
5. A = map the 2015 and 2016 datasets using select command;
6. B = combine 2015 & 2016 data using full outer joins;
7. C = reduce the y result as t2;
8. D = map the t1 and t2 using select command;
9. E = combine the t1 & t2 using full outer join;
10. F = reduce the result of E as t3;
11. G = map the t3 and 2017 datasets using select command;
12. H = combine the t3 & 2017 dataset using full outer join;
13. I = reduce the result of H as final table;
14. Store the output;