2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)

Chicago Crime Data Analysis Using PIG in Hadoop

Anupam Mukherjee
Department of Computer Science & Engineering
Siliguri Institute of Technology
Siliguri, West Bengal, India
anupamsit@gmail.com

Sourav De
Department of Computer Science & Engineering
Cooch Behar Government Engineering College
Cooch Behar, West Bengal, India
dr.sourav.de79@gmail.com

Siddhartha Bhattacharyya
Faculty of Electrical Engineering and Computer Science
VSB Technical University of Ostrava
Ostrava, Czech Republic
siddhartha.bhattacharyya@vsb.cz

Jan Platos
Faculty of Electrical Engineering and Computer Science
VSB Technical University of Ostrava
Ostrava, Czech Republic
jan.platos@vsb.cz

***Abstract* –The need for analyzing the different crime dataset has significantly increased over the past few decades. It is necessary to detect the wide variety of crimes and the corresponding place of occurrences accurately. Government bodies throughout the world maintained Open Data initiatives, which is a large collection of heterogeneous dataset. This enables the government agencies to maintain the law and order of the society. The prime objective of this paper is to analyze the Chicago crime dataset to extract the significant crime information over the years. The proposed analysis is performed to bring out the crime information depending on some predetermined criteria, e.g. total crimes of different types, narcotic crime cases, an offense involving children, analysis of hourly theft cases and location identification where the major theft cases have occurred. The proposed analysis is performed in fully distributed Hadoop cluster. Moreover, we have extracted crucial minuscule statistics on theft-related criminal activities and their latitude or longitude.**

***Keywords*—**Fully Distributed Hadoop Cluster, Hadoop Distributed File System, Pig Latine, Crime data, Open Data initiatives, Name Node, Data Nodes.

I. INTRODUCTION

Crime data analysis and prediction have become the most formidable task now a day as the human population is increasing day by day. The crime tendency among the people is also increasing, depending on the socio-economical status and other related factors. National judicial report of Chicago shows that, on an average, out of 100 peoples, 12 peoples affected by crime-related activities [1]. Government investigation agencies are currently using big data analysis to predict the wide verities of crimes across the globe. Predictive crime data analysis is popular in many countries, such as US, China [2]. 'Predictive policing' has already been implemented by US, UK, China. The predictive policing model has been built by using a statistical algorithm. This paper analyses some of the significant section of the Chicago crime information for extracting different crime information. Crime data have been taken from the city of Chicago in between the year January 2001 to April 2018 from a US government site and the data has been recorded by the Chicago Police Department [3].

978-1-5386-7638-7/18/$31.00 ©2018 IEEE

Data mining techniques have a great significance in the field of crime problem detection and analysis [4]. Bruin, et al. [5] presented toolbox for changing criminal behavior. In their work, they have incorporated many important factors including, frequency, seriousness, duration etc. Almanie, T., et al. [6] focused on finding spatial and temporal criminal hotspots and they have explained the utilization of Decision Tree Classifier and Naïve Bayesian Classifier in order to forecast potential crime types. To identify the crime location or region and determine the necessary action is a challenging task. Crime prediction and prevention is an important field of work, to generate a significant solution for general citizens and police forces. Many works have already implemented to identify necessary information from a large database [7]. They have implemented and discussed many existing methods; several maps application also exists to identify the location along with the crime types and city [8].Chicago crime record has tracked by the Chicago police department on regular basis. Analysis of this paper provides a comprehensive overview based on historical data.

In this study, we have presented a detailed analysis of wide verity of crimes by means of Chicago crime dataset. In order to maintain law and order at the society, we need to detect wide verity of crimes and analyze them to take effective preventive mechanism.

The contribution of this article is as follows.

- The analyses are much more challenging due to the heterogeneous data presents in the dataset.
- The first analysis includes the total number of crimes happened between January 2001 to April 2018. The outcome of this analysis clearly demonstrates that the intensity of theft-related crimes is extreme.
- Secondly, the analysis in case of narcotic crime gives a positive conclusion, i.e. narcotic crime rates are decreasing day by day. It has been observed that 38,998 numbers of narcotic crimes have been decreased in last 17 years. This paper also introduces a graphical analysis of narcotic crimes in between

the year of 2001 to 2017, which clearly reflects the positive intensity of Chicago police.

- Furthermore, we have to provide special attention to the child related criminal activities [9]. From our analysis, we may say that the offense involves in children was extremely increased in between 2003 to 2007. Till now, the number is quite alarming. We have also made a crime intensity wise offense analysis and make a comparison in between 2001 to 2017. Different attributes involve in this comparisons are Contribute delinquency of a child, aggregate sex assault of a child by family members, sex assault of children by family members, endangering life or health of a child, crime sex abuse by family member, child pornography, child abduction, harbor runway, child abuse and other offense. Cases of child abduction, endangering life and sexual assault of a child by the family member have been decreasing but cases of child pornography and uncategorized offenses have been rapidly increasing.
- As various types of crimes are happening everywhere in Chicago, this paper has tried to make a deep level analysis of "theft" related crimes. Theft related cases are extreme in between 12 PM to 1 PM. The maximum number of theft cases is held during office hours i.e. from 9 AM to 5 PM and at midnight.
- Moreover, the focus of this paper is to identify time level and location/ region based theft level analysis, which is quite challenging tasks. In the case of predictive data analysis, this two information (time and location) will play a significant role [10].

The structure of this paper is organized as follows. Section II gives a brief introduction about problem identification. Section III discusses proposed work flow architecture. Section IV shows the observation of experimental results and analysis. Section V concludes the paper.

II. PROBLEM IDENTIFICATION

Crime data analysis has a large significant impact on public safety and national security. For this purpose, Government bodies throughout the world make Open Data Initiatives [11]. The crime record of Chicago city contains many important attributes, which is required in data processing applications. Size of the Chicago crime record file is around 1.5 GB and that contained several types of crime information from January 2001 to April 2018. The main objective is to find different crime activities more accurately and analyze them [12]. Here we make a deep level analysis of 'theft' related criminal activities. Moreover, the focus of this paper is to identify time level and location/ region based theft level analysis, which is quite challenging tasks. The major challenges in these steps are to identify and classify this unstructured dataset to a structured format. In case of predictive data analysis, these two information's (time and location) will play a significant role [13].

III. WORK FLOW ARCHITECTURE

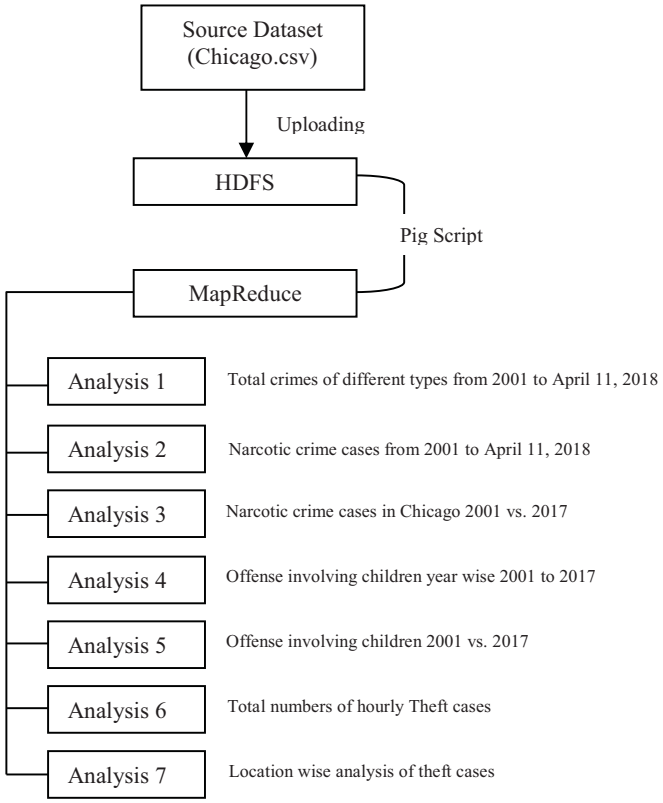

Fig.1. Work Flow Architecture

The entire work of this paper is performed on Fully Distributed Hadoop Cluster mode [14]. Gradually execution of the above workflow model has depicted in Figure 1. Chicago data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system [3]. This data set represents different crime incidents. In the next phase, downloaded dataset uploaded into the Hadoop distributed file system (HDFS) [15]. HDFS has designed to run on commodity hardware. It has major significant advantages compared to other distributed file system. HDFS is most suitable for applications that have dealt with voluminous datasets. The

proposed workflow model has been executed in master/slave architecture. Our proposed HDFS clusters consist of a single Name Node (Server) and several data nodes (Slave) [16].

In this paper, we have used pig-scripting language for data analysis. Pig is a high level programming language required for data analysis in Apache Hadoop. Pig scripting language [17] support both structured and unstructured data. Major Advantages of Pig is that It can translate the large dataset to a Map-Reduce [18] [19] framework in a Hadoop Cluster.

In this proposed workflow framework, Analysis 1 is performed to extract the total number of crimes by its type, occurred in-between January 2001 to April 2018 in the city of Chicago. In addition to this, proposed analysis 2 brings out the narcotic crime cases from 2001 to 2018. Furthermore, we have presented a comparative analysis in-between two years (2001 and 2017), which is performed by proposed analysis 3. The analysis 4 focuses on different child-related crime offenses involved in-between 2001 to 2017. Thereafter the analysis 5is performed in such a way to get proper observation in between 2001 and 2017. Furthermore, Analysis 6 makes a detailed level observation on hourly theft-related activities. Henceforth this paper also presented the location wise theft case analysis in analysis 7. Location, latitude, or longitude of a region is always having considerable impact in case of predictive analysis [20].

IV. EXPERIMENTAL RESULTS AND ANALYSIS

i) Analysis 1: Proposed analysis 1 produces the total crimes of different types committed in between January 2001 to April 2018.

Analysis1: Analysis of *Total* crimes of different types from 2001 to April 11, 2018

```
1:    Procedure  Run PIG Latin in Map Reduce Mode
                 (Crime_type_by_Year)
2:      Crimes = LOAD Chicago Crime dataset using
        CSV Loader
3:      Select = FOREACH Crimes GENERATE
                 Year, 1 as no
4:    grps = GROUP Select BY year;
5:      counts = FOREACH grps GENERATE
      group,
          SUM (Select.no);
6:      dump counts;
7:      STORE counts INTO the output directory
8:          end procedure
```

We have presented a comparison over wide verity of criminal activities those happened at the Chicago city in between Jan. 2001 to April 2018 by the analysis 1. The resultant analysis has depicted in figure 2. Majorly involved crime types are – prostitution, sex offenses, ritualism, obscenity, narcotics, burglary, offense involving children, kidnapping, stalking, gambling, robbery, battery crime, assault and theft. After analyzing the figure 2, we came to a positive conclusion that the major crimes offenses are involved in the city of Chicago are "Theft – (1377571)" and then followed by "Battery related crimes" – (1201266). Some other parts also need proper attention like – "Narcotics" - (703138), "Assault" - (405813), "Burglary" - (380708) and "Robbery" - (249902).

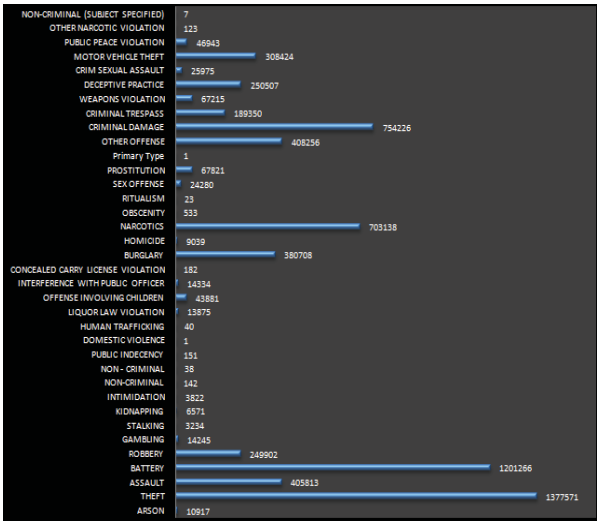

Fig.2. Total crimes of different types from 2001 to April 11, 2018

ii) Analysis 2: Proposed analysis 2 produces the total Narcotic crimes committed in between 2001 to April 2018.

Analysis2: Analysis of Total Narcotic crimes from 2001 to April 11, 2018

```
1:    Procedure  Run PIG Latin in Map Reduce Mode
                 (Narcotic_Crime_by_Year)
2:        Crimes = LOAD Chicago Crime dataset
          using CSV Loader
3:        filter_data = filter crimes by
          NARCOTICS
4:        final_data = FOREACH filter_data2
          GENERATE group, COUNT(final_data)
5:        dump final_data
6:        STORE final_data INTO the output
          directory
7:    end procedure
```

It has been observed on year wise investigation of narcotic crime from the analysis 2 that overall, narcotic

crime study has produced an optimistic conclusion. The rate of narcotic crime tendency is decreasing in the city of Chicago. It has been observed that almost 38,998 numbers of narcotic crimes have been diminishing in last 17 years. Detail graphical analysis has shown in figure 3.

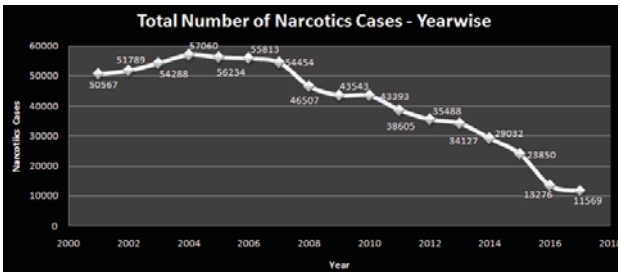

Fig.3. Total Number of Narcotics Cases from 2001 to April 11, 2018

iii) Analysis 3: Proposed analysis 3 generates a comparative analysis of narcotic crimes in between 2001 vs. 2017.

Analysis3: Analysis of Narcotic crimes in Chicago 2001 vs. 2017.

1: **Procedure** Run PIG Latin in Map Reduce Mode (Narcotic_Crime_inbetween_2001 vs. 2007)
2: Crimes = LOAD Chicago Crime dataset using CSV Loader
3. data01 = FILTER crimes BY 2001 and NARCOTICS
4. grp01 = GROUP data01 BY description
5. grp01_counts = FOREACH grp01 GENERATE group, COUNT(data01) as c2001
6. data17 = FILTER crimes BY 2017 and NARCOTICS
7. grp17 = GROUP data17 BY description
8. grp17_counts = FOREACH grp17 GENERATE group, COUNT(data17) as c2017
9. joined_data = JOIN grp01_counts BY group,grp17_counts by group
10. final_data = FOREACH joined_data GENERATE group, c2001, c2017
11. STORE final_data INTO the output directory
12.**end procedure**

We have presented a comparative analysis between two different datasets (2001 and 2017). The resultant output of this analysis depicted in figure 4. After analyzing figure 4, we came to a solution that the overall narcotic crime rate is decreasing. Chicago Police and there Government administration showing there positive intent in some major areas, like – Poss:Crack, White heroin, Poss: Cannabis 30 gms or less.

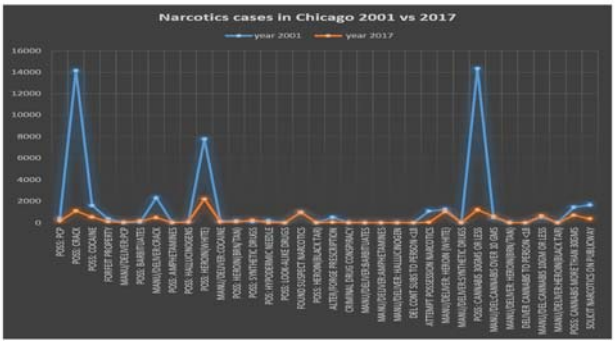

Fig.4. Analysis of Narcotic crimes in Chicago 2001 vs. 2017.

iv) Analysis 4: Proposed analysis 4 produces the total offense involving children in between 2001 to 2017.

Analysis4: Offense involving children year wise 2001 to 2017

1: **Procedure** Run PIG Latin in Map Reduce Mode (Children_Offense_by_Year)
2: Crimes = LOAD Chicago Crime dataset using CSV Loader
3: filter_data = FILTER crimes by 'OFFENSE INVOLVING CHILDREN'
4: filter_by_year = GROUP filter_data BY year
5: final_data = FOREACH filter_by_year GENERATE group, COUNT(filter_data)
6: dump final_data
7: STORE final_data INTO the output directory
8: **end procedure**

Child related offense is extremely important issue for a country. Chicago government should be more proactive at this point. It is clearly observed that from 2001 onward, rate of child related crime tendency is monotonically increasing. In 2004, that value increased up to 3,077. Presently the number is around 2,200, but still now, it is a serious point of concern for the city of Chicago. Analytical report has shown in figure 5.

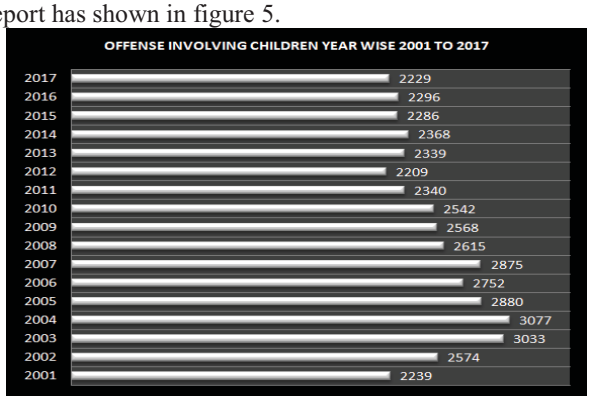

Fig.5.Offense involving children in year wise from 2001 to 2017

v) Analysis 5: Proposed analysis 5 generates a comparative analysis of offense involving children in between 2001 vs. 2017.

Analysis5: Analysis of Offense involving children 2001 vs. 2017.

```
1:    Procedure  Run PIG Latin in Map Reduce Mode
      (Children_Offense_Comparison_2001_2017)
2:          Crimes = LOAD Chicago Crime dataset
            using CSV Loader
3:          data01 = FILTER crimes BY 2001 and
            'OFFENSE INVOLVING CHILDREN'
4:          grp01 = GROUP data01 BY description
5:    grp01_counts = FOREACH grp01 GENERATE
      group, COUNT(data01) as c2001
6:    data17 = FILTER crimes BY 2017 and 'OFFENSE
      INVOLVING CHILDREN'
7:    grp17 = GROUP data17 BY description
8:    grp17_counts = FOREACH grp17 GENERATE
      group, COUNT(data17) as c2017
9:    joined_data = JOIN grp01_counts BY
      group,grp17_counts by group
10:   final_data = FOREACH joined_data GENERATE
      group, c2001, c2017
11:   STORE final_data INTO the output directory
12: end procedure
```

We made a comparative analysis of child related criminal activities in between 2001 and 2017. Many important facts are coming out here. Child abuse is increasing at a staggering speed and it is a matter of great concern for the Government. The child abduction rate is relatively reduced. The Child health related issues are still a deep matter of concern.

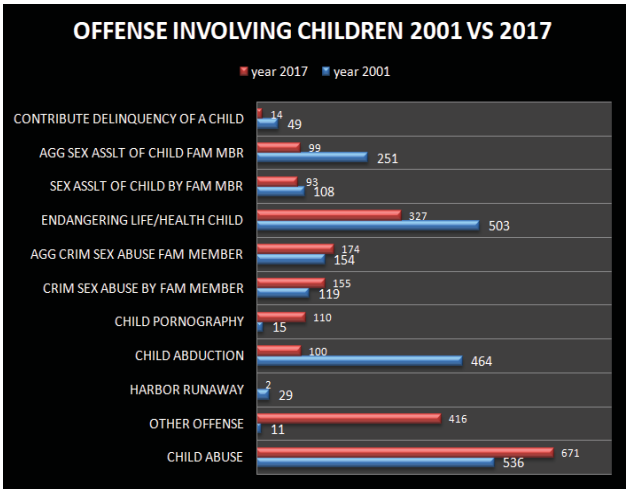

Fig.6. Analysis of Offense involving children 2001 vs. 2017.

vi) Analysis 6: Proposed analysis 6 creates a deep level analysis of total number of hourly-based theft cases.

Analysis6: Total number of hourly theft cases.

```
1:    Procedure  Run PIG Latin in Map Reduce Mode
      (Hourly_Based_Theft)
2:          Crimes = LOAD Chicago Crime dataset
            using CSV Loader
3:     t = FOREACH crimes GENERATE
      id,primarytype, FLATTEN(STRSPLIT(date, ' ',
      3)) AS (date:chararray,time:chararray,
      ampm:chararray)
4:     t1 = FOREACH t GENERATE id,primarytype,
      FLATTEN(STRSPLIT(time, ':', 3)) AS
      (h:chararray,m:chararray,s:chararray), ampm
5:     t2 = FILTER t1 BY THEFT
6:     t3 = GROUP t2 BY (h, ampm)

7:     t4 = FOREACH t3 GENERATE
      CONCAT(group.h,' ',group.ampm) AS (h_ampm :
      chararray),COUNT(t2)
8:     STORE t4 INTO the output directory
9:  end procedure
```

In this section, we have presented a deep level analysis of theft related crimes and the resultant output is depicted in figure 7. Theft cases are the highest incident in the city of Chicago. Henceforth, we made a detailed level (hourly based) theft analysis. Theft related cases is maximum in between 12 PM to 1 PM. Maximum number of theft cases are held on office hours i.e. 9 AM to 5 PM and on midnight. Furthermore, we try to generate the location or region where the theft cases have majorly occurred.

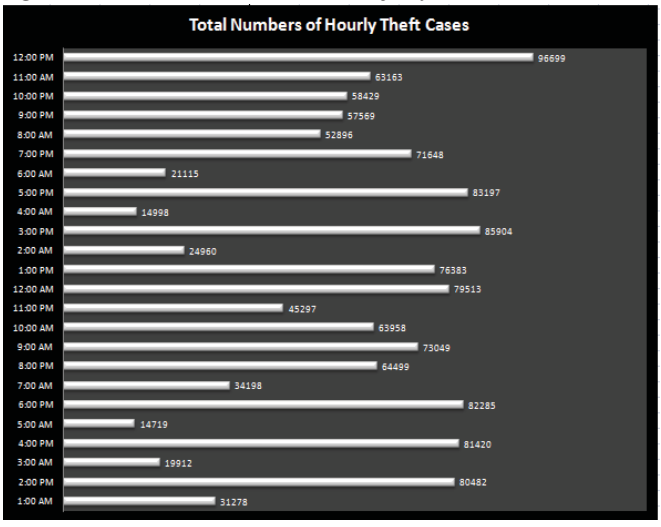

Fig.7. Total number of hourly theft cases.

vii) Analysis7: Proposed analysis 7 generates the region or location where the major theft cases occurred.

Analysis7: Location wise analysis of theft cases.

```
1:  Procedure  Run PIG Latin in Map Reduce Mode
        (Location_wise_analysis)
2:      Crimes = LOAD Chicago Crime dataset using
            CSV Loader
3:      t = FOREACH crimes GENERATE
            id,primarytype,FLATTEN(STRSPLIT(da
            te, ' ', 3)) AS
            (date:chararray,time:chararray,ampm:char
            array),xcoordinate,ycoordinate,year;
4:      t1 = FOREACH t GENERATE
            id,primarytype,FLATTEN(STRSPLIT(ti
            me, ':', 3)) AS
            (h:chararray,m:chararray,s:chararray),
            ampm,xcoordinate,ycoordinate,year;
5:      t2 = FILTER t1 BY primarytype=='THEFT' and
            h=='12' and ampm=='PM' and
            year==2017;
6:      t3 = FOREACH t2 GENERATE
            id,xcoordinate,ycoordinate;
7:      STORE t3 INTO 'output_dir/loc1x/' USING
            PigStorage(',');
8:  end procedure
```

The above analysis generates around 4500 co-ordinates values or latitude/longitude values, from which we can identify the location of crime region. That will help to predict the crime region in advance. This analysis is one of the important challenging tasks of this paper to solve the crime tendency in the city of Chicago.

CONCLUSION

This paper analyzes Chicago crime data from 2001 to April, 2018. The analysis includes different types of crimes, like prostitution, sex offenses, narcotics, burglary, kidnapping etc. In addition to this, we have analyzed the database by extracting narcotic and child abuse related crimes. In case of narcotic crimes, the analysis shows that the rate of some major narcotic crime (poss:crack, white heroin, poss. :cannabis 30gm) is decreasing. In this context, it is to be noted that the rate of child abuse is highly increasing and the rate of child abduction has relatively reduced. Moreover, the hourly-based theft analysis reflects that the maximum number of theft cases occurred in between 12P.M to 1 P.M., along with their place of occurrence (latitude & longitude). All analyses are implemented on a fully distributed Hadoop cluster.

REFERENCES

[1] Ennis, P. H. (1967). *Criminal victimization in the United States: A report of a national survey* (No. 2). Chicago: National Opinion Research Center, University of Chicago.

[2] Bonta, J., Law, M., & Hanson, K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders: a meta-analysis. *Psychological bulletin*, *123*(2), 123.

[3]https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2. [Accessed: 20- June-2018].

[4]Krishnamurthy, R., & Kumar, J. S. (2012). Survey of data mining techniques on crime data analysis. *International Journal of Data Mining Techniques and Applications*, *1*(2), 117-120.

[5] De Bruin, J. S., Cocx, T. K., Kosters, W. A., Laros, J. F., &Kok, J. N. (2006, December). Data mining approaches to criminal career analysis. In *Data Mining, 2006. ICDM'06. Sixth International Conference on* IEEE, 171-177.

[6] Almanie, T., Mirza, R., &Lor, E. (2015). Crime prediction based on crime types and using spatial and temporal criminal hotspots. *arXiv preprint arXiv:1508.02050*.

[7] Tayebi, M. A., Frank, R., &Glässer, U. (2012, November). Understanding the link between social and spatial distance in the crime world. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems,* ACM, 550-553.

[8] Haining, R., &Haining, R. P. (2003). *Spatial data analysis: theory and practice*. Cambridge University Press.

[9] Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y., & Chau, M. (2004). Crime data mining: a general framework and some examples. *computer*, *37*(4), 50-56.

[10] Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 121-130.

[11] Nath, S. V. (2006, December). Crime pattern detection using data mining. In *Web intelligence and intelligent agent technology workshops, 2006. wi-iat 2006 workshops. 2006 ieee/wic/acm international conference on* IEEE, 41-44.

[12] Kassen, M. (2013). A promising phenomenon of open data: A case study of the Chicago open data project. *Government Information Quarterly*, *30*(4), 508-513.

[13] Beck, C., & McCue, C. (2009). Predictive policing: what can we learn from Wal-Mart and Amazon about fighting crime in a recession?. *Police Chief*, *76*(11), 18.

[14] Leverich, J., &Kozyrakis, C. (2010). On the energy (in) efficiency of hadoop clusters. *ACM SIGOPS Operating Systems Review*, *44*(1), 61-65.

[15] Borthakur, D. (2008). HDFS architecture guide. *Hadoop Apache Project*, *53*, 1-13.

[16] Kaushik, R. T., & Bhandarkar, M. (2010, June). Greenhdfs: towards an energy-conserving, storage-efficient, hybrid hadoop compute cluster. In *Proceedings of the USENIX annual technical conference*, 109, 34-38.

[17] Gates, A., & Dai, D. (2016). *Programming pig: Dataflow scripting with hadoop*. " O'Reilly Media, Inc.", Gravenstein Highway North, Sebastopl, USA.

[18] Patel, A. B., Birla, M., & Nair, U. (2012, December). Addressing big data problem using Hadoop and Map Reduce. In *Engineering (NUiCONE), 2012 Nirma University International Conference on* IEEE, 1-5.

[19] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

[20] Kahle, D., & Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2. *R Journal*, *5*(1), 144-161.