A project report on

A SYSTEMATIC APPROACH FOR ANALYZING CYBERCRIME OCCURENCES USING BIG DATA

Submitted in partial fulfilment for the award of the degree

of

M.Tech Software Engineering

by

HARISH KUMAR S (13MSE0029)

SCHOOL OF INFORMATION TECHNOLOGY AND ENGINEERING

January, 2018

DECLARATION

I here by declare that the thesis entitled “A SYSTEMATIC APPROACH FOR ANALYZING CYBERCRIME OCCURENCES USING BIG DATA” submitted by me, for the award of the degree of M.S Software Engineering, VIT is a record of bonafide work carried out by me under the supervision of Prof. Kavitha B R .

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Vellore **Signature of the candidate**
Date: 23/01/1201

CONTENTS

LIST OF FIGURES

Chapter I

Introduction

1.1.INTRODUCTION

Big data is the voluminous and complex collection of data that comes from different sources such as sensors, content posted on social media website, sale purchase transaction etc. Such voluminous data becomes tough to process using ancient processing application. There are various tools and techniques in the market for big data analytics. With continually increasing population, crimes and crime rate analyzing related data is a huge issue for governments to make strategic decisions so as to maintain law and order. This is really necessary to keep the citizens of the country safe from crimes. The best place to look up to find room for improvement is the voluminous raw data that is generated on a regular basis from various sources by applying Big Data Analytics (BDA) which helps to analyze certain trends that must be discovered, so that law and order can be maintained properly and there is a sense of safety and well-being among the citizens of the country.

1.2.OVERVIEW

This project will analyse all the crimes that happened in the particular place with the help of big data tools. Here we are reviewing and analysing the features of crime incidents, their respective elements and propose a combinatorial incident description schema. Based on the result we are providing list of recommended actions, corresponding measures and effective policies. This will enable better monitoring, handling and moderate crime incident occurrences.

1.3.CHALLENGES

Data Size – The system will analyse more than 1 lakh of records.

1.4.PROJECT STATEMENT

The Proposed system will analyse all the crime occurrences in a city and give better understanding of crime occurrences of the particular place in the city.

1.4.1.EXISTING SYSTEM

The existing system consisted of two categories.

- Type I offences characterize singular or discrete events facilitated by the introduction of malware programs such as keystroke loggers, viruses, and rootkits.
- Type II offences are facilitated by programs that are not classified as crime ware, and they are generally repeated contacts or events from the perspective of the user.

Existing System concept deals with providing backend by using only MySQL which contains lot of drawbacks i.e. data limitation and processing time of huge data is very high and once data is lost we cannot recover.

1.4.2.PROPOSED SYSTEM

The proposed system was a schema-based cybercrime incident description that:

1) Identifies the features of a cybercrime incident and their potential elements

2) Provides a two-level offence classification system based on specific criteria.

3) Proposed concept deals with providing database by using Hadoop with Spark we can analyse unlimited data's and simply add number of machines to the cluster so the results are produced in less time.

4) The proposed schema can be extended with a list of recommended actions, corresponding measures and effective policies that counteract the offence type and subsequently the particular incident.

5) Proposed concept deals with providing database by using Hadoop tool so we can analyse unlimited data and simply add number of machines to the cluster(based on the requirements) and results are generated with less time, high throughput and cost of maintenance is also very less.

1.5.OBJECTIVE

To propose a system which analyse and identifies the features of crime incidents, their respective elements and proposes a combinatorial incident description schema. The system will act as a guide where the high frequency of cybercrime occurrences.

1.6.SCOPE OF THE PROJECT

In this system we are analyzing crime data in different areas in a city by using Hadoop framework along with Hadoop ecosystems like HDFS, MapReduce, SQOOP, Hive and Pig. By using these tools we can process Unlimited data, no data lost problem, high throughput, maintenance cost was also very less and it is an open source software, it is compatible on all the platforms since it is Java based.

Chapter II

Background

2.1. LITERATURE SURVEY

Title: Crime Pattern Detection Using Data Mining

Author: Shyam Varan Nath

Year: 2006

Description:

Data mining can be used to model crime detection problems. Crimes are a social nuisance and cost our society dearly in several ways. Any research that can help in solving crimes faster will pay for itself. About 10% of the criminals commit about 50% of the crimes. Here we look at use of clustering algorithm for a data mining approach to help detect the crimes patterns and speed up the process of solving crime. We will look at k-means clustering with some enhancements to aid in the process of identification of crime patterns. We applied these techniques to real crime data from a sheriff's office and validated our results. We also use semi-supervised learning technique here for knowledge discovery from the crime records and to help increase the predictive accuracy. We also developed a weighting scheme for attributes here to deal with limitations of various out of the box clustering tools and techniques. This easy to implement data mining framework works with the geospatial plot of crime and helps to improve the productivity of the detectives and other law enforcement officers. It can also be applied for counter terrorism for homeland security.

Title : Using Machine Learning Algorithms to Analyze Crime Data

Author: Lawrence McClendon and Natarajan Meghanathan

Year : 2015

Description:

Data mining and machine learning have become a vital part of crime detection and prevention. In this research, we use WEKA, an open source data mining software, to conduct a comparative study between the violent crime patterns from the Communities and Crime Normalized Dataset provided by the University of California-Irvine repository and actual crime statistical data for the state of Mississippi that has been provided by neighborhoodscout.com. We implemented the Linear Regression, Additive Regression; and Decision Stump algorithms using the same finite set of features, on the Communities and Crime Dataset. Overall, the linear regression algorithm performed the best among the three selected algorithms. The scope of this project is to prove how effective and accurate the machine learning algorithms used in data mining analysis can be at predicting violent crime patterns.

Title : Crime Prediction Based on Crime Types and Using Spatial and Temporal criminal Hotspots

Author : Tahani Almanie, Rsha Mirza and Elizabeth Lor

Year : 2015

Description:

This paper focuses on finding spatial and temporal criminal hotspots. It analyses two different real-world crimes datasets for Denver, CO and Los Angeles, CA and provides a comparison between the two datasets through a statistical analysis supported by several graphs.

Then, it clarifies how we conducted Apriori algorithm to produce interesting frequent patterns for criminal hotspots. In addition, the paper shows how we used Decision Tree classifier and Naïve Bayesian classifier in order to predict potential crime types. To further analyze crimes' datasets, the paper introduces an analysis study by combining our findings of Denver crimes' dataset with its demographics information in order to capture the factors that might affect the safety of neighbourhoods. The results of this solution could be used to raise people's awareness regarding the dangerous locations and to help agencies to predict future crimes in a specific location within particular time.

Title : Data mining Techniques to Analyze and Predict Crimes

Author **:** S.Yamuna and N.Sudha Bhuvaneswari

Year : 2012

Description:

Data mining can be used to model crime detection problems. Crimes are a social nuisance and cost our society dearly in several ways. Any research that can help in solving crimes faster will pay for itself. About10% of the criminals commits about 50% of the crimes. Data mining technology to design proactive services to reduce crime incidences in the police stations jurisdiction. Crime investigation has very significant role of police system in any country. Almost all police stations use the system to store and retrieve the crimes and criminal data and subsequent reporting. It became useful for getting the criminal information but it does not help for the purpose of designing an action to prevent the crime. It has become a major challenge for police system to detect and prevent crimes and criminals. There haven't any kind of information is available before happening of such criminal acts and it result into increasing crime rate. Detecting crime from data analysis can be difficult because daily activities of criminal generate large amounts of data and stem from various formats. In addition, the quality of data analysis depends greatly on background knowledge of analyst, this paper proposes a guideline to overcome the problem.

Title : Application for Analysis and Prediction of Crime Data Using Data Mining

Author : anisha agarwal and dhanashree chougule

Year : 2016

Description :

Today, time is a concerning factor for sentencing criminals. Many a time a criminal released on bail may yet bea potential threat to the society, even after they have served their sentence. This threat can be reduced if a prediction analysis is done on the concerned person to determine if he is about to do the crime or not. This aspect can be beneficial both for law enforcement and the safety of our country. Data mining is an approach that can handle large voluminous datasets and can be used to predict desired patterns. Our sole users will be the police officers who from time to time shall be able to predict the possibility of the crime a

criminal is probable to commence in the nearest future as well as which particular crime he will be committing. In this paper, we look at the use of frequent pattern mining with association rule mining to analyze the various crimes done by a criminal and predict the chance of each crime that can again be performed by that criminal. This analysis may help the law enforcement of the country to take a more accurate decision or may help in safeguarding an area if criminal released on bail is very much likely to perform crime. We will concentrate on Apriori algorithm with association rule mining technique to achieve the result.

Title : Survey of Crime Analysis and Prediction

Author : Lenin Mookiah,William Eberle and Ambareen Siraj

Year : 2014

Description:

Crime analytics and prediction have long been studied among research communities. In recent years, crime data from different heterogeneous sources have given immense opportunities to the research community to effectively study crime pattern and prediction tasks in actual real data. In this survey paper, we will discuss research that takes into account a variety of crime related variables, and shows where in some cases, information that has been widely accepted as influencing the crime rate, actually does not have an effect.

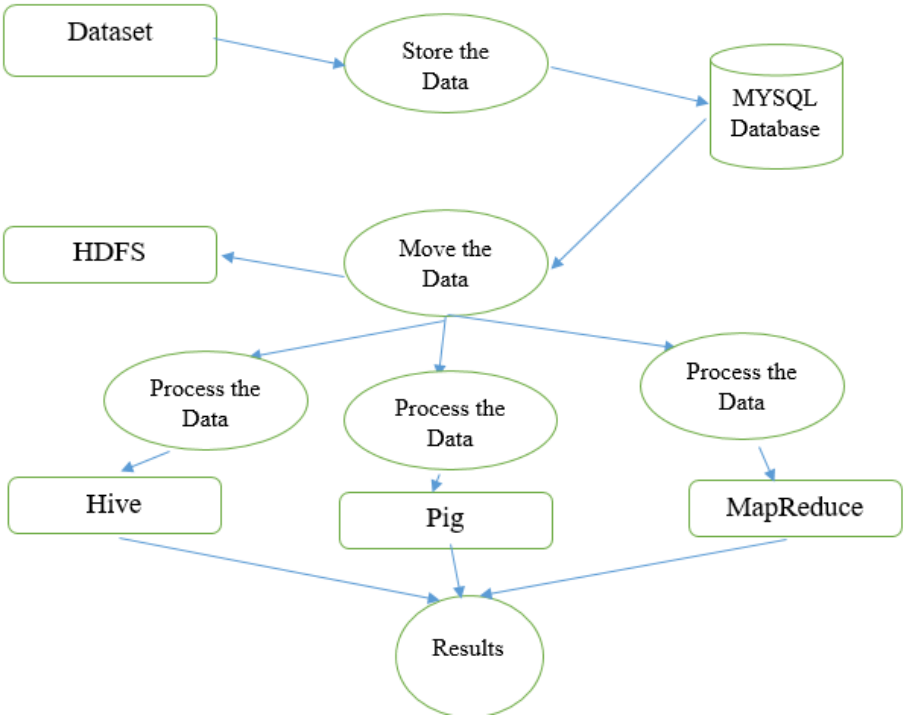
Chapter III

System Design

3.1.HIGH LEVEL DESIGN

High-level design (HLD) explains the architecture that would be used for developing a software product. The architecture diagram provides an overview of an entire system, identifying the main components that would be developed for the product and their interfaces. The HLD uses possibly nontechnical to mildly technical terms that should be understandable to the administrators of the system.

3.1.1. USE CASE DIAGRAM



3.2.PURPOSE

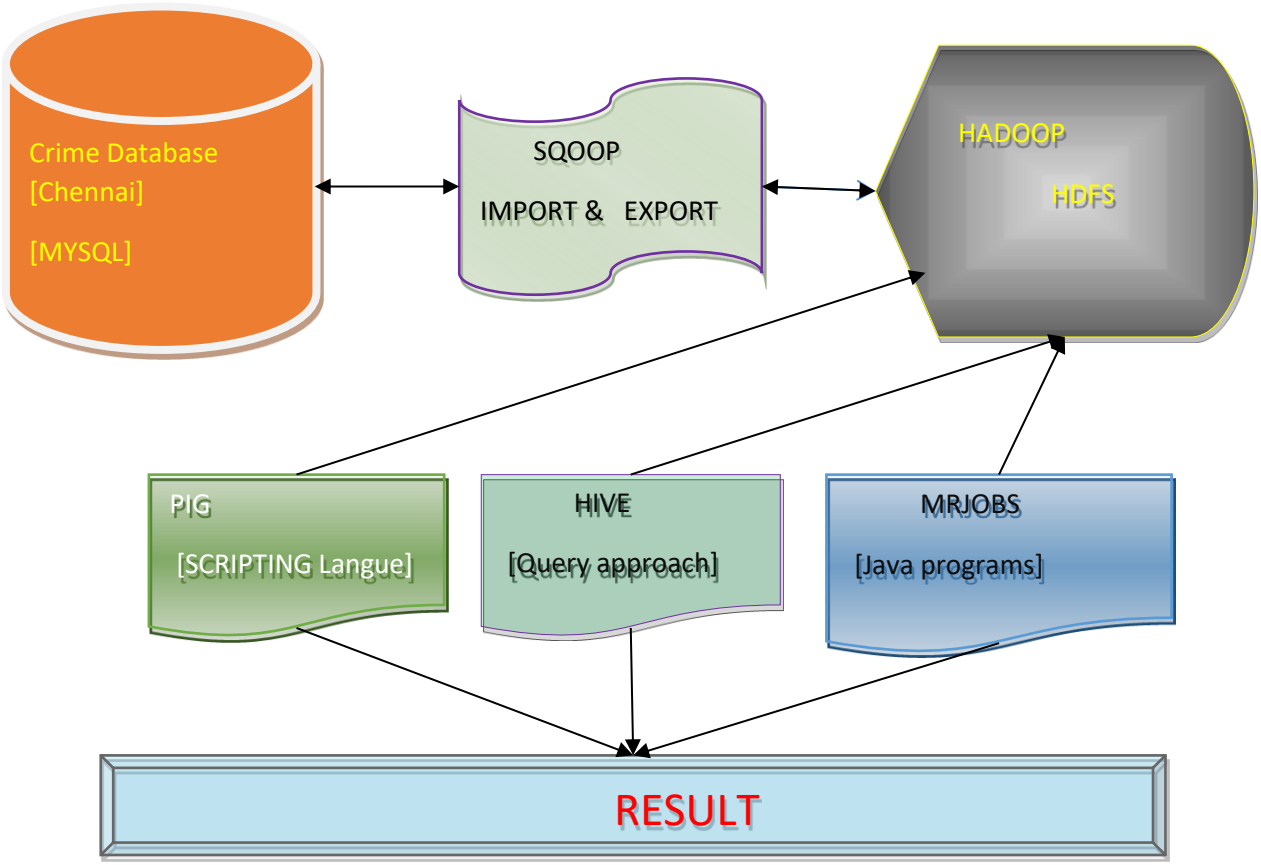
Preliminary design — In the preliminary stages of a software development, the need is to size the project and to identify those parts of the project that might be risky or time consuming.

Design overview — As the project proceeds, the need is to provide an overview of how the various sub-systems and components of the system fit together.

3.3.IMPORTANCE

It provides an overview of a solution, platform, system, product, service or process. Such an overview is important in a multiproject development to make sure that each supporting component design will be compatible with its neighbouring designs and with the big picture. The highest-level design briefly describe all platforms, systems, products, services and processes that it depends on and include any important changes that need to be made to them.

3.4.ARCHITECTURE DIAGRAM



3.4.1.Architecture Diagram

Chapter IV

Module Description

4.1.Data Pre-processing Model

This Module involves transforming raw **data** into an understandable format. Real-world **data** is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. **Data pre-processing** is a proven method of resolving such issues. The Crime Data Set will be provided into MySQL Database.

4.2.Data Migration Model with Sqoop

Transfer the Crime dataset into Hadoop (HDFS), with the help of Sqoop tool. Sqoop is a command-line interface application for transferring data between relational databases and Hadoop. Using Sqoop we have to perform lot of the function, such that if we want to fetch the particular column or if we want to fetch the dataset with specific condition that will be support by Sqoop Tool and data will be stored in Hadoop (HDFS).

4.3.Data Analytic Module with HIVE

Hive is a data ware house system for Hadoop. It runs SQL like queries called HQL (Hive query language) which gets internally converted to map reduce jobs. Hive was developed by Facebook. Hive supports Data definition Language (DDL), Data Manipulation Language (DML) and user defined functions. In this module we have to analysis the dataset using HIVE tool which will be stored in Hadoop (HDFS).For analysis dataset, HIVE using HQL Language. Using hive we perform Tables creations, joins, Partition, Bucketing concept. Hive analysis the only Structure Language.

4.4.Data Analytic Module with PIG

Apache Pig is a high level data flow platform for execution Map Reduce programs of Hadoop. The language for Pig is pig Latin. Pig handles both structure and unstructured language. It is also top of the map reduce process running background. In this module also used for analyzing the Data set through Pig using Latin Script data flow language.in this also we are doing all operators, functions and joins applying on the data see the result.

4.5.Data Analytic Module with MapReduce

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce.In this module we are analyzing the data set using MAP REDUCE. Map Reduce Run by Java Program.

REFERENCES

[1] S. Sathyadevan, M. Devan, and S. Surya Gangadharan, “Crime analysisand prediction using data mining,” in Networks Soft Computing(ICNSC), 2014 First International Conference on, Aug 2014, pp. 406–412.

[2] T. Pang-Ning, S. Michael, and K. Vipin, Introduction to Data Mining,1st ed. Pearson, 5 2005.

[3] S. Kaza, Y. Wang, and H. Chen, “Suspect vehicle identification forborder safety with modified mutual information,” in Proceedings ofthe 4th IEEE International Conference onIntelligence and SecurityInformatics, ser. ISI’06. Berlin, Heidelberg: Springer-Verlag, 2006,pp. 308–318.

[4] V. Vaithiyanathan, K. Rajeswari, R. Phalnikar, and S. Tonge, “Improvedapriori algorithm based on selection criterion,” in ComputationalIntelligence Computing Research (ICCIC), 2012 IEEE InternationalConference on, Dec 2012, pp. 1–4.

[5] C. Chu-xiang, S. Jian-jing, C. Bing, S. Chang-xing, and W. Yun-cheng,“An improvement apriori arithmetic based on rough set theory,” in Circuits, Communications and System (PACCS), 2011 Third Pacific-Asia Conference on, July 2011, pp. 1–3.

[6] S. Kaza, T.Wang, H. Gowda, and H. Chen, “Target vehicle identificationfor border safety using mutual information,” in Intelligent TransportationSystems, 2005. Proceedings. 2005 IEEE, Sept 2005, pp. 1141–1146.

[7] W. Huang, M. Krneta, L. Lin, and J. Wu, “Association bundle - anew pattern for association analysis,” in Data Mining Workshops, 2006.ICDM Workshops 2006. Sixth IEEE International Conference on, Dec 2006, pp. 601–605.

[8] N. Sasaki, R. Nishimura, and Y. Suzuki, “Audiowatermarking basedon association analysis,” in Signal Processing, 2006 8th International Conference on, vol. 4, Nov 2006.

[9] A. Ben Ayed, M. Ben Halima, and A. Alimi, “Survey on clustering methods: Towards fuzzy clustering for big data,” in Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of, Aug 2014, pp. 331–336.

[10] A. Thammano and P. Kesisung, “Enhancing k-means algorithm for solvingclassification problems,” in Mechatronics and Automation (ICMA),2013 IEEE International Conference on, Aug 2013, pp. 1652–1656.

[11] Y. Zhao and G. Karypis, “Evaluation of hierarchical clusteringalgorithms for document datasets,” in Proceedings of the Eleventh International Conference on Information and

Knowledge Management, ser. CIKM ’02. New York, NY, USA: ACM, 2002, pp. 515–524.[Online]. Available: http://doi.acm.org/10.1145/584792.584877

[12] C.-N. Hsu, H.-S. Huang, and B.-H. Yang, “Global and componentwiseextrapolation for accelerating data mining from large incomplete datasets with the em algorithm,” in Data Mining, 2006. ICDM ’06. SixthInternational Conference on, Dec 2006, pp. 265–274