*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 hereby 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 *ii*

[Chapter I 1](#_Toc509372381)

[Introduction 1](#_Toc509372382)

[1.1. INTRODUCTION 1](#_Toc509372383)

[1.2 Background 1](#_Toc509372384)

[1.3 Importance 1](#_Toc509372385)

[Chapter II 2](#_Toc509372386)

[2 . Overview and Planning 2](#_Toc509372387)

[2.1 OVERVIEW 2](#_Toc509372388)

[1.2. CHALLENGES 2](#_Toc509372389)

[1.3. PROJECT STATEMENT 3](#_Toc509372390)

[1.3.1. EXISTING SYSTEM 3](#_Toc509372391)

[1.3.2. PROPOSED SYSTEM 3](#_Toc509372392)

[1.4. OBJECTIVE 3](#_Toc509372393)

[1.5. SCOPE OF THE PROJECT 4](#_Toc509372394)

[Chapter II 4](#_Toc509372395)

[Background 4](#_Toc509372396)

[2.1. LITERATURE SURVEY 4](#_Toc509372397)

[2.2 Literature Summary 8](#_Toc509372398)

[Chapter III 9](#_Toc509372399)

[System Design 9](#_Toc509372400)

[3.1. HIGH LEVEL DESIGN 9](#_Toc509372401)

[3.1.1. USE CASE DIAGRAM 9](#_Toc509372402)

[3.2. PURPOSE 10](#_Toc509372403)

[3.3. IMPORTANCE 10](#_Toc509372404)

[3.4. ARCHITECTURE DIAGRAM 11](#_Toc509372405)

[Chapter IV 12](#_Toc509372406)

[IMPLEMENTATION AND TESTING 12](#_Toc509372407)

[4.1 Modular Description 12](#_Toc509372408)

[4.2 Modules 12](#_Toc509372409)

[1. Data Pre-processing Model 12](#_Toc509372410)

[2. Data Migration Model with SQOOP 12](#_Toc509372411)

[3. Data Analytic Model with HIVE 12](#_Toc509372412)

[4. Data Analytic Model with PIG 12](#_Toc509372413)

[5. Data Analytic Model with MapReduce 12](#_Toc509372414)

[6. Data Visualization Model with R 12](#_Toc509372415)

[**4.3 Module Description** 12](#_Toc509372416)

[4.3.1 Data Pre-processing Model 12](#_Toc509372417)

[4.3.2 Data Migration Model with SQOOP 13](#_Toc509372418)

[4.3.3 Data Analytic Module with HIVE 14](#_Toc509372419)

[4.3.4 Data Analytic Module with PIG 15](#_Toc509372420)

[4.3.5 Data Analytic Module with MapReduce 15](#_Toc509372421)

[4.3.6 Data Visualization Module with R 16](#_Toc509372422)

[4.4 Table Description 17](#_Toc509372423)

[4.5 Code and Standard 17](#_Toc509372424)

[CHAPTER-V 17](#_Toc509372425)

[**TESTING AND IMPLEMENTATION** 17](#_Toc509372426)

[5.1 Testing 17](#_Toc509372427)

[5.2 TESTING METHODS 18](#_Toc509372428)

[5.2.1 Architecture Testing 18](#_Toc509372429)

[5.2.2 Performance Testing 18](#_Toc509372430)

[5.3 Test Cases 19](#_Toc509372431)

[5.4 Implementation 20](#_Toc509372432)

[REFERENCES 23](#_Toc509372433)

**LIST OF FIGURES**

|  |  |
| --- | --- |
| TITLE | PAGE NO |
| USE CASE DIAGRAM | 8 |
| ARCHITECTURE DIAGRAM | 9 |
| DATA PREPROCESSING DIAGRAM | 11 |
| SQOOP DIAGRAM | 12 |
| HIVE DIAGRAM | 12 |
| PIG DIAGRAM | 13 |
| MAPREDUCE DIAGRRAM | 14 |
| R DIAGRAM | 14 |
| PERFORMANCE TESTING DIAGRAM | 17 |

**Executive Summary:**

Law authorization administrators are progressively perceiving that they are no longer in a data poor world: information and data about the criminal condition and criminal movement flourish. The test is to corral this abundance of information into learning that can upgrade basic leadership, enhance methodologies to battle wrongdoing, and increment wrongdoing aversion benefits. At the end of the day, the point is to change over information and data into significant knowledge.

In several cases, however, this increase in knowledge has not essentially translated to a rise in data. For a lot of the history of enforcement, criminal intelligence information that relates to the activities of criminal people or teams of offenders was maintained by specialised units or by individual detectives. Even with the introduction of intelligence units, these analytical teams usually unbroken their info among the slender orbit of their specific unit.

The project “A SYSTEMATIC APPROACH FOR ANALYZING CRIME OCCURNCES USING BIG DATA” analysing the crime incidents across the city which are present in large quantities of data (Big data) using the tool such as Hadoop, Hive, Pig and R to provide meaningful results for the police to implement change and embrace the information huge environment of modern policing.

The proposed project analysing the huge data of crime occurrences of a particular city and its all information. The data was huge so we can’t able to process the traditional database or software. For processing the huge amount of data we are going to use big data tool to process in less time and produce the meaningful results so that the government should necessary steps to ensure the protection of people. The system also provides the visualization of crime trends across the city so that people can understand easily.

# Chapter I

# Introduction

## 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 analysing 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 analyse 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 Background

Day by day the crime rate is increasing considerably. Crime cannot be predicted since it is neither systematic nor random. Also the modern technologies and hi-tech methods help criminals in achieving their misdeeds. Even though we cannot predict who all may be the victims of crime but can predict the place that has probability for its occurrence. The anticipated outcomes can't be guaranteed of 100% precision yet the outcomes demonstrates that our application helps in diminishing wrongdoing rate to a specific degree by giving security in wrongdoing touchy regions. So to build such an intense wrongdoing examination instrument we need to gather wrongdoing records and assess it

## 1.3 Importance

Crime analysis is a law enforcement function that involves systematic analysis for identifying and analysing patterns and trends in crime and disorder. Data on examples can help law authorization organizations send assets in a more viable way, and help criminologists in distinguishing and capturing suspects. Wrongdoing examination additionally assumes a part in contriving answers for wrongdoing issues, and planning wrongdoing avoidance systems. Quantitative sociology information examination strategies are a piece of the wrongdoing investigation process, however subjective techniques, for example, inspecting police report accounts likewise assume a part.

## 1.4 Organisation of Report

In the Report, the chapter one explains about the introduction describing the system importance. In the chapter two, the report explains about the overview and planning of proposed system with the challenges and assumptions of our project along with the hardware and software requirements. The chapter also explains about the Gantt-chart and workbench structure of our project. In chapter three we have discussed about the Literature Survey and summary of it.

In the chapter four, we have discussed about the system design which includes high level design , Low level design along with the codes, standards ,constraints and trade-offs of our project. Chapter five describes about the system implementation which includes codes with unit testing and integration testing. Outputs and results are described in the chapter six. Chapter Seven describes about the conclusion and future scope of our project. Finally references are clearly described in chapter eight.

# Chapter II

# . Overview and Planning

## 2.1 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.

## 2.2 CHALLENGES

Data Collection:

To collect the relevant data so that they will give enough information. Collection of irrelevant data won’t give any necessary information. So the collection of information is very important.

Data Analysis:

Analysis of data in a very less time is very important for the project.so that we are able to process large amount of data to infer actionable knowledge.

Data Movement:

Move the data from MySQL to Hadoop for processing the data. We are using the SQOOP tool to move the large data from MySQL to Hadoop.

## 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. It will analyse the huge amount of data using big data tools such as

## 2.4 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.

## 2.5 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.

## 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.

## SCOPE OF THE PROJECT

In this system we are analysing 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.

* 1. Architecture specifications

|  |  |
| --- | --- |
| **User Action** | **System Response** |
| Load the Dataset | Dataset loaded to disk |
| Perform the Query | Show results based on the Query |
| Import the database | Database loaded into File Storage |
| Request for result | Show the Result in server |

* 1. **HARDWARE REQUIREMENTS**

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. They are used by software engineers as the starting point for the system design. It shows what the system does and not how it should be implemented.

PROCESSOR : PENTIUM IV 2.6 GHz, Intel Core 2 Duo.

RAM : 4GB DD RAM

MONITOR : 15” COLOR

HARD DISK : 40 GB

* 1. **SOFTWARE REQUIREMENTS**

The software requirements document is the specification of the system. It should include both a definition and a specification of requirements. It is a set of what the system should do rather than how it should do it. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating cost, planning team activities, performing tasks and tracking the team’s and tracking the team’s progress throughout the development activity.

Framework : Hadoop

Database : MY SQL 5.5

Language : Pig, Hive, Java

Tool : Sqoop

Operating System : cent os

IDE : Eclipse

# Chapter III

# Literature Survey and Summary

## 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** F: 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.

**Title** :An intelligent Analysis of a City Crime Data Using Data Mining

**Author** :Malathi. A, Dr. S. Santhosh Baboo and Anbarasi. A

**Year**  : 2011

**Description:**

Data mining applied in the context of law enforcement and intelligence analysis holds the promise of Alleviating crime related problem. In this paper we use a clustering/classify based model to anticipate crime trends. The data mining techniques are used to analyze the city crime data from Tamil Nadu Police Department. The results of this data mining could potentially be used to lessen and even prevent crime for the forth coming years.

**Title** :Using Data Mining Techniques to Analyze Crime patterns in the Libyan

National Crime Data

**Author** :Dr: Zakaria Suliman Zubi,Ayman Altaher Mahmmud

**Year**  : 2011

**Description:**

The proposed system use a clustering/association rules based model to analyze crime and criminals. This model could MLRC. By this model, we aim to explore the applicability of data mining technique in the efforts of crime prevention with particular emphasis to the dataset which was collected from some police departments. Applying some algorithms will demonstrate the overall results of using both algorithms to perform better results rather in association rules mining or clustering. The rules generated by association rule mining could be easily presented in human language which might be used by police officers to help them decided a crime prevention strategy.

**Title** :Propose Data Mining AR-GA Model to Advance Crime analysis

**Author** :Emad K. Jabar, Soukaena H. Hashem, Enas M. Hessian

**Year**  : 2013

**Description:**

This paper presents a proposed three correlated dimensional model AR-GA of identifying a criminal, in the absence of witness or any clue by the forensic experts. In these situations, here we have tried to identify the criminal by correlate the criminal with crime and location of the crime using mining and genetic. The advantages of using KNN imputation in pre-processing are; KNN can predict both qualitative attributes (the most frequent value among the k nearest neighbours) and quantitative attributes (the mean among the k nearest neighbours). It can easily treat instances with multiple missing values. It takes in consideration the correlation structure of the data.

**Title** :Analysis and Prediction of Crimes by Clustering and Classification

**Author** :Rasoul Kiani, Siamak Mahdavi, Amin Keshavarz

**Year**  : 2015

**Description:**

This paper presents a new framework for clustering and predicting crimes based on real data. Examining the methods proposed for crime prediction shows that the parameters such as the effect of outliers in the data mining pre-processing, quality of the training and testing data, and the value of features have not been addressed before. In this framework, the GA was used to improve outlier detection in the pre-processing phase, and the fitness function was defined based in accuracy and classification error parameters. In order to improve the clustering process, the features were weighted, and the low-value features were deleted through selecting a suitable threshold. The proposed method was implemented, and the results of the optimized and non-optimized parameters were compared to determine their quality and effectiveness

## 2.2 Literature Summary

Analysis of large amount of datasets using big data tool. The clustering and classification of data’s will give the inference of the data’s. Data mining was the important technique to be applied on the datasets to find the hidden patterns and to infer new knowledge.

.

.

# Chapter III

# System Design

## 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

MYSQL

Database

Dataset

HDFS

MAPREDUE

PIG

HIVE

Results

Fig 3.1.1 Use Case Diagram

## 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.

## 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.

## ARCHITECTURE DIAGRAM

Crime Database [Chennai]

[MYSQL]

HADOOP

HDFS

SQOOP

IMPORT & EXPORT

HIVE

[Query approach]

RESULT

MRJOBS

[Java programs]

PIG

[SCRIPTING Langue]

3.4.1.Architecture Diagram

# Chapter IV

# IMPLEMENTATION AND TESTING

## 4.1 Modular Description

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and it’s constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

## 4.2 Modules

1. Data Pre-processing Model

2. Data Migration Model with SQOOP

3. Data Analytic Model with HIVE

4. Data Analytic Model with PIG

5. Data Analytic Model with MapReduce

6. Data Visualization Model with R

## 4.3 Module Description

### 4.3.1 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.

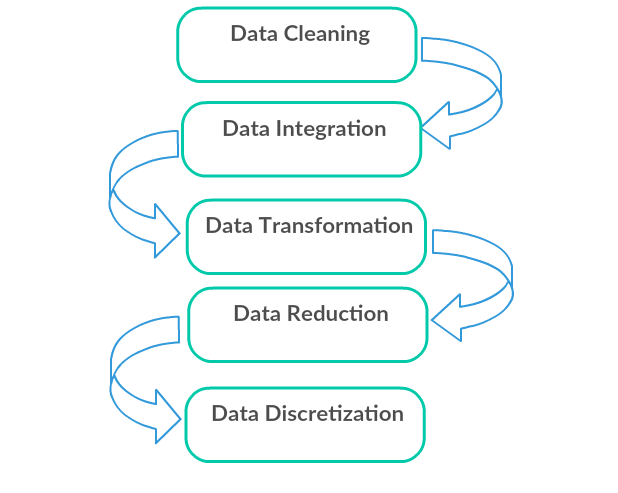


Fig 4.1 Data Pre-processing Diagram

### 4.3.2 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).

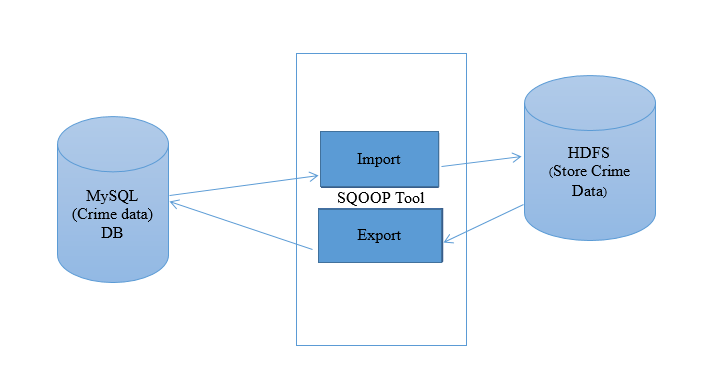


Fig 4.2 SQOOP Diagram

### 4.3.3 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.

Meta-store

(JDBC browse, Query, DDL)

CSV file (crime)

Hive Query language

Execution engine

HDFS (crime data)

[Text file format]

To store the schema/ metadata of tables, databases, columns in a table, their data types, HDFS mapping

MapReduce

Fig 4.3 HIVE Diagram

### 4.3.4 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.

Load (Relational name)

HDFS (Crime Data)

[Text file format]

Store

Fig 4.4 PIG Diagram

### 4.3.5 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.

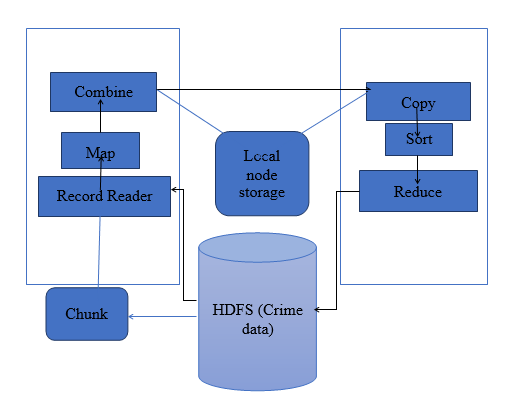


Fig 4.5 MapReduce Diagram

### 4.3.6 Data Visualization Module with R

Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. With interactive visualization, you can take the concept a step further by using technology to drill down into charts and graphs for more detail, interactively changing what data you see and how it’s processed. In this project data can be visualized using R studio R language provides wide variety of tools to visualize the dataset and provide results graphically means.

R studio (use libraries)

Datasets

Fig 4.6 R Diagram

## 4.4 Table Description

Table Name: Crime Data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dc\_Dist | PSA | Date | Time | Pol\_Dis | Type\_Crime | Hour | Area | Year | Month |

Field Description:

1. Dc\_Dist - District Id
2. PSA - Police service Area
3. Date - Date of crime
4. Time - Time when crime occurred
5. Pol\_Dis - Distance of police station from crime occurred
6. Type\_crime - Type of crime
7. Hour - Hour when crime occurred
8. Area - Area where crime occurred

## 4.5 Code and Standard

The application done on VM Ware work station using HADOOP framework, SQOOP, HIVE, PIG ,Map reduce algorithm and R studio.

4.6 Techniques and algorithm

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. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

Algorithm:

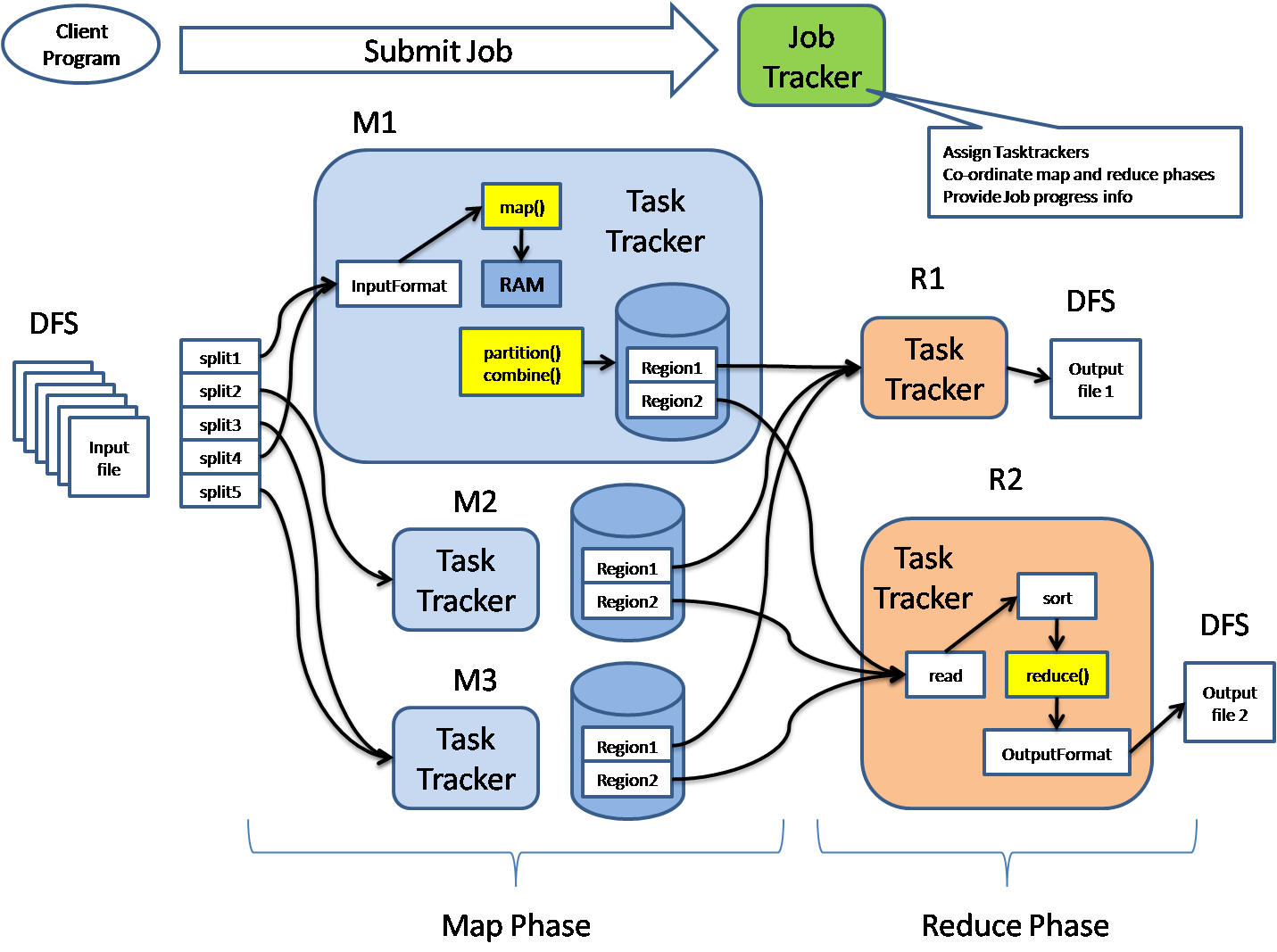
**MapReduce** is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster. MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

**Map stage**: The map or mapper’s job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

**Shuffle stage**: worker nodes redistribute data based on the output keys (produced by the map function), such that all data belonging to one key is located on the same worker node.

**Reduce stage:** This stage is the combination of **Shuffle** stage and the **Reduce** stage. The Reducer’s job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

MapReduce Architecture:



**Hadoop:**

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

Hadoop Architecture

At its core, Hadoop has two major layers namely:

* Processing/Computation layer **(MapReduce**), and
* Storage layer (**Hadoop Distributed File System**).

**HDFS:**

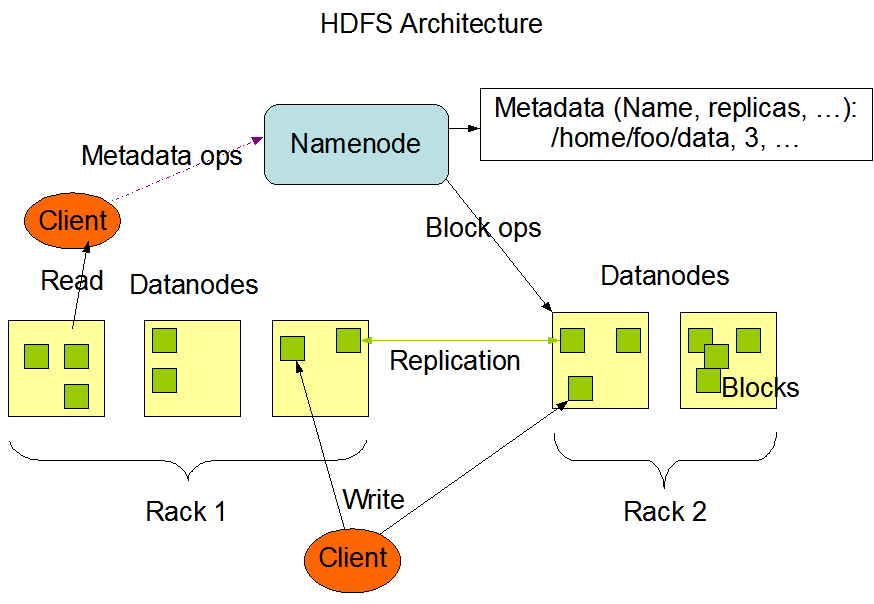
Hadoop File System was developed using distributed file system design. It is run on commodity hardware. Unlike other distributed systems, HDFS is highly fault tolerant and designed using low-cost hardware. HDFS holds very large amount of data and provides easier access. To store such huge data, the files are stored across multiple machines. These files are stored in redundant fashion to rescue the system from possible data losses in case of failure. HDFS also makes applications available to parallel processing.

## Features of HDFS

* It is suitable for the distributed storage and processing.
* Hadoop provides a command interface to interact with HDFS.
* The built-in servers of name node and data node help users to easily check the status of cluster.
* Streaming access to file system data.
* HDFS provides file permissions and authentication.

HDFS Architecture

Given below is the architecture of a Hadoop File System. HDFS follows the master-slave architecture and it has the following elements.



**NameNode**

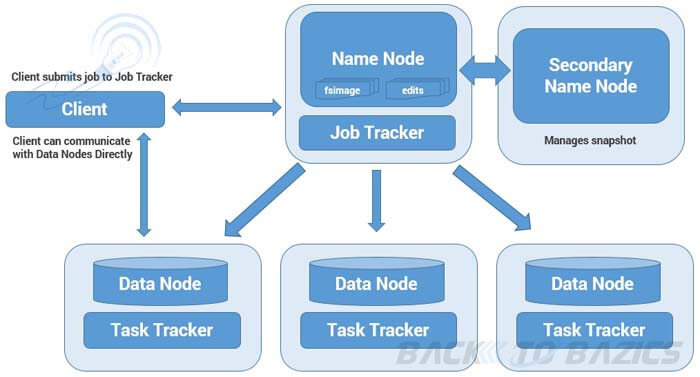
The namenode is the commodity hardware that contains the GNU/Linux operating system and the namenode software. It is software that can be run on commodity hardware. The system having the namenode acts as the master server and it does the following tasks:

* Manages the file system namespace.
* Regulates client’s access to files.
* It also executes file system operations such as renaming, closing, and opening files and directories.

### Datanode

The datanode is a commodity hardware having the GNU/Linux operating system and datanode software. For every node (Commodity hardware/System) in a cluster, there will be a datanode. These nodes manage the data storage of their system.

* Datanodes perform read-write operations on the file systems, as per client request.
* They also perform operations such as block creation, deletion, and replication according to the instructions of the namenode.



### Block

Generally the user data is stored in the files of HDFS. The file in a file system will be divided into one or more segments and/or stored in individual data nodes. These file segments are called as blocks. In other words, **the minimum amount of data that HDFS can read or write is called a Block.** The default block size is 64MB, but it can be increased as per the need to change in HDFS configuration.

**Jobtracker**

It is a master service and it allocates tasks to the task trackers and also monitoring the job scheduling.

**Tasktracker**

A TaskTracker is a node in the cluster that accepts tasks - Map, Reduce and Shuffle operations - from a JobTracker

**Secondary NameNode**

It is a separate physical machine which acts as a helper of name node. It performs periodic check points. It communicates with the name node and take snapshot of meta data which helps minimize downtime and loss of data. It is a master service and backup for metadata not for Namenode.

SQOOP:

SQOOP is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases. It is provided by the Apache Software Foundation. SQOOP IMPORT imports individual tables from RDBMS to HDFS. Each row in a table is treated as a record in HDFS. All records are stored as text data in text files or as binary data in Avro and Sequence files.SQOOP EXPORT exports a set of files from HDFS back to an RDBMS. The files given as input to SQOOP contain records, which are called as rows in table. Those are read and parsed into a set of records and delimited with user-specified delimiter.

HIVE:

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive. It is used by different companies.Hive stores schema in a database and processed data into HDFS.It provides SQL type language for querying called HiveQL or HQL.

Architecture of HIVE:



Description:

|  |  |
| --- | --- |
| **Unit Name** | **Operation** |
| User Interface | Hive is data warehouse infrastructure software that can create interaction between user and HDFS. The user interfaces that Hive supports are Hive Web UI, Hive command line, and Hive HD Insight (In Windows server). |
| Meta Store | Hive chooses respective database servers to store the schema or Metadata of tables, databases, columns in a table, their data types, and HDFS mapping. |
| HiveQL Process Engine | HiveQL is similar to SQL for querying on schema info on the Megastore. It is one of the replacements of traditional approach for Map Reduce program. Instead of writing Map Reduce program in Java, we can write a query for Map Reduce job and process it. |
| Execution Engine | The conjunction part of HiveQL process Engine and Map Reduce is Hive Execution Engine. Execution engine processes the query and generates results as same as Map Reduce results. It uses the flavour of Map Reduce. |
| HDFS or HBASE | Hadoop distributed file system or HBASE are the data storage techniques to store data into file system. |

HIVE working:



Interaction of HIVE with Hadoop:

|  |  |
| --- | --- |
| **Step No.** | **Operation** |
| 1 | **Execute Query**  The Hive interface such as Command Line or Web UI sends query to Driver (any database driver such as JDBC, ODBC, etc.) to execute. |
| 2 | **Get Plan**  The driver takes the help of query compiler that parses the query to check the syntax and query plan or the requirement of query. |
| 3 | **Get Metadata**  The compiler sends metadata request to Metastore (any database). |
| 4 | **Send Metadata**  Metastore sends metadata as a response to the compiler. |
| 5 | **Send Plan**  The compiler checks the requirement and resends the plan to the driver. Up to here, the parsing and compiling of a query is complete. |
| 6 | **Execute Plan**  The driver sends the execute plan to the execution engine. |
| 7 | **Execute Job**  Internally, the process of execution job is a MapReduce job. The execution engine sends the job to JobTracker, which is in Name node and it assigns this job to TaskTracker, which is in Data node. Here, the query executes MapReduce job. |
| 7.1 | **Metadata Ops**  Meanwhile in execution, the execution engine can execute metadata operations with Metastore. |
| 8 | **Fetch Result**  The execution engine receives the results from Data nodes. |
| 9 | **Send Results**  The execution engine sends those resultant values to the driver. |
| 10 | **Send Results**  The driver sends the results to Hive Interfaces. |

PIG:

Apache Pig is a high level data flow platform for execution Map Reduce programs of Hadoop. The language for Pig is pig Latin. The Pig scripts get internally converted to Map Reduce jobs and get executed on data stored in HDFS. Every task which can be achieved using PIG can also be achieved using java used in Map reduce.

Architecture of PIG:

The language used to analyze data in Hadoop using Pig is known as **Pig Latin**. It is a highlevel data processing language which provides a rich set of data types and operators to perform various operations on the data.

To perform a particular task Programmers using Pig, programmers need to write a Pig script using the Pig Latin language, and execute them using any of the execution mechanisms Grunt Shell, UDFs, and Embedded. After execution, these scripts will go through a series of transformations applied by the Pig Framework, to produce the desired output.

Internally, Apache Pig converts these scripts into a series of MapReduce jobs, and thus, it makes the programmer’s job easy. The architecture of Apache Pig is shown below.



# CHAPTER-V

## **TESTING AND IMPLEMENTATION**

### 5.1 Testing

Testing is the process where the test data is prepared and is used for testing the modules individually and later the validation given for the fields. Then the system testing takes place which makes sure that all components of the system property functions as a unit. The test data should be chosen such that it passed through all possible condition. Actually testing is the state of implementation which aimed at ensuring that the system works accurately and efficiently before the actual operation commence. The following is the description of the testing strategies, which were carried out during the testing period.

### 5.2 TESTING METHODS

### 5.2.1 Architecture Testing

Architectural testing is crucial to ensure success of your Big Data project. Poorly or improper designed system may lead to performance degradation, and the system could fail to meet the requirement.

### 5.2.2 Performance Testing

Performance Testing for Big Data includes two main action

* **Data ingestion and Throughout**: In this stage, the tester verifies how the fast system can consume data from various data source. Testing involves identifying different message that the queue can process in a given time frame. It also includes how quickly data can be inserted into underlying data store for example insertion rate into a Mongo and Cassandra database.
* **Data Processing**: It involves verifying the speed with which the queries or map reduce jobs are executed. It also includes testing the data processing in isolation when the underlying data store is populated within the data sets. For example, running Map Reduce jobs on the underlying HDFS
* **Sub-Component Performance**: These systems are made up of multiple components, and it is essential to test each of these components in isolation. For example, how quickly message is indexed and consumed, MapReduce jobs, query performance, search, etc.

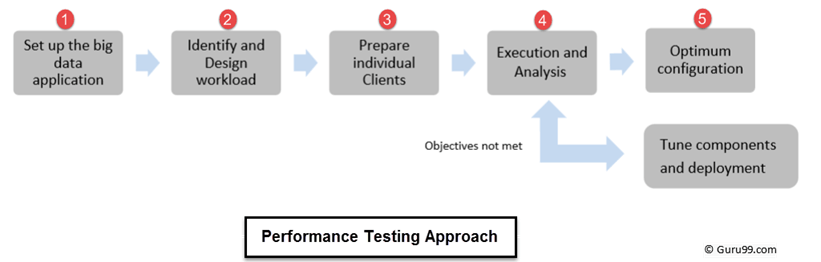


Fig 5.1 Performance Testing Diagram

### 5.3 Test Cases

**TC01:** Data Staging Validation

|  |  |
| --- | --- |
| Test Case ID | TC01 |
| Test Case Description | Pre-Hadoop stage involves process validation |
| Expected Output | Structured HDFS |
| |  |  |  | | --- | --- | --- | | Step | Input | Actual Output | | 1. | Data from various source like RDBMS, weblogs, social media, etc. | Structured HDFS |   **RESULT: TEST CASE PASSED** | |

**TC02:** MapReduce Validation

|  |  |
| --- | --- |
| Test Case ID | TC02 |
| Test Case Description | Data aggregation or segregation rules implemented on the data |
| Expected Output | Key Value pairs are generated |
| |  |  |  | | --- | --- | --- | | Step | Input | Actual Output | | 1. | Business logic validation on every node. | Key Value pairs are generated |   **RESULT: TEST CASE PASSED** | |

**TC03:** Data Visualization in R

|  |  |
| --- | --- |
| Test Case ID | TC03 |
| Test Case Description | Visualize the data based on the libraries |
| Expected Output | Plots and Graphs |
| |  |  |  | | --- | --- | --- | | Step | Input | Actual Output | | 1. | Csv file which has the data | Plots and Graphs |   **RESULT: TEST CASE PASSED** | |

## 5.4 Implementation

5.4.1 MYSQL

1) **Connect to mysql**

mysql connect -u training -p

training

2) Databases operations

show databases

use movielens;

3)Tables operations

show tables;

desc chennai;

4) Retrieve values from Table

select \* from cdata limit 10;

5.4.2 SQOOP

**1)** Import database from mysql to HDFS

sqoop import --connect jdbc:mysql://localhost/movielens --username training --password training --table cdata --m 1 --target-dir crime\_data

2) Databases

sqoop eval --connect jdbc:mysql://localhost --username training --password training --query "show databases"

sqoop eval --connect jdbc:mysql://localhost/movielens --username training --password training --query "show tables"

sqoop eval --connect jdbc:mysql://localhost/movielens --username training --password training --query "select \* from cdata limit no.of rows"

3) Retrieve particular values from databases

sqoop import --connect jdbc:mysql://localhost/movielens --username training --password training --table cdata --m 1 --fields-terminated-by '\t' --columns "area" --target-dir area\_crime

5.4.3 HIVE

1) Enter into HIVE

hive

2) Create Database

create database show;

use show;

3)Create Table

create table crime\_1(type\_crime string,area string)

> row format delimited fields terminated by ',' ;

4) Load values from HDFS to HIVE

load data inpath '/user/training/directoryname/part-m-00000' into table crime;

5) Retrieve values

select \* from crime limit 10;

6) Retrieve particular values

select type\_crime from crime where area=’madapakkam’;

5.4.4 PIG

1) Enter PIG

pig

2) Load from csv file to directory

hadoop fs -put crime cri1

3) Import to a dir

cri1 = LOAD ' /user/training/crime ' using PigStorage(',') AS (type\_crime:chararray,area:chararray);

desc cri1;

4) STORE cri1 into 'cri1result';

DUMP cri1;

g2=GROUP cri1 by (type\_crime);

desc g2;

dump g2;

STORE g2 into 'cri1result'

5.4.5 R Language

library(sqldf)

library(png)

crime <- read.csv('I:\\crime\_analysis\\r\_part\\crimed.csv',TRUE)

head(crime)

DF=sqldf(" SELECT type\_crime,count(\*) as total FROM crime group by type\_crime order by total desc limit 10")

print(DF)

DF$Crime\_type=as.character(DF$type\_crime)

DF$total=as.numeric(DF$total)

barplot(DF$total,names.arg=DF$Crime\_type,col=rainbow(10),ylab="tot\_crimes",cex.names=0.9,las=2,main=" station Road total crimes ",ylim=c(5000,357187))

print(DF)

dev.off() #to shut down all running devices anrd return to this current window

# 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