DEPARTMENT OF COMPUTER ENGINEERING & APPLICATIONS
INSTITUTE OF ENGINEERING & TECHNOLOGY

B.Tech. IV Year CSE(DA)

Project Report

On

'A Data Mining Framework to Analyze Road Accident Data'

Under the supervision of

Mr. Rahul Pradhan, Assistant Professor

Submitted by

1. Rajiv Yadav (H-64/161500434)
2. Hemlata Sarswat (H-57/161500235)
3. Anujay Jain (H-51/161500106)
4. Shahaban Ali(H-70/161500496)

Group No: G-06

Odd Semester, 2019-20

Index

Chapter 1

Introduction

1.1 Motivation and Overview

In the developed as well as developing countries, Infrastructure development is one of the major investments by the government, while safety of passengers on roads is of utmost importance. A road optimization during the construction or during maintenance phase, requires that the engineers analyze all the parameters that play a crucial role in ensuring safety for the passengers and preventing accidents. One of the key objectives in accident data analysis is to identify the main factors associated with road accidents.

This is a research-based data analysis project in which we try to analyze a large data set not capable of being analyzed by typical database or data analysis software like Excel.

To overcome this, we try to implement distributed processing using Hadoop and pipe the result with Apache Zeppelin to analyze and visualize the data set and generate a decision tree.

1.2 Objective

One of the key objectives in accident data analysis to identify the main factors associated with a road and traffic accident. However, heterogeneous nature of road accident data makes the analysis task difficult. Data segmentation has been used widely to overcome this heterogeneity of the accident data. In this paper, we proposed a framework that used K-modes clustering technique as a preliminary task for segmentation of 11,574 road accidents on road network of Dehradun (India) between 2009 and 2014 (both included). Next, association rule mining are used to identify the various circumstances that are associated with the occurrence of an accident for both the entire data set (EDS) and the clusters identified by K-modes clustering algorithm. The findings of cluster based analysis and entire data set analysis are then compared. The results reveal that the combination of k mode clustering and association rule mining is very inspiring as it produces important information that would remain hidden if no segmentation has been performed prior to generate association rules. Further a trend analysis have also been performed for each clusters and

EDS accidents which finds different trends in different cluster whereas a positive trend is shown by EDS. Trend analysis also shows that prior segmentation of accident data is very important before analysis.

1.3 Scope

Traffic Engineers and Government agencies can:

- Identify the basic nature of accidents happening in the selected highway stretches
- Identify the root cause of accidents based on the collected data.
- Identify the features of the road causing the accident.
- Identify and Compare various road segments for optimization.
- Identify road intersection types and the frequency of accidents.
- Run instructional recommendation system.

Chapter 2

Literature Survey

Review of literature is important in any research work. Many researchers have carried out research work in the area of road accidents. Some of them have analyzed accident data in different ways. Some of them Identification of Black spot zone. Some of them have developed accident models for forecasting future accident trends. They have also proposed strategies for road safety.

In the present chapter literature review is carried out covering the different issues related to road accident and road safety.

Yannis T.H. (2014) was presented A Review of The Effect of Traffic and Weather Characteristics on Road Safety. Despite the existence of generally mixed evidence on the effect of traffic parameters, a few patterns can be observed. For instance, traffic flow seems to have a nonlinear relationship with accident rates, even though some studies suggest linear relationship with accidents. Regarding weather effects, the effect of precipitation is quite consistent and leads generally to increased accident frequency but does not seem to have a consistent effect on severity. The impact of other weather parameters on safety, such as visibility, wind speed and temperature is not found straightforward so far. The increasing use of real-time data not only makes easier to identify the safety impact of traffic and weather characteristics, but most importantly makes possible the identification of their combined effect. The more systematic use of these real-time data may address several of the research gaps identified in this research.

K. Meshram and H.S. Goliya (2013) were presented an analysis of accidents on small portion NH-3 Indore to Dhamnod. The data for analysis is collected for the period of 2009 to September 2011. More accidents occurred in Manpur region by faulty road geometry. The trend of accidents occurring in urban portion (Indore) is more than 35 % to rate of total accidents in each year. This may due to high speeds and more vehicular traffic. In the present study area the frequency of fatal accidents are 2 in a week and 6 for minor accidents in a week. More number of accidents observed in 6 p.m. to 8 p.m. duration because in that time more buses are travels between villages and city. One fatal and five casualties are occurring per km per year in the study

area. The volume of the trucks passing through study corridor is increasing by year. At Rajendra Nagar from 2000 onwards the traffic is reduced due to the construction of by passes in that area.

Rakesh Mehar and Pradeep Kumar Agarwal(2013) were highlighted the deficiencies in the present state of the art and also presents some basic concepts so that systematic approach for formulation of a road safety improvement program in India can be developed. The study presents basic concepts to develop an accident record system, for ranking of Safety hazardous locations, for identification of safety improvement measures and to determine priorities of safety measures. It is expected that this study will provide a systematic approach for development of road safety improvement program in India and thus pave the way for improving safety on Indian roads.

E.S.Park (2012) studies the safety effect of wider edge lines was examined by analyzing crash frequency data for road segments with and without wider edge lines. The data from three states, Kansas, Michigan, and Illinois, have been analyzed. Because of different nature of data from each state, a different statistical analysis approach was employed for each state: an empirical Bays, before-after analysis of Kansas data, an interrupted time series design and generalized linear segmented regression analysis of Michigan data, and a cross sectional analysis of Illinois data. Although it is well-known that causation is hard to establish based on observational studies, the results from three extensive statistical analyses all point to the same findings. The consistent findings lend support to the positive safety effects of wider edge lines installed on rural, two-lane highways. In conclusion, this study lends scientific support to the positive safety effects of wider edge lines installed on rural two-lane highways. Although the magnitudes of crash reductions were somewhat different from state to state, the results point in the same direction.

Amir H. Ghods et al. (2012) Differential speed strategies increased the number and rate of car-truck overtakes over the range of volumes considered in this analysis. This suggests a negative effect on safety resulting from differential speed strategy applied to two-lane rural highways. On a positive side DSL and MSL strategies have reduced the number of car-car overtakes at different volumes, hence increasing safety. This latter relationship suggests a calming effect of slower trucks on the speed of the traffic stream, which results in fewer interactions between cars. No significant effect was observed concerning differential speed control strategies and both average TTC and PTDO. The effect on TTC was due to volume; highest TTC for car-car and car-truck interactions at very low volumes, decreasing to a minimum in the range between 500 vph to 800 vph and increasing slightly thereafter. This indicator suggests the highest head-on risk

is experienced in the mid volume region. The average speed of traffic decreases in a nonlinear fashion with volume with differential speed strategies indicating a downward shift in this relationship.

Michael Williamson and Huaguo Zhou (2012) were the development of calibration factors for crash prediction models in the new Highway Safety Manual (HSM) for rural two-lane roadways in Illinois. The crash prediction modes (so called Safety Performance Functions (SPF)) in the HSM were developed using data from multiple states, therefore the models must be calibrated to account for local factors, such as weather, roadway conditions, and drivers' characteristics. In this study, two calibration factors were developed for two different SPFs to give a better prediction of crash frequencies on rural two lane roadways in Illinois. This study determined the SPF that best predicts the crashes was developed specifically for rural two-lane Two-way roadways in Illinois. It is recommended that local SPFs be developed and compared to the HSM SPF when evaluating the safety of a roadway.

R.R. Dinu, A. Veeraragavan (2011) was presented Random Parameter Models for Accident Prediction on Two-Lane Undivided Highways in India. Based on three years of accident history, from nearly 200 km of highway segments, is used to calibrate and validate the models. The results of the analysis suggest that the model coefficients for traffic volume, proportion of cars, motorized two-wheelers and trucks in traffic, and driveway density and horizontal and vertical curvatures are randomly distributed across locations. They have concluded with a discussion on modeling results and the limitations of the present study.

Chapter 3

Proposed Model

3.1 System Architecture & Modules

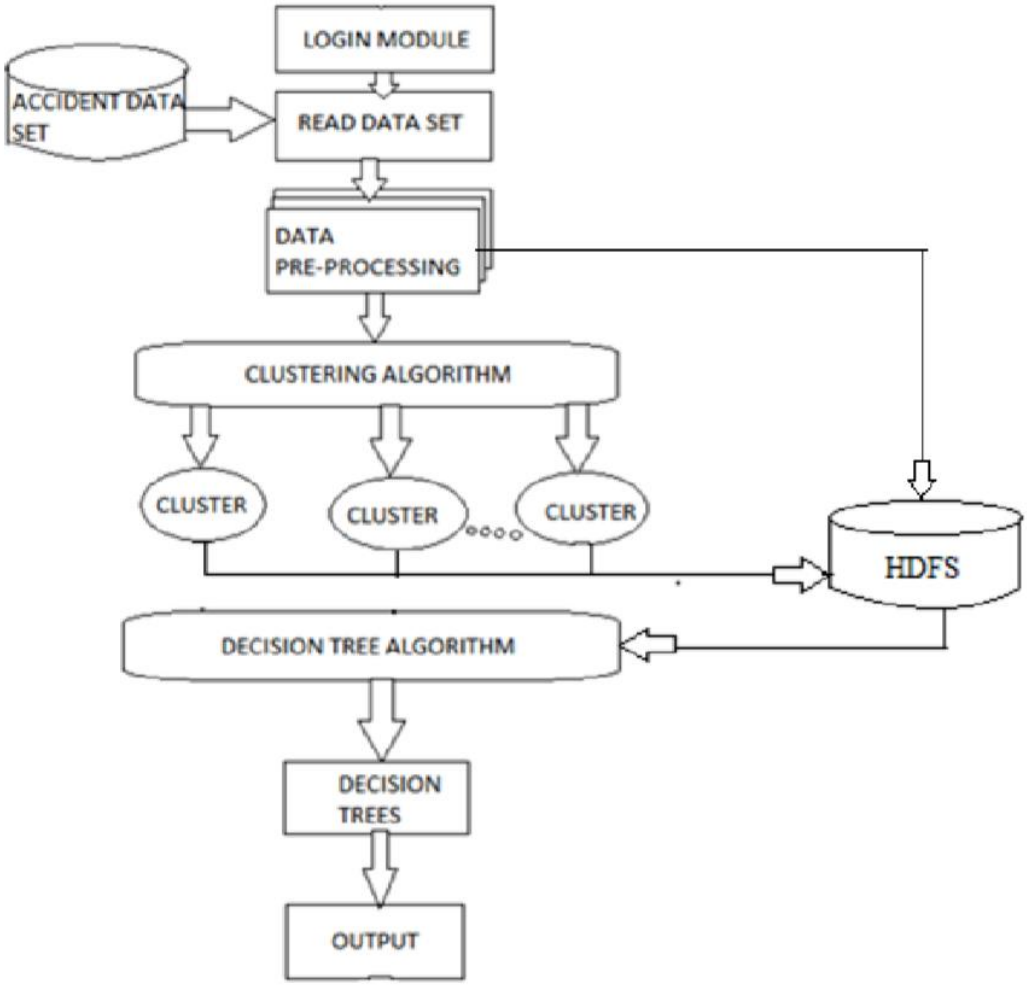


Figure 3.1 Architecture of Proposed System

Basically, there are five modules that complete the project as listed below:

Module 1: Login Module

- Username and Password authentication
- Starting Hadoop distributed file system(HDFS)

Module 2: Data Pre-Processing

- Converting Unstructured data to Structured data for Pre-Processing
- Data cleaning
- Removing Missing Values
- Removing Noisy data
- Removing duplicate records
- Integration of data sets
- Fragmentation and replication for Hadoop

Module 3: Clustering Module

- Setting Up Master and Slave nodes.
- Processing MapReduce jobs in a parallel environment.
- Fetching MapReduce Gain output to Zepellin for tree induction.

Module 4: Attribute Selection and Tree Induction

- A Data Mining functionality for generating Decision Tree.
- Preparing Training Data set.
- Preparing Validation Data set.

Module 5: Visualization with Apache Zeppelin

- Data Visualization.
- Other Statistical analysis

3.1 Login Modules

This phase of the project involves login part of the particular data set. In this, we can login using a particular username and password generated for the particular data.

If we want to login to a particular data set, we can login using that username and password that uses Hadoop technology.

- **Username and Password Authentication**

  The ***su*** command is used to enter the username and password to gain privileges into the Windows account.

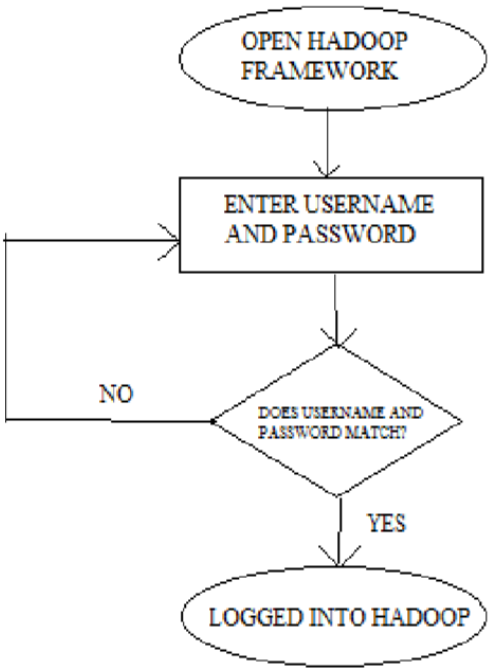


Figure 3.1.1 Flowchart of Login Module

- **Starting Hadoop Distributed File System(HDFS)**

  ssh localhost command is used to enable secure shell. This is required to start all the daemon process on all the nodes from one machine. start -dfs.sh command, start -yarn.sh

command is used to start the Hadoop distributed file system with the Namenode and YARN daemons. stop -dfs.sh command, stop -yarn.sh command is used to terminate the Hadoop distributed file system and YARN daemons.

3.2 Data Pre-Processing

- **Converting Unstructured data to Structured data for Pre-Processing**

Three words to describe Big Data are:

- Volume
- Velocity
- Variety

The concept of developing processes to manage the increasing 'volumes' and 'velocity' of data almost seems conceivable.

Structured data is relatively simple and easy to use in process improvements as the data generally resides in databases in the form of columns and rows. It is grouped into relations or classes based upon shared characteristics. The data is generally allocated attributes (data descriptions) related to the classes within each group to help in ordering and logically grouping. Finally, it can be described by predefined formats (string or value) with predefined lengths of characters.

This makes structured data a good starting point for anyone looking for robust data to create information upon which to form meaningful insights. Structured data can be queried and analyzed to sort, group, filter, count and sum in order to answer business questions or measure process capability. Whilst this doesn't account for the validity of the data it does enable relatively easy processing to verify and observe the data. Structured data forms a large part of the data used by many in process improvements, however this trend is quickly changing as the dominance of unstructured data increases.

Unstructured data is a generic term used to describe data that doesn't sit in databases and is a mixture of textual and non-textual data. Unstructured non-textual data generally relates to media such as images, video and audio files. As the volumes of this type of data increases through the use of smart phones and mobile Internet the need to analyses and understand it grows too. Slightly less unwieldy are unstructured textual data made up of media files (documents, spreadsheets, presentations), email messages and an array of other files generated and stored on corporate networks.

As unstructured data resides on corporate networks, within collaboration tools and in the cloud, it can be extremely difficult to interrogate, or event locate. In order to, search the data, processes need to be in place to help tag and sort it. This step is key to allow for semantic searching against key words or contexts. Unstructured data is being utilized in a big way for social media companies wanting to understand their markets and customers in more depth. This presents the same opportunities to many of our businesses to help understand not only its customers better, but operations within.

The challenge for businesses is to develop processes to apply structure to the unstructured nature of the data. For example, determining the level of satisfaction of customers by analyzing emails and social media may involve searching for words or phrases. Words and phrases may be grouped into positive, negative or neutral classifications.

At this stage, the unstructured data is transformed to structured data where the groups of words found based upon their classification are assigned a value. A positive word may equal 1, a negative -1 and a neutral 0. This unstructured data can now be stored and analyzed as you would with structured data. Much more work is needed in this area to analyses the unstructured non-textual data and many of the big vendors are working on solutions.

- **Data cleaning**

Data preprocessing is one of the important tasks in data mining. Data preprocessing mainly deals with removing noise, handle missing values, removing irrelevant attributes in order to make

the data ready for the analysis. In this step, our aim is to preprocess the accident data in order to make it appropriate for the analysis.

- **Integration of data sets**

Data integration involves combining data residing in different sources and providing users with a unified view of them. This process becomes significant in a variety of situations, which include both commercial (such as when two similar companies need to merge their databases) and scientific domains. Data integration appears with increasing frequency as the volume and the need to share existing data explodes. It has become the focus of extensive theoretical work, and numerous open problems remain unsolved.

- **Fragmentation and replication for Hadoop**

In some operating system's file systems, a data file over a certain size is stored in several chunks or fragments rather than in a single contiguous sequence of bits in one place on the storage medium, a process that is called fragmentation. This allows small unused sections of storage (for example, where old data has been deleted) to be reused.

Replication is the process of making a replica (a copy) of something. A replication (noun) is a copy. The term is used in fields as varied as microbiology (cell replication), knitwear (replication of knitting patterns), and information distribution (CD-ROM replication).

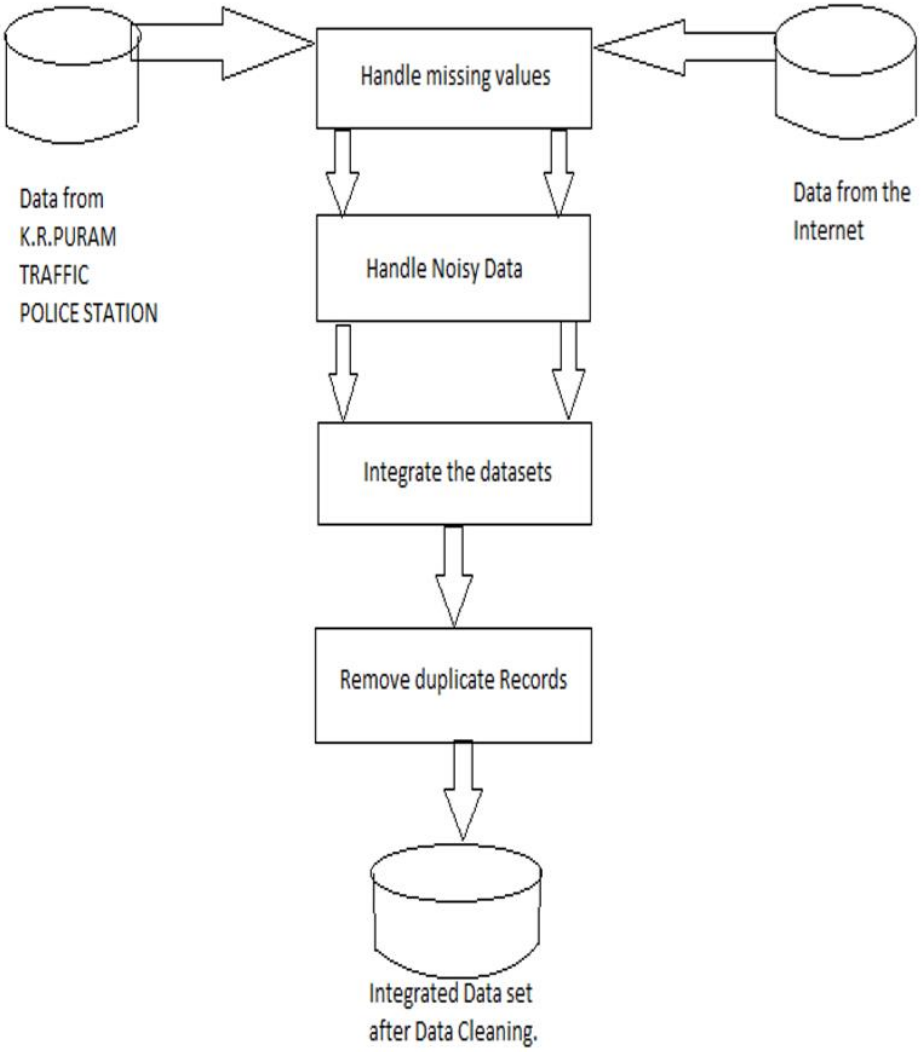


Figure 3.2.1: Data Pre-Processing

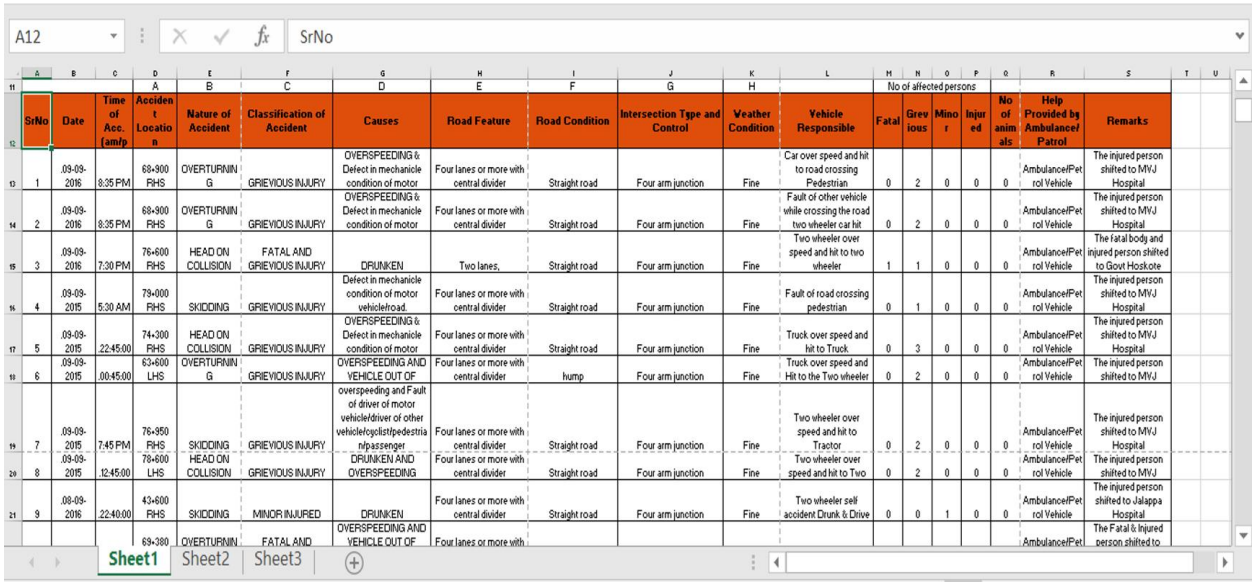
Pre-processing is a huge task in any data analysis project as the data input to the system is not in polished format and needs to be cleaned and prepared well using a series of steps before it can be used for the actual data mining process. The initial dataset is in a raw state which is also known as unstructured data format. It is meaningless and cannot be analyzed before structuring it using a metadata dictionary.

Some of the challenges faced in this project related to the data pre-processing are handling missing values and handling noisy values as a part of data cleaning, integrating the data from 3 different datasets using vertical joins and removing duplicate records for multiple index entries. The first phase of data cleaning includes removing the missing values. It is necessary

to remove missing values from the data as the values not present can actually affect the results of the analysis in a negative way. So, to the file system we need to omit the records that have missing values.

After removing the missing values from the dataset, we have achieved one of the cleaning tasks. The next task is to remove the noisy values from the dataset. The noisy values in the dataset can be anything from -1 to infinity values garbage integers in place of expected ranges or negative values. According to that approach we handle the noisy values from some attributes by assigning the mean values of the attributes that belong to the same class as the attributes with the noisy values. This ensures that there is no over fitting of data. The next step in preprocessing is that of Integration. Integration in this case mainly has the issue of vertical join and also of removing duplicate records due to multiple indices.

Unstructured data is a generic label for describing data that is not contained in a database or some other type of data structure . If left unmanaged, the sheer volume of unstructured data that's generated each year within an enterprise can be costly in terms of storage. The information contained in unstructured data is not always easy to locate. It requires that data in both electronic and hard copy documents and other media be scanned so a search application can parse out concepts based on words used in specific contexts. This is called semantic search. The below figure shows the Unstructured Accident Data.



SrNo	Date	Time of Acc. (am/p	Accident Location	Nature of Accident	Classification of Accident	Causes	Road Feature	Road Condition	Intersection Type and Control	Weather Condition	Vehicle Responsible	No of affected persons: Fatal	Grevious	Minor	Injured	No of animals	Help Provided by Ambulance/ Patrol	Remarks
1	.09-09-2016	8:35 PM	68+900 RHS	OVERTURNING	GRIEVIOUS INJURY	OVERSPEEDING & Defect in mechanicle condition of motor	Four lanes or more with central divider	Straight road	Four arm junction	Fine	Car over speed and hit to road crossing Pedestrian	0	2	0	0	0	Ambulance/Petrol Vehicle	The injured person shifted to MVJ Hospital
2	.09-09-2016	8:35 PM	68+900 RHS	OVERTURNING	GRIEVIOUS INJURY	OVERSPEEDING & Defect in mechanicle condition of motor	Four lanes or more with central divider	Straight road	Four arm junction	Fine	Fault of other vehicle while crossing the road two wheeler car hit	0	2	0	0	0	Ambulance/Petrol Vehicle	The injured person shifted to MVJ Hospital
3	.09-09-2016	7:30 PM	76+600 RHS	HEAD ON COLLISION	FATAL AND GRIEVIOUS INJURY	DRUNKEN	Two lanes,	Straight road	Four arm junction	Fine	Two wheeler over speed and hit to two wheeler	1	1	0	0	0	Ambulance/Petrol Vehicle	The Fatal body and injured person shifted to Govt Hoskote
4	.09-09-2015	5:30 AM	79+000 RHS	SKIDDING	GRIEVIOUS INJURY	Defect in mechanicle condition of motor vehicle/road,	Four lanes or more with central divider	Straight road	Four arm junction	Fine	Fault of road crossing pedestrian	0	1	0	0	0	Ambulance/Petrol Vehicle	The injured person shifted to MVJ Hospital
5	.09-09-2015	22:45:00	74+300 RHS	HEAD ON COLLISION	GRIEVIOUS INJURY	OVERSPEEDING & Defect in mechanicle condition of motor	Four lanes or more with central divider	Straight road	Four arm junction	Fine	Truck over speed and hit to Truck	0	3	0	0	0	Ambulance/Petrol Vehicle	The injured person shifted to MVJ Hospital
6	.09-09-2015	00:45:00	63+600 LHS	OVERTURNING	GRIEVIOUS INJURY	OVERSPEEDING AND VEHICLE OUT OF	Four lanes or more with central divider	hump	Four arm junction	Fine	Truck over speed and Hit to the Two wheeler	0	2	0	0	0	Ambulance/Petrol Vehicle	The injured person shifted to MVJ
7	.09-09-2015	7:45 PM	76+950 RHS	SKIDDING	GRIEVIOUS INJURY	overspeeding and Fault of driver of motor vehicle/driver of other vehicle/cyclist/pedestrian/passenger	Four lanes or more with central divider	Straight road	Four arm junction	Fine	Two wheeler over speed and hit to Tractor	0	2	0	0	0	Ambulance/Petrol Vehicle	The injured person shifted to MVJ Hospital
8	.09-09-2015	12:45:00	78+600 LHS	HEAD ON COLLISION	GRIEVIOUS INJURY	DRUNKEN AND OVERSPEEDING	Four lanes or more with central divider	Straight road	Four arm junction	Fine	Two wheeler over speed and hit to Two	0	2	0	0	0	Ambulance/Petrol Vehicle	The injured person shifted to MVJ
9	.08-09-2016	22:40:00	43+600 RHS	SKIDDING	MINOR INJURED	DRUNKEN	Four lanes or more with central divider	Straight road	Four arm junction	Fine	Two wheeler self accident Drunk & Drive	0	0	1	0	0	Ambulance/Petrol Vehicle	The injured person shifted to Jalappa Hospital
			69+380	OVERTURNING	FATAL AND	OVERSPEEDING AND VEHICLE OUT OF	Four lanes or more with										Ambulance/Petrol	The Fatal & Injured person shifted to

Figure 3.2.2: Unstructured Data Set

Structured data refers to any data that resides in a fixed field within a record or file. This includes data contained in relational databases and spreadsheets. Structured data first depends on creating a data model – a model of the types of business data that will be recorded and how they will be stored, processed and accessed. This includes defining what fields of data will be stored and how that data will be stored: data type and any restrictions on the data input.

Structured data has the advantage of being easily entered, stored, queried and analyzed. At one time, because of the high cost and performance limitations of storage, memory and processing, relational databases and spreadsheets using structured data were the only way to effectively manage data. The figure below shows structured Accident data.

accdate	acctime	acclocation	nature	classification	causes	roadfeature	roadcondition	fatal	grevious	minor	injured
09-09-2016	8:35 PM	68+900 RHS	OVERTURNING	GRIEVIOUS INJURY	OVERSPEEDING & De...	Four lanes or mor...	Straight road	0	2	0	0
09-09-2016	8:35 PM	68+900 RHS	OVERTURNING	GRIEVIOUS INJURY	OVERSPEEDING & De...	Four lanes or mor...	Straight road	0	2	0	0
09-09-2016	7:30 PM	76+600 RHS	HEAD ON COLLISION	FATAL AND GRIEVIO...	DRUNKEN	Two lanes	Straight road	1	1	0	0
09-09-2015	5:30 AM	79+000 RHS	SKIDDING	GRIEVIOUS INJURY	Defect in mechani...	Four lanes or mor...	Straight road	0	1	0	0
09-09-2015	10:45 PM	74+300 RHS	HEAD ON COLLISION	GRIEVIOUS INJURY	OVERSPEEDING & De...	Four lanes or mor...	Straight road	0	3	0	0
09-09-2015	12:45 AM	63+600 LHS	OVERTURNING	GRIEVIOUS INJURY	OVERSPEEDING AND ...	Four lanes or mor...	hump	0	2	0	0
09-09-2015	7:45 PM	76+950 RHS	SKIDDING	GRIEVIOUS INJURY	overspeeding and ...	Four lanes or mor...	Straight road	0	2	0	0
09-09-2015	12:45 PM	78+600 LHS	HEAD ON COLLISION	GRIEVIOUS INJURY	DRUNKEN AND OVERS...	Four lanes or mor...	Straight road	0	2	0	0
08-09-2016	10:40 PM	43+600 RHS	SKIDDING	MINOR INJURED	DRUNKEN	Four lanes or mor...	Straight road	0	0	1	0
08-09-2016	10:45 PM	69+380 RHS	OVERTURNING	FATAL AND GRIEVIO...	OVERSPEEDING AND ...	Four lanes or mor...	Slight Curve	1	3	0	0
08-09-2014	7:30 PM	21+600 RHS	OVERTURNING	FATAL AND MINOR I...	OVERSPEEDING	Four lanes or mor...	Straight road	1	0	1	0
06-09-2015	3:50 PM	12+800 RHS	RIGHT TURN COLLISION	MINOR INJURED	OVERSPEEDING	Three lanes or mo...	Straight road	0	0	0	0
06-09-2014	12:40 PM	6+050 LHS	RIGHT TURN COLLISION	GRIEVIOUS INJURY ...	OVERSPEEDING	Four lanes or mor...	Straight road	0	2	1	3
05-09-2015	11:15 AM	62+350 RHS	RIGHT TURN COLLISION	FATAL	OVERSPEEDING & Ve...	Four lanes or mor...	Slight Curve	1	0	0	0

Figure 3.2.3: Structured Data After Pre-Processing

3.3 Clustering Module

A Hadoop cluster is a special type of computational cluster designed specifically for storing and analyzing huge amount of unstructured data in a distributed computing environment.

Clustering analysis is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.

This Phase of the project involves the distributed processing of large dataset. It is not possible to load the dataset of 1 GB and above size into the memory of a simple computer which

we use at desktop. Hence to process such large amount of data we use the Map Reduce paradigm of Hadoop framework.

The main focus in this phase is to analyze 1GB of data from the dataset and calculate their Information gain, entropy, split info for the attributes.

In a Hadoop cluster, a MapReduce program is known as a job. A job is run by being broken down into pieces, known as tasks. These tasks are scheduled to run on the nodes in the cluster where the data exists.

MapReduce jobs are executed by YARN in the Hadoop cluster. The YARN Resource Manager spawns a MapReduce Application Master container, which requests additional containers for mapper and reducer tasks. The Application Master communicates with the Namenode to determine where all of the data required for the job exists across the cluster. It attempts to schedule tasks on the cluster where the data is stored, rather than sending data across the network to complete a task. The YARN framework and the Hadoop Distributed File System (HDFS) typically exist on the same set of nodes, which enables the Resource Manager program to schedule tasks on nodes where the data is stored.

As the name MapReduce implies, the reduce task is always completed after the map task. A MapReduce job splits the input data set into independent chunks that are processed by map tasks, which run in parallel. These bits, known as tuples, are key/value pairs. The reduce task takes the output from the map task as input and combines the tuples into a smaller set of tuples.

Each MapReduce Application Master monitors its spawned tasks. If a task fails to complete, the Application Master will reschedule that task on another node in the cluster.

This distribution of work enables map tasks and reduce tasks to run on smaller subsets of larger data sets, which ultimately provides maximum scalability. The MapReduce framework also maximizes parallelism by manipulating data stored across multiple clusters.

MapReduce

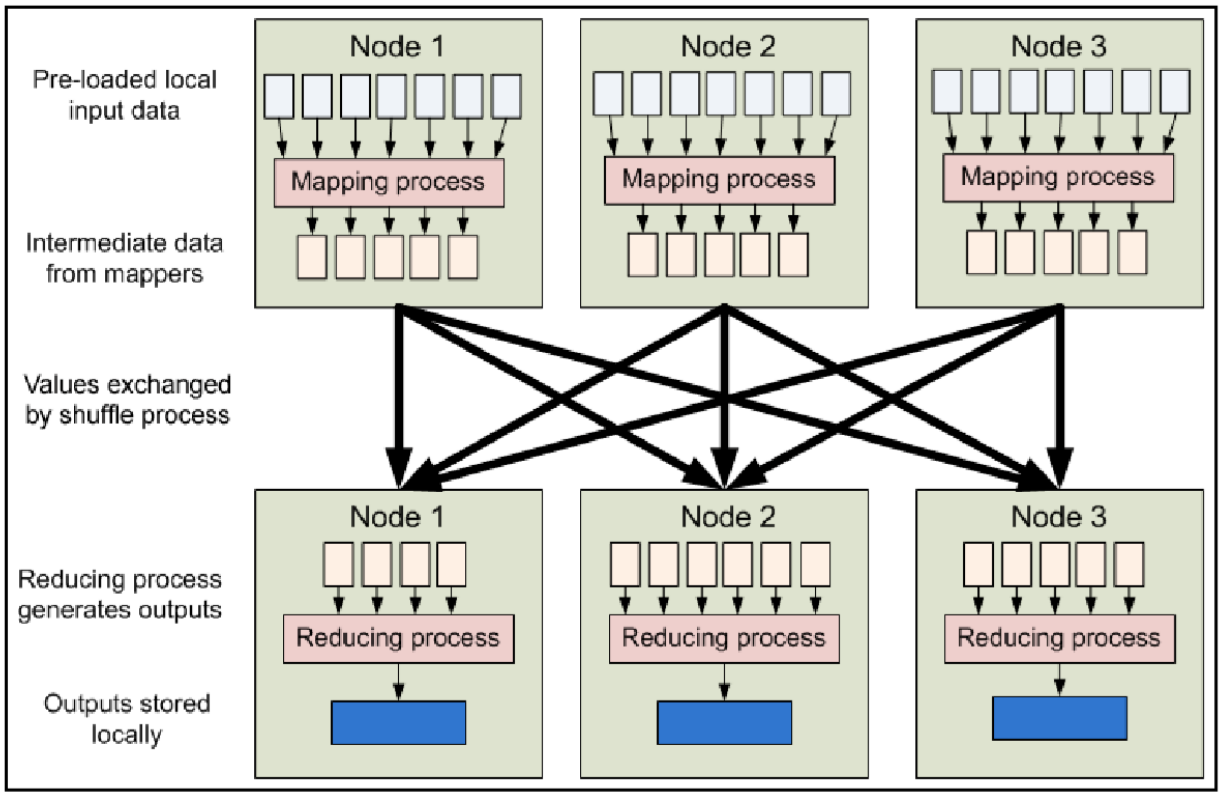


Figure 3.3.1: Clustering Using MapReduce

MapReduce is a programming model and an associated implementation for processing and generating Big Data sets with parallel distributed algorithm on a cluster.

A MapReduce program is composed of Map( ) procedure that performs filtering and sorting. The Map function in the Map Reduce program reads each row one by one and does the preliminary processing of counting and calculating. It will then write the output simultaneously into intermediate storage on HDFS or Local file system as per run configurations and a Reduce( ) method that performs a summary operation. The reduce function takes each row from the output of Map function and then aggregates them based on key-value pair, calculates the final gain and outputs it to the file system.

At the end, a simple file which lists the attributes with their respective information details is obtained. This analysis is the result of distributed processing and takes several hours on a

small Hadoop cluster of 3-4 nodes. Increasing the number of data nodes in the Hadoop cluster significantly reduces the processing time.

This model is a specialization of *Split-Apply-Combine* strategy for data analysis. It is inspired by the Map and Reduce functions commonly used in Functional programming.

BIBLIOGRAPHY

1. Sachin Kumar , Durga Toshniwal ,"Analyzing Road Accident Data Using Association Rule Mining, International Conference on Computing, Communication and Security (ICCCS)", IEEE 2015.

2. An Shi,Zhang Tao, Zhang Xinming, Wang Jian, "Evolution of Traffic Flow Analysis under Accidents on Highways Using Temporal Data Mining", Fifth International Conference on Intelligent Systems Design and Engineering Applications,2014.

3. Eyad Abdullah, Ahmed Emam, "Traffic Accidents Analyzer Using Big Data", International Conference On Computational Science and Computational Intelligence, 2015.

4. Seoung-hun Park ,Young-guk Ha, "Large Imbalance Data Classification Based on MapReduce for Traffic Accident Prediction", Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing,2014.

5. Lokesh Hebbani, "Road Safety Scenario in India Problems & Solutions", 5th Foundation Day Lecture CiSTUP, IISC January 10, 2014.

6. Costabilea. J., Walla, J., Vecovskia, V & Baileya, "The rapid deployment of an effective road safety counter measure through a smart phone application- The story of Speed Adviser", Proceedings of the Australasian Road Safety Research, Policing & Education Conference November,2014.