CAMBRIDGE INSTITUTE OF TECHNOLOGY

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

DATA MINING FRAMEWORK TO ANALYZE ROAD ACCIDENT DATA

PROJECT GUIDE

Prof .Bharani B.R.

(Asst. Professor)

TEAM MEMBERS

Aishwarya Saseendran(1CD13IS005)

Aishwarya Vaishali (1CD13IS006)

Ajitha AS (1CD13IS007)

Swetha D (1CD13IS053)

CONTENTS

ABSTRACT

- In the developed as well as developing countries, Infrastructure development is one of the major investment 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.

Cont.

- The data to be analyzed(both structured and unstructured) is collected from various sources and has several attributes.It is a challenge to gather all such relevant data, detect and analyze it together to generate decision trees that give insights on previous accidents.

- For this purpose,we propose to harness the power of Data Mining technologies like Hadoop.The analysis will be represented in the form of Decision tree which can be represented graphically.

INTRODUTION

- Road accidents are uncertain and unpredictable incidents and their analysis requires the knowledge of the factors affecting them.
- The major problem in the analysis of accident data is its Heterogenous nature.
- Thus, heterogenity must be considered during analysis of the data, otherwise some relationship between the data may remain hidden.
- Although, researchers used segmentation of the data to reduce this heterogeneity using some measures such as expert knowledge, but there is no guarantee that this will lead to an optimal segmentation which consists of homogeneous groups of road accidents.
- Therefore, cluster analysis can assist the segmentation of road accidents.

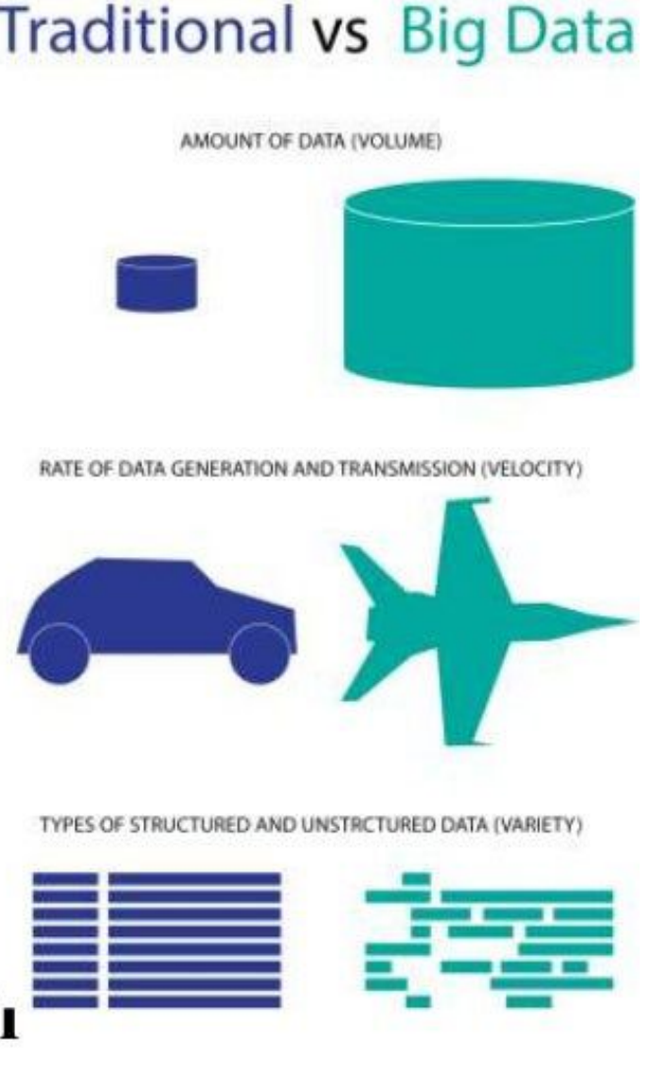
PROBLEM DEFINITION

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 Zepellin to analyze and visualize the data set and generate a decision tree.

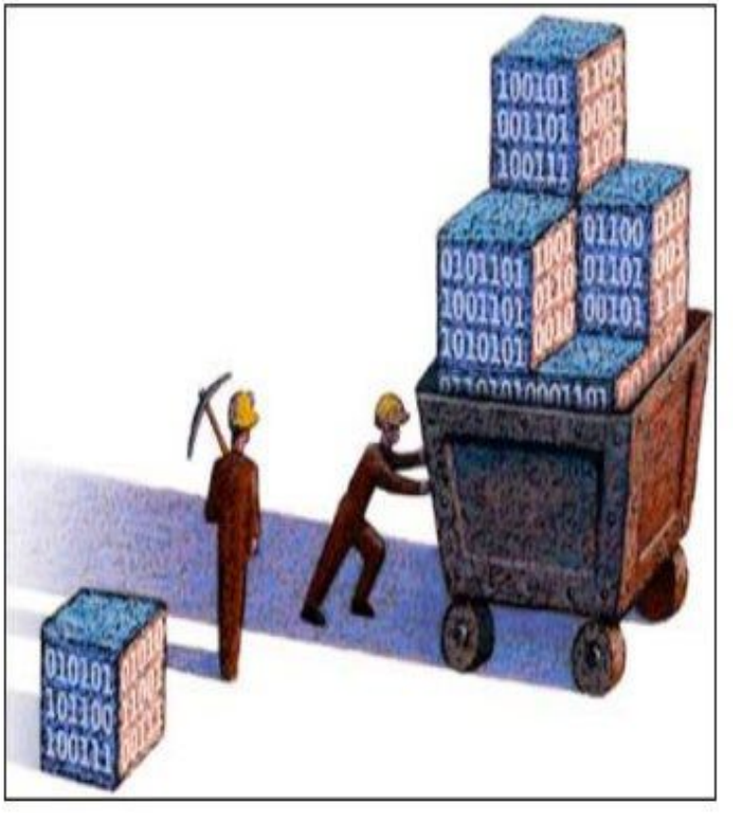
EXISTING SYSTEM

- **The traditional analysis method mainly depends on database system and the education of customers.**
- **The database system are limited to size , inaccurate and takes more time for huge data set.**
- **Database systems can process only structured data.**
- **Therefore using a traditional database will not be efficient**



PROPOSED SYSTEM

- **The proposed system overcomes the above mentioned issues in an efficient way.**
- **The proposed system uses Data mining approach to compare various road accidents occurred in last 5 years and identify highly accident prone areas.**
- **Government agencies can run instructional recommendation systems based on this analysis.**



LITERATURE SURVEY

Author	Objective	Data Mining Techniques	Accuracy
Chaozhonget.al (2009)	To identify the factors significantly influencing single vehicle crash severity.	Random Forest, Rough set theory	0.73%
Ali et.al (2010)	To identify Most important factors which affect injury severity	Classification & Regression tree	72.49%
Liping et.al (2010)	To predict Traffic accident duration of incident and driver information system	Artificial neural Networks	85.35%
DipoT.Akomolafe, Akinbola Olutayo (2012)	To predict causes of accidents and accident prone locations.	Decision tree: Id3, Functional tree	70.27%
Tibebe et.al (2013)	To Explore the possible application of data mining technology for developing a classification model	Classification & Regression tree	87.47%

REQUIREMENTS

HARDWARE REQUIREMENTS

PROCESSOR TYPE: Intel® Core™ i5-6200U CPU

PROCESSOR SPEED: 2.30 GHz

HARD DISK: 1 TB

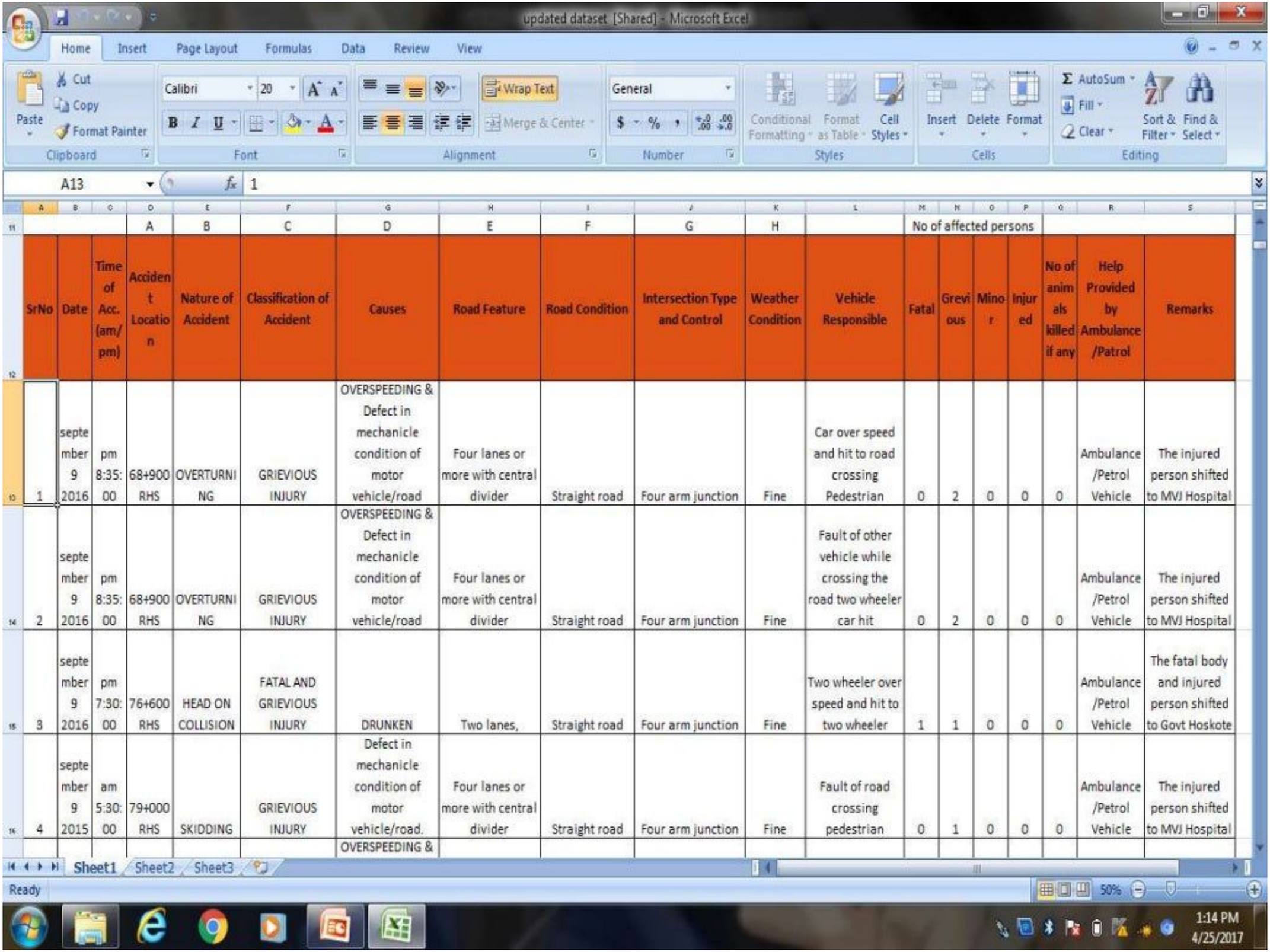
RAM: 8 GB

SOFTWARE REQUIREMENTS

OPERATING SYSTEM: UBUNTU

FRONT END: APACHE ZEPPELIN

BACK END: SCALA-SPARK



updated dataset [Shared] - Microsoft Excel

Home Insert Page Layout Formulas Data Review View

Cut Copy Format Painter Paste Clipboard

Calibri 20 Font

Wrap Text Merge & Center Alignment

General Number

Conditional Formatting Format as Table Cell Styles Styles

Insert Delete Format Cells

AutoSum Fill Clear Sort & Filter Find & Select Editing

A13 1

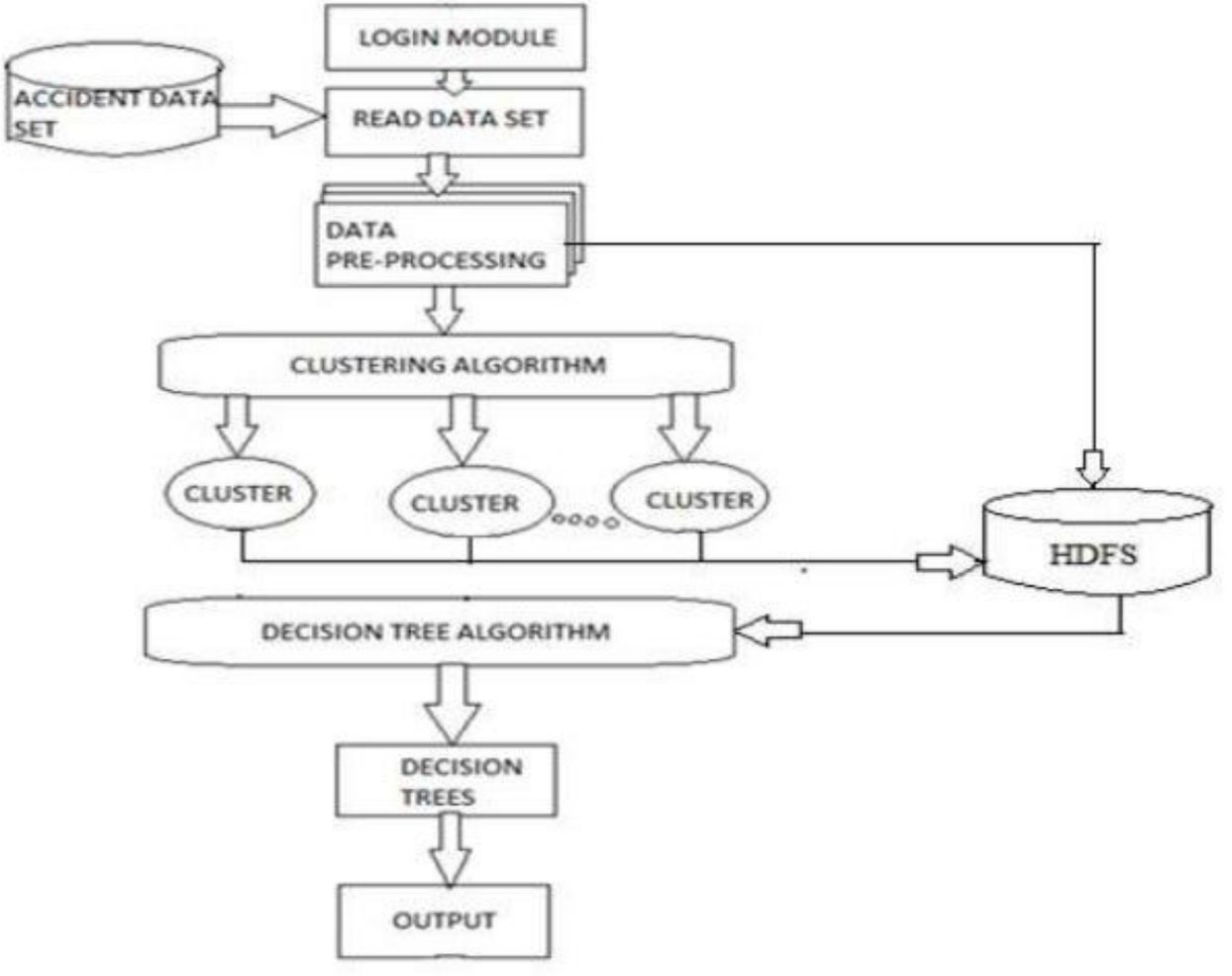
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
11				A	B	C	D	E	F	G	H		No of affected persons						
12	SrNo	Date	Time of Acc. (am/ pm)	Accident Location	Nature of Accident	Classification of Accident	Causes	Road Feature	Road Condition	Intersection Type and Control	Weather Condition	Vehicle Responsible	Fatal	Grevious	Minor	Injured	No of animals killed if any	Help Provided by Ambulance /Patrol	Remarks
13	1	september 9 2016	pm 8:35:00	68+900 RHS	OVERTURNING	GRIEVIOUS INJURY	OVERSPEEDING & Defect in mechanicle condition of motor vehicle/road	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
14	2	september 9 2016	pm 8:35:00	68+900 RHS	OVERTURNING	GRIEVIOUS INJURY	OVERSPEEDING & Defect in mechanicle condition of motor vehicle/road	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
15	3	september 9 2016	pm 7:30:00	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
16	4	september 9 2015	am 5:30:00	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
							OVERSPEEDING &												

Sheet1 Sheet2 Sheet3

Ready 50%

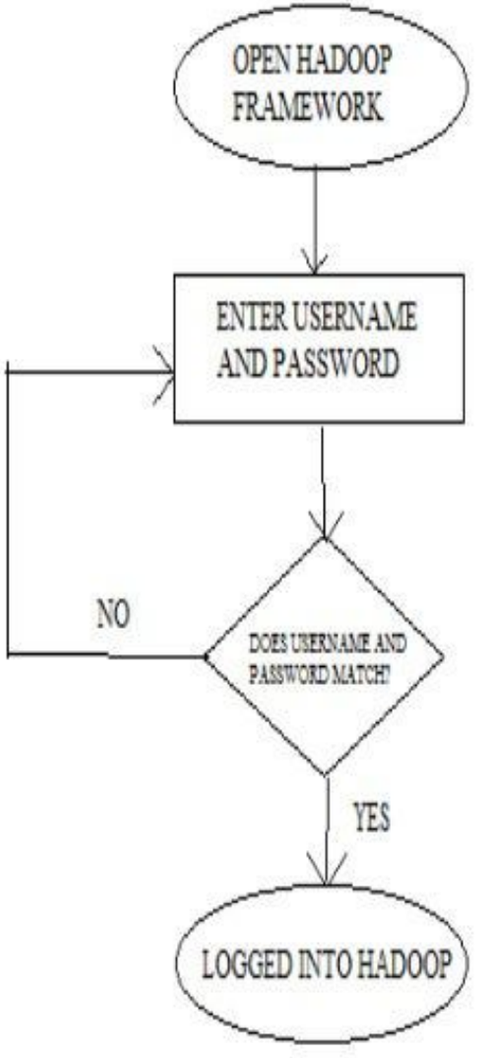
1:14 PM 4/25/2017

DESIGN AND IMPLEMENTATION



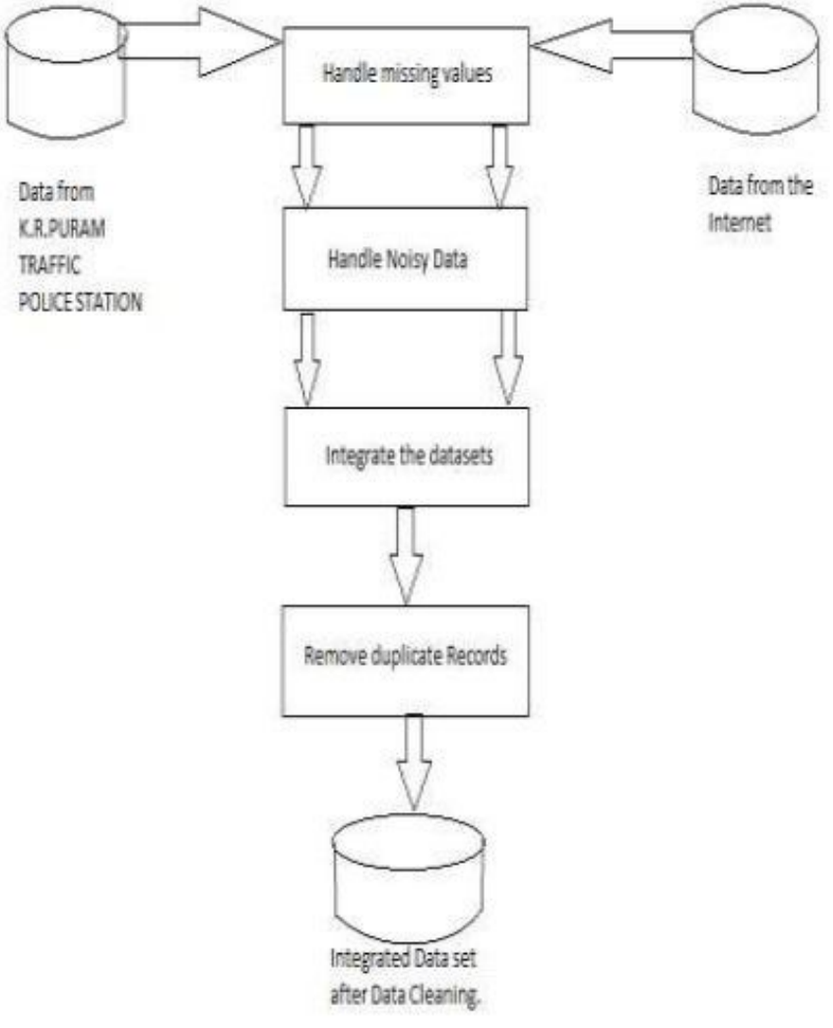
ARCHITECTURE OF THE SYSTEM

LOGIN MODULE



FLOWCHART OF LOGIN MODULE

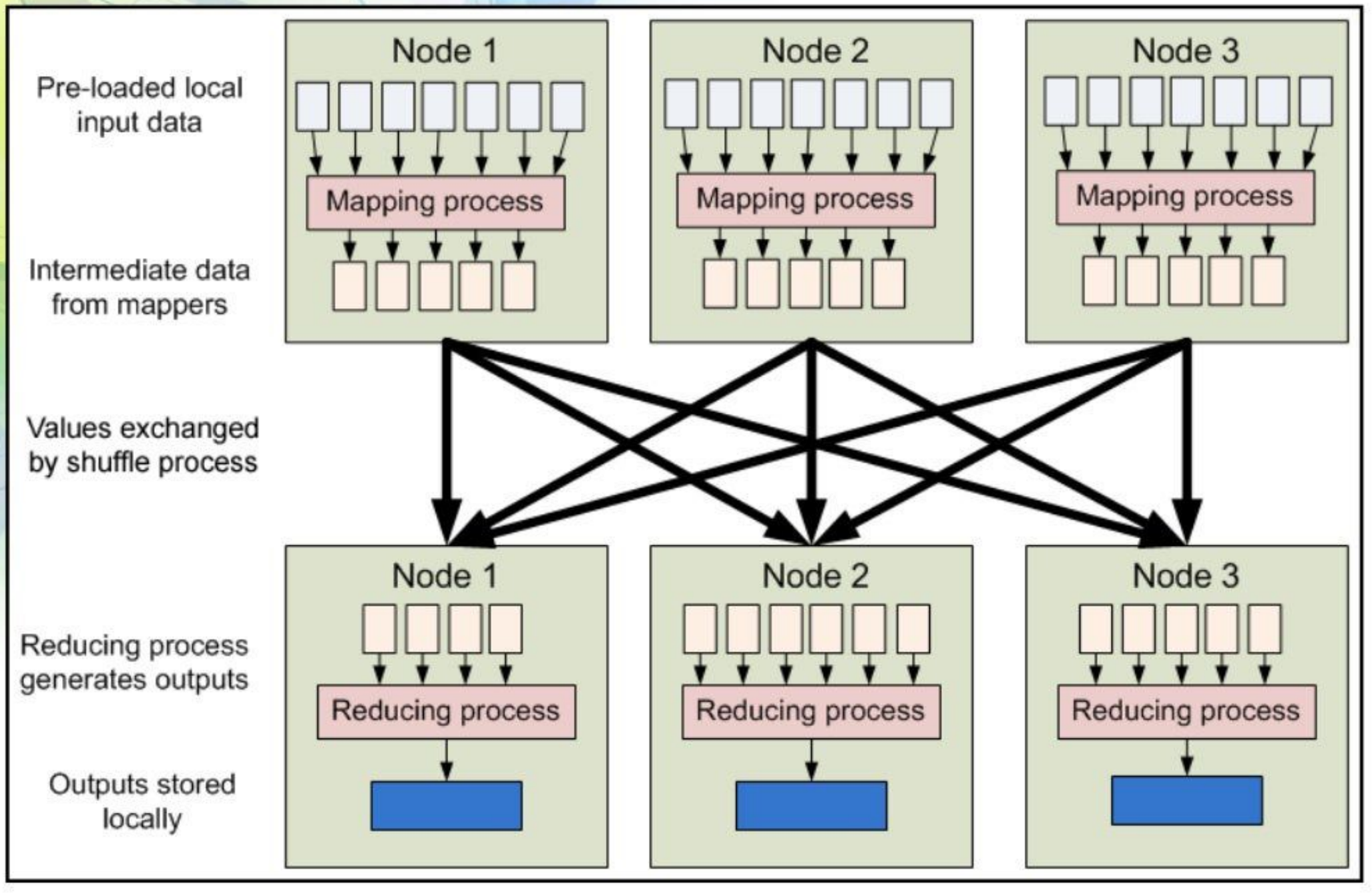
DATA PRE-PROCESSING MODULE



Pre-processing is a huge task in any data analysis

Some of the challenges faced in this project related to Data Pre-processing are:

- ***Handling missing values***
- ***Handling noisy data***
- ***Integrating the data from 2 different datasets***
- ***Removing duplicate records for multiple index entries.***



CLUSTERING USING MAPREDUCE

COMPILED RESULTS (EXPECTED)

- **Distribution of Accidents across Attributes**

- **Decision Tree Visualization**

- **Severity Vs User Defined Attribute**

CONCLUSION

- The tree generated is pruned to large extent due to memory restrictions and varied type of data.
- Further room for improvement exists by adding more clusters to the distributed processing module & using more user friendly visualizations.

References

1. https://hadoop.apache.org/docs/r0.18.3/

2. http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-nodecluster/

3. http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-nodecluster/

4. https://developer.yahoo.com/hadoop/tutorial/module4.html


THANK YOU