**ABSTRACT**

Considering the ever-growing enormous data with its wide variety and veracity, processing such data has become a challenge. The large volume of data and its complexity has forced us to improve the processing and analytical tools that handle such data. As a result, traditional methods are ineffective and horizontal scaling methods have been applied to process such clusters. Several frameworks have been developed from which Hadoop is the most commonly used one for Big Data analytics. Hadoop framework consists of many components such as HDFS, Apache Hive, Apache Pig, HBase, Zookeeper, Sqoop, Flume etc. that can be used to analyse and extract information from large amount of data. It is crucial to understand, pre-process and extract patterns since datasets can be semi-structured or unstructured clusters. In this project, we shall analyse the bank data and derive a behavioural pattern within customers to understand them better so that the banks can provide several schemes and strategies and identify the target customers. There is a two step procedure to analyse the data that involves Hadoop and Pig to handle the large data in the first step and using RStudio to perform K-means algorithm over the data and derive clusters that can be used for decision making.

**LIST OF FIGURES**

**FIGURE NO FIGURE NAME PAGE NO**

3.1 VMware workstation icon 15

3.2 VMware workstation platform 16

3.3 Cloudera Icon 16

3.4 Apache Hadoop Icon 17

3.5 Hadoop Architecture 18

3.6 Apache PIG Icon 19

3.7 Apache PIG Architecture 20

3.8 RStudio Icon 20

3.9 Eclipse Icon 21

3.10 Eclipse Platform 22

4.1 System Architecture 23

4.2 Data Flow Diagram 24

4.3 Pre Processing Data 25

4.4 Establishing Relationships

and Creating Tables 26

4.5 Extracting Data 27

4.6 Clustering Data 28

5.1 ASCII File 29

5.2 .csv File 30

5.3 Loading Data into HDFS 30

5.4 Inserting Fields 31

5.5 Forming Tuples 31

5.6 Flattening Tuples 32

5.7 Using UDFs 32

5.8 Input Files in HDFS 33

5.9 Customized Data in HDFS 33

5.10 Extracted Files 34

5.11 Boxplot of numerical data that uses transaction 35

5.12 Boxplot of numerical data that

does not use transaction 35

5.13 Plot of Transaction used data 36

5.14 Plot of Transactionless data 36

5.15 Plot of Average Salary and Transaction data 37

5.16 Plot of Average Salary and Loan amount 37

5.17 Cluster points of three clusters 38

5.18 Cluster points on transactionless data 38

5.19 Cluster plot of transaction based classification 39

5.20 Mean Cluster plot of transaction based

Classification 40

5.21 Cluster plot of transaction based classification 40

5.22 Mean cluster plot of transactionless classification 41

5.23 Pie Chart of Transactionless data 42

5.24 Pie Chart of Transactional data 42

**LIST OF TABLES**

**TABLE NO TABLE NAME PAGE NO**

6.1 Pre Processing Data 44

6.2 Establishing Relationships

and Creating Tables 45

6.3 Extracting Data 47

6.4 Clustering Data 48

6.5 System Testing 49

**LIST OF ABBREVIATIONS**

**ACRONYMS** **ABBREVIATIONS**

HDFS Hadoop Distributed File System

ASCII American Standard Code for International

Interchange

API Application Program Interface

AWS Amazon Web Services

CSV Comma Separated Values