FRAUD DETECTION ON BIG TAX DATA USING BUSINESS INTELLIGENCE, DATA MINING TOOL. A CASE OF ZAMBIA REVENUE AUTHORITY

By

Memorie Mwanza

A Dissertation submitted in partial fulfilment of the requirements for the degree of Masters of Engineering in Information Communication Technology, Security.

The University of Zambia

School of Engineering

January, 2017.

DECLARATION

I, the undersigned, declare that this has not previously been submitted in candidature for any degree. The dissertation is the result of my own work and investigations, except where otherwise stated. Other sources are acknowledged by given explicit references. A complete list of references is appended.

Signature: …….

Date: 3rd January, 2017

CERTIFICATE OF APPROVAL

This document by MEMORIE MWANZA is approved as fulfilling the requirements for the award of the degree of MASTERS OF ENGINEERING IN INFORMATION COMMUNICATION TECHNOLOGY, SECURITY of the University of Zambia

Examiner's Signature……..… Date: …….

Examiner's Signature: …… Date: …….

Examiner's Signature: …… Date: …….

ABSTRACT

Tax collecting in the developing countries has been associated with a lot of fraud which is a challenge to detect. This is because of the growth in size of data and also the absence of fully automated business processes. Zambia's tax administration is not an exception to such challenges. Zambia Revenue Authority houses huge sizes of data that need complex mechanisms in order to extract useful tax information. The purpose of the study was to establish the magnitude of the challenges in fraud detection on bulk tax data, to come up with a model which will be used to design a prototype for detection of fraud on tax data for ZRA and further to design the tool which will help to detect fraud on the bulk tax data. Our baseline study showed that currently ZRA uses traditional methods such as Targeted Audits, Random Audits, and whistle blowing to detect fraud on tax data. The baseline also showed that it takes long, above 7 days to detect anomalies or fraud on the bulk tax data. This method is tedious, time consuming and is prone to error. A model which implements data mining, outlier algorithms for fraud detection and is based on, Continuous Monitoring of Distance Based and Distance Based Outlier Queries was then designed.

Further, the prototype of ZRA Fraud detector was developed in java using weka Java libraries and NetBeans IDE which implements numerous data mining algorithms. The back end was implemented using MySQL and workbench 6.3 CE a unified visual tool for database architects and developers was used to interact with the Database

To implement the prototype, both algorithms were used, underpayments and overpayments according to business rules were detected and were marked as outliers. The results produced by our tool showed improvement in terms of speed of detecting fraud. It took 2 milliseconds to detect anomalies on 1000 bulk tax records as compared to the traditional method which takes above 7 days to detect anomalies on one record. The results of the fraud detection tool also showed the capability of clustering the tax payers into meaningful groups based on the business rules.

Keywords: Business Intelligence, Data mining, fraud detection, outlier algorithm**.**

DEDICATION

I dedicate this thesis to my God and His Only Beloved Son, for my good-health throughout this research study, and also for keeping my mind sound and focused during the long hours, days and nights, spent on the research work. Jehovah God you are the star to which am looking, and the rock on which I stand, forever be honored.

There are so many persons in my circle of connections whom I appreciate and treasure. It is a great privilege to be surrounded by such people who love and care about me. I also dedicate this work to my Late Dad Gilbert Kapole Mwanza and Mum Jane Mwanza for giving me a solid foundation, always my super hero, and to my husband Frazier Joe Makani, thank you for the shoulder to lean on. Brothers and Sisters, my children, Kraus, Pascal, Duncan, Mbaakani and Sibajane I will always appreciate you for being a tower of my strength.

To all my close friends and the entire family I greatly appreciate for being an inspiration.

ACKNOWLEDGEMENT

Firstly, I wish to thank the University of Zambia, School of Engineering, Department of Electrical and Electronics for the foundation leading to this Research work. To my Research Supervisor, Dr. Jackson Phiri, from the School of Natural Sciences in the Computer Science Department, I wish to thank him for guiding my research work all the way up to the finishing point. His advice and assistance has been instrumental on this research journey. I appreciate very much the valuable comments and the time spent reviewing the research work.

The support and cooperation of Zambia Revenue Authority particularly, Research and Planning team, Domestic Taxes team, Customs Services team and Information Technology team during the Research has been very significant for which I am thankful. To all of them I extend my warm and special thank you.

Lastly to all those who have contributed to my academic progress in various ways I say thank you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF KEYWORDS

Business Intelligence

Data Mining

Outlier Algorithm

Tax Administration

Fraud

Fraud detection

Data warehouse

Extract Transform Load

Continuous Monitoring of Distance based Algorithm

Distance based Outlier

LIST OF ABREVIATIONS

BI	Business Intelligence
CMDB	Continuous Monitoring of Distance-based Algorithm
DM	Data Mining
DC	Data Cube
DW	Data Warehouse
DBO	Distance Based Outlier Algorithm
ETL	Extraction Transformation Loading
IRS	Internal Revenue Service
LT	Large Taxpayer
MT	Medium Taxpayer
ST	Small Taxpayer
ZRA	Zambia Revenue Authority

CHAPTER ONE: INTRODUCTION TO THE RESEARCH

This chapter, provides an introduction for this research. Business Intelligence, Data mining tasks (Outlier Algorithm) and fraud were defined. The chapter also describes the research problem which leads to the need to establish the challenges Zambia Revenue Authority face regarding fraud detection on taxes.

1.1 Introduction

Fraud in the Tax administration sector has increased [1] extremely in the recent years. This presents a cost to the economy by reducing the revenue for the government [1] thereby affecting the running of programs. Despite technological advancement providing efficiency in conducting business, these improvements have also brought about an explosion in the amount of electronic data creating a challenge to detect fraud on tax data.

The term 'fraud' commonly includes activities such as theft, corruption, money laundering, bribery and extortion. The legal definition varies from country to country. Fraud essentially involves using deception to dishonestly make a personal gain. Whilst the Oxford English Dictionary [2] defines fraud as wrongful or criminal deception intended to result in financial or personal gain, in the academic literature fraud has been defined as leading to the abuse of a profit organization's system without necessarily leading to direct legal consequences. It has also been defined as obtaining something of value from someone through deceit. Currently, even though a universally accepted definition of Tax fraud is missing in the literature, America's government agency responsible for tax collection and tax law enforcement indicates that it is often defined as an International wrong doing on the part of a Taxpayer with the specific purpose of evading tax known or believed to be owing [3], or omits from a return any sum which should have been included, making any false statement or entry in a return and also making false statement in connection with a claim for any deduction or

allowance [4].

Although definitions vary, most are based around these general themes.

Fraud can mean many things and result from many varied relationships between offenders and victims such as [5];

i. Crimes by individuals or businesses against government, for example, tax evasion, grant fraud; social security benefit claim frauds;
ii. Employee fraud against employers, for example, payroll fraud; falsifying expense claims; thefts of cash, assets or intellectual property (IP);
iii. Crimes against financial institutions, for example, using lost and stolen credit cards; cheque frauds; fraudulent insurance claims.

This research study, focuses on fraud by individuals or businesses against the government, in particular tax evasion and tax avoidance for example, through underpayments, nonpayment and underreporting in Zambia Revenue Authority.

Although fraud is prevalent across organisations of all sizes and in all sectors and locations, research shows that certain business models will involve greater levels of fraud risk than others. By nature, Fraud is very difficult to detect but nevertheless, there has been a good deal of progress in terms of research and also implementation of tools, technologies and techniques that can be used to combat fraud. Worldwide, the Tax Administrators have different ways of detecting fraud or anomalies on the Tax Payers data. The strategies are based on efficiency, effectiveness and for others it is based on the extent of automation of the processes within the Tax Administration.

For this research study, Data mining, outlier algorithm which is based on distance Continuous monitoring has been considered. Outlier is an important task in many applications like fraud detection, plagiarism, computer network management, event detection (for example, in sensor networks), to name a few [6]. An object is considered an outlier, if it deviates from the “typical case” or from the normal range significantly. Figure 1 shows outliers in a dataset, the fact that most of the observations fall into clusters N1 and N2, they are considered two “normal” regions; while points in region O1 as well as points o2 and o3 are outliers (in red), due to their far distance from the “normal” regions [7]. An exact definition of an outlier normally depends on hidden assumptions concerning the data structure and the associated detection method, though some definitions are universal enough to cope with varieties of data and methods.

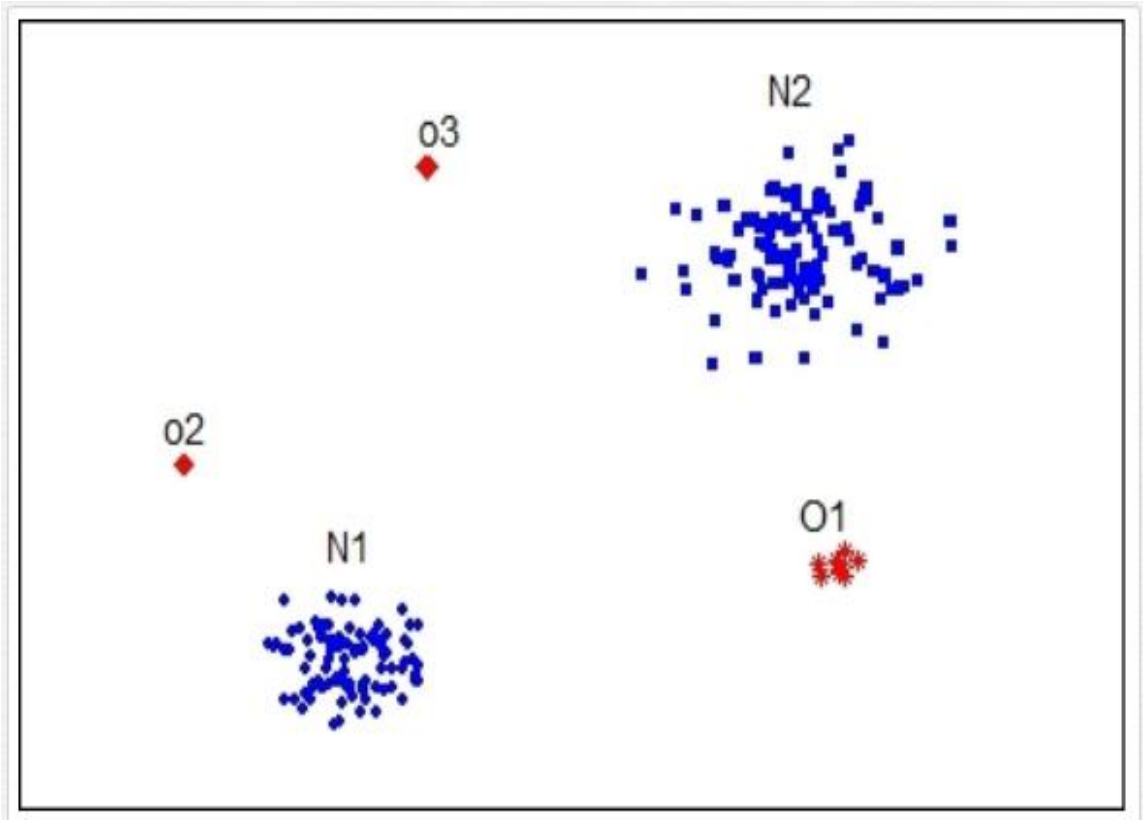

Figure 1: A pictorial view of Outlier Detection [7].

A typical example, Figure 1, is where outliers can be considered as patterns in data that do not conform to a distinct view of "normal" behavior, or as observations in a data set which appear to be inconsistent with the remainder of that set of data [8] [7]. Identifying observations inconsistent with the "normal" data, or detecting previously unobserved emergent patterns is commonly referred to as outlier detection or novelty detection. The distinction between novel patterns and outliers is that the novel patterns are often incorporated into the "normal" model after being detected whereas outliers are typically removed or corrected. These outliers are suspicious and possibly concern fraudulent transaction, and therefore may be flagged for further investigation.

Johnson [9] defines "an outlier as an observation in a data set which appears to be inconsistent with the remainder of that set of data". The process of outlier detection may be seen as the complement of clustering, in the sense that clustering tries to form groups of objects whereas outlier detection tries to spot objects that do not participate in a group.

The amount of business data that is generated has risen steadily every year [10]and more and more types of information are being stored in unstructured or semi structured formats. Traditional data mining has no power anymore to deal with the huge amount of unstructured and semi structured written materials based on natural languages.

According to Ranjan [11], it is the cumulative standards, automation, and technologies have led to enormous quantities of data becoming available in modern businesses. As the demands for Business Intelligence and also data mining continue to grow, data warehouse technologies have set up repositories to store this data. Improved Extract, Transform, Load (ETL) and even recently Enterprise Application Integration tools have increased the speedy collecting of data. OLAP reporting technologies have allowed faster generation of new reports which analyze the data. Business intelligence and data mining has now become the art of sifting through large amounts of data, extracting pertinent information, and turning that information into knowledge upon which actions can be taken such as detecting frauds on Tax information also defined as [12] a broad category of applications, such as the mixture of the gathering, cleaning and integrating data from various sources, and introducing results for the purpose of helping enterprise users make better business decisions.

1.2 Motivation and Significance of the Dissertation

A principal motivation behind this research study is the belief that Fraud in the financial sector which includes the tax administration sector, [13] is increasingly becoming a serious problem and as a result, this dishonest performance of taxpayers influences negatively the incomes available to public services as well as creating harm on the honest taxpayers. This has been fueled by the changes in the business dimension, the vast amounts of data accumulated in the tax administrations and that the data sources or repositories are currently underutilised. Therefore using Business Intelligence and data mining could provide powerful techniques for tax administrations to discover useful knowledge in support of their compliance enhancing agendas and Enhanced fraud detection requirement.

The data mining model will provide better Control measures on Tax frauds, underpayments and underreporting. It will also help detect any anomalies in the Tax Data. ZRA will therefore be better informed for the purpose of making sound decisions.

1.3 Scope

The scope of this research is designing and developing a data mining model as a business intelligence solution for fraud detection on bulky tax data in ZRA. This research will focus on data from domestic taxes only.

1.4 Problem Statement

There is a problem in all Tax Administration sectors. All governments, regardless of whether they are large or small, public or private, local or multinational have not been spared or freed from Fraud in its various manifestations.

Despite putting up various Technologies, techniques, strategies to fight fraud such as Planned and Targeted Audits, Random Audits, whistle blowing, Tax Fraud and evasion have continued to be a challenge because it remains a limiting factor to the capacity of the government in raising revenues to carry out their economic policies. Just like other fields and industries, the Zambian Tax Administration sector has not been spared by this phenomenon.

Zambia Revenue Authority is a quasi-governmental organization with a mandate to collect revenue on behalf of the Government of the Republic of Zambia. Some of the main responsibilities of the Authority include:

i. To properly assess and collect taxes and duties at the right time
ii. To ensure that all monies collected are properly accounted for and banked
iii. To provide statistical information on revenue to the Government

In order for ZRA to meet its objective, it has automated several processes including, eRegistration, eFilling, ePayments which generates enormous Business data commonly known as Taxes and Taxpayer Information. This Information sits across different data silos,

Knowledge discovery from different Data Silos

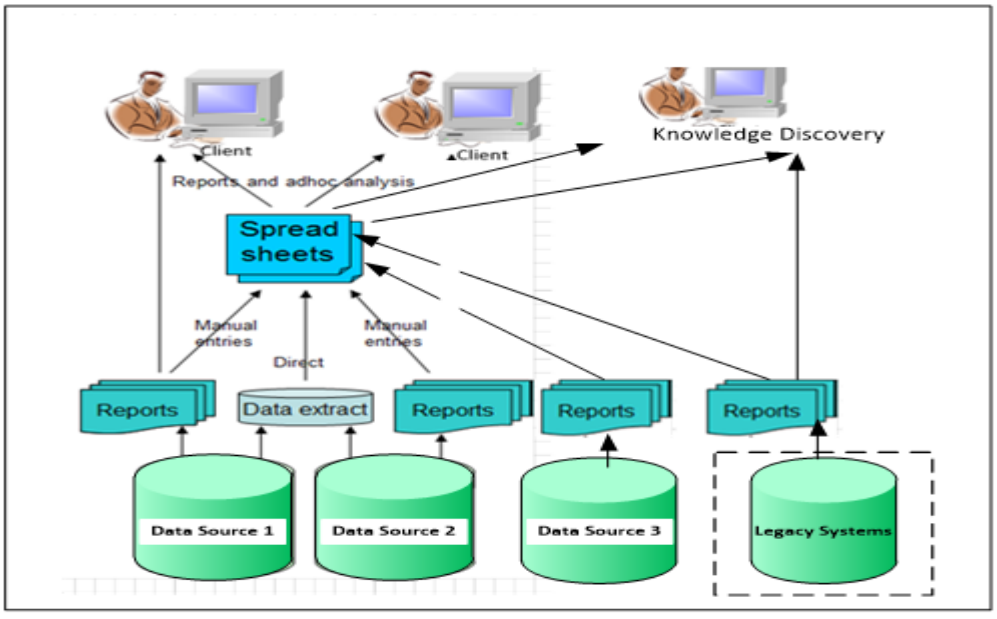

Figure 2: Current setup of Information Discovery. [14]

In this day and age, with the growth of technology which is supported by the comprehensive and speedy communication network, it is easier and faster to do fraud which can even be much more complex to be detected.

These fraudulent activities lead to larges revenue losses worldwide for the States. For this reason it is extremely important at the present in time of economic crises to detect and combat fraud in order to possibly recover the revenue for the government and its people.

One way to effectively and efficiently combat fraud is by using Business Intelligence techniques and applying Data Mining, which is the exploratory phase of gaining knowledge in databases process. It makes use of a set of techniques that help to reach at vital information that may lead to fraud data. One of these techniques is the construction of a classification model, which attempts to identify indicative patterns of fraud in different areas of tax data. Business Intelligence as defined by Ventana Research is the process of integrating data and applying analytics to yield meaningful information that can guide actions and inform decisions to accomplish a broad range of business objectives.

Applying data mining to fraud detection will also help come up with a model which will be able to identify those records for the tax payer that seem to be inappropriate which will then be tagged as being fraudulent or not because of their deviation from the normal distribution of the dataset. These records are finally known as outliers and can, in fact, be indicators of potential fraudulent activities.

In conclusion, the use of Business Intelligence, Data Mining techniques in order to detect unusual or peculiar characteristics or patterns for this research is very cardinal for the Tax Administration system.

1.5 Aim of the study

The aim of the study was to establish the challenges Zambia Revenue Authority face regarding fraud detection on Taxes and later design a data mining model which will be able to detect fraud on taxes based on taxpayer data. This study focused on two areas;

i. The baseline study that helped to establish the extent of the challenges in fraud detection for the tax payers in Zambia Revenue Authority
ii. The automation and development of the fraud detection tool using the results from the baseline study.

1.6 Objectives

i. To investigate the challenges Zambia Revenue Authority face regarding fraud detection on Taxes.
ii. To design a Business Intelligence data mining model for detecting Fraud on Bulk Tax Data.
iii. To develop a prototype based on the model in (2) that will detect Fraud on Tax Data.

1.7 Research Questions

This research will be guided by the following research questions;

i. What challenges is ZRA currently facing with regard to fraud detection on Taxes?
ii. How can we design a data mining model as a Business Intelligence tool to detect Fraud on Tax Data to enable ZRA meet its business objective?
iii. How can we develop a prototype based on the model in (2) that will detect Fraud on Tax Data to enable ZRA meet its business objectives?

1.8 Research Contributions

In this study, two major contributions were made. The first contribution is the baseline study that showed that the tax administration in Zambia (ZRA) has challenges in connection with detection of fraud.

Our second contribution is the algorithm used in implementing the prototype that was used to detect fraud in tax payers as outliers.

1.9 Organization of the Thesis

The thesis is organised into five chapters. Chapter 1 is the Introduction to the Research. In this chapter, we give a brief overview of Business Intelligence, data Mining and Fraud detection. The problem statement is also outlined. The aims and motivation of this thesis is provided. This chapter concludes by the giving an outline of the thesis.

Chapter 2 looks at the review of literature on Business Intelligence, Data mining and Fraud detection different sectors and related works. In this chapter, we begin by providing a comprehensive review and the background of Business Intelligence and Data Mining as well Fraud detection.

The methodology for designing and developing the ZRA Fraud detection tool is

highlighted in detail in Chapter 3. This chapter, begins by looking at the baseline study and the study design for the first objective of this research. Later the methodology used to come up with the design and the architecture for the model is highlighted.

Chapter 4 gives the results of the research for both the baseline study and also the implementation results of the model. Finally an in-depth discussion of the results of the research in comparison to the current situation.

1.10 Summary

This chapter provided the basic introduction of the work in this dissertation which is the business intelligence and data mining (Outlier Algorithm. The definition of Fraud from a general perspective and according to academic literature and also a definition according to the Tax Administration sector was given.

The motivation, significance and scope of the work in this study are then outlined. Finally the problem was stated in the problem statement necessitating this study. The aims and the research contributions were also outlined. The chapter was closed with the outline of the dissertation.

CHAPTER TWO:

LITERATURE REVIEW

2.1 Introduction

This chapter firstly highlights the concerns about fraud in the tax administration sector and discusses methodological and technological characteristics about business Intelligence and Data mining in relation to fraud detection as a solution for this problem. Further a review of related works in areas of application such as the banking Sector, Insurance, the Health, Education, and in the Telecommunication sector.

2.2 Review of the Literature

Although there's been some technological advancements in the way fraud is detected in the Tax Administration sector, particularly in Zambia Revenue Authority, the capacity and performance of such methods in reporting and extraction of knowledge and patterns from databases shows that the level of adoption of Business intelligence and Data Mining to discover knowledge in the tax administration sector is still very key.

According to Gonzalez C., [15] tax fraud and tax evasion have been a constant concern for tax administrations, especially when pertaining to developing countries . Data mining, according to Silwattananusarn, T., & Tuamsuk, K [16], also known as kowledge discovery is a key phase in the process of discovering interesting patterns in databases that are useful in decision making.

Data mining is a discipline of growing interest and importance, and an application area that can provide significant competitive advantage to an organisation such as the tax administartion by exploiting the potential of large data warehouses.The task of finding patterns such as underpayments, underreporting, outliers and generally the behaviour of tax payers are not new. Traditionally it was the responsibility of Business analyst who generally used stastical techniques to uncover such patterns. The scope of such an activity has however changed because of the advancement in technologies which has created large electronic databases which store huge business Data. Transactipons can be analysed to identify tax payment patterns of individual taxpayers

as well as a group of tax payers.

Data mining involves integration of techniques from multiple disciplines such as database and data warehousing technology, statistics, machine learning, pattern recognition, artificial neural networks, data visualization, information retrieval, image and signal processing [17].

Through the whole process of Knowledge discovery, organisations can unlock the wealth contained its data thus justifying the cost of development and implementation of expensive automated systems that may just dump data to a repository or database.

Business Intelligence systems are developed and designed to help the organizations understand their customers, financial situation, operations, performance and trends. The amount of business data that is generated has risen steadily every year, echoes Manicka R. et al [18] and more and more types of information are being stored in unstructured or semi structured formats. Traditional data mining has no power anymore to deal with the huge amount of unstructured and semi structured written materials based on natural languages.

According to Nandakumar A., [19] in the era where organizations are rich in data, the true value lies in the ability to collect this data, sort and analyze it such that it derives actionable business intelligence (BI). To analyze the data, traditional data mining algorithms like clustering, classification form the basis for machine learning activities in the business intelligence support tools. As business intelligence technologies provide historical, current and predictive views of business operations [20], It has been proved today according to Aziz A., that enterprises [21] rely on a set of automated tools for knowledge discovery to gain business insight and intelligence.

According to Singh H., [22] business intelligence has continued to gain popularity in large and medium-sized organizations. The hasty evolution of structured and unstructured data and the intent to apply the data for redefining decision-making and operational processes has shaped the need to harness the big data. Moreover, there is a growing demand for data utilization to discover insights and predict future outcomes.

Business Intelligence systems are the ones, they are developed and designed to help the organizations understand their customers, financial situation, operations, performance and trends. BI systems act as business measurements units. Poor intelligence leads to poor decision making. Business intelligence raises issues like data quality, which an organization has to deal with, in order to improve its decision making.

In the present day, and according to Ramanigopal C., [23] it is important to make sound

business decisions based on complete data. With the proper Business Intelligent implementation, businesses can make decisions and feel comfortable that they are provided with the proper tools and data needed to believe in their decisions. Without the correct business intelligence solution, even well planned and executed data warehouse architectures can fail. Business Intelligence can provide professionals with the information they need to make the most effective decisions for their organization. It also provides data about the organizations details, such as customers, products, and services.

In my opinion, business intelligence coupled with data warehousing, Data mining and its technologies acts like "***a one stop shop solution***", such that, if implemented in an organization, it brings with it advantages such as;

Reduced information bottlenecks because of the relevant information at hand. Better decisions made every day based on the rich, exact and up-to-date information. BI systems enable fast decisions. Alignment of the organization towards its business objectives.

All this enables organisations to understand their customers, financial situation, operations, performance and trends.

2.2.1 Data Mining Models

A Model is defined by several authors differently, Cambridge Dictionary [24] defines it as a simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions, whilst Stachowiak H., [25] describes the fundamental properties that makes a model as one that needs to possess three features such as based on an original, should reflects a relevant selection of the original's properties and lastly needs to be usable in place of the original with regards to some purpose. Oxford dictionary [26] defines it as a three-dimensional representation of a person or thing or of a proposed structure, typically on a smaller scale than the original: In the context of this study, the definition of the model will be based on Stachowiak's [25].

Literature review indicates that there are several existing BI models [27]. Various authors try to give an over-all structure of a BI model. These models are different in their structures such as layers, components, processes, and relationships. Though different in structures, there are some common components among these BI models identified by a number of authors which consists of four different layers such as [28]

[29] [30]; Data source(s), Integration services, Data repositories, Analytical facilities. There are other models described, [27] for example Spruijt R., distinguish a BI framework consisting of three layers:

i. **Access layer** - for bringing all appropriate components and functions from the logic layer together and to present them to the user in an integrated and personalized fashion.
ii. **Logic layer** – focuses on the compilation, processing and distribution of management support data. Two types of systems are distinguished at this layer: systems for data analysis and components for knowledge distribution.
iii. **Data layer** – containing the data for the analysis, often a data warehouse.

Below these three layers are the operational systems that deliver the data to the data layer.

A good BI model therefore, will take into consideration the layers such as data sources which are a most obvious with most BI models, it is also necessary to consider data staging area in the ETL (Extract- Transform-Load) process because most data from data source require cleansing and transformation data warehouse, as well as analysis reporting and data mining.

Data mining refers to extracting or "mining" knowledge from large amount of data [31]. It is the fetching of hidden and necessary information from the data and the knowledge discovered by data mining is previously unknown, potentially useful, and valid and of high quality [32]. Data mining is used for wide variety of purposes both in private and in public sectors. Its applications are successfully implemented in various fields such as Fraud detection, health care, finance, retail, telecommunication, risk analysis, education [33] [34] [35] [36].

Data mining carries a collection of a multitude of disciplines, such as database systems, statistics, artificial intelligence, data visualization, and others. As alluded to by Kumara P., et al [37], the knowledge discovered can be applied to Information Management, Query Processing, Decision-Making, Process Control and many other applications. According to Gonzales M., [38] Data mining is about choosing the right tools for the job and then using them skillfully to discover the information and issues in the data where there is an identified problem to which an answer is needed or where it is suspected, or known that the answer is buried somewhere in the data but are not sure where exactly.

According to Wua R. et al [39], data mining is a process to discover uncertain,

unknown, and hidden information from the existing data. He furthers defines data mining as a unique method of finding new facts and relationships in the existing data that have not as yet been discovered by experts. Wua R. et al [39], regards data mining as an analysis method using automatic or semi-automatic tools to discover the meaningful relationship or rules from a huge amount of data.

Data mining requires algorithms or methods to analyze the data of interest, currently there are several existing algorithms for data mining that can be used for anomaly detection [40] [41] but only a few are applicable where a vast amount of information has to be processed.

2.2.2 Components in BI and Data mining.

BI system and components are grouped to compose a completely integrated system that works as a backbone to allow organization sustain and to survive with correct response to environmental pressures. They play a major role in decision making process. Each component, Figure 3, has a direct impact on decision making in a varied technique [42] [43].

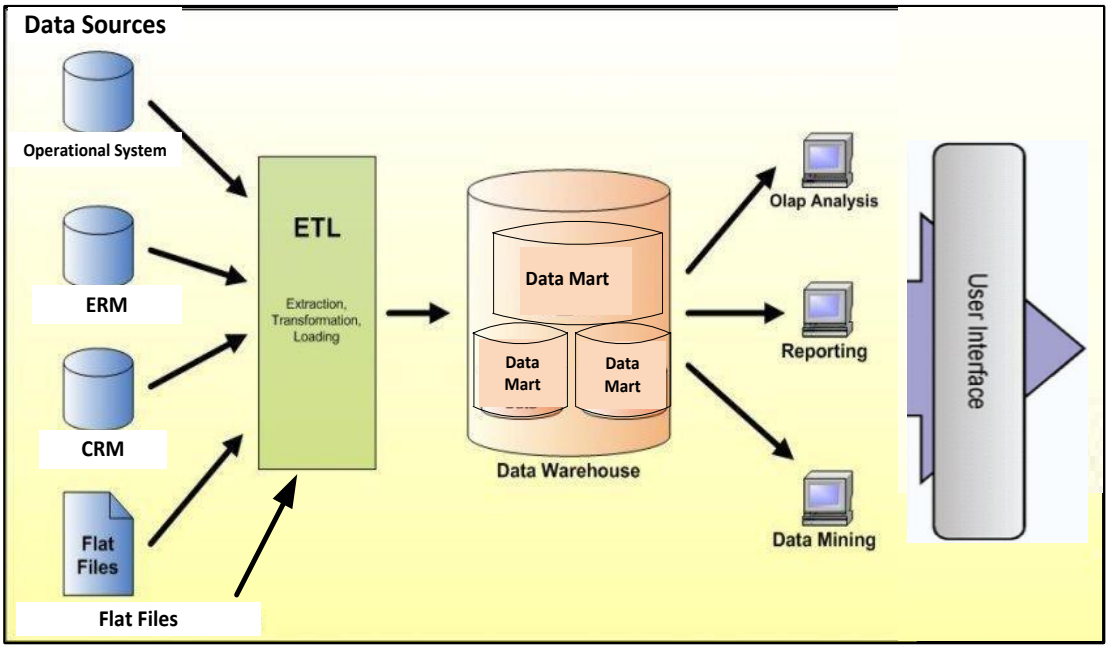

Figure 3: Components of a BI System [28]

i. Data Sources

It is important for organizations to clearly identify their data sources states Lih Ong [44]. For tax administrators including Zambia Revenue Authority, knowing where the required data can be obtained is useful in addressing specific business questions and requirements, thereby resulting in significant time savings and greater speed of information delivery. Furthermore, the knowledge [45] can also be used to

facilitate data replication, data cleansing, and data extraction. Even though there are many existing data sources, some of them might be inaccessible, unreliable or irrelevant to current business needs. With correct identification of data sources, as well as the business requirements and business needs, problems such as inconsistent information, difficulty in finding root causes, and pattern recognition on data can easily be handled.

Data sources can be operational databases, historical data, and external data for example, from the taxpayers, or information from the already existing data warehouse environment. The data sources can be relational databases or any other data structure that supports the line of business applications. They also can reside on many different platforms and can contain structured information, such as tables or spreadsheets, or unstructured information, such as plaintext files or pictures and other multimedia information.

ii. Process of extracting data and loading into Data warehouse

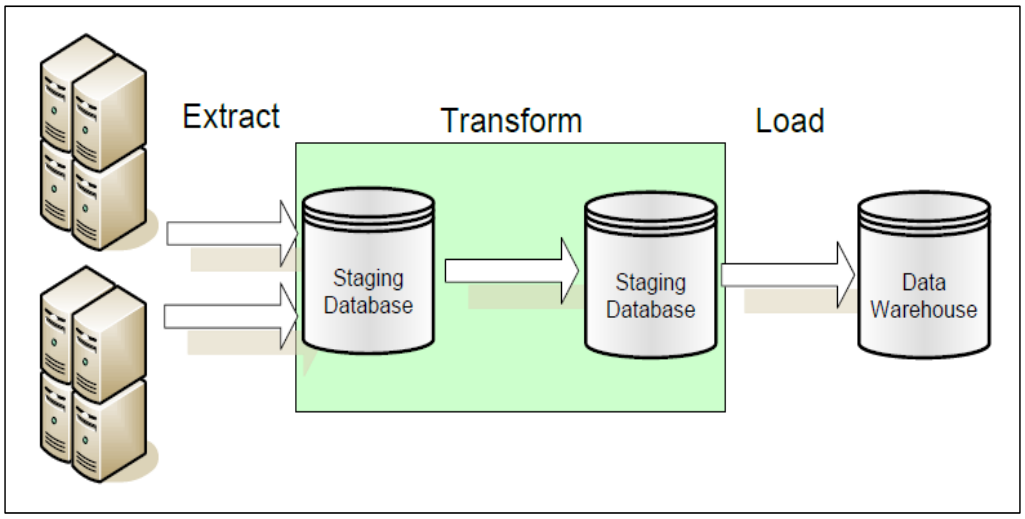

Figure 4: ETL Process [46]

Data is “extracted” from the existing data sources based on the business applications using a data extraction tools. The extracted data is then transformed using a series of transformation routines. The data format of the output will largely dictate the transformation process. Data quality and integrity checking is performed as part of the transformation process, and corrective actions are built into the process. Transformations and integrity checking are performed in the data staging area. Finally, once the data is in the target format, it is then loaded into the data warehouse ready for presentation [38].

iii. Data warehousing

Talwar K and Gosain A, [47] defines data warehouse as a systematic data storage system, and is used to support the resolution of merging the heterogeneous data, it is considered as [48], a set of combined historical data that help in decision-making. Also a technical base that operate on the bases of quick and flexible responses to the activities of business. Data warehouse also provides several benefits for the organization, some of which are, Data Integration, because of the multiplicity of sources. Data analysis to make it easy to understand, which helps in the decision-making. Reduce the cost through access to historical data.

The foundation of the data warehouse is to attain integration among heterogeneous information in diverse databases, systems transaction processing and Legacy System, as well as external data sources that are relevant to the subject [49].

iv. OLAP (On-line analytical processing)

OLAP is utilized to extract knowledge from the Data warehouse. According to Sharma P., et al, [50] it refers to the way in which business users can slice and dice their way through data using sophisticated tools that allow for the navigation of dimensions such as time or hierarchies. Online Analytical Processing or OLAP provides multidimensional, summarized views of business data and is used for reporting, analysis, modeling and planning for optimizing the business. OLAP techniques and tools can be used to work with data warehouses or data marts designed for sophisticated enterprise intelligence systems. These systems process queries required to discover trends and analyze critical factors. It also creates models that make life easier to users to categorize problem in simple forms and shows all variables affect the problem model [51]

v. Advanced Analytics

Hurwitz & Associates [52]describes advanced analytics as providing algorithms for complex analysis of both structured and unstructured data. It includes sophisticated statistical models, machine learning, neural networks, text analytics, and other advanced data mining techniques. Some of the specific statistical techniques used in advanced analytics include decision tree analysis, linear and logistic regression analysis, social network analysis, and time series.

Advanced analytics takes advantage of statistical analysis techniques to predict or provide certainty measures on facts and it cares more about the future prediction which is based on statistics processed and delivered OLAP [53]. Companies use

advanced analytics to discover patterns and anomalies in large volumes of data, and then use this insight to predict the outcomes of future events and interactions. In addition, advanced analytics is used for optimization and complex event processing and analysis. With advanced analytics, your organization can adjust its plans and strategies to become more competitive, minimize potential risk and optimize decision-making in real time.

vi. Data Warehouse.

The data warehouse is the significant component of business intelligence. According to Wua R. S. [39] more and more companies are using BI tools to analyze sales and other related transactional data to detect fraud. He further states [39], that software companies, such as SAS Institute Inc., SPSS Inc., NCR's Teradata, and IBM's Cognos Business Intelligence all of them provide fraud-detection oriented BI tools. Some US government agencies are also employing BI tools to detect tax evasion. According to Lisa McCormack [39], a manager in the audit division in Austin, the controller's office in Texas relies on a data warehouse tool to check for sales tax compliance.

A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile dataset. It is utilized to collect data from different sources that can then be organized into semantic and integrated data storage. Data warehousing is used to support structured and specific queries. The results of queries are mostly analytical reports that can be employed to support decision making. Data warehouse supports the physical propagation of data by handling the numerous enterprise records for integration, cleansing, aggregation and query tasks. It can also contain the operational data which can be defined as an updateable set of integrated data used for enterprise wide tactical decision-making of a particular subject area.

Palak Gupta et al [54] also defines Data Warehouse as a repository of business or enterprise databases which gives a picture of historical and current organization's operations. It provides [55] a historical, integrated view of an organization's operations.

In any type of organization, information is one of the greatest central assets. This asset is almost always well-maintained by an organization in dual forms, that is, the operational systems of record and the data warehouse. Roughly, the operational systems are where the data is placed in, and the data warehouse is where we get the data out.

A good BI architecture should include the layers such as Data sources which are a most obvious with most BI Models, it is also necessary to consider data staging area in the ETL (Extract- Transform-Load) process because most data from data source require cleansing and transformation data warehouse, as well as Analysis, reporting, visualization and Data mining. In any type of organization, information is one of the greatest central assets. This asset is almost always well-maintained by an organization in dual forms, that is, the operational systems of record and the data warehouse. Roughly, the operational systems are where the data is placed in, and the data warehouse is where we get the data out.

2.2.3 Data mining process

Knowledge discovery is iterative process as one may need to go back and forth in order to obtain the required knowledge from data. The process of knowledge discovery is described pictorially and it indicates that data mining is only a section in the whole process [56].

CRISP-DM (CRoss Industry Standard Process for Data Mining) [57] [58] [59] [60] proposed a complete process model for carrying out data mining projects. This process model is independent of both the industry sector and the technology used. It is one of the most widespread and broadly adopted model for knowledge discovery. It has already been acknowledged and relatively extensively used particularly in the fields of research as well as industrial. The main purpose of CRISP-DM is to deliver an efficient process that can be used to produce knowledge from the vast data repositories.

The value of such vast amounts of data needs to be unlocked by the use of data mining in order to utilise its potential since it is not enough just to store data without using it to learn from previous occurrences. The Data mining phase can be divided into the following;

i. Business understanding

The first phase ensures that all participants understand the project goals from a business

Or organizational perspective. These business goals are then incorporated in a data mining problem definition and detailed project plan. For a tax auditing agency, this would involve understanding the audit management process, the role and functions performed, the information that is collected and managed, and the specific challenges to improving audit efficiency. This information would be incorporated

into the data mining problem definition and project plan.

CRoss Industry Standard Process for Data Mining (CRISP-DM)

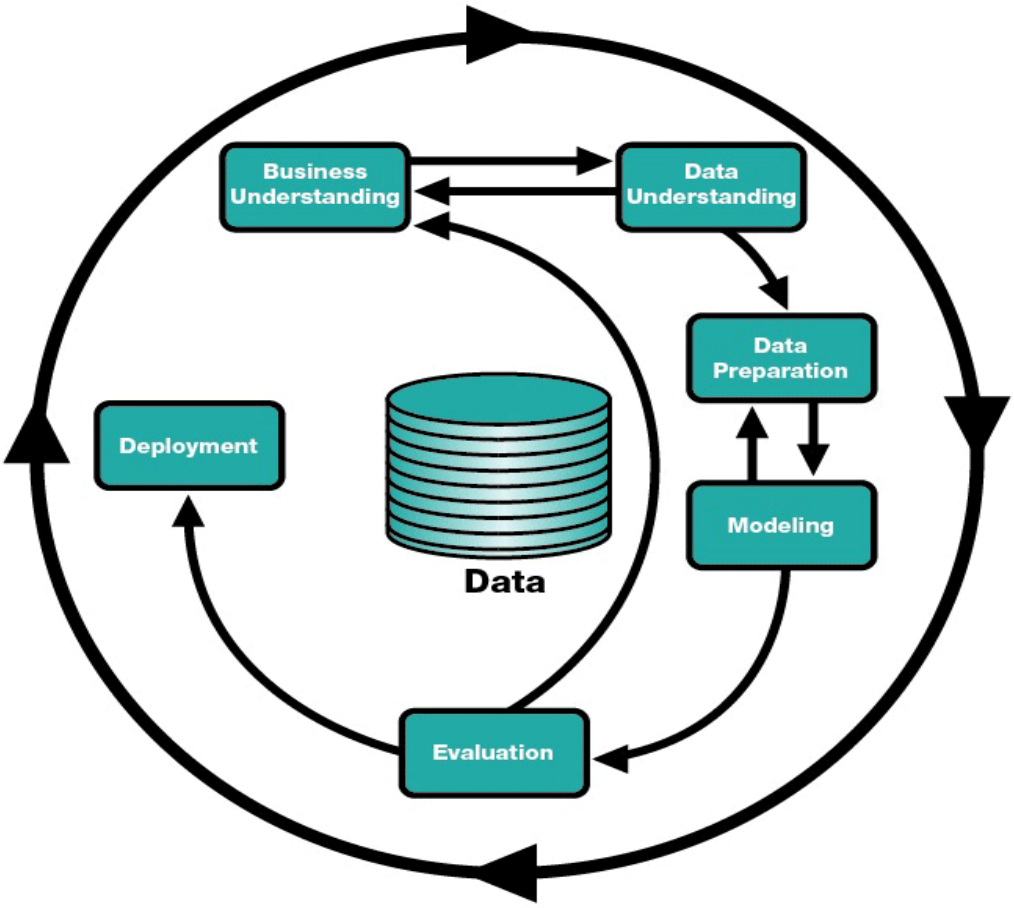

Figure 5: Generic process of Knowledge Discovery [57] [58] [59] [60]

ii. Data understanding

The second phase is designed to assess the sources, quality and characteristics of the data. This initial exploration can also provide insights that help to focus the project. The result is a detailed understanding of the key data elements that will be used to build models. This phase can be time-consuming for tax agencies that have many data sources, but it is critically important to the project.

iii. Data preparation

The next phase involves placing the data in a format suitable for building models. The analyst uses the business objectives determined in the business understanding step to determine which data types and data mining algorithms to use. This phase also resolves data issues uncovered in the data understanding phase, such as missing data.

It must be noted that in reality, though stored on the database, there are many types

of problems on bulk Tax data that affect business and its processes. In general these can be divided into two groups; defective data and inconsistent data [22] [20]. For all organisations it is necessary to take the efforts to single out problems on data in order to address them before the need arises.

a) Defective Data

There are many types of data defects. Inaccurate, incomplete, unavailable or obsolete data is defined as defective data. For example, manually entered data is frequently full of spelling errors, typographic errors. At times, people just neglect to fill in all the fields or enter correct data in the wrong fields. Data defects can also appear when a system is moved from one platform to another, an old application is replaced with a new one, or data is moved from one application to another on an automated and scheduled basis. Or indeed a need arise to integrate data from several different sources, a scenario such as the one being addresses by this study [22] [20].

The problem with defective data is that it is difficult to find out once it enters into the system. The best way to prevent defective data is to prevent it from being entered in the first place, so for this the businesses should invest in systems that validate and fix data at the source when it enters into the system or moved between systems through an application interface. To fix defective data already in the system, the business must invest in data profiling and cleansing tools to cleanse and validate data sets before they are loaded into data warehouses for further use in BI [22] [20].

b) Inconsistent Data

Inconsistency is another problem that can be found on data. This happens when data occurs over time as the data is either duplicates or out dated. For example, customer data degrades over time as people marry, divorce, die, move, or change their names, so, this results in inconsistent data [22] [20].

iv. Modeling

The modeling phase involves building the data mining algorithms that discover interesting patterns from the data. There are a variety of data mining techniques; each is suitable for discovering a specific type of knowledge. A tax agency would use classification or regression models, for example, to discover the characteristics of more productive tax audits. Each technique requires specific types of data, which may require a return to the data preparation phase. The modeling phase

produces a model or a set of models that contains the discovered knowledge in an appropriate format.

v. Evaluation

This phase focuses on evaluating the quality of the model or models. Data mining algorithms can uncover an unlimited number of patterns; many of these, however, may be meaningless. This phase helps determine which models are useful in terms of achieving the project's business objectives. In the context of audit selection, a predictive model for audit outcome would be assessed against a benchmark set of historical audits for which the outcome is known.

vi. Deployment

In the deployment phase, the organization incorporates the data mining results into the day-to-day decision making process. Depending on the significance of the results, this may require only minor modifications, or it may necessitate a major reengineering of processes and decision support systems. The deployment phase also involves creating a repeatable process for model enhancements or recalibrations. Tax laws, for example, are likely to change over time. Analysts need a standard process for updating the models accordingly and deploying new results. The appropriate presentation of results ensures that decision makers actually use the information. This can be as simple as creating a report or as complex as implementing a repeatable data mining process across the enterprise. It is important that project managers understand from the beginning what actions they will need to take in order to make use of the final models.

The six phases described are integral to every data mining project. Though each phase is important, the sequence is not rigid; certain projects may require you to move back and forth between phases. The next phase or the next task in a phase depends on the outcome of each of the previous phases. The inner arrows indicate the most important and frequent dependencies between phases. The outer circle symbolizes the cyclical nature of data mining projects, namely that lessons learned during a data mining project and after deployment can trigger new, more focused business questions. Subsequent data mining projects, therefore, benefit from experience gained in previous ones.

To create the models, the analyst typically uses a collection of techniques and tools. Data mining techniques come from a variety of disciplines, including machine learning, statistical analysis, pattern recognition, signal processing, evolutionary

computation and pattern visualisation.

2.2.4 Data Mining Techniques used to Detect Fraud Patterns.

This study reviewed some of the available data mining techniques that can handle different classes of problems including Fraud detection. The classes are presented as;

i. **Classification** - Classification builds up (from the training set) and utilizes a model (on the target set) to predict the categorical labels of unknown objects to distinguish between objects of different classes. These categorical labels are predefined, discrete and unordered [61].

   The research literature describes that classification or prediction is the process of identifying a set of common features (patterns), and proposing models that describe and distinguish data classes or concepts. Common classification techniques include neural networks, the Naïve Bayes technique, decision trees and support vector machines. Such classification tasks are used in the detection of credit card, healthcare and automobile insurance, and corporate fraud, among other types of fraud, and classification is one of the most common learning models in the application of data mining in fraud detection.

ii. **Clustering** - Clustering is used to partition objects into previously unknown conceptually meaningful groups (i.e. clusters), with the objects in a cluster being similar to one another but very dissimilar to the objects in other clusters. Clustering is also known as data segmentation or partitioning and is regarded as a variant of unsupervised classification [61]. Cluster analysis decomposes or partitions a data set (single or multivariate) into dissimilar groups so that the data points in one group are similar to each other and are as different as possible from the data points in other groups [61]. It is suggested that data objects in each cluster should have high intra-cluster similarity within the same cluster but should have low inter-cluster similarity to those in other clusters [61]. The most common clustering techniques are the K-nearest neighbour, the Naïve Bayes technique and self-organizing maps.

iii. **Prediction** - Prediction estimates numeric and ordered future values based on the patterns of a data set [61]. It is noted that, for prediction, the attribute, for which the value being predicted is continuous-valued (ordered) rather than categorical (discrete-valued and unordered). This attribute is referred as the predicted attribute [61]. Neural networks and logistic model prediction are the most commonly used

prediction techniques.

2.2.5 Outlier Algorithms in Fraud Detection.

Today, most of the work existing on outlier detection lies in the arena of statistics. In statistics, one can find over 100 outlier detection techniques [62] [63]. These have been developed for different data distributions, parameters, desired number of outliers and type of expected outliers. However, all of those tests suffer from the following two serious problems. First, almost of them are univariate (i.e., single attribute). This restriction makes them unsuitable for multidimensional datasets. Second, all of them are distribution-based.

Outlier detection is employed to measure the distance between data objects to detect those objects that are grossly different from or inconsistent with the remaining data set [61] [62] [63]. Data that appear to have different characteristics than the rest of the population are called outliers [61]. The problem of outlier/anomaly detection is one of the most fundamental issues in data mining.

For the purpose of overcoming these problems, several outlier detection techniques have been proposed in the data mining including nearest-neighbour based, density based, Clustering-based and distance-based and the discounting learning algorithm [61]. In this dissertation, our focus is the distance-based approach because the characteristics involved such as the speed, Accuracy, Expandability, and also because it has been most commonly used. Further, it can be used as pre-processing before applying more sophisticated application dependent outlier detection techniques.

2.3 Related Works

Various studies have demonstrated successful use of outlier detection and this is also shown through the studies broadly conducted by the statistics community [62] [63], where the objects are modelled as a distribution, and objects are marked as outliers depending on their deviation from this distribution. The problem of fraud detection using outlier detection has also been addressed by the database and data mining communities [62].

In the Business Intelligence community, different scholars have defined outliers differently. Lekhi N., [31] defines an outlier as an observation that deviates from other observations as to arouse suspicion that it was generated by a different mechanism. Whilst Aggarwa C. C., defines outlier [64] as a data point which is significantly

different from the remaining data, which is also referred to as abnormalities, discordant, deviants, or anomalies in the data mining and statistics literature.

Business Intelligence and data mining for the purpose of outlier detection has been applied in different areas with specific problems for each area of the fraud being addressed [63]. Therefore, it is not surprising that fraud detection is a complex task where diverse systems may be needed for different kinds of frauds. Each of these systems have different procedures and parameters to detect fraud.

The problem of fraud is primarily caused by extremely large databases. There are several areas where this problem occurs, such as in Tax Administration, Banking, Insurance, Health, Education and Telecommunications sector [63].

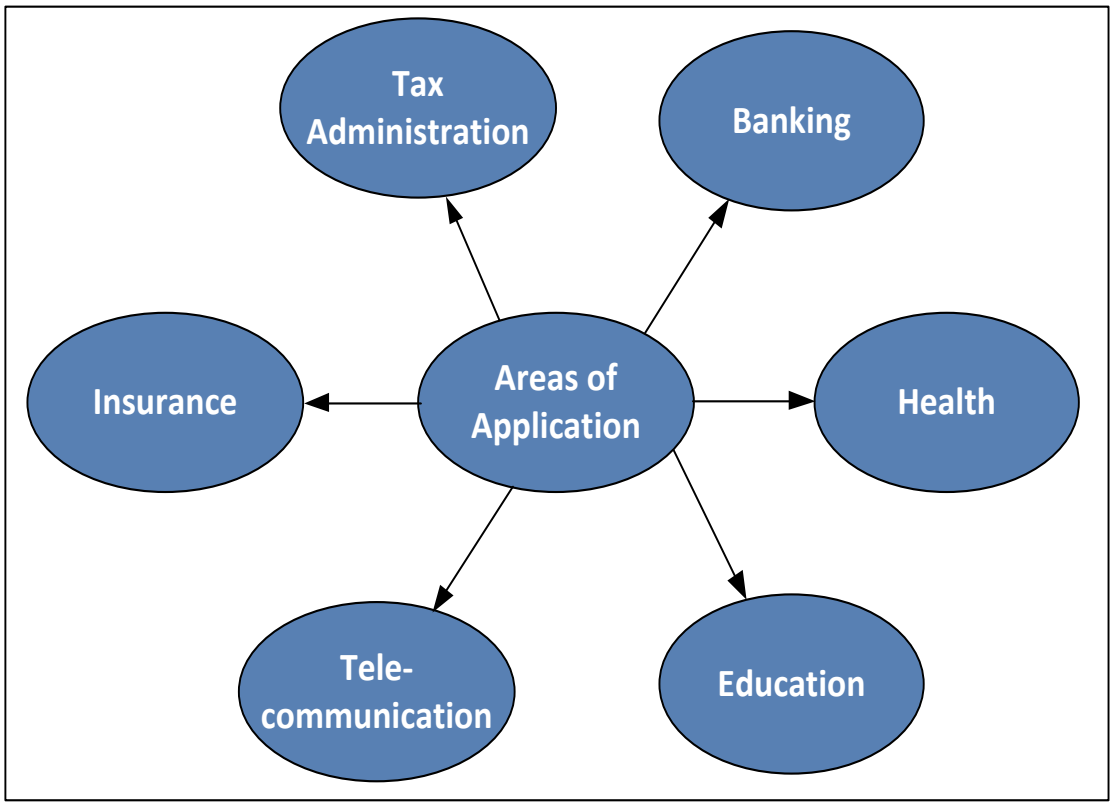

Figure 6: Areas of Application of Fraud Detection Techniques [65]

2.3.1 Data Mining in the banking Industry

As banks grew in size, channels and geographical footprint, their transactions jumped multifold. Manual reporting, which was time consuming error and filled redundancy proved unequal to the task and made way for automated systems. Banks began to increasingly depend on technology to manage their huge volumes of data. With the introduction of computers in banks, branch wise computerized reports, mostly Microsoft excel spreadsheets were consolidated at a bank level. However, they were not very comprehensive and being limited to banking transactions, did not support

decision making.

Just like the other sectors or industries, banking Industry has not been spared from the need to keep up with their constantly changing industry to stay viable and competitive Padhy N., et al states [66]. Systems previously involved manual recording of branch transactions and the generation of reports from manual ledgers according to Sahu R. et al, [67] these had to be consolidated with other branches into a final report. With such a far-reaching scope of services and concerns in the banking industry according to MicroStrategy [68] banks handle immense amounts of information. It's a big challenge to keep track of significant information and even to know which information is valuable. Its further stated that [68] the banking enterprises need the tools to take advantage of the vast information at their disposal and that the information technology available today allows these companies to make better business decisions and to better target performance goals so as to reduce costs, increase revenue, and maximize the value of information.

Suppose that bank management wants to establish the characteristics of clients that have been Insolvent in the past [69], echoes Bogdan U., et al, Such information can usually be requested from IT personnel at the bank, who, in such cases, must spend a considerable amount of time to produce the requested report, on top of their regular workload. By the time the report reaches the manager's desk, it may be too late for decision making. According to Ngai E., et al [65], nowadays, auditing practices have to be conducted in a timely manner to cope with an increasing number and occurrence of financial statement fraud cases. The novel techniques such as data mining, claims that it has advanced classification and prediction capabilities and can be employed to facilitate auditors' role in terms of successfully accomplishing the task of fraud detection. There has been a limited use of data mining techniques for the detection of financial statement frauds [70]. Data mining plays an important role in financial fraud detection, as it is often applied to extract and uncover the hidden truths behind the very large quantities of data [65]. Lin et al. [70], conducted an experts' questionnaire survey to evaluate the fraud factors using Lawshe's approach. The result of this expert questionnaires shows that 32 factors can be regarded as the suitable measurements for the continuing assessment of fraud detection.

2.3.2 BI and Data mining in the Education Sector

Though faced with challenges of implementing BI in higher education, Guster D et al

[71] states that BI is explicitly regarded as a solution with considerable promises in regard to adding the much needed efficiency on an operational level within higher education. Cheng M., [72] states that in higher education BI are becoming an inevitable trend as seen from the fact that some universities have applied BI systems as there educational systems while others focus on the research of BI.

Applying data mining techniques to educational data for knowledge discovery is significant to educational organizations as well as students. Knowledge driven data supports educational decision support system. Educational data mining enhance our understanding of learning by finding educational trends which includes improving student performance, course selection, in-house trainings and faculty development. Data mining techniques helps in increasing student's retention rate, increase educational improvement ratio, and increase student's learning outcome [73]. Thus, data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationship which help in effective decision making says Bala M., [73].

Today, specifically within higher education BI is viewed as a solution with much promise in regard to adding much needed efficiency on an operational level. However, there appears to be some confusion as to what actually constitutes a BI system. According to Guster D., et al [74], any kind of Higher Ed reporting systems is labeled as a business intelligence solution. She further states that one large impetus for BI in Higher Ed is the amount of disparate data sources and the time required to process integrated reports. Guster D., et al [74]. In contrast Florida State University (FSU) [74] is integrating BI metrics and thresholds in conjunction with workflow processes and has been able to adopt BI to sift and filter out at risk students that end up being "dead ends" and then focus attention on students with adequate academic skills.

There have also been successes related to the use of BI technology within higher education. Specifically, data mining and analytics have been used to analyze student data, guide course redesign and for retooling assessment as well as to encourage new communication models between instructors and student learning, Baepler et al. [75]. In summary, the authors examined how analytics and data mining can shape the effectiveness of teaching and learning.

2.3.3 Data Mining in the Health Sector

Healthcare organizations are under ever increasing pressure to do more with less and are continuously seeking ways to ensure that resources are deployed as efficiently as

possible while ensuring high quality patient care [76]. Charite T. [77] claims that Doctors don't typically have a lot of time, so any tool that attempts to change the way they do things to reduce readmissions needs to be streamlined and that knowing who your patients are, what conditions they have and how to keep them healthy and prevent them from needing expensive follow-up care are now top priorities.

Healthcare providers have known for a while that applying predictive analytics to the problem of patient readmissions was going to be important to their financial future. Now that they've had a few years to think about the problem, they're finding ways to make meaningful improvements in their readmission rates.

According to Charite T., [77] University of Tennessee Medical Center hospitals have introduced a new tool that takes data from the hospital's electronic health record system and clusters patients into different risk levels. It also assesses historic data to determine care strategies that have worked in the past for certain kinds of patients. One of the biggest reasons for the success of the tool is that doctors don't have to be too involved in its use [78]. They get recommendations each morning and then are able to apply their clinical skills. This predictive analytics tool has brought together quality health with financial stability.

Information is essential to meeting these goals – it has been referred to as the life blood of healthcare as it is essential for effective clinical and administrative decision making [45]. Healthcare decision making is complex and requires access to a wide array of high-quality information [79]. It is widely acknowledged that BI can provide benefits to health-care organizations including improved patient care and outcomes, effective utilization of human resources improved process efficiency.

Managing data in healthcare organizations has become a challenge as a result of healthcare managers having considerable differences in objectives, concerns, priorities and constraints [14]. The planning, management and delivery of healthcare services included the manipulation of large amounts of health data and the corresponding technologies have become increasingly embedded in all aspects of healthcare. Information is one of the most factors to an organization success that executive managers or physicians would need to base their decisions on, during decision making..

Healthcare organizations typically deal with large volumes of data containing valuable information about patients, procedures and treatments. These data are stored in operational databases that are not useful for decision makers or executives [14].

According to Ashrafi et al [80], health providers must be able to readily access and use

the right information at the right time and patients should be able to access their health information in order to be able to self- manage their conditions. This statement confirms that the Health Sector has equally not been able to resist the need for Business Intelligence as delivering quality healthcare requires the integration of patient health information from many different sources and availing a diverse set of users. The amount of data generated by and for the healthcare industry is overwhelming says Ashrafi et al [80]. He further states that it is business intelligence capabilities that deliver value by pulling data from various sources and bringing them into a common repository, enabling a thorough analysis of data, and creating insights into routine operations while providing decision support mechanism.

The literature search show how the capabilities of BI, use data and information to generate knowledge that serves as input for decision making in health care industry. Without a business intelligence solution in place, the physician would be tasked with manually sifting through vast amounts of data to hopefully make an accurate diagnosis. In this instance, business intelligence software helped address complications arising from cranial surgery, and was able to make the hospital more efficient and improve the treatment of critically ill patients.

Healthcare sector according to Batko K., et al is [81] one of the most dynamic sectors of the economy and health has become one of the major priorities of the individual countries. Moreover, healthcare is one of the fundamental tasks of modern states, and issues relating to them are the subject of interest in various scientific disciplines. The transformation of the healthcare sector has therefore become the recent subject of research in many fields, especially medicine, psychology, social policy. Recently, it is the subject of interest of economists, as well as management and IT specialists.

Interest in applications of Business Intelligence (BI) in different areas of the economy has been growing from year to year [81]. In recent years, it has been increasing in Poland as well. A relatively new area of using this systems is the healthcare area. Intelligent techniques provide an effective computational methods and robust environment for business intelligence in the healthcare domain. It seems to be very important, due to the fact, that much of the data storage in all kinds of system used in healthcare organizations resides in proprietary silos which makes access difficult more so considering that, data were previously stored and organized in the traditional way (both paper and digital), which was time consuming and difficult to ensure the desired level of efficiency. What is worth noting is that use of BI systems is determined by the

efficiency of the intelligent techniques, methodologies and tools.

2.3.4 Data Mining in the Tax Administration

Fraud in its various manifestations is a phenomenon that no modern society is free of. All governments, regardless of whether they are large or small, public or private, local or multinational, are affected by this reality, which seriously undermines the principles of solidarity and equality of citizens before the law and threatens business. There are many fields and industries affected by this phenomenon. In many cases there are even companies that are not known to have been victims of fraud.

As a general concept, Tax administration refers to the entire range of operations that a mandated government entity runs in order to implement and enforce the tax laws and regulations. Tax administrations have varying mandates, structures and naming conventions across different countries.

A tax administration's core business is, generally speaking, to get the right tax at the right time from the right taxpayers, and to make the funds timely available for the right tax recipients (the state, municipalities, congregations, and others). Tax laws and regulations determine from whom, how much, and when, tax is due. The laws and regulations set forth certain registration, filing, reporting and payment obligations that the taxpayers must observe.

Tax laws and regulations are, however, not always simple and easy to comply with. On the other hand there are always citizens and organisations deliberately seeking ways to avoid or evade taxes.

Many fraud detection problems involve a large amount of information Says Siehl E., [82] [15] considering that Revenue Authorities tend to have vast amounts of data accrued over time from both active and inactive taxpayers. Much of this information is known only in terms of the amount of data it consumes within the tax administration storage systems. Much of the intelligence value of this data remains untapped within many revenue authorities and tax administrators as long as data mining techniques are not employed to mine the irregularities on this data [83].

Processing of these data in search of fraudulent transactions requires a statistical analysis which needs fast and efficient algorithms, among which data mining provides relevant techniques, facilitating data interpretation and helping to improve understanding of the processes behind the data. These techniques have facilitated the detection of tax evasion and irregular behavior in other areas such as banking,

insurance, telecommunications, IT, money laundering, and in the medical and scientific fields, among others. Today, there are many fields and industries affected by this phenomenon.

The possibility to detect and prosecute tax violators depends crucially on data availability and data quality. Hence, actions taken against tax fraud relate to an improvement of the data quality available to tax officers.

In other economies, taxation is based on tax returns and other data submissions that the Taxpayers must give regularly or in the occurrence of certain events to the tax administration. The tax administration's systems process the data and calculate or validate the taxes due. Tax administrations generally receive data also from certain third parties for comparison purposes and for prepopulating the return forms for certain taxpayer groups. For instance in Finland the natural persons‟ income tax returns come pre-filled with data from third parties such as employers, banks and labor unions. In corporate taxation the tendency is towards self-assessment where the taxpayer him-/her-/itself calculates the tax, and it is the tax administration's task to validate it, either as such or corrected, pursuant to the verification and control measures.

According to Martikainen J., [84] two particularly prominent uses for data mining are identified within a tax administration's operational framework and these are;

Tax administrations can build up truly risk-based workflows for processing the registration, filing, reporting and payment transactions that the taxpayers make or should make. A comprehensive risk rating, based on data mining modeling, can be applied to each transaction so that all available relevant data are utilised. As a result, high-risk transactions can be flagged for case-specific treatment while low-risk transactions can move on to automated routine processing.

Using data mining, Tax administrations can segment the taxpayers and identify segment specific compliance profiles in terms of diverse abilities and tendencies to comply. This helps tax administrations better design and target their services and compliance actions.

Castellón P et al [15] indicates that previously to detect tax fraud, tax institutions began using random selection audits or focusing on those taxpayers who had no previous audits in recent periods and selecting cases according to the experience and knowledge of the auditors.

Later methodologies were developed based on statistical analysis and construction of

financial or tax ratios [15] which evolved into the creation of rule-based systems and risk models [15]. These transform tax information into indicators which permit ranking of taxpayers by compliance risk.

In the recent years, data mining and artificial intelligence techniques have been incorporated into the audit planning activities in several of Government organization with an objective mainly to detect patterns of fraud or evasion, which are being used by tax authorities for specific purposes.

Table 1: DM Techniques used by various Tax Administrations to detect Fraud [15].

No.	TECHNIQUE APPLIED	USA	CANADA	AUSTRALIA	UK	BULGARIA	BRAZIL	PERU	CHILE
	CLASSIFICATION	✓	✓		✓	✓		✓	✓
1	Neural Networks	✓	✓	✓	✓	✓		✓	✓
2	Decision Tree	✓	✓	✓				✓	✓
3	Bayesian Networks								
	CLUSTERING			✓	✓	✓			
4	SOM			✓					✓
5	K-means			✓					✓
	Naïve Bayes	✓					✓		
7	Visualisation Techniques	✓					✓		
3	Logistic Regression	✓		✓	✓	✓			
9	K-Nearest Neighbor			✓					
10	Association Rules							✓	
11	Fuzzy Rules							✓	
12	Outlier						✓		
13	Time Series		✓						
14	Regression				✓				

i. In the United States of America

According to Castellon P. [15], The Internal Revenue Service (IRS), the institution responsible for administering taxes in the United States, has used data mining techniques for various purposes, among which are measuring the risk of taxpayer compliance, the detection of tax evasion and criminal financial activities, electronic fraud detection, detection of housing tax abuse, detection of fraud by taxpayers who receive income from tax credits and money laundering.

During the presentation at the Government Big Data Symposium Butler J. [85],

confirms of the large service and enforcement which Internal Revenue Service (IRS) has as shown in the table below;

According to Padhy N., et al [66] the data mining system implemented at the Internal Revenue Service to identify high-income individuals engaged in abusive tax shelters show significantly good results The major lines of investigation included visualization of the relationships and data mining to identify and rank possibly abusive tax avoidance transactions. To enhance the quality of product data mining techniques can be used effectively. The Internal Revenue Service relies on technology more than ever to sniff out tax cheats using data mining as shown in Table 2.

Table 2: Tax Behavior and Analytics initiatives (IRS) [85].

Tax Behavior and Analytics Initiatives		
No.	**Behavior**	**Analytics Initiatives**
1	Failure to file or pay	Identity patterns of filing and
2	Abusive tax shelters	payment non-compliance
3	Identity Theft	Predict and prevent ID theft and refund fraud
4	Return preparer Compliance	Estimate US tax gap Measure taxpayer burden
5	Misreporting income or deductions	Optimise case inventories and treatment strategies
6	Refund Fraud	Simulate effects of tax changes
7	Off-shore transactions	Analyse criminal networks
8	Financial Crimes	

Source: Butler, J., 2013.

IRS extends the use of data mining to include the personal data of millions more taxpayers [86]. Its sophisticated data-matching and pattern-recognition technology, largely developed by IBM over the past decade, will reach up the income ladder to include more middle-income and small-business filers who itemize deductions.

According to McGuire P., [87] the U.S. Internal Revenue Service (IRS) is recognized

in the analytics and business intelligence industry for having one of the most complex yet accurate predictive analytics frameworks.

Table 3: IRS Enforcement and Services [85].

Internal Revenue Service Enforcement and Services		
1	Tax Return Processing	234 million tax returns filed 1.8 billion third-party information returns
2	Account Management	$2.4 trillion in gross receipts 122 million refunds totaling $415 billion
3	Customer Service	319 million visits to IRS website 83 million toll-free telephone calls
4	Enforcement	223 letters or notices sent to tax payers $116 billion in accounts receivable

She further states [87] that their focus is on finding pockets of revenue not reported or under-reported, and also looking at millions of returns quickly to see variations against constraint-based models. The IRS uses [85] their predictive analytics tools and applications as a first step in investigating potential tax evasion, fraud, under-reporting, tax preparer noncompliance, and money-laundering.

Business intelligence systems at IRS has a model which enables highly flexible queries against one of the largest databases in the world. IRS researchers can now search and analyze hundreds of millions or even billions of records at one time using a centralized source of accurate and consistent data instead of having to reconcile information from multiple inconsistent sources. Analysts are able to determine patterns in groups of people most likely to cheat on their taxes.

The data warehouse reduced the time it takes to trace mistakes in claims and analyze data from six to eight months to only a few hours.

The DW is more secure than the old legacy system storage tapes, thereby better protecting taxpayer data. Literature also reviews that the Norwegian government implemented BI and Data mining in 2014, on a broader basis [8], whilst focusing on data analysis, business intelligence, and risk-based tax auditing.

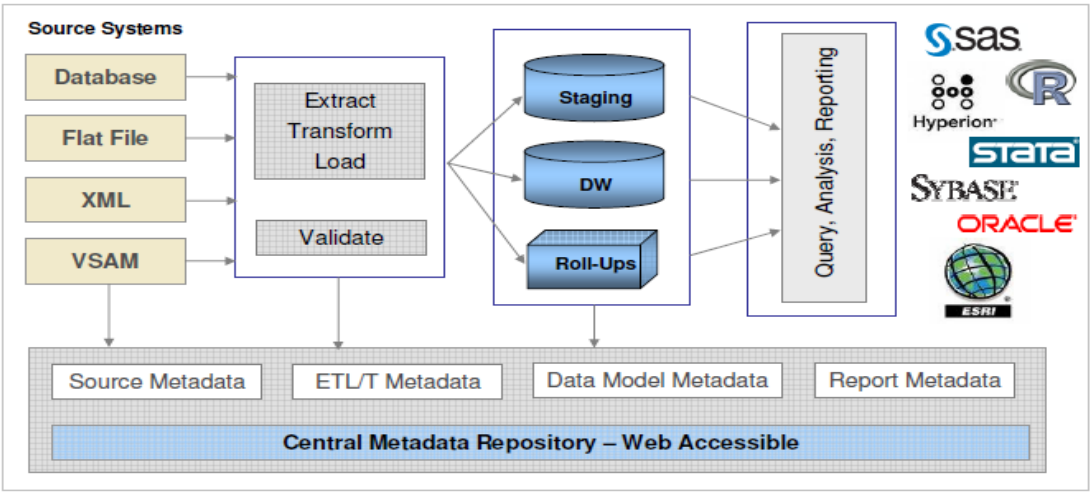

Figure 7: BI and Data Mining Model, IRS [85].

ii. In Peru

Peru was one of the first to apply these techniques to detecting tax evasion [15], adding to the selection system of the Maritime Customs of Callao an artificial intelligence tool based on neural networks. This model was improved through the application of fuzzy rules and association for pre-processing variables and classification and regression trees (CART) to select the most relevant variables.

iii. In Brazil

Has developed project risk analysis and applied artificial intelligence (HARPIA) jointly with the Brazilian Federal Revenue and universities in the country [15]. This project consists of developing a detection system of a typical points to help the regulators to identify suspicious transactions based on a graphic display of information on historical imports and exports and a system of export product information based on Markov chains, to help importers in the registration and classification of their products, avoid duplication and to calculate the probability that a string is valid in a given domain.

iv. In case of Chile

Data mining has been in Chile. Literature shows that the first trial was developed in 2007 [15] using the SOM and kmeans to segment VAT taxpayers according to their statements and characteristics. Later, in 2009, following the international trend, risk models were built of different stages of the life cycle of the taxpayer, in which neural networks, decision trees and logistic regression techniques are applied. The first trial was further developed to identify potential users of false invoices through artificial neural networks and decision trees, mainly using

information from tax and income declarations in micro and small enterprises.

v. In the Australian Tax Office

According to a report on a review into the Australian Taxation Office's compliance approach to individual taxpayers [88], the ATO currently holds 19.3 million active tax file numbers (TFNs) for individual taxpayers. Under Australia's self-assessment system, individual taxpayers are responsible for lodging annual income tax returns in which all assessable income is to be declared and only deductions, offsets and credits to which that taxpayer is entitled are to be claimed. From this information, [88] the ATO processes and determines the net amount of tax payable or refundable to the taxpayer.

Each year, the ATO receives approximately 12.4 million individual income tax returns, over 10 million of which contain claims for deductions, offsets or credits totaling $54.1 billion.

The ATO acknowledges that within this system there is a high risk of taxpayers and tax return being lodged incorrectly considering the high volumes of data being generated by the lodging of the returns. His has therefore created the need for ATO to employ complex analytical models to risk assess and select certain returns for manual checking [15]. This Compliance Program is based on a risk model which uses statistical techniques and data mining in order to make comparisons, to find associations and patterns by logistic regression, decision trees and SVM [15].

vi. In Africa:

Whilst Drummond et al, [89] identifies that raising more domestic revenue is a priority for most sub-Saharan African countries and that organizing this revenue is a way for governments' to create economic space in order to provide essential public services such as building of roads infrastructure, school infrastructure, health infrastructure and reduce foreign aid from developed countries.

It has been evidenced that [90] the domestic tax bases in most African countries are weakened by widespread tax frauds. This shows that today, African governments and their public sector agencies everywhere have not been spared from the pressure to perform more efficiently and effectively as Corruption in tax administration in Africa remains a fundamental barrier to effective and fair taxation and to building trust between government and citizens.

Previously, the traditional methods for addressing risk have served many authorities well, but there is now a need to use more advanced technologies to

combat fraud, error and waste in the Tax Administration sector such as Business Intelligence, Data Warehouse and Data Mining with specific algorithms such as Outlier algorithms to detect pattern in the huge data.

There are very few recent studies assessing the extent to which and how fraud affects tax administration in Africa, but surveys on citizen experience [91] and perceptions of corruption within tax administration paint a worrying picture, with more than 50% of respondents who were in contact with tax administrations having reported experiencing corruption when dealing with tax and custom officials in several African countries. Studies and anecdotal examples also demonstrate that corruption in tax administration takes different forms, from bribery to patronage, to revolving doors and regulatory capture.

Approaches to fighting corruption in tax administration undertaken by governments in Africa often aim at addressing the main drivers of corruption [91]. They include measures to enhance the autonomy and capacity of tax agencies, for example through the establishment of semi-autonomous tax agencies, higher salaries, measures to improve tax services and reduce tax-payers interactions with tax officials, by for instance investing in technology and tax-payer education, as well as measures to improve internal control and oversight and encourage informants to report corruption.

The fact that not so much has been studied in Africa in terms of fraud in Tax Administration, technologies to detect fraud has not extensively been studied as well. This is evidently indicated by the Google, Google scholar and Science Direct Search where every research returns very little about the Data Mining and the applied algorithms in the Tax Administration sector.

Botswana, a rapidly developing country with many new organisations establishing presence every year also acknowledges the challenges to analyse data effectively and efficiently in order to gain important strategic and competitive advantage [11].

a. Bi and Data mining in Morocco

Studies [92] indicate that Moroccan tax authorities have not been spared from tax evasion including underreporting and underpayment which is seen as the set of behaviors of the taxpayers that aim to reduce the tax which must be paid normally. This is where tax liability owed by a taxpayer is reduced or eradicated entirely. Strategies of tax fraud and tax evasion are diversifying in an international environment and becoming more and more complex. However, based on the fact

that the detection of fraudulent taxpayers is the most difficult step, morocco [92] established that a new approach using Data mining process in order to stop such behavior, thereby improving fiscal control. This approach is based on Data mining techniques which are classification and prediction through Tanagra software.

b. BI and Data mining in Kenya

In Kenyan, studies review [93] that evasion and schemes are growing in complexity. The dynamism in their adaptation, modification and transformation as well as the exponential growth of the information available to tax administrators has added pressure to the Tax administration in terms of fraud and tax evasion considering that taxpayers complete most of their procedures and compliance actions on the web. Kenya tax administration however recognize the fact that Data mining based on statistics and artificial intelligence allowing extraction of useful knowledge have been in use for a long time in the industries such as banking, telecommunication and health giving them confidence that the use of such techniques can be implemented in the Tax Administrations to improve the detection of Fraud.

c. BI and Data mining in Tanzania

In Tanzania, Arusha region, a research study indicates and suggest that Data mining is very important as it can improve the industry as well as the well-being of the citizens. To arm themselves for this battle, [94] more and more tax authorities in Africa are recommending turning to Business Intelligence, data mining and analytics to improve their business processes, which will result in better compliance.

d. In Zambia

Unindustrialized countries like Zambia are experiencing barriers to attaining key objectives of taxation. These objectives are worldwide, but countries attempt to achieve them in very different ways. To be precise, developing countries, face great obstacles in achieving these aims as the need for high government expenditure is so greater in developing countries where the capital stock (for example, schools, hospitals and roads) is low [95]. During a presentation at a workshop jointly organized by the World Bank Institute, PRMPS in South Africa it was confirmed that developing countries which includes Zambia has not been spared from obstructions to achieving its key objectives of taxation by intelligently exploiting data (data mining and predictive analysis) [96].

In Zambia, tax revenues fund a lot but not all of expenditure. The shortfall is mainly covered by foreign support from cooperating partners and by government borrowing, both locally and internationally.

2.4 Summary

In this chapter a comprehensive overview of the background theory, concepts, and technologies has been given. Examples of the area of application and related works indicating where and how Business Intelligence data mining particularly detection of fraud using the Outlier Algorithms has been given.

CHAPTER THREE:

METHODOLOGY

3.1 Introduction

This chapter firstly defines the research methodology of this research study giving details of the sample selection and a description of the procedure used in designing the instrument as well as the collection of the data. The procedures or techniques used to analyze the data is also described.

This study endeavors to investigate the challenges faced by ZRA regarding fraud detection on Taxes and to design a Business Intelligence data mining Fraud detection model as well as develop a prototype based on the model which will detect fraud on tax data to enable the organization achieve its desired goals.

3.2 Study Design

This research study used a Mixed Methods Research (MMR), involving both Quantitative and Qualitative methodology [97] [98] [99] [100] as it is acknowledged that a greater depth of understanding of the study is generally gained by qualitative research than by quantitative research, while better objectivity and generalizability is obtained by quantitative research.

Mixed methods requires that qualitative and quantitative data are collected simultaneously in a research.

The rationale for using both qualitative approach and quantitative approach known as Mixed Mode Research (MMR) lies in what Creswell J., [101] specifies that combining quantitative and qualitative research designs is very advantageous with the objective to preserve the strengths and reduce the weaknesses in both quantitative and qualitative designs. It offers extra and thorough understandings into the phenomenon being studied such as the data mining and fraud detection systems on tax data in ZRA. This approach also permits the capture of information which might be missed if only one research approach is going to be utilized which is either quantitative or qualitative.

3.3 Baseline Study

The baseline study for this research involved establishing the extent of the challenges in fraud detection for the taxes and tax payers, and examining the methodologies, processes, architectures, and technologies currently used in ZRA to transform raw data into meaningful and useful information. A survey using a structured questionnaire and unstructured interviews were used to collect data from the participants.

In this study 200 Questionnaires [102] [103] were distributed to employees from the three (3) different business layers within ZRA namely, Domestic Taxes, Customs Services and Information Technology. Additionally, oral interviews were also used as a way of obtaining information.

Further this research study only focuses on the historical data from Domestic Taxes. However, ZRA has two divisions [104] (Customs Services Division and Domestic Taxes Division) which are directly involved in Tax collection. In carrying out their mandates, these two

Divisions are supported by the following; the Corporate Services Division, Finance Division, Human Resource Department, Research and Planning Department, Project Management Department, Internal Audit Department, Investigations Department and Information Technology Department. ZRA has its presence across the country as shown in Figure 8 [105].

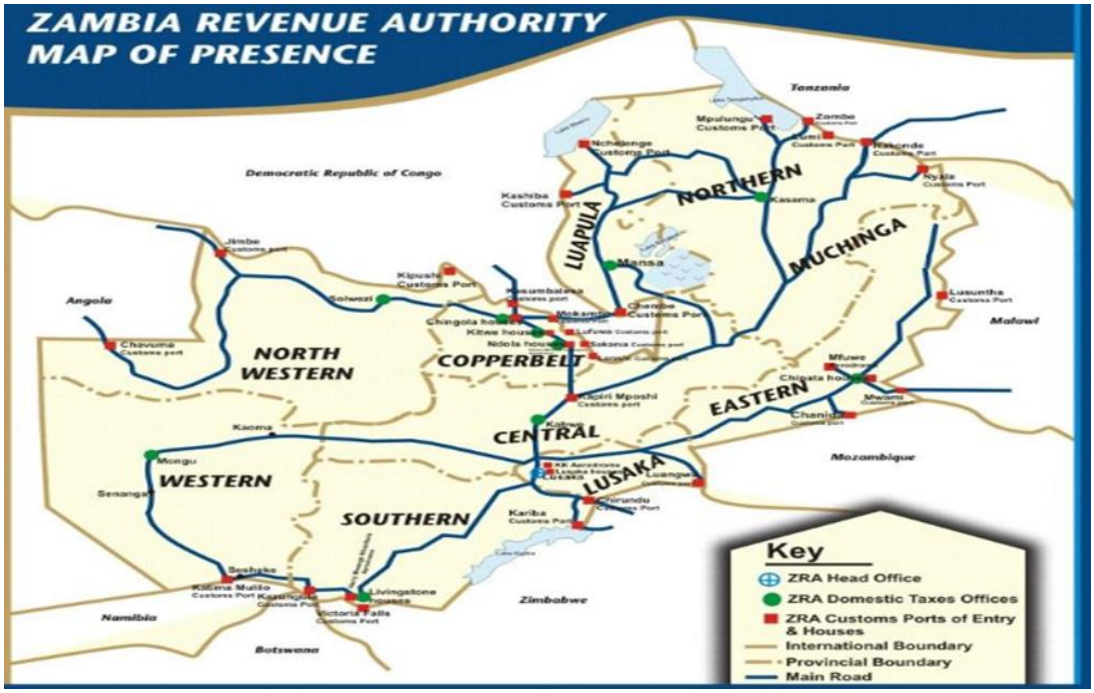

Figure 8: Points of Presence, ZRA [105].

For the baseline study, the structured questionnaire and the unstructured interviews were used particularly;

i. To understand the extent of challenges ZRA is facing in terms of fraud detection on the tax data.
ii. To examine the current technologies used in ZRA to transform raw Data a structured questionnaire was used

The baseline study also assisted to understand the history of ZRA and its mandate in terms of Tax administration. Business documents such as the ZRA Act and the corporate strategy. This gives a guide on the government's expectation in terms of Tax Collection.

The Zambian tax system according to Nhekairo W., [95] broadly comprises of income taxes, consumption taxes, property taxes and trade taxes. These taxes are collected by the Zambia Revenue Authority (ZRA), the agent whose mandate is to ensure revenue is collected on behalf of the government.

Table 4: Tax Categories in Zambia [95]

NO.	TAX CATEGORY	TYPE OF TAX
1	**Income taxes**	Company income tax Pay As You Earn (PAYE)
		Withholding tax Mineral royalty Property Transfer Tax
2	**Property taxes**	
3	**Consumption taxes**	Import and domestic VAT Excise duties Customs duty
4	**Trade taxes**	
		Export duty

3.3.1 Sampling

For this research study, Stratified Sampling technique was applied in order to obtain a representative sample [106]. Results from a Stratified sampling technique are more reliable and gives much detailed information and this therefore justifies the sample selection from the employees of the Zambia Revenue Authority with consideration of all Divisions with common characteristics such as Tax administration or management issues.

In this study the method used to acquire the sample of participants was based on inclusion criterion.

Sample size per stratum = $n_i = n.\ P_i = n.\ \left(\frac{N_i}{N}\right)$ (Equation 1) [106]

Source: Kothari, C.R., 2009.

Using equation 1, the sample size shown in Table 6 were determined;

Where:

i. n is the sample size or total number of samples to be drawn from the sample space.
ii. n_i is the number of samples to be drawn from the i^{th} stratum.
iii. N_i is the population per i^{th} stratum.
iv. N is the total population in a sampling frame or unit.
v. P_i is the ratio of the total number of samples in the i^{th} stratum

Table 5: Selection Criteria of Sample Size

Inclusion criteria
Types of participants
This research included only ZRA employees
The study included those that work in the Division involved in Tax Collection, Tax Administration and also those involved in Data Management.
Types of Stations of interest
The research study considered Employees from all ZRA stations as long as they have Domestic Tax, Customs and Support Services presence.

a) Domestic Taxes = 200 (900/1800) = **100**
b) Customs Services= 200 (700/1800) = **78**
c) Support Services = 200 (200/1800) = **22**

Table 6: Sample Size

No.	Division	Population N_i	Sample Size Calculation (n_i)	Responses per stratum.
1.	Domestic Taxes	900	100	86
2.	Customs Services	700	78	55
3.	Support Services	200	22	15
	Total	**1,800 (N)**	**200 (n)**	**156**

Total Population of ZRA is 1800 employees

3.3.2 Data Sources

The sources of data for this research was primary and secondary source. The primary source of information was from the constituted sample from the research and was captured from the end Users of the systems through the questionnaire. The secondary sources of data was from the published materials, mainly hard copy books, eBooks on the internet, reports and journals. White papers presented at local and international forums were also used. Some informal interviews were also held in order to acquire the information on the history and the mandate of ZRA.

The research also utilised data held by the database. This data was extracted from the Domestic Taxes Databases using the ETL process as shown in Figure 8.

3.3.3 Types of Data Problems

Internally, an organization can also create section data due to modernization in order to suit their needs. For example, the Tax Administration in Zambia has mordenised its Tax Administration functions into Small Tax Office, Medium Tax Office and Large Tax Office.

For this study, data from the Data source had some defects and was inconsistent as defined in section 2.2.3.

i. Missing values/attributes**,** when data was extracted into the Excel files, some values were missing such that it was difficult to even import them into the database. Some fields like date were left blank which needed to be filled with the correct data.

ii. Introduction of new attributes, this was introduced to help censor the data so that the information is not displayed as it is originally from the data sources. Modification of data such as unique Identifiers TPIN was done.

3.3.4 ETL Tools and the Process

Extract, transform and load (ETL) is the set of functions combined into one tool or solution that enables companies to "extract" data from numerous databases, applications and systems, "transform" it as appropriate, and "load" it into another database, a data mart or a data warehouse for analysis, or send it along to another operational system to support a business process [107]. For this study, two (2) java

classes, PaymentDataNormaliser and ProfileDataNormaliser were created to extract data from Excel to the MySQL database.

The goal of this task expressed in Figure 9 was to ensure that the data had formats which are conforming to the database requirement and also to understand the format of the data, assess the overall quality of the data and to extract the data from its source so it can be manipulated in the next task.

This database contains information relating to profiles of taxpayers that is both importers and exporters. Additionally the database also contains the total amounts in connection with imported and exported goods.

i. **Extracting.**

   The raw data coming from the Data source was extracted directly and exported to Excel Sheet. This allows the extract to be as simple and as fast as possible and allows greater flexibility to restart the extract if there is an interruption [108].

ii. **Cleaning**.

   In most cases, the level of data quality acceptable for the source systems is different from the quality required by the main database [108]. Data quality processing involved many distinct steps, including checking for valid values such as the Tax payers Identification Number (TPIN) and whether it is in the range of valid values, ensuring consistency across values such as the TPIN's and the Tax Payer profiles consistency, removing duplicates where the same Tax Payer appear twice with slightly different attributes, and checking whether the business rules and procedures for Administering taxes have been enforced such as any Tax payer with a turnover tax above 50 million to be classified as Large Tax Payer.

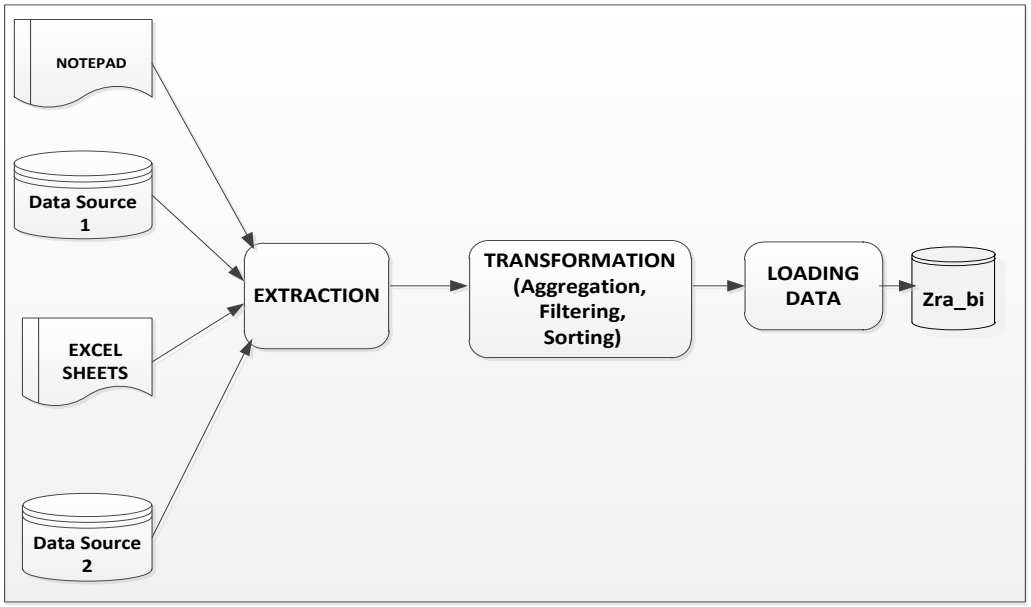

Figure 9: Extract, Transform, Loading [108].

iii. **Conforming**.

Data conformation is required whenever two or more data sources are merged into one data source [108]. Separate data sources for example Data Source 1 and Data Source 2 cannot be queried together unless some or all of the fields in these sources have been made identical.

iv. **Loading**.

Ultimately loading data into the main data source allows easy querying because all of the fields in these sources have been made identical. Chillar R. S. [109]. This also make Analysis of data using application software easier. This data is then used to discover patterns and anomalies and also create visualisation such as a table or graph [110] [19].

v. Confidentiality of Extracted Data.

To maintain confidentiality of tax payers Data that was extracted from the existing Data Sources, the technique and science in Figure 10, of creating non readable data or cipher so that only authorized person is only able to read the data [111] [112] was used . This technique uses the three methods in which each letter in the plaintext is replaced by some fixed number of position down the alphabet. This art and science of protecting information from being reasoned by converting it into a form non-recognizable in this research [113] [112] . Caesar cipher was adopted for this research. It is a substitution type where each letter in the plaintext is replaced by a letter some fixed number of

positions down the alphabet [114].

For this research study, the cipher used a left shift of three, so that each occurrence of E in the plaintext becomes B in the cipher text. The action of a Caesar cipher was to exchange each plaintext letter with a different one a fixed number of places down the alphabet for the purpose of scrambling data.

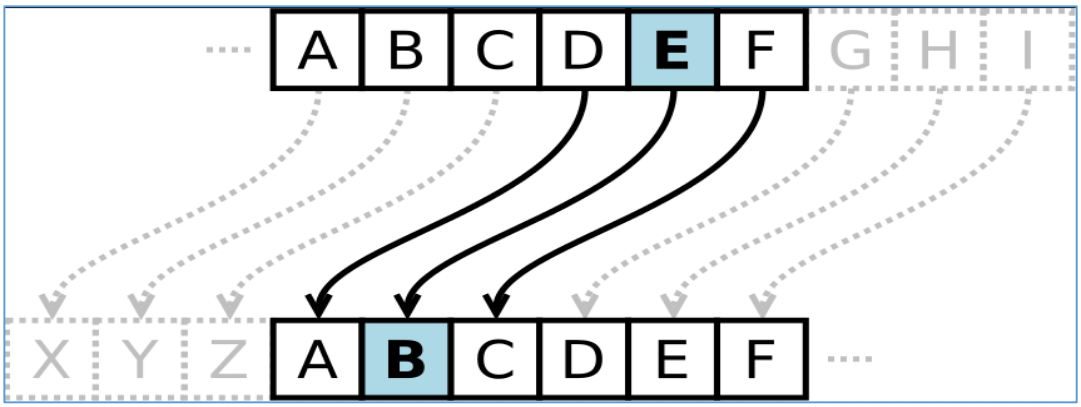

Figure 10: Caesar Cipher for Confidentiality with a shift of 3 [114]

3.3.5 Data Collection Tools

There are several [106] methods of collecting appropriate primary data particularly in surveys or Experiments. The instrument used to collect data for this research was the structured questionnaire and the informal interview. This method of data collection is quite popular, particularly in the case of big enquiries [106], and is being used by private individuals, research workers, private and public organisations and even by governments.

A questionnaire was therefore sent to the sample (200) selected from the population with a request to answer the questions and return the questionnaire response [95] [97] a questionnaire is a 'tool' which consists of a number of questions printed or typed in a definite order on a form or set of forms.

3.3.6 Data Collection Procedures

The procedure of collecting data for this research involved an Iterative process of designing questions and instructions to balance flexibility and specificity. Two associates were engaged to review and revise the questions and instructions to ensure correctness. To come up with a model, a systematic review of the literature was done. This procedure reviewed different model in existence. The model for this research was then built based on the existing ones.

In order to design and develop a prototype for this research study, authorization to

access the ZRA Data was required. An application letter addressed to the Commissioner General was used. Data was extracted from the Data Sources by the authorised Database Administrator and the extracted Data was scrambled in order to achieve confidentiality. Taxpayer's identity both Business Names and the TPIN's were scrambled to achieve secrecy.

3.3.7 Data Analysis

Though there are a variety of packages used for analysis of data today such as [115] NIVA, SPSS, SAS and Stata packages, formal SPSS computer programme version 21 was preferred for data analysis for this research.

SPSS is a widely used program for statistical analysis and has been used most researchers such as Market researchers, Health Researchers Survey Companies, the governments education researchers and Data Miners [116]. All statistical tests were at 5% significance level. The Pearson's chi-squared test and Fisher's exact test were used for comparison of proportions between groups. The Fisher's exact test was used when one or more of the cells had an expected frequency of five or less. Both SPSS and MS Excel were utilized for analysis. The collected data was categorised into essential and appropriate groups. This process also involved Coding the question, tabulation, Data Cleaning, analysis as well as examining the Questionnaires for correctness. Existing duplicate records within the database were removed and all incomplete surveys were discarded from the analysis. According to Ghosh B. [117], data analysis requires logical organisation of data in order for the logical results to be achieved. Generally, the idea of statistical analysis is to summarise and examine data in order for it to be useful for decision-making.

3.4 Implementation of BI and Data mining model for ZRA Fraud detection.

A fraud detection model was then designed based on the literature review in chapter two and also the findings of the baseline study covered under objective one in chapter four. Further, Microsoft Visio 2013 was used as a tool to come up with the model.

A five layered BI and Data mining model to help with Fraud detection on Bulky Tax Data was proposed and designed. This five layered model in Figure 13, takes into consideration the compliance enablement, Taxpayer processes importance and quality of data as well as information flow that contribute to the bulk tax data on the data

sources, ETL (Extract-Transform-Load), Integrated data store (Data Warehouse), data mining and End user component.

3.4.1 Implementation of the ZRA Fraud detector prototype.

Based on the model shown in section 3.4 The ZRAFraud detection tool was therefore proposed and implemented for Zambia Revenue Authority and shall provide functionality for detecting fraud on the bulky tax data. The tool has two components, namely;

i. **Data Extraction and cleaning component**: This component extract Payment Data and Profile Data from a Microsoft Excel spreadsheet, transform and load data into MySQL database. This components implements two java classes' paymentDataNormaliser and profileDataNormaliser to do the task.

ii. **Detection component**: Outlier detection was used to detect the Fraud on Tax Data and the distance between data objects to detect those objects that are grossly different from or inconsistent with the remaining data set [61]were considered. Data that looked to have different characteristics than the rest of the population were regarded as Outliers. Two algorithms (Continuous Monitoring of Distance Based and Distance based algorithm) that uses the Weka Java libraries to detect all fraudulent transaction in the database were used.

Both components uses the MySQL-Connector to connect to the Database.

Java programming language using NetBeans Integrated Development Environment (IDE) was used to develop the prototype. It is open source hence there are no licences required. The Database was created in MySQL. Workbench 6.3 CE was used to interact with the Database.

The hardware used is a laptop with 2.80GHz intel core Duo CPU, 4GB RAM and 500GB hard drive space.

3.4.2 Business Process Mapping

Tax administrations today tend to have their operations organised in processes. While taxation is the obvious one intersecting all core business process, there are several other important processes around to make taxation possible, efficient and effective.

Boyer et al. [118], confirms that in order to achieve business effectiveness, BI requires to be linked to the goals of the organization and aligned with the business and corporate strategy. He further states that organisations that successfully align their BI initiatives with business and corporate strategy can collect the benefits of overall performance improvements. Successful linking the corporate and business strategy to BI allows visibility into the relevant information and has the ability to improve decision-making capabilities.

It is vital that BI initiatives are aligned with corporate business strategy to enable successful execution of the initiatives, which is the first critical step in a strategic approach to BI. It is also relevant to understand how the organization monitors and measures outcomes on the journey towards successful business alignment strategy [118].

In order to come up with the model and the prototype, it was quite necessary to understand the main business processes as well as the corporate strategy of the Zambian Tax Administration (ZRA) as shown in Figure 11. A brief description of each later follows;

i. **The Taxpayer Education and Outreach process** determines how the tax administration seeks to increase the taxpayers' ability to deal with their taxation matters correctly. Typical tax administration's measures here include information dissemination in various forms and channels, targeted educational campaigns, and personalised guidance in complicated situations. The principal aim is to prevent the taxpayers' mistakes before they occur.

ii. **The Data Flow Management process**, is the tax administration's "logical" process. It defines how inbound data, the tax administration's "raw material", are received from various sources and interfaces, and how these data are transmitted to operational systems. The outbound end, the process defines how the outputs of the Taxation and the Taxpayer Education and Outreach processes are prepared for delivery in the appropriate channels to the recipients.

   The Taxation process covers the tax administration's procedures to attend to a taxpayer from the moment a ground for taxation emerges until it discontinues. The process lays down the workflow in customer registration issues, in validating the taxable income and taxes due**,** in conducting tax audits, as well

as in overseeing that taxes are duly paid and in pursuing collection measures for the indebted taxpayers.

Business Process Mapping for Zambian Tax Administration

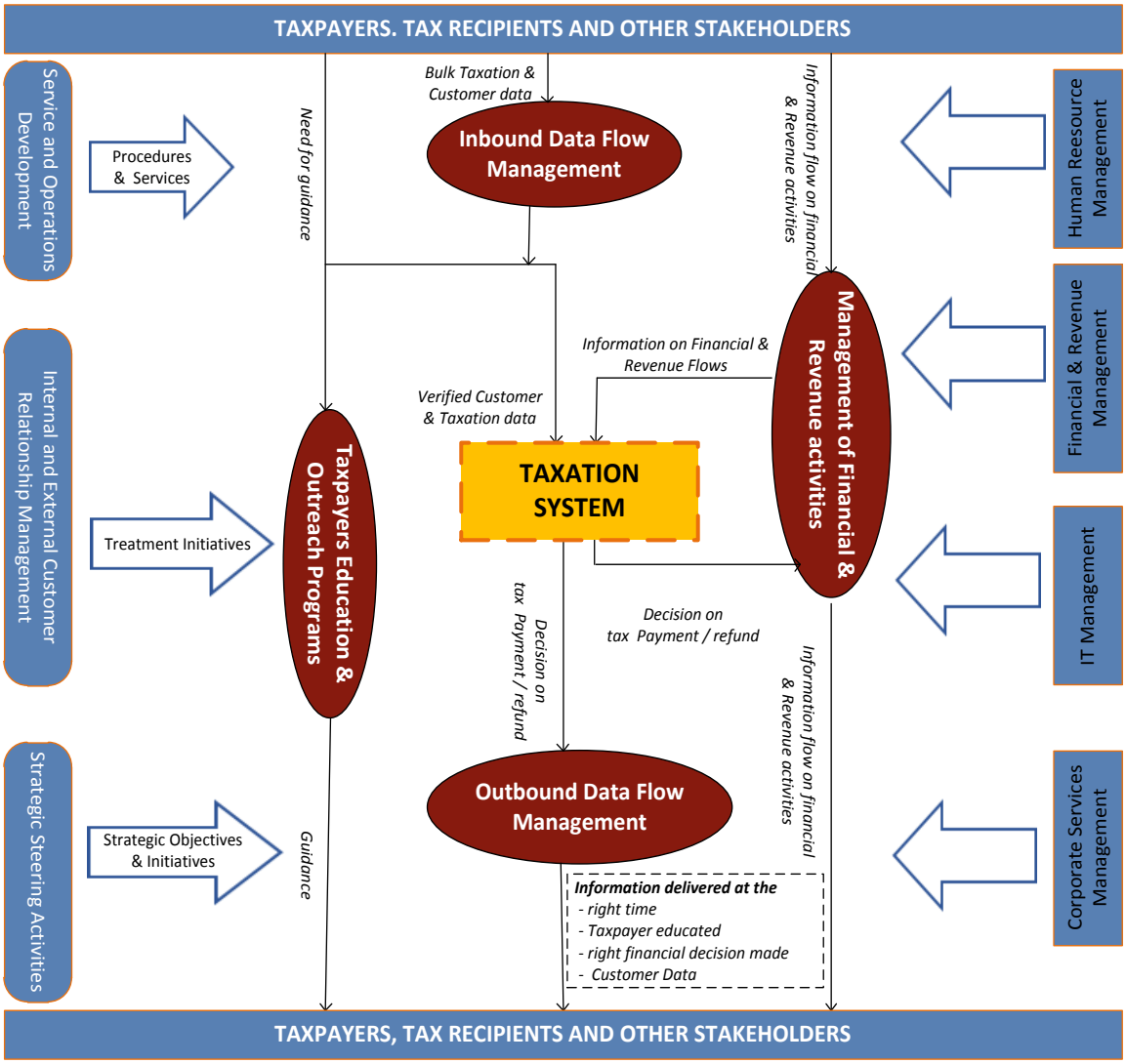

Figure 11: Business Process Mapping of the Zambian Tax Administration

iii. **The Financial and Revenue Flows Management process** defines the procedures in distributing the tax revenue among the tax recipients (the state, municipalities, congregations, and others) and transmitting the respective funds to them.

iv. **The Strategic Steering Process** covers the environmental scanning, strategic planning, medium-term business planning, short-term operational planning, resources planning, operational target setting, as well as the follow-up of the strategy implementation and effectiveness.

v. **The Customer Relationship Management process**, determines how taxpayer behaviour is observed and analysed, how these analyses are used for taxpayer segmentation, and how segment-specific treatment strategies are designed. The treatment strategies are deployed in the Taxation and the Taxpayer Education and Guidance processes.

vi. **The Services and Operations Development process** outlines how the tax administration's services and operations are developed. The starting point here is typically a development initiative together with a corresponding needs assessment. The outputs can be, for instance, new internal procedures, new services for taxpayers, or new treatment measures to address certain tax risks.

3.4.3 Current Taxation process of a Zambian Tax Administration

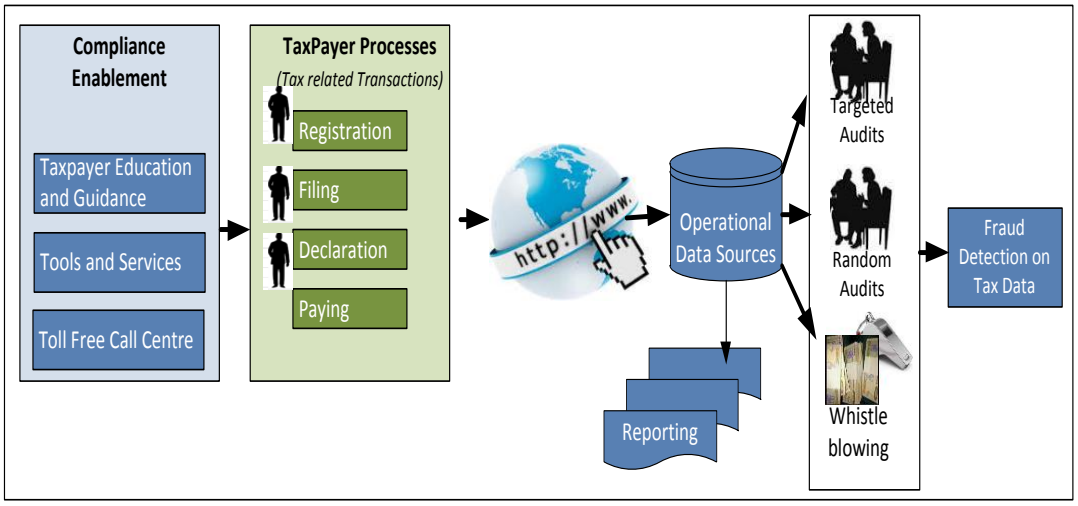

Figure 12: Taxpayer from the moment a ground for taxation emerges

i. Compliance Enablement

This Compliance Enablement in Figure 12, aims at creating a framework within which all taxpayer education activities by the Zambia Revenue Authority (Authority) shall be optimized with the sole aim of helping the organization meet its strategic business objectives through building excellent stakeholder relations, both internally and externally.

Compliance Enablement such as taxpayer education and call center serves as the conduit for effective and efficient revenue collection in any revenue authority. It plays a vital role in promoting voluntary compliance, disseminating critical

information such as rights and obligations of taxpayers and ensuring taxpayers understand tax laws and how they are expected to manage their tax obligations.

ii. Taxpayer's processes

The taxpayer shall obtain a return from ZRA offices or Portal. The taxpayer shall Register, make declarations, make payment or file the return manually at any ZRA office or through the portal **[119]**.

The DEO shall capture all the data on the return as entered by the taxpayer and submit the task. The system shall then create a validation task for the Validating officer.

The Validating Officer shall verify the data captured against the data on the physical return.

Based on the transaction details, the data will be stored in the data sources as shown in Figure 13. Audits are then effected based on the tasks allocated to the officers and details of the transaction. Audits are also effected based on the need arising. It is at this point that fraud or any anomaly on the bulk Tax data get to be detected and later flagged for further investigation.

3.4.4 Implemented model of a BI for the Zambian Tax administration.

This model implemented encompasses the compliance enablement and the tax payer processes which provides a framework and a channel for effective and efficient revenue collection in the Tax Administration. It will play a vital role in promoting voluntary compliance and in turn generate big data for the tax administration. The central repository is another crucial component for all or significant parts of the data that the tax administration collect. Data mining algorithms are then introduced from which fraud or anomalies are detected. Analytics for reporting purposes will also be carried out.

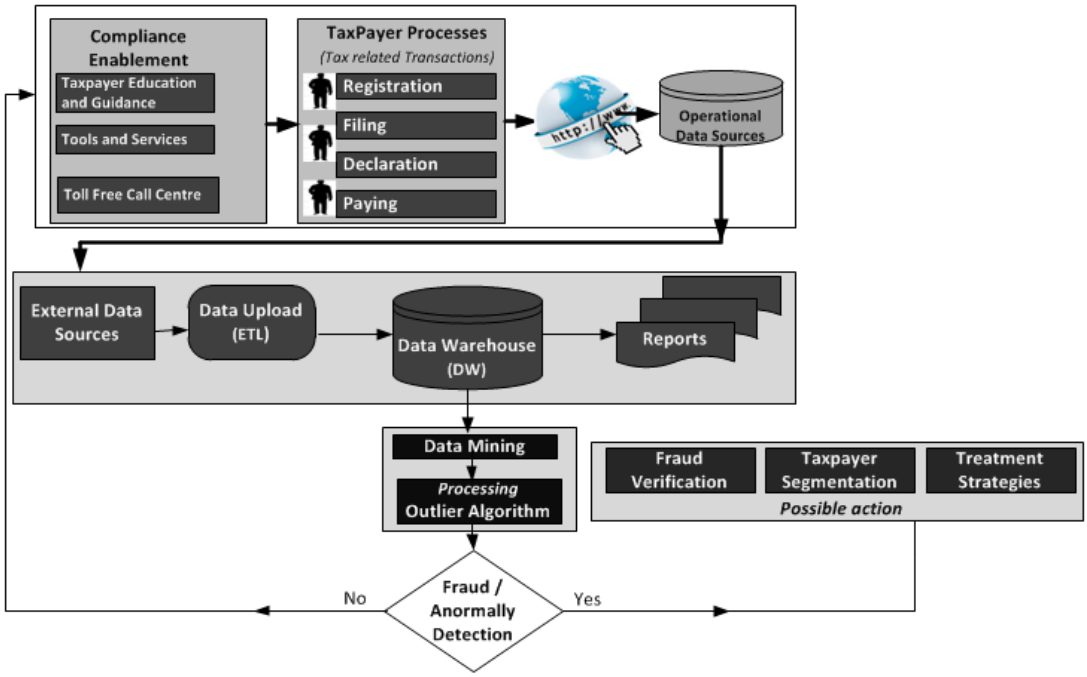

Figure 13: Implemented model for Zambian Tax Administration

3.4.5 User requirement specification

A rating has been provided for each requirement as shown below;

D - **Desirable,** meaning the function rated as such is appropriate.

M - **Mandatory,** meaning the function rated as such is required compulsory.

Table 7: Functional User Requirements Specification.

No		Function	System Features	M/D
1.0		**Data Sources**		
			a) The system should be able to read and/or extract data from all internal systems; (i) Data Sources 1 (ii) Data Sources 2 (iii) Data Sources 3	M
			b) The system should be able to read and/or extract data from all ZRA external partners	M
			c) The system should be able to read and extract data from semi-structured and unstructured data (BIG Data)	M
			d) The system should be able to transform data as required by the authority	M
			e) The system should be able to handle large volumes of data (i.e. at least a minimum of 50 terabytes)	M
			f) The system should have capability to clean or transform unstructured data to required	M

No		Function	System Features	M/D
			format and then load this data from all the sources into the data warehouse	
2.0	**Data Warehouse**			
			g) The system should have conformed dimensions	M
			h) The system should have capability to configure new dimensions and fact tables	M
			i) The star schemas should be designed at an appropriate granularities	M
			j) The system should be able to handle slow changing dimensions (SCD) without losing history	M
			k) The system must ensure that data integrity is maintained at all levels (i.e. Database level, ETL process level, and Access level).	M
			l) The system must have OLAP (On-Line Analytical Processing) capabilities	M
3.0		**Data Analysis**		
			m) The system should be able to generate various reports using different dimensions (for example, comparisons across Divisions and tax regimes for: non-compliance, revenue compliance, audit outcomes, tax type performance by trader, sector and other groupings.)	M
			n) The system should be able to generate various moving average reports	M
			o) The system should be able to generate rolling charts/reports	M
			p) The system should have capability to visualise data as specified by the user. (For example via pie chart, bar charts, graphs.)	M
			q) The system should have configurable Key Performance Indicators (KPI) and Key Performance Indicator (KPI) action lists	M
			r) The system should provide for centralised risk-based selection for audit and investigation through risk profiling	M
			s) The system should have capability to perform trend and pattern analysis	M
			t) The system should be able to handle calculated measures	M
			u) The system should be able to perform centralised transaction verifications (CTV) on VAT related taxpayers.	M

No		Function	System Features	M/D
			v) The system must have the provision of on-line warning or alert messages via the system itself and also through email	M
			w) The system should be able to send standard reports via email to appropriate recipients at configurable time periods	M
			x) The system should have dash boards and scorecards and capability to configure new ones	M
			y) The system must have an easy-to-use report generator facility, which must allow for the generation of ad-hoc reports.	M
			z) The system must provide a facility to convert an ad-hoc report into a standard report.	M
4.0	**Bulk Data Analysis**			
			aa) The system should be able to capture and analyse unstructured data (for example emails, memos, newspapers.)	M
			bb) The system should be able to visualise the output from unstructured data analysis (for example using pie charts, graphs.)	M
			cc) The system should be able to plug into the social media, extract the data and provide capability to analyse it	M
5.0		**Data Mining**		
			dd) The system should have data mining algorithms that will help the authority to come up with data mining models to analyse Taxpayers, detect fraud, and other business challenges	M
			ee) The system should provide for data mining models (for example, Fraud Detection, trend analysis, Taxpayer Churn) that can be used by different departments.	M
			ff) The data mining model(s) must be able to predict outcomes better than chance. i.e. the model must find some useful input variables that it can use to predict future events. (for example, macro analysis to provide assurance on tax take (expected tax) and to inform tax policy formulators)	M
			gg) The system should be able to perform analysis to shape forthcoming annual plans	M

Table 8: Non Functional User Requirements

No	System Features	M/D
	a) The system must have a graphical user interface and be menu driven with screen navigation and screen selection facilities.	M
	b) Only officers working on a particular case or analysis should be authorised to access it at different levels based on defined privileges.	M
	c) The system must be able to support multiple concurrent users.	M
	d) The system must be accessible from multiple locations seamlessly with centralized processing and a centralized database.	M
	e) The system must have a facility to archive data after a predefined period. A facility to enable enquiry on archived data and retrieval of archived records should be made available.	M
	f) The system must maintain up to at least six years (configurable) prior history on-line data before archiving.	M
	g) The system must have an online help facility for each user screen.	M
	h) The system must have a multi-level approach to system security based on the recognized audit controls of 'authentication' and 'verification', including: i. **Level 1** – Unique usernames and passwords should be required for all users at the operating system or domain level for example, Windows user accounts. ii. **Level 2** – Once the user has gained authorized and authenticated access to the operating system, unique usernames and passwords should be required for all users at the application level. iii. **Level 3** – Once the user has gained authorized and authenticated access to the application, there should be configurable access to specific modules/transactions within the application. iv. **Level 4** – The final level of secure access control is based on implementing configurable parameters to control the user's actions once access has been gained to authorised modules or transactions.	M
	i) Where the system enables external users for transactions the system should ensure that ZRA data cannot be compromised.	M
	j) The system must have the facility to control user/group access at user-defined levels/functions including the following: <u>Functions</u> i. User id/password ii. Systems Administration and Control iii. Application/Module/Sub-module iv. Forms (screens) and Reports v. Buttons/Functions	M
	k) The system must have a facility to log out the user after specified time of inactivity.	M
	l) The system must have end to end application encryption.	M
	m) System passwords must be user-defined based on the ZRA ICT password policy, for example, i. Complex and alpha numeric ii. Minimum 8 characters in length	M

	iii. Held in an encrypted form.	
	n) The system must require periodic password changes at regular intervals defined by management.	M
	o) The system must lock account after user-defined unsuccessful attempts.	M
	p) The system must provide for a clear audit trail for all transactions with a time and user account stamp. Authorized staff/roles must be able to view audit trail logs online or print hardcopy.	M
	q) The system must support automatic archiving of audit trail logs per user-configurable intervals. Audit trail logs must be read-only and security-enabled to prevent unauthorized access or modifications.	M
	r) The system must have facility for output/reports to be directed to either printer, screen or file and reports must be user-defined characteristics including: i. A title/ description ii. Detail the period/ date iii. Page numbering iv. End of report message v. Nil report message where appropriate vi. Default spooling where output size exceeds user-defined limit. vii. No limit on number of print queues	M
	s) The system must seamlessly integrate all modules/functions within the proposed solution, both on a one-to-one basis and also on a one-to-many.	M
	t) The system must have complete package/module integration running under a recognized, Open Standards compliant hardware independent operating system (to be determined by the application requirements).	M
	u) The system must have the facility to print any information displayed on a screen.	M
	v) The system must have the provision of on-line enquiry/interrogation for transaction level record retrieval.	M
	w) The system must properly display, calculate and transmit date data, including, but not restricted to 21st Century dates.	M
	x) The system must meet the Minimum tolerable Mean Time between Failures (MTBF) in any one calendar year to achieve an uptime of 98%.	M
	y) The system suppliers must provide the users and ZRA with copies of all documentation pertaining to the software including user manuals, technical specifications.	M
	z) All the software and hardware and associated utilities installed must be well documented.	M
	aa) The system suppliers must commit to providing on-going technical support for the first 2 years.	M
	bb) The system suppliers must test all the system functions before implementation.	M
	cc) The system suppliers must correct any system error that is detected in the course of testing and thereafter re-test again as necessary.	M
	dd) The system must be able to handle run-time errors correctly and give agreed or defined error messages to the users (i.e. ZRA users).	M

	ee) The system must be thoroughly tested to ensure user acceptance with production volume workloads in a production environment to determine overall performance in terms of throughput, response time, run-time and resource utilization.	M
	ff) The system must be tested for scalability.	M
	gg) The system must be tested to ascertain that interfaces are functional and correct by uploading/downloading data from interfaced systems.	M
	hh) The system must be flexible enough to accommodate any future changes and/or enhancements that may arise.	D
	ii) The system must provide for a development, testing, quality assurance, training, and disaster recovery and production environments.	M
	jj) The Supplier must provide ZRA with the source code	M
	kk) The Supplier must ensure end users are adequately trained in the use of the system as well as technical training for system administrators and second level support staff.	M
	ll) The system should be accessible through the web browser and other tools like MS Excel	M
	mm) The system should provide for all reports to be exported in different formats (for example PDF, Word, Excel)	M

3.4.6 Fraud Detection Use Case

This use cases in Figure 14 depict the activities that a user will be able to carry out on the ZRA Fraud detection application.

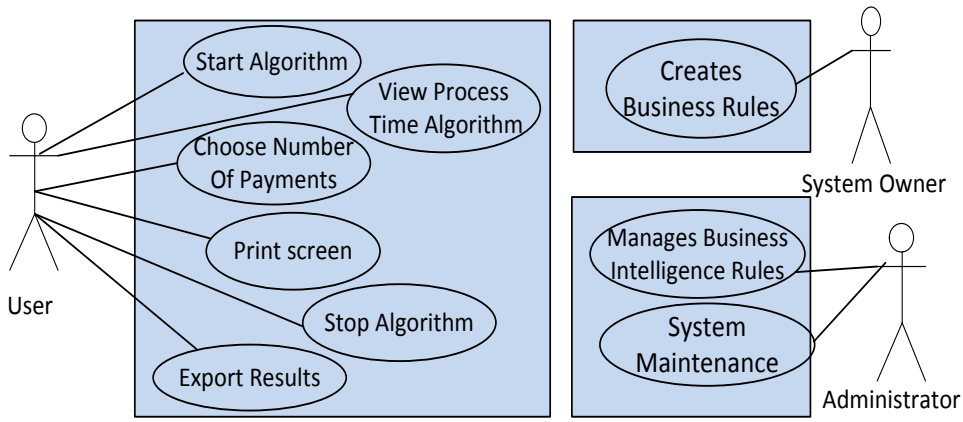

Figure 14: Fraud Detection Use Case Diagram

i. User Use Case

User will be able to view the process time, Start the algorithm, Stop the algorithm, Print the Screen and export the results into a different file format.

ii. Administrator Use Case

This use cases describe the activities an administrator will be able to carry out

on the ZRA Fraud detection application such as; manage the business rules and also carry out system maintenance.

iii. **System Owner Use Case**

This use cases describe the activities which the system owner will be able to carry out on the ZRA Fraud detection application. The system owner will be able to create business rules that will be used to measure against the behavior of Taxpayers.

3.4.7 System Modelling

The ZRA Fraud detection model was implemented based on a three-tier architecture design. The design segments an application's components into three tiers of services in Figure 15, namely Presentation, Logic and lastly Data access. This architecture has been used before as shown in the literature [120]

i. **Presentation**

The presentation tier, also called user services layer, gives a user access to the application. This layer presents data to the user. It also permits data manipulation and data entry such as the Outlier Algorithm Detection setup where a desired number of records to be analysed can be defined. The start and stop function and also the visualisation function of the application.

ii. **Logic**

The middle tier, also known as business services layer, consists of business and data rules. The two algorithms implemented in this design (Continuous Monitoring of Distance Based and Distance Based Outlier Queries) exists at this Tier.

iii. **Data Access**

The data tier, or data services layer, implements the data storage in this case the zra_bi database which contains the Tax Payer Profile and the payments activities.

3.4.8 Definition of Class for ZRA Fraud detection Application

i. **ZRAFDMain Frame**

This is the main frame of the window of the ZRA Fraud Detection Application.

ii. MainTab Panel

This tab holds the Outlier Algorithm User functionality

iii. SetupTab

This tab allows a desired number of payments to be setup for the purpose of analysing the records.

iv. VisualiserTab

This tab will enable the algorthms to be run and thereafter will be able to show the visual results

v. ZRAFD Controller

The Controller will get input from the presentation, for example the number of records set (1000). This then gets data from the data access based on the input and presents it to the presentation layer.

vi. Continuous Monitoring Distance Based (CMDB)

This is the class that implements the continuous monitoring of distance based algorithm

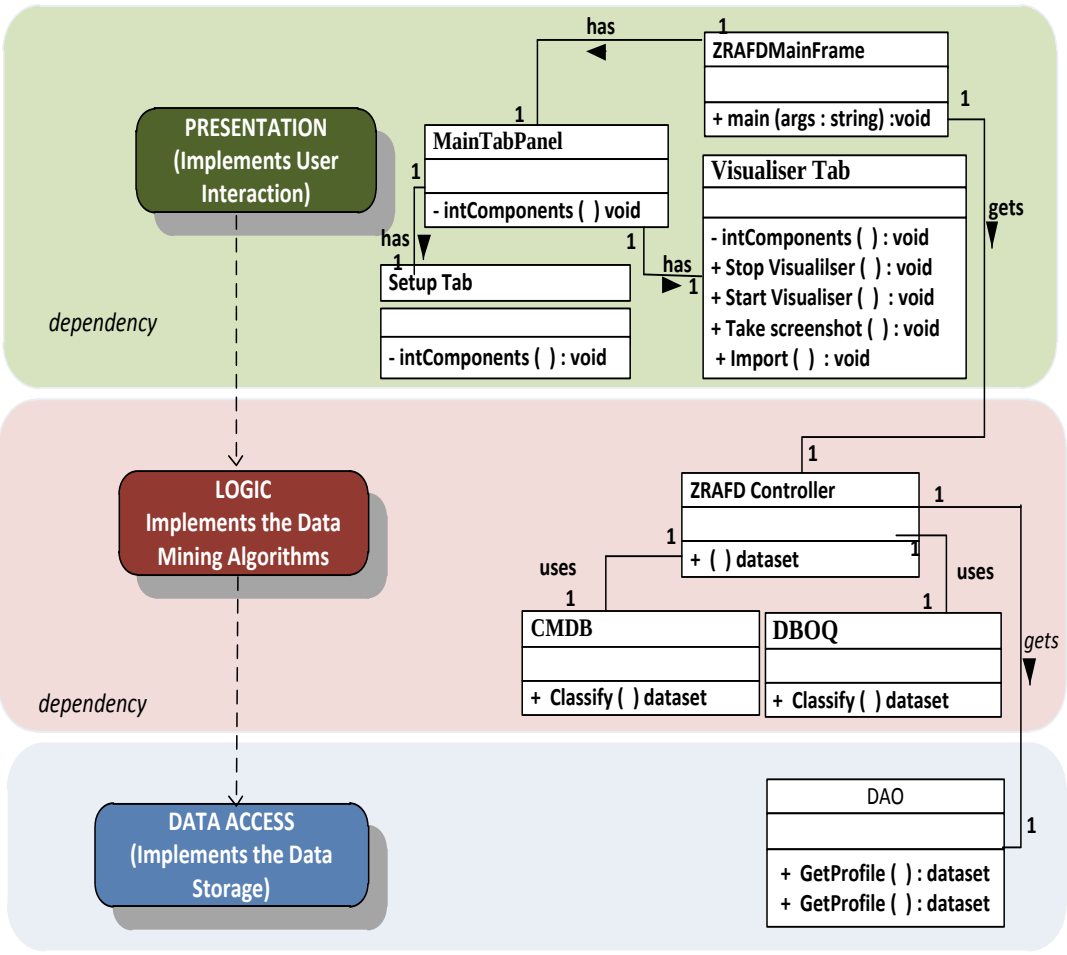

Figure 15: Three Tier Architecture, Class diagram Fraud detection System *[120]*

vii. **Distance Based Outlier Query**

This class implements the distance based Outlier Queries

viii. **Data Access**

This will return data from MySQL database

3.4.9 ZRA Fraud Detection Entity Relationship Diagram (ERD).

An Entity Relationship Diagram (ERD) is a graphical representation of an information system which shows the various entities and their attributes. Figure 16 shows the Entity Relationship Diagram (ERD) of the ZRA Fraud Detection comprising of the profiles and the payment entities and their attributes.

3.4.10 Profile and Payment Entities

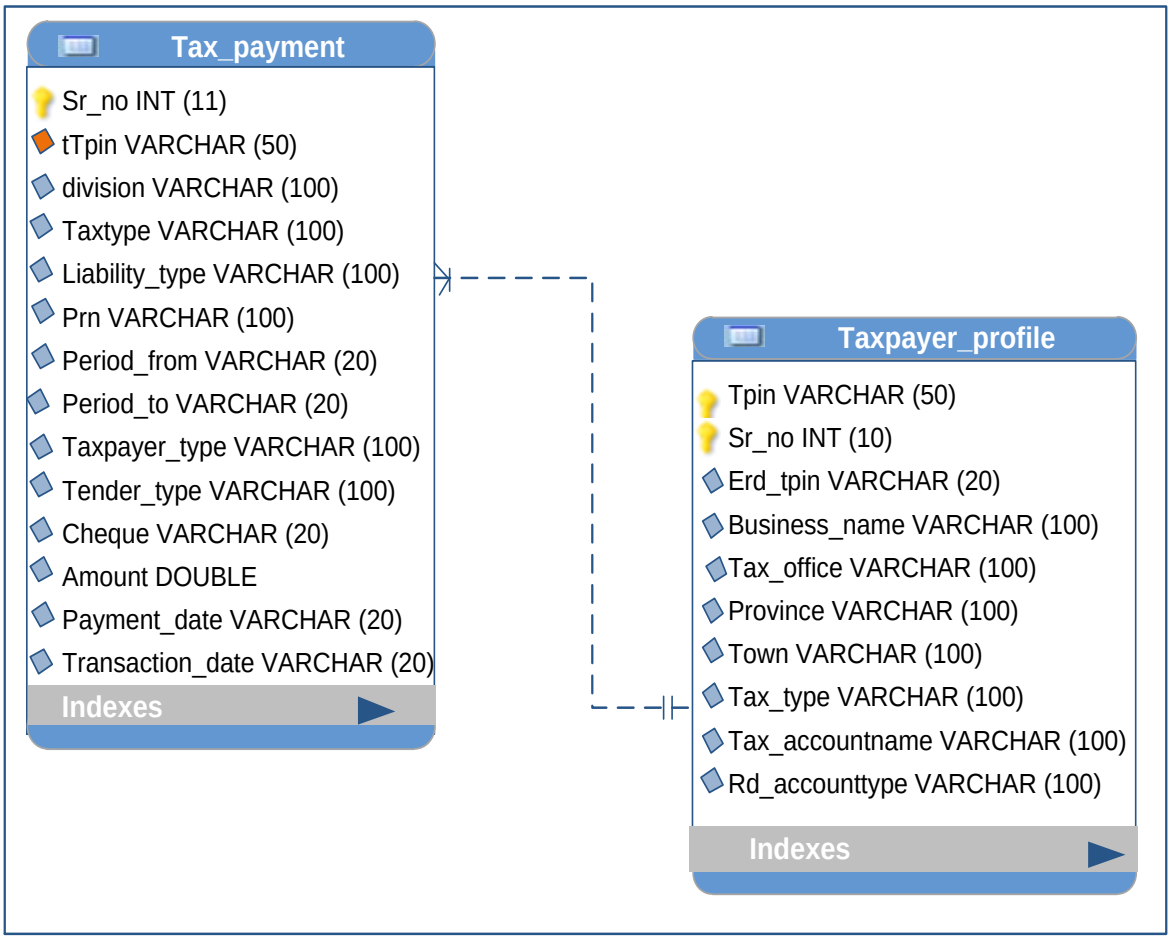

Figure 16: Entity Relationship, ZRA Fraud Detection

3.4.11 Database and Tables

A database zra_bi in table 9 and table 10 was created using MySQL Workbench 6.3

running on Port: 3306 with Login User: root. Two (2) tables namely; tax_payment and taxpayer_profile were created for this study.

Table 9: Structure, TAX PAYMENT Table

TAX PAYMENT						
SR. NO	COLUMN NAME	DATE TYPE	LENGTH	KEY	NULL	COMMENTS
1	tpin	varchar	50		Y	MANUAL
2	division	varchar	100		N	MANUAL
3	taxtype	varchar	100			MANUAL
4	liability_type	varchar	100			MANUAL
5	prn	varchar	100			MANUAL
6	period_from	varchar	20			MANUAL
7	period_to	varchar	20			MANUAL
8	taxpayer_type	varchar	100			MANUAL
9	tender_type	varchar	100			MANUAL
10	cheque	varchar	20			MANUAL
11	amount	double				MANUAL
12	payment_date	varchar	20			MANUAL
13	transaction_date	varchar	20			MANUAL

Table 10: Structure, TAX PAYMENT table

TAXPAYER PROFILE						
SR. No.	COLUMN NAME	DATE TYPE	LENGTH	KEY	NULL	COMMENTS
1	tpin	varchar	50	Y	N	
2	sr_no	varchar	10	Y	N	
3	erd_tpin	varchar	100			MANUAL
4	business_name	varchar	100			MANUAL
5	tax_office	varchar	100			MANUAL
6	province	varchar	20			MANUAL
7	town	varchar	20			MANUAL
8	tax_type	varchar	100			MANUAL
9	tax_accountname	varchar				MANUAL
10	erd_accounttype	varchar				MANUAL

3.4.12Sequence Diagram

The Sequence Diagram in Figure 17 shows a detailed flow for data mining sequence of events showing how fraud is detected on bulk tax data. The diagram shows the sequence of events and interactions for a Fraud detection system.

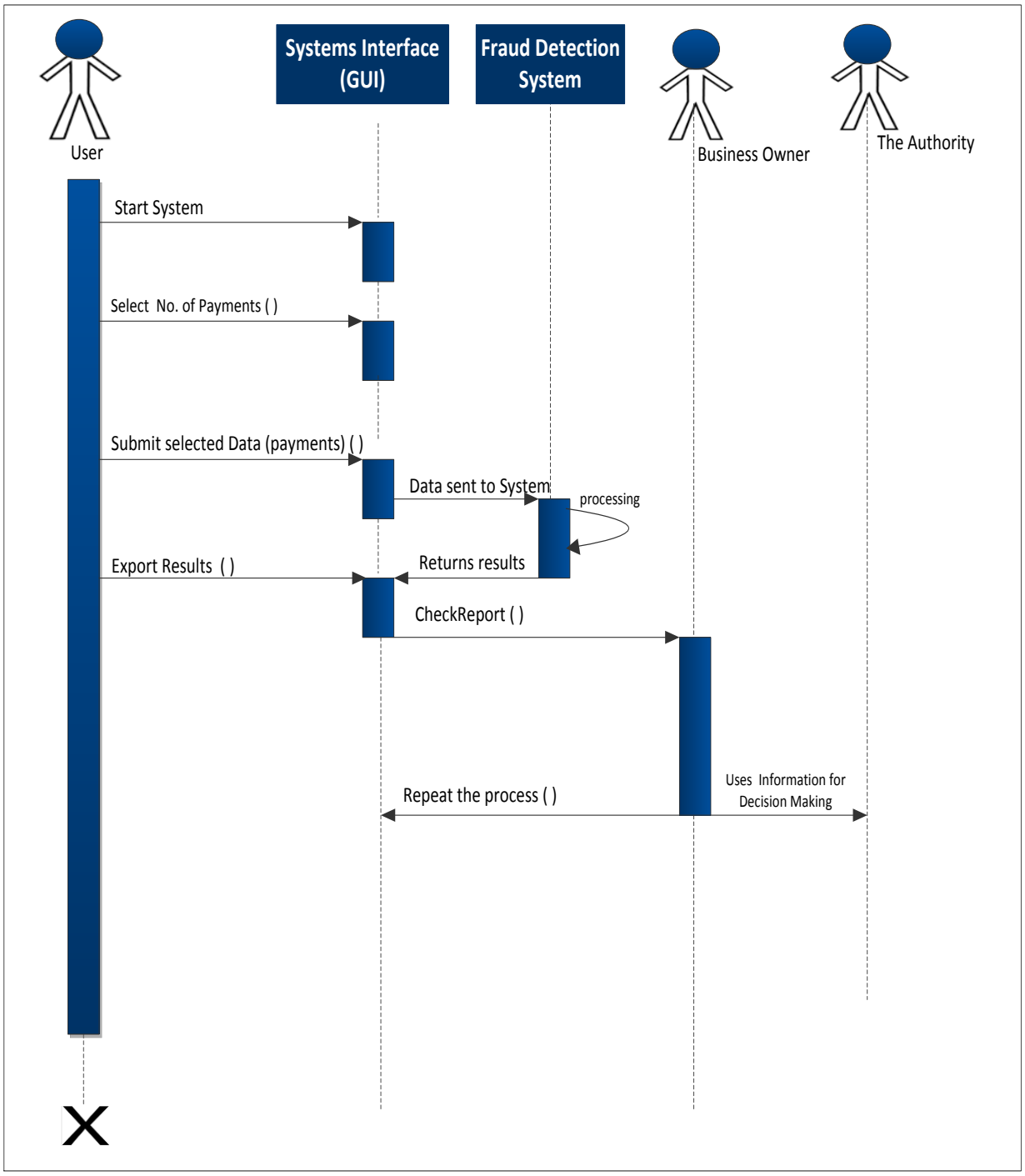

Figure 17: Fraud Detection Sequence Diagram

3.4.13Deployment Diagram

The deployment diagram in Figure 18 shows how a system will be physical deployed in the hardware environment and the purpose is to show the connection between the components and how they will physically communicate with each other. A node will represent either a physical machine or a virtual machine. For this study, two Laptops were used running each windows 64 bit Operations System, One was used as a backup and the other one was used both as an application and Database server. Further Data was extracted from the existing Data sources.

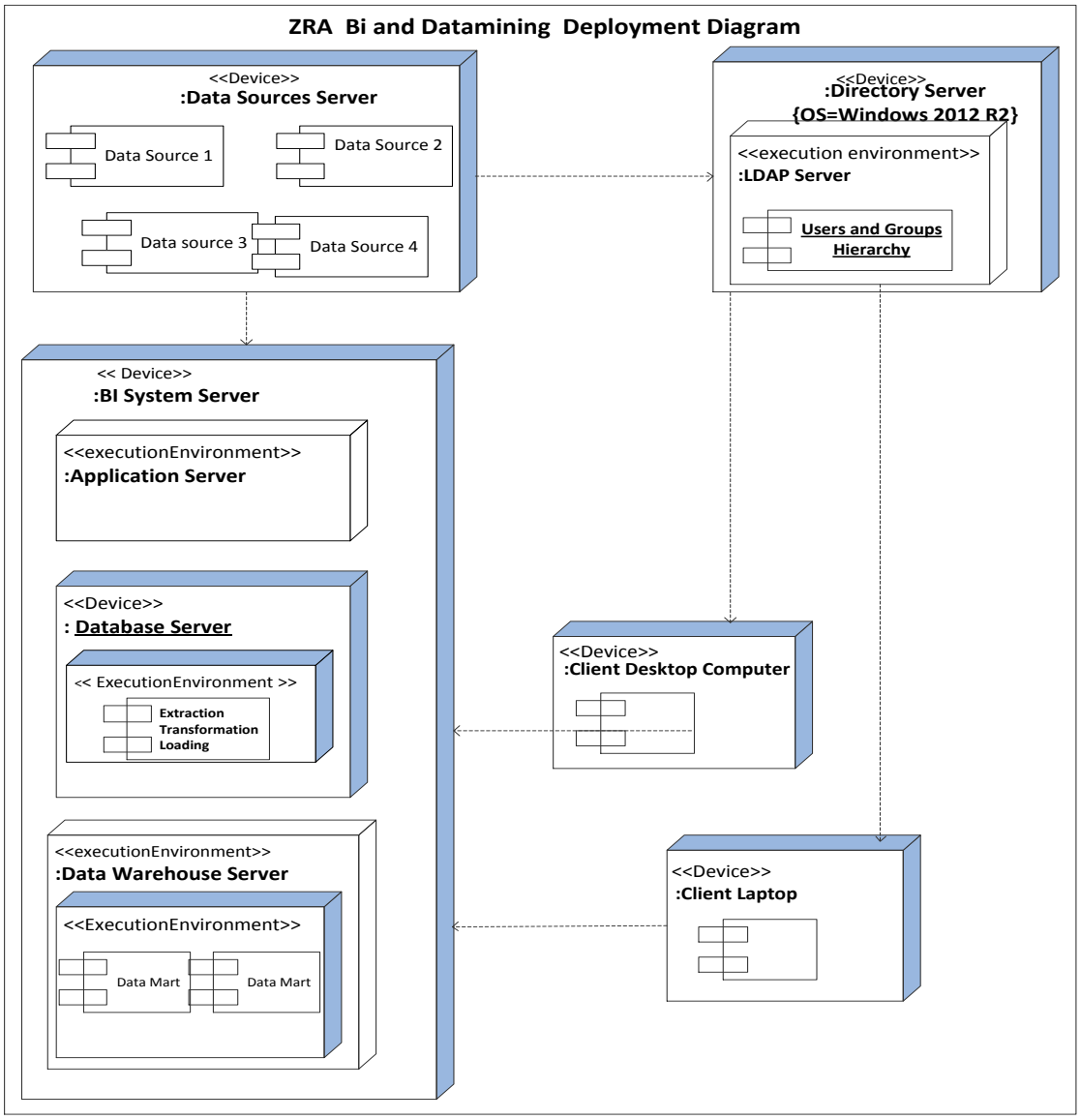

Figure 18: Deployment Diagram

3.5 Summary

This chapter provides firstly definition of the research methodology of this study giving details of the sample selection and a description of the procedure used in designing the instrument as well as the collection of the data and the procedure.

Further, details of what was done in order to design the model which was later used to come up with the prototype for the Fraud detection on bulk Data is given. The need to understand the business process of the Tax Administration was emphasised and this was provided diagrammatically. For the purpose of implementing the tool, the diagrams were also used to depict the flow of events of the system. The chapter ended with the technology description of the implemented solution as well as the structure of the system.

CHAPTER FOUR

RESULTS

4.1 Introduction

The results in this section looks at the outcome of the survey based on the questionnaires and the oral interviews. The results are organised according to various and specific areas that the study set out to answer.

Further, this section looks at the implementation of a ZRA Fraud detection tool. The ZRA Fraud detector was developed in java using weka java library (NetBeans Integrated Development Environment). Weka java library implements numerous Data mining algorithms. The NetBeans IDE used is a free, open-source, cross-platform integrated development environment (IDE) with software development built-in support for Java Programming Language.

4.2 Baseline Line Study

The results in this section looks at the outcome of the survey based on the questionnaires and the oral interviews. The results are organised according to various and specific areas that the study set out to answer.

4.2.1 Implementation of BI, data mining and in ZRA

To fully understand whether ZRA has implemented BI Technologies (Data Mining and Data Warehouse), a questionnaire was distributed to employees of ZRA in the operating Divisions and also the supporting Department.

Figure 19 shows that ZRA had not implemented Business Intelligence and Data mining. More than half of the respondents as indicated in the results show that 110 out of 156 respondents, representing (71%) indicated that BI and Data mining has not yet been implemented in ZRA. 46 out of 156 respondents, representing (29%) indicated that

BI and Data mining has been implemented in ZRA.

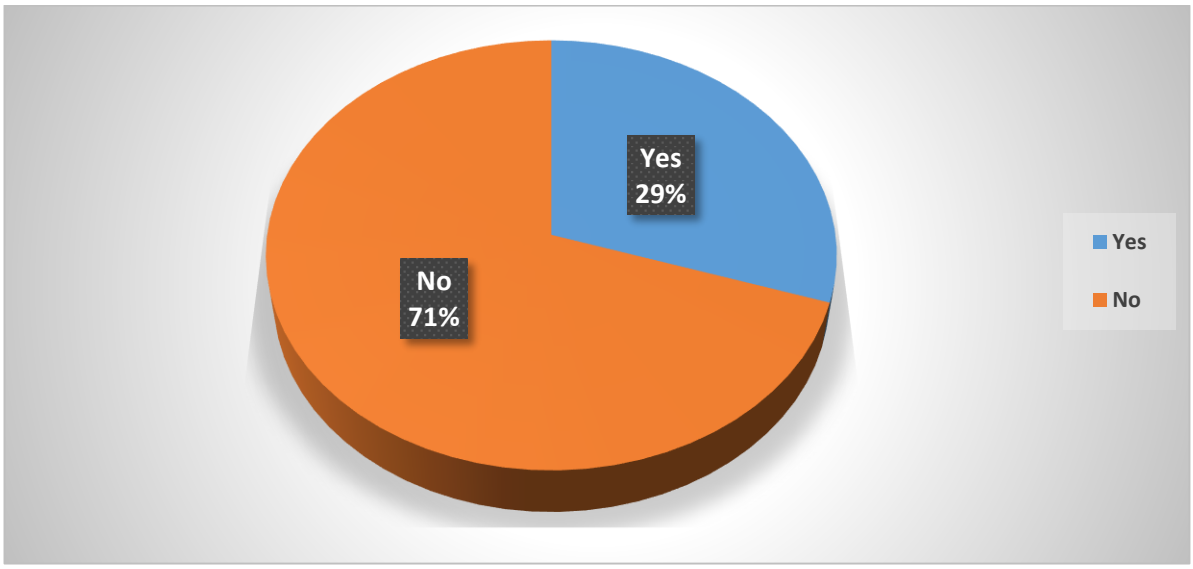

Figure 19: Implementation of Business Intelligence in ZRA

4.2.2 Tools used to analyse data currently, in ZRA

A greater majority of the respondents as shown in Figure 20, indicated that the data was currently analysed and turned into something that can be interpreted using Spreadsheet Analysis, 78%. There were 11% that indicated software that allow own specific queries was used, 4% indicated that software that takes raw data and creates visualization was used.

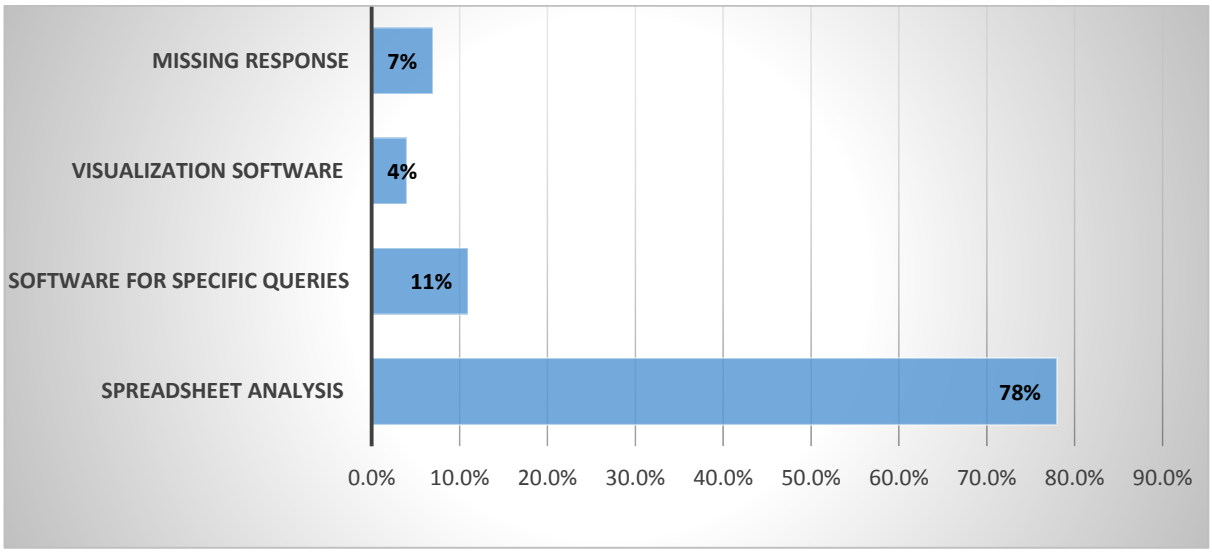

Figure 20: How Data is currently analysed.

4.2.3 Current Data Management challenges in ZRA.

The most challenge faced by the institution in terms of data management, according to the respondents, was data sources not being integrated. There were 65/156 (42%) End

users that cited this challenge, whereas, 50/156 (32%) cited data cleaning from Legacy systems not easy, 30/156 (19%) cited data not recorded properly, and less than 6/156 (4%) cited data not stored properly, 3/156 (2%) too many data sources, and 2/156 (1%) rampant data growth. Lack of interface with outside data sources, poor quality data, and inflexible system were some of the other challenges reportedly facing the institution in terms of the provision of the reports.

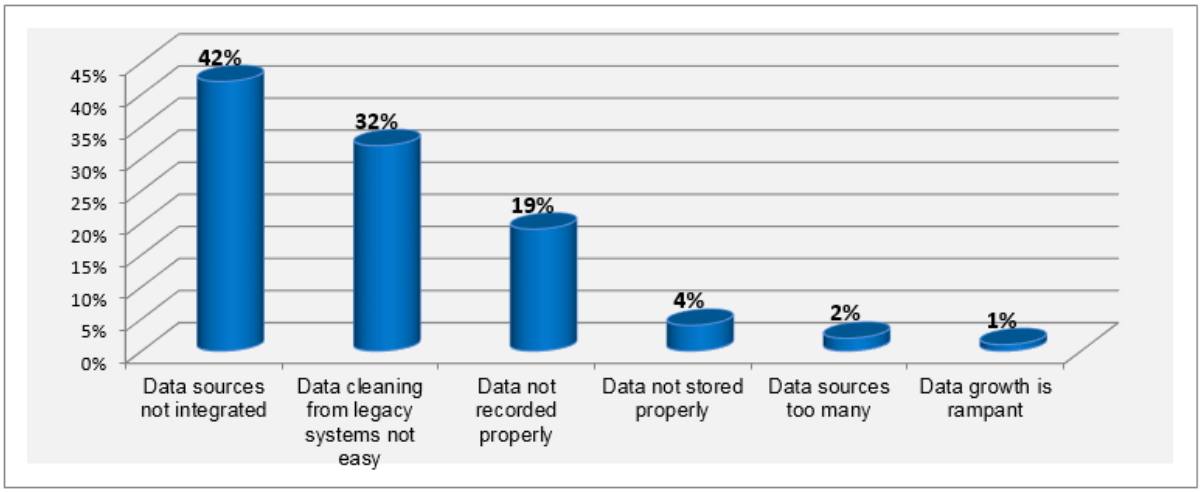

Figure 21: Data Management Challenges

4.2.4 Challenges with current reporting.

When asked to describe challenges facing their current reporting environment, 62/156 (40%) indicated that current reports are not interactive and do not provide enough information for analysis, 60/156 (38%) indicated too much time was spent preparing reports rather than analyzing the data, 20/156 (13%) indicated people outside IT want to create their own reports but are not able to, 2/156 (1%) indicated Excel spreadsheet is not too flexible to an extent, and 0/156 (0%) indicated IT develops all reports and the backlog is too long. System used to analyse reports being manual is 12/156 (8%).

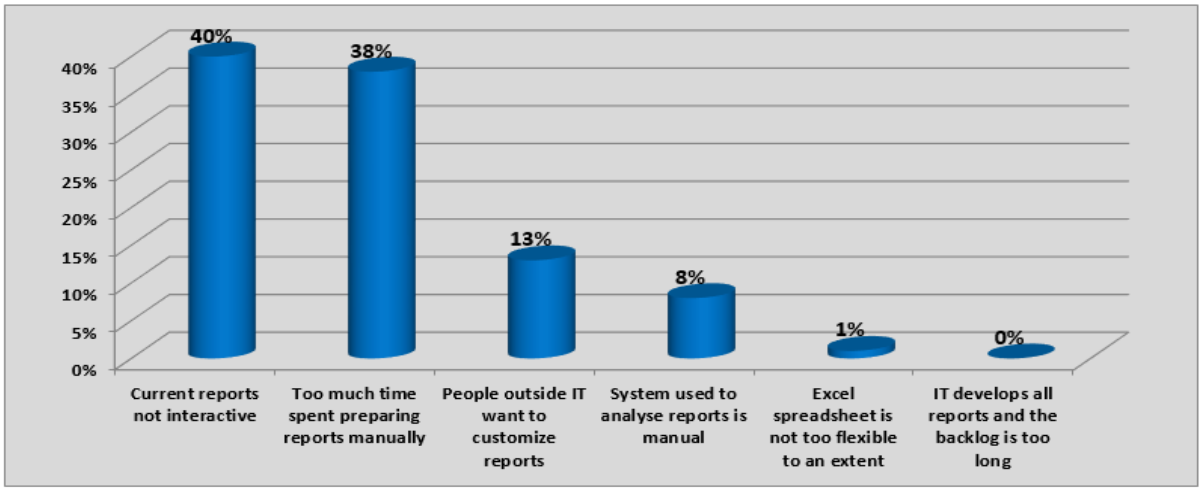

Figure 22: Challenges with current reporting

4.2.5 Methods used to detect fraud on Taxes

Majority of respondents indicated that they detected fraud mostly through targeted and random audits, 90/156 (58%). There were only 0/156 (0%) respondents that indicated they detected fraud through data mining techniques (Figure 23). Using informants to detect fraud was indicated by 30/156 (19%) respondents and using under-cover operations by 36/156 (23%).

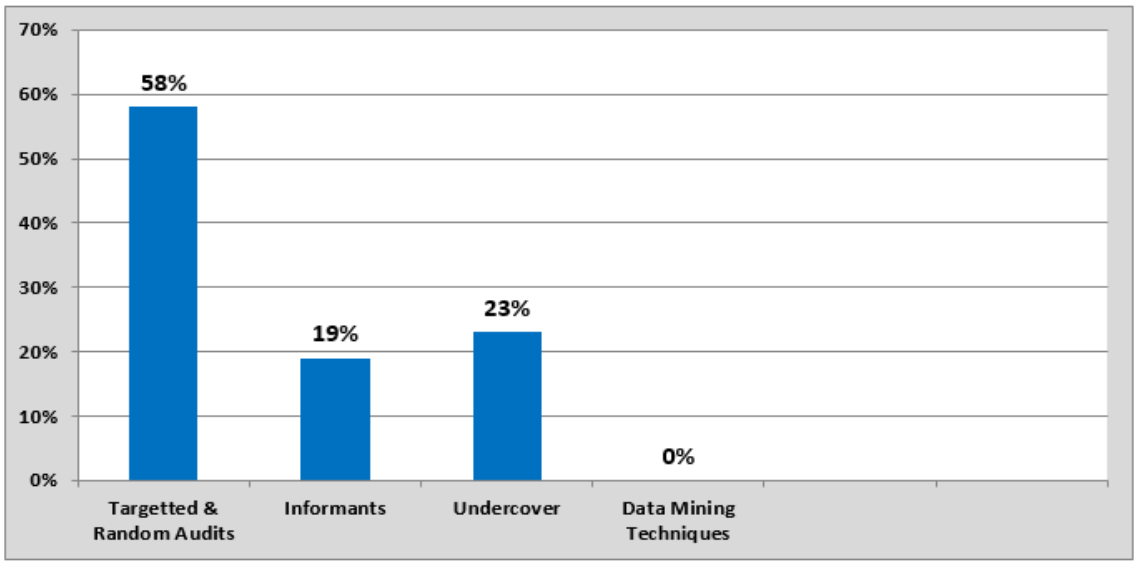

Figure 23: Fraud detection Methods in ZRA

4.2.6 What users want to mine or explore on the complex and large data

Further, 120/156 (77%) of the respondents indicated they would want to mine or explore on this complex and large data in order to detect frauds as shown in Figure 24. 15/156 (10%) indicated "Trends on compliance" whilst 10/156 (6%) indicated

"Patterns on Filing and Non-Filings". Furthermore, about 10/156 (6%) indicated they would want to mine "Underpayments" and 0/156 (0%) wanted to mine "Relationships between groups and 1/156 (1%) "Pattern on Registration".

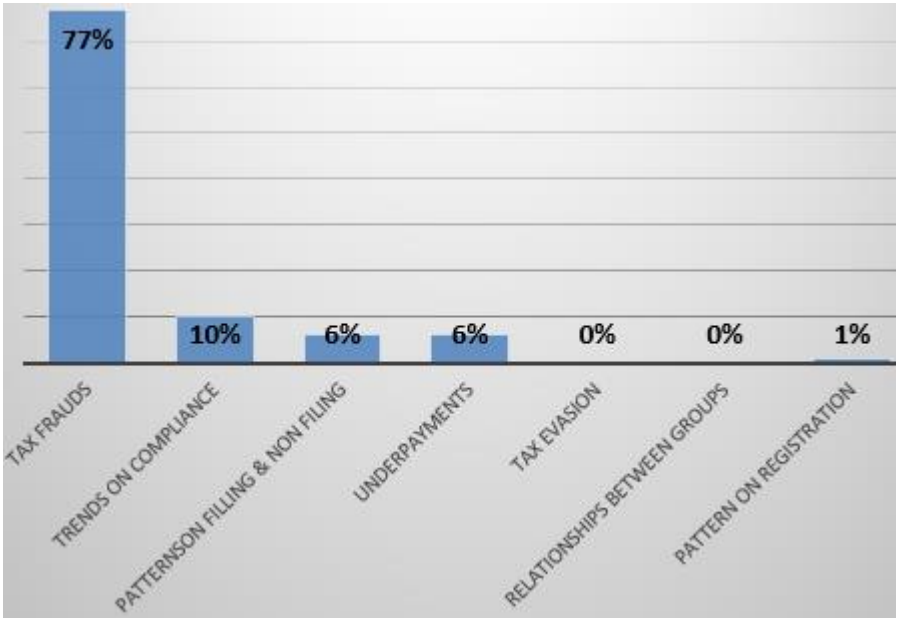

Figure 24: What to mine on the bulk Data

4.2.7 Speed of detecting fraud on tax data in the Zambia's Tax Administration Sector.

Based on the survey results in Figure 26, 140 out of 156 respondents, representing 90% indicate that it takes long, above 7 days to detect fraud on tax data using the traditional methods whilst 16 out of 156 respondents, representing 10% indicates that it takes below 7 days to detect fraud on tax data using the traditional methods.

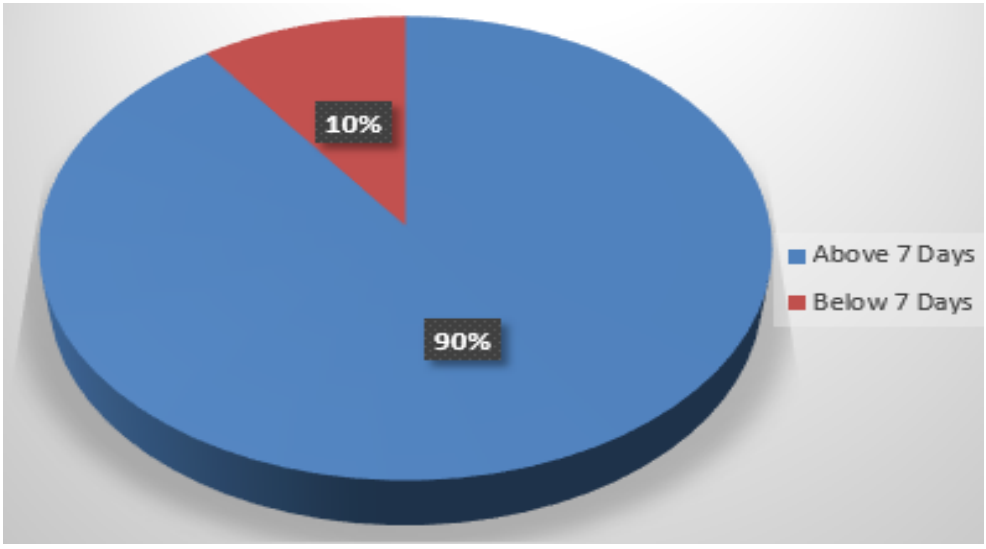

Figure 25: Speed of detecting Fraud on Bulk Data

4.3 System Implementation

This section outlines the system design for the ZRA Fraud detection tool and gives a description of how it works. The section also describes the hardware and the software requirements for this implementation.

4.3.1 Technology Description

Implementation of the ZRA Fraud detection tool will consist of two (2) Laptops were used and both of them were running on Intel (R) Core (TM) 2 Duo CPU 2.80 GHz. Both Laptops were using a minimum of 4GB of RAM and 500 GB Hard Disk Drive.
The Application and the Database were both setup on the first Laptop to run the Application software and to run the MySQL Database system for storage of data in the database formats.
The other Laptop was used as a Backup in case of any failures.
This backup also contained all configurations for application and database similar to that in the first Laptop. Both laptops were running Windows 7 Professional with Service Park 1 64-bit operating systems.
The ZRA Fraud detector was developed in java using weka Java library. Weka java library implements numerous data mining algorithms. NetBeans IDE was used for development. NetBeans IDE is a free, open-source, cross-platform integrated development environment (IDE) with software development built-in support for Java Programming Language.

The Database, zra_bi was created in MySQL and workbench 6.3 CE was used to interact with the Database. MySQL Workbench is a unified visual tool for database architects and developers. MySQL Workbench was use to provide SQL Development, and comprehensive administration tools for server configuration and, user administration.

4.3.2 System Structure.

The system has four packages, namely; unza.logic, unza.presentation, unza.data access and unza.utils as shown below.

i. Data Access Object class (DAO).

This DAO class accesses the database zra_bi using root as username and root as password and gets the payments or financial records from the table, tax_payment. The details such as sr_no, tpin, division, and amount will be processed and in cases where the password for Mysql has changed, such changes may also be effected in the DAO. Class.

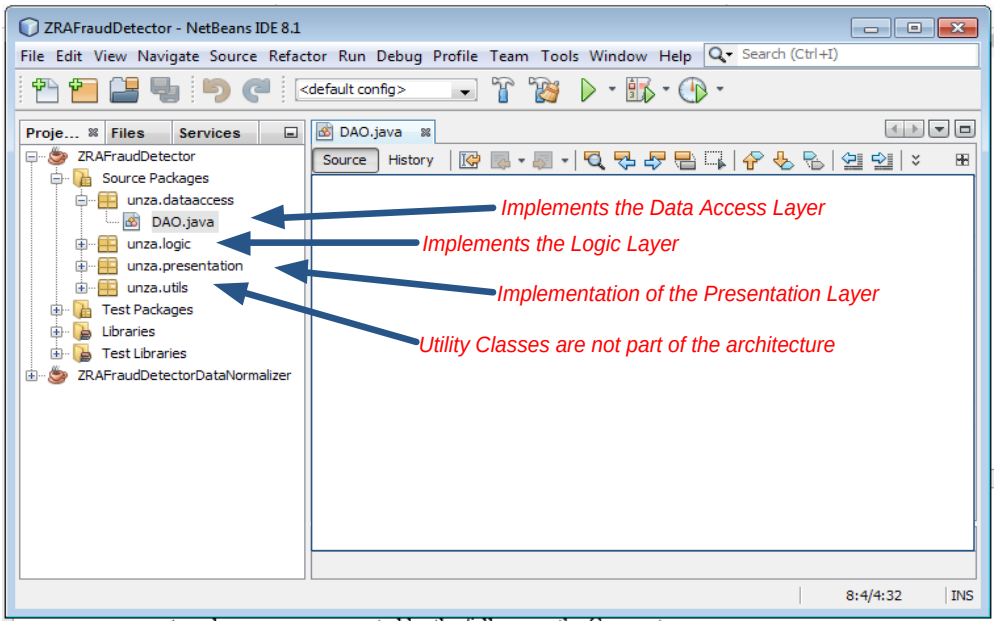

Figure 26: System Structure

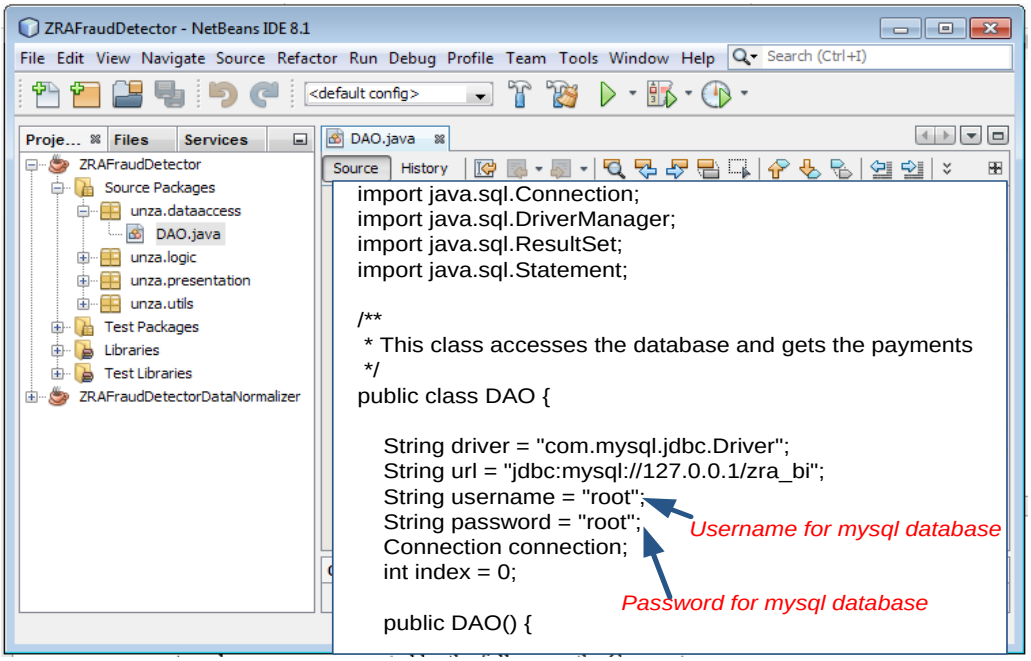

Figure 27: Data Access Object class information

ii. ZRAFD tool Classes being used.

Classes appear under packages when expanded.

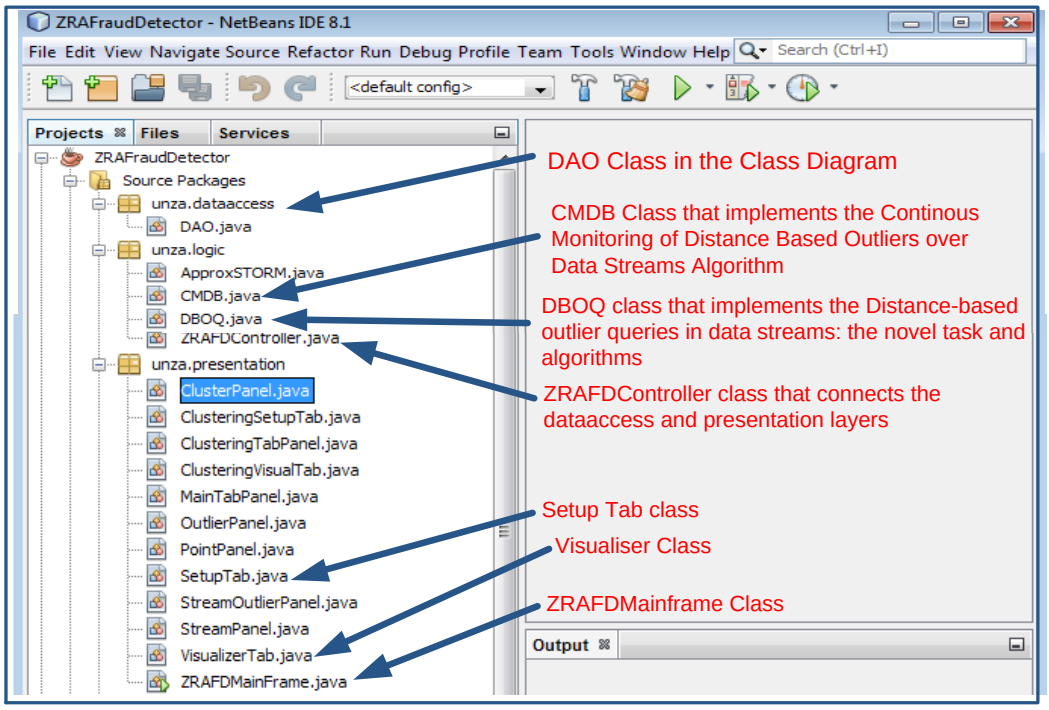

Figure 28: ZRAFraud detection Visualisation Tab

iii. Libraries used by ZRAFraudDetector

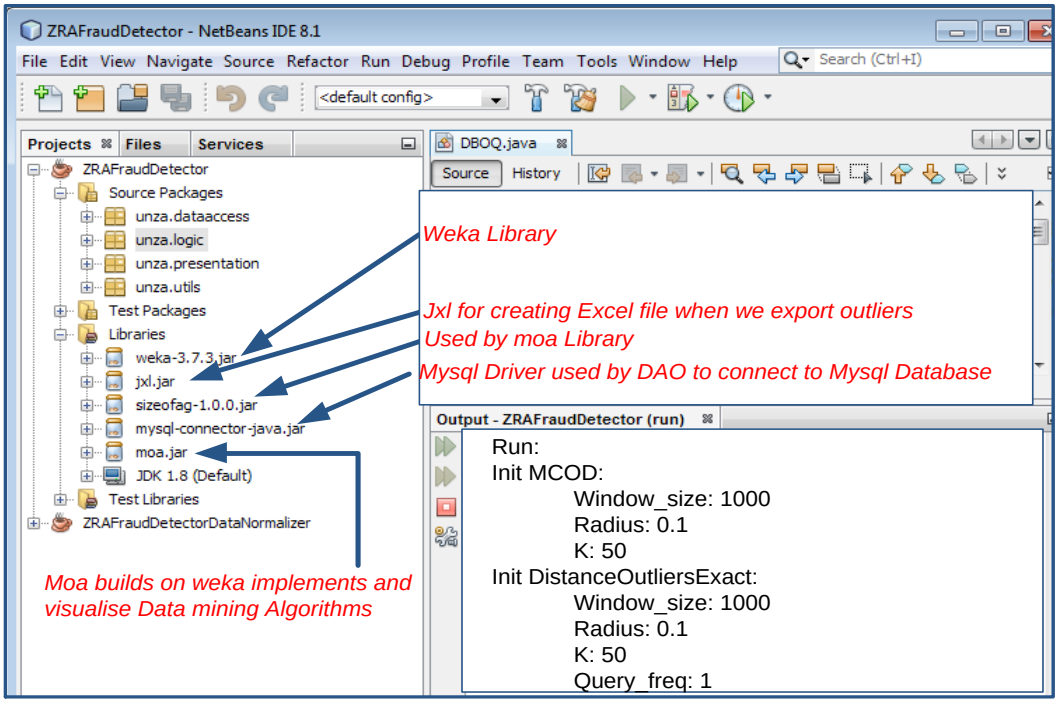

Figure 29: ZRA Fraud detection Libraries

In this ZRAFraudDetector prototype, the Distance Based Outlier Data point o is a

distance-based outlier if less than K neighbors in a window lie within distance R from o.

K = 50;

R = 0.1

We have three sliding windows as indicate;

LT – Large Taxpayer

MT – Medium Taxpayer

ST – Small TaxPayer

iv. Continuous monitoring of distance-based code.

CMDB is the main code that implements the Continuous monitoring of distance-based outliers over data streams. For this research, ZRA Fraud detection tool was designed based on the code written by Kontaki et al [62].

```
public class CMDB extends MCODBase {
    public FloatOption radiusOption = new FloatOption("radius", 'r', "Search
radius.", 0.1);
    public IntOption kOption = new IntOption("k", 't', "Parameter k.", 50);

    public CMDB()
    {
        // System.out.println("MCOD: created");
    }

    @Override
    public void Init() {
        super.Init();

        m_WindowSize = windowSizeOption.getValue();
        m_radius = radiusOption.getValue();
        m_k = kOption.getValue();

        Println("Init MCOD:");
        Println("   window_size: " + m_WindowSize);
        Println("   radius: " + m_radius);
        Println("   k: " + m_k);
```

Figure 30: Code for Continuous Monitoring of Distance Based Outliers

v. Distance-Based Outlier Query code.

DBOQ is the main code that implements the algorithm presented in "Distance-based outlier queries in data streams: the novel task and algorithms [121].

```
public class DBOQ extends STORMBase {

    public class ISBNodeExact extends ISBNode {

        public int count_after;
        // nn_before:
        //   A list that needs O(login) time for ordered insertion and search.
        //   It must be able to perform a search in the list using e.g. <=.
        private ArrayList<Long> nn_before;

        public ISBNodeExact(Instance inst, StreamObj obj, Long id, int k) {
            super(inst, obj, id);
            m_k = k;
            count_after = 0;
            nn_before = new ArrayList<Long>();
```

Figure 31: Code, Distance Based Outlier

vi. ZRA FD Main frame class

This is the main class of the System that creates the "ZRA-FD Outlier Graphical User Interface" Using this interface, we are able to start and stop the algorithm, we are able to export the outliers from this interface set the number of payment and also create the visualization as shown in Figure 32

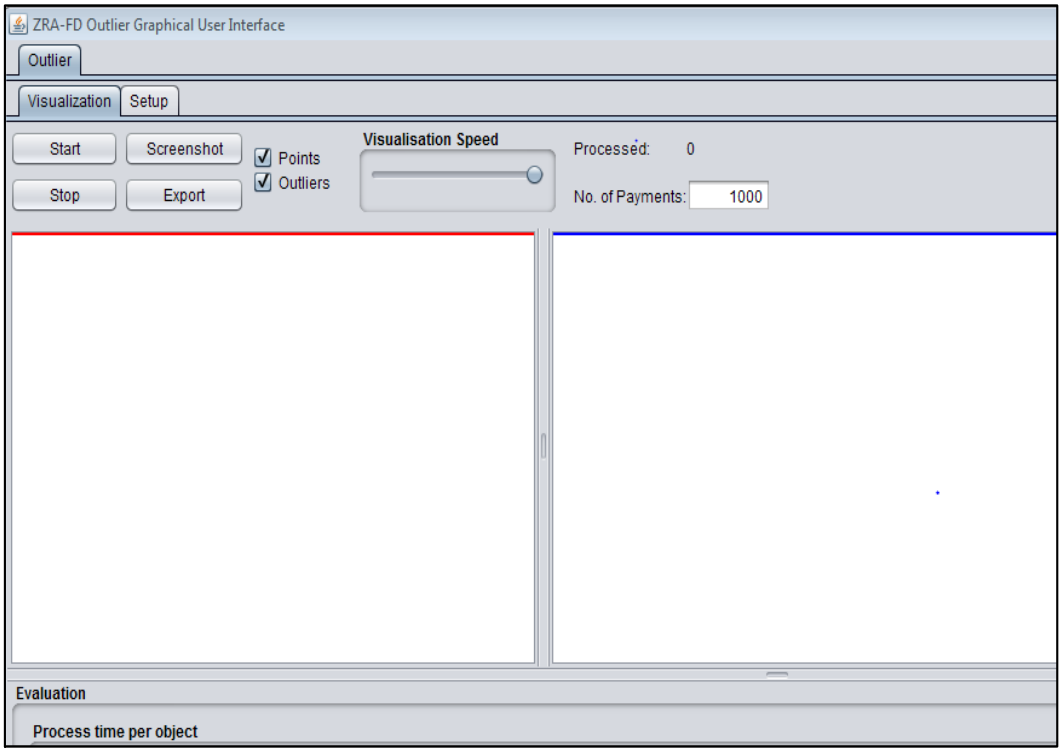

Figure 32: ZRA Main Frame

vii. ZRAFD Outlier graphical User Interface code

```
package unza.presentation;
import java.awt.BorderLayout;
import java.awt.Dimension;
import java.awt.Insets;
import java.awt.Toolkit;
import javax.swing.JFrame;
import javax.swing.JPanel;
public class ZRAFDMainFrame extends JFrame {
   private static final long serialVersionUID = 1L;
   private javax.swing.JTabbedPane panel;
   public ZRAFDMainFrame() {
      setTitle("ZRA-FD Outlier Graphical User Interface");
      initGUI();
      Dimension screenSize = Toolkit.getDefaultToolkit().getScreenSize();
```

Figure 33: Code showing the Main Frame

4.3.3 System detection of outliers

The implementation of the system shows values that "lie outside" the normal region (much smaller or larger than) than the values in a set of tax data payment range which may be due to inconsistency in the payment of taxes, errors on filling of the tax returns or indeed as a result of fraud. The business rules are defined as follows;

i. For Large Tax payer

If Payment is LT and that payment is less than a minimum set for LT or greater than the maximum set for LT, then such a payment is flagged as an outlier, which is data laying outside the normal region.

ii. Medium tax payer

If Payment is MT and that payment is less than a minimum set for MT or greater than the maximum set for MT, then such a payment is flagged as an outlier which is data laying outside the normal region.

iii. Small tax payer

If Payment is ST and that payment is less than a minimum set for ST or greater than the maximum set for ST, then such a payment is flagged as an outlier which is data laying outside the normal region.

For this study, the business rules were set as follows;

Table 11: Business rules as set in the algorithm

Class	Range	Class	Range
LT Min	1,001,000	LT Max	200,000,000
MT Min	501,000	MT Max	1,000,000
ST Min	100	ST Max	500,000

Figure 40 shows a report of payments picked by the algorithm as outliers detected based on the business rules defined by the User, Table 11, Figure 35 and Figure 37. For example, data presented in record number 1 in Figure 40 and Figure 41 with TPIN 3113648373 belonging to class MTO with amount 421, 217.88 was picked as an outlier because it lies outside the normal region (MT Min 501 and MT Max 1,000,000). This record is then regarded as suspicious and is further subjected to scrutiny in order to establish the reasons for its class as depicted in Figure 40. Further, the system, using the implemented algorithm, Figure 37 shows that 1000 records of tax data is processed in a shorter period of time.

i. Setup Tab

This interface will enable the setup of the parameter of the algorithm for it to run. The screen shows the set minimum and the maximum payment for each range.

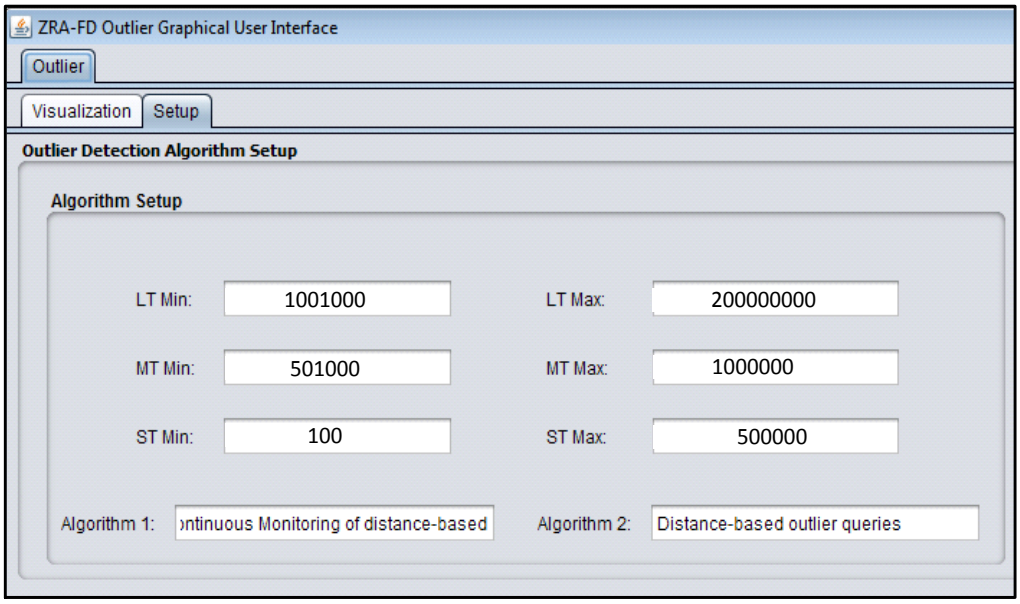

Figure 34: Setup Tab Screenshot

```
package unza.presentation;
import moa.clusterers.outliers.MyBaseOutlierDetector;
import moa.evaluation.MeasureCollection;
import moa.gui.outliertab.OutlierAlgoPanel;
import moa.gui.outliertab.OutlierEvalPanel;
import moa.gui.outliertab.TextViewerPanel;
import moa.streams.clustering.ClusteringStream;
public class SetupTab extends javax.swing.JPanel {

 private OutlierAlgoPanel outlierAlgoPanel0;
             private OutlierEvalPanel outlierEvalPanel1;
             private TextViewerPanel logPanel;
    public SetupTab() {
        initComponents();
        //outlierAlgoPanel0.renderAlgoPanel();
```

Figure 35: Code for the Setup Tab

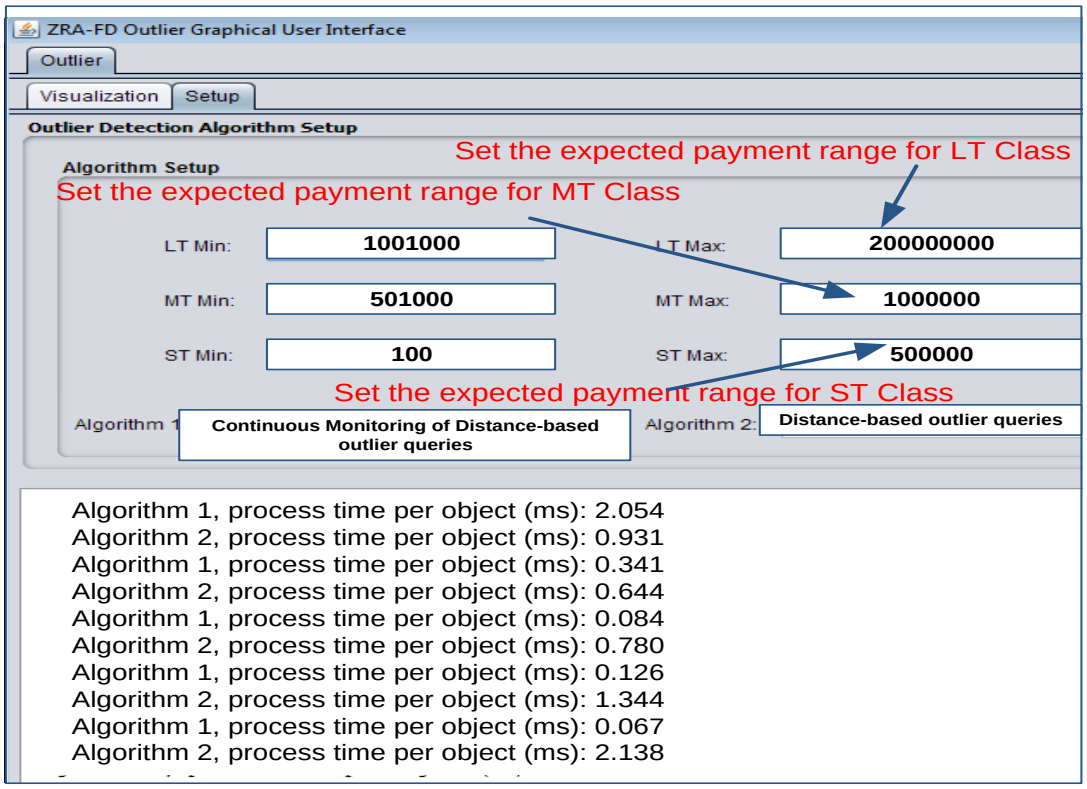

Figure 36: Outlier detection, range of expected payment.

ii. Results after the Algorithm is run

Figure 37 shows the results after the algorithms is run, the outliers can then be exported to an excel file by clicking the export button and an excel file called, "ZRA outlier Payments.xsl" is created. Clicking an outlier point gives information about that payment.

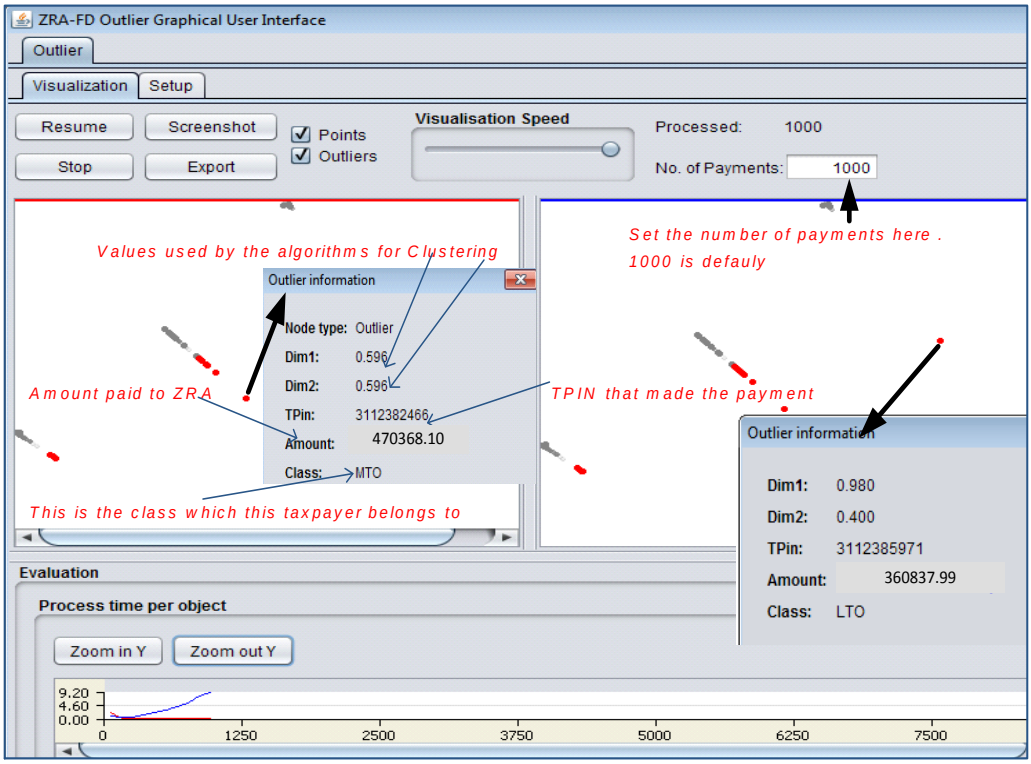

Figure 37: Outlier Information

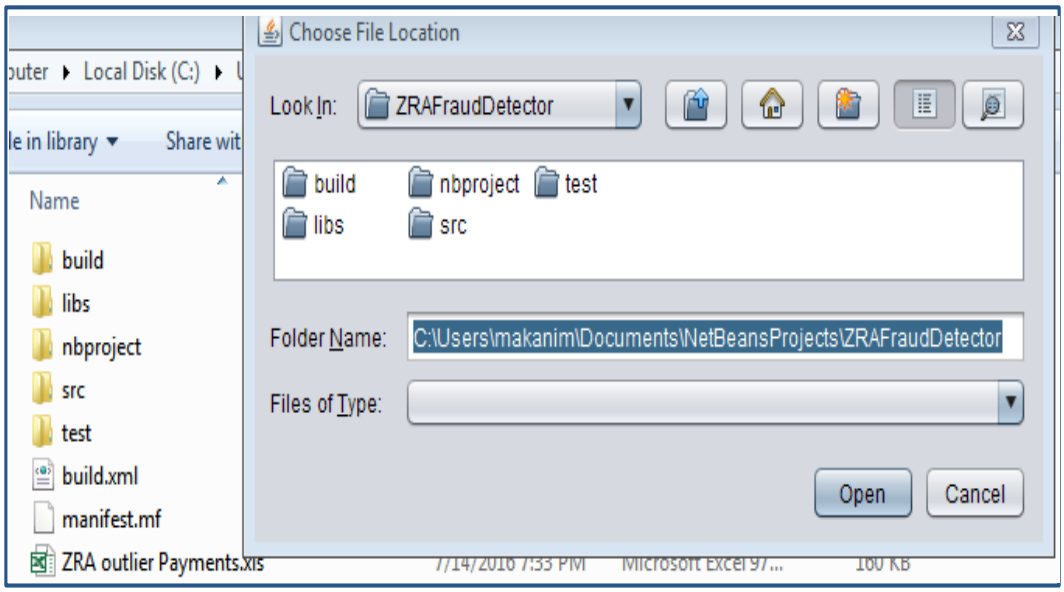

Figure 38: Exporting Outliers detected

A report for all payments that seem suspicious. This report will later be used for further investigation.

No.	TPin	Calss	Amount
1	3113648373	MTO	421217.88
2	3112331391	MTO	360837.99
3	3113613228	MTO	339851.25
4	3113721641	MTO	289459.0
5	3112382466	MTO	670368.1
6	3112329216	MTO	671540.0
7	3113667391	STO	47616.21
8	3113739373	MTO	315731.0
9	3112332482	MTO	294159.53
10	3112413193	STO	44282.97
11	3113193863	STO	50000.0
12	3112331635	STO	51653.07
13	3113648373	MTO	421217.88
14	3112331391	MTO	360837.99
15	3113613228	MTO	339851.25
16	3113721641	MTO	289459.0
17	3112385971	LTO	360837.99
18	3112382466	MTO	670368.1
19	3112329216	MTO	671540.0
20	3113667391	STO	47616.21
21	3113739373	MTO	315731.0
22	3112332482	MTO	294159.53
23	3112413193	STO	44282.97
24	3113193863	STO	50000.0
25	3112331635	STO	51653.07
26	3113648373	MTO	421217.88
27	3112331391	MTO	360837.99
28	3113613228	MTO	339851.25
29	3113721641	MTO	289459.0

Figure 39: A report on outlier detection

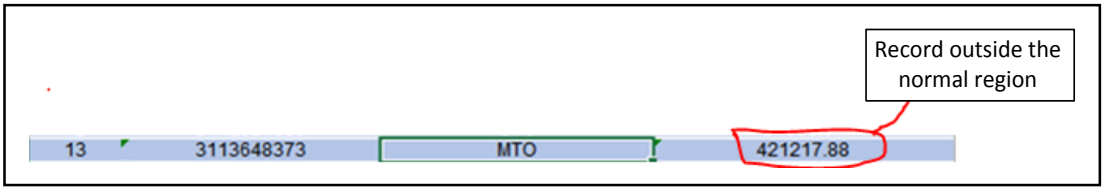

Figure 40: Example outlier

4.3.4 Evaluation of the two Algorithms and the traditional methods of detecting fraud

The implemented fraud detection tool gives advantages in terms of speed as shown in figure 36, accuracy, capability to cluster and expandability.

i. **Speed:** Figure 36, under section 4.3 gives the results of both the CMDB and the DBO after the algorithms are set to run. Given the same parameters such as the number of records to be processed for the same in the same clusters such as LT, MT, ST, The results show that Algorithm 1 (CMDB) took lesser time (0.067 minutes) as compared to Algorithm 2 (DBO) with 2.138 minutes. This therefore means that using Algorithm 1 (CMDB) will be more advantageous

because of the speed of processing it takes per object. However, comparing both algorithms and the traditional methods for which the baseline study indicates that it takes longer, more than 7 days to process and detect fraud on tax data on the same given number of records, the implemented algorithm is preferred.

ii. **Accuracy**: Figure 37 shows the information for each outlier. The taxpayer information including the amount, the dimension for the outlier and also the group to which the outlier belong. Algorithm 1 shows lesser distance denoted by the Dim 0.596, (Dimension) between the outlier and the group as compared to Algorithm 2 with Dim 0.980. When compared with the algorithm implemented, the traditional methods is not as accurate as the automated methods. It is prone to human error.

iii. **Capability to Cluster:** As discussed in section 2.1.2, Clustering is used to partition objects into previously unknown conceptually meaningful groups. For this study, both Algorithms 1 and Algorithms 2 showed the strength or capability to put the objects into meaningful groups such as LT, MT, ST.

iv. **Expandability**: This is a requirement to expand in future for the purpose of allowing more records to be processed at a given time. This was measured by adjusting the number of records being processed at a given time from 1000 to 1300, in the setup tab.

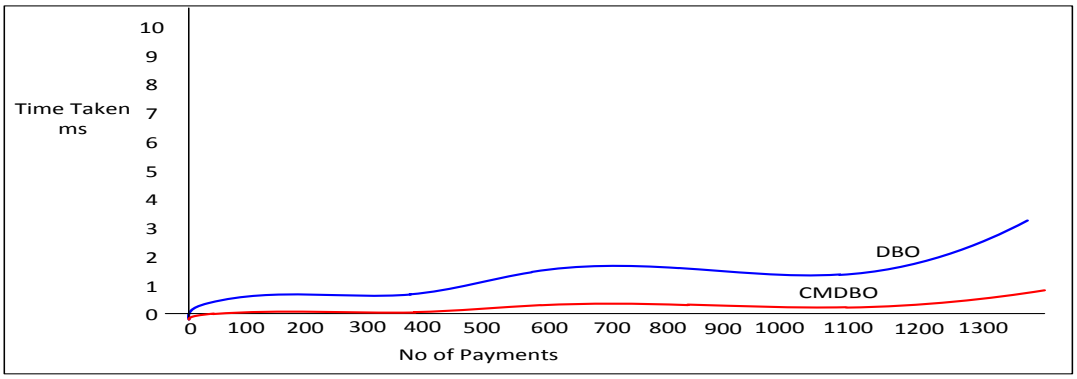

Figure 41: Evaluation of the two Algorithms

Algorithm 1 – CMDBO. **Algorithm 2** - DBO

4.4 Summary

In this chapter, the ZRA fraud detection prototype was successfully implemented. The tool was able to successfully detect the payments which are suspicious and are known as Outlier records which are then exported to Excel for further analysis.

CHAPTER FIVE

DISCUSSION AND CONCLUSION

5.1 Introduction

This chapter discusses the challenges that Zambia Revenue Authority faces in terms of fraud detection on taxes. We also looks at the solution proposed for fraud detection in ZRA based on the challenges established whilst conducting the survey. Finally a prototype based on the model on fraud detection is discussed. A conclusion is then made from the study and a summary given for the research study.

5.2 Discussion

An extensive review of the literature on business intelligence, data mining technologies and fraud detection in various sectors or industries and finally in the tax administrations sector as well as the various studies conducted by the statistics community demonstrating successful use of outlier in fraud detection shown in Section 2.3 of the literature review [62] [63]. Anomaly detection is an important data mining task aiming at the selection of some interesting objects, called outliers that show significantly different characteristics than the rest of the data set.

5.2.1 Challenges ZRA face regarding fraud detection on Taxes.

This research objective was achieved by engaging respondents from the three (3) different business layers within ZRA namely, Domestic Taxes, Customs Services and Information Technology to answer a questionnaire. Oral interviews were also used as a way of obtaining information.

Therefore, it was observed that ZRA uses excel to manipulate and analyse data on taxes. It was further discovered that detection of fraud on taxes is achieved through targeted & random audits, this method is time consuming, inefficient and is prone to error. A Random Audit is a follow up system used to verify information about taxes. This type of an audit is not announced in advance and makes everyone or every part of the tax system subject to the possibility of being selected for audit. Whereas the targeted audit is one where the participants in the audit are aware in advance in order for them to prepare for audit task in advance. Based on the survey results in Figure 26, indicate that it takes above 7 days to detect fraud on tax data using the traditional

methods (90%) whilst 10% indicates that it takes below 7 days to detect fraud on tax data using the traditional methods given the same number of records to process like the parameter defined in the system.

Though there is already an indication demonstrating the power of analytics in the Tax system and other applications which are based on Oracle, the study however indicates that there is a growing acknowledgement of the need to automate fraud detection in ZRA by implementing business intelligence and data mining in a comprehensive way which will bring about efficiency, accuracy and a convenient way of operations particularly on detection of Fraud.

Additionally, this study also recorded challenges that ZRA faces when it comes to data management and also reporting, this indicated a preference to implement data warehouse technologies in order to integrate the most critical databases holding Bulk tax data in ZRA.

Whilst the Authority faces these challenges, it has moderated its operations by putting up and maintaining Units that support the functions of audits and data manipulation such as Business Support System unit and Business intelligence Unit both in domestic Taxes and also Data Management Intelligent Unit in Customs Services Division. These units are involved in data management and manipulation which includes analysis and reporting at a higher level.

Though literature also gives evidence that the tax bases in most African developing countries are weakened by widespread of tax frauds [71] and that the developing countries have not been spared from the pressure to perform more efficiently and effectively, not so much research and also implementation has been done on these BI, data mining technologies to help out with fraud detection.

5.2.2 Business Intelligence, Data mining model to detect Fraud on Tax Data.

According to the results of the baseline study in section 4.2 of this dissertation, there is a problem in terms of how fraud is detected in ZRA.

In order to come up with the data mining model which will help with detection of fraud in the Tax Administration in Zambia (ZRA) a thorough review of the literature in section 2.2 on different existing models was done. Further, the survey results also reviewed the need to have an automated way of detecting fraud. Though there are some differences that can be seen in different models of BI and Data mining, there are

components which are quite common. These features are important and should be included in a model.

It has further been discovered through the review of literature that some existing BI and Data mining models lack support on metadata management. A good BI architecture should include the layer of metadata.

Kimball R., [122], looks at metadata as, all the information that defines and describes the contents, Structures, and operations of the DW/BI system. It describes the contents of the warehouse and defines the structures that hold those contents and the processes that brought those contents into being.

A metadata repository is therefore critical for business users for the purpose of storing and standardize metadata across different systems. Organisations will be able to track and monitor data flows within their BI environment if well-structured metadata is in place. Further, organisation will be able to guarantee the consistency of definitions and descriptions of data that support BI components and thus avoid misunderstanding and misinterpretation of data.

Implementation of operational data store (ODS) to take care of the needs of an organization is very important. It provides the users with current and integrated information that can be accessed or updated directly. Krishna G. et al [123], model contains only data warehouse and data marts whereas model by Madhuri J., [124] include only data warehouse but without ETL.

5.2.3 Prototype for Fraud on Tax Data

Finally, in order to develop a prototype for Fraud detection on bulk tax data, a Model in Section 3.4.4 and Figure 13 was used. Section 5.2.1 indicates that mostly, fraud in the Zambia's tax administration is detected through targeted and random audits as well as using the informants and under-cover operations commonly known as traditional methods.

The study also reviewed that Business users extract data from various sources systems independently, thereafter, this data is later loaded into spreadsheets for performing separate manipulations without centrally sharing the end result throughout the organization. Further Data sources are many, multi-structured and originating from different and several systems. The data sources are not integrated and data is growing rampantly. This has proved a challenge when it comes to data management, data manipulation and knowledge discovery.

As shown in the literature review Section 2.2, fraud on Bulk tax data is becoming an increasingly serious problem hence effective detecting of tax fraud has always been an important but complex task for the tax administration. The detection of tax fraud using traditional Audit methods is difficult though an important task. While knowing the limitations of an audit, the Tax administration has concluded that traditional and standard auditing procedures are insufficient to detect frauds. These limitations suggest the need for additional tools for effective detection of falsified tax information.

According to the baseline study undertaken, the Zambia's Tax Administration has not fully automated the detection of fraud on Bulk tax Data. Manual systems have in them inherent risks such as risks associated with inefficiency, poor record keeping and possible manipulation of processed information, hence the need for an automated solution. This will be able to combine data from various sources which includes the legacy systems and later consolidate them into easy-to-interpret, actionable classifications and predictions.

Anomaly detection is also not comprehensively done. However, some of these existing systems have some BI components within them such as all systems that are implemented based on Oracle and all SAP related systems.

Business Intelligence, Data Warehouse and Data Mining are therefore underlying strategic techniques for tax administrations to discover useful knowledge in support of their Tax Frauds detection, Tax Evasion and compliance enhancing agendas.

Based on the three tier Client Server Architecture, Figure 15 and the model designed and implemented, Figure 13, a fraud detection tool was developed. This Data was extracted from two data sources. Two java classes namely; PaymentDataNormaliser and ProfileDataNormaliser were created in order to transform, clean and load the Data into one Data Source zra_bi.

This procedure was done because the data had challenges such as spelling and missing values/attributes as described in Section 3.3.3

The ZRA Fraud detector tool was developed in java using weka Java library (NetBeans Integrated Development Environment). Weka java library implements numerous Data mining algorithms. The NetBeans IDE used is a free, open-source, cross-platform integrated development environment (IDE) with software development built-in support for Java Programming Language.

The two algorithms to detect the outlier were used as shown in Figure 35 and Figure 37. The results showed that algorithm1 provided better results in terms of the speed of processing and also the accuracy of the modelling as compared to Algorithim 2 despite both algorithms being subjected to the same parameters such as the minimum and maximum number of records for each cluster or category such as LT, MT, ST and also given the same data.

Further, traditional methods of detecting fraud as illustrated by the baseline study indicates that it takes longer in terms of speed of processing and also the accuracy of detecting fraud on tax data.

5.3 Comparison with Other Similar Works

Business Intelligence has been around and has been used primarily to generate standard reports, but today it has come to stand for a variety and diverse activities [125]. According to the literature review, Various Tax Administrations today are making a lot of efforts to implement BI and DM. Similar works in the literature is given as examples. Spruijt R., [27] proposed a BI and Data mining model consisting of three layers such as, Access layer for bringing components and functions from the logic layer in order to present them to the user in an integrated and personalized fashion, Logic layer focuses on the compilation, processing and distribution of management support data, and finally Data layer containing the data for the analysis, often a data warehouse.

The IRS [85], is currently using the compliant Data Ware house (CDW), this model enables highly flexible queries, it provides a centralized source of accurate and consistent data, further, and it is more secure than the old legacy system storage tapes, thereby better protecting taxpayer data. Therefore, our approach to identifying possible frauds is based on a five (5) layered BI and data mining model. It is the technology solution of large-scale adoption of Business Intelligence and data mining to help detect fraud based on ZRA's bulk tax data. It takes into consideration the importance and quality of data as well as information flow in the system. The five layered BI and Data mining model to help with Fraud detection on Bulky Tax Data was implemented. This five layered model in Figure 13, takes into consideration the compliance enablement, Taxpayer processes importance and quality of data as well as information flow that contribute to the bulk tax data on the data sources, ETL (Extract-Transform-Load), Integrated data store (Data Warehouse), data mining and End user component.

Outlier detection has been in existence for quite some time now [126], and literature indicates that there is no single universally applicable or generic detection approach and that authors have applied a wide variety of techniques in order to get the good and desired results.

ZRA Tax Expert views gathered in relation to this study of fraud detection using BI and data mining during the questionnaire time mainly endorses the above described model confirming that implementing such technologies could serve as a main solution.

5.4 Possible Application

Today, there are several areas where BI and data mining using outlier algorithms has been applied. This is seen in the Tax Administration, Insurance, Health, Education and Telecommunications sector [63].

Section 2.2.1 of the literature review of this study indicates that in the banking sector, the innovative techniques such as data mining, with advanced classification and prediction capabilities using outliers indicates that it can be employed to simplify auditors' role in terms of successfully accomplishing of the task of fraud detection.

In the education sector, there also has been achievements associated to the use of BI and Data mining technologies. A review was made in the literature, Section 2.2.2 where specifically data mining and analytics have been used to analyze student data, guide course redesign and for retooling assessment as well as to encourage new communication models between instructors and student learning,

According to the literature review of Section 2.2.3, hospitals have also experienced the pressure to introduce new tools (BI and Data Mining) that takes data from the hospital's electronic health record system, then clusters patients into different risk levels. Using Data mining also assesses historic data to determine care strategies that have worked in the past for certain kinds of patients in order to be applied on the new patients.

In the Tax Administration, there are several possible applications of BI and Data mining. One of them is the segmentation of Taxpayers based on the characteristics and the tax payer profiles as well as the Business rules and outlier detection rules. The Tax administration can also use Data mining to analyse the current taxpayers, their connection to other types of businesses as well as the types of Taxes which are filed by the taxpayers. In Zambia, there is a possibility that a taxpayers can have several businesses which pay different tax types but are owned by one taxpayer. In such a case, Data mining techniques will be able to cross match the data within the data ware house

in order to find the associations and some interesting patterns.

The literature review in chapter 2 of this report indicate that the possibility to detect and prosecute tax violators rests critically on data availability and data quality, therefore, actions taken against tax fraud relate to an improvement based on data quality available to tax administration. The Tax Administration, the governments and the private sectors are advocating for implementation of BI and data mining technologies because of the many and strong advantages presented in the literature.

5.5 Conclusion

This study considered the opportunities of using Business intelligence, data mining, to detect Fraud on bulk tax data in ZRA. As evidenced from the literature, data mining is a process used in order to discover uncertain, unknown, and or hidden information from the existing data. Fraud detection in the tax administration using data mining and outlier algorithm, is an important task which is grounded on the selection of some objects that are quite interesting, these outliers show significantly different characteristics than the rest of the data set.

Therefore, from this research, it can be observed that Business Intelligence and Data mining are key technologies to many of the shortcomings of the traditional approach in combating error and fraud, and it also gives reasons to believing that data mining could meaningfully contribute to making the detection of fraud or anomaly more effective for tax administration in Zambia (ZRA).

Using Data mining, Outlier algorithms can detect patterns of data for taxpayer's payments and their profiles and are able to pick the anomalies in the data such as the underpayments, non-payment. Data mining using outliers can also help segment the Tax payers according to their Business Profiles and their business turnover, and values.

Implementation of Business Intelligence, data mining in the Tax Administration will enable the organization make well informed business decisions, there will be a well-defined position of the organization in terms of Revenue. Implementing Data mining technologies will also facilitate Tax payers behavior positively and payment patterns.

Further, it was observed that in outlier detection, Organisations through their developers should select algorithms based on the suitability for their data set, in terms of the correct distribution model, the capabilities to increment in order to allow more data to be stored, the accuracy of the results of the processing when the algorithm is run.

There is also need to further come up with a plan of how to handle the outliers when detected. The purpose of this study was to develop a data mining model and prototype as a Business Intelligence tool for detection of fraud on tax data and analysis of bulk data for ZRA and this study focused on data sources for Domestic Taxes only and achieving this objective was made possible by firstly, examining the challenges that ZRA is facing with regard to detection of Fraud.

5.6 Future Works

In Future research, there are also great opportunities to get deeper into improving upon this model in various ways such as advanced predictive analytics as shown in Section 4.3.2. This is the best strategic reporting and basic forecasting with additional operational intelligence and decision making. Through the eyes of a taxpayer and other stakeholders, ZRA must be looked at as one, therefore, introduction of a comprehensive data warehouse solution will also simply give a consolidation of data from a variety of sources that is designed to support strategic and tactical decision making. ZRA houses data from several sources such as imports, exports and stakeholders profiles. Consideration of a data warehouse will enable the organisation analyse its bulk tax data over time. A typical example of a typical query submitted to a data warehouse might be: "What was the total revenue produced for the newly commissioned ZRA station in the new Chinsali district for Domestic Taxes particularly 'Value Added Tax' during the first quarter of 2015"?

The other enhancements that can be incorporated is the automated reporting with interactive dashboards feature which indicates the status of things at a specific point in time and a scorecard, on the other hand, to display progress over time towards specific goals of the organization. This therefore mean that ZRA will be able to provide the revenue status to its stakeholders such as the government, parliamentarians and the Board of Directors.

5.7 Summary

This chapter gives the summary of the chapter five (5). It provides the conclusions made from the study in terms of the importance of BI and data mining system as one of the solutions currently seen as the most effective way to detection of fraud and anomalies that exist on bulk tax data. This chapter further gives a comparison of BI and data mining with other similar works as well as the possible application.

REFERENCES

[1] Y. Wua, "Using data mining technique to enhance tax evasion detection performance, Expert Systems with Applications," *An International Journal*, vol. 39, p. 8769–8777, 2012.

[2] *Oxford Concise English Dictionary*, *11th Edition*, Oxford University Press, 2009.

[3] IRS, "irm_25-001-001.html (IRS Fraud Handbook)," 23 January 2014. [Online]. Available: https://www.irs.gov/irm/part25/irm_25-001-001.html. [Accessed 16 May 2016].

[4] Y. Chu, "Combating Tax Fraud," 1 February 2012. [Online]. Available: https://www.imf.org/external/np/seminars/eng/2012/asiatax/pdf/chu.pdf. [Accessed 16 May 2016].

[5] Fraud and Risk Management Working Group, "Fraud Risk Management: A Guide to Good Practice," *CIMA: Fraud Risk Management*, 2008.

[6] A. Gounaris, "Continuous Monitoring of Distance-Based Outliers over Data Streams," in *Proceedings of the 27th IEEE International Conference on Data Engineering (ICDE)*, Hannover, Germany., 2011.

[7] C. Zwilling and M. Yongmei Wang, Multivariate Computing and Robust Estimating for Outlier and Novelty in Data and Imaging Sciences, S. P. A, Ed., Science, Technology and Medicine open access publisher., 2015.

[8] V. Bart and V. van, fraud analytics using descriptive, social network techniques. A guide io data science for fraud detect detection, prevvention and analytics, North California,: usa, 2015.

[9] R. Johnson, Applied Multivariate Statistical Analysis, Prentice Hall, 1992.

[10] C. Kanakalaksmil, "A concise study on Text Mining for Business Intelligence.," International Journal of Advanced Research in Computer and Communication Engineering," vol. 4, no. 6, 2015..

[11] C. Rajan, "Business Intelligence for competence in Consumer packaged good industry," *International Journal of Marketing and Technology.*, vol. 2, no. 5, May 2012.

[12] A. Karim, "The value of Competitive Business Intelligence System (CBIS) to Stimulate Competitiveness in Global Market," *International Journal of Business and Social Science*, vol. 2, no. 19, p. 196, October 2011.

[13] F. Ameur and T. Mohamed, "Taxpayers Fraudulent Behavior Modeling the Use of Datamining in Fiscal Fraud Detecting Moroccan Case," *Scientifc Research, Applied Mathermatics*, 2012.

[14] O. Sheta and A. N. Eldeen, "The Technology Of Using a Data Warehouse to Support Decision-Making in Health Care," *International Journal of Database Management Systems (IJDMS)*, vol. 5, no. 3, June 2013.

[15] P. Castellón, A. González and D. Velásquez, "Characterization and detection of taxpayers with false invoices using data mining techniques," *Expert Systems with Applications,* p. 1427–1436., 2013.

[16] T. Silwattananusarn and K. Tuamsuk, "Data Management and Its Application or Knowledge Management: A Literature Review from 2007 to 2012.," *International Journal of Data Mining & Knowledge Management Process,* pp. 13 - 24., 2012.

[17] H. Wahbeh, H. Al-Radaideh, Al-Kabi and a. et, "A Comparison Study between Data Mining Tools over some Classification Methods.," *International Journal of Advanced Computer Science and Applications (Special Issue on Artificial Intelligence).,* pp. 18 - 26, 2011.

[18] C. Kanakalaksmil and C. Manicka, "A concise study on Text Mining for Business Intelligence.," *International Journal of Advanced Research in Computer and Communication Engineering.,* vol. 4, no. 6, 2015.

[19] A. Nandakumar, N. Jallapa and Yambem, "A Survey on Data Mining Algorithms on Apache Hadoop Platform.," *International Journal of Emerging Technology and advanced Engineering.,* vol. 4, no. 1, January 2014..

[20] N. Kumari, "Business Intelligence in a Nutshell," *Internation Journal of Innovative Research in Computer Communication Engineering,* vol. 1, no. 4, pp. 969 - 975, June 2013.

[21] A. Aziz and A. Rashidi, "Data Text and Web Mining for Business Intelligence," *International Journal of Data Mining and Knowledge Management Process,* vol. 3, no. 2, pp. 1 - 21, March 2013.

[22] H. Singh and S. Bhagat, "Systems are developed and designed to help the organisatdions understand their Customs," *International Journal of Science, Technology & Management,* vol. Volume No.04, no. Issue No. 01, January 2015.

[23] C. Ramanigopal, "Business Intelligence for competence in Consumer packaged good industry," *International Journal of Marketing and Technology,* vol. 2, no. 5, May 2012..

[24] Cambridge University , "Model," Cambridge University Press, 2016. [Online]. Available: http://dictionary.cambridge.org/dictionary/english/model. [Accessed 9 July 2016].

[25] M. Seidl, M. Scholz, C. Huemer and G. Kappel, UML @ Classroom: An Introduction to Object-Oriented Modeling, New York: Springer International Publishing, 2014.

[26] Oxford University Press, "model," Oxford University Press, 2016. [Online]. Available: http://www.oxforddictionaries.com/definition/english/model. [Accessed 5 July 2016].

[27] R. Spruijt, "BI Architecture that fits Organisationanis Requirements," 2014.

[28] T. Jaakko, "How can the Business Intelligence system be developed at YIT," 2011.

[29] R. Mansour, "Business Intelligence Strategies and Implementation," (2012)..

[30] J. Madhuri, "Significance of Data Warehousing and Data Mining in Business Applications," *International Journal of Soft Computing and Engineering (IJSCE),* vol. 3, no. 1, 2013.

[31] N. Lekhi, M. Mahajan and Landran, "Improving Cluster Formulation to ReduceOutliers in Data Mining," *International Journal of Innovative Research in Computer and Communication Engineering,* vol. 2, no. 6, 6 June 2014.

[32] D. Pachgade and S. Dhande, "Outlier Detection over Data Set Using Cluster-Based and Distance-Based Approach," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 2, no. 6, June 2012.

[33] D. Kumar and D. Bhardwaj, "Rise of Data Mining: Current and Future Application," *International Journal of Computer Science,* vol. 8, no. 5 No. 1, September 2011.

[34] A.N. Paidi, "Data Mining: Future Trends and Applications," *International Journal of Modern Engineering Research (IJMER),* vol. 2, no. 6, pp. pp.4657-4663, November - December 2012.

[35] L. Singh, "Data Mining: Review, Drifts and Issues," *International Journal of Advance Research and Innovation,* Vols. 44-48, 2013.

[36] M. Bharati, B. Desai and Ramageri, "Role of Data Mining in Retail sector," *International Journal on Computer Science and Engineering (IJCSE),* vol. 5, no. 1, January 2013.

[37] P. Kumara and A. Ranjanet, "Data Mining And Its Significance In Industrial Applications," *International Journal of Advanced Research in Computer Science,,* vol. 3, no. 2, pp. 275-278, March-April 2012 https://www.academia.edu/Download.

[38] M. Gonzales, IBM Data Warehousing with IBM Business Intelligence Tools., Indianapolis.: Joe Wikert, Wiley Publishing., 2003.

[39] R. Wua, B. Lin, B. Chang, C. David and C. Yen, "Using data mining technique to enhance tax evasion detection performance," *Expert Systems with Applications International Journal,* p. 8769–8777, 2012.

[40] M. Gupta, J. Gao, C. Aggarwal and J. Han, "Outlier Detection for Temporal Data: A Survey," in *Knowledge and Data Engineering, IEEE Transactions*, 2014.

[41] F. Wolfgang M, "Expected similarity estimation for large-scale batch and streaming anomaly detection.," in *International Joint Conference on Neural Networks*, 2015.

[42] A. Mohammad and Al-ma'aitah, "The Role of Business Intelligence Tools in Decision Making Process," *International Journal of Computer Applications,* vol. Volume 73, no. No.13, July 2013.

[43] Ranjan J., "Business Intelligence: Concepts, Components, Techniques and Benefits," *Journal of Theoretical and Applied Information Technology,* vol. 9, no. 1.

[44] O. Lih, "Five Layered, Business Intelligence Architecture.," *Communications of the IBIMA,* vol. Vol. 2011, 2011.

[45] J. Maeda and C. Hanlon, "Harnessing the power of enhanced data for healthcare quality improvement:Lessons from a Minnesota hospital association pilot Project Practitioner Application.," *Journal of Healthcare Management,* vol. 57, no. 6, p. 406–418, 2012.

[46] R.J. Davenport, *ETL vs ELT A Subjective View Part of the series of the Insource Commercial Aspects of BI discussion papers*, 2008.

[47] K. Talwar and A. Gosain, "Hierarchy classification for Data Warehouse: A Survey,," *International Conference on Communication, Computing & Security (ICCCS)*, vol. 6, pp. 460-468, 2013.

[48] A. Shadi and H. Abu, "The Role of Data Warehouse in Decreasing the Time of Decision Taking," *Australian Journal of Basic and Applied Sciences*, vol. 9, no. 5, pp. 216-219, 2015.

[49] M. Laxmaiah and A. Govardhan, "A conceptual metadata framework for spatial data Warehouse," *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol. 3, no. 3, pp. 63-73, 2013. .

[50] P. Sharma and P. Girdhar, "Online Analytical Processing (OLAP)," *International Journal for Research in Computer Science*, vol. VOL 2 July, no. ISSUE 7, 2015.

[51] Joomag, " IT BITS, IT Concepts Revisited, Big Data, Business Model, ERP CRM ECM BI Internet of Things," no. November issue,, 2014.

[52] M. Kaufman and D. Kirsch, "Advanced Analytics : The Hurwitz Victory Index Report," Hurwitz & Associates, 2014.

[53] K. Kakhani, S. Kakhani and S.R. Biradar, "Research Issues in Big Data Analytics," *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, vol. Volume 2, no. issue 8, August 2013.

[54] P. Gupta and B. Narang, "Role of Text Mining in Business Intelligence," *Gian Jyoti E-Journah*, vol. 1, no. 2, 2012 2012.

[55] L. Gao, E. Chang and S. Han S, "Powerful Tool to Expand Business Intelligence: Text Mining," in *Proceedings Of World Academy Of Science, Engineering And Technology*, 2005.

[56] I. B. Sowan, "Enhancing Fuzzy Associative Rule Mining Approaches for Improving Prediction Accuracy.," Bradford., 2011.

[57] U. Shafique and H. Qaiser, "A comparative study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)," *International Journal of Innovation and Scientic Research*, vol. 12, no. 1, pp. 217-222, November 2014.

[58] U. Keshavamurthy and H.S. Guruprasad, "Learning Analytics: A survey," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 18, no. 6, December 2014.

[59] D. Larose and C. Larose, Discovering Knowledge in Data: An Introduction to Data mining, 2nd ed., 2014.

[60] K. Singh and Chandandeep, "Predicting student attrition and factors affecting attrition in Higher Education," *International Journal of Engineering Research and Applications (IJERA)*, 29 March 2014.

[61] A. Sharma and K. PanigrahiPrabin, "A Review of Financial Accounting Fraud Detection based on Data Mining Techniques," *International Journal of Computer Applications*, vol. 39, no. 1, p. 0975 – 8887, 2012.

[62] M. Kontaki, A. Gounaris, N. Apostolos, K. Tsichlas and Y. Manolopoulos, "Continuous Monitoring of Distance-Based Outliers over Data Streams," in *Proceedings of the 27th IEEE International Conference on Data Engineering (ICDE)*, Hannover, Germany, 2011.

[63] K. Singh and S. Upadhyaya, "Outlier Detection: Applications And Techniques," *IJCSI International Journal of Computer Science,* vol. 9, no. 1, Issues, Vol. 9, Issue 1, No 3, January 2012, ISSN (Online): 1694-0814, www.IJCSI.org January 2012.

[64] C. Aggarwa, "Outlier Analysis," in *Outlier Analysis,* New York, Boston: Kluwer Academic Publishers.

[65] E. Ngai, Y. Hu, Y. Wong and X. S. Y. Chen, The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature, 2011, p. 559–569.

[66] N. Padhy, M. Pragnyaban and R. Panigrahi, "The Survey of Data Mining Applications And Feature Scope," *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT),* vol. 2, no. 3, June 2012.

[67] R.K. Sahu, "Application of Business Intelligence in the Banking Industry," *Management Information Systems,* vol. 6, no. 4, 10 July 2011.

[68] MicroStrategy, , "Major applications of business Intelligence software in the banking industry".

[69] U. Bogdan and E. Đurković, "Application of Business Intelligence in the Banking Industry," *Management Information Systems,,* vol. 6, no. 4, 2011.

[70] L. Chi-Chen, L. Chiub, Yan, D. C. Shaio, Shaio and Huang, "Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments," *Knowledge-Based Systems,* vol. 89, p. 459–470, November 2015.

[71] D. Guster and C. G. Brown, "The Application of business intelligence to higher education: technical and managerial perspectives," *Journal of Information Technology Management,* vol. XXIII, no. 2, 2012.

[72] M. Cheng, "Application of business Intelligence in higher Education sector," 2012.

[73] M. Bala and D.B.Ojha, "Study of applications of Data Mining Techniques in Education," *International Journal of Research in Science and Technology,* vol. 1, no. 6, pp. 135 - 146, 2012.

[74] D.Guster and G. C. Brown, "Technical And Managerial Perspectives," *Journal Of Information Technology Management,* vol. Volume Xxiii, no. Number 2, 2012.

[75] J. Cynthia and Baepler, "Academic Analytics and Data Mining in Higher Education," *International Journal for the Scholarship of Teaching and Learning,* Vols. Vol. 4,, no. No. 2, 2010.

[76] R. Hanson, "Good health information – An asset not a burden," *Australian Health Review,,* vol. 35, no. 1, pp. 9-13, 2011.

[77] E. Burns, "How predictive analytics in healthcare can lower readmissions," TechTarget, 2015.

[78] E. Burns, "EHR systems holdingback healthcare data analytics efforts," 2015.

[79] N. Foshaya and C. Kuziemsky, "Towards an implementation framework for business intelligence in health care," *International Journal of Information Management,* p. 20– 27, 2014.

[80] N. Ashrafi, L. Kelleher and J. P. Kuilboe, "The impact of business intelligence on healthcare delivery in the USA," *Interdisciplinary Journal of Information, Knowledge, and Management,* vol. 9, pp. 117-130, 2014.

[81] K. Batko and M. Olsza, "The Use of Business Intelligence Systems in Healthcare Organizations in Poland," in *Proceedings of the Federated Conference on Computer Science and Information Systems,* 2012.

[82] E. Siehl, "Addressing tax evasion and avoidance," Addressing-tax-evasion-and-avoidance, 2011.

[83] Atos, "tackling fraud with a big data approach. The common error of under-utilizing Revenue Authority data," 2015.

[84] J. Martikainen, "Data Mining in Tax Administration - Using Analytics to Enhance Tax Compliance," 2012.

[85] J. Butler, "Big Data and Analytics at the IRS: Perspectives and Initatives," 5-6 March 2013. [Online]. Available: https://www-01.ibm.com/events/wwe/grp/grp004.nsf/vLookupPDFs/Jeff%20Butler's%20Presentation/$file/Jeff%20Butler's%20Presentation.pdf. [Accessed 4 July 2016].

[86] R. Satran, "irs-data-web-snares-mostly-low--and-middle-income-taxpayers," 2013.

[87] P. McGuire, "Internal Revenue Service Predictive Analytics," in *Predictive Analytics World for Government conference*, Washington, 2011.

[88] A. Noroozi, "Review into the Australian Taxation Office's compliance approach to individual taxpayers – income tax refund integrity program," http://igt.gov.au/51C40C55-06D6-4035-AD9C-2840FBF475DD/FinalDownload/DownloadId-565D0EF630B952E1D2E07F5AC5C7A5ED/51C40C55-06D6-4035-AD9C-2840FBF475DD/files/2014/11/income-tax-refund-integrity-program.pdf, Sydney, 2013.

[89] P. Drummond, D. W., N. Srivastava and L. E. Oliveira, "Mobilizing revenue in sub-Saharan Africa: empicial norms and key determinants," International Monetary Fund, Washington DC, 2012..

[90] Fiscal Affairs Department, "Revenue mobilization in developing countries," International Monetary Fund, Washington D. C, 2011.

[91] M. Martini, "Approaches to curbing corruption in tax administration in Africa," *Transparency International, U 4 Expert answer,* 25 June 2014.

[92] F. Ameur and M. Tkiouat, "Taxpayers Fradulent BehaviourModelling the Use of Datamining in Fiscal Fraud Detecting Moroccan Case," *Scientific Research Journal,* pp. 1207-1213, 10 September 2012.

[93] CIAT , "Prevention and control of tax evasion," in *CIAT Technical Conference*, Nairobi, 2013.

[94] D. Cleary, "irish-tax-and-customers," SAS Institute, [Online]. Available: http://www.sas.com/da_dk/customers/irish-tax-and-customers.html. [Accessed 9 March 2016].

[95] W. Nhekairo, "The Taxation System in Zambia," Jesuits Centre for Theological Reflection:, Lusaka, Zambia.

[96] B. Msiska, "Planning a Change Strategy for Tax Administration: A case of ZRA," [Online]. Available:

http://siteresources.worldbank.org/PSGLP/Resources/5ZambiaSA.pdf. [Accessed 3 March 2016].

[97] P. B. Jonker J., The Essence of Research Methodology: A concise guide for Master and PhD Students in Management, London: Library of Congress Control, 2010.

[98] T. Lund, "Combining qualitative and quantitative approaches: Some arguments for mixed methods research.," *Scandinavian Journal of Educational Research.,* vol. 5, no. 2, pp. 155-165, 2012.

[99] V. Venkatesh, S. Brown and H. Bala, "Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems," *MIS Quarterly,* vol. 37, no. 1, pp. 21-54, 2013.

[100] R. Frels and A. J. Onwuegbuzie, "Administering quantitative instruments with Qualitative Interviews: A mixed research approach.Journal of," *Journal of Counseling & Development, 91(2), 184-194. doi: 10.1002/j.1556-6676.2013.00085.x,* vol. 91, no. 2, pp. 184-194, 2013.

[101] J. W. Creswell, Educational research: Planning, conducting, and evaluating quantitative and qualitative research, Boston, MA: Pearson Education, Inc, 2012.

[102] Corporate Research and Consultation Team, "Questionnaires," 2010.

[103] B. Acharya, "Questionnaire Design," Tribhuvan, Nepal, 2010.

[104] Resource and Planning Department, "Zambia Revenue Authority Website," ZRA, 2012 - 2013. [Online]. Available: https://www.zra.org.zm/commonHomePage.htm?viewName=organizationStructure. [Accessed 10 March 2016].

[105] Research and Planning Department, "PointsofPresence," Zambia Revenue Authority, 2012 - 2013. [Online]. Available: https://www.zra.org.zm/commonHomePage.htm?viewName=PointsofPresence. [Accessed 8 March 2016].

[106] C. Kothari, Research Methodology, Methods and Techniques., University of Rajasthan, Jaipur, New Deihli: New Age International (P) Limited Publilshers, New Delhi, ISBN, 2009, pp. 63 - 64.

[107] R. Katragadda, D. Nandigam and S. T. Sreenivas, "ETL tools for Data Warehousing: An empirical study of Open Source Talend Studio versus Microsoft SSIS," New Zealand, 2015.

[108] R. Kimball and J. Caserta, The Data Warehouse ETL Toolkit, Practical techniques for Extracting, Cleaning, Conforming and Delivery of Data, Indianapolis: Wiley Publishing, Wiley Publishing, pp. 55 - 160.

[109] R.S. Chillar, "Extraction Transformation Loading – A Road to Data warehouse," in *2nd National Conference Mathematical Techniques: Emerging Paradigms for Electronics and IT Industries*, India,.

[110] H. Kaur, "A Review of Applications of Data Mining in the Field of Education," *International Journal of Advanced Research in Computer and Communication Engineering,* vol. Vol. 4, no. Issue 4, April 2015.

[111] S. Gaurav and K. Ajay, "Cryptography Algorithms and approaches used for data security," *International Journal of Scientific & Engineering Research,* vol. 3, no. 6, 2012.

[112] I. E. Ismael and F. Abdulameerabdulkareem, "Enhancement Caesar Cipher for Better Security," *IOSR Journal of Computer Engineering (IOSR-JCE),* vol. 16, no. 3, pp. 01-05, May-June 2014.

[113] O. Hamdan, B. Alanazi, B. Zaidan and A. Zaidan, "New Comparative Study Between DES, 3DES and AES within Nine Factors," *Journal of Computing,* vol. 2, no. 3, pp. 152-157, 2010.

[114] G. Goyal and S. Kinger, "Modified Caesar Cipher for Better Security Enhancement," *International Journal of Computer Applications,* vol. 73, no. 3, July 2013.

[115] S. MacDonald and N. Headlam, Research Methods Handbook: Introductory guide to research methods for social research, Manchester, 1986.

[116] KDnuggets, "KDnuggets Annual Software Poll: Analytics , Data Mining software used," KDnuggets, Chicago, 2013.

[117] B. Ghosh, Scientific Method and Social Research, New Delhi: Sterling Publishers Private Limited, 2000.

[118] Boyer, John, Frank and Kay, Business Intelligence Strategy – Practical Guide for Achieving BI, 2010.

[119] Zambia RevenueAuthority, "E-services," ZRA, [Online]. Available: http://www.zra.org.zm/main.htm?actionCode=showHomePageLnclick. [Accessed 24 June 2016].

[120] M. Mwiya, J. Phiri and G. Lyoko, "Public Crime Reporting and Monitoring System Model Using GSM and GIS Technologies: A Case of Zambia Police Service," *International Journal of Computer Science and Mobile Computing,* vol. 4, no. 11, pp. 207-226, 2015.

[121] F. F. Angiulli F., "Distance-based outlier queries in data streams: the novel task and algorithms," *Data Mining and Knowledge Discovery,* vol. 20, no. 2, pp. 290-324, 2010.

[122] R. Kimball, J. Mundy and W. Thornthwaite, The Microsoft Data Warehouse Toolkit: With SQL Server 2008 R2 and the Microsoft Business Intelligence Toolset, Indianapolis: Wiley Publishing, 2011.

[123] G. K. Tiruveedula, A. Mohamed and A.H. Sabahaldin, "Towards OLAP - Based Data Mining Using Multidimensional Database and Fuzzy Decision Trees," *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* vol. 3, no. 6, November 2013.

[124] J. V. Madhuri, "Significance of Data Warehousing and Data Mining in Business," *International Journal of Soft Computing and Engineering (IJSCE),* vol. 3, no. 1, 2013.

[125] T. DavenPort, "The New World of Business Analytics," International Institute for Analytics, 2010.

[126] H. Kriegel, P. Kroger and A. Zimek, "Outlier Detection Techniques," Colombus, Ohio, 2010.

[127] G. C. PhridviRaja M. B., "Data mining – past, present and future – a typical survey on data," in *The 7th International Conference Interdisciplinarity in Engineering (INTER-ENG 2013)*, 2014.

[128] R. &. R. M. Mikut, Data mining tools. Wiley interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011, pp. 1 - 13.

[129] B. N. C. S. Blackburn K., "Tax Evasion, the underground economy and finacial development," *Journal of economic Behaviour and Organisation*, vol. 83, pp. 243-253, 2012.

[130] Zambia Revenue Authority, "ZRA Map of Presence," ZRA, Lusaka, 2012 - 2013.

[131] Elsevier, "journals-overview," [Online]. Available: http://www.elsevier.com/solutions/sciencedirect/content/journals-overview.html.. [Accessed 7 May 2016].

[132] S. Malik, "A comparative study of two major search engines google and yahoo," *An International Research Journal of Computer Science and Technology*, vol. 7, no. 1, 2016.

[133] M. Mwanza and J. Phiri, "Fraud Detection on Bulk Tax Data Using Business Intelligence, Data Mining Tool: A case of Zambia Revenue Authority," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5, no. 3, March 2016.

APPENDIX

Source Code A: Data Access

DAO.java, the class used to access the database to get payment data.

```
package unza.dataaccess;
import java.sql.Connection;
import java.sql.DriverManager;
import java.sql.ResultSet;
import java.sql.Statement;
/**
 * This class accesses the database and gets the payments
 */
public class DAO {
    String driver = "com.mysql.jdbc.Driver";
    String url = "jdbc:mysql://127.0.0.1/zra_bi";
    String username = "root";
    String password = "root";
    Connection connection;
    int index = 0;
    public DAO() {
        try {
            Class.forName(driver);
            connection = DriverManager.getConnection(url, username, password);
        } catch (Exception ex) {
            ex.printStackTrace();
        }
    }
    /**
     * Get financial records
     *
     * @return
     */
    public String getNextFinancialRecord() {
        String result = "";
        String result_bk = "";
        try {
            Statement statement = connection.createStatement();
            String query = "SELECT * FROM tax_payment WHERE sr_no > " + index + " Limit 1";
            ResultSet resultSet = statement.executeQuery(query);
            if (resultSet.next()) {
                index = resultSet.getInt("sr_no");
                result = result + resultSet.getString("tpin")+",";
                result = result + resultSet.getString("division")+",";
                result = result + resultSet.getDouble("amount");
                result_bk = result;
            } else {
                result = result_bk;
            }
        } catch (Exception ex) {
            ex.printStackTrace();
        }
        return result;    }}
```

Source Code B: Logic, RunVisualiser.java

```
package unza.logic;

import moa.gui.outliertab.DataPoint;
import moa.gui.outliertab.GraphCanvas;
import java.awt.Color;
import java.awt.event.ActionEvent;
import java.awt.event.ActionListener;
import java.io.BufferedWriter;
import java.io.FileWriter;
import java.io.IOException;
import java.io.PrintWriter;
import java.util.ArrayList;
import java.util.Iterator;
import java.util.LinkedList;
import java.util.List;
import java.util.logging.Level;
import java.util.logging.Logger;
import moa.cluster.Cluster;
import moa.cluster.Clustering;
import moa.clusterers.AbstractClusterer;
import moa.clusterers.ClusterGenerator;
import moa.evaluation.MeasureCollection;
import moa.gui.outliertab.TextViewerPanel;
import unza.presentation.ClusteringSetupTab;
import moa.gui.clustertab.ClusteringVisualEvalPanel;
import unza.presentation.ClusteringVisualTab;
import moa.gui.visualization.WekaExplorer;
import moa.streams.clustering.ClusterEvent;
import moa.streams.clustering.ClusterEventListener;
import moa.streams.clustering.ClusteringStream;
import unza.presentation.StreamPanel;
import weka.core.Attribute;
import weka.core.DenseInstance;
import weka.core.FastVector;
import weka.core.Instance;
import weka.core.Instances;

public class RunVisualizer implements Runnable, ActionListener, ClusterEventListener{
   /** the pause interval, being read from the gui at startup */
   public static final int initialPauseInterval = 5000;

   /** factor to control the speed */
   private int m_wait_frequency = 1000;

   /** after how many instances do we repaint the streampanel?
    *  the GUI becomes very slow with small values
    * */
   private int m_redrawInterval = 100;
   /* flags to control the run behavior */
   private static boolean work;
   private boolean stop = false;

   /* total amount of processed instances */
   private static int timestamp;
   private static int lastPauseTimestamp;
```

```
    /* amount of instances to process in one step*/
    private int m_processFrequency;

    /* the stream that delivers the instances */
    private final ClusteringStream m_stream0;

    /* amount of relevant instances; older instances will be dropped;
       creates the 'sliding window' over the stream;
       is strongly connected to the decay rate and decay threshold*/
    private int m_stream0_decayHorizon;

    /* the decay threshold defines the minimum weight of an instance to be relevant */
    private double m_stream0_decay_threshold;

    /* the decay rate of the stream, often reffered to as lambda;
       is being calculated from the horizion and the threshold
       as these are more intuitive to define */
    private double m_stream0_decay_rate;

    /* the clusterer */
    private AbstractClusterer m_clusterer0;
    private AbstractClusterer m_clusterer1;

    /* the measure collections contain all the measures */
    private MeasureCollection[] m_measures0 = null;
    private MeasureCollection[] m_measures1 = null;

    /* left and right stream panel that datapoints and clusterings will be drawn to */
    private StreamPanel m_streampanel0;
    private StreamPanel m_streampanel1;

    /* panel that shows the evaluation results */
    private ClusteringVisualEvalPanel m_evalPanel;

    /* panel to hold the graph */
    private GraphCanvas m_graphcanvas;

    /* reference to the visual panel */
    private ClusteringVisualTab m_visualPanel;

    /* all possible clusterings */
    //not pretty to have all the clusterings, but otherwise we can't just redraw clusterings
    private Clustering gtClustering0 = null;
    private Clustering gtClustering1 = null;
    private Clustering macro0 = null;
    private Clustering macro1 = null;
    private Clustering micro0 = null;
    private Clustering micro1 = null;

    /* holds all the events that have happend, if the stream supports events */
    private ArrayList<ClusterEvent> clusterEvents;

    /* reference to the log panel */
    private final TextViewerPanel m_logPanel;

    public RunVisualizer(ClusteringVisualTab visualPanel, ClusteringSetupTab
clusteringSetupTab){
```

```
        m_visualPanel = visualPanel;
        m_streampanel0 = visualPanel.getLeftStreamPanel();
        m_streampanel1 = visualPanel.getRightStreamPanel();
        m_graphcanvas = visualPanel.getGraphCanvas();
        m_evalPanel = visualPanel.getEvalPanel();
        m_logPanel = clusteringSetupTab.getLogPanel();

        m_stream0 = clusteringSetupTab.getStream0();
        m_stream0_decayHorizon = m_stream0.getDecayHorizon();
        m_stream0_decay_threshold = m_stream0.getDecayThreshold();
        m_stream0_decay_rate                                                    =
(Math.log(1.0/m_stream0_decay_threshold)/Math.log(2)/m_stream0_decayHorizon);

        timestamp = 0;
        lastPauseTimestamp = 0;
        work = true;

        if(m_stream0 instanceof InstanceCreator){
           ((InstanceCreator)m_stream0).addClusterChangeListener(this);
           clusterEvents = new ArrayList<ClusterEvent>();
           m_graphcanvas.setClusterEventsList(clusterEvents);
        }
        m_stream0.prepareForUse();

        m_clusterer0 = clusteringSetupTab.getClusterer0();
        m_clusterer0.prepareForUse();

              m_clusterer1 = clusteringSetupTab.getClusterer1();
        if(m_clusterer1!=null){
           m_clusterer1.prepareForUse();
        }
        m_measures0 = clusteringSetupTab.getMeasures();
        m_measures1 = clusteringSetupTab.getMeasures();

        /* TODO this option needs to move from the stream panel to the setup panel */
        m_processFrequency = m_stream0.getEvaluationFrequency();

        //get those values from the generator
        int dims = m_stream0.numAttsOption.getValue();
        visualPanel.setDimensionComobBoxes(dims);
        visualPanel.setPauseInterval(initialPauseInterval);

        m_evalPanel.setMeasures(m_measures0, m_measures1, this);
        m_graphcanvas.setGraph(m_measures0[0], m_measures1[0],0,m_processFrequency);
     }

  public void run() {
           runVisual();
     }

     public void runVisual() {
        int processCounter = 0;
        int speedCounter = 0;
        LinkedList<DataPoint> pointBuffer0 = new LinkedList<DataPoint>();
        LinkedList<DataPoint> pointBuffer1 = new LinkedList<DataPoint>();
        ArrayList<DataPoint> pointarray0 = null;
        ArrayList<DataPoint> pointarray1 = null;
```

```
while(work || processCounter!=0){
   if (m_stream0.hasMoreInstances()) {
      timestamp++;
      speedCounter++;
      processCounter++;
      if(timestamp%100 == 0){
         m_visualPanel.setProcessedPointsCounter(timestamp);
      }

      Instance next0 = m_stream0.nextInstance();
      DataPoint point0 = new DataPoint(next0,timestamp);

      pointBuffer0.add(point0);
      while(pointBuffer0.size() > m_stream0_decayHorizon){
         pointBuffer0.removeFirst();
      }

      DataPoint point1 = null;
      if(m_clusterer1!=null){
            point1 = new DataPoint(next0,timestamp);
            pointBuffer1.add(point1);
              while(pointBuffer1.size() > m_stream0_decayHorizon){
                 pointBuffer1.removeFirst();
              }
      }

      if(m_visualPanel.isEnabledDrawPoints()){
         m_streampanel0.drawPoint(point0);
         if(m_clusterer1!=null)
            m_streampanel1.drawPoint(point1);
         if(processCounter%m_redrawInterval==0){

m_streampanel0.applyDrawDecay(m_stream0_decayHorizon/(float)(m_redrawInterval));
            if(m_clusterer1!=null)

m_streampanel1.applyDrawDecay(m_stream0_decayHorizon/(float)(m_redrawInterval));
         }
      }

      Instance traininst0 = new DenseInstance(point0);
      if(m_clusterer0.keepClassLabel())
         traininst0.setDataset(point0.dataset());
      else
         traininst0.deleteAttributeAt(point0.classIndex());
      m_clusterer0.trainOnInstanceImpl(traininst0);

      if(m_clusterer1!=null){
         Instance traininst1 = new DenseInstance(point1);
         if(m_clusterer1.keepClassLabel())
            traininst1.setDataset(point1.dataset());
         else
            traininst1.deleteAttributeAt(point1.classIndex());
         m_clusterer1.trainOnInstanceImpl(traininst1);
      }

      if (processCounter >= m_processFrequency) {
         processCounter = 0;
```

```
                for(DataPoint p:pointBuffer0)
                   p.updateWeight(timestamp, m_stream0_decay_rate);

                pointarray0 = new ArrayList<DataPoint>(pointBuffer0);

                if(m_clusterer1!=null){
                   for(DataPoint p:pointBuffer1)
                      p.updateWeight(timestamp, m_stream0_decay_rate);

                            pointarray1 = new ArrayList<DataPoint>(pointBuffer1);
                }

                processClusterings(pointarray0, pointarray1);

                int pauseInterval = m_visualPanel.getPauseInterval();
                if(pauseInterval!=0 && lastPauseTimestamp+pauseInterval<=timestamp){
                   m_visualPanel.toggleVisualizer(true);
                }

             }
          } else {
             System.out.println("DONE");
             return;
          }
          if(speedCounter > m_wait_frequency*30 && m_wait_frequency < 15){
             try {
                synchronized (this) {
                   if(m_wait_frequency == 0)
                      wait(50);
                   else
                      wait(1);
                }
             } catch (InterruptedException ex) {

             }
             speedCounter = 0;
          }
       }
       if(!stop){
          m_streampanel0.drawPointPanels(pointarray0,timestamp,m_stream0_decay_rate,
m_stream0_decay_threshold);
          if(m_clusterer1!=null)
             m_streampanel1.drawPointPanels(pointarray1,  timestamp,  m_stream0_decay_rate,
m_stream0_decay_threshold);
          work_pause();
       }
    }

   private void processClusterings(ArrayList<DataPoint> points0, ArrayList<DataPoint>
points1){
       gtClustering0 = new Clustering(points0);
       gtClustering1 = new Clustering(points1);

       Clustering evalClustering0 = null;
       Clustering evalClustering1 = null;

       //special case for ClusterGenerator
```

```
        if(gtClustering0!= null){
            if(m_clusterer0 instanceof ClusterGenerator)
                ((ClusterGenerator)m_clusterer0).setSourceClustering(gtClustering0);
            if(m_clusterer1 instanceof ClusterGenerator)
                ((ClusterGenerator)m_clusterer1).setSourceClustering(gtClustering1);
        }

        macro0 = m_clusterer0.getClusteringResult();
        evalClustering0 = macro0;


        //TODO: should we check if micro/macro is being drawn or needed for evaluation and skip
otherwise to speed things up?
        if(m_clusterer0.implementsMicroClusterer()){
            micro0 = m_clusterer0.getMicroClusteringResult();
            if(macro0 == null && micro0 != null){
                //TODO: we need a Macro Clusterer Interface and the option for kmeans to use the non
optimal centers
                macro0 = moa.clusterers.KMeans.gaussianMeans(gtClustering0, micro0);
            }
            if(m_clusterer0.evaluateMicroClusteringOption.isSet())
                evalClustering0 = micro0;
            else
                evalClustering0 = macro0;
        }

        if(m_clusterer1!=null){
            macro1 = m_clusterer1.getClusteringResult();
            evalClustering1 = macro1;
            if(m_clusterer1.implementsMicroClusterer()){
                micro1 = m_clusterer1.getMicroClusteringResult();
                if(macro1 == null && micro1 != null){
                        macro1 = moa.clusterers.KMeans.gaussianMeans(gtClustering1, micro1);
                }
                if(m_clusterer1.evaluateMicroClusteringOption.isSet())
                    evalClustering1 = micro1;
                else
                    evalClustering1 = macro1;
            }
        }

        evaluateClustering(evalClustering0, gtClustering0, points0, true);
        evaluateClustering(evalClustering1, gtClustering1, points1, false);

        drawClusterings(points0, points1);
    }

    private void evaluateClustering(Clustering found_clustering, Clustering trueClustering,
ArrayList<DataPoint> points, boolean algorithm0){
        StringBuilder sb = new StringBuilder();
        for (int i = 0; i < m_measures0.length; i++) {
            if(algorithm0){
                if(found_clustering!=null && found_clustering.size() > 0){
                    try {
                        double msec = m_measures0[i].evaluateClusteringPerformance(found_clustering,
trueClustering, points);
                        sb.append(m_measures0[i].getClass().getSimpleName()+"    took    "+msec+"ms
```

```
(Mean:"+m_measures0[i].getMeanRunningTime()+")");
                    sb.append("\n");

                } catch (Exception ex) { ex.printStackTrace(); }
            }
            else{
                for(int j = 0; j < m_measures0[i].getNumMeasures(); j++){
                    m_measures0[i].addEmptyValue(j);
                }
            }
        }
        else{
            if(m_clusterer1!=null && found_clustering!=null && found_clustering.size() > 0){
                try {
                    double msec = m_measures1[i].evaluateClusteringPerformance(found_clustering,
trueClustering, points);
                    sb.append(m_measures1[i].getClass().getSimpleName()+"    took    "+msec+"ms
(Mean:"+m_measures1[i].getMeanRunningTime()+")");
                    sb.append("\n");
                }
                catch (Exception ex) { ex.printStackTrace(); }
            }
            else{
                for(int j = 0; j < m_measures1[i].getNumMeasures(); j++){
                    m_measures1[i].addEmptyValue(j);
                }
            }
        }
    }
    m_logPanel.setText(sb.toString());
    m_evalPanel.update();
    m_graphcanvas.updateCanvas();
  }
public void drawClusterings(List<DataPoint> points0, List<DataPoint> points1){
    if(macro0!= null && macro0.size() > 0)
        m_streampanel0.drawMacroClustering(macro0, points0, Color.RED);
    if(micro0!= null && micro0.size() > 0)
        m_streampanel0.drawMicroClustering(micro0, points0, Color.GREEN);
    if(gtClustering0!= null && gtClustering0.size() > 0)
      m_streampanel0.drawGTClustering(gtClustering0, points0, Color.BLACK);

    if(m_clusterer1!=null){
      if(macro1!= null && macro1.size() > 0)
          m_streampanel1.drawMacroClustering(macro1, points1, Color.BLUE);
      if(micro1!= null && micro1.size() > 0)
          m_streampanel1.drawMicroClustering(micro1, points1, Color.GREEN);
      if(gtClustering1!= null && gtClustering1.size() > 0)
        m_streampanel1.drawGTClustering(gtClustering1, points1, Color.BLACK);
    }
  }
public void redraw(){
    m_streampanel0.repaint();
    m_streampanel1.repaint();
  }
public static int getCurrentTimestamp(){
    return timestamp;
  }
```

```
    private void work_pause(){
        while(!work && !stop){
            try {
                synchronized (this) {
                    wait(1000);
                }
            } catch (InterruptedException ex) {

            }
        }
        run();
    }

    public static void pause(){
        work = false;
        lastPauseTimestamp = timestamp;
    }

    public static void resume(){
        work = true;
    }

    public void stop(){
        work = false;
        stop = true;
    }

    public void setSpeed(int speed) {
        m_wait_frequency = speed;
    }

    public void actionPerformed(ActionEvent e) {
        //reacte on graph selection and find out which measure was selected
        int selected = Integer.parseInt(e.getActionCommand());
        int counter = selected;
        int m_select = 0;
        int m_select_offset = 0;
        boolean found = false;
        for (int i = 0; i < m_measures0.length; i++) {
            for (int j = 0; j < m_measures0[i].getNumMeasures(); j++) {
                if(m_measures0[i].isEnabled(j)){
                        counter--;
                    if(counter<0){
                        m_select = i;
                        m_select_offset = j;
                        found = true;
                        break;
                    }
                }
            }
            if(found) break;
        }
        m_graphcanvas.setGraph(m_measures0[m_select],
m_measures1[m_select],m_select_offset,m_processFrequency);
    }
```

```
public void setPointLayerVisibility(boolean selected) {
   m_streampanel0.setPointVisibility(selected);
   m_streampanel1.setPointVisibility(selected);
}
public void setMicroLayerVisibility(boolean selected) {
   m_streampanel0.setMicroLayerVisibility(selected);
   m_streampanel1.setMicroLayerVisibility(selected);
}
public void setMacroVisibility(boolean selected) {
   m_streampanel0.setMacroLayerVisibility(selected);
   m_streampanel1.setMacroLayerVisibility(selected);
}
public void setGroundTruthVisibility(boolean selected) {
   m_streampanel0.setGroundTruthLayerVisibility(selected);
   m_streampanel1.setGroundTruthLayerVisibility(selected);
}

public void changeCluster(ClusterEvent e) {
   if(clusterEvents!=null) clusterEvents.add(e);
   System.out.println(e.getType()+": "+e.getMessage());
}
public void exportCSV(String filepath) {
   PrintWriter out = null;
   try {
      if(!filepath.endsWith(".csv"))
         filepath+=".csv";
      out = new PrintWriter(new BufferedWriter(new FileWriter(filepath)));
      String del = ";";

      Iterator<ClusterEvent> eventIt = null;
      ClusterEvent event = null;
      if(clusterEvents!=null && clusterEvents.size() > 0){
         eventIt = clusterEvents.iterator();
         event = eventIt.next();
      }

      //raw data
      MeasureCollection measurecol[][] = new MeasureCollection[2][];
      measurecol[0] = m_measures0;
      measurecol[1] = m_measures1;
      int numValues = 0;
      //header
      out.write("Nr"+del);
      out.write("Event"+del);
      for (int m = 0; m < 2; m++) {
         for (int i = 0; i < measurecol[m].length; i++) {
            for (int j = 0; j < measurecol[m][i].getNumMeasures(); j++) {
               if(measurecol[m][i].isEnabled(j)){
                  out.write(m+"-"+measurecol[m][i].getName(j)+del);
                  numValues = measurecol[m][i].getNumberOfValues(j);
               }
            }
         }
      }
      out.write("\n");

      //rows
```

```
        for (int v = 0; v < numValues; v++){
          //Nr
          out.write(v+del);

          //events
          if(event!=null && event.getTimestamp()<=m_stream0_decayHorizon*v){
            out.write(event.getType()+del);
            if(eventIt!= null && eventIt.hasNext()){
              event=eventIt.next();
            }
            else
              event = null;
          }
          else
            out.write(del);

          //values
          for (int m = 0; m < 2; m++) {
            for (int i = 0; i < measurecol[m].length; i++) {
              for (int j = 0; j < measurecol[m][i].getNumMeasures(); j++) {
                if(measurecol[m][i].isEnabled(j)){
                        double value = measurecol[m][i].getValue(j, v);
                        if(Double.isNaN(value))
                                out.write(del);
                        else
                                out.write(value+del);
                }
              }
            }
          }
          out.write("\n");
        }
        out.close();
      } catch (IOException ex) {
        Logger.getLogger(RunVisualizer.class.getName()).log(Level.SEVERE, null, ex);
      } finally {
        out.close();
      }
    }
    public void weka() {
          try{
                  Class.forName("weka.gui.Logger");
          }
          catch (Exception e){
                  m_logPanel.addText("Please add weka.jar to the classpath to use the Weka
explorer.");
                  return;
          }
    Clustering wekaClustering;
       if(m_clusterer0.implementsMicroClusterer()                                    &&
m_clusterer0.evaluateMicroClusteringOption.isSet())
          wekaClustering = micro0;
       else
          wekaClustering = macro0;

       if(wekaClustering == null || wekaClustering.size()==0){
          m_logPanel.addText("Empty Clustering");
```

```
            return;
        }
        int dims = wekaClustering.get(0).getCenter().length;
        FastVector attributes = new FastVector();
        for(int i = 0; i < dims; i++)
              attributes.addElement( new Attribute("att" + i) );

        Instances instances = new Instances("trainset",attributes,0);

        for(int c = 0; c < wekaClustering.size(); c++){
           Cluster cluster = wekaClustering.get(c);
           Instance inst = new DenseInstance(cluster.getWeight(), cluster.getCenter());
           inst.setDataset(instances);
           instances.add(inst);
        }

        WekaExplorer explorer = new WekaExplorer(instances);
    }
}
```

Source Code C: Presentation

ZRAFDMainFrame.java, Main class of the System that creates the GUI

```
package unza.presentation;
import java.awt.BorderLayout;
import java.awt.Dimension;
import java.awt.Insets;
import java.awt.Toolkit;
import javax.swing.JFrame;
import javax.swing.JPanel;

/**
 * The main class of the System that creates the GUI
 */
public class ZRAFDMainFrame extends JFrame {

   private static final long serialVersionUID = 1L;
   private javax.swing.JTabbedPane panel;

   public ZRAFDMainFrame() {
      setTitle("ZRA-FD Outlier Graphical User Interface");
      initGUI();
      //set the frame to fill the screen
      Dimension screenSize = Toolkit.getDefaultToolkit().getScreenSize();
      Insets scnMax = Toolkit.getDefaultToolkit().getScreenInsets(getGraphicsConfiguration());
      int taskBarSize = scnMax.bottom;
      setSize(screenSize.width, screenSize.height - taskBarSize);
      this.setResizable(false);
      setLocationRelativeTo(null);
   }
   private void initGUI() {
      setLayout(new BorderLayout());

      // Create and set up tabs
      panel = new javax.swing.JTabbedPane();
      add(panel, BorderLayout.CENTER);
```

```
        //create the outlier panel
        MainTabPanel tabPanel = new MainTabPanel();
        panel.addTab("Outlier", null, (JPanel) tabPanel, "ZRA-FD Outlier");
    }
    public static void main(String[] args) {
        try {
            javax.swing.SwingUtilities.invokeLater(new Runnable() {
                @Override
                public void run() {
                    //change look and feel to nimbusw
                    try {
                        for (javax.swing.UIManager.LookAndFeelInfo info :
javax.swing.UIManager.getInstalledLookAndFeels()) {
                            if ("Nimbus".equals(info.getName())) {
                                javax.swing.UIManager.setLookAndFeel(info.getClassName());
                                break;
                            }
                        }
                    } catch (Exception ex) {
                        ex.printStackTrace();
                    }
                    // Create and set up the window
                    ZRAFDMainFrame gui = new ZRAFDMainFrame();
                    gui.setDefaultCloseOperation(JFrame.EXIT_ON_CLOSE);
                    gui.setVisible(true);
                }
            });
        } catch (Exception e) {
            e.printStackTrace();
        }
    }
}           g.setColor(default_color);
        }

        int drawSize = DRAW_SIZE;
        int drawStart = 0;

        g.fillOval(drawStart, drawStart, drawSize, drawSize);
        g.drawOval(drawStart, drawStart, drawSize, drawSize);
    }

    public void highlight(boolean enabled){
        highligted = enabled;
        repaint();
    }

    public boolean isValidCluster(){
        return (center!=null);
    }
    public int getClusterID(){
        return (int)cluster.getId();
    }
    public int getClusterLabel(){
        return (int)cluster.getGroundTruth();
    }
    public String getSVGString(int width){
        StringBuffer out = new StringBuffer();
```

```
        int x = (int)(center[x_dim]*window_size);
        int y = (int)(center[y_dim]*window_size);
        int radius = panel_size/2;
         out.append("<circle ");
        out.append("cx="'+x+"' cy="'+y+"' r="'+radius+"'");
        out.append(" stroke='green' stroke-width='1' fill='white' fill-opacity='0' />");
        out.append("\n");
        return out.toString();
    }
    public void drawOnCanvas(Graphics2D imageGraphics){
        int x = (int)(center[x_dim]*window_size-(panel_size/2));
        int y = (int)(center[y_dim]*window_size-(panel_size/2));
        int radius = panel_size;
        imageGraphics.drawOval(x, y, radius, radius);
    }
}
```