UNIVERSITY OF NAIROBI
SCHOOL OF COMPUTING & INFORMATICS

TOWARDS JUDICIAL DATA WAREHOUSING AND DATA MINING

By

MBURU RAHAB WANJIRU

(P58/63702/2011)

Supervisor

CHRISTOPHER MOTURI

A project report submitted in partial fulfilment of the requirement for the Masters of Science in Computer Science of the University of Nairobi.

September 2014

DECLARATION

The research project presented in this project report is my original work and has not been presented in any other institution. Reference is hereby made from works of other researchers that may have more insight into this project.

Mburu Rahab Wanjiru: _____   Date: _____
(P58/63702/2011)

This project report has been submitted in partial fulfillment of the requirement of the Master of Science Degree in Computer Science of the University of Nairobi with my approval as the University supervisor

Christopher A Moturi: _____   Date: _____
Deputy Director
School of Computing and Informatics

DEDICATION

To all scholars who are interested in research on ETL process, data warehousing and data mining for judicial data.

ACKNOWLEDGEMENT

DEFINITION OF TERMS

JTF: Judicial Transformation Framework which is a roadmap on how to make continuous improvement on the performance of the Kenya Judiciary.

JTF Book: a book that describes the framework for transforming the judiciary

ETL: Extract Transform Load-a three step process that is used.

ETL Process: A process that is used for populating a data warehouse by extracting data from source systems.

ETL Tools: Pieces of software responsible for the extraction of data from several sources.

Data Warehouse: A database used to store data. It is a central repository of data in which data from various sources is stored.

Data warehousing: The process of compiling and organizing data into one common database. Data warehousing is the process of compiling information or data into a data warehouse.

Data Mining: A way of making further analysis to data collected in order establish any hidden formats, trends, with an aim of improving decision making process.

Relational Database: A computer database where data is collected and organized in tables which give flexibility for manipulation in many different formats

Judicial Intelligence: The process of analyzing judicial data with an aim of quickening the decision making process.

Case Management System (CMS): An application that is used to collect case data in some courts.

ABSTRACT

Institutions have often adopted the use of Extraction, Transformation and Loading (**ETL**) of data into a data warehouse. A data warehouse is a central repository of information which can be retrieved later for analytics or other data mining related activities. **ETL** is a process that is used to take information from one or more sources, normalize it in some way to some convenient schema, and then insert it into some other repository for future use. This is usually achieved using automated tools in order to make this process easier and efficient. The choice of the ETL tool to use may be dictated by many factors among them the DBMS in use, compatibility with existing data source and hardware infrastructure among other factors.

Even thou there are major efforts in development of databases for use in various departments, currently no existing data warehouse in the Kenya Judiciary. This makes it challenging to mine data. Thus, each department with such initiative operates independently with their preferred relational database that suits only the purpose for the particular department. In other cases, some information is stored in different formats and data types or as flat files. There is need to create a central repository for data from the various departments. However, this would call for major preparation for the data collected as well as the infrastructure. One way to make preparations is to adopt the ETL process which utilizes ETL tools to carry out activities such as cleansing, customization, reformatting, integration, and insertion of data into a data warehouse. The Kenya judiciary could benefit from this process where all data collected from operational databases is collected using standardized data collection tools, cleansed and inserted into a data warehouse. The main objective of this study is to demonstrate how the Judiciary would benefit from ETL process to migrate data from a sample source system, make changes to the data and load into a specified destination ready for further exploration or analysis using other data mining tools. This research will adopt a descriptive research design as it entails systematic collection of information and careful selection of the units to be studied as well as careful measurement of each variable or unit of data collected either as entities, fields or attributes of data. In this research data will be collected   using document analysis and interviews. Presentation of results will be done by use of tables, Csv and xls files.

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION

1.1 Background

The Kenya Judiciary is structured into various departments such as Performance Management, Finance, Human Resources, Public Relations, Communication and Information Technology. The content of the Kenya Judiciary Transformation Framework (JTF) 2012-2016 call for quick administration of justice to the Republic of Kenya - thus supportive pillars have been adopted to boost these efforts where each department is expected to yield to a certain level of performance. In the JTF book, pillar 4 is the use of Information and Communications Technology (ICT) as an enabler to quick administration of justice. The proper adoption of ICT in the justice system will boost all the departmental operations with the end result being acceleration towards achieving expeditious delivery of justice. As part of efforts geared towards enhancing quick delivery of justice through adoption of ICT, the Kenya Judiciary ICT Department hopes to initiate efforts towards building a data warehouse. The warehouse once successfully implemented will act as a repository of all vital information from various relational databases for the purpose of further analysis or even generating judicial intelligence. From such centralized data repository, various reports can be drawn and further refinement and querying can be done with an aim for gaining more insight into the data collected. There is need to make improvements on how information has been collected and handled in the past.

Initiatives in data collection have yielded creation of relational databases such as the Case Management System. With such operational systems in place it is important to consider building a data warehouse in order to boost information storage and backup processes. The information collected in the data warehouse can be used to build intelligence in the data in a process known as data mining.

Performing data mining for the judicial data would imply extraction of knowledge from a repository of information or a knowledge base with an aim of discovering trends, improving decision making, developing performance indexes, understanding the judicial operations, and providing judicial data for integration with other organizations. Mining data for information entails analyzing the vast amounts of data accumulated overtime from automated processes, manual processes and electronic databases. In order to be effective in this process there is

need to consider the tools that can be used appropriately in the area of data modeling, data surveying and data preparation for the mining process.

Data preparation implies building the right model and preparing the miner implies ensuring that the right model is built. Correct data preparation prepares both the miner and the data. When this data is well prepared the data miner is able to gain proper understanding of its contents, use and applicability, limitations of the data as well as come up with quality models that can be adopted for the mining process. Success in mining judicial data calls for a three step process known as ETL (Extract, Transform and Load). This means that data is extracted from an identified source system, loaded into a staging area where various transformations are done to ensure that data is in a format that is acceptable, and loading it into a destination ready for exploration with a data mining tool. Such a source would be a relational database or otherwise. Success in data warehousing and data mining requires various considerations in the area of data preparation. This implies examining the existing data and its infrastructure, preparing this data in a format that is ready for a data miner for future references. The aim of this project was therefore to demonstrate how the ETL process can be adopted as a data pre-processing and movement initiative with the the overall objective to achieve successful data warehousing and mining.

1.2 Problem Statement

The characteristics of Kenya judicial data as it originates during the process of capturing, documentation and storage in operational databases, can make data warehousing and data mining operations challenging. The Kenya Judiciary needs to make major changes in data collection, preprocessing and transformation procedures it has adopted in the past. Before one can apply data mining to the current judicial data the nature and format of data that is collected in various operational systems and even flat files may need major transformation. There needs to be a way of extracting and storing data to allow easier reporting as well as perform predictive analytics onto these data. This process of collecting and transforming data once improved would boost the efforts geared towards achieving high levels of efficiency as outlined in the JTF book. However, this should be done without slowing down or interfering with the operational systems.

1.3 Purpose of Study

This project aims to demonstrate how the ETL process can be utilized to solve the Kenyan judiciary problem. It showcases how data can be collected using a standardized tool and

transformed with an aim of loading it into a database. The demonstration shows sample data that is referred to as dirty being extracted using a sample ETL tool, transformed and is clean awaiting loading into the final repository which is the data warehouse. For the purpose of this project the source data will be from registry files and MS Excel files. However, the data will be manipulated so as not to reflect the specific details. Using information extracted from the judiciary data files, this project proposes to demonstrate how ETL process can be utilized as a vital tool for data preparation as well as for populating a data warehouse.

1.4 Objectives

The objectives of the research were:

1. Determine common data sources and operational systems in judicial environment and display the kind of data available and the pre-processing needs.
2. Demonstrate, using case returns data and flat files data, how the ETL process can be used to migrate data from sources to a data warehouse.
3. Propose a conceptual model for the judicial ETL process with its associated architectures.
4. Develop a guideline for creating a data warehouse for the Kenya Judiciary.

1.5 Research Questions

1. What are the operational systems available in the judiciary that can serve as data sources?
2. What data is generated by the existing data sources or operational systems within the Kenya Judiciary?
3. What data that require pre-processing in order to meet the operational needs?
4. What ETL tools and techniques available for use in populating a data warehouse?
5. What frameworks that can be adopted for implementation of an ETL environment?

1.6 Project Justification

The quality of data available in the organization's data warehouse is influenced by the underlying data collection procedures, transformation, integration and extraction procedures which are aspects of the ETL process. Therefore, the following issues arising from the ETL process are important to look at keenly: The nature and quality of data received and entered from the source systems; Cleansing of bad data that may consist of missing attributes, inconsistency, and failure to conform to a given standard among others.

The guarantee to building a quality data mining model is dependent on existence of properly prepared data; which acts as a guideline to the data miner in understanding the contents of data. Inadequate data preparation and inaccurate data survey makes it difficult for data mining thus time-consuming models are built and rebuilt in an effort to understand data. Correct data preparation prepares both the miner and the data. To prepare the data implies that the model is built right, while for the miner means the right model is built. Data preparation and the data survey lead to an understanding of the data that allows the right model to be built, and built right the first time.

1.7 Scope

The project sought to utilize the data generated from the operational systems within the Kenya judiciary and data collected from a few courts in selected regions. Data collected from the registries and manually entered in the registers will be the major point of focus as it captures the real situation on the ground of how case details are captured and recorded. However this may be extended to some raw data captured from existing past records from registry files where data clerks have been hired to record it in electronic spreadsheets and store it as Microsoft Excel (.xls) and even pdf files.

1.8 Assumptions

The specific assumptions included: the registry files contains case data that is detailed enough, there would be cooperation from the parties and relevant authorities that will be sought to assist in capturing data. The application of ETL in the Kenya Judiciary will reveal new ways of collecting and storing data in a central repository where useful knowledge can be retrieved to serve the purpose of building judicial knowledge base.

CHAPTER TWO
LITERATURE REVIEW

This chapter provides an overview of ETL process, ETL tools and techniques, previous research works in data integration in judicial systems.

2.1 The ETL Process

The extraction, transformation and loading (ETL) process provides a mechanism for populating a data ware house using data from existing data sources e.g. relational databases, flat files, virtual storage access methods (VSAM) or even external sources through the network or scanning systems. All the information could be different from each other thus, before one begins the process of analysing data, it is important to consider the transformations required to the data so that it meets the operational needs. The transformation has been automated and embedded as a step of the ETL process. Ahmed kabiri and dalilachiadm (2013) illustrates the architecture of DW system in a data warehouse as having four levels (Figure 1).
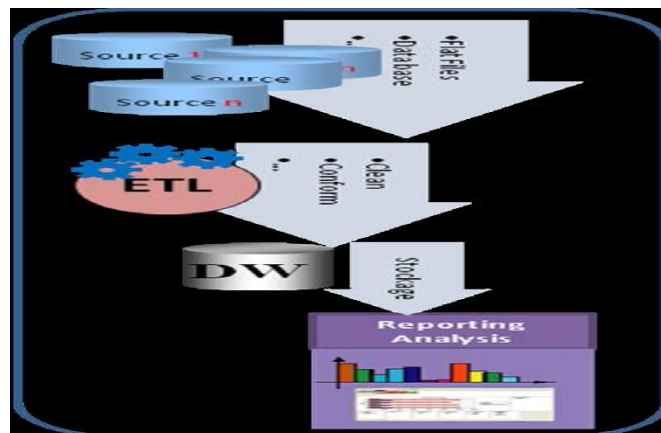
Figure 1: Data Warehouse Environment

Data Sources: These encompass all types of data sources and they act as data provider. They include databases and flat files.

ETL: The Extraction, Transformation and Loading (ETL) integration layer in DW environment. ETL tools pull data.

DW: A central repository to save data produced by ETL layer. DW is a database including fact tables besides dimension tables. Together these tables are combined in a specific schema which may be star schema or snowflake schema.

Reporting and Analysis: has the mission to catch end-user request and translate it to DW.

Collected data is served to end-users in several formats. For example data is formatted into reports, histograms and dashboard. ETL is used to migrate data from one data base to another with an aim of forming a data warehouse. ETL is a critical component in the development of a data warehouse as it consumes up to 70% of the resources. The degree of data accuracy is highly dictated by the ETL process and may to the great extent dictate the success or failure of a data warehouse. The purpose of ETL is to populate a data warehouse.

Sweety Patel et al…(2012),demonstrates an overview of the ETL process. This issue considers it very important to pay attention to data file formats, operating systems and hardware while integrating different heterogeneous platforms. ETL is a combination of process and technologies that require skills from various professionals like database designers, analysts, application developers. It's a process that keeps on changing in order to accommodate new inventions and operational needs. Therefore it should be aimed at performing as an automated process that runs as per a well-defined schedule to make updates and other necessary changes. It's also seen as a pre-decessor to data warehouse or data mart development efforts. Also it should be an independent process performed on a separate relational database server.

According to Kufandirimbwa and Kuranga (2012) - The paper argues for the importance of adopting mining of judicial data. In this issue, they illustrate how computer technology can be used to derive a data mining model. ETL tools have been used to help in the data cleansing or preprocessing to arrive at the required field. Illustrated are the data mining techniques that can be used like decision trees and online analytical processing (OLAP) to perform data analysis, perform automatic summaries and help discover hidden patterns as well as display the results to the end user. Their approach adopted use of Microsoft Structured Query Language server to create the OLAP and Microsoft excel (browser) to present the information to the end user. The data mining technique used was decision trees.

Construction of a data warehouse is seen as a three step procedure. These include acquiring data from source data where standard models may be applied such as entity relationship diagrams. The second vital step is mapping of attributes from the source data to fit the destination data attributes. These are a challenging task usually referred to as ETL process. In this area there are no standard models yet due to the nature of data complexity while

attempting to cleanse, transform or manipulate data using other operators. The third step of data warehouse construction is designing the destination area where several models have been suggested e.g. star schema, snowflake among others Shaker H. et al..,(2011).

The ETL process has been defined as a vital step in the early stages of designing a data warehouse and thereafter data mining. Due to its complexity, there are no standard models yet for adoption while designing these processes. This is because of its difficulty and lack of formal model for representing ETL activities that map the incoming data from different data sources to be in a suitable format for loading to the target Data Warehouse Kimball and Caserta, 2004; Demarest, 1997; Oracle Corp., 2001; Inmon (1997) in Shaker H. Ali et al..(2011).

Without ETL tools,making sense of raw data can be very challenging.ETL tools are very vital for business intelligence as the input provided to the BI tools dictate the correctness of the output Nils Schmidt.et al..(2011).This paper performs an evaluation of some ETL tools available in the market. The evaluation is based on a framework that uses criteria such as Price, Functionality, Ease of Use and Architecture. The findings show that Analytical reports would be lost without ETL. This dictates that data be extracted from the source, data to be transformed into a unified format and finally data to be loaded into a database or data warehouse.

While attempting to make an evaluation of ETL tools,Nils Schmidt,et al..(2011), compare and evaluate Microsoft SQL Integration Service, Pentaho Integration Services a leading open source ETL tool, and ETI High Performance Corporate Data Integration. Their objective was to use their suggested framework to rate these tools. Comparison between the two indicate that SIS has a higher demand considering the market trends and while providing an alternative solution for the high cost most people have adopted use of open source ETL tools. According to Ranjit Singh,Dr and Kawaljeet Singh (2010), the causes of data quality problems in data warehousing are attributed to the ETL process. Data sources, data integration & data profiling, Data staging and ETL, data warehouse modeling & schema design have been mentioned as the key sources of problems. This paper highlights the dimensions of data quality to include accuracy, reliability, importance, consistency, precision, timeliness, fineness, understandability, conciseness and usefulness. Data quality problems may occur when there are poor handling processes, lack of well-defined data entry and

maintenance procedures, errors emanating from the ETL process, external data that does not fit the expected organization standards.

Efforts have been done to help conceptualize the process. Several architectural designs have been defined as different people try to model an ideal ETL design. Such include the solution by Alkis et al…2008) where a conceptual model is defined using entities that are derived from UML notation. This has been used to illustrate concepts, attributes, relationships, transformations, relationships between entities as well as constraints associated with the ETL process. Their approach suggests adoption of a four step methodology which includes identifying the appropriate data sources, active candidates for data stores, attributes mapping and annotating the diagram with runtime constraints.

In their paper,Alkis et al…(2008) ,proposes a Methodology for the Conceptual Modeling of ETL Processes. This will help in analyzing the contents of the existing data sources and how they will map into the warehouse model developed. The methodology consists of well-defined steps that will assist to identify the proper dataset stores, assist in mapping the attributes while putting into consideration the existing constraints emanating from attributes of data.The purpose of ETL process is to help identify relevant information from the data source, extraction of these data, customization and integration of this data into a common format and finally load the data into a data warehouse or data mart. The information is cleansed while bearing in mind the required rules. According to these issues the framework for ETL process is as shown below:
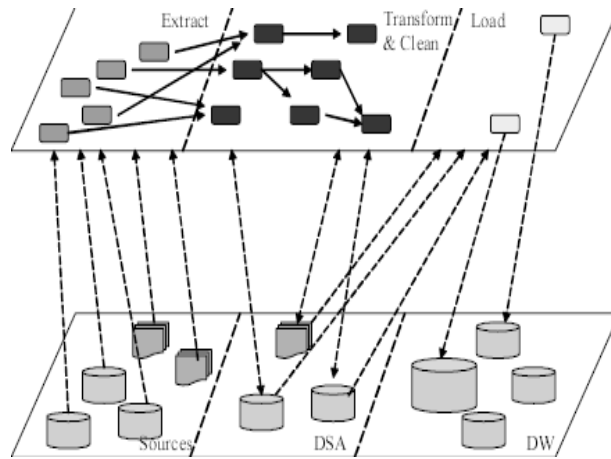
Figure 2:The Environment of ETL

In the past there have been tremendous efforts in design and development of data warehouses with ETL being an integral of the process. Such efforts have yielded the common (CRISP-DM) model.Lukasza. Kurgan, petr musilek (2006) makes a comparison of these models.Their discovery led to the illustration shown in the table below:

Figure 3: Data Warehouse Environment

Table 1 Side-by-side comparison of the major existing KDDM models

Model	Fayyad et al.	Cabena et al.	Anand & Buchner	CRISP-DM	Cios et al.	Generic model
Area	Academic	Industrial	Academic	Industrial	Academic	N/A
No of steps	9	5	8	6	6	6
Refs	(Fayyad et al., 1996d)	(Cabena et al., 1998)	(Anand & Buchner, 1998)	(Shearer, 2000)	(Cios et al., 2000)	N/A
Steps	1 Developing and Understanding of the Application Domain	1 Business Objectives Determination	1 Human Resource Identification	1 Business Understanding	1 Understanding the Problem Domain	1 Application Domain Understanding
			2 Problem Specification			
	2 Creating a Target Data Set	2 Data Preparation	3 Data Prospecting	2 Data Understanding	2 Understanding the Data	2 Data Understanding
			4 Domain Knowledge Elicitation			
	3 Data Cleaning and Preprocessing		5 Methodology Identification	3 Data Preparation	3 Preparation of the Data	3 Data Preparation and Identification of DM Technology
	4 Data Reduction and Projection		6 Data Preprocessing			
	5 Choosing the DM Task					
	6 Choosing the DM Algorithm					
	7 DM	3 DM	7 Pattern Discovery	4 Modeling	4 DM	4 DM
	8 Interpreting Mined Patterns	4 Domain Knowledge Elicitation	8 Knowledge Post-processing	5 Evaluation	5 Evaluation of the Discovered Knowledge	5 Evaluation
	9 Consolidating Discovered Knowledge	5 Assimilation of Knowledge		6 Deployment	6 Using the Discovered Knowledge	6 Knowledge Consolidation and Deployment

This provides information about the developing party (academic of industry), number of steps, main reference, and a comparison of steps across all models.In their paper they attempt to discuss the importace of deriving a methodology that supports the process that leads to finding knowledge.According to this paper,the main reason for establishing and using process models is to organize the Knowledge Discovery and Data Mining (KDDM) projects within a common framework. As a result,such models help organizations to understand the Knowledge Discovery process and provide a road map to follow while planning and carrying out data mining projects. This in turn results in time and cost savings, and in a better understanding and acceptance of such projects.The unique discovery is that all the models

discussed above include data preparation as a key step towards achieving successful data mining. This steps acts as a key step before modeling which calls for emphasis on proper approach towards use of tools, techniques and approaches to the data preparation process. The modern approach to the data preparation is through adoption of ETL tools that help to automate the preparation process.

(Pete Chapman (NCR) ,et al…,2002) in his article describes how a data mining project undergoes several phases. These are described in detail in the Cross Industry Standard Process for Data Mining (CRISP-DM).Hence forth other participants have made tremendous efforts in developing other models that integrate the ETL process as a stage in data warehouse development.

2.2 ETL Tools

There have been tremendous efforts in automation of the data integration process for the data warehouse. Such efforts have yielded to a variety of tools for automating this critical process. In their paper, Ahmed and Dalila showcase these tools as open source and commercial categories. Each tool come with its own notations, metadata and GUI and is not interchangeable. Further the categories are subdivided as those that are free with purchase of DBMS software and those that are purely payable. Selecting a tool for use is an individual preference depending on the nature of the data source and the capability of the tool to acquire data from the existing data source. Different people have adopted either the open source or commercial tools depending on what they want to achieve while populating the data warehouse.

There are other ETL tools available in the market and can be adopted for data integration based on the operational needs of the organization. The following table provides a listing of these tools with their respective companies. However the order does not necessarily dictate their popularity rating.

	ETL Tool	Version	Vendor
1.	Oracle Warehouse Builder (OWB)	11gR1	Oracle
2.	Data Services	XI 4.0	SAP Business Objects
3.	IBM Infosphere Information Server	9.1	IBM
4.	SAS Data Integration Studio	9.4M1	SAS Institute
5.	PowerCenterInformatica	9.5	Informatica
6.	Elixir Repertoire	7.2.2	Elixir
7.	Data Migrator	7.7	Information Builders
8.	SQL Server Integration Services	10	Microsoft
9.	Talend Studio for Data Integration	5.2	Talend
10.	DataFlow Manager	6.5	Pitney Bowes Business Insight
11.	Pervasive Data Integrator	10.0	Actian (Pervasive Software)
12.	Open Text Integration Center	7.1	Open Text
13.	Oracle Data Integrator (ODI)	11.1.1.5	Oracle
14.	Data Manager/Decision Stream	8.2	IBM (Cognos)
15.	Clover ETL	3.4.1	Javlin
16.	Centerprise	6.0	Astera
17.	DB2 Infosphere Warehouse Edition	9.1	IBM
18.	Pentaho Data Integration	4.1	Pentaho
19	Adeptia Integration Suite	5.1	Adeptia
20.	DMExpress	5.5	Syncsort
21.	Expressor Data Integration	3.7	QlikTech

2.3 Use of Data Mining in Judicial System

Scott and Thinyane (2013) designed an electronic Judiciary (e-Judiciary) system with the aim of addressing the existence and continuation of digital divide in traditional justice administration that governs traditional authorities for rural and marginalized communities of South Africa. Such system practices customary law as a method of providing justice to the community at large but one of the major disadvantage of such a system is the lack of a proper safe-keeping environment for judicial information making dissemination of vital information to the community difficult. The web-based system provides a better platform for the administration of minor offences that are solved by the traditional courts thereby communities with understanding and knowledge about judiciary operations.

Thammaboosadee and Silparcha (2008) proposed a framework for criminal judicial reasoning system using data mining techniques to determine reasons from court verdicts.
They developed a system with modules based on legal rules and principles that are used to construct knowledge system that uses data from the Thailand Supreme Court verdicts on criminal cases as training data set. They demonstrate the use of XML standard in document structuring and in supporting judicial decision support system that guides in judgment

supported by law theories and principles. This paper addresses one of the most complex legal activities in court level which is judicial reasoning.

Verma, Srinivasan, and Sapra (2014) proposed a conceptual model for Decision Support Systems based on Data Mining and Case-Based Reasoning comprising a knowledge base, case base reasoning and data mining subsystems. The model has capacity to self-learn, identify association between data, classifying and clustering of the data based on the characteristics, suggest recommended actions to users, all these factors incorporate intelligence into the system thus increasing the capacity of problem solving and improve suggestion accuracy. They demonstrate a relation between the Data Mining, Case-Based Reasoning System and Decision Support Systems.

Criminology is one of the most important fields where the application of data mining techniques can produce important results. Hussain, Durairaj and Farzana (2012) proposed criminal behavior analysis using data mining techniques. Their model interlinks the common set of universal principles with the attributes of the individuals for profiling the criminal behavior including personality, criminal activity, extrinsic factors and the attributes of individuals.

Ni et al. (2004) examined the possibility of using data mining in a case-based system and propose a case-based system framework based on data mining techniques.

Costake (2001) highlighted some key problems of the automation of the judicial system as part of the e-Government. They proposed a model of the judicial system on the basis of which information and communications technology can be used to transform judicial systems.

Xie (2008) analyzed case guiding system and the necessity of setting up such a system with the aim of saving judicial resources, promoting judicial justice, and improving judicial efficiency. The paper outlines the management of the overall design of case guiding system including the outline of the system, the composition of every subsystem, each function module, and the basic technical requirement of the information database and systematic operation characteristics.

Annapoorani and Vijaya (2013) developed a tool for sorting and summarizing judicial data simultaneously into various groups based on the title and thus making it simpler and easier in searching for the required data.

Ticom, de Souza and de Lima (2007) present a labor- related judicial application of unstructured data mining techniques using probabilistic methodologies, linear regression and rule induction in text categorization. This text mining is essential in discovering previously unknown patterns useful for judicial applications.

-Hamin, Othman and Mohamad (2012) outlined the technical applications available in the Malaysian courts and highlighted how the growing adoption of ICT by the courts in many jurisdictions raises several issues of security risks. They examine the emergence of non-legal risks closely connected to security and the legal risk raised by courtroom technology and how it can be mediated through the understanding basic security control such as authentication, non-repudiation, confidentiality, data integrity and privacy encroachment. The paper argues that despite the common notion that law will always lag behind technology, the courts would need to understand these legal and non-legal security risks and manage them in the most efficient and effective manner if judicial business were to be continually advanced by the ICT.

Zernik (2010) discusses how data mining was conducted through the online public access system to examine the validity and integrity of records and of the system as a whole. They report how many records were not verified, records were universally missing their authentication counterparts as required by law to render them valid and effectual. The system as a whole was deemed invalid. The public was unable to discern the difference. The author asserts that case management systems of the courts must be subjected to certified, functional logic verification.

Winn (2009) affirms that the aggregate data in court files should be used to promote better management of the judicial system, promote greater understanding and opportunity for civic involvement in the judicial process, managed to reduce the cost of access to core judicial records, facilitate empirical research into operation of the judicial system, and to enhance positive network effects of the information. Information management policy should focus on encouraging public participation in the judicial process and discouraging practices which

undermine the administration of justice. Potential solutions are multifaceted, interrelated, require new and creative ways, and must be coordinated by using all the tools available such as rules, training, and technology, in order to further the twin goals of public access and privacy within judicial information management.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

The first task involved the collection of user requirements. The second task involved the analysis of the structure and content of the existing data sources and their intentional mapping to the common data warehouse model. Related literature and personal experience suggest that the design of an ETL process aims towards the production of a crucial deliverable: the mapping of the attributes of the data sources to the attributes of the data warehouse tables. The proposed conceptual model for ETL processes complement a set of design steps, which lead to the basic target, i.e., the attribute interrelationships. These steps constitute the methodology for the design of the conceptual part of the overall ETL process and could be summarized as follows:

1. Data Extraction and preparation
2. Data Transformation
3. Loading data into the warehouse

Stage	Description	Tasks involved	Operations
Data Extraction/Data Preparation	This stage involves getting the data and describing it using metadata references. It also calls for further exploration with various tools as one attempts to verify its integrity.	Getting the data from the identified source systems.	Read Excel Read CSV Read Database
Data Transformation	This stage involves: Collecting and categorizing data into respective regions they came from. Cleaning the data and reconstructing it such that it's ready for exploration with an ETL tool. The data is formatted using appropriate criteria	Inserting and removing unnecessary columns and rows. Filling in the gaps with the missing attributes. Introducing new attributes that will contribute to the completeness of data. Refining the metadata reference tables. Creating	Data cleansing Replace missing values Fill Data Gaps Remove unused/unnecessary values Filtering -Remove duplicates Sorting -Sort Ascending/Descending -Shuffle
Loading	This involves transferring data	Write Database Write Excel Write CSV	Exporting the transformed data to an a different destination.

3.2 Research Design

This project adopts a descriptive study and its main objective was to provide accurate and valid representation of data. The research is well structured to include all the aspects of the variables described in order to answer the research questions outlined in the research question section. The research design will entail generating primary data through use of interviews as well as analyzing existing data that has been collected overtime in form of court case details.

3.3 Data Sources

There have been tremendous efforts to introduce independent systems such as the case management, audio visual recording of court proceedings, and the judicial help desk system all these as efforts geared towards improvement of court operations as well as information management within the Kenya Judiciary. Judicial data is mainly accumulated in many different formats depending on which system is in operation. It's however necessary to do a detailed analysis to find out what type of data is accumulated in each system and the possibility of extracting data from each, making necessary modifications and loading it into an analytical database or even warehouse in preparation for mining. Possible sources of data include the Case Management System as a source of data which in its current status, stores information in tables and generates various reports based on the specifications of the user. However this system has not been adopted in all the court station as it does not provide a standardized way of collecting data that is applicable to all court registries. The reports generated are usually in form of pdf and excel formats and their creation are limited to qualified IT personnel who log into the system using a user name and password and generate the report based on what one wants to look at a particular time. However, it is important to improve on the data collection procedure and standards so that it is simple and clear to users who may not necessarily be ICT experts. Also standardization of data collection will have an overall effect on the reporting capabilities of these operational system(s) with an aim of providing top decision makers and ICT department staff capabilities in accessing, analyzing and reporting data with little or even no help from the IT staff. These can be achieved using an ETL tool.

3.4 Data Collection Techniques

In order to accumulate enough data for use in this project the data was collected from two areas.

- Daily Case returns entries. This is data collected from the registries in the court register books on daily basis. Comprises of details of cases

- Historical data available in the performance department as excel files that have been sent through email for their analysis.

Two major techniques were used in order to collect the daily case returns data from the registries. They include:

Interviews

Court registries are charged with the responsibility of collecting and storing data as they register the court cases details on daily basis. The staff working in the registry is target for interview. This includes the court clerks, registrars, executive officers and other relevant authorities. Unstructured interviews will be employed in obtaining information from the judicial officers who may have slightly more time at their disposal. Whenever possible, or agreed upon the interview responses shall be recorded in writing. Such officers will include the registrars and the head of departments. In addition a set of pre-determined questions will be put to the head of departments ICT director and Deputy Director, and their responses recorded in writing. The interview questions will be intentionally ordered such that succeeding questions build on preceding ones.

Document reviews

This method will entail thorough scrutiny of any existing documentation on the existing systems such as the registers that are used to enter case application on daily basis.The documents include user manuals,system documentation or any other relevant document which hosts information that relates to an operational system.

Observation

Frequent visits will be done to the various registries to see how registers are filed.This also will entails attending court proceedings so as to identify the possibility of recording data in real time by the court clerks.

3.5 ETL Tasks

This comprises of three steps process as described below.

Data Preparation and Extraction

This task involves gathering data from external sources and bringing this data to your systems and databases. The data in question is located in spreadsheets that aren't integrated with any master database. The goal of this task is 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. Data required for this project was obtained from registers that are entered daily from the courts registries. This data was then recorded in excel worksheets. This data was collected together, merged into one excel sheet. Where applicable the data was merged using Rapid Miner which is a data mining tool with ETL capabilities. Prior to extracting data, it needs to be formed in such a way that it can be acceptable for extraction using any preferred ETL tool. Before the proposed solution the scenario in the judiciary was such that data was collected and sent through an email system to the respective destination at the headquarters as shown below.
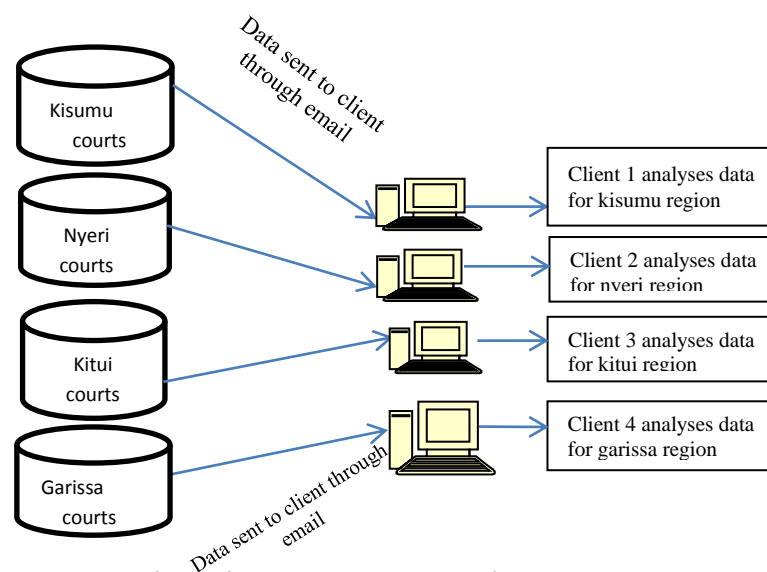
Figure 4: Data Warehouse Environment

The Staging Area

Before running an ETL process it's important to have an intermediate storage area whose main purpose is to help in the data processing. This sits between the data sources and the final destination for the data which is the warehouse. This area keeps on being refreshed by new updates of files and data before, during and after a successful ETL process. There are staging area architectures, however, which are designed to hold data for extended periods of time for archival or troubleshooting purposes. These architectures can be implemented in form of tables in relational databases, text-based flat files e.g. XML files stored in file systems or proprietary formatted binary files stored in file systems. The staging area used for this project is SQL server and MS Excel application. Data collected from the registers and recorded in excel worksheets have a lot of inconsistences because the person in charge has to count the case of a given instance manually. Therefore it generates possible errors such as:

- Modification either intentionally or non-intentional by a third party.
- Sending of these data is through email system and when the email system is down there is a lot of delay in sending data.
- The data collected goes through many third parties thus possible introduction of chances of modification.

Data Transformation

This stage calls for use of a variety of software tools and even custom programming to manipulate the data so that it integrates with data you already have. For example, consider data that is collected with errors such as missing attributes, data gaps, unnecessary values, and duplicates among others. It's necessary to transform this data and make the corrections so that it matches up with the data that currently resides in your databases and systems. Data that is collected from the registry can be referred to as "dirty". This is because it is subject to a lot of modification before it's loaded into the staging area. The following are erroneous issues relating to the data:

- Missing values/attributes-When the excel files are sent they are found to be lacking come values such that it makes it difficult to even import them into the database. e.g some fields like date are left open which translates to data irrelevance as an application has to be filed on a certain date. Such fields need to be filled with the correct data probably from the physical register.

- Introduction of new attributes-This will be introduced to achieve to help censor the data so that the information is not displayed as it is originally from the register.
- A possible modification is as shown below:-

| NAI 17 (UR 9) |
| NAI 22 (UR 13) |
| NAI 27 (UR 14) |
| NAI 36 (UR 19) |

The above is a unique number indicating the application number. These will be transformed into an auto generated number that acts as a unique identifier to the field.

Grouping entities-This is done in the field JUDGE where instead of their names, CASE DETERMINANT ID will be assigned and displayed instead of their names.

Format type-The date field will be assigned to a specific format which is dd/mm/yyyy.

Loading Data

The transformed data will be moved from the staging area to the warehouse which is a database that is in SQL server ready for mining. When data is successfully transformed it's important to load it into your system or database. However, before you load the data, it's important to have a backup of your current system or that you can roll back or undo the import if it fails. After loading the data, it's common to run audit reports so you can review the results of the merged databases and systems to make sure the new data hasn't caused any errors or problems.

3.6 Hardware and Software Requirements

Extraction, Transformation and Loading Tools –They are designed to aid in the process of extracting data from the different sources, convert it into useful forms and thereafter loading the data into a data warehouse. Also they can be used to update the data warehouse periodically. Some tools offer all inclusive services meaning they have built in capacity to complete an ETL cycle without need for combination with other applications.

Hardware Requirements

This may be determined based on different parameters which include:

1. The available number of data sources

2. The actual file size

3. RAM demands and usage pattern

For hardware, any server or computer cable of running a minimum of the 64 bit edition of Windows 7 operating system is recommended. Also, other operating systems like Linux versions can also be used as long as there are no compatibility issues with the software requirements.

Software Requirements

The judicial data warehouse will require use of a strong database that has the capacity to accommodate integration and analysis services. For example, Microsoft SQL Server, Oracle among others. The integration services are necessary for performance of data extraction, transformation and loading into the data warehouse. The analysis services engine is necessary for performance of actual data mining and reporting tasks on the data mining process. For this project the software that will be adopted for use are MS SQL Server 2005/2008 as the DBMS with Rapid Miner as the ETL tool. Microsoft Excel 2010 is the preferred application for use in uploading data into a centralized database. Ubuntu operating system and Windows Operating system are the preferred platform for operation. However several other platforms are going to be tried to experiment which will work optimally for judicial data. Such will include possible explorations with postgres data base and Rapid Miner as the ETL tool.

3.7 Exploring the data from the data sources

For this project the main sources of data is the data collected from the handwritten register that is usually stored in the courts registry. Part of this data is captured differently by various court registries and entered in excel file on daily basis in form of case returns. However, vital information oftenly goes uncollected thus remains in the registry files.As part of computerization efforts the data has been collected and filled in excel sheets which consist of a series of columns and rows organized into a tabular format. The sample excel sheet is as shown below

APPLICATION NO	DATE FILED/ R	URGENCY	APPLICANT	RESPONDENT	NATURE OF APPLICATION	STATUS	JUDGE (S)
					2005		
NAI 331	9. Dec. 2005	URGENT	CHURA (KENYA) LIMITED	SOSPETER	EXTENSION OF TIME	ALLOWED COSTS IN THE APPEAL ON 30/3/2006 AT KISUMU	SUGARS, JA (IN CHAMBERS)
NAI 336 (UR 205)	16. Dec. 2005	URGENT	MASTER BOND KENYA LTD	JULIUS	STAY OF EXECUTION	DISMISSED WITH COSTS ON 31/3/2006 AT KISUMU	KOPIYO, KINYAJUI, JJA.
NAI 339 (UR 208)	22. Dec. 2005	URGENT	INVESTERS	CORNELIUS	FILE AN APPEAL	DISMISSED ON 30/3/2006 AT KISUMU	SUGARS, JA (IN CHAMBERS)
NAI 349	20. Jan. 2006	URGENT	THE WORKAHOLICS	MWANIKI MWANIKI	FILE NOTICE & RECORD OF	ALLOWED ON 19/6/2006 AT KISUMU	JAMES, JA (IN CHAMBERS)
NAI 350	20. Jan. 2006	NOT URGENT	ROAD BUSTERS	CUCU	DISPENSE WITH SERVICE	ALLOWED ON 27/3/2006 AT KISUMU	WAKIAMA, JA (IN CHAMBERS)
NAI 351	20. Jan. 2006	NOT URGENT	ESTHER NZILANI	MUSA	SUBSTITUTION	STRUCK OUT WITH NO ORDER AS TO COSTS ON 17/5/2006 AT NY:	MINNEH, JA (IN CHAMBERS)
NAI 352	25. Jan. 2006	NOT URGENT	COLLINS ODHIAMBO	INDIMULI	EXTENSION OF TIME	DISMISSED WITH COSTS ON 19/6/2006 AT KISUMU	LINET, JA (IN CHAMBERS)
NAI 353 (UR 214)	25. Jan. 2006	NOT URGENT	COFFEE BOND	POPPER LTD	EXTENSION OF TIME	ALLOWED ON 30/3/2006 AT KISUMU	WAKIHURI, JA (IN CHAMBERS)

Figure 5:Sample raw Data from an excel sheet

The columns include: *APPLICATION NO*(a unique number that identifies the applied); *DATE FILED*(the date when that suit was filed); URGENCY (used to show the nature of the application);RESPONDENT(the person answerable pertaining to the file application);NATURE OF APPLICATION(further description to the suit);STATUS(indicates progress);JUDGE(person making the ruling on the suit).

The above data has been separated using delimiters. If the data was to be used as is for the ETL process it would incur some overhead in locating them every time they are processed (unlike fixed-width formatting), which may have performance implications. However, use of character delimiters (especially commas) is also a crude form of data compression which may assist overall performance by reducing data volumes — especially for data transmission purposes. Use of character delimiters which include a length component (Declarative notation) is comparatively rare but vastly reduces the overhead associated with locating the extent of each field. Therefore, the above data as is requires additional considerations. For example

- **Data types:** The data requires to be assigned a specific data type. Also there should be specific constraints which restrict these data to a specific data type. Such restrictions are usually established by convention, but not formally indicated unless the data is transferred to a relational database system.

- **Separated columns:** In the above example, individual columns are separated using whitespace characters. This is also called indentation or "fixed-width" data formatting. Another common convention is to separate columns using one or more delimiter characters.

3.8 Conceptual Framework

The following diagram illustrates how the ETL process can be carried out in the judiciary using the operational systems existing as data sources. It further includes all other possible sources from systems that are being implemented or are in the process of implementation. The extraction, transformation and loading tools will be chosen based on the user preference from a list available. The data once transformed is deposited in a knowledge repository or a data warehouse that will house all the different types of data. Thereafter, data mining and analysis tools will be used to generate further insight or gather more intelligence.

CM   JHD   IFMIS

Data Extraction using choice tools.e.g Rapid miner

Staging Area Database running SQL server

Transformation

Splitting

cleansing

Case returns data collected at court stations and shared over intranet to the fileserver

Fileserver

ETL Tool

Central repository for all Data from the client

Case returns data collected at courtstation Y

The Judicial Data warehouse

Data mining

Figure 6: Conceptual model for Judicial ETL Process

CHAPTER FOUR
RESULTS AND DISCUSSION

The chapter outlines the results of the activities that were set out to be achieved in the objectives. This includes identification of data sources available in the judiciary that can provide a platform for data extraction, determination of data pre-processing needs in preparation for mining, and identification of an ETL tool for collecting data and analyzing it.

4.1 Available Data Sources

Court Registry Books

Different court registries have a register where applications filed on daily basis are written down manually. The registries include
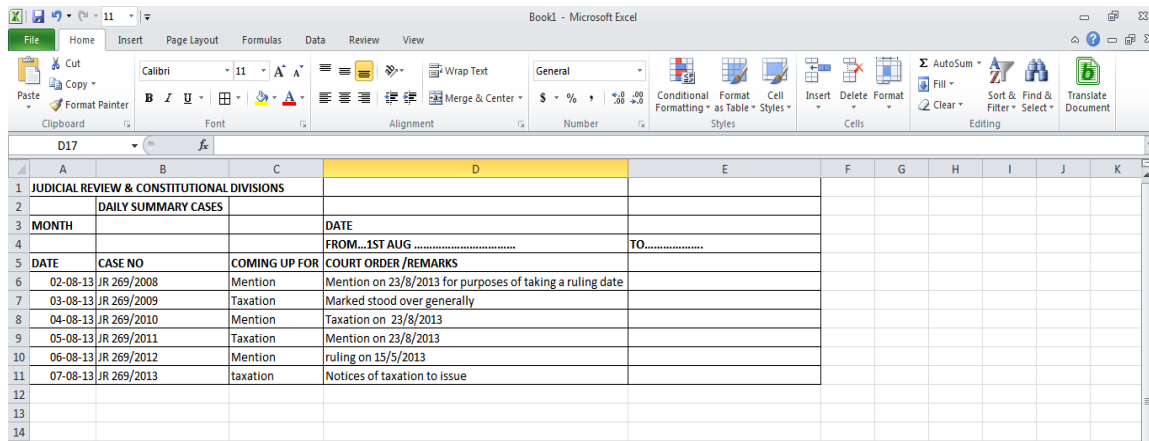
   i.   Civil registry
   ii.   Criminal registry
   iii.   Judicial Review and constitutional matters
   iv.   Family registry
   v.   Traffic registry

Different court clerks in various registries record data differently.Sample data collected from the the judicial review and constitution registry book has previously been recorded in a register as shown below.

Case No	Plaintiff	Defendant	Nature Of Claim	Filing Date	Amount Paid	Date of hearing
133/12	Rahab	XYZ	Land related	12/01/2013	500	2/03/2013

Excel Sheets

There has been tremendous attempts by the staff to record case details in the computer using MS Excel application. Such efforts have yielded many different excel sheets each with its own format depending on the activities they are involved in. A sample is as shown in Figure 7 below

Figure 7: Sample from judicial review and constitutional divisions

Registry Files

Thes are physical files that are located in the court registries.They contain a collection of vital documents that describe how a case is progressing from the time it was initiated. Usually the files are bulky, sometimes thay are old and tattered especially the ones belonging to a case that has that been there for long.

Case Management System

The Kenya Judiciary handles various categories of cases such as civil cases, criminal cases, civil applications, civil appeals, criminal appeals and election petitions. This information is captured in a database for easy management and manipulation. The Case Management System (CMS) system was initiated to help in the management of election petition cases. However, efforts are still underway to see its continuous improvement to help capture case details and record them in a relational database. The database has been developed using Postgress database. Thus, an efficient CMS has the potential to record important information relating to cases, proceedings, rulings and judgment and generate basic reports on some of the queries that may emanate from the management or even users of the system. Then information gathered thereof can be useful for future efforts geared towards performing

evaluation of scale efforts, monitor progress towards achieving set goals like finishing backlogs as well as put measures in place to improve in the manner of delivery of justice.

Judicial Help Desk System

This sytem was initiated in the office of the ombudsperson as an effort by the Chief Justice to engage with the public. This is a complains system where any person with a issue to do with the judiciary raises it. The system provide a forum where clients and members of staff in the Judiciary could report complaints including those of unethical conduct, harassment and conflict of interest.

IFMIS

The Integrated Financial Management System (IFMIS) is an information system that tracks financial events and summarizes financial information in the judiciary. The system provides critical support for management and budget decisions, fiduciary responsibilities, and the preparation of financial reports and statements.

Document Management System

This is an open source system that was initiated to boost the process of recording documents. The system allows scanned documents to be stored in a document library hosted in a server machine.

4.2 Data Collection

The data recorded in Excel sheets is sent to the Head Of Station or Executive Officer in the name of daily court returns. The Executive Officer compiles the returns by merging the sheets into a monthly court return. The excel workbooks are sent via email to the Perfomance Department by a third party who could be the Stations Executive Officer, ICT Officer or other person delegated to the task.

The following challenges relating to data collection were identified:
1. Lack of a standard way of collecting data across all registries and thus the process has been rejected by some court stations.

2. Some field computations are done manually thus a possibility of producing eroneous information.

3. Physical files pertaining to a case application are usually stored in the registry department.These files contain alot of information that is oftenly not captured in the registers. Mostly this has to do with financial aspect of the cases eg. fines, bails etc. In addission, the details such as the charge sheet information are not captured.

4. Any judicial staff is given the responsibility of updating the data on daily basis.This data is hand written as its acquired over the counter meaning it is subject to modification without trace.

5. The data collection chain is too long because of the many people involved.

4.3 STAT 1 and STAT 2

Notably the Perfomance Management Department have initiated efforts to provide a template for use across the registries commonly known as STAT 1 and STAT 2. The data propagation procedure entails acquiring non standardized data from the data sources which is the excel files. The excel files are sent to the departmental heads who later send them as email attachments to the judicial officer in the performance management department's client workstation. The data has to be downloaded by login into the email account and each person downloads the email that is associated with the courts in his region assigned. The analysis procedure is carried out using Excel application. And the reports are written down using Microsoft word. However there is no available database even in the performance management department thus there are massive excel sheets in the individuals client machines depending with the region they are handling. The template is as shown in Figure 8.

REPUBLIC OF KE

THE JUDICIARY STAT

SUMMARY STATISTICAL RETURN O:

Kadhi court at ....District...County........

Name(s) of the Kadhi(s)..

A. Summary of cases for the month of .... Year ....

Type of case	Number of cases outstanding at the start of the month (Age of outstanding cases are in months)							Number of cases initiated during the month	Number of Male Litigants (defendants and plaintiffs where applicable depending on whether its a criminal/ civil case)	Number of Female Litigants (defendants and plaintiffs where applicable depending on whether its a criminal/ civil case)	Number of cases initiated during the month and resolved	Number of cases outstanding at outstanding		
	< 3	3 to < 6	6 to < 9	9 to < 12	12 to < 24	24 to < 60	60 and above					< 3	3 to < 6	6 to < 9
Marriage														
Divorce														
Succession														
Other cases														
Total	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 8:Template from performance department

The following challenges were identified with regard to the template

i. The template is complicated to the users and its use requires expertise in data analysis.

ii. The template is tedious as it requires one to count cases manually from the court register in order to fill the fields required thereof.

iii. The template does not capture all the case details pertaining to all cases

iv. There is no existing database to store the templates data before and after manipulation

v. Merging the independent Excel sheets from all regions and all the court stations for analysis is often a tedious job.

4.4 Data Pre-Processing

Data preparation is a key aspect of transforming this data into a manner that can be transformed using an appropriate ETL tool. Since there is no agreed upon standard of collecting and recording the data, it becomes hard to export the data or import it because of the following challenges:

1. Omitted fields
2. Blank fields

3. Repetition

4. Typing errors and spelling mistakes

4.5 Proposed Data Collection and Analysis Solution

In order to curb the above challenges the following approaches may be adopted while attempting to solve the problems associated with data collection and analysis.

1. Acquire a standardized data collection tool that can be used to collect case details across all the registries. This tool should be sharable between the Performance Department and the originating registries so as to avoid delays in sending data especially when the email server is not operational.

2. Acquire a proper ETL tool that is applicable to the scenario e.g. Rapid Miner server software.

3. Procure a database management system that can be used as a repository for data.(considerations could be done to acquire an all-inclusive database that comes with Business Intelligence tools like e.g. SQL server.

4.6 Standardized Data Collection Tool

The collection tool is an Excel worksheet whose fields are organized to reflect the contents of the documents as they appear in the original files. Vital information will be extracted and filled in this worksheet. To avoid errors data validation will be enhanced at input so that there is a drop down list that acts as a guideline to selecting fields that are pre-known. This application will ease the data collection and retrieval process. Once filed the worksheet is shared using share point over the intranet architecture to the end-user who is the performance analyzer.

Figure 9a: The standardized data collection tool

	O	P	Q	R	S	T	U	V	W	X	Y
1	Def_Adv	Date _of_plea	Court_station desc	Coming up for	Judicial Officer	Designation	Witnesses	Court Decision	Next Date	Amount	Receipt_No
2	Null	12-01-13	Milimani High Court	Hearing	Hon Justice Mueni	senior magistrate	3	Fined or 3months in prison	0	1,000.00	RCPT1
3	Null	02-01-13	Milimani Magistrate cour	plea	mnn	Seniour resident magistrate	0	cash bail	12-07-14	20,000.00	RCPT2
4	Null	03-01-13	Kibera Magistrate court	Judgement	Justice mueni	senior magistrate	0	life sentence	0	-	0
5	Null	04-01-13	Kibera Magistrate court	plea	js.kioko	senior magistrate	0	case withdrawn u/d 210	0	-	0
6	Null	05-01-13	Milimani Magistrate cour	Hearing	TALINYA ABDALLA ISSA	senior magistrate	1	case withdrawn u/d 211	0	-	0
7	Null	06-01-13	Kibera Magistrate court	ruling	TALINYA ABDALLA ISSA	Seniour resident magistrate	1	case withdrawn u/d 212	0	-	0
8	Null	07-01-13	Milimani High Court	Judgement	BAIBA BHIDHA MJIDHO	principal magistrate	1	case withdrawn u/d 213	0	-	0
9	Null	08-01-13	Milimani High Court	Judgement	MOHAMED SHEIGH	senior principal magistrate	1	case withdrawn u/d 214	0	-	0
10	Null	09-01-13	Milimani High Court	plea	PETER KINYUA MUCHERU	senior magistrate	1	case withdrawn u/d 215	0	-	0
11	Null	10-01-13	Milimani Magistrate cour	Judgement	honJustice Richard	senior magistrate	0	case withdrawn u/d 216	0	-	0
12	Null	11-01-13	Kibera Magistrate court	Judgement	honJustice Joslene	senior magistrate	0	case withdrawn u/d 217	0	-	0
13	Null	12-01-13	Kibera Magistrate court	Judgement	honJustice Richard	senior magistrate	0	case withdrawn u/d 218	0	-	0
14	Null	13-01-13	Milimani High Court	Judgement	honJustice Oliech	senior magistrate	0	case withdrawn u/d 219	0	-	0
15	Null	14-01-13	Kibera Magistrate court	Judgement	honJustice Mwero	senior magistrate	0	case withdrawn u/d 220	0	-	0
16	Null	15-01-13	Kibera Magistrate court	Judgement	honJustice Laliel	senior magistrate	0	case withdrawn u/d 221	0	-	0
17	mueri & co adv	01-01-14	Kibera Magistrate court	Hearing	sgfdhgfhf	magistrate	Null	dismissed with costs on 13/10/2014 at eldoret	13-10-14	15000	RCPT3
18	mutula & sons adv	02-01-14	Milimani High Court	Hearing	gdrddfgffg	magistrate	Null	marked as withdrawn on 20/9/2014 at eldoret	20-09-14	0	RCPT4
19	wakili adv	03-01-14	Milimani High Court	Hearing	frgdtfdtfduy	magistrate	Null	allowed costs in the appeal on 20/9/2014 at eldoret	20-09-14	10000	RCPT5
20	loloo co& adv	04-01-14	Milimani High Court	Hearing	jffffyfuy	magistrate	Null	allowed costs in the appeal on 27/4/2014 at nairobi	27-04-14	9000	RCPT6

Figure 9b: The standardized data collection tool

Characteristics of the Tool

  i.   Simplicity -The application used to prepare the tool is already widely used by the court staff therefore they do not have to go through intensive training to adopt and use the tool.

  ii.   Costs - Since it is already available and installed in the clients' workstations no much additional costs will be incurred thus reduce budgetary concerns.

  iii. All inclusive - The tool has attempted to capture all court processes pertaining to criminal and civil cases.

  iv.   Contains data validation constraints and drop down lists that allow one to select some fields from a pre-defined list to avoid typing errors.

  v.   Includes added fields that help to capture detailed data from the registry files which could be useful even to other court users or third parties such as intelligence department, police among others.

4.7 The ETL Tool

The Extract, Transform, Load (ETL) tool allow the user to receive files or other data from vendors or other third parties which you need to manipulate in some way and then insert into your own database. In the context of this project, the main role of this tool is to help extract the data that was collected using the standardized excel sheet and perform possible manipulations to the data and thereafter save the work in a database for a future miner. Rapid Miner has been adopted as the ETL tool of choice. The standardized template can be extracted directed using ETL process operators available in Rapid Miner. These include processes such as Read Excel, Readcsv, Read Database

Reasons for adoption of Rapid Miner

  i.   Rapid Miner supports all steps of the data mining process from data loading, pre-processing, visualization, interactive data mining process design and inspection, automated modeling, automated parameter and process optimization, automated feature construction and feature selection, evaluation, and deployment.

  ii.   User friendliness - Rapid Miner can be used as stand-alone program on the desktop with its graphical user interface (GUI).

  iii. Rapid Miner can be used as a data mining engine for your own products and Java library for developers.
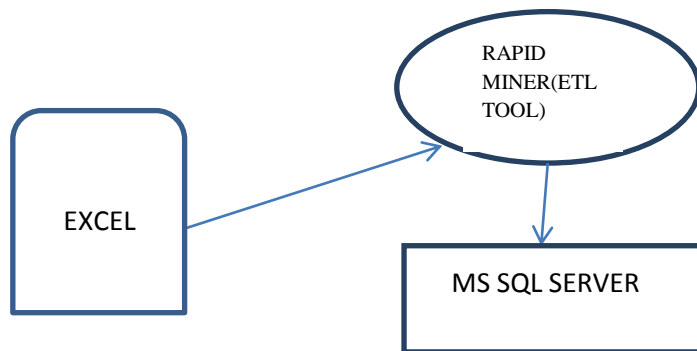
Figure 10:ETL Flow Chart

4.8 Data Description

Data that was earlier collected using non standardized data collection tool requires further modification so that it's in the right format for exploration using the ETL tool. The format of the table is organized such that each row is an example and each column represents an attribute. From the data the first row of the Excel sheet is used for attribute names. For the Read Excel Example Process, an Excel file is needed first. The data set was copied on sheet 1 of the Excel file thus the sheet number parameter is given value 1. It is important to provide the correct location of the file in the Excel file parameter. The first cell of the sheet is **A1** and last required cell is **J2370**, thus the imported cell range parameter is provided value **'A1:J2370'**. As the first row of the sheet contains names of attributes, the first row as names

	APPLICATION N	DATE FILED RE	URGENCY	APPLICANT	RESPONDENT	NATURE OF APPLICATION	STATUS	JUDGE (S)
1	APPLICATION N	DATE FILED RE	URGENCY	APPLICANT	RESPONDENT	NATURE OF APPLICATION	STATUS	JUDGE (S)
2								
3				**2005**				
4	NAI 331	9. Dec. 2005		CALTEX (KENYA) LIMITED	JOHANNES OKELLO OMBOTO	EXTENSION OF TIME	ALLOWED COSTS IN THE APPEAL ON 30/3/2006 AT KISUMU	OMOLO, JA (IN CHAMBERS)
5	NAI 336 (UR 205)	16. Dec. 2005		BROKE BOND KENYA LTD	JAMES NYABIOSI	STAY OF EXECUTION	DISMISSED WITH COSTS ON 31/3/2006 AT KISUMU	BOSIRE, OKUBASU, WAKI, JJA.
6	NAI 339 (UR 208)	22. Dec. 2005		BROKE BOND KENYA LTD	JAMES NYABIOSI	FILE AN APPEAL	DISMISSED ON 30/3/2006 AT KISUMU	OMOLO, JA (IN CHAMBERS)
7	NAI 349	20. Jan. 2006		SEVENTH DAY ADVENTIST CHURCH & OTHERS	MASUGA CONSTRUCTION C	FILE NOTICE & RECORD OF APPEAL	ALLOWED ON 19/6/2006 AT KISUMU	WAKI, JA (IN CHAMBERS)
8	NAI 350	20. Jan. 2006		AKAMBA PUBLIC ROAD SERVICES LTD	GAUDENCIA ATIENO AKIRA	DISPENSE WITH SERVICE	ALLOWED ON 27/3/2006 AT KISUMU	OKUBASU, JA (IN CHAMBERS)
9	NAI 351	20. Jan. 2006		ESTHER MWEMBU NYOIKE	SAMWEL NYOIKE NDRIAH	SUBSTITUTION	STRUCK OUT WITH NO ORDER AS TO COSTS ON 17/5/2006 AT NYERI	GITHINJI, JA (IN CHAMBERS)
10	NAI 352	25. Jan. 2006		GERPHAS ALPHONSE ODHIAMBO	FELIX ANDEGO	EXTENSION OF TIME	DISMISSED WITH COSTS ON 19/6/2006 AT KISUMU	WAKI, JA (IN CHAMBERS)
11	NAI 353 (UR 214)	25. Jan. 2006		KENYA TEA DEVELOPMENT AGENCY	BENSON ONDIMU MOSESE T	EXTENSION OF TIME	ALLOWED ON 30/3/2006 AT KISUMU	OMOLO, JA (IN CHAMBERS)
12								
13				**2006**				
14	NAI 10 (UR 5)	13. Jan. 2006		RICHARD A. ONDITI	KENYA COMMERCIAL FINAN	STAY OF EXECUTION & EXTENSION OF TIME	ALLOWED ON 31/3/2006 AT KISUMU	OMOLO, TUNOI, OKUBASU, JJA.
15	NAI 13 (UR 6)	20. Jan. 2006		KENYA TEA DEVELOPMENT AGENCY LTD	KENNETH ONDIEK OBARE T	EXTENSION OF TIME	DISMISSED WITH COSTS ON 14/6/2006 AT KISUMU	BOSIRE, JA (IN CHAMBERS)
16	NAI 33	8. Mar. 2006		THE BOARD OF MANAGEMENT OF KENYA MED	DR. MAURICE ONYIMBI	STRIKE OUT CIVIL APPEAL NO.34 OF 2005	ALLOWED WITH COSTS ON 19/6/2008 AT KISUMU	OMOLO, GITHINJI, AGANYANYA, JJA
17	NAI 48 (UR 29)	21. Feb. 2006		THE KENYA REVENUE AUTHORITY	HON. DAVID MWIRARIA, MD	STAY OF ANY PROCEEDINGS	WITHDRAWN ON 24/11/2006 AT KISUMU	OMOLO, BOSIRE, ONYANGO OTIENO
18	NAI 65 (UR 34)	8. Mar. 2006		DAMJI PRAGJI MANDAVIA	SARA LEE HOUSEHOLD AND	STAY OF EXECUTION	WITHDRAWN ON 29/3/2006 AT KISUMU	OMOLO, TUNOI, BOSIRE, JJA.
19	NAI 75 (UR 41)	17. Mar. 2006		KERESCENT MASINDE	CONCEPT NAMASEMBA & A	EXTENSION OF TIME	DISMISSED WITH COSTS ON 1/12/2006 AT NAIROBI	OMOLO, GITHINJI, WAKI, JJA.
20	NAI 81 (UR 45)	27. Mar. 2006		MT. ELGON HARDWARE LTD & 400 OTHERS	KENYA REVENUE AUTHORIT	STRIKE OUT APPEAL	ALLOWED ON 21/6/2006 AT KISUMU	BOSIRE, OKUBASU, WAKI, JJA.
21	NAI 101 (UR 62)	11. Apr. 2006		SUSAN DUDU & ANOTHER	HOUSING FINANCE COMPAN	INJUNCTION	WITHDRAWN ON 15/6/2006 AT KISUMU	TUNOI, BOSIRE, OKUBASU, JJA.
22	NAI 111 (UR 67)	24. Apr. 2006		LIUCAS MWAURA NDUATI	JOHN HENRY OKEYO	EXTENSION OF TIME	ALLOWED WITH COSTS TO THE RESPONDENT ON 1/12/2006 AT KISUM	ONYANGO OTIENO, JA. (IN CHAMBE
23	NAI 117	3. May. 2006		SAVINGS & LOAN KENYA LIMITED	MAYFAIR HOLDINGS LIMITE	EXTENSION OF TIME	ALLOWED COSTS IN THE APPEAL ON 19/6/2006 AT KISUMU	WAKI, JA (IN CHAMBERS)
24	NAI 118	3. May. 2006		MUHURONI SUGAR COMPANY LTD	KASAMANI & CO. ADVOCAT	EXTENSION OF TIME	WITHDRAWN ON 20/6/2006 AT KISUMU	BOSIRE, JA (IN CHAMBERS)
25	NAI 119	3. May. 2006		ERIC OKEYO	BEDROCK HOLDINGS LTD	EXTENSION OF TIME	DISMISSED WITH COSTS ON 30/6/2006 AT NAIROBI	DEVERELL, JA (IN CHAMBERS)
26	NAI 131	11. May. 2006		GEOFFREY MATOKE	KWAMBOKA MASASE & OTI	STRIKE OUT RECORD OF APPEAL	ALLOWED ON 1/12/2006 AT KISUMU	OMOLO, WAKI, ONYANGO OTIENO,
27	NAI 137	15. May. 2006		WILLIAM BARASA OBUTITI	MUMIAS SUGAR COMPANY	LEAVE	ALLOWED COSTS IN THE APPEAL ON 12/6/2006 AT KISUMU	GITHINJI, JA (IN CHAMBERS)
28	NAI 146	29. May. 2006		RICHARD AKWESEA ONDITI	KENYA COMMERCIAL BANK	EXTENSION OF TIME	ALLOWED COSTS IN THE APPEAL ON 1/12/2006 AT KISUMU	TUNOI, GITHINJI, ONYANGO OTIENO
29	NAI 147	29. May. 2006		JOHANNES OKELLO OMBOTO	CHEVRON (K) LTD & ANOTH	STRIKE OUT RECORD OF APPEAL	ALLOWED WITH COSTS ON 29/11/2006 AT KISUMU	OMOLO, GITHINJI, WAKI, JJA.
30	NAI 148	29. May. 2006		KHAN MOHAMMED SALIM T/A KHAN & ASSOC	NATIONAL INDUSTRIAL CRE	STRIKE OUT RECORD OF APPEAL	ALLOWED ON 21/3/2007 AT KISUMU	OMOLO, BOSIRE, DEVERELL, JJA.
31	NAI 161	16. Jun. 2006		BENSON ONDIMU MASESE T/A B.O. MASESE &	KENYA TEA DEVELOPMENT	STRIKE OUT CIVIL APPEAL NO.16 OF 2006	DISMISSED ON 20/6/2007 AT KISUMU	BOSIRE, WAKI, ONYANGO OTIENO,
32	NAI 162	16. Jun. 2006		KENNETH ONDIEK OBAE T/A K.O. OBAE & CO.	KENYA TEA DEVELOPMENT	STRIKE APPEAL	DISMISSED ON 20/6/2007 AT KISUMU	BOSIRE, WAKI, ONYANGO OTIENO,
	NAI 171 (UR 100)	23. Jun. 2006		DOO INVESTMENTS AND OFFICE MANAGEMEN	HOUSING FINANCE COMPAN	INJUNCTION	DISMISSED WITH COSTS ON 14/7/2006 AT NAIROBI	TUNOI, BOSIRE, OKUBASU, JJA.

94							
95		**2008**					
96	NAI 18 (UR 120)	25. Feb. 2008		PAUL OJIJO OMANGA	JAPHETH ANGILA	STAY OF RULING & ORDER	WITHDRAWN WITH COSTS ON 26/6/2008 AT KISUMU	OMOLO, BOSIRE, OKUBASU, JJA.
97	NAI 35	13. Mar. 2008	URGENT	JAPHETH ANGILA	PAUL OJIGO OMANGA	STRIKE OUT APPEAL	ALLOWED COSTS IN THE APPEAL ON 27/6/2008 AT KISUMU	OMOLO, OKUBASU, ALUOCH, JJA.
98	NAI 57	8. Apr. 2008		SAMSON GETANGITA NKWEGE & 2 OTHERS	ANISON NYAHIRI MUHINDI	EXTENSION OF TIME	ALLOWED COSTS IN THE APPEAL ON 20/6/2008 AT KISUMU	GITHINJI, JA (IN CHAMBERS)
99	NAI 61	14. Apr. 2008		PAUL OJIGO OMANGA	JAPHETH ANGILA	LEAVE	WITHDRAWN ON 27/6/2008 AT KISUMU	OMOLO, OKUBASU, ALUOCH, JJA.
100	NAI 73	25. Apr. 2008		HEZEKIAH MICHOKI	ELIZABETH ONYANCHA OME	LEAVE	WITHDRAWN ON 24/6/2008 AT KISUMU	OMOLO, JA (IN CHAMBERS)
101	NAI 74 (UR 40)	29. Apr. 2008	URGENT	THE HON JOEL OMAGWA ONYANCHA	SIMON NYAUNDI OGARI & A	STAY OF PROCEEDINGS	WITHDRAWN WITH COSTS ON 20/5/2008 AT NAIROBI	WAKI, AGANYANYA, ALUOCH, JJA.
102	NAI 75 (UR 41)	29. Apr. 2008	NOT URGEN	ABDI NURU OMAR & ANOTHER	MOHAMED ADEN ABDI	EXTENSION OF TIME	ALLOWED COSTS IN THE APPEAL ON 26/6/2008 AT KISUMU	OKUBASU, JA (IN CHAMBERS)
103	NAI 78 (UR 43)	2. May. 2008	URGENT	THE HON JOEL OMAGWA ONYANCHA	SIMON NYAUNDI OGARI & A	EXTENSION OF TIME	WITHDRAWN WITH COSTS ON 9/7/2008 AT NAIROBI	OMOLO, JA (IN CHAMBERS)
104	NAI 83	6. May. 2008		HOSKIN INDIMULA OLAKA	JOHN KAYELI OLAKA	AN ORDER		
105	NAI 98 (UR 58)	19. May. 2008		THE HON JOEL OMAGWA ONYANCHA	SIMON NYAUNDI OGARI & O	EXTENSION OF TIME	WITHDRAWN WITH COSTS ON 15/10/2008 AT NAIROBI	TUNOI, JA (IN CHAMBERS)
106	NAI 117 (UR 72)	5. Jun. 2008	URGENT	DELPHIS BANK LIMITED	CHANNAN SINGH CHATTHE	EXTENSION OF TIME	ALLOWED COSTS IN THE APPEAL ON 11/7/2008 AT KISUMU	BOSIRE, JA (IN CHAMBERS)
107	NAI 118 (UR 73)	5. Jun. 2008	URGENT	DELPHIS BANK LIMITED	CHANNAN SINGH CHATTHE	EXTENSION OF TIME	STRUCK OUT WITH NO ORDER AS TO COSTS ON 27/6/2008 AT KISUMU	BOSIRE, JA (IN CHAMBERS)
108	NAI 119 (UR 74)	5. Jun. 2008	URGENT	DELPHIS BANK LIMITED	CHANNAN SINGH CHATTHE	EXTENSION OF TIME	ALLOWED COSTS IN THE APPEAL ON 11/7/2008 AT KISUMU	BOSIRE, JA (IN CHAMBERS)
109	NAI 160 (UR 101)	4. Jul. 2008	NOT URGEN	DELPHIS BANK LIMITED	CHANNAN SINGH CHATTHE	EXTENSION OF TIME	WITHDRAWN WITH COSTS ON 25/3/2009 AT KISUMU	OMOLO, GITHINJI, AGANYANYA, JJA
110	NAI 166	9. Jul. 2008		SMALL ENTERPRISES FINANCE COMPANY LTD	THOMAS ADONGO T/A KISU	EXTENSION OF TIME	ALLOWED COSTS IN THE APPEAL ON 5/12/2008 AT KISUMU	OKUBASU, JA (IN CHAMBERS)
111	NAI 167	9. Jul. 2008		MOKUA OTWORI ALIAS RICHARD MEROKA MO	MOSOTA OTWORI	LEAVE	STRUCK OUT WITH COSTS ON 16/1/2009 AT KISUMU	OMOLO, JA (IN CHAMBERS)
112	NAI 191	27. Jul. 2008		NATIONAL BANK OF KENYA LTD	KATANGI DEVELOPERS LTD	EXTENSION OF TIME	DISMISSED WITH COSTS ON 17/7/2009 AT NAIROBI	TUNOI, JA (IN CHAMBERS)
113	NAI 202 (UR 129)	28. Jul. 2008	NOT URGEN	DAVID ONYANGO OBANGE	ROSE AKUKU AOKO & 3 OTH	EXTENSION OF TIME	ALLOWED COSTS IN THE APPEAL ON 27/3/2009 AT KISUMU	AGANYANYA, JA. (IN CHAMBERS)
114	NAI 203 (UR 130)	28. Jul. 2008	NOT URGEN	ABDI NURU OMAR & ANOTHER	MOHAMED ADEN ABDI	STAY OF EXECUTION	ALLOWED COSTS IN THE APPEAL ON 3/12/2008 AT KISUMU	OMOLO, BOSIRE, ALUOCH, JJA.
115	NAI 212	30. Jul. 2008		HEZEKIA MICHOKI	ELIZABETH ONYANCHA OME	LEAVE	DISMISSED WITH COSTS ON 5/12/2008 AT KISUMU	ONYANGO OTIENO, JA. (IN CHAMBE
116	NAI 214	30. Jul. 2008		FRANCIS ORIOSA ORANGO	JOSEPH MATO NGOKO	EXTENSION OF TIME	WITHDRAWN WITH COSTS ON 2/12/2008 AT KISUMU	ALUOCH, JA (IN CHAMBERS)
117	NAI 215	30. Jul. 2008		DR. MANASE ONYIMBI	THE BOARD MANAGEMENT	EXTENSION OF TIME	ALLOWED COSTS IN THE APPEAL ON 16/1/2009 AT KISUMU	ALUOCH, JA (IN CHAMBERS)
118	NAI 217 (UR 137)	1. Aug. 2008	NOT URGEN	NAGENDRA S. AXENA	MIWANI SUGAR MILLS LIMIT	STAY OF EXECUTION	DISMISSED WITH COSTS ON 16/1/2009 AT KISUMU	OMOLO, BOSIRE, ALUOCH, JJA.
119	NAI 240 (UR 153)	28. Aug. 2008	URGENT	DELPHIS BANK LIMITED	CHANNAN SINGH CHATTHE	EXTENSION OF TIME	ALLOWED COSTS IN THE APPEAL ON 17/10/2008 AT NAIROBI	TUNOI, JA (IN CHAMBERS)
120	NAI 241 (UR 154)	28. Aug. 2008	URGENT	DELPHIS BANK LIMITED	CANNON ASSURANCE KENY	STAY OF PROCEEDINGS	ALLOWED COSTS IN THE APPEAL ON 11/12/2008 AT NAIROBI	BOSIRE, GITHINJI, ALUOCH, JJA.
121	NAI 251 (UR 161)	8. Sep. 2008	NOT URGEN	WALTER ENOCK NYAMBATI	MOSE NYAMBEGA	STAY OF EXECUTION	WITHDRAWN ON 23/10/2008 AT NAIROBI	ALUOCH, JA (IN CHAMBERS)
122	NAI 291	3. Nov. 2008		KENNEDY BOSIRE, GICHANA	THE REGISTERED TRUSTEE	INJUNCTION	WITHDRAWN ON 25/3/2009 AT KISUMU	OMOLO, GITHINJI, AGANYANYA, JJA
123	NAI 299	13. Nov. 2008		OTINGA WARONGA OMUNGARA	JAMES OYIER OGOLA	LEAVE	WITHDRAWN WITH NO ORDER AS TO COSTS ON 24/3/2009 AT KISUMU	OMOLO, JA (IN CHAMBERS)
124	NAI 300	13. Nov. 2008		KENNEDY BOSIRE, GICHANA	THE REGISTERED TRUSTEE	INJUNCTION	DISMISSED WITH COSTS ON 8/5/2009 AT KISUMU	OMOLO, GITHINJI, AGANYANYA, JJA
125	NAI 301	18. Nov. 2008		KENYA AGRICULTURAL RESEARCH INSTITUTE	WILLIAM OKONGO & 2 OTHE	STRIKE OUT NOTICE OF APPEAL	DISMISSED WITH COSTS ON 8/5/2009 AT KISUMU	BOSIRE, ONYANGO OTIENO, AGANY
126	NAI 339 (UR 225)	30. Dec. 2008		BILLIAH KEMUNTO ONDABA	LYDIA JAHENDA AGOI ONDA	STAY OF PROCEEDINGS	ALLOWED COSTS IN THE APPEAL ON 13/2/2009 AT NAIROBI	GITHINJI, ONYANGO OTIENO, ALUO
127								

Figure 11:Non –standard Excel sheets

Read Excel

This operator was used to load data from a sample Microsoft Excel spreadsheets that was collected and entered by one of the court clerk in the civil registry. The spreadsheet above was used to test the effect of non-standard data. The data path is provided as *C:\Users\user\Desktop\COMBINED CIVIL APPLICATIONS REGISTER*. Process was created by copying the data set present in the civil registers from various regions and combining this data into a new Excel file which was named **COMBINED CIVIL REGISTER**. In order to load the data into the ETL tool using the process Read Excel, the following steps were followed up:

1. From the start menu open the Rapid miner software:
2. Create a new process from File->New.
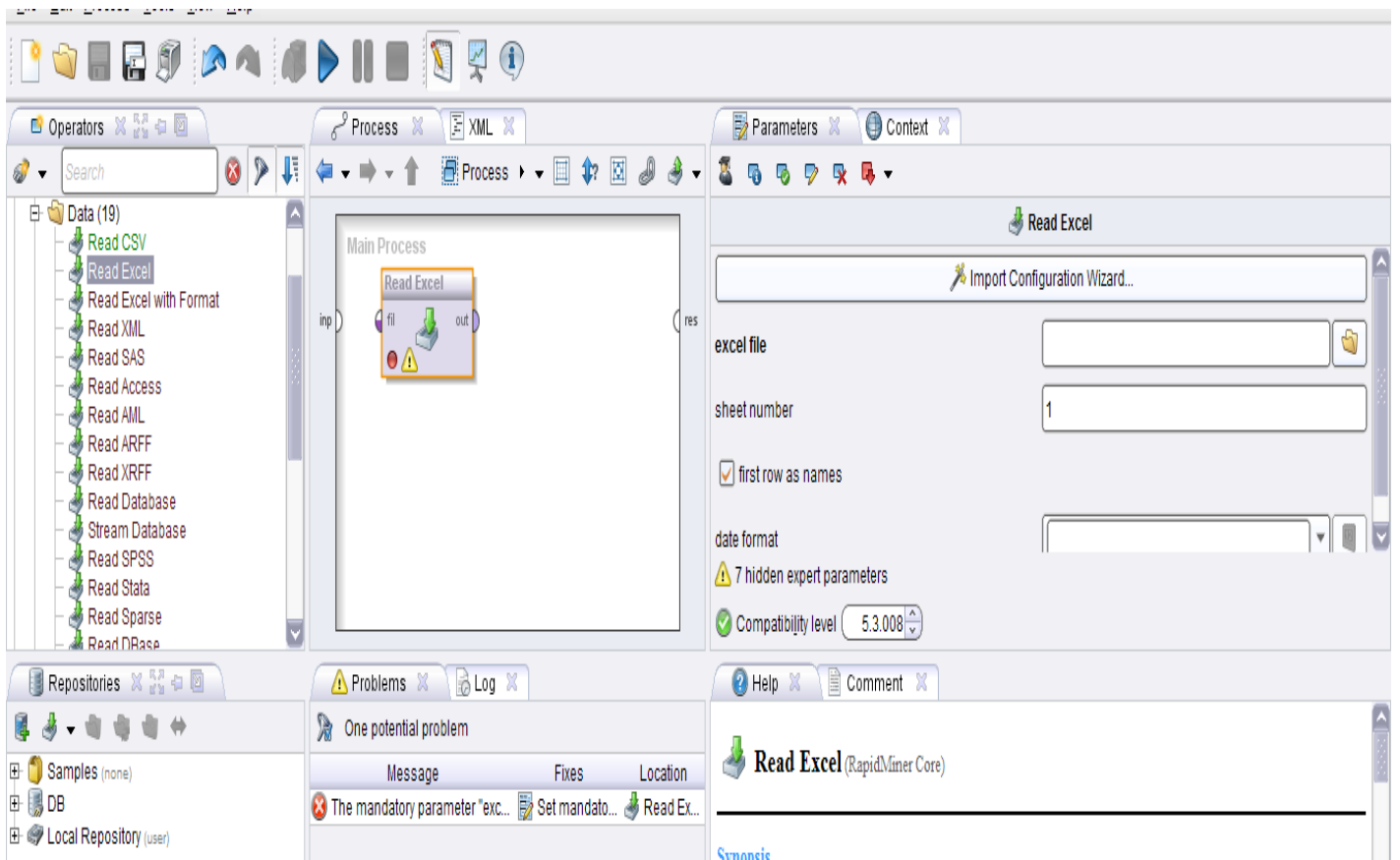3. Click Operators->Import->Data->Read Excel/Database

Figure 12:Read Excel

Impact of Extracting non-standardized Data

Upon running the process was run and the results generated were as follows:-

Figure 13:Non-standardized data extracted

4.9 Discussion of Results

From the above records it is evident that the data that was run had errors which have introduced question marks in the 'urgency' column. This is as a result of rows left blank without any data filled. To solve the above problem, the following recommendations are necessary:

1. Data should have been cleansed first which entails removing blank spaces, deleting repetition and typing errors.

2. ETL tool requires data to be merged into one excel sheet as it can only load one sheet at a time. It is also important to save the file as .CSV to make it acceptable to most ETL tools.

Working with Standardized Data

However, upon standardization of the data collection process, the ETL tool should be able to read from the standardized data collected through excel sheets or CSV files. The following example illustrates extracted data from the standardized template.

Figure 14:Standardized data extracted

Sample Transformation to the Extracted Data

Sorting

This uses an operator sort. It utilizes the civil repository data Kisumu as the input set in and sorts the data in ascending or descending order according to the specified single attribute 'date_of_plea'. This attribute is specified using the *attribute name* parameter. Sorting is done in increasing or decreasing direction depending on the setting of the *sorting direction* parameter. The diagram below is used to demonstrate this transformation.
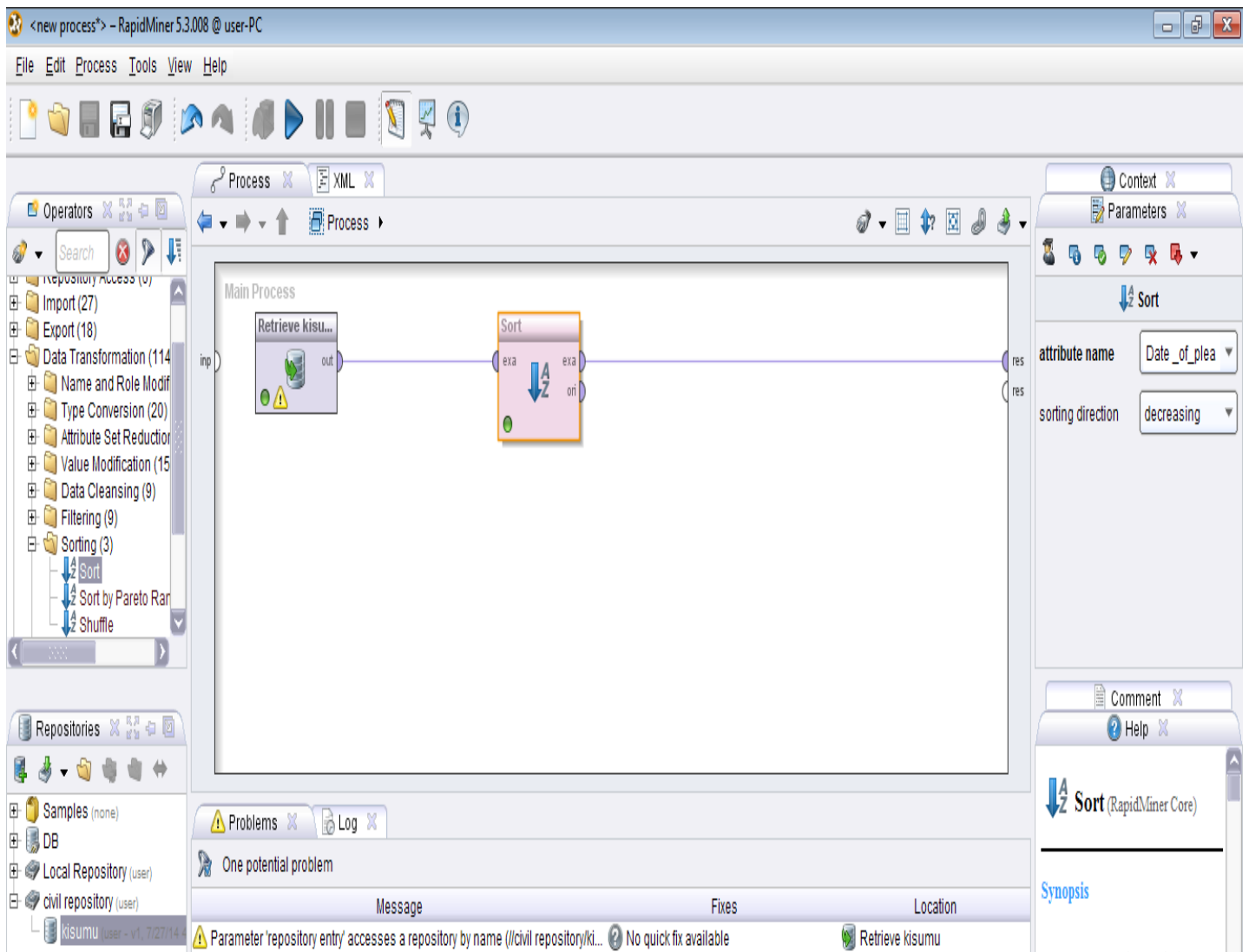
Figure 15: Sort operator

Output to a file or database

Operators such as Write CSV,Write database are used to export the sorted data to a database or a specified file format for future reference. An example is as shown below

Figure 16: write csv process

Figure 17:Output data in csv format

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.0   Achievements

This project has managed to expose the data sources and operational systems in judicial environment and display the kind of data available .It has also demonstrated using case returns data and flat files data how the ETL process can be used to migrate data from sources to a data warehouse. Within this study is a proposed conceptual model for the judicial ETL process with its associated architectures. At the appendices section there is an attached guideline for creating a data warehouse for the Kenya Judiciary.

5.1   Recommendations

In order to prepare the Kenyan Judiciary for data warehousing and data mining, the following recommendations are proposed.

Setting up of an Intranet

An Intranet will enhance sharing of data between the Performance Department and the originating registries and avoid delays in sending data and reduce reliance on the email server.

Backup Storage and Centralized Data Repository

There is need to procure an all-inclusive database management system that comes with Business Intelligence tools.

Standardization of Data Collection Procedures

To eliminate the challenges associated with data collection it is important to adopt a standard way of collecting data across the registries. This should be such that only relevant information is collected. Where possible there should be a template that provides users with instructions on how to enter data and the specific details to include.

User Training

A comprehensive training should be undertaken for the court users and data entry clerks who are charged with the responsibility of recording the data. The court register should also be modified such that the information recorded thereof is reflected in the standardized template. Expertise should be sought while acquiring a template so that the fields included can be used to capture case details as much as possible across all registries. This template should also have drop down lists where the data entry clerk can choose from a pre-defined list instead of typing data.

Building a Data Warehouse

Data mining solutions can be adopted for use by the Performance Management Department as a useful tool to help analyze overall performance of staff in the judiciary. To further improve this process it would be recommendable to adopt some of the guidelines outlined in Appendix section.

Use of Automation in Populating the Warehouse

It is important to automate the process of loading data into a database in order to save on time usage. There are many tools available in the market that helps in automation of data loading processes. Any tool applicable if acquired would go a long way in simplifying this process.

5.2   Conclusion

This project has shown that ETL process can be adopted to solve the data integration process for the Kenya judiciary. Data from various operational systems has been extracted using a sample ETL tool and modifications applied therefore the resultant data is without errors and is in a suitable format to be populated into a data warehouse. Thereafter the Kenya Judiciary can adopt data mining solutions .Further to these attached at the appendices section are more guidelines that are vital to the preparation of data warehousing and data mining with regard to data integration.

REFERENCES

1. Sweety Patel, Piyush Patel, Saumil Patel.(2012). "Overview of ETL process with its important".Department of Computer Science, Fairleigh Dickinson University, NJ- 07666, USA,Department of Computer Science, Rajasthan Technical University, India.

2. Sweety Patel, Piyush Patel, Saumil Patel / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 2,Mar-Apr 2012, pp.906-908 906 | P a g e.

3. Shaker H. Ali El-Sappagh, Abdeltawab M. Ahmed Hendawi,AliHamed El Bastawissy(2011). "A proposed model for data warehouse ETL processes". Mathematics Department, College of Science, King Saud University, Saudi Arabia Information Systems Department, Faculty of Computers and Information, Cairo University, Cairo, Egypt Journal of King Saud University –Computer and Information Sciences.Http://www.ksu.edu.sa,Http:// www.sciencedirect.com

4. Nils Schmidt(nilsschmidt@gmx.de),Mario Rosa(marioarosa1@gmail.com), Rick Garcia(rjgarcia5@utpa.edu), Efrain Molina(emoli81@hotmail.com), Ricardo Reyna(rreyna11@utpa.edu), John Gonzalez,(jegonzalez11@gmail.com),(2011). "ETL Tool Evaluation – A Criteria Framework".,University of Texas-Pan American,1201 W. University Drive, Edinburg, TX (956) 381-UTPA.

5. Alkis Simitsis and PanosVassiliadis.( ) "A Methodology for the Conceptual Modeling of ETL Processes". National Technical University of Athens, Dept. of Electrical and Computer Eng.Computer Science Division, Iroon Polytechniou 9, 157 73, Athens, Greeceasimi@dbnet.ece.ntua.grUniversity of Ioannina,Dept. of Computer Science, 45110, Ioannina, Greecepvassil@cs.uoi.gr.

6. LUKASZ A. KURGAN and PETR MUSILEK .(2006). " A survey of Knowledge D iscovery and Data Mining process models". The Knowledge Engineering Review, 21, pp 124 doi:10.1017/S0269888906000737.

7. Owen Kufandirimbwa and Cry Kuranga.(2012)."Towards Judicial Data Mining: Arguing for Adoption in the Judicial System", Published in the Online Journal of Physical and Environmental Science Research, ISSN 2315-5027; Volume 1, Issue 2, pp. 15-21.

8. Rüdiger Wirth Daimler &Jochen Hipp.(2000).(CRISP-DM: Towards a Standard Process Model for Data.

9. Sotarat Thammaboosadee and UdomSilparcha.(2009) "A GUI Prototype for the Framework of Criminal Judicial Reasoning System", Journal of International Commercial Law and Technology ,Vol. 4, Issue 3.

10. Ranjit Singh and Dr. Kawaljeet Singh(2010) "A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing ",IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 2, May 2010

11. Jau-Huei Lin, M.D. and Peter J. Haug(2006) " Data Preparation Framework for Preprocessing Clinical Data in Data Mining", PMCID: PMC1839316, M.D.(AMIA AnnuSymp Proc. 2006; 2006: 489–493.).

12. Website:www.pic.dhe.ibm.com/infocenter/data tools.datamining.doc

13. Website:http://www.**etltool**.com/**etl-tools**-comparison/

14. http://www.ipcdesigns.com/etl_metadata/ETL_Architecture_Guide.html

15. Judicial Transformation Framework manual

16. www.technet.microsoft .com

17. Thammaboosadee, S., Silparcha, U., (2008). A framework for criminal judicial reasoning system using data mining techniques, Proceedings of the International Conference on Digital Ecosystems and Technologies, DEST 2008. 2nd IEEE, pp 518 – 523

18. Ni, Z.W., Yang, S.L., Yang, Y. Li, F., (2004). Case-based reasoning framework based on data mining technique, Proceedings of International Conference Machine Learning and Cybernetics, Vol 4, pp 2511 – 2514

19. Verma, L., Srinivasan, S., Sapra, V. (2014). Integration of rule based and case based reasoning system to support decision making, Proceedings of International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), pp 106 - 108

20. Hussain, K.Z., Durairaj, M., Farzana, G.R.J. (2012). Criminal behavior analysis by using data mining techniques, Proceedings of International Conference on Advances in Engineering, Science and Management (ICAESM), pp 656 - 658

21. Xie, R., (2008). The Application of Data Mining in Judicial Judgment, Proceedings of the 4th International Conference on on Wireless Communications, Networking and Mobile Computing, WiCOM '08, pp 1 – 4

22. Costake, N., (2001). E-governance and the judicial system - A point of view, Proceedings of the 12th International Workshop on Database and Expert Systems Applications, pp 423 – 427

23. Annapoorani, V., Vijaya, A., (2013). A prevailing judicial package for clustering and sorting information extraction, International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), pp 241 - 244

24. Scott, M.S., Thinyane, M., (2013). The e-Judiciary system: Obliteration of the digital divide through ICT4D in traditional justice systems, International Conference on Adaptive Science and Technology (ICAST), pp 1 - 8

25. Ticom, A.A.M, de Souza, B., de Lima, L.P., (2007). Text Mining and Expert Systems Applied in Labor Laws, Seventh International Conference on Intelligent Systems Design and Applications, ISDA 2007, pp 788 - 792

26. Hamin, Z., Othman, M.B., Mohamad, A.M., (2012). ICT adoption by the Malaysian high courts: Exploring the security risks involved, International Conference on Innovation Management and Technology Research (ICIMTR), pp 285 – 289

27. Zernik, J., (2010). Data Mining of Online Judicial Records of the Networked US Federal Courts, International Journal on Social Media, MMM: Monitoring, Measurement, and Mining, pp 69 – 83

28. Winn, P.A., (2009). Judicial Information Management in an Electronic Age: Old Standards, New Challenges, The Federal Courts Law Review, Vol 3, Issue 2, pp 135 - 176

APPENDICES

Guide to Design of Data Warehouse for the Kenya Judiciary

<u>Introduction</u>

A data warehouse is a database designed to enable business intelligence activities: It exists to help users understand and enhance their organization's performance. It is designed for query and analysis rather than for transaction processing, and usually contains historical data derived from transaction data, but can include data from other sources. Data warehouses separate analysis workload from transaction workload and enable an organization to consolidate data from several sources. This helps in:

- Maintaining historical records
- Analyzing the data to gain a better understanding of the business and to improve the business

The data warehouse works with data collected from multiple sources .These source data may come from internally developed systems, purchased applications, third-party data syndicators and other sources. For the Kenya judiciary such data sources would include Case management system (CMS), judicial help desk system (JHDS), Integrated Financial Management Information system (IFMIS), Flat files.

Importance of data warehouse to Kenya Judiciary

i. To centralize data. This means that the judiciary will have a central depository of data gathered from various sources.

ii. Promote consistency of data thus create an atmosphere of fairness in judgments and efficiency in court decision making process.

iii. Provides an efficient way of storing information for long periods of time.

The data in a data warehouse is typically loaded through an extraction, transformation, and loading (ETL) process from multiple data sources. Modern data warehouses are moving toward an extract, load, and transformation (ELT) architecture in which all or most data transformation is performed on the database that hosts the data warehouse. It is important to note that defining the ETL process is a very large part of the design effort of a data warehouse. Similarly, the speed and reliability of ETL operations are the foundation of the data warehouse once it is up and running.

Data Warehouse Tasks

These are the day to day activities executed in a typical data warehouse

  i. Configuring a database for use as a data warehouse

  ii. Designing the data warehouse

  iii. Performing upgrades of the database and data warehousing software to new releases

  iv. Managing schema objects, such as tables, indexes, and materialized views

  v. Managing users and security

  vi. Developing routines used for the extraction, transformation, and loading (ETL) processes

  vii. Creating reports based on the data in the data warehouse

  viii. Backing up the data warehouse and performing recovery when necessary

  ix. Monitoring the data warehouse's performance and taking preventive or corrective action as required.

Hardware Requirements

The data warehouse and the stage area are servers which can be hosted in one platform with the following specifications:

Network

10/100/1000M Ethernet card with IPv4 or IPv6 protocol

10/100/1000M Ethernet card with IPv4 and IPv6 protocol

Network connection to connect to a remote report server or ETL machine

Platforms

Microsoft Windows 2008 Standard Edition, Enterprise Edition with Service Pack 1

Red Hat Enterprise Linux® 5 GA with Update 1

Database servers

Minimum Microsoft SQL Server 2005 with Service Pack 2

Oracle 10g release 2

Oracle 11g release 1

Processor:

Minimum Intel core i5

PSU: Two hot swappable for redundancy

Location of the Equipment

The server should be housed in the server room at Milimani Law courts and similar set up should exist at the other high courts across the country. Currently the judiciary possesses servers with the following specifications.

Available hardware

Currently the judiciary possesses IBM servers with the following specifications:

Processor: Intel Xeon CPU e31240@3.3(8 CPUs)

Memory:8 GB RAM

Storage: 2TB hard drive

Operating systems supported: Microsoft Windows Server 2012/Microsoft Windows Server 2008 R2/; Microsoft Windows Server 2008, Red Hat Linux, SUSE Linux, Vsphere 5.0

Power supply:(std/max):300 W(1/1) or redundant 460 W (1/2)

Network interface: Dual Gigabit Ethernet

Cost inter-operability and staff training considerations.

The cost of the servers keep changing so it should be determined by getting quotes from the vendors at the time of implementation. It's important to have in-house personnel with skills on database design and administration.

Data Sources, Preparation and Loading

The main sources of data in the Kenya judiciary are:

Court management system (CMS)

- Judicial help desk (JHD)
- Integrated financial information system (IFMIS)
- Flat files

To extract data from the above sources the following key steps should be followed:-

i. Collect acceptable data that is in a standardized format: This can be achieved by creating a template for the manual records (flat files) and standard reports from the CMS, JHD and IFMIS.

ii. Collected data should be channeled to a staging area (for cleansing and manipulation).

iii. Data should be loaded into the warehouse which is a data base created in SQL server or otherwise as preferred by the organization.

The Data Warehouse Staging Area

This is the area that will provide intermediate storage for data processing during the ETL process. Its usually a database that sits between the data sources and the datawarehouse.There function of the staging area is to boost in the efficiency of the ETL process through consolidation of data from different sources as it is temporarily stored as it awaits further processing. This area also helps in taking control of concurrency issues during processing which is a result of connection problems in the cause of retrieving data from multiple sources. The staging areas also support hosting of data to be processed on independent schedules or data that is targeted to multiple targets.

Implementation

This is a can be implemented in form of tables which is created using either SQL Server database, oracle or postgress.Also,the data can be stored in XML format which is text based. It should have the capability to split data, manipulate, cleansing data through rules that enforce the integrity of data. Cleansing entails removing unwanted fields, merging fields, inserting new fields with an aim of achieving standardization. The following diagram demonstrates how a staging area can be implemented to accommodate the various departments' data within the Kenya judiciary.
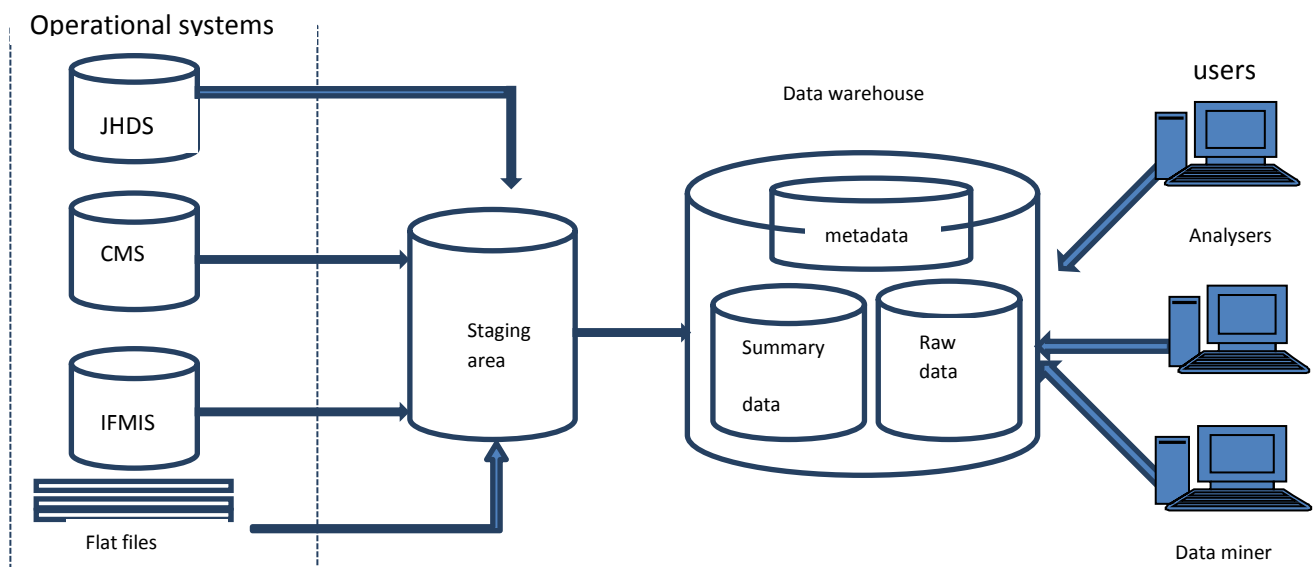

Figure 17:Archutecture of Kenya judiciary data warehouse architecture with a staging area

Build the Judicial data warehouse

When designing a data warehouse there are many considerations one should put in mind which include:

i.   Purpose: The major aim of building the warehouse is to provide a platform for rapid access to information for the purposes of analysis and reporting.

ii.   Dimensional modeling: This produces dimension tables in which each table contains fact attributes that are independent of those in other dimensions. This implies organizing information into structures that enhance the efficiency of queries such that the analysts are able to query data and get the exact information in a more summarized manner.

iii. Dimension tables: This entails organizing information in its own dimensional table e.ga dimensional model of case details might include a fact table named cases_details that consists of fields or records such as case type, division, and judicial officer. Each of this information is organized into its own dimension table. Case type is placed in a case dimension table, judicial officer in the judicial officer dimension table and division in its own division dimension table.

iv.   Fact table: A data warehouse should include at least one fact table that captures data that measures the organizations operations. They consist of large number of rows and millions of records depending with the organizations historical data. Contains numerical data i.e. facts that can be summarized to provide information about the history of the operation of the organization. Contains foreign key and primary keys and attributes of the fact records. e.g case details table that may comprise of case_id,case_type,date_registered,judicialofficer_id,division_id which captures the details of registered cases.

v.   Aggregation tables: They are used to help reduce the size of fact tables by aggregating data into summary records when the fact table is created. Contain summaries of fact table information.

vi.   Indexes: They help in improving the performance of the data warehouse. They help to identify levels in hierarchical structure and fact tables indexing.

Building a data warehouse is a significant project and may involve a number of steps. This are illustrated in the table below

Rollout Map for creating a Data warehouse

Step description	Activities involved	Objectives to be achieved	Implementation details
Design data warehouse	Designing a data warehouse project plan	Familiarization with the project task and schedule	Develop the work plan and time lines
Define technical requirements and specifications for the warehouse	Identify and Document aspects of the project scope, capacity of data and functionality.	Setting up specific milestones and associated deliverables (data, functionality, structures)	Setting agreements with stakeholders, department involved and other interested parties to the task
Data Acquisition	Identify, evaluate, and design elements pertinent to data acquisition. This include visiting registries and talking to staff who are working on the data source systems	Document and agree on data warehouse design components (design standards, data source mapping, ETL process, load, refresh and purge modules, as well as data mart design)Identify the available source system data in the judiciary that will be used to populate the warehouse	Develop the Data Acquisition Plan. This includes the standardized data collecting tool for all the departments. Each department uses a template so that there is uniformity in collecting data. Implementation Strategies
Quality Assurance	Assess Data Quality, Develop Design & Build Standards. This may entail performing tests such as unit testing by the developer that is the ETL logic, business rules and high risk areas. System testing where all jobs are executed and data processed to ensure no issues arise during job execution. Data validation is done to data in the warehouse using queries, Excel.	Define requirements for data (cleansing, error handling, audit & control) assess quality of source data, identify data management procedures, and finalize tool selection. Generate report validation tools for accuracy usually done by the data warehouse team.	Develop Data Quality Components Design and Test Plans
Data Warehouse Technical Architecture	Identify, plan, and design capacity, hardware requirements, software components and test criteria	Ensure adequate capacity, identify process flow of data, plan integration of modules, data, access, and meta data	Plan Implementation and Test Environments for Warehouse Architecture Components
Warehouse Administration	Develop, Design and Build Standards for Warehouse Administration.This include appointing an administrator who is an ICT expert to	Design and plan version control, data archiving, scheduling, usage, data governing, backup, restoration, query profiles, and security	Plan for Development and Test Environments for Warehouse Administration Components
Meta data Management	Identify and document Meta data Requirements, as well as design and build standards	Collect meta data needed by technical and user communities, identify required meta data tools, and develop standards to whichmeta data should adhere	Develop a Meta data Management Plan to manage and monitor meta data, Develop Design and Build Standards (or identify appropriate tool)
Data Access	Determine Access Requirements and	Collect specific access requirements to support	Develop Access Plan and Determine access tools.

	develop design and build standards	analysis capabilities, data manipulation functionality, and user interface. Develop, design and build standards to support query and reporting and query criteria, user security, and confirmation of data availability	Develop Standards and Confirm Data Availability
Design and Build	Design and Build Development Standards and Modules for all Components of Data Warehouse Solution	Design, build, load and test Data Acquisition Modules (ETT, load, refresh, purge, data mart) Design, develop, and generate or implement Data Quality Modules (cleansing, error handling, audit & control) Design, build, load and test Architecture Components (Multidimensional, Test and production Databases) Design, build, load and test Warehouse Administration Modules (versioning, scheduling, backup, restoration, performance, security) Build or Implement Meta data Modules Design, build, load and test Data Access Components (reports and query criteria, user security, user access and style specifications)	Design, Build, Populate and Test: Data Acquisition Modules Data Quality Modules Databases and other Schema's Warehouse Administration Modules Meta data Modules Data Access Modules
Documentation	Define and Produce Documentation Standards, Procedures and Environment	Specify and develop documentation deliverables (User, Technical, Operational, Reference)	Develop Documentation Requirements, Standards, and Delivery Strategy and Produce Final Documentation
Testing	Define testing strategies, develop test procedures and perform testing specific to the scope of the Data Warehouse solution	Design, develop and implement testing plans and strategies for ETT, Performance, Interfaces, Integration, Volume, Query Profiles and other Components	Develop and Implement Testing Strategy, Plan, Models and Integration requirements
Training	Define and develop increment training requirements and plans	Identify and document training requirements for technical and end user staff, Identify specific roles who should receive training, and create training databases	Develop Training Strategy, Requirements, and Class Material. Create Training Databases
Installation	Define and develop installation plan	Develop an installation plan to support the	Develop Sequential or repeatable (Step by Step)

		production, test and other maintenance environments for the data warehouse solution	Installation Plan
Transition	Define transition strategy to production environment	Identify how the transition to production will occur including planning for data acquisition, preparation of production database(s), developer preparations and other cut-over issues	Develop Cut-Over Plan, Implement Maintenance, Production and Regression Environments
Production Support	Measure and support the production systems	Evaluate and audit the system for performance, faults, use, growth, recovery and tuning issues	Develop Library for production support metrics, corrections, enhancements, results
End Phase	Prepare for and complete phase end activities	Secure acceptance of phase end deliverables, release resources, assess and audit deliverables	Prepare Phase End Report and perform Quality Assessment activities
Post Implementation Support	Evaluate implemented increment, and non-implemented requirements, data warehouse architecture and plans	Assess responsiveness of solution to stated need, assess performance of data warehouse, identify next increment	Document Evaluation of Data Warehouse Solution and Architecture, Identify Next Increment Opportunities, Assess Performance of Project to Plan
End DW Increment	Prepare for and complete final project activities	Secure final acceptance of project deliverables	Conduct Discovery Meetings with Business Community

I/O requirements

Middleware connectivity

For successful implementation of data warehouse in judiciary there should be a reliable connection between the remote sites and the main server and the users to the servers. This can be done by having a wide area network (WAN) between Milimani law courts and the other high court stations. There should also be connectivity to the magistrate courts over the WAN.in all the courts at all levels there should a standard local area network (LAN) to enable the users to collect and post data to the servers.

Tool Required

The best tools for building a data warehouse would be the ones with known stability and ease of use coupled with reliable support and resilient security. Some of the best tools include:

  i.   Oracle
 ii.   Microsoft SQL server
iii.   Postgres

Users

The users of the data warehouse include:-

Top Management: This include departmental heads, registrars, judges and senior and administrative staff who may be interested in acquiring critical information for decision support system

End users: This is mostly staff who may be interested in data analysis. Such include the staff in the PMD department.

Data entry clerks: Point out who are the users and maybe categorize them (pmd,data entry etc.)Describe how the data users need to prepare to cope with the implementation…e.g. training needed in capturing new data to match new DW requirements.

Applications (DM)

Rapid miner SSIS etc.

Management of the DW

This is mostly done by the Database administrators who should have thorough knowledge of data bases.

Deployment & system management

Security Issues

Security of Data should be highly prioritized. This can be done by implementing standard security policy across the 147 court stations. Some of the items of the policy include:

- Having a centralized intrusion detection, malware and antivirus software.
- centralized internet access
- high standard firewall
- Security awareness campaigns for users
- security and network security

Backup and recovery

This is the task of the system administrator. However, during system design its necessary toplan a backup strategy and consider what is to be backed up and how frequently this should be done. The backup can be done locally, remotely, in tapes or otherwise as the organization may require.

The following activities should be considered:

i. Build and integrate the backup recovery plan. NOLOGGING(direct path loads and inserts, index and table creation) operations should be taken into account when designing a backup and recovery plan.

ii. Plan for the strategy to adopt or combination of strategies to adopt.e.g. The ETL strategy which recovers a backup that does not contain non-recoverable transactions and replay the ETL that has taken place between the backup and the failure. The incremental backup strategy: performs a backup immediately after an otherwise non-recoverable transaction has taken place.

iii. Consider the minimal resources required to support the backup scenario.

iv. Take a weekly backup of the data warehouse

v. Every night store or backup a copy of the raw data files which are loaded into the data warehouse and then re-run the nightly ETL processes to roll the data warehouse process.