



THE UNIVERSITY OF NAIROBI
SCHOOL OF COMPUTING AND INFORMATICS

A METHODOLOGY FOR THE IMPLEMENTATION OF A DATA WAREHOUSE USING
AN ETL PROCESS MODEL FOR IMPROVED DECISION SUPPORT

ANNE WANJIKU KIBUGU
REGISTRATION NUMBER: P53/73060/2014

SUPERVISOR
PROF. WILLIAM OKELO ODONGO

The research project submitted in partial fulfilment of the requirement for the award of the
Degree of Master of Science in Distributed Computing Technology of the University of Nairobi.

NOVEMBER 2016

DECLARATION

This research project is my original work and has not been presented or is due for presentation for any award at any learning institution.

SIGNATURE: _____ **DATE:** _____

ANNE WANJIKU KIBUGU

P53/73060/2014

The research project has been submitted for examination with my approval as the University's Supervisor.

SIGNATURE: _____ **DATE:** _____

PROF. WILLIAM OKELO ODONGO

SCHOOL OF COMPUTING AND INFORMATICS

THE UNIVERSITY OF NAIROBI

DEDICATION

I dedicate this research project to my beloved mother, who taught me the values of discipline, integrity and hard work. Although I cannot share with her the fulfilment of this project's success, I greatly appreciate the fact that she instilled the importance of education and closely monitored the progress of this research. The project is also dedicated to my daughter Tamia Wambui, who gives me reasons to forge ahead in life and is a constant reminder that even the largest and hardest tasks can be accomplished, one step at a time. All glory and honor to the Almighty God, for His sufficient provision and for placing the right people, at the right time throughout the project.

ACKNOWLEDGEMENT

I would like to acknowledge the invaluable and consistent instructions and guidance of my supervisor Prof. William Okelo Odongo, the lecturers' insightful criticisms and patient encouragement, and the entire teaching and non-teaching staff of The University of Nairobi, School of Computing and Informatics, to include among others; Dr. Evans A. Miriti, Dr. Christopher Chepken and Dr. Stephen Mburu. The knowledge and expertise acquired from this great team is exceptional and transformational to me and by extension, the Society.

I would also like to thank the entire staff of Higher Education Loans Board, especially the senior management and ICT staff, who found time out off their busy schedules to interact with me during data collection and granted access to data sources that enabled me come up with the research findings. I also thank those who participated during the system's performance evaluation, who willingly shared their observations and comments by filling out the questionnaires.

I would like to thank my loved ones, friends and colleagues at work and at the University, who have supported me throughout the entire process.

DEFINITION OF TERMS/ ABBREVIATION AND ACRONYMS

HELB – Higher Education Loans Board

ETL – Extract, Transform and Load

EMM - Entity-Mapping Methodology

UML – Unified Modelling Language

OLTP - OnLine Transaction Processing

OLAP - OnLine Analytical Processing

TVET – Technical and Vocational Education and Training

UG – Undergraduate

ABSTRACT

The operational processes and functionalities of a data warehouse constitute of a resource-intensive workflow, constituting an important part of the back-end of the System designs. To resolve the intense workflow and to manage the data warehouse operational processes, extraction-transformation-loading (ETL) processes are used.

A methodology used for the implementation of an organization's data warehouse was described in this project. The study's approach was informed by the need of a data warehouse as a decision support system for Higher Education Loans Board. The purpose of the project was to provide a solution for mining data from several sources; operational, historical and external databases to a central location for reporting and analyzing. The literature review investigated existing methodologies applied by other researchers. The evaluation of the studies and related surveys aimed at identifying the key elements and characteristics of an effective ETL process model towards the implementation of a data warehouse. A prototype system was developed based on an Entity Mapping Methodology, with each phase of the ETL and data warehouse implementation processes discussed. The research approach employed descriptive and case study designs, where the former pointed at providing perceptions into the research problem by describing the various variables of interest. Data collection was conducted to establish the technological state of affairs at HELB, detailing the existing divergent data sources and System architecture and data management. The project documented the research findings. The study's target population was twenty-seven (27) respondents. Questionnaires administered to ICT staff, senior and middle level managers was used as a tool to collect data. Twenty-one (21) respondents filled-in the questionnaires, which constituted a response rate of 77.77%. Data analysis is done using Excel and findings demonstrated with tables and charts. From the research findings, it is evident that the goal and specific objectives are met. The methodology ensured that the implementation of the data warehouse reflected HELB's long-term strategic plan of integrating and automating its processes and improving decision-making. This was evident from users' response during the system performance evaluation. The ETL process met the expectations of management in analyzing and generating reports in a flexible, simple and friendly way.

To realize the full benefit of a data warehouse implementation, it was recommended that HELB would continuously encourage and support staff in embracing data warehousing to reinforce decision-making.

TABLE OF CONTENTS

DECLARATION	ii
DEDICATION	iii
ACKNOWLEDGEMENT	iv
ABSTRACT.....	vi
TABLE OF CONTENTS.....	vii
LIST OF FIGURES	ix
LIST OF TABLES	x
CHAPTER ONE	1
1.0 INTRODUCTION	1
1.1. Background.....	1
1.2. Statement of the Problem.....	3
1.3. Proposed Solution	3
1.4. Objectives	3
1.5. Research Questions.....	4
1.6. Justification of the study	4
CHAPTER TWO	5
2.0 LITERATURE REVIEW	5
2.1. Introduction.....	5
2.2. The Datawarehousing Concept	6
2.3. Data Warehouse Modelling	7
2.4. ETL Tools	8
2.4.1. Surveys and Reports from Consulting Firms	8
2.4.2. Commercial Tools.....	8
2.5. How ETL Works.....	11
2.6. Conceptual models for ETL processes.....	12
2.7. Proposed ETL Conceptual Model.....	13
2.7.1. The Entity Mapping Methodology (EMM) Conceptual Model	14
2.7.2. Primitives of EMM Constructs	15
2.8. The Proposed Entity Mapping Methodology (EMM) Framework	21
CHAPTER THREE	23
3.0 RESEARCH METHODOLOGY	23
3.1. Introduction.....	23
3.2. Research design	23
3.3. Target Population.....	23
3.4. Data Collection Methods	24
3.5. Data Analysis	25
3.6. Reliability and Validity.....	26
CHAPTER FOUR.....	27
4.0 RESULTS AND DISCUSSIONS	27
4.1. Introduction.....	27
4.1.1. Response Rate	27
4.1.2. Gender of the respondents.....	27
4.1.3. Age of the respondents.....	28
4.1.4. Department of work	28
4.1.5. Managerial Position	29
4.1.5. Reports Generation	30
4.2. System Architecture and Implementation.....	31
4.2.1. System Design	31
4.2.2. System Architecture	31
4.2.3. Requirements Analysis	32

4.2.3.1.	System Users and Their Roles	32
4.2.3.2.	System Requirements.....	32
4.2.4.	Use Cases for HELB Data warehouse System.....	33
4.2.5.	Database Design.....	34
4.2.6.	Class Design.....	35
4.2.6.1.	The Data Tier Class Design	35
4.2.6.2.	The Business Tier Class Design.....	36
4.2.6.3.	The presentation Tier Class Design	37
4.3.	ETL Prototype.....	38
4.3.1.	The System's Main Interface	38
4.3.2.	Interface for Managing Databases	39
4.3.3.	Interface for managing tables extraction.....	40
4.3.4.	Interface for ETL Function	40
4.3.5.	Interface for Target Data warehouse.....	41
4.4.	Sample Reports from the data warehouse.....	41
4.5.	Performance Evaluation.....	43
4.5.1.	Evaluation of Proposed Model.....	43
4.5.2.	System Evaluation.....	43
4.5.2.1.	Usability.....	43
CHAPTER FIVE		47
5.0 CONCLUSION AND RECOMMENDATIONS		47
5.1.	Introduction.....	47
5.2.	Achievements of the Research.....	47
5.3.	Impact of the Research.....	48
5.4.	Limitations of the Research	49
5.5.	Recommendations.....	49
REFERENCE.....		51
APPENDICES		53
Appendix I: Data Collection Authorization Letter.....		53
Appendix II: Research Questionnaire		54
Appendix III: Performance Evaluation Questionnaires		56
Appendix IV: Code Snippets		57

LIST OF FIGURES

Figure 1: EMM Conceptual Model.....	15
Figure 2: Graphical constructs for the proposed EMM (Vassiliadis et al., 2002a)	17
Figure 3: Relational schema DS1 for undergraduate-loans database	18
Figure 4: Relational schema DS2 for TVET-loans database	18
Figure 5: DW_HELB Star schema for the proposed data warehouse	19
Figure 6: EMM scenario for building Loan Products	20
Figure 7: General framework for EMM model.....	21
Figure 8 : Gender of the respondents	27
Figure 9: Age of the Respondents.....	28
Figure 10: Respondent’s Department	29
Figure 11: Level of Position	30
Figure 12: The n-tier Architecture for System Development (Bradley & Millspaugh, 2009)	31
Figure 13: System level use-case diagram for the ETL tool.....	34
Figure 14: Database design for the ETL tool.....	35
Figure 15: Data layer classes for the ETL tool.	36
Figure 16: The business tier classes.....	37
Figure 17: Presentation layer classes for ETL tool.....	38
Figure 18: The Login Interface for the ETL tool	38
Figure 19: Main User interface for the ETL Tool.....	39
Figure 20: The ETL tool interface for managing databases	39
Figure 21: The ETL tool interface for managing schema tables	40
Figure 22: The ETL tool interface for ETL Function	40
Figure 23: The ETL tool interface for the target data warehouse.....	41
Figure 24: TVET Allocation per institution Data Warehouse Reports.....	41
Figure 25: TVET Allocation per individual Data Warehouse Reports.....	42
Figure 26: Undergraduate Allocation per institution Data Warehouse Reports	42
Figure 27: Undergraduate Allocation per individual Data Warehouse Reports	42
Figure 28: Respondents on System GUI Acceptance	44
Figure 29: Respondents on ease to Navigate	45
Figure 30: Respondents on System Visibility.....	45

LIST OF TABLES

Table 1: Data Collection Methods	25
Table 2: Gender of the respondents	27
Table 3: Age of the Respondents	28
Table 4: Respondent's Department.....	29
Table 5: Level of Position.....	30
Table 6: No. of Staff generating reports	30
Table 7: Functional requirements of the Data Warehouse Prototype	33
Table 8: Non- Functional requirements of the Data Warehouse Prototype.....	33
Table 9: Models Comparison and Evaluation.....	43
Table 10: Respondents on System GUI Acceptance	44
Table 11: Respondents on ease to Navigate	44
Table 12: Respondents on application Visibility	45

CHAPTER ONE

1.0 INTRODUCTION

1.1. Background

Quality and consistent data is required as a key foundation for any finance-based organization to remain competitive in today's business world. Higher education institutions and government agencies in Kenya are adopting emerging and sophisticated technology to ensure strategic and tactical decision-making is achieved. Creation of data warehouses in these institutions produce authoritative source of decision support. Data warehouse can be referred to as a descriptive high-level term - as a collection of tools and methodologies. It aims at allowing the knowledge-base expert make informed, timely and accurate results. This is achieved by exploring integrated data from diverse information systems within an organization.

Definitions of data warehouses vary as found in ("Inmo96," n.d.). "In some instances a data warehouse is defined as a subject-oriented, integrated, time-variant, non-volatile collection of data as used by most enterprises. It is defined as a central point of data integration where various know experts can manipulate various aspect of business intelligence and data marts". Data warehouses maintain critical historical data, mined from operational data storage and converted into formats available to the organization's analytics and management (Anne Marie, 2009). Historical data found in the data warehouse is cleansed, integrated and organized (Oketunji & Omodara 2011).

For successful implementation of a data warehouse, ETL tool is used to perform three key jobs:

1. data extraction from divergent data sources,
2. propagation to data staging environment for transformation and cleaning, and
3. data loading to the target data warehouse.

The problems associated to cleaning, transforming and loading information to the data warehouse is achieved through the special ETL tools (Shilakes and Tylman, 1998).

The Higher Education Loans Board (HELB), a state corporation was established through legislation in 1995. It had the authority to disburse students' loans, bursaries and scholarship to advancing their higher education in local and regional learning institutions. HELB holds huge records of students, employees, employers and other customers such as suppliers.

Voluminous historical and current data (about 2Million records) received from heterogeneous sources are stored in four main databases and servers in the Board; Web Portal PHP MySQL database; Loans Management database (Oracle 11g); HR and Payroll (Oracle 9i) used for processing statements and SQL Server 2008 for ACCPAC applications; Inventory and Fixed Assets. The Board continues to collect data more than ever before because of the increased enrollment numbers of students in higher education institutions locally and in the region. This project aimed at finding a formal and systematic representation framework for capturing the Extract, Transform and Load functionalities of a data warehouse. Data mapping from different data stores to a favourable format to realised the goal of loading data to the target data warehouse. A conceptual model was proposed that was used to model the various stages of the ETL processes.

The conceptual model reviewed the design of the ETL processes, customization, and mapping between the attributes of the sources of data together with their correspondings to the target data warehouse. This research used HELB as a case study for the development of the ETL process model in its data warehouse implementation for the purpose of improving decision making capabilities that ultimately result in enhancing the Board's service delivery and customer satisfaction.

Existing data sources at HELB

Since its inception, the Board has been using standalone applications: in-house and off-the-shelf systems for transaction processing and other operations across the organizations. They include;

1. Loans Management System: The Board's core system has two in-house sub-systems; the Lending System and Recovery and Repayment System used for loans disbursements and recoveries, respectively.
2. Human Resource and Payroll System by E-Horizon
3. Financial System and Procurement – Sage ACCPAC
4. Team Mate used for Audit operations

Data required to make informed decisions are within fragmented systems not properly integrated and not fully utilized. Synthesizing and optimizing information from these data sources has been challenging and time-consuming. In its current strategic plan, the Board has a strategic initiative to enhance its business process automation, integration and adoption by coming up with a data ware house for decision making enhancement.

1.2. Statement of the Problem

To increase efficiency in service delivery and improve timely decision-making, the Board has crucial task to integrate all workflows and processes across the Board. Much of the Board's data exist in silos. Information is not readily available across the spectrum to help the various business units of HELB access consistent and accurate data and maintain competitive edge. There is need to identify and resolve the powers and abilities data warehousing has, as a decision support strategy. The existing legacy systems and the ERP System under implementation lack the feature of storing and retrieving data centrally from historical, operational and external databases without probing the transactional databases. To counter the problem and achieve the specific goal of coming up with a data warehouse to integrate all data sources, an ETL process model needs to be used perform the ETL processes for decision making. Notwithstanding the significance of these processes, minimum research in this discipline has been conducted owing to the complexity and density. Adoption of a standard model is lacking appropriate - to represent the ETL states. The is emphases to present the ETL process in a more standard and formal method.

1.3. Proposed Solution

This project proposes to present a comprehensive ETL process model that will guide mapping of information from heterogenous data stores to the intended data warehouse in a practical manner, with a focus on the mapping and modelling phases of the ETL process. The proposed solution uses an ETL process model for data-warehousing implementation suitable for integrating all fragmented data in HELB for reporting and analytical goals. Enhancement of earlier models was used to develop the proposed model. Support was built on some previous missing mapping features.

1.4. Objectives

- 1.To investigate the methodologies employed by other researchers in coming up with ETL process models.
- 2.To identify gaps in previous research and present a conceptual design of the proposed ETL processes model.
- 3.To develop a prototype based on the identified mapping techniques that will perform the ETL functionalities and generate reports for decision support.

4. To evaluate the performance of the prototype by performing validation checks to confirm effectiveness and demonstrate that it is an improved method of mining data for decision support.

1.5. Research Questions

1. Based on existing literature and research what are functional and technical requirements critical in coming up with an ETL Process model?
2. What gaps identified in previous research will be supported in coming up with an ETL process model conceptual design?
3. Will implementing the new ETL process model enhance data mining for purposes of reporting and improved decision strategy?

1.6. Justification of the study

Data warehousing is crucial in a business world where accurate, secure and up-to-date information is required for business operations.. The implementation of a comprehensive ETL process will help mine data from heterogenous sources to the data warehouse in a practical manner, with a focus on the mapping and modelling phases of the ETL process. Such a model could not be identified in literature. The ETL functionalities will facilitate ensuring consistent data is available for generating reports to make better decisions in HELB. The proposed implementation will be in line with the Board's strategic plan of enhancing its business process automation, integration and adoption, replacing older reporting platforms with modern - day single version of the truth.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1. Introduction

Providing top level management with information, historical and current, is the key role of a data warehouse system, in order to make informed decision without interrupting the daily operations of On -Line Transaction Processing (OLTP) systems such as ERP, CRM and legacy applications. Higher education institutions and state organs have experienced competition and pressures calling for the intervention of data warehouse as the solution to engage and keep its customers, having the chance of knowing details about their needs (Han and Kamber, 2006). The use of data warehouses systems have helped to explore and store large amount of information from various sources of data (Alenazi et al. 2014). Data warehouse is considered to be an active mechanism to information integration necessary as strategic approach for government to make decisions through the use of a storage area of exact data across the value chain (“Architecture For Real-Time Analytical Data Integration And Data,” 2015”).

A data warehouse is a collection of technologies aimed at enabling the decision maker to make better and faster decisions. Data warehouses differ from operational databases in that they are subject oriented, integrated, time variant, non volatile, summarized, larger, not normalized, and perform OLAP. The generic Data warehouse architecture consists of three layers (data sources, DSA, and primary data warehouse) (Simitsis & Vassiliadis n.d.). The use of data warehouses systems have helped to explore and store large amount of information from various sources of data (Alenazi et al. 2014).

Data warehousing is considered an active mechanism to information integration necessary, as strategic approach for government to make decisions by using a storage area of exact data across the value chain (Alenazi et al. 2014). Data warehouse activities include data sourcing, data staging (ETL) and development of decision support oriented end-user application(Senapati & Kumar 2014). The ETL process is important for data warehousing efforts since it is the process through which data is loaded to the warehouse. A data warehouse cannot exist without the ETL processes, which contribute to quality of the data.

ETL tools are specialized tools that deal with data warehouse heterogeneity, cleaning and loading. The extraction phase converts the data into a single format necessary for transformation

processing. Through data cleaning, duplicated records are eliminated, inconsistencies detected and sources of errors found in data.

The data used in ETL processes can come from any source: a mainframe application, database tables, an ERP application, a CRM tool, the internet, a flat file or an Excel spreadsheet. Although ETL processes area is very important, it has little research. 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 (Mawilmada 2011). To build a Data warehouse one must run the ETL tool which has three tasks:

- (1) data is extracted from different data sources,
- (2) propagated to the data staging area where it is transformed and cleansed, and then
- (3) loaded to the data warehouse.

ETL tools are a category of specialized tools with the task of dealing with data warehouse homogeneity, cleaning, transforming, and loading problems (“Shilakes and Tylman,” n.d.). This research will try to find a formal representation model for capturing the ETL processes that map the incoming data from different data sources to be in a suitable format for loading to the target data warehouse.

2.2. The Datawarehousing Concept

Data warehousing technology aims to structure the data in an appropriate way to access the data, and use it in an efficient and effective manner (Mawilmada 2011). As stated by Kerkri, Quantin, Allaert, Cottin, Charve, Jouanot and Yétongnon (2001), the Data warehouse is responsible for the consistency of information. The integration of tools such as query tools, reporting tools and analysis tools provide opportunity to handle the coherence of information. The aim of data warehousing is to organise the gathering of a wide range of data and store it in a single repository (Kerkri et al., 2001). Currently, data warehousing plays a major role in the business community at large. It is also relevant to healthcare as mentioned in del Hoyo-Barbolla and Lees (2002, p. 43), “in a competitive climate, if healthcare organisations are to keep their customers, knowing and managing information about them is essential and organisations realized that it is crucial to access viable and timely data.” Furthermore, integrating data from the different sources and converting them into valuable information is a way to obtain competitive advantage (del Hoyo-Barbolla & Lees, 2002).

Data warehousing is “a collection of decision support technologies aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions” (Parmar et al. 2016). According to Inmon (2005, p.29) Data warehouse is a “subject-oriented, integrated, time-variant and non-volatile collection of data in support of management decisions”. March and Hevner (2007) argued that the three components of intelligence namely understanding, adaptability and profiting from experience are important considerations when designing the data warehouse.

Also, these authors mentioned that the Data warehouse should allow managers to gather information such as identifying and understanding different situations and the reasons for their occurrence. Further, they have argued that the, Data warehouse should “enable a manager to locate and apply the relevant organizational knowledge and to predict and measure the impact of decision over time” (March & Hevener, 2007, p.1035). However, as mentioned by March and Hevner (2007), these arguments forms the challenges that need to be considered when implementing a data warehouse.

2.3. Data Warehouse Modelling

Data warehouse modelling is generally divided into three major processes; requirements analysis, conceptual design and the ETL processes. Raw data sent to the data warehouse has different formats and the processing needs for each of these format types require accurate approaches in order to be accommodated to the specific mappings. Identity management, secured transactions, XML or text parsing, and digital signal processing need to arrive to the data warehouse with regard to its requirements. Requirements definition is an important part of designing a data warehouse. Data needed already exists within multiple transaction systems. Necessary information based on common dimensions is gathered from the sources, then mined and transformed to meet the needs of the Users.

A number of brainstorming sessions to develop the OLAP queries needs to be conducted to capture the business questions, all assigned to categories. Samples of data from divergent sources; the legacy systems, external sources, ERP Systems should be mined to the data warehouse. Conversion and cleaning issues are addressed through an ETL tool. The process had defined stages: Database Stage; ETL Stage; Warehousing stage; Tools stage and Interface Stage. ETL tool (s) will be used.

2.4. ETL Tools

2.4.1. Surveys and Reports from Consulting Firms

In a recent study (Russom n.d.), the authors reported that due to the diversity and heterogeneity of data sources, ETL was unlikely to become an open commodity market. The ETL market had reached a size of six hundred million dollars for year 2001; still the growth rate had reached a rather low 11% (as compared with a rate of 60% growth for year 2000). That was explained by the overall economic downturn environment. In terms of technological aspects, the main characteristic of the area was the involvement of traditional database vendors with ETL solutions built in the DBMS's.

The three major database vendors that practically shipped ETL solutions 'at no extra charge' were pinpointed: *Oracle* with Oracle Warehouse Builder [Orac03], *Microsoft* with Data Transformation Services [Micr03] and *IBM* with the Data Warehouse Center [IBM03]. Still, the major vendors in the area were Informatica's Powercenter [Info03] and Ascential's DataStage suites [Asce03, Asce03a] (the latter being part of the IBM recommendations for ETL solutions). The study went on to propose future technological challenges/forecasts that involved the integration of ETL with (a) XML adapters, (b) EAI (Enterprise Application Integration) tools (e.g., MQ-Series), (c) customized data quality tools, and (d) the move towards parallel processing of the ETL workflows. The aforementioned discussion was supported from a second recent study [Gart03], where the authors noted the decline in license revenue for pure ETL tools, mainly due to the crisis of IT spending and the appearance of ETL solutions from traditional database and business intelligence vendors. The Gartner study discussed the role of the three major database vendors (IBM, Microsoft, Oracle) and pointed that they would take a portion of the ETL market through their DBMS-built-in solutions. By 2007, more than 50 percent of new data warehouse deployments had used ETL tools provided by the major DBMS vendors, IBM, Microsoft, and Oracle (0.7 probability) [Frie04].

2.4.2. Commercial Tools

In an overall assessment commercial ETL tools were responsible for the implementation of the data flow in a data warehouse environment. Most of the commercial ETL tools were of two kinds: engine-based or code-generation based. The former assumed that all data had to go through an engine for transformation and processing.

Moreover, quite often the engine took over the extraction and loading processes, making the ExtractTransform Load processes one big process, where the intermediate steps were transparent to the user. On the other hand, in code-generating tools all processing took place only at the target or source systems.

Although, there were a variety of ETL tools in the market the researcher elaborated more on the major vendors.

IBM. DB2 Universal Database offered the Data Warehouse Center [IBM03], a component that automated data warehouse processing, and the DB2 Warehouse Manager that extended the capabilities of the Data Warehouse Center with additional agents, transforms and metadata capabilities. Data Warehouse Center was used to define the processes that move and transform data for the warehouse. Warehouse Manager was used to schedule, maintain, and monitor these processes. Within the Data Warehouse Center, the warehouse schema modeler was a specialized tool for generating and storing schema associated with a data warehouse. Any schema resulting from this process could be passed as metadata to an OLAP tool. The process modeler allowed user to graphically link the steps needed to build and maintain data warehouses and dependent data marts. DB2 Warehouse Manager included enhanced ETL function over and above the base capabilities of DB2 Data Warehouse Center. Additionally, it provided metadata management, repository function, as such an integration point for third-party independent software vendors through the information catalog.

Microsoft. The tool that was offered by Microsoft to implement its proposal for the Open Information Model was presented under the name of Data Transformation Services[Micr03, BeBe99]. Data Transformation Services (DTS) were the data-manipulation utility services in SQL Server (from version 7.0) that provided import, export, and data-manipulating services between OLE DB [Micr03a], ODBC, and ASCII data stores. DTS were characterized by a basic object, called a package, that stored information on the aforementioned tasks and the order in which they need to be launched. A package could include one or more connections to different data sources, and different tasks and transformations that were executed as steps that define a workflow process [GSB+01]. The software modules that support DTS were shipped with MS SQL Server. These modules included: –DTS Designer: A GUI used to interactively design and execute DTS packages –DTS Export and Import Wizards: Wizards that eased the process of defining DTS packages for the import, export and transformation of data –DTS Programming

Interfaces: A set of OLE Automation and a set of COM interfaces to create customized transformation applications for any system supporting OLE automation or COM.

Oracle. Oracle Warehouse Builder [Orac01, Orac02a, Orac03] was a repository-based tool for ETL and data warehousing. The basic architecture comprised two components, the design environment and the runtime environment. Each of these components handled a different aspect of the system; the design environment handled metadata, the runtime environment handled physical data. The metadata component revolved around the metadata repository and the design tool. The repository was based on the Common Warehouse Model (CWM) standard and consisted of a set of tables in an Oracle database that were accessed via a Java-based access layer. The front-end of the tool (entirely written in Java) features wizards and graphical editors for logging onto the repository. The data component revolved around the runtime environment and the warehouse database. The Warehouse Builder runtime was a set of tables, sequences, packages, and triggers that were installed in the target schema.

The code generator that bases on the definitions stores in the repository, it creates the code necessary to implement the warehouse. Warehouse Builder generates extraction specific languages (SQL*Loader control files for flat files, ABAP for SAP/R3 extraction and PL/SQL for all other systems) for the ETL processes and SQL DDL statements for the database objects. The generated code is deployed, either to the file system or into the database. Ascential Software. DataStage XE suite from Ascential Software [Asce03, Asce03a] (formerly Informix Business Solutions) is an integrated data warehouse development toolset that includes an ETL tool (DataStage), a data quality tool (Quality Manager), and a metadata management tool (MetaStage). The DataStage ETL component consists of four design and administration modules: Manager, Designer, Director, and Administrator, as such a metadata repository, and a server.

DataStage. The DataStage Manager is the basic metadata management tool. In the Designer module of DataStage, ETL tasks execute within individual 'stage' objects (source, target, and transformation stages), in order to create ETL tasks. The Director is DataStage's job validation and scheduling module. The DataStage Administrator is primarily for controlling security functions. The DataStage Server is the engine that moves data from source to target.

Informatica. Informatica PowerCenter [Info03] is the industry-leading (according to recent studies [FrGa04, Gart03, Giga02]) data integration platform for building, deploying, and managing enterprise data warehouses, and other data integration projects. The workhorse of Informatica PowerCenter is a data integration engine that executes all data extraction, transformation, migration and loading functions in-memory, without generating code or requiring developers to hand-code these procedures.

2.5. How ETL Works

During the ETL process, data was extracted from an OLTP database, transformed to match the Data warehouse schema, and loaded into the Data warehouse database (Sharma & Gupta 2012). Many data warehouses also incorporated data from non-OLTP systems, such as text files, legacy systems, and spreadsheets. ETL was often a complex combination of process and technology that consumed a significant portion of the Data warehouse development efforts and required the skills of business analysts, database designers, and application developers. The ETL process was not a one-time event.

As data sources changed, the Data warehouse would be periodically updated. Also, as business changed the Data warehouse system needed to change – in order to maintain its value as a tool for decision makers, as a result of that the ETL also changed and evolved. The ETL processes were designed for ease of modification. A solid, well-designed, and documented ETL system was necessary for the success of a Data warehouse project. An ETL system consisted of three consecutive processes described as follows;

1. Extract - the process of reading data from a specified source database and extracting a desired subset of data.
2. Transform - the process of converting the extracted/ acquired data from its previous form into the form it needed to be in so that it could be placed into another database. Transformation occurred by using rules or lookup tables or by combining with other data.
3. Load - the process of writing the data into the target database.

2.6. Conceptual models for ETL processes

The ETL process, in data warehouse, was a hot point of research because of its importance and cost in data warehouse project building and maintenance. Although the ETL processes were critical in building and maintaining the data warehouse systems, there was a clear lack of a standard model that could be used to represent the ETL scenarios. This section navigated through the efforts done by other researcher to conceptualize the ETL processes.

There were few attempts around the specific problem of this work. The research mentioned [BoFM etl, 1999] as the first attempt to clearly separate the data warehouse refreshment process from its traditional treatment as a view maintenance or bulk loading process. Still, the proposed model was informal and the focus was on proving the complexity of the effort, rather than the formal modeling of the processes themselves.

[CDL+ 1998, CDL+ 1999] introduced the notion of ‘intermodel assertions’, in order to capture the mappings between the sources and the data warehouse. However, any transformation was dereferenced for the logical model where a couple of generic operators were supposed to perform the task.

In [StMR etl, 1999] the authors proposed an UML-based metamodel for data warehouses. The metamodel covered the two basic parts of the data warehouse architecture, that were, the backstage and the front-end. Different kinds of metadata applied to these two parts. For the front end, Business Concepts were introduced, in order to cover the reference business model of the data warehouse. OLAP features like cubes, dimensions and facts were also introduced for the front-end, too. For the backstage of the data warehouse, the authors covered the workflow from the sources towards the target data stores. Entities like Mapping (among entities) and Transformation (further classified to aggregations, filters, etc.) were employed to cover the inter-concept relationships in the workflow environment. The overall approach was a coherent, UML-based framework for data warehouse metadata, defined at a high-level of abstraction. Specialized approaches for specific parts (like definitions of OLAP models, or ETL workflows) could easily be employed in a complementary fashion to the [StMR etl, 1999] framework (possibly through some kind of specialization) to add more detail to the metadata representation of the warehouse.

Clearly, the authors provided a framework for the modeling of data warehouses; but still was too general to capture the peculiarities of a more detailed level such as the modeling of ETL processes.

The authors in [TrLu etl, 2003] presented a conceptual model based on the Unified Modeling Language (UML) [OMG01] for the design of ETL processes which dealt with several technical problems. They presented a UML-based mechanism that represented common ETL operations such as aggregation, conversion, filtering, join, loading, incorrect, merge, wrapper, and surrogatekeys. Still, their method inherited the advantages and the disadvantages of UML modeling: it was based on a well-established modeling technique, but its deliverable was not clear and easy to understand at the early stages of a data warehouse project, where people with different knowledge background (business managers and database administrators) were met. And this was realized, especially because of the encapsulation into classes of very crucial elements of an ETL process, such as attributes and relationships among them.

In terms of industrial approaches, the model that stemed from [KRRT etl, 1998] would have been an informal documentation of the overall ETL process. Moreover, [KRRT etl,1998] presented the only methodology that was known, mainly grouped as a set of tips and technical guidelines. The major deficiency of the [KRRT etl, 1998] methodology was the lack of conceptual support for the overall process. As far as inter-attribute mappings were concerned, tables of matching between source and warehouse attributes were suggested. The spirit of [KRRT etl, 1998] was more focused towards the logical design of ETL workflows; thus, rather than systematically tracing system and user requirements, transformations and constraints, the proposed methodology quickly turned into a description of data flows for common cases of ETL transformations(e.g., surrogate keys, duplicate elimination and so on).

2.7. Proposed ETL Conceptual Model

In this section, the research focused on the conceptual part of the definition of the ETL processes. More specifically, the project dealt with the earliest stages of the data warehouse design. During this period, the research was concerned with two tasks that were practically executed in parallel:

- i. The first of these tasks involved the collection of requirements from the part of the users.
- ii. The second task, which was of equal importance for the success of the data warehousing project, involved the analysis of the structure and content of the existing data sources and their intentional mapping to the common data warehouse model.

The design of an ETL process aimed at the production of a crucial deliverable: *the mapping of the attributes* of the data sources to the attributes of the data warehouse tables.

The production of this deliverable involved several interviews that resulted in the revision and redefinition of original assumptions and mappings; thus it was imperative that a simple conceptual model was to be employed in order to facilitate the smooth redefinition and revision efforts and to serve as the means of communication with the rest of the involved parties.

To conceptualize the ETL processes used to map data from sources to the target data warehouse schema, the researcher studied the previous research projects, made some integration, and added some extensions to the approaches mentioned above. As indicated in the previous research projects, the front end of the data warehouse had monopolized the research on the conceptual part of data warehouse modeling and most research efforts was dedicated to capturing of the conceptual characteristics of the star schema of the warehouse and the subsequent data marts and aggregations.

2.7.1. The Entity Mapping Methodology (EMM) Conceptual Model

EMM was the proposed conceptual model for modeling the ETL processes which was required to map data from sources to the target data warehouse schema. In Fig.1 abstractly describes the general model for ETL processes. The bottom layer depicts the data stores that are involved in the overall process. On the left side, the original data providers can be observed. Typically, data providers are relational databases and files. The data from these sources are extracted (as shown in the upper left part of Fig.1) by extraction routines, which provide either complete snapshots or differentials of the data sources.

Then, these data are propagated to the Data Staging Area (DSA) where they are transformed and cleaned before being loaded to the data warehouse. The data warehouse is depicted in the right part of the data store layer and comprises the target data stores. The loading of the central warehouse is performed from the loading activities depicted on the upper right part of the figure (Load).

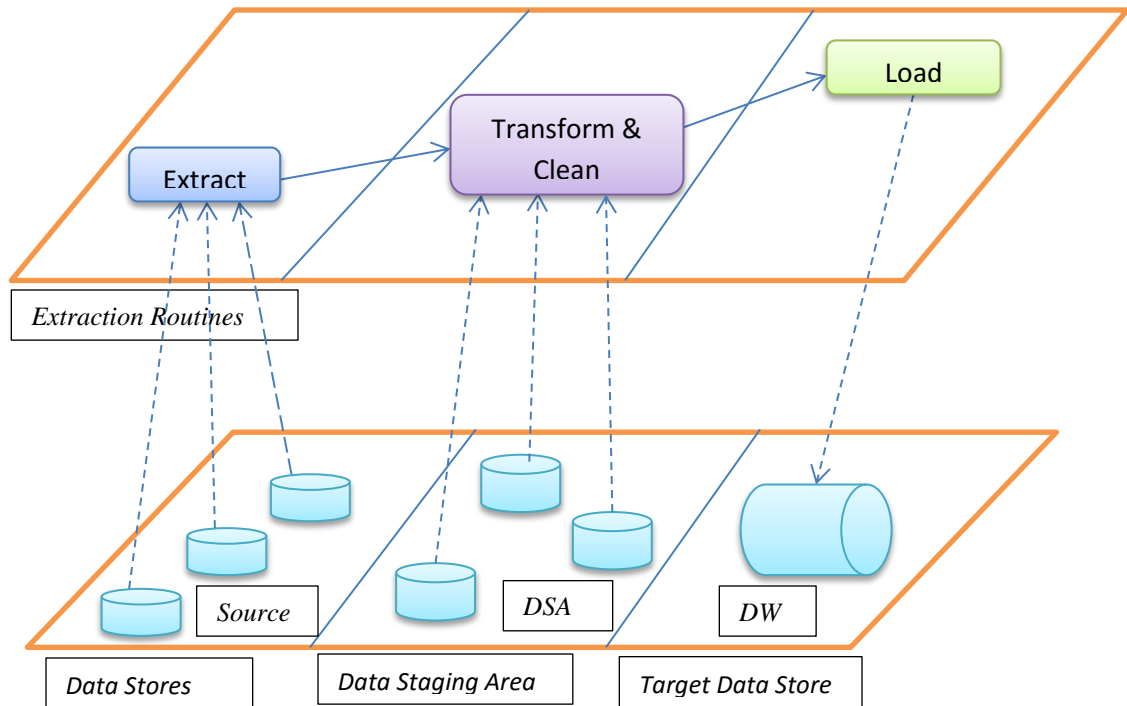


Figure 1: EMM Conceptual Model

In this paper, the researcher complemented this model in a set of design steps, which lead to the basic target, i.e., the attribute interrelationships. These steps constituted the methodology for the design of the conceptual part of the overall ETL process and could be summarized as follows:

- (a) identification of the proper data stores;
- (b) candidates and active candidates for the involved data stores;
- (c) attribute mapping between the providers and the consumers, and
- (d) annotating the diagram with runtime constraints (e.g., time/event based scheduling, monitoring, logging, exception handling, and error handling).

2.7.2. Primitives of EMM Constructs

The basic set of constructs that is used in the proposed Entity Mapping Methodology is shown in Fig. 2. In this section, some explanation about the usage of the constructs of the proposed Entity Mapping Methodology were given, as follows:

- *Loader relationship*: was used when the data was moved directly from the last source element (the actual source or the temporary one) to the target data element. The actual

source; was the base source from which the data were extracted, on the other hand, the temporary source; was the one that was resulted during the transformation operations.

- *Optional loader relationship*: was used to show that the loaded data to the output attribute could be extracted from candidate source element x or candidate source element y.
- *Convert into structure*: represented the conversion operations required to restructure the non-structured base source into structured one (relations as tables and attributes). The conversion operation saves its result into temporary tables, so the transformation operation could be applied to the new temporary source.
- *Entity transformation operation*: this kind of transformations usually resulted in a temporary entity. There were standard operators that were used inside this construct.
- *Attribute transformation operation*: standard operations were used with this construct. User defined function (UDF) as a transformation operation: user could use his defined operations, so any kind of transformation could be added, such as currency conversion functions, packages (units) conversions, and so on.
- *Non-structured source*: represented any source that was not in the relational structure. The non-structured source may be semi-structured or unstructured source such as XML files, web logs, excel workbook, object oriented database, etc.






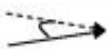





Mapping Construct		To Represent
Name	Shape	
Cylinder		Schema
Rectangle		Entity
Oval		Attribute
Diamond with rounded arrow		Convert into structure
Solid arrow		Loader Relationship
Connected arrows		Optional Loader Relationship
Square with rounded edge		Attribute Transformation
Square with triangle edge		User Defined Function (UDF)
Hexagon		Entity Transformation operation
Document		Non-structured source
Rectangle with folded corner		User Note

Figure 2: Graphical constructs for the proposed EMM (Vassiliadis et al., 2002a)

The transformation functions that took place in the proposed model EMM were classified into built-in or standard functions, such as join, union, and rename, and user defined functions as mentioned above, like any formula defined by the user. Another classification for the transformation functions according to the level of transformation was entity transformations functions, and attributes transformations functions.

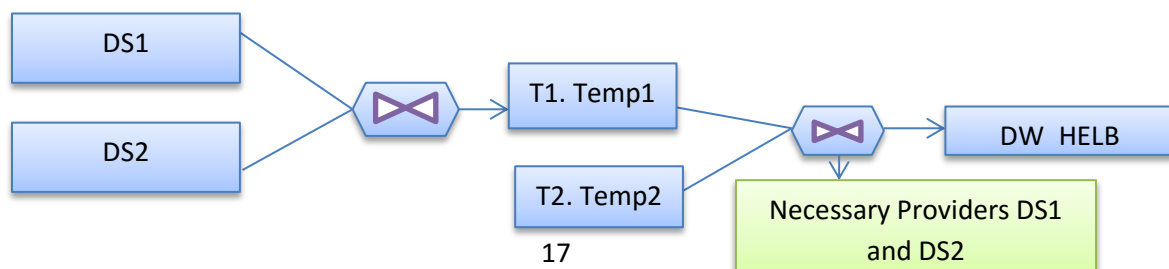
2.7.3. Methodology for the usage of the conceptual model

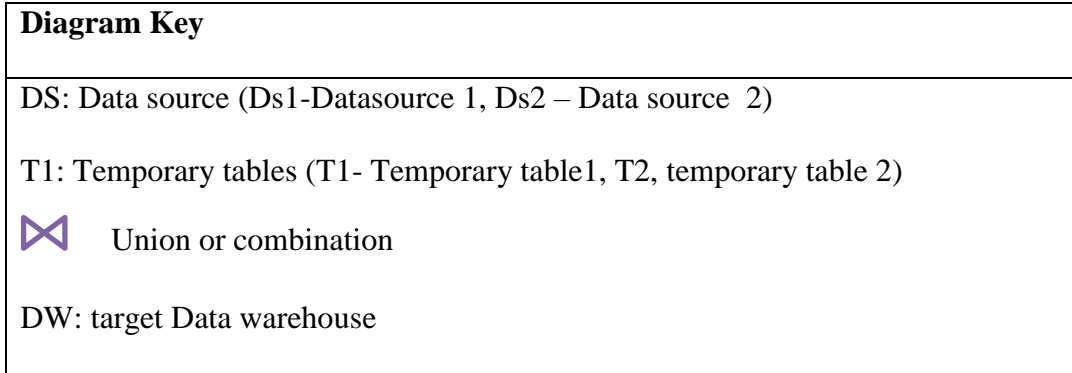
In this section, the researcher presented the use of the methodology, by giving the sequence of steps for a designer to follow, during the construction of the data warehouse. Each step of the methodology was presented in terms of the example of HELB case scenario (with the hope of clarifying the nature of the employed modeling entities). As already mentioned, the ultimate goal of the design process was the production of inter-attribute mappings, along with any relevant auxiliary information.

The ultimate goal was to build HELB a data warehouse for monitoring loan uptake for its two major products (undergraduate loans for chartered universities and Diploma loans for TVET institutions) in order to make informed decisions on projected budgeting for loans in relation to students' number and loan uptake. It has a relational data source described by schema DS1 for undergraduate loans, shown in Fig. 3, another relational data source described by schema DS2 for TVET loans, shown in Fig. 4. A relational data warehouse was designed to capture student loans data from the two predefined data sources. The star schema in Fig. 5 shows the design of the proposed data warehouse, which consists of one fact table and four dimensions tables.

Step 1: Identification of the proper data stores.

The first thing that a designer faces during the requirements and analysis period of the data warehouse process is the identification of relative data sources. Assume that for a particular subset of the data warehouse, we have identified the concept DW_HELB which is a fact table of how loans were distributed according to respective allocations to students.





Source: Author (2016)

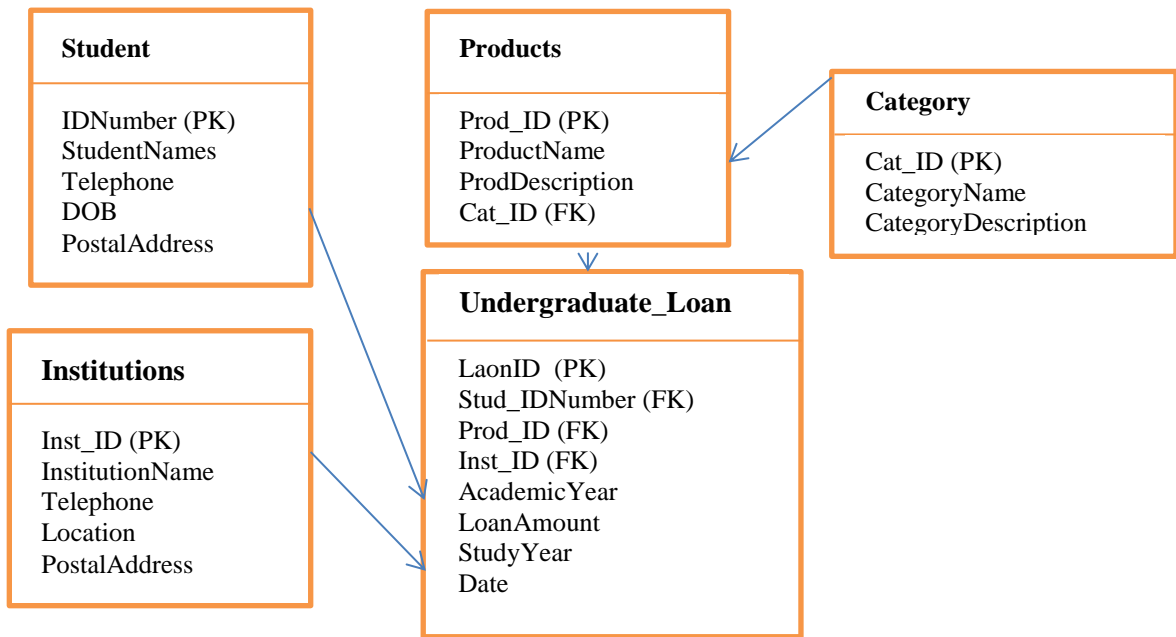


Figure 3: Relational schema DS1 for undergraduate-loans database

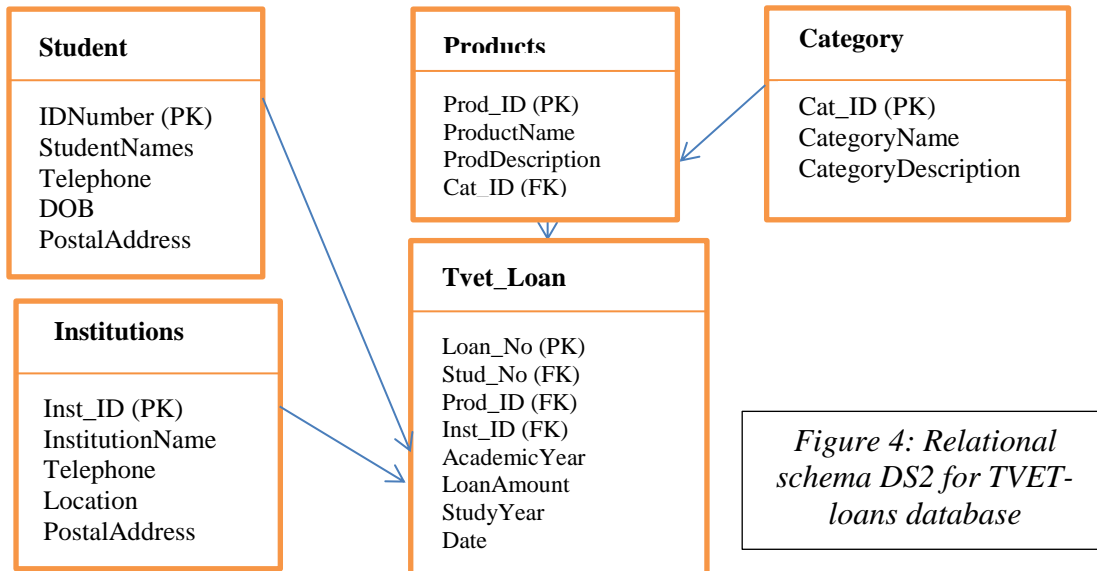


Figure 4: Relational schema DS2 for TVET-loans database

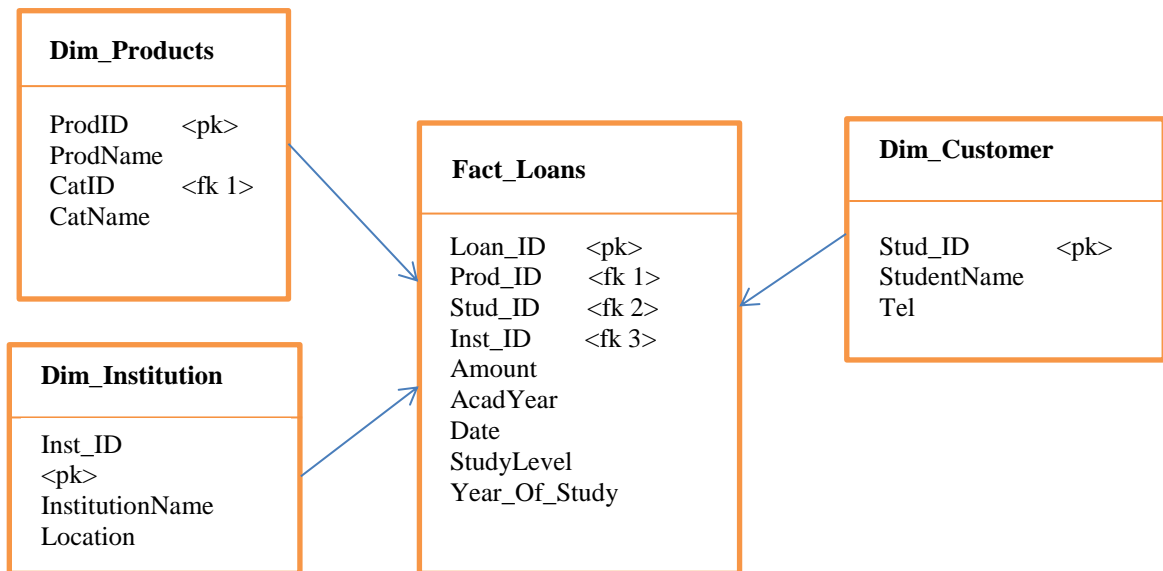


Figure 5: DW_HELB Star schema for the proposed data warehouse

Step 2: Attribute mapping between the providers and the consumers.

The most difficult task of the data warehouse designer is to determine the mapping of the attributes of the sources to the ones of the data warehouse. This task involves several discussions with the source administrators to explain the codes and implicit rules or values, which are hidden in the data and the source programs. Moreover, it involves quite a few ‘data preview’ attempts (in the form of sampling, or simple counting queries) to discover the possible problems of the provided data. For each target attribute, a set of provider relationships must be defined. In simple cases, the provider relationships are defined directly among source and target attributes. In the cases where a transformation is required, the researcher passes the mapping through the appropriate transformation. In the cases where more than one transformations are required, the researcher can insert a sequence of transformations between the involved attributes, all linked through composition relationships. ETL constraints are also specified in this part of the design process.

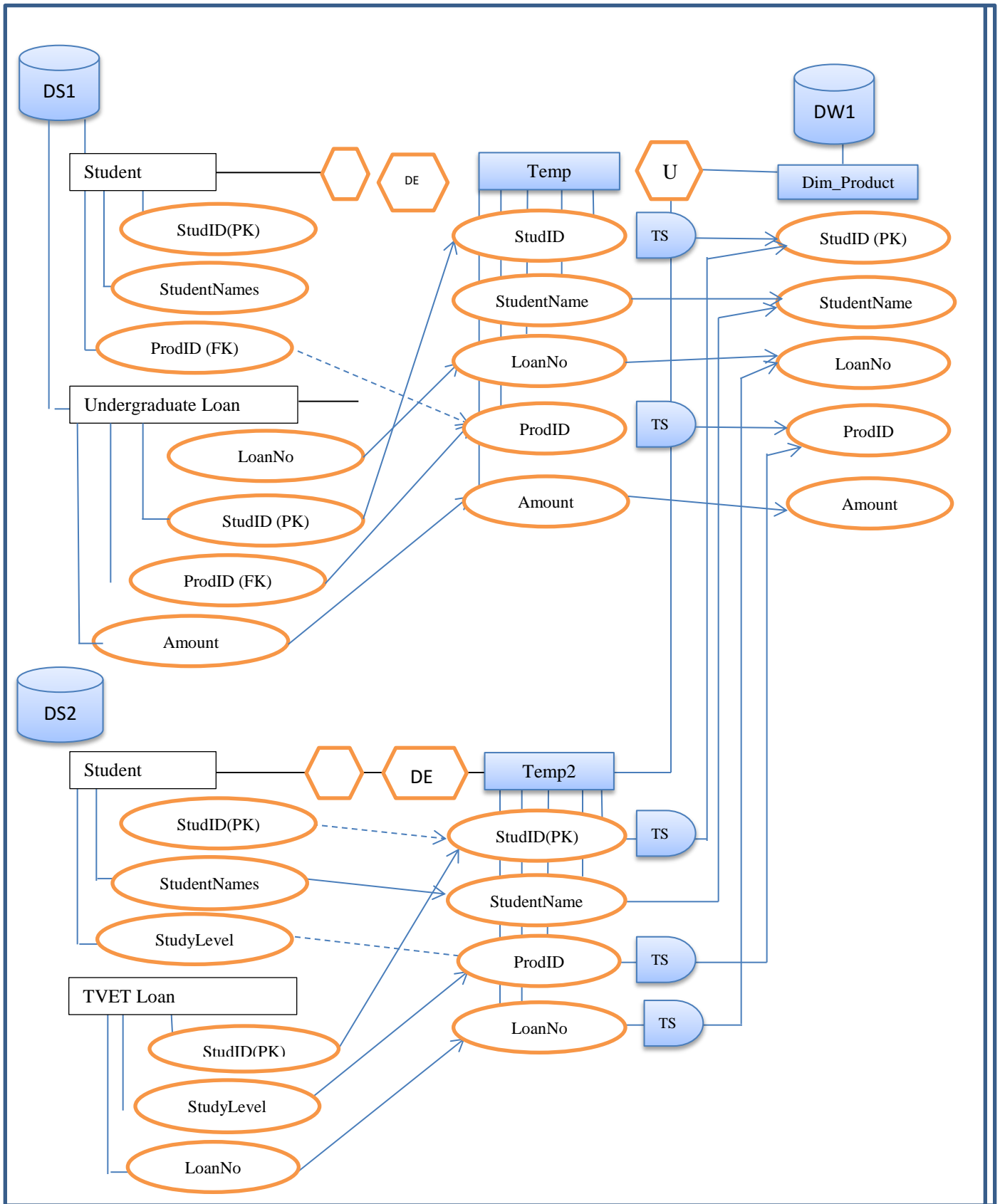


Figure 6: EMM scenario for building Loan Products

Diagram Key
DS: Data source
TS: Transformation
Temp: Temporary tables
DW: target Data warehouse

Source: Author (2016)

Step 3: Annotating the diagram with runtime constraints

Apart from the job definition for an ETL scenario, which specifies how the mapping from sources to the data warehouse is performed, along with the appropriate transformations, several other parameters possibly need to be specified for the runtime environment. This kind of runtime constraints include:

- **Time/Event based scheduling:** The designer needs to determine the frequency of the ETL process, so that data are fresh and the overall process fits within the refreshment time window.
- **Monitoring:** On-line information about the progress/status of the process is necessary, so that the administrator can be aware of what step the load is on, its start time, duration, etc. File dumps, notification messages on the console, e-mail, printed pages, or visual demonstration can be employed for this purpose.

2.8. The Proposed Entity Mapping Methodology (EMM) Framework

Fig. 7 below shows the general framework of the proposed entity-mapping diagram.

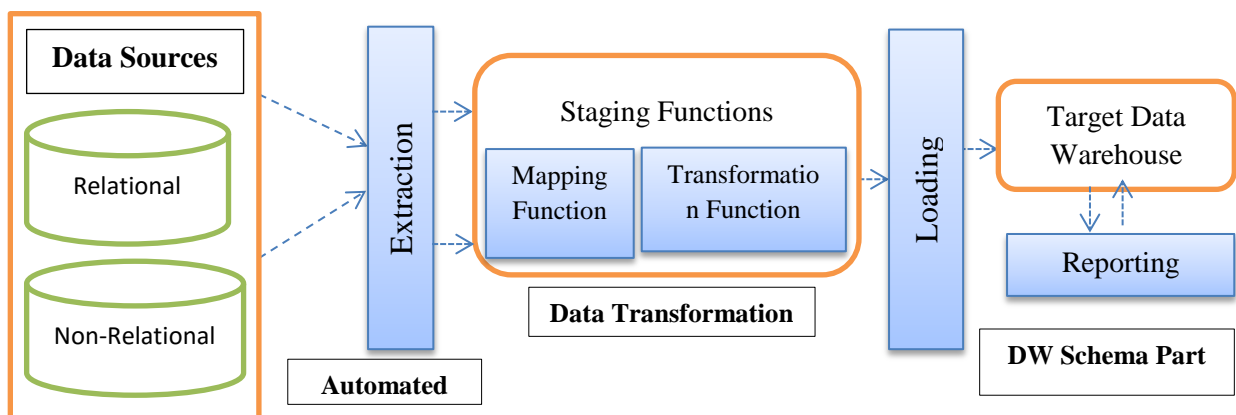


Figure 7: General framework for EMM model

In the data source(s) part: the participated data sources were drawn. The data sources may be structured databases or non-structured sources. In case of structured sources; the participated databases and their participated tables and attributes were used directly as the base source, and in case of non-structured sources; a conversion step was applied first to convert the non-structured source into structured one (tables and its attributes). From the design view, there was one conversion construct that could convert any non-structured source into structured (relational) database, but from the implementation view, each type of non-structured source had its own conversion module which was called wrapper.

Wrappers are specialized program routines that automatically extract data from different data sources with different formats and convert the information into a structured format. The typical tasks of a wrapper are: (a) fetching data from a remote resource, (b) searching for, recognizing and extracting specified data, and (c) saving this data in a suitable structured format to enable further manipulation (Ferreira & Furtado n.d.).

1. **Extraction:** during the extraction process some temporary tables were created to hold the result of converting non-structured sources into databases. The extraction process included initial extraction and refresh. The initial extraction took place when the ETL scenario executed for the first time while there was no data in the destination data warehouse. The refresh extraction took place to capture the delta data (difference between old data in the DW and updated data in the data sources). It was preferred to separate the ETL scenario with initial extraction from the ETL scenario with refresh extraction. This means that the user needed to build two EMM models for the same ETL scenario; one for the initial extraction, and the other for the refresh extraction using the old data in the temp tables found in the staging area.
2. **In the DW schema part:** the data warehouse schema table (fact or dimension) was drawn. In spite of that the fact table and the dimension table were clearly different in their functionalities and features but all of them were data containers. Basically the data warehouse was stored as relational structure not as multidimensional structure. The multidimensionality occurs in the online analytical processing (OLAP) engines.
3. **In the mapping part:** the required transformation functions were drawn. The transformation operations took place on the incoming data from both the base source and/or the temporary source in the staging area. Some transformation operations lead to temporary results which were saved in temporary tables in the staging area.
4. **The staging area:** a physical container that contained all temporary tables created during the extraction process or resulted from the applied transformation functions.
5. **Loading:** as the data reached the final appropriate format, it was loaded to the corresponding data element in the destination DW schema from the staging area.

CHAPTER THREE

3.0 RESEARCH METHODOLOGY

3.1.Introduction

The aim of this chapter was to gather data, the methods and requirements needed for the architecture modelling to resolve the existing problem. Research design embraced throughout the study and together with justifications, is provided in this chapter. The target group from which data was obtained to answer the research questions, is provided. The project adopted the approach of analyzing the problem, making a proposal to solve the problem, development of the model/prototyping, evaluation and closure.

3.2.Research design

The “5Ws” and “H” are addressed; what, when, where, how much and by what means in relation to a research study. The conditions for collection and analysis of data are well arranged with the ultimate goal of achieving relevance to the project’s purpose. According to (Selltiz et al, 1962) research design is a conceptual structure that acts as a guiding roadmap and plan for the collection, measurement and analysis of data.

The study adopted descriptive and case study designs. The descriptive research design intended to provide insights into the research problem, achieved through describing the variables of interest. Descriptive research design was used when collecting information with the main purpose of illustrating the state-of-affairs at HELB, detailing the existing divergent data sources and the methods used for decision-making. The adopted designs was preferred since it is simpler to use and convenient.

3.3.Target Population

Population is defined as the group to which a researcher would like the results of the study to be generalizable, in addition to possibly been a set of all cases of interest (Quality & Line n.d.). The size and geographical location of a population may vary and take any form (Gay and Diehl, 1992).

HELB technical ICT staff members were interrogated to explain the existing system architecture and data management systems in place. They included the System Administrators, Database administrators, System analysts and Systems Security administrator, Network Administrator and senior technical Support staff. The technical experts helped determine the warehouse design based on classifications of OLAP queries collected. They described the existing System architectures and data quality.

HELB management and Business Units champions were interviewed to determine the business processes of the Board. The aim was to prioritize and categorize data. Data collected from Management was used to key business decisions; the type and size of data to load to the data warehouse. Four senior heads of departments will be interviewed; Operations, Finance, ICT and Research, Planning and Strategy. The Board's Management determined the cut-off years for transactional history. Other target population was members of the Records and Digitization Committee, responsible for the digitization of records and information in the Board. The target population was 27 respondents.

3.4.Data Collection Methods

The primary data collection sources were questionnaires, both closed and open-ended questions. While effectively using closed ended questions time and other resources were saved and facilitated an easier analysis. Open-ended questions gave in-depth responses and better view of all relevant information of existing System architecture and status of data quality and types. The use of open-ended questions gave an insight to respondent's reaction based on their feelings, background, hidden motivation, interest and decisions.

Interviews were conducted on representatives target groups. Management were interviewed too involved in the decision making process of HELB. The Users will be engaged to seek response and their scale of satisfaction. The review of existing literature relevant to this project will be thoroughly done with key focus on decision support, data warehouse development techniques in higher education. Efforts will be made to review these secondary sources as the most recent contributions.

Type of Data/Information	Source of Data	Target Group	Population	Data/Information Collected
Historical data and trends	<ul style="list-style-type: none"> Legacy System External Sources 	<ul style="list-style-type: none"> System Analysts Records Digitization Committee 	2 2	<ul style="list-style-type: none"> Past and existing technologies and tools at the Board. Data volumes and architecture. Nature of data (accuracy and completeness). Data collected through interviews and questionnaires
Business/Operational Needs	<ul style="list-style-type: none"> HELB Strategic Plan ICT Plan and documentation 	<ul style="list-style-type: none"> Heads and Management Departmental Champions 	4 4	<ul style="list-style-type: none"> Business/Operational needs, policies and processes Business/Operational decisions and procedures Data collected through interviews and questionnaires
Existing System Architecture	<ul style="list-style-type: none"> ICT Plan; System Security and Infrastructure Plan 	<ul style="list-style-type: none"> Head of ICT Technical staff 	1 6	<ul style="list-style-type: none"> Architecture of ERP System and other existing platforms. ETL Process requirements Data collected through surveys and recent studies
Functional Requirements of DWH	<ul style="list-style-type: none"> DWH Proposed Model/LR 	<ul style="list-style-type: none"> Business Experts 	2	<ul style="list-style-type: none"> Business Needs and requirements Information gathered through surveys and recent studies and models
Technical Requirements of DWH	<ul style="list-style-type: none"> DWH Proposed Model/LR 	<ul style="list-style-type: none"> System Analysts System and Network Administrators Database Administrator Other technical staff 	2 2 2	<ul style="list-style-type: none"> Infrastructure Requirements Logical and Conceptual design requirements Physical design requirements Information gathered through surveys and recent studies and models

Table 1: Data Collection Methods

3.5.Data Analysis

After a description of the data collection methodology and the general characteristics of the respondents, a descriptive analysis was conducted. To the extent that data integration to a data warehouse is not currently prevalent in HELB, the substantive findings from the descriptive analysis enrich the body of existing knowledge on data integration and warehousing and its adoption in the Board for decision support.

Data analysis is the method used to describe fact, detect patterns, develop explanations and test hypotheses. The questions and responses from the schedules were coded and entered into the computer using Microsoft Excel 2010 software. Microsoft Excel was used for the data analyses.

3.6. Reliability and Validity

By definition, reliability refers to how consistent a given set of measurements are achievable. On the other hand, validity states the degree to which the research mirrors to the given actual research problems.

Evaluating and comparing the validity of results was emphasized. It can also be defined as the repeatability of findings where same results are received upon each single test run or while observing behavior of some events. On the other hand, validity refers to the credibility or believability of the research project.

There is key relationship between reliability and validity. For instance if data is valid it can deduced that they must be reliable. However, it is worth noting that reliability is necessary but not sufficient, it been a condition for validity.

CHAPTER FOUR

4.0 RESULTS AND DISCUSSIONS

4.1. Introduction

4.1.1. Response Rate

This project targeted 27 respondents during data collection with regard to the development of the ETL process for data warehouse implementation. The outcome indicated that 21 out of the 27 respondents were presented with questionnaires and were able to return them, making a response rate of 77.77%. As documented by Mugenda (2003) if an above 60% response rate is achieved, the same is considered appropriate for credible results. During this research project the reasonable response rate was achieved.

4.1.2. Gender of the respondents

This research project established the gender of the respondents. The findings indicated that 67% of the respondents were male while 33% constituted female respondents: This inferred that majority of the employees dealing with Systems were majorly male as depicted by the findings.

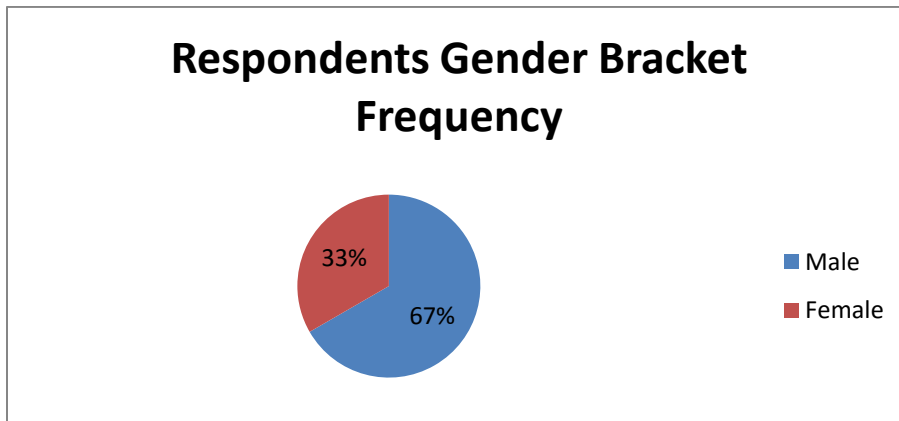


Figure 8 : Gender of the respondents

Respondents Gender Bracket		
Gender	Frequency	Percentage
Male	14	67%
Female	7	33%
Total	21	100%

Table 2: Respondents' Gender

4.1.3. Age of the respondents

This research project established the age of the respondents. The findings indicated that, 52 % of the respondents were aged 30-40 years, 29% aged 18-30 years, and 19% of the respondents were above 40 year. From the findings it can be deduced that majority of the staff in the organization are made up of old employees of age between 30-40 years. It can be drawn that most of these employees have had experience in accessing data from the Information Systems in the Board.

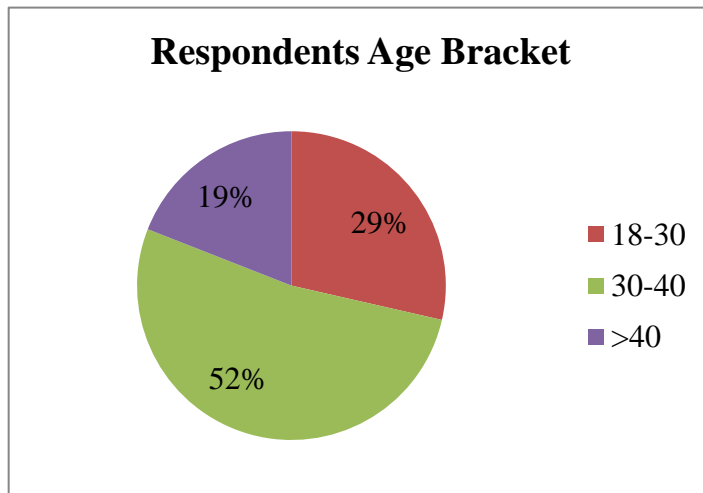


Figure 9: Age of the Respondents

Respondents Age Bracket		
Age Bracket	Frequency	Percentage
18-30	6	29%
30-40	11	52%
>40	4	19%
Total	21	100%

Table 3: Age of the Respondents

4.1.4. Department of work

This research project established the departments the respondents work in. Based on the findings, 19 % of the respondents worked in the organization's operations department (Recovery and Disbursement), 10 % in Research and Strategy department, 29% in ICT department and 10 % in Technical department. 19% of the respondents are management while 14% as the section heads.

From the findings, it's evident that majority of the employees were from the ICT department this can be attributed to the fact that the core implementation of System is mostly affecting the ICT department.

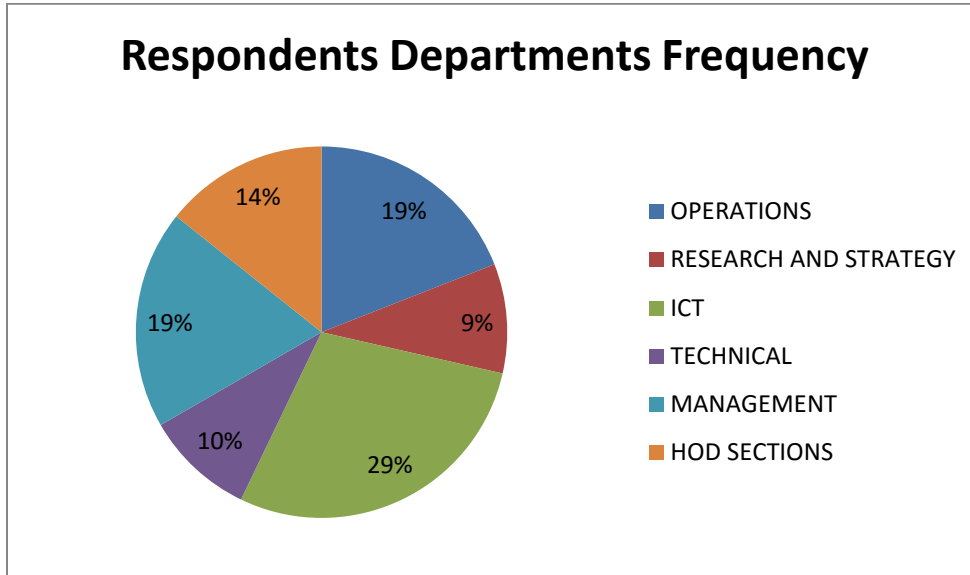


Figure 10: Respondent's Department

Respondents Departments		
Department/Committee	Frequency	Percentage
OPERATIONS	4	19%
RESEARCH AND STRATEGY	2	10%
ICT	6	29%
TECHNICAL	2	10%
MANAGEMENT	4	19%
HOD SECTIONS	3	14%

Table 4: Respondent's Department

4.1.5. Managerial Position

Decision-making was a critical role for middle and senior management. Management defined the need to improve the process of analyzing and reporting on information relevant to the Board's decision to fund students pursuing higher education. Previous process resulted to incomplete or inaccurate analysis, limiting timely policy decision. Middle and senior managers constituted 71% against staff offering administrative and support work. Therefore, the need to develop a system capable of performing the ETL processes for the majority.

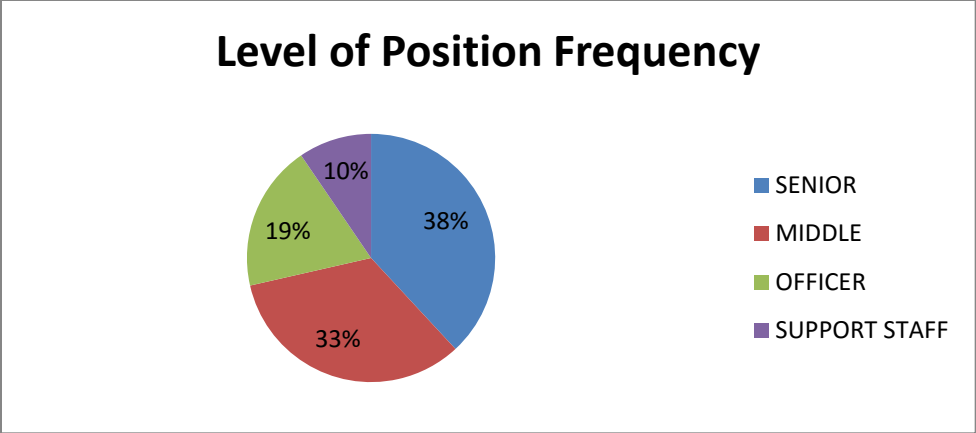


Figure 11: Level of Position

Level of Position		
POSITION	Frequency	Percentage
SENIOR	8	34%
MIDDLE	7	38%
OFFICER	4	19%
SUPPORT STAFF	2	9%
Total	21	100%

Table 5: Level of Position

4.1.5. Reports Generation

The number of staff generating reports for decision-making constituted 71%. This informed the research on the need to offer user-friendly and easy to use system to staff that were more or less non-technical but business experts.

No. of Staff generating reports		
Response	Frequency	Percentage
Yes	15	71.4%
No	6	28.6%
Total	21	100.0%

Table 6: No. of Staff generating reports

4.2 System Architecture and Implementation

4.2.1. System Design

The system that has both a front-end and back-end applications was built. The front-end application was a web-based system built using the PHP language and Codeigniter Framework. Dreamweaver CS5 IDE was used in the development of the application. The system's database on the other hand, was designed using open source SQL Server Database management system. The system database was implemented using Toad data modeler. This database design tool was preferred because of its ability to allow users to visually, create, maintain and documents for new and existing database systems.

4.2.2. System Architecture

The system was implemented using the n-tier application architecture, in which the system was composed of independent components that worked in multiple 'tiers' or layers. Writing of multi-tier applications is a common practice in writing independent components that may be stored and run in different machines (“ Bradley & Millspaugh,” n.d.) they contend that the three-tier application model (shown in figure 12), is the most widely used multi-tier approach. The tiers of the model are: Presentation, Business, and the Data tier.

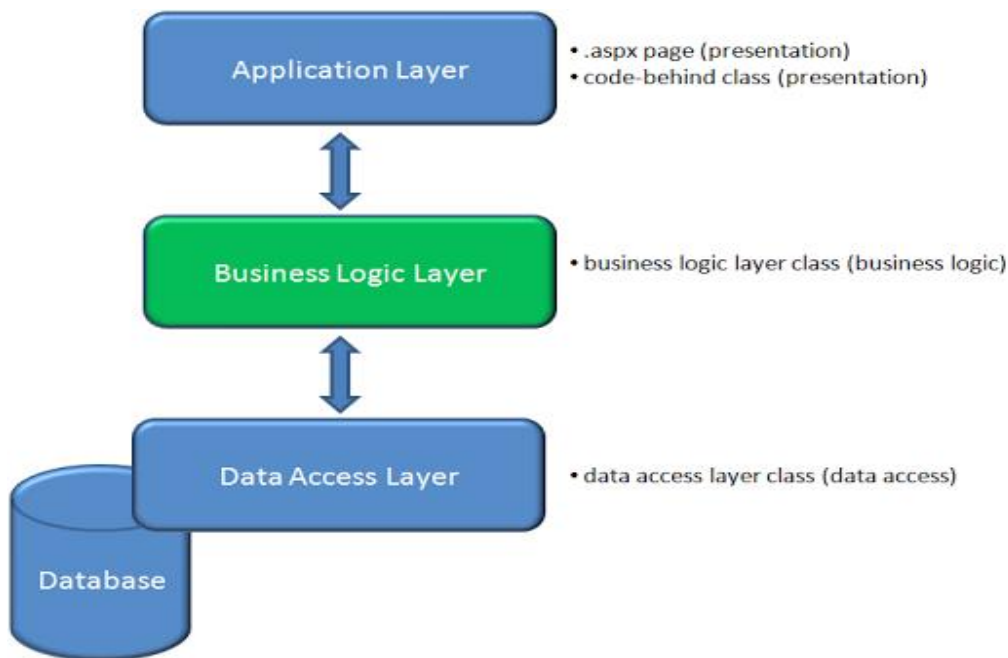


Figure 12: The n-tier Architecture for System Development (Bradley & Millspaugh, 2009)

The presentation tier also referred to as the client layer is a composition dedicated components that present data to the users (the user Interface). The user interface for the reusability assessment system consists of windows application forms that comprise of graphical icons. The encapsulation of the business logic/rules of the application occurs on the business tier. Data manipulation and transformation is done through the use of the business rules. The other key responsibility is processing of retrieved data and sending it to the presentation layer. Finally, the data layer comprises of the database components.

4.2.3. Requirements Analysis

This section describes the different users of the system and their roles, and subsequently presents system requirements—(both functional and non-functional), based on user needs and roles.

4.2.3.1. System Users and Their Roles

- a) **System Administrator:** This User has administrative rights of the System. The roles of the administrator include:
 - Creating User accounts; includes setting system privileges to users.
 - Managing user accounts (editing and deleting user accounts).
- b) **Database Administrator:** This User is involved in the mining of data using the ETL tool to the target data warehouse. The specific functions of the system manager include:
 - Data extraction from the various databases and files
 - Transforming extracted data from divergent sources in preparation for loading.; including data cleansing and
 - Loading data to the target data warehouse.
- c) **Management:** This User is in management and is interested in decision-making activity by facilitation for the data warehouse reporting services. The specific functions of the manager include:
 - Reporting activities
 - Decision making activities based on mined data residing in data warehouse
 - Strategic functions.

4.2.3.2. System Requirements

The System functional and non-functional requirements are summarized in tables 7 and 8 respectively.

a. Functional requirements

ID	Requirement
FR-1	Ability to extract data from heterogeneous sources.
FR-2	Data cleansing ability for detecting and correcting (or removing) corrupt or inaccurate records from mined table.
FR-3	Map entities using entity modelling.
FR-4	Load data into the target data warehouse schema for data warehousing activities
FR-5	Decision support capabilities by the data warehouse
FR-6	Reporting functionality of the system

Table 7: Functional requirements of the Data Warehouse Prototype

b. Non-functional requirements

ID	Requirement
NFR-1	The system should guard against accidental deletion and erroneous update of stored data.
NFR-2	The system should provide for user authentication.
NFR-3	The system should check and verify that entered data is in the appropriate format
NFR-4	The system should have adequate understandability, testability, maintainability, and reusability.

Table 8: Non- Functional requirements of the Data Warehouse Prototype

4.2.4. Use Cases for HELB Data warehouse System

The system will have a system administrator, who will have the overall administrative rights of the system. The roles of the three system users are depicted in the system-level use-case diagram shown in figure 13.

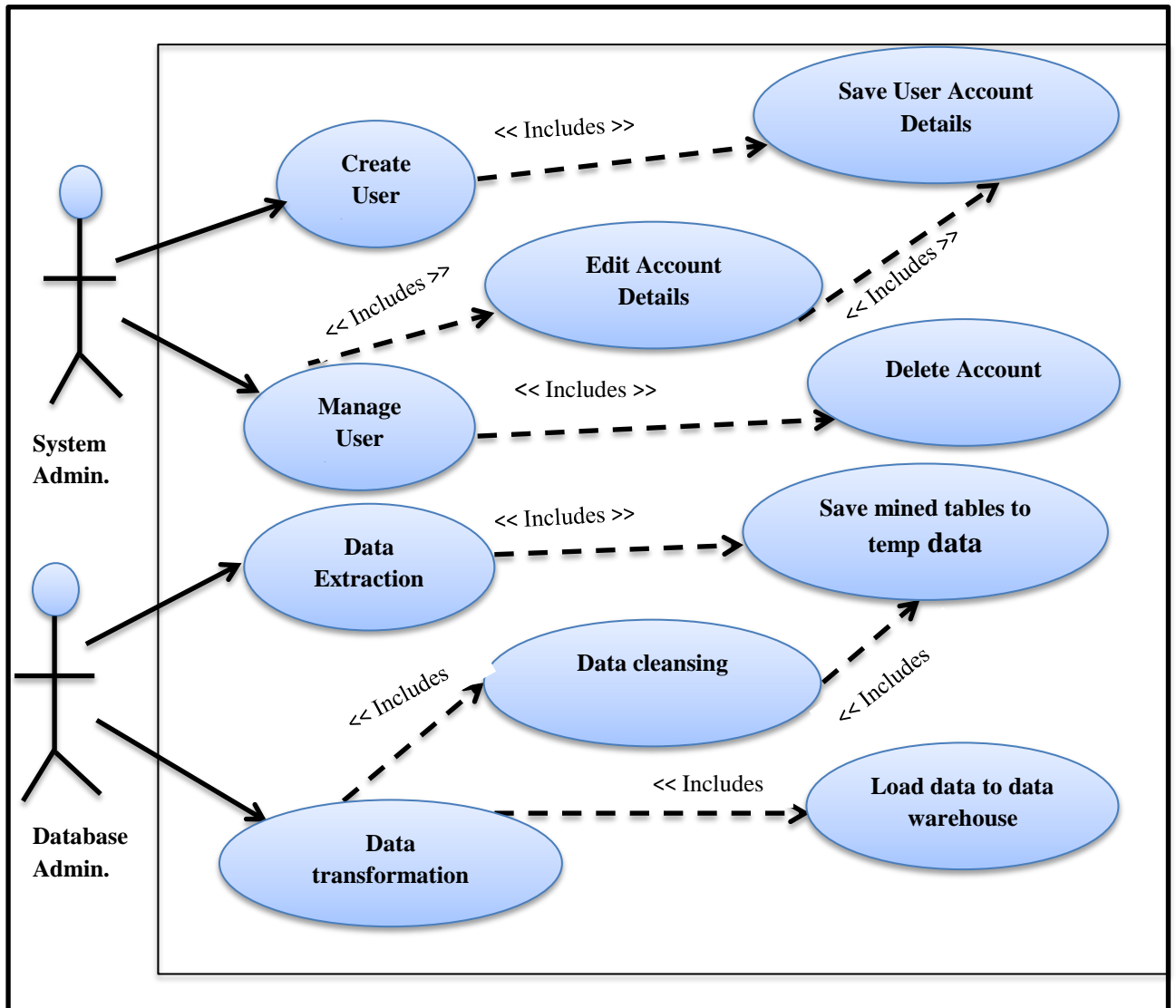


Figure 13: System level use-case diagram for the ETL tool

4.2.5. Database Design

A database for storing students' information as well as users' accounts was built using MYSQL—which is a relational database management system. The identified entities and attributes for the database are shown in figure 14.

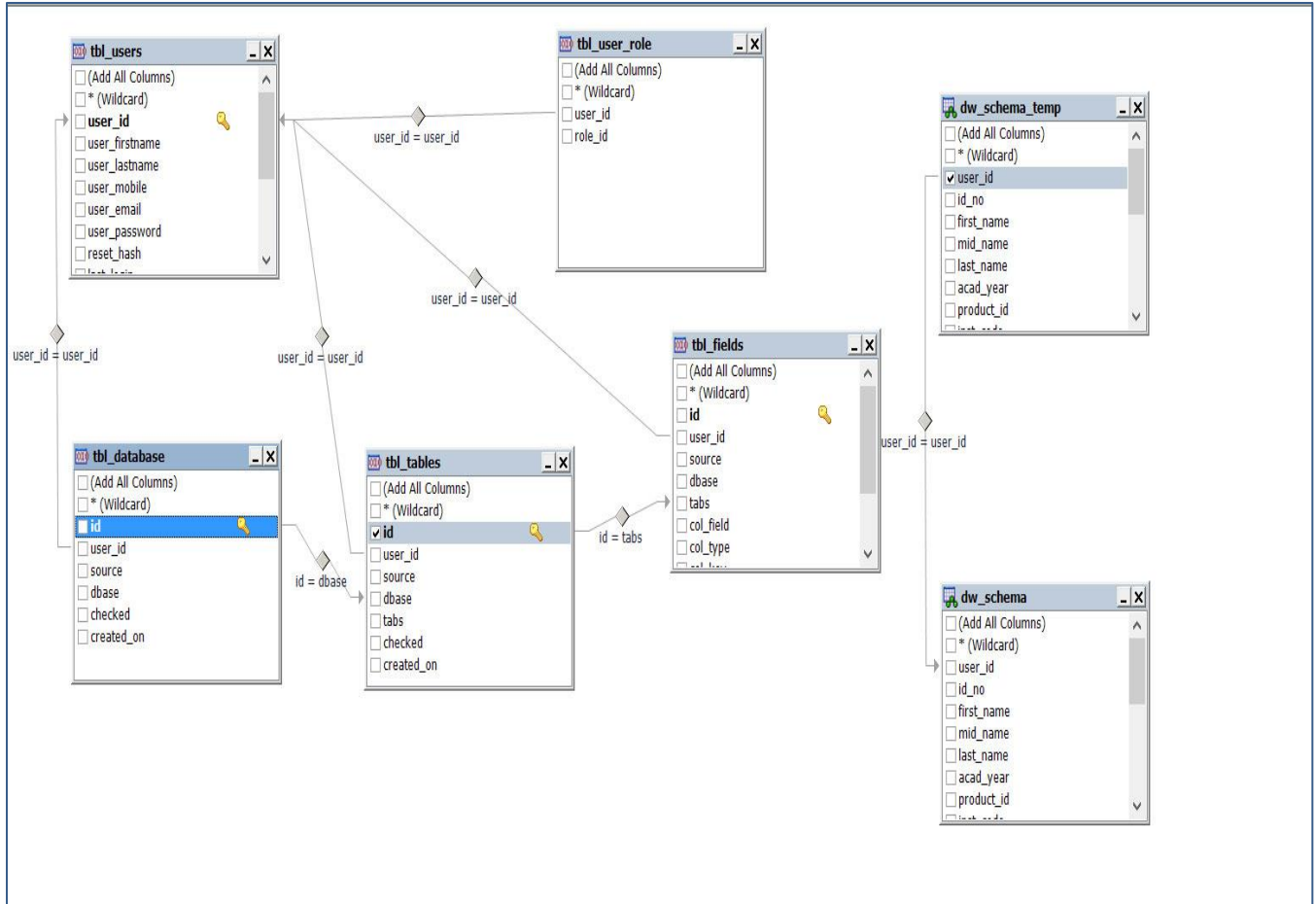


Figure 14: Database design for the ETL tool

4.2.6. Class Design

4.2.6.1. The Data Tier Class Design

The data layer for the application is comprised of two public classes, i.e. Database, Auth_model, Auth_lib, and form_validation. The inheritance hierarchy for the data layer classes is shown in figure 15.

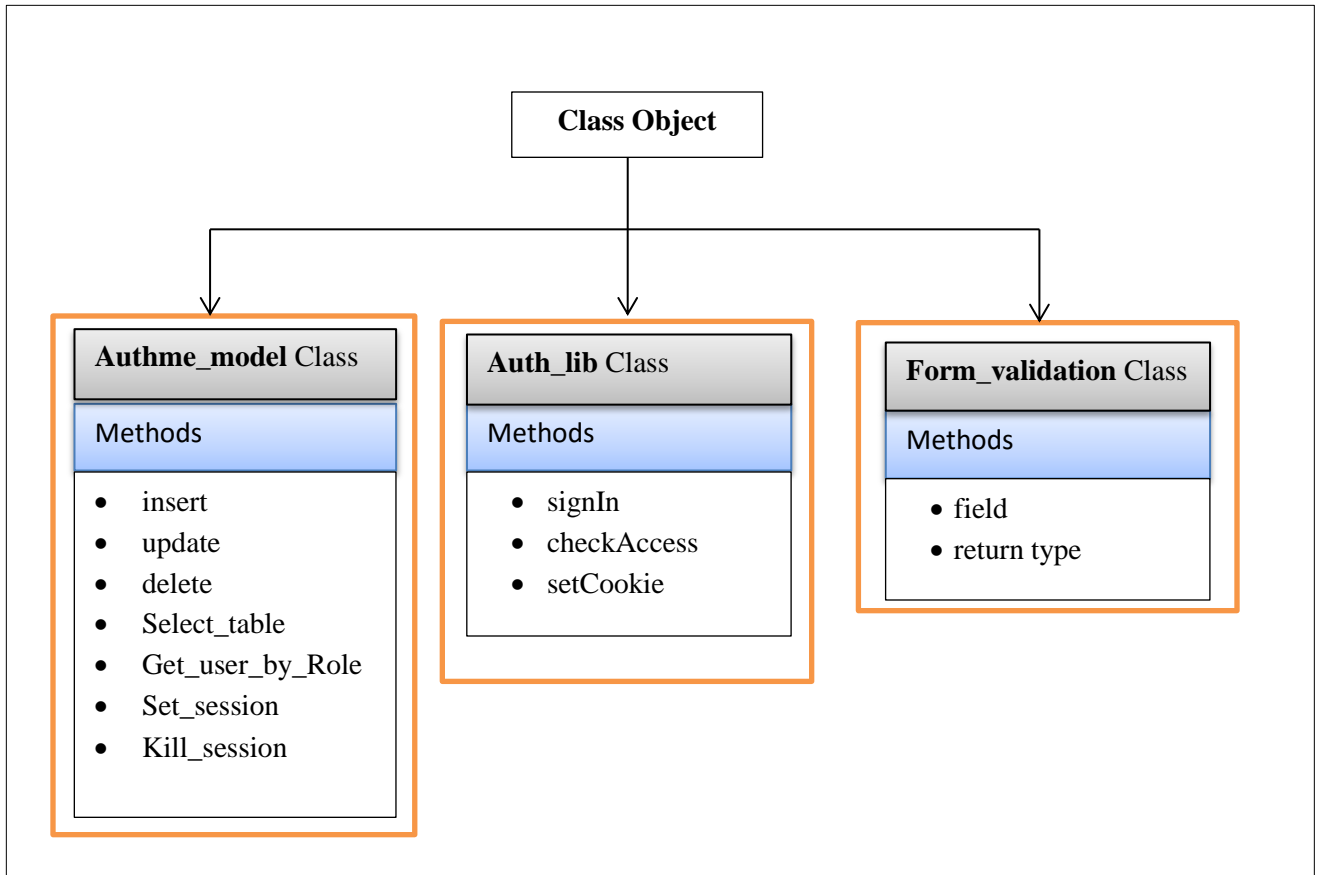


Figure 15: Data layer classes for the ETL tool.

- i. **Public Class Authme_model:** this class performs database connection and Database objects creation i.e. tables and views.
- ii. **Public Class form_validation:** comprises of methods that ensures that all user input is provided as required. That is, it validates if all required fields are provided.
- iii. **Public Class auth_lib:** comprises of methods that performs password encryption and decryption in addition to session's management.

4.2.6.2. The Business Tier Class Design

The Business layer (Layer 2) for the application encapsulates business logic for data manipulation and transformation of the data into information. Processing the data retrieved from the database is also performed at this layer where data is sent to the presentation layer. In addition, this Layer for the system has one class, namely dss, which inherits from class, Authme_model. The business layer class and its members are shown in the figure below.

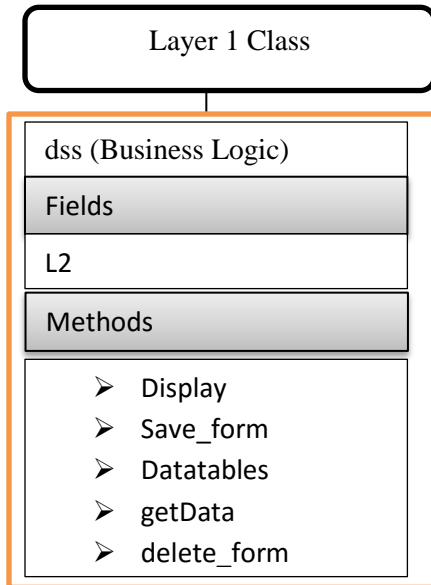


Figure 16: The business tier classes

4.2.6.3. The presentation Tier Class Design

The role of this layer is to describe the components responsible for presenting data to the users- the user interface are described in this layer. It also presents user input to the business layer. The major presentation tier (Layer 3) classes for system are:

- i. Class frmUser: This Layer 3 component handles the creation and management of user accounts. It includes methods for displaying user account information that exists in the system database, as well as methods that enable the user to create new user accounts.
- ii. Class frm_database: This Class consists of methods that display existing database from the various heterogeneous sources. A user selects specific databases to use.
- iii. Class frm_tables: This class includes methods that list all tables existing in selected database from various sources specified by the DBA.
- iv. Class frm_staging: This class includes methods that enable the user to supply temporary tables after mapping the various entities from the specified schemas.
- v. Class frm_extraction: This class includes methods that extract, transforms and loads data to the target warehouse.
- vi. Class frm_datawarehouse: This class includes methods that enable the user perform search and reporting capabilities based on data loaded to the data warehouse.

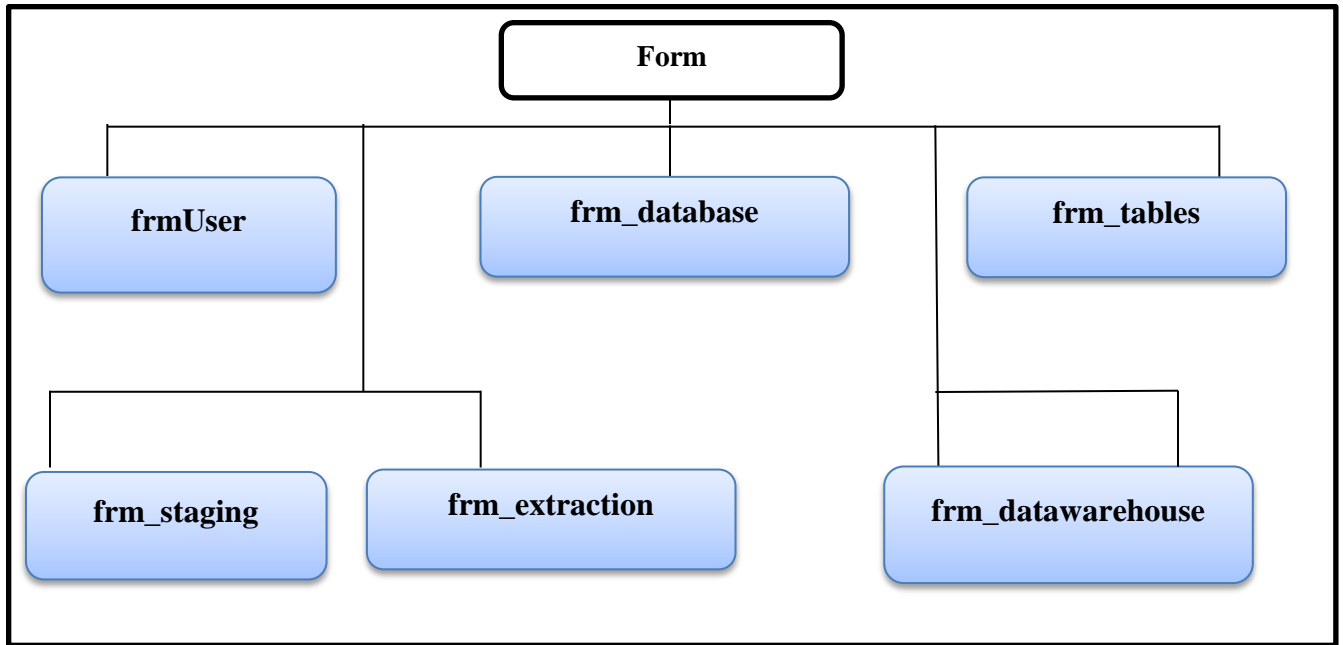


Figure 17: Presentation layer classes for ETL tool

4.3. ETL Prototype

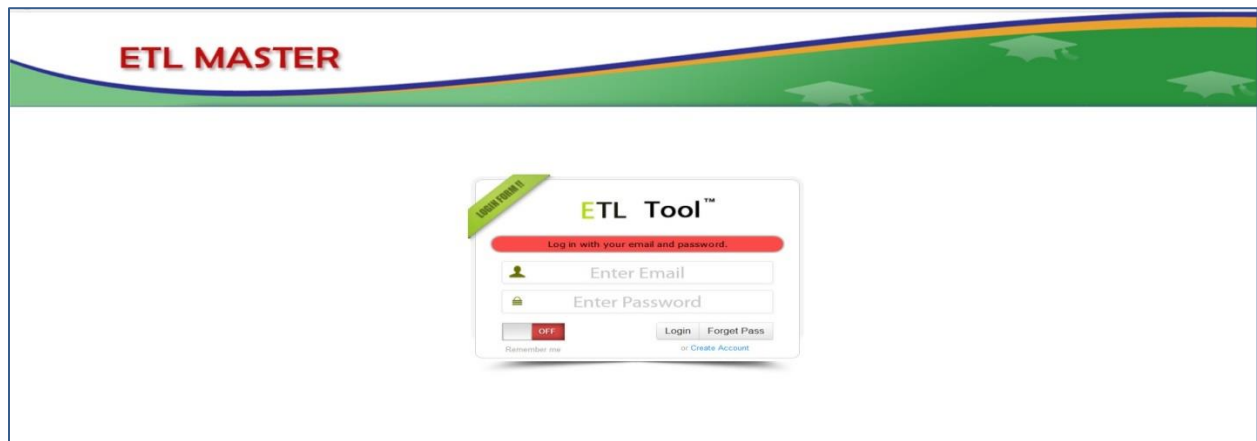


Figure 18: The Login Interface for the ETL tool

4.3.1. The System's Main Interface

After a user had successfully logged into the system, the main system user interface was displayed. The interface, displayed the major tasks that the user can perform. The choice of a particular task displayed the relevant corresponding sub-interface. The main user interface for system is displayed in below.

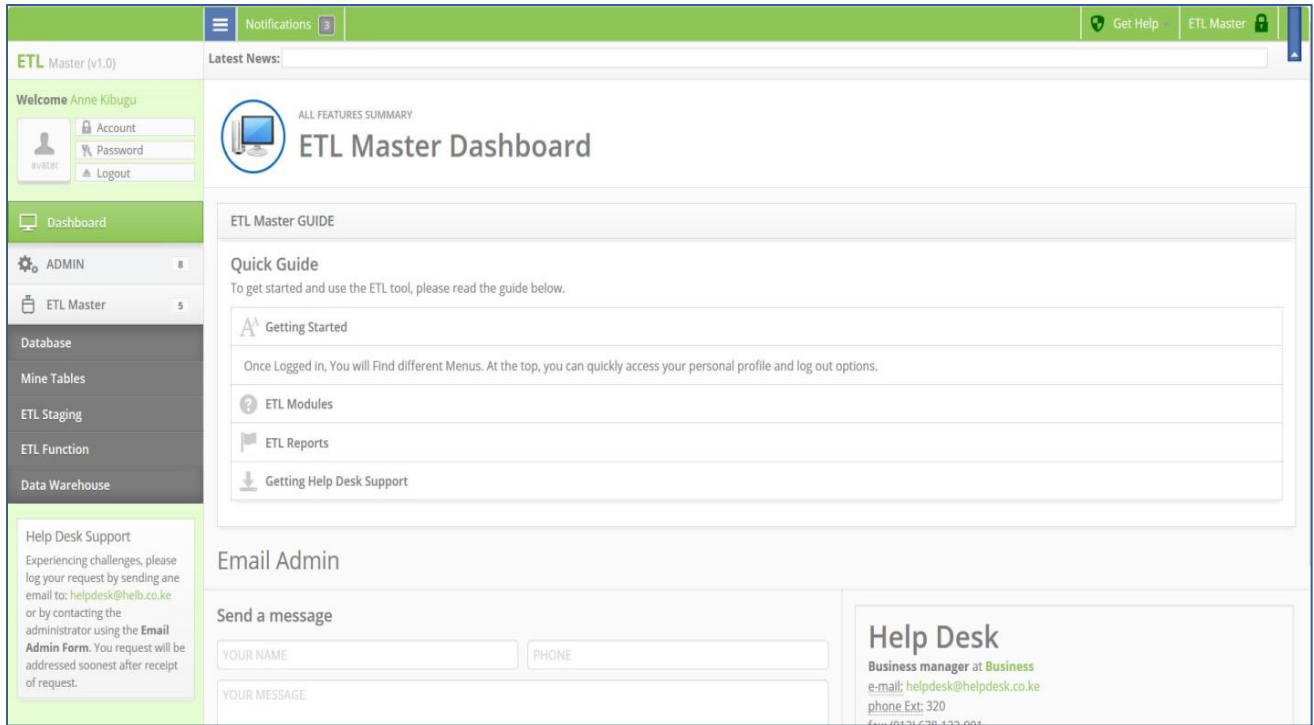


Figure 19: Main User interface for the ETL Tool

4.3.2. Interface for Managing Databases

When the user who is logged in as an administrator he/she chooses various options from the menu ‘ETL Maser’ (shown in figure 20). From this interface, the user can select database schema(s) from different relational databases required for creation of the data warehouse.

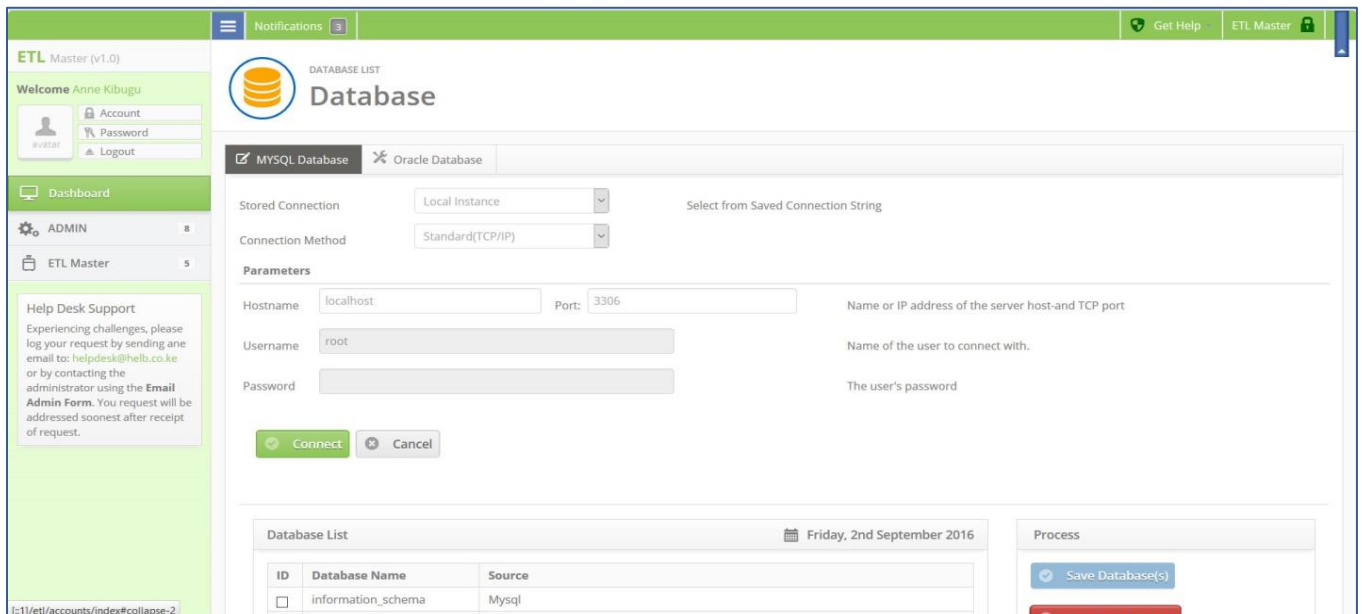


Figure 20: The ETL tool interface for managing databases

4.3.3. Interface for managing tables extraction

The major tasks that can be performed from this interface include; selection of the various tables needed for creation of the target data warehouse. Figure 21, shows the screenshot of the said interface.

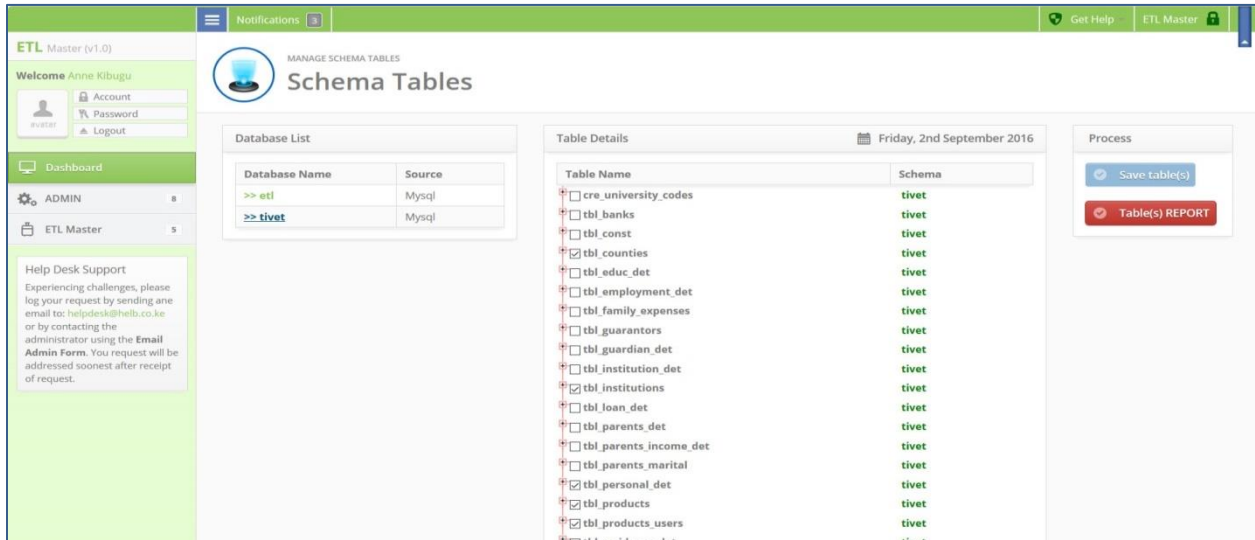


Figure 21: ETL tool interface for managing schema tables

4.3.4. Interface for ETL Function

The major tasks that can be performed from this interface include ETL processes to the data warehouse. Figure 22, demonstrates the screenshot of the said interface.

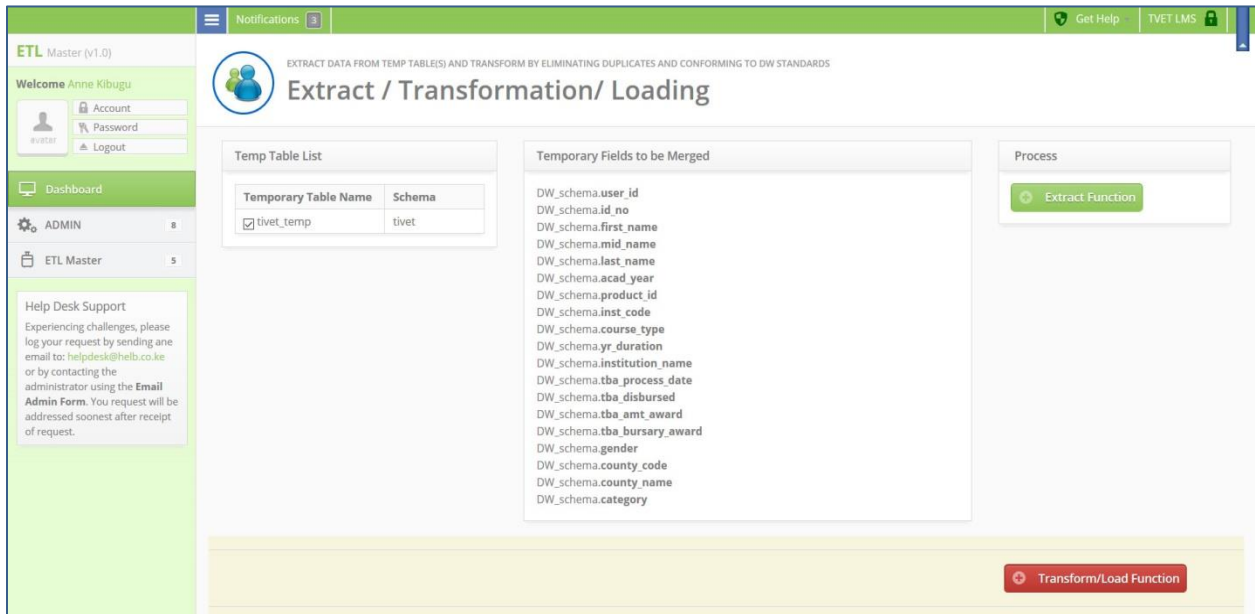


Figure 22: ETL tool interface for ETL Function

4.3.5. Interface for Target Data warehouse

The major tasks that can be performed from this interface include; searching records based on search parameters and reporting capabilities as required. Figure 23, shows the screenshot of the said interface.

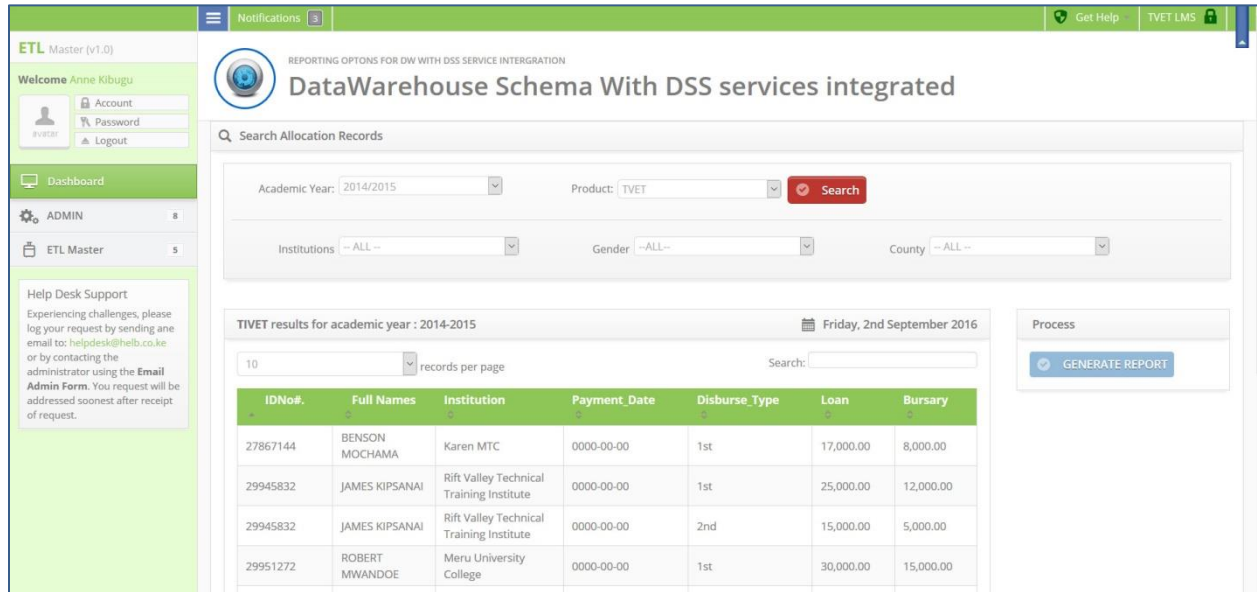


Figure 23: The ETL tool interface for the target data warehouse

4.4. Sample Reports from the data warehouse

Reporting capability is shown in fig. 24 below.

HIGHER EDUCATION LOANS BOARD		
TIVET Loan Allocation Awards	2014/2015	Date: 05/09/2016
	TIVET LOAN Allocation Summary	
Institution Name	Successful	Total Awarded
Karen MTC	1	17,000.00
Meru University College	1	30,000.00
Michuki Technical Training Institute	1	30,000.00
Murang'a College of Technology	1	30,000.00
Nairobi MTC	1	20,000.00
Nyeri MTC	1	17,000.00
Ramogi Institute of Advanced Technology	1	30,000.00
Rift Valley Technical Training Institute	2	40,000.00
No. Of students Awarded	9	Total Amount Awarded 214,000.00
Prepared by:	Date:

Figure 24: TVET Allocation per institution Data Warehouse Reports

HIGHER EDUCATION LOANS BOARD					
TIVET Loan Awards		2014/2015		Date: 20/09/2016	
TIVET LOAN Allocation Summary					
Institution Name: Rift Valley Technical Training Institute					
#	IDNumber	Applicant Name	Disburse_Year	Gender	Total Loan Awarded
1	29945832	JAMES CHERUIYOT KIPSANAI	1st	MALE	25,000.00
2	29945832	JAMES CHERUIYOT KIPSANAI	2nd	MALE	15,000.00
No. Of students Awarded:		2	Total Amount Awarded:		40,000.00

Figure 25: TVET Allocation per individual Data Warehouse Reports

HIGHER EDUCATION LOANS BOARD		
UNDERGRADUATE Loan Allocation Awards		Date: 20/09/2016
UNDERGRADUATE LOAN Allocation Summary		
Institution Name	Successful	Total Awarded
EGERTON UNIVERSITY	1	30,000.00
JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNO	1	35,000.00
KISII UNIVERSITY	1	30,000.00
MOI UNIVERSITY	1	35,000.00
NAROK UNIVERSITY COLLEGE	1	55,000.00
UNIVERSITY OF NAIROBI	2	70,000.00
No. Of students Awarded	7	Total Amount Awarded 255,000.00
Prepared by:		Date:
Verified by:		Date:
Manager / Assistant Manager Disbursement/Allocation		

Figure 26: Undergraduate Allocation per institution Data Warehouse Reports

HIGHER EDUCATION LOANS BOARD					
UNDERGRADUATE Loan Awards		2014/2015		Date: 20/09/2016	
UNDERGRADUATE LOAN Allocation Summary					
Institution Name: UNIVERSITY OF NAIROBI					
#	IDNumber	Applicant Name	Disburse_Year	Gender	Total Loan Awarded
1	30678966	FELICITY NDUKU NYAMBURA	1st	FEMALE	40,000.00
2	32230259	SIMON MAINA GITOBU	1st	MALE	30,000.00
No. Of students Awarded:		2	Total Amount Awarded:		70,000.00

Figure 27: Undergraduate Allocation per individual Data Warehouse Reports

4.5. Performance Evaluation

4.5.1. Evaluation of Proposed Model

The matrix below compared the various approaches for modelling ETL processes and evaluated the proposed model against other models. *P* demonstrates that the model supported the matching parameters, partially.

Comparison and evaluation matrix				
Measure	Models			
	UML environment	Conceptual constructs	Mapping expressions	EMM
Design aspects				
Complete graphical model	x	✓	x	✓
New constructs	x	✓	x	✓
Object Oriented concept independent	✓	P	✓	✓
DBMS independent	✓	✓	✓	✓
Mapping operations	✓	✓	✓	✓
User defined transformation	x	x	x	✓
Mapping relationship	✓	✓	✓	✓
Source independent	x	x	x	x
Source converting	x	x	x	✓
Flat model	✓	✓	✓	✓
Implementation aspects				
Generate mapping document	x	x	x	✓
Non-relational handling	x	x	x	x
Generate SQL	x	x	✓	✓
Develop a tool		✓	✓	✓
Evaluation	4	7.5	7	13
✓ = 1; X=0; P: partial=0.5; Grand Total=13.				
Key: ✓ =YES; X= NO				

Table 9: Models Evaluation and Comparison

4.5.2. System Evaluation

4.5.2.1. Usability

The project research further reviewed the performance of the ETL tool, as to whether it was user friendly and how receptive it was to the HELB Users. It sought to establish the user satisfaction response in relation to GUI, Navigation and Visibility of the system. The tables below depict some of the feedback from users as per the usability parameters.

a) Graphical User Interface Acceptance

Respondents on GUI Acceptance		
Response	Frequency	Percentage
User Friendly	15	71%
Not user Friendly	4	19%
Don't know	2	10%

Table 10: Respondents on System GUI Acceptance

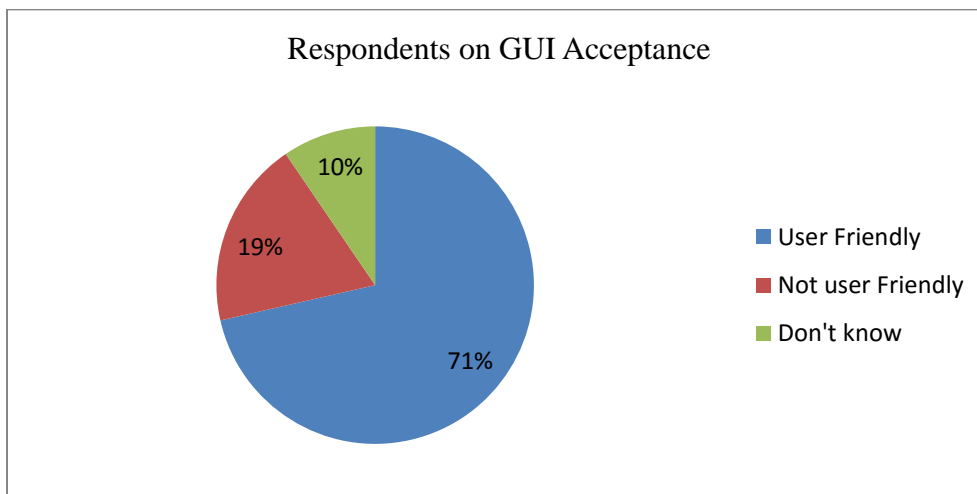


Figure 28: Respondents on System GUI Acceptance

b) Ease to Navigate

Ease to Navigate		
Response	Frequency	Percentage
Easy	16	76%
Complex	2	10%
Not sure	3	14%

Table 11: Respondents on ease to Navigate

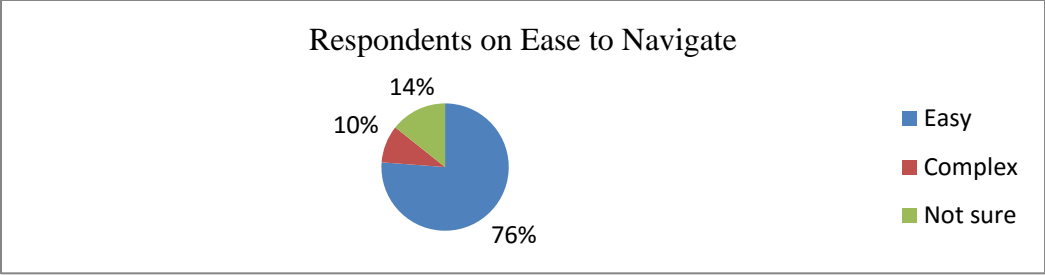


Figure 29: Respondents on ease to Navigate

c) Visibility

Ease to Navigate		
Response	Frequency	Percentage
Visible	11	53%
Fairly Visible	6	29%
Not Visible	4	19%

Table 12: Respondents on application Visibility

Figure 4.10: Respondents on System Visibility

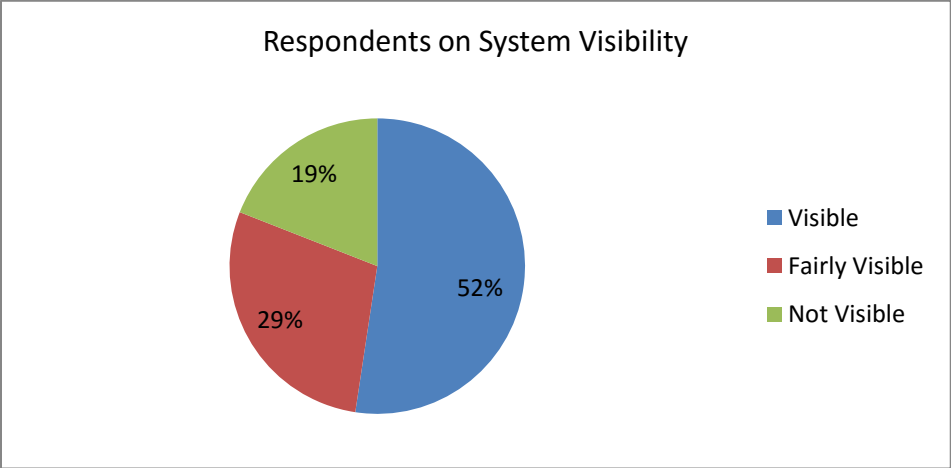


Figure 30: Respondents on System Visibility

4.5.2.2. System Usability Discussion

From the above usability parameters administered to twenty one (21) system user it can be deduced that from the GUI 71% of the users find the overall system to have a good user interface that is easy to interact, 19% of the system users are of the opinion that the overall system does not have a good interface to interact while 10% of the users have don't know.

From this assessment, the research can easily conclude that majority of the users are comfortable using the system without ignoring the remarks raised by the 19% system users' i.e. color scheme of the interface interactive screen and some highlighted that the system is congested.

76% of the users find the system easy to navigate from on point to another while 10% of them find challenges navigating from one point to another, 14 % aren't sure as they still require more time to analyze the system.

On the visibility of the system especially for the ETL tool 52% of the users have no problem with font size used, 29% are partially comfortable with the screen visibility while 19% have issues with screen visibility. 29% and 19% of the system users is as a result of aging workforce of the employees.

CHAPTER FIVE

5.0 CONCLUSION AND RECOMMENDATIONS

5.1. Introduction

The implementation of a data warehouse in HELB, through an ETL process model was the main research objective. The achievement was based on the mining of data from heterogeneous data sources, for improved timely decision-making and increased efficiency in service delivery.

The research project sort to address the main challenge of data existing in silos, and lack of integrated historical and current data across HELB's spectrum to help the business access accurate and consistent data to maintain competitive edge. The overriding goal of the research was to develop an ETL model with the capability for decision making. The key objective of the research was to develop a System prototype that demonstrated the capabilities and functionalities of the ETL by mapping inbound data and loaded to the traget Data warehouse for HELB.

The literature review investigated existing methodologies employed by other researchers. The evaluation of the studies and related surveys aimed at identifying the characteristics and key elements of an effective ETL process model on the implementation of a data warehouse. The information played a crucial role, identifying the limitations of existing methodologies and frameworks, found in the literature and hence led to the development of the Entity Mapping Methodology, used in this research. Review of existing methodologies from the studies and existing tools and technologies, provided in-depth understanding of functional and non-functional requirements for a simple and effective ETL process model.

In Chapter one, the research provided background information of exiting divergent data sources in HELB. This research evaluated the challenges posed by heterogeneous sources of information and introduced the concept of data warehousing. The HELB staff was involved during data collection and System performance evaluation. Subsequent chapters of the project described the same.

5.2. Achievements of the Research

Generally, the study was aimed at investigating existing methodologies by addressing issues related to ETL process models, in relation to HELB scenario.

The first objective of the project was to identify and examine the strategies and methods ICT technical experts use to enable non-technical Users extract data from different sources in an accurate but simple way.

Through the evaluation of documented studies and existing tools, two major models were identified and assessed. They included; the ‘intermodel assertions’, mapping and conceptual UML-based metamodel in relation to the backstage and the front-end data warehouse architecture. Three major international database vendors that provided ETL solutions were evaluated: Oracle, Microsoft and IBM.

The second objective focused on identifying the shortcomings and gaps of the ETL process strategies/methods. The objective was achieved by comparing methods that exist in industry with what literature describes as ideal techniques. It was established that the models identified were highly subjective, complex and too technical for end users; hence unreliable in extracting data for timely and improved decision-making.

The third specific objective was to develop a prototype based on the identified mapping techniques capable of performing the ETL functionalities and generate reports for decision support. The research achieved this objective by executing two overall tasks; system design and development.

A front-end and a back-end application were built. The front-end application was a web-based system built using the PHP language and Codeigniter Framework. Dreamweaver CS5 IDE was used in the development of the application. The system's database on the other hand, was implemented using open source SQL Server Database management system. The system database was developed using Toad data modeler.

The last objective was to test the working of the ETL process model/the prototype and show that it was ideal to the existing methods. This objective was achieved through system performance evaluation. Users were asked to fill out questionnaires to demonstrate that an improved method of extracting data from difference sources, for decision-making, was achieved. Lastly, a comparison between the new model and the ones existing in literature and in industry was conducted for validating the efficiency of the new model.

5.3.Impact of the Research

This research makes important contributions to the academia as well as corporate computing industry- public and private. The research establishes the relationship between software development and real problem solutions such as corporate decision-making. It also provides means for defining and experimentally validating model in a precise and formal manner.

The research allows comparison, selection and modification of an ETL process models that can provide a theoretical support for various software tools.

To HELB business, this research is of value addition and significant since the daily operational environment and the data warehouse are separated. Historical data is a snapshot data, integrated and subject oriented. The research further present a system in an environment characterized as read-only, therefore not resource intensive and very large data sets can be analyzed timelessly.

5.4.Limitations of the Research

Data warehousing is generally a relatively new field in state organs in Kenya and much research needs to be done. Few research studies and opportunities existed locally, relevant to this study and that could further present practical examples and scenarios. The need to bench mark with related agencies was hindered, lacking a commonly agreed procedures and standards to perform the ETL processes.

The extraction of data was a major problem, mainly because of the nature and confidentiality of students' bio-data. Limited data was used to develop the prototype (three financial years).

Further, optimization to improve system performance was not achieved and resumption problems were experienced. Data warehousing testing and 100% data verification was not achieved during this research since data quality from sources was not assured.

5.5.Recommendations

- The initial cost of implementing a data warehouse based on any existing commercial ETL methods are high and the process requires commitment at all levels. However, the benefits of a data warehouse override the cost and return on investment can be realized if the opportunities are reviewed. The implementation of a data warehouse is a worthwhile investment project that most organization handling voluminous data cannot overlook.
- There is need for completeness and freshness of data and a near real time ETL process is recommended.
- Based on the changing environment in distributed computing, databases and other related technological user needs, further recommendation is made for the algorithmic and theoretical results in data warehousing and underlying components and processes such as the ETL.

- To realize the full benefit of a data warehouse implementation, through the Entity Mapping Methodology, HELB should continually encourage and support staff in embracing data warehousing as a way of improving decision-making, using divergent historical and current data.

The following recommendations regarding future research and practice are made:

- Automation of some steps of the methodology; further validation is needed, particularly redundancy checks and thorough implementation evaluation.
- Further research is required in order to put forward a standard integrated framework that state organizations can adopt in achieving successful ETL processes to enhance data accuracy, consistency and transparency.
- Researchers should study non-technical factors that impede the successful implementation of data warehouse in state corporations in the country and give recommendations on the best ways of overcome them.

REFERENCE

- Alenazi, S.R.A. et al., 2014. Prototyping Data Warehouse System for Ministry of Higher Education in Saudi Arabia. , 7(4), pp.74–81.
- Anon, 2015. ARCHITECTURE FOR REAL-TIME ANALYTICAL DATA INTEGRATION AND DATA. , (July).
- Ferreira, N. & Furtado, P., Near Real-Time with Traditional Data Warehouse Architectures : Factors and How-to.
- Mawilmada, P.K., 2011. IMPACT OF A DATA WAREHOUSE MODEL FOR IMPROVED DECISION - MAKING. , (October).
- Oketunji, T. & Omodara, O., 2011. Design of Data Warehouse and Business Intelligence System. , (June).
- Parmar, V., Yadav, R. & Sharma, M., 2016. REVIEW ARTICLE A DATA CLEANING MODEL FOR DATAWARE HOUSE. , 4(1).
- Quality, D. & Line, B., Data quality.
- Russom, P., TDWI CHECKLIST REPORT Data Integration for Real-Time Data Warehousing and Data Virtualization TDWI CHECKLIST REPORT Data Integration for Real-Time Data Warehousing and Data Virtualization.
- Senapati, R. & Kumar, D.A., 2014. A Survey on Data Warehouse Architecture. , pp.5235–5239.
- Sharma, N. & Gupta, S.K., 2012. DESIGN AND IMPLEMENTATION OF ACCESS THE CONTENTS IN THE. , 6(1), pp.61–64.
- Simitsis, A. & Vassiliadis, P., A Methodology for the Conceptual Modeling of ETL Processes.
- Bernstein, P., Rahm, E., 2000. Data warehouse scenarios for model management. In: Proceedings of the 19th International Conference on Conceptual Modeling (ER'00), LNCS, vol. 1920, Salt Lake City, USA, pp. 1–15.
- Berson and Smith, 1997A. Berson, S.J. Smith Data Warehousing, Data Mining, and OLAP.
- Dobre, A., Hakimpour, F., Dittrich, K.R., 2003. Operators and classification for data mapping in semantic integration. In: Proceedings of the 22nd International Conference on Conceptual Modeling (ER'03), LNCS, vol. 2813, Chicago, USA, pp. 534–547.
- El Bastawesy, A., Boshra, M., Hendawi, A., 2005. Entity mapping diagram for modeling ETL processes. In: Proceedings of the Third International Conference on Informatics and Systems (INFOS), Cairo.
- Muneer Alsurori, Juhana Salim,” Information and Communication Technology for Decision-Making in the Higher Education in Yemen: A Review” 2009 International Conference on Electrical Engineering and Informatics, 5-7 August 2009, Selangor, Malaysia.
- Wang Aihua, Guo Wenge, Xu Guoxiong, Jia Jiyou, Wen Dongmao, 2009. “GIS-Based Educational

- Decision- Making System” Proceedings of 2009 IEEE International Conference on Grey Systems and Intelligent Services, November 10-12, 2009, Nanjing, China., 2009 IEEE, pp 1198-1202.
- Qiusheng Liu, Guofang Liu,” Research on the Framework of Decision Support System Based on ERP Systems”, 2010 Second International Workshop on Education Technology and Computer Science, 2010 IEEE.
- Manjunath T.N, Ravindra S Hegadi, Ravikumar G K."Analysis of Data Quality Aspects in DataWarehouse Systems", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (1) , 2010, 477-485.
- W. H. Inmon. 2012. “*Building the Data Warehouse.*” John Wiley & Sons, Third edition.
- Rajan Vohra1 & Nripendra Narayan Das Intelligent decision support systems for admission management In higher education institutes- International Journal of Artificial Intelligence & Applications (IJAA), Vol.2, No.4, October 2011.
- Glorio, O., Mazon, J. N., Garrigoz, I., & Trujillo, J. (2010). Using Web-based Personalization on Spatial Data Warehouses (pp. 1-8). EDBT-ACM.
- Hamdan, A. (2005). Women and Education in Saudi Arabia: Challenges and Achievements. International Education Journal, 6(1), 42-64.
- Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques (2nd ed.). Morgan Kaufmann Publishers.
- Inmon, W. H. (1992). Building the Data Bridge: The Ten Critical Success Factors of Building a Data Warehouse. Database Programming & Design.
- Inmon W. H. (1996), "The data warehouse and data mining", Communications of the ACM, 39(11), 49 – 50.
- Mohammed, M. A., Hasson, A. R., Shawkat, A. R., & Al-Khafaji, N. J. (2012). *E-Government Architecture Uses Data Warehouse Techniques to Increase Information Sharing in Iraqi Universities* (pp. 1-5). IEEE Explore.
- Wang, F. (2009). Application Research of Data Warehouse and Its Model Design (pp. 798-801). IEEE computer society.
- Yebai, L., & Li, Z. (2009). *Interactive Modeling of Data Warehouse on E-business System* (pp. 80-83). IEEE computer society.
- Yen, D.C., Chou, D.C. and Chang, J. (2002), “A synergic analysis for Web-based enterprise resources-planning systems”, Computer Standards & Interfaces, 24(4), 337-46.

APPENDICES

Appendix I: Data Collection Authorization Letter



**UNIVERSITY OF NAIROBI
COLLEGE OF BIOLOGICAL AND PHYSICAL SCIENCES
SCHOOL OF COMPUTING AND INFORMATICS**

Telephone: 4447870/4444919/4446544
Telegrams: "Varsity" Nairobi
Telefax: 254-2-4447870
Email: director-sci@uonbi.ac.ke

P.O. Box 30197
Nairobi
Kenya

Our Ref: UON/CBPS/SCI/MSC/DCT/2014

5 July 2016

TO WHOM IT MAY CONCERN

Dear Sir/Madam

RE: ANNE WANJIKU KIBUGU ; REG. NO. P53/73060/2014

This is to confirm that the above named is a bona fide student of the University of Nairobi, School of Computing and Informatics.

She is pursuing a M.Sc. course in Distributed Computing Technology. She would like to collect data for her project entitled: ***"Data Warehouse Implementation Model for Improved Decision Support at Higher Education Loans Board."*** Under the supervision of Prof. W. Okelo-Odongo.

Any assistance accorded to her will be highly appreciated.

Yours faithfully

A handwritten signature in blue ink, appearing to be 'W'.

**PROF. W. OKELO-ODONGO
DIRECTOR
SCHOOL OF COMPUTING & INFORMATICS**

**School of Computing & Informatics
University of NAIROBI
P. O. Box 30197
NAIROBI**

Appendix II: Research Questionnaire

I am a postgraduate student at University of Nairobi, School of Computing & Informatics undertaking Master degree in Distributed Computing Technology (MSc. DCT). I am carrying out an academic research on: the development of an ETL tool for implementing of a data warehouse in HELB to enhance decision making by generating reports from divergent data sources. This research is purely aimed for academic purpose and the information provided will be treated with utmost confidentiality. Kindly fill this questionnaire to the best of your ability Thank you.

Sample Questionnaires

Key

✓ Please put [X] where appropriate

Section A: Personal Data

1. Gender
 - Male
 - Female
2. Age bracket
 - 18 – 30
 - 30 – 40
 - Over 40
3. Level of Position
 - Senior Management
 - Middle Level Management
 - Officer
 - Support staff
4. Department/Committee:-.....

Section B: General Questionnaires

- a) Are you involved in preparing reports for decision support in HELB?
 - Yes
 - No
- b) How do you get the right data to generate your reports in order to make informed decision?
 - ICT Department personnel
 - External Sources
 - Legacy Systems
 - Hard files and records
 - Others Business Units
- c) How many sources do you often use to get data to generate your reports?
 - 1-3
 - 3-5
 - Others

- d) How do you authenticate the source of data?
 - Against past activities
 - Through request to ICT
 - Verifying origin of data
- e) What amount of data do you require to generate reports?
 - Departmental data
 - Data from various business units
 - Big data
 - Not sure
- f) How fast to you access the data from the sources identified in (a)?
 - Within minutes
 - Within an hour
 - Within 24 hours
 - In a day
 - Within days
- g) What kind of reports do you think can help you to make better decisions?
 - Daily reports
 - Weekly- monthly reports
 - Historical reports
 - From various fact tables
- h) Do you think you always work with the right data to generate the reports?
 - Yes
 - No
 - Not sure
- i) Is there an electronic central repository to access HELB integrated data and generate reports?
 - Yes
 - No
 - Don't Know
- j) Do you believe data extracted from the sources mentioned in (a) is accurate and consistent?
 - Yes
 - No
 - Don't Know
- k) How do you resolve problems encountered during the preparation of the reports?
 - Manually
 - Through ICT Support team
 - Automatically
- l) What is your general view about data extraction, reporting and analysis in HELB?

.....

.....

.....

.....

Appendix III: Performance Evaluation Questionnaires

1) How is the Graphical User Interface of the system?

Parameters	Cross (X) Where appropriate
User Friendly	
Not user Friendly	
Don't know	

2) How is the Navigation of the system from one point to another?

Parameters	Cross (X) Where appropriate
Easy	
Complex	
Not sure	

3) Do you have challenges seeing the content on the system screen?

Parameters	Cross (X) Where appropriate
Visible	
Fairly Visible	
Not Visible	

Appendix IV: Code Snippets

1. Login

```
/**
 * Login page
 */
public function login($message="")
{
    // Redirect to your logged in landing page here
    if(logged_in()) redirect('accounts/index');

    $data['title'] = ETL | LOGIN FORM;
    $data['message'] = 'Log in with your email and password.';
    $data['error'] = false;
    if($message != ""){
        $data['message'] = $message;
    }
    $post = $this->input->post();
    if($post == FALSE){
        $this->display('signIn',$data);
    }
    else{
        $this->form_validation->set_rules('email_add', 'Email', 'required|valid_email');
        $this->form_validation->set_rules('password', 'Password',
'required|alpha_numeric|min_length[6]|max_length[20]');
        $this->form_validation->set_error_delimiters(" ", "");

        if ($this->form_validation->run()===FALSE) {
            $data = array(
                'email_add' => form_error('email_add'),
                'password' => form_error('password')
            );
            echo json_encode($data);
        }
        else{
            if($this->authme->login($this->input->post('email_add'), $this->input-
>post('password'))){
                if($this->session->userdata('status') == 0){
                    $data = array('success' => 'inactive');
                }else if($this->session->userdata('status')== 1){
                    $data = array('success' => 'deactivated');
                }else{
                    $data = array('success' => 'success');
                }
                //$data = array('success' => $this->session->userdata('status'));
                echo json_encode($data);
            }else {
                $data = array(
                    'error' => 'Your email address and/or password is
incorrect.'
                );
            }
        }
    }
}
```

```

        echo json_encode($data);
        // echo '{"error": "Your email address and/or password is
incorrectss.", "nome_erro": "" . form_error('nome') . ""}';
    }
}
} //End of post evaluation
}

/**
 * Forgot password page
 */
public function forgot()
{
    // Redirect to your logged in landing page here
    if(logged_in()) redirect('accounts/index');

    #Validate firds
    $this->form_validation->set_rules('email_reset', 'Email',
'required|valid_email|callback_email_exists');

    if ($this->form_validation->run()===FALSE) {
        $data = array(
            'email_reset' => form_error('email_reset')
        );
        echo json_encode($data);
    }else{
        $this->load->model('Authme_model');
        $email = $this->input->post('email_reset');
        $user = $this->Authme_model->get_user_by_email($email);
        $slug = md5($user['user_id'] . $user['user_email'] . date('Ymd'));
        $path=base_url().'auth/reset/'. $user['user_id'] .'.' . $slug;
        ///////////////////////////////////////////////////////////////////
        date_default_timezone_set('Africa/Nairobi');
        $config['protocol'] = 'smtp';
        $config['smtp_host'] = '192.168.1.1';
        $config['smtp_port'] = '25';
        $config['smtp_timeout'] = '7';
        $config['charset'] = 'iso-8859-1';
        $config['wordwrap'] = TRUE;
        $config['newline'] = "\r\n";
        $config['mailtype'] = 'text'; // or html
        $config['validation'] = TRUE; // bool whether to validate email or not
        $date=date ("d M Y, h:i A ");

        $this->email->initialize($config);
        $this->email->from('emacharia@helb.co.ke', 'Account Registration:'); // Change
these details

        $this->email->to($email);
        $this->email->subject('Reset your Password:');

```

\$message="To reset your password please click the link below and follow the instructions: \n".\$path."\n
If you did not request to reset your password then please just ignore this email and no changes will occur.

Note: This reset code will expire after, ". date('j M Y')."***DO NOT RESPOND TO THIS EMAIL****

Regards,
HELB Technical Team
";

```
$this->email->message($message); //Set email message
$val=$this->email->send(); //Send Email
if($val==true){
    $data = array( 'success' => 'success');
}
else{
    $data = array('success' => 'error');
}
//$data = array( 'success' => $message);
echo json_encode($data);
}
}
```

2. Business Logic

#function to save modules

```
function save_form($page,$action,$id=",$code="){
    switch($page)
    {
        case 'frm_database':
            #Extract form data
            $val="";
            $post = $this->input->post();
            $post['user_id'] = $this->session->userdata('user_id'); //Get category by
            role
```

```
            if($action=='insert'){
                unset($post['cid']);
                $array=count($post['dbase']);
                $myVal = array();

                for($i=0; $i<$array; $i++){
                    $myVal['user_id']=$post['user_id'];
                    $myVal['source']=$post['source'];
                    $myVal['dbase']=$post['dbase'][$i];
                    $myVal['created_on']=date("Y-m-d");
                    $myVal['checked']=1;

                    $rec=$this->accounts_model-
                    >select_table('tbl_database','dbase'," WHERE dbase='".$post['dbase'][$i]."' AND
                    user_id='".$post['user_id'],'row');

                    if(!$rec){
```

```

                                $val = $this->accounts_model-
>_insert($myVal,'tbl_database');
                                }else{
                                    $val=true;
                                }
                            }
                        }

                        /***** output message *****/
                        if($val){
                            echo json_encode(array('info'=>'success'));
                        }
                        else{
                            echo json_encode(array('info'=>'error'));
                        }
                    }
                    break;

                    case 'frm_tables':
                        #Extract form data
                        $val="";
                    $post = $this->input->post();
                    $post['user_id'] = $this->session->userdata('user_id'); //Get category by
                    role

                    if($action=='insert'){
                        unset($post['cid']);unset($post['baseurl']);

                        $array=count($post['tabs']);
                        $myVal = array();

                        for($i=0; $i<$array; $i++){
                            $myVal['user_id']=$post['user_id'];

                            if($post['schema']=='ORCL'){
                                $myVal['source']='Oracle';
                            }else{
                                $myVal['source']=$post['source'];
                            }

                            $myVal['dbase']=$post['schema'];
                            $myVal['tabs']=$post['tabs'][$i];
                            $myVal['created_on']=date("Y-m-d");
                            $myVal['checked']=1;

                            $rec=$this->accounts_model-
>select_table('tbl_tables','tabs'," WHERE tabs="".$post['tabs'][$i]."" AND dbase="".$post['schema'].""
AND user_id="".$post['user_id'],'row');

                            if(!$rec){
                                $val = $this->accounts_model-
>_insert($myVal,'tbl_tables');

```

```

        }else{
            $val=true;
        }
    }

}

}else if($action=='update'){
    /*$array=count($post['dbase']);
    $myVal = array();

    for($i=0; $i<$array; $i++){
        $myVal['source']=$post['source'];
        //$myVal['dbase']=$post['dbase'][$i];
        $myVal['created_on']=date("Y-m-d");
        $myVal['checked']=0;
        $rec=$this->accounts_model-
>select_table('tbl_database','dbase'," WHERE dbase='".$post['dbase'][$i]."' and
user_id='".$post['cid'],'row');

        if($rec){

            $arrays=array('dbase'=>$post['dbase'][$i],'user_id'=>$post['cid']);
            echo json_encode($myVal);
            $val = $this->accounts_model-
>_update($myVal,'tbl_database','user_id',$post['cid']);

            //$val = $this->accounts_model-
>_insert($myVal,'tbl_database');

        }else{
            $val=true;
        }
    }

}*/

}

/***** output message *****/
if($val){
    echo json_encode(array('info'=>'success'));
}
else{
    echo json_encode(array('info'=>'error'));
}
break;
case 'frm_fields':
    #Extract form data
    $val="";
    $post = $this->input->post();
    $post['user_id'] = $this->session->userdata('user_id'); //Get
category by role

```

```

        if($action=='insert'){
            unset($post['cid']);unset($post['baseurl']);

            #Check if table exists
            $rec=$this->accounts_model-
>select_table('tbl_fields','temp_name'," WHERE dbase=".$post['dbase']."' AND
user_id=".$post['user_id'],'row');

            if($rec){
                $val=$this->accounts_model->delete_table(
'tbl_fields','user_id',$post['user_id'] );
            }

            #Split table
            $array=count($post['col_field']);
            $myVal = array();

            for($i=0; $i<$array; $i++){

                $keywords = explode('-', $post['col_field'][$i]);

                $table=trim($keywords[0]);
                $col=trim($keywords[1]);
                if(!empty($keywords[2])) {
                    $type=trim($keywords[2]); }else{ $type=""; }
                $key=trim($keywords[3]); }else{ $key=""; }

                $myVal['user_id']=$post['user_id'];

                if($post['dbase']=='tivet'){
                    $myVal['source']='Mysql';
                }else{
                    $myVal['source']='Oracle';
                }

                $myVal['dbase']=$post['dbase'];
                $myVal['tabs']=$table;
                $myVal['col_field']=trim($col);
                $myVal['col_type']=$type;
                $myVal['col_key']=$key;
                $myVal['created_on']=date("Y-m-d");
                $myVal['checked']=1;
                $myVal['temp_name']=$id;

                //echo json_encode($myVal);
                //die();
                $rec=$this->accounts_model-
>select_table('tbl_fields','tabs'," WHERE tabs=".$table."' AND dbase=".$myVal['dbase']."' AND
col_field=".$col."' AND user_id=".$post['user_id'],'row');

```



```

                                if(!$rec){
                                    $val = $this->accounts_model-
>_insert($myVal,'tbl_fields');
                                }else{
                                    $val=true;
                                }
                            }
                        }else if($action=='update'){
                            /*$array=count($post['dbase']);
                            $myVal = array();

                            for($i=0; $i<$array; $i++){
                                $myVal['source']=$post['source'];
                                //$myVal['dbase']=$post['dbase'][$i];
                                $myVal['created_on']=date("Y-m-d");
                                $myVal['checked']=0;

                                $rec=$this->accounts_model-
>select_table('tbl_database','dbase'," WHERE dbase='".$post['dbase'][$i]."' and
user_id='".$post['cid'],'row');

                                if($rec){

                                    $arrays=array('dbase'=>$post['dbase'][$i],'user_id'=>$post['cid']);
                                    echo json_encode($myVal);
                                    $val = $this->accounts_model-
>_update($myVal,'tbl_database','user_id',$post['cid']);

                                    //$val = $this->accounts_model-
>_insert($myVal,'tbl_database');

                                }else{
                                    $val=true;
                                }
                            }*/
                        }
                    }

                    /***** output message *****/
                    if($val){
                        echo json_encode(array('info'=>'success'));
                    }
                    else{
                        echo json_encode(array('info'=>'error'));
                    }
                }
                break;
                default :
                $page = 'frm_modules';
                break;
            }
        }
    }
}

```