Design and Implementation of a Data Warehouse: Case of Ethiopian Ministry of Science and Higher Education

Mengistu Kumilachew
HiLCoE, Software Engineering Programme, Ethiopia
Information Network Security Agency, Ethiopia
mengekume@gmail.com

Fekade Getahun
HiLCoE, Ethiopia
Department of Computer Science, Addis Ababa University, Ethiopia
fekade.getahun@aau.edu.et

Abstract

The reporting and sharing of information have been synonymous with databases as long as there have been systems to host them. Now more than ever, users expect the sharing of information in an efficient and secure manner. However, due to the sheer number of databases, getting the data in an effective way requires a coordinated effort between the existing systems. There is a need today to have a data warehouse and business intelligence system that users can easily utilize to make improved business decisions, rather than trying to traverse the multiple sources. A data warehouse integrates heterogeneous data sources and provides information for analysis and decision support. Business Intelligence (BI) helps users to make quality business decisions to resolve the business needs of an organization.

The problem at Ministry of Science and Higher Education involves a description of data warehousing techniques, design, expectations, and challenges regarding data cleansing and transforming existing data, as well as other challenges associated with extracting from transactional databases. It also includes a technical piece discussing database requirements and technologies used to create and refresh the data warehouse. Finally, there are explanations for how users will consume the data in the data warehouse, such as through reporting and other BI.

The proposed solution advocates the use of design science methodology. The system is developed taking into account data warehouse and business intelligence approaches and related technologies, identify the specific requirement of data warehouse and business intelligence system, propose and design an architecture for the system, and develop and implement a prototype to demonstrate the practicability of the designed system. Then information about system requirements is gathered. Based on the assessment and system requirement, a system is designed and a prototype is implemented. The proposed prototype is evaluated by subject matter experts and the analysis shows acceptable result.

Keywords: Data Warehouse; Business Intelligence; ETL; OLAP; Data Mart; Business Decisions

1. Introduction

Businesses look forward to store critical data in easy-to-find and centralized locations. Employees at all levels would be able to tap that rich data to make decisions based on concrete and analytical facts instead of gathering scattered information from different sources or using plain intuition [1].

In modern organizations there exists database and an On-line transaction processing system. The on-line transaction system supports day-to-day business operations and accumulates a huge amount of data.

One of the drawbacks of on-line transaction system is that it can only provide a basic day-to-day business process. Many transactions are produced every day and stored in multiple physical locations.

Data warehouse is one of the emerging technologies that helps to manage data effectively and can deliver information to the decision-maker under a reasonable amount of time. Data warehouse concepts are created for the purpose of integrating and organizing large amounts of data for the ease of retrieval and analysis. Historical and current data from various sources must be cleansed and verified before it is deposited into the data warehouse.

The characteristic of data warehouse database is different from on-line transaction processing system.

Inmon [8], the father of data warehouse, describes data warehouse as a subject-oriented, integrated, non-volatile, and time-variant collection of data in support of management decision.

Using multidimensional data model, star and snowflake schema is almost always different and much simpler than the schema of an on-line transaction processing system. The data warehouse contains a derived data from on-line transaction processing system and other external data.

There are two approaches to build a data warehouse [8]: top-dwon and bottom-up. This paper focuses on bottom-up approach to build the individual warehouse because it is

- faster and easier to implement,
- less risk failure, and
- suitable for the mission of pilot project and create a prototype.

A data warehouse is a database of unique data structure that allows relativity quick and easy performance of complex quires over large amount of data. A classical production information system is preliminarily adapted to input data [2].

Business Intelligence (BI) is the delivery of accurate and useful information to decision makers with necessary time frame to support effective decision making [3].

Online analytic processing (OLAP) database is a technology for storing, managing, and querying data specifically designed to support business intelligence uses.

Extract, Transformation, and Load (ETL) System is a set of processes that cleans, transforms, combines, de-duplicates, archives, conforms, and structures data for use in a data warehouse.

This paper contains the description of the process development of the data warehouse and business intelligence. Snowflake schema is used to build the data warehouse model.

2. Related Work

Chumnungid [4] developed a data warehouse in a statistical environment which is a labor force survey case study. It is concerned with the creation of a new conceptual database, termed data warehouse, to provide information accessibility across different database platforms. The building of the data warehouse aims to store the most required data for delivery to the end user. The data warehouse that stores statistical data is the same as that developed for business use. The application tool that allows user access to data in the data warehouse is the OLAP tool. OLAP was designed to support data warehouse structure and access information across different dimensions or different levels of data in each dimension. The data warehouse provides easy and convenient statistical data when compared to the old traditional method of data access. Furthermore, a user can apply the data warehouse results to a decision support system.

Oketunji and Omodara [7] developed Data Warehouse and BI System in the case study of a Retail Industry in order to assist the management of the company in making better decision using the historical data available within the organization. The business users (decision makers) lack the ability to access data easily when needed. In an attempt to address this shortcoming, several departments within the company find their own resources, use different data available and hired consultants to solve their individual short-term data needs. In many cases, the same data was extracted from the same source systems to be accessed by separate departments without any strategic overall information-delivery strategy. The management realized the negative effect the different sources of the data has on the reports presented by the managers as the lack of integration. Given the importance of the information for the company, management was motivated to deal with the problem of data inconsistency by introducing a central data warehouse and to ensure that data is available to all users irrespective of their department. The data harmonization and the need for consistent and quality report gave birth to the project of data warehouse in the company.

Leonard and Edward [6] implemented Enterprise Data Warehouse within the context of a higher education environment. The reason for such an implementation was to give users at all levels of the university an integrated, secure and consistent data source from which they could report on and set their business needs more efficiently.

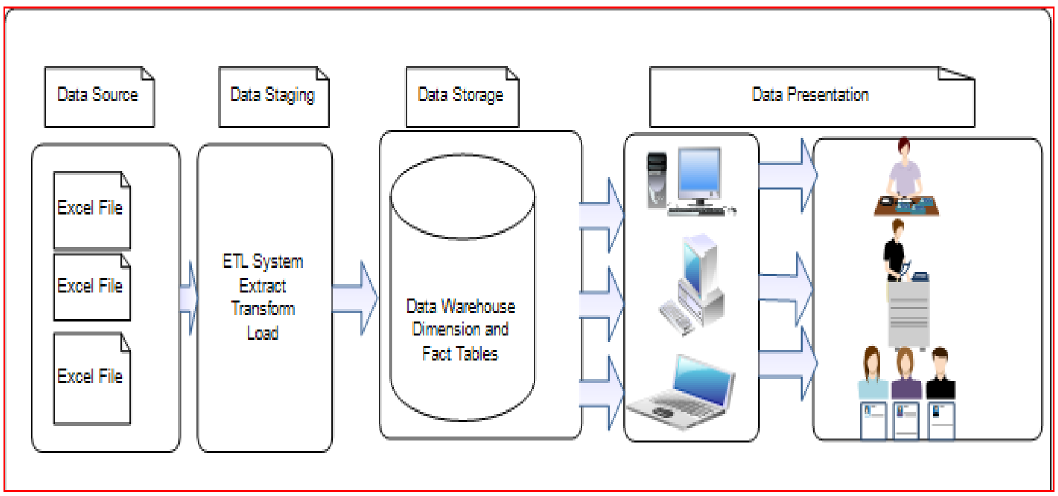
Abdullah and Obaid [5] designed and implemented educational data warehouse using OLAP. The decision making of an organization needs an extensive perspective on all parts and numerous organizations have along these lines created consolidated data warehouses that contain data drawn from various sources. Organizations have to transform their documents into a knowledge source. Therefore, the unified and consolidated perspective of data of the organization is introduced to the utilization of the data warehouse and it is most effective approach to achieve this purpose. In fact database with abilities of analyses is a set of subject-oriented, integrated, dependent on dissimilar time interims and fixed, that is utilized for administrative or decision support. Over a relatively long period, Management Information Systems (MIS) had a major effect on the management of centers of the training that are responsible for tasks similar to Decision Support Systems.

Recently, educational systems are encountering new difficulties and requirements of the dynamic economy to increase efficiency and quality. This poses the need to establish and institutionalize comprehensive systems for supporting decision making.

3. The Proposed Solution

3.1 System Architecture

One of the main aims of a data warehouse is to extract data from different sources and consolidate them in a repository for easy access. In the case of Ministry of Science and Higher Education, the data source is Microsoft Excel. The loading of the data warehouse was done through the use of ETL (Extract, Transformation and Load) process.



Figurer 1: Data Warehouse and Business Intelligence Architecture Design

As shown in Figure 1, there are four components: data source, data staging, data storage, and data presentation. The data source is an Excel file. The data staging area is the step for the data from the Excel file where data comes together through the ETL and meets its transformations before it continues to the actual data warehouse. The data storage area is the actual data warehouse full of data that is in dimension and fact tables, the standard formatting method of data warehouses. Finally, the presentation layer provides users with the data in a format they need to make decisions effectively.

3.2 The Data Warehouse Data Model

There are three levels of data modeling: conceptual, logical, and physical. The conceptual model is a concise and precise consolidation of all goal-relevant structural and behavioral features of the system presented in a predefined format.

The logical model aims at translating the conceptual representation of the database obtained in the conceptual model into a particular logical model. It defines the tables, attributes, keys, mandatory role constraints, and referential integrity. It also provides a foundation for the basis of a physical data model.

Physical modeling comprises the actual design of a database based on the requirements that were done in logical modeling. During physical modeling, tables and columns are created based on entities and attributes that were defined during logical modeling.

Data warehouses are based on the multidimensional model, which views data in an n-dimensional space. MultiDim model is powerful enough to represent the conceptual level, logical and physical elements required in data warehouse applications.

3.3 Transformation Rules

The following are the transformation rules used to transform from logical to physical model.

Rule 1: A level L is mapped to a dimension table that contains all attributes of the level. The identifier of the level will be the key of the table.

Rule 2: A fact F is mapped to a fact table that includes as attributes all measures of the fact. Further, a primary key is added to the table.

Rule 3: A relationship between a fact F and a dimension level L or between dimension levels (parent and child levels), can be mapped into two different ways, depending on its cardinalities.

Rule 3a: If the relationship is one-to-one or one-to-many, the table corresponding to the fact table or to the child level is extended with the primary key of the table corresponding to the dimension level or the parent level, respectively, that is, there is a foreign key in the fact or child table pointing to the other table.

Rule 3b: If the relationship is many-to-many, a new bridge table is created that contains as attributes the primary keys of the tables corresponding to the fact table and the dimension level, or the parent table and child levels, respectively. The physical design of the data warehouse is shown in Figure 2.

4. Implementation

4.1 Loading Stage, Dimension and Fact Tables

Before loading data into dimension and fact tables, the source data is loaded into the stage database. The aim of the staging is to load the data without much transformation.

Source System: Ministry of Science and Higher Education accepts the data from universities in Excel file. So the source of data is the Excel file.

Data staging: The source data is loaded into the stage database. The aim of the staging is to load the data without much transformation.

Data Extraction: Identifies the entire data source. It also specifies the source files from which data is to be extracted. During this stage, data is extracted from the source Excel file and prepared for transformation.

Data Transformation: Data is prepared for loading into the data warehouse data staging repository.

*Data Loading: E*ach major group of data must be kept up-to-date in the data warehouse. The data is loaded into the data warehouse staging tables as shown in Figure 3.

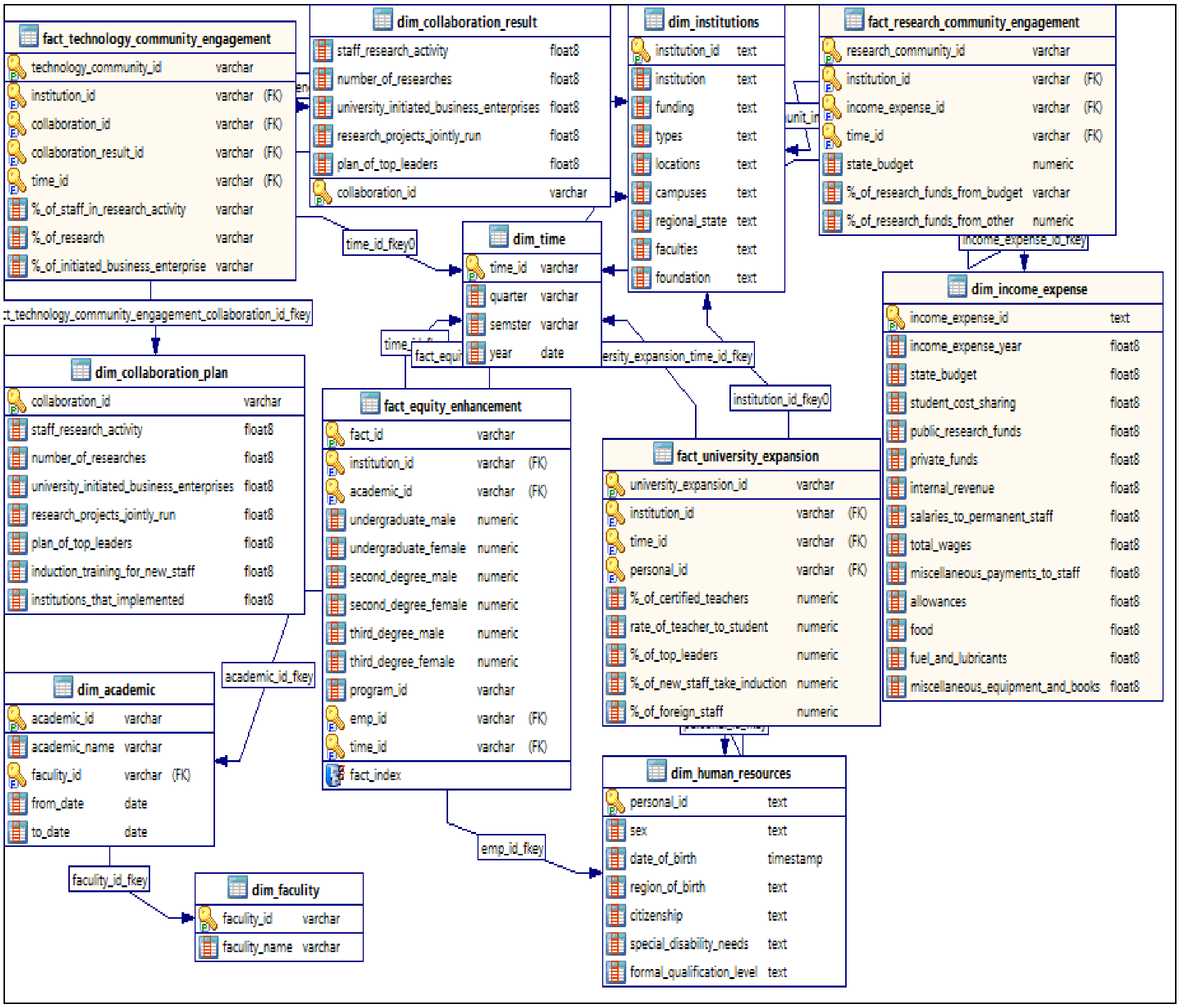


Figure 2: Physical Design

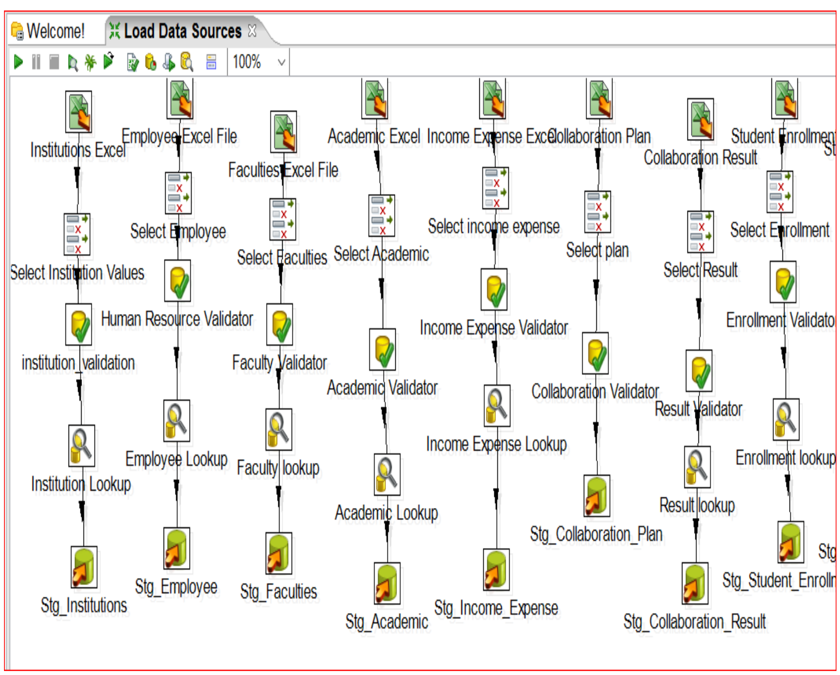


Figure 3: Loads Data into Staging Tables

Next is loading both the dimension and the fact tables as it is required. The following ETL design loads the data into the dimensional and fact tables. Data Extraction: Data is extracted from stage tables.

Data Transformation: After extraction data is transformed. To transform pentaho has select values transform design which is used to select, alter and to remove unnecessary fields.

Data loading: After selecting and removing the fields, running the transformation loads data into dimension tables as shown in Figure 4.

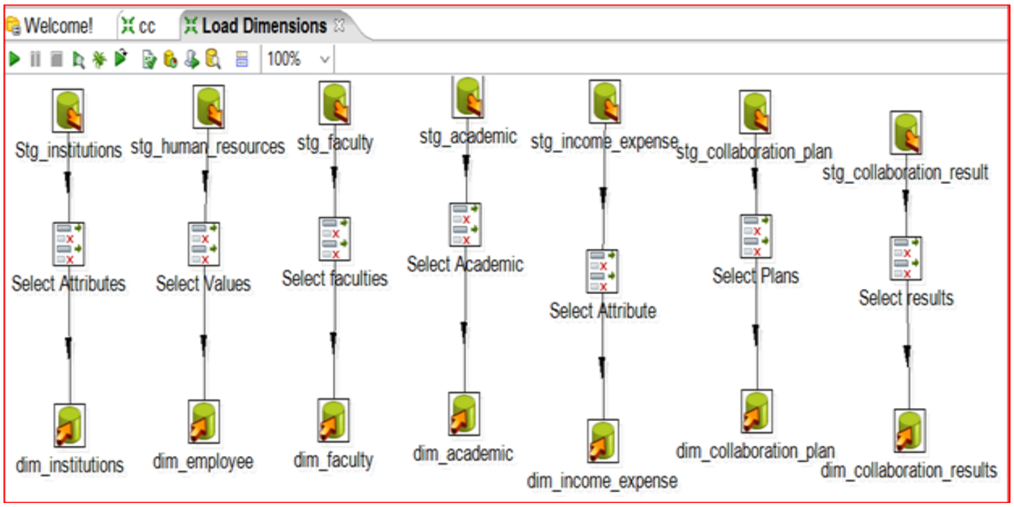


Figure 4: Loading Data into Dimension Tables

After loading the dimensions tables, the next step is loading of the fact tables as shown in Figure 5.

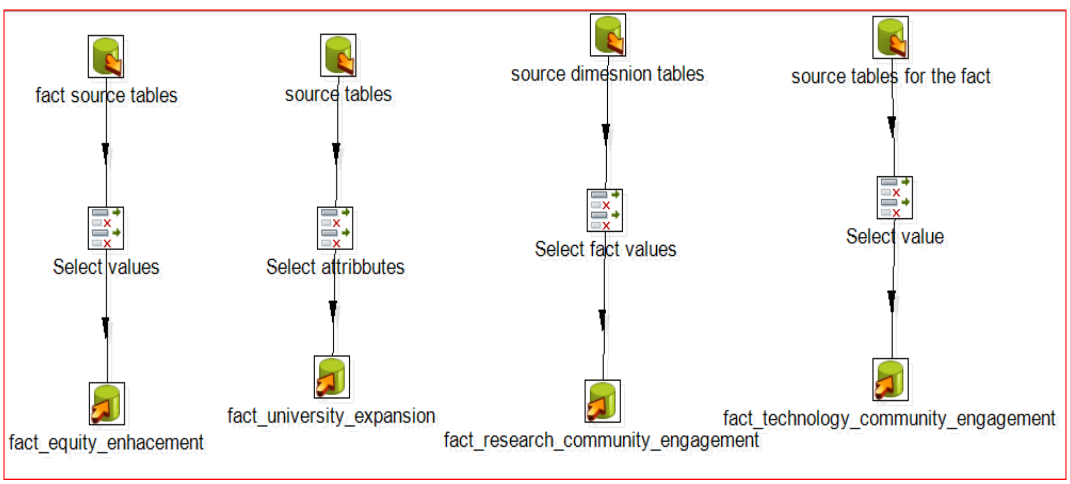
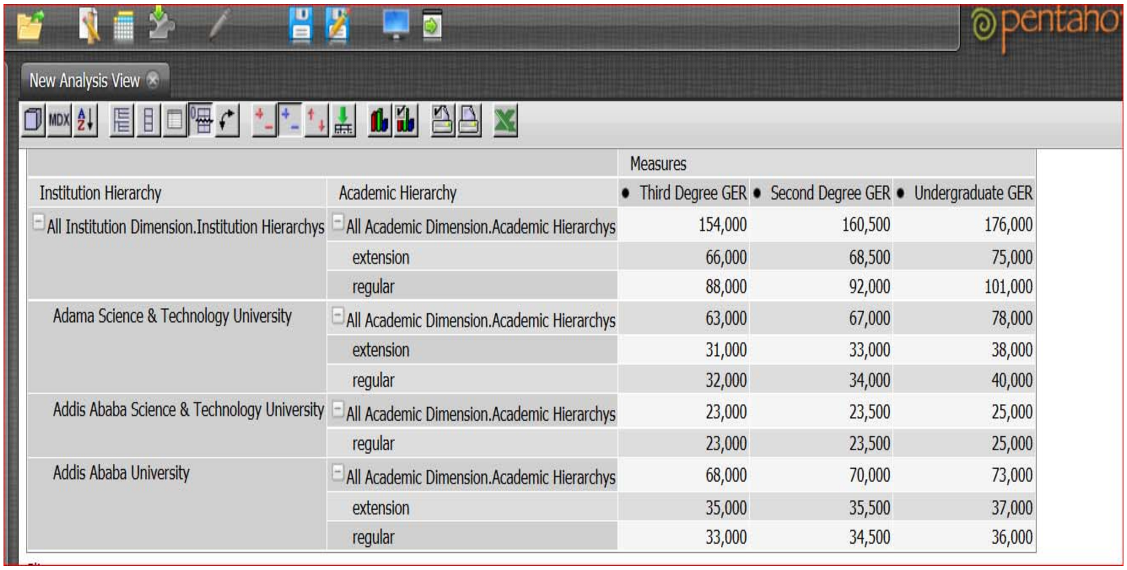


Figure 5: Fact Table Load

4.2 Prototype

The data warehouse has been loaded with the right data. It is now time to make use of the data stored in the data warehouse. Business intelligence is all about making best use of the available data in order to make better decision by Ministry of Science and Higher Education.

Figure 6 shows how to analyze gross enrolment information of undergraduate, second degree and third degree students in a given institution. By linking through the dim_faculity, dim_academic and dim_institutions tables in the schema and further link to fact_equity_enhacement.



Institution Hierarchy	Academic Hierarchy	Third Degree GER	Second Degree GER	Undergraduate GER
All Institution Dimension.Institution Hierarchys	All Academic Dimension.Academic Hierarchys	154,000	160,500	176,000
	extension	66,000	68,500	75,000
	regular	88,000	92,000	101,000
Adama Science & Technology University	All Academic Dimension.Academic Hierarchys	63,000	67,000	78,000
	extension	31,000	33,000	38,000
	regular	32,000	34,000	40,000
Addis Ababa Science & Technology University	All Academic Dimension.Academic Hierarchys	23,000	23,500	25,000
	regular	23,000	23,500	25,000
Addis Ababa University	All Academic Dimension.Academic Hierarchys	68,000	70,000	73,000
	extension	35,000	35,500	37,000
	regular	33,000	34,500	36,000

Figure 6: Graduated and Enrolled Student Information

The user can also analyse the data by bar graph and a slicer as shown in Figure 7. Pentaho biserver has a feature called a slicer, which makes the analysis more interactive for the user.

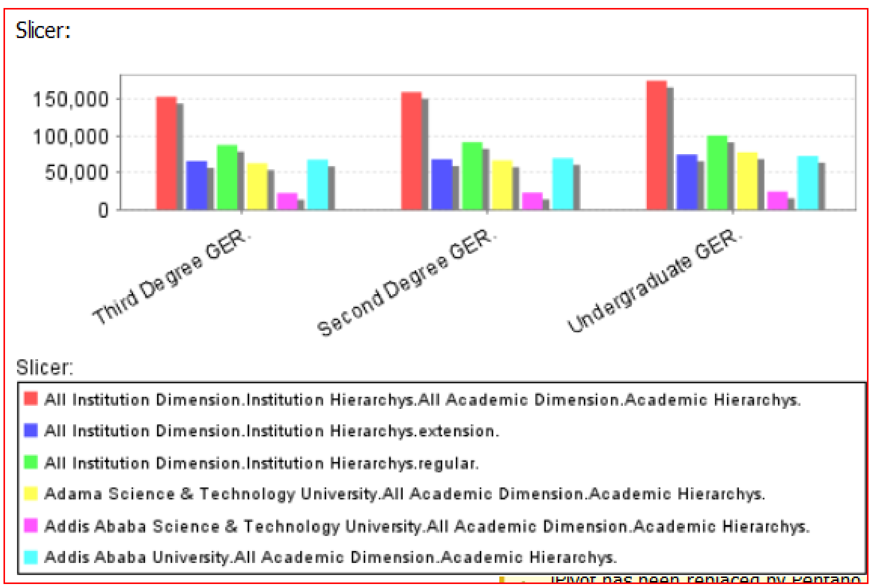


Figure 7: Graduated and Enrolled Student Information

5. Conclusions and Future Works

The paper focused on designing and implementing a data warehouse for the Ministry of Science and Higher Education to better integrate the system for simpler and improved reporting. Efforts were made to consider all the procedures of data warehouse and business intelligence. Data can be integrated from different sources to a single repository called data warehouse which is used for delivering business intelligence to the end users. Historical and current data from various sources must be cleansed and verified before it is deposited into the data warehouse. The ETL process is the significant step of data quality before data analysis, data summary and presentation.

Future development of the application should focus on new analysis functionality for end user as well as optimizing performance. Continuous performance testing and redesign of the cube together with its access layer is likely to be needed in the future as well. Extended functionality and increased performance must however not restrict the current scalability of the system due to the fact that requirements are expected to change in the future as well. Possible follow-up work should therefore be aiming at reviewing different parts, requirements specification, the requirements architecture itself and the verification with practitioners who can contribute a lot with deep implementation experience. The verification should be extended to include this experience. Possibilities are reviews of the document and description of the reference architecture by practitioner, interviews to focus on certain points of the reference architecture and to gather scenarios to apply the reference architecture to and test it fits.

References

[1] Miriam Tuerk, “The Open Source Data Warehouse Revolution", Seth Grimes Alta Plana Corporation, June 10, 2008.

[2] Sweety Patel, “Original Research Articles”, Department of Computer Science, Fairleigh Dickinson University, USA, Vol. 1, No. 4, 2012.

[3] Larson, B Delivering Business Intelligence with Microsoft SQL Server 2005, McGraw-Hill/Osborne

[4] Kuntarod Chumnungid, "Developing a Data Warehouse in a Statistical Environment: Labor Force Survey Case Study, Faculty of Graduate Studies, Mahidol University, 2002.

[5] Zina A. S. Abdullah and Taleb A. S. Obaid, "Design and Implementation of Educational Data Warehouse Using OLAP", College of Information Technology, University of Basra, Iraq, Vol. 5, Issue 5, October 2016.

[6] Leonard and Edward M., "Design and Implementation of an Enterprise Data Warehouse", Marquette University, December 2011.

[7] Temitope Adeoye Oketunji and Raufu Olalekan Omodara, “Design of Data Warehouse and Business Intelligence System”, School of Computing, Blekinge Institute of Technology, Karlskrona, Sweden, June 2011.

[8] Inmon W. H., “Building the Data Warehouse”, A Wiley QED Publication, John Wiley and Sons, Inc., New York.