

Design of a Data Warehouse Model for a University Decision Support System

Ibrahim Inuwa[1] Dr. N. D. Oye[2]

1. Department of Information Technology Modibbo Adama University of Technology Yola, Adamawa State - Nigeria.
2. Department of Computer Science Modibbo Adama University of Technology Yola, Adamawa State - Nigeria.

Abstract
Data Warehouse (DW) can be a valuable asset in providing a stress-free access to data for reporting and analysis. Regrettably, building and preserving an active DW is usually associated with numerous hitches ranging from design to maintenance. Research in the field of data warehousing has led to the emergence of vital contemporary technologies to aid design, management, and use of information systems that is capable of conveying a Decision Support System (DSS) to organizations. Nevertheless, in the face of persistent achievement and evolution of the field, abundant research is still left unturned across many diverse areas of the data warehousing. The objective of the paper therefore, is to design a DW database model for a University DSS using a dimensional modeling and techniques. A proposed DW database model with specific focus on modeling and design has been realized in this study. The researchers have demonstrated on how a DW database model can be realized using the dimensional modeling and technique.
Keywords: Data Warehouse, Modeling, Decision Support System, Decision Making.

1. Introduction

A typical university often comprises a lot of subsystems crucial for its internal processes and operations. Examples of such subsystems include the student registration system, the payroll system, the accounting system, the course management system, the staff system, and many others. In essence, all these systems are connected to many underlying distributed databases that are employed for every day transactions and processes. However, universities rarely employ systems for handling data analysis, forecasting, prediction, and decision making (Youssef, 2012). Over the last few years, organizations have increasingly turned to data warehousing (DW) to improve information flow and decision support. A DW can be a valuable asset in providing easy access to data for analysis and reporting. Unfortunately, building and maintaining an effective DW has several challenges (Güzin, 2007). According to Inmon (1992) a DW is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process. Inmon and Hackathorn (1994) stated that a DW is a distinct data extracted from one or more production databases to produce an authoritative source for decision support. It can be considered a decision support system. Reporting data is another important aspect of data warehousing because the main output from DW systems are either queries with minimal formatting or formal reports.

Research in data warehousing and Online Analytical Processing (OLAP) has produced important technologies for the design, management, and use of information systems for decision support. Much of the interest and success in this area can be attributed to the need for software and tools to improve data management and analysis given the large amounts of information that are being accumulated in corporate as well as scientific databases. Meanwhile, industry has adopted these established technologies and developed many of them to commercial maturity. However, despite the continued success and maturing of the field, much research remains to be done across many different areas of data warehousing (Manifesto of a Dagstuhl Perspectives Seminar, 2004).

In other to pave the way for future research in the whole field of DW, Stefano, R., Alberto A., Jens L. and Juan T. (2006) specifically focus on modeling and design of DW, trying to answer the following question: "Has research on this topic come to an end? If not, what's left to do?" they believed that, Multidimensional modeling requires specialized design techniques. Though a lot has been written on how a data warehouse should be designed, there is no consensus on a design method yet. Thus, overall, research on DW modeling and design is far from being dead, partly because more sophisticated techniques are needed for solving known problems, partly because of the new problems raised during the adaptation of DWs to the peculiar requirements of today's business. This paper suggests a DW design for a university information system whose aim is to support decision making. The objective of the paper therefore, is to design a database model for a University Decision Support System using a dimensional modeling and techniques.

2. Literature Review

Demarest (2013) was explicit when it say that planning the developing and deployment of a standard data warehouse should be taken as an IT project, hence what made IT project fail applies also applies when

developing data warehouse; thus the need for Project Planning and following the system development life cycle. There is the need for careful planning, requirements specification, design, prototyping and implementation.

Data warehousing is the process of collecting data to be stored in a managed database in which the data are subject-oriented and integrated, time variant, and nonvolatile for the support of decision-making (Inmon, 1993). Data from the different operations of a corporation are reconciled and stored in a central repository (a DW) from where analysts extract information that enables better decision making (Cho and Ngai, 2003). Data can then be aggregated or parsed, and sliced and diced as needed in order to provide information (Fox, 2004). There are two main authors that are known in the world of DW design, their approaches to some area of the data warehousing are different; William Inmon and Ralph Kimball. The approach by Inmon is top down design while that of Kimball is bottom up design. Most of the practitioners of DW subscribe to either of the two approaches. According to Inmon (1993) a DW is a subject-oriented, integrated, time-variant, non-volatile collection of data used in support of decision making processes. “Subject Oriented” means that a DW focuses on the high-level entities of the business (Chan, 1999) and the data are organized according to subject (Zeng, 2003 and Ma, 2000) ”Integrated” means that the data are stored in consistent formats, naming conventions, in measurement of variables, encoding structures, physical attributes of data, or domain constraints (O’Leary, 1999). For example, whereas an organization may have four or five unique coding schemes for ethnicity, in a DW there is only one coding scheme (Chan, 1999).

According to Kimball and Ross (2002) DW is the conglomerate of all Data Marts within the enterprise. Information is always stored in the dimensional model. Kimball views data warehousing as a constituency of Data Marts. Data Marts are focused on delivering business objectives for departments in the organization. And the DW is a conformed dimension of the Data Marts. Kimball *et al*., (1996) described a Data Mart as a subset of DW. The DW is the sum of all the Data Marts, each representing a business process in organization by a means of a star schema, or a family of star schemas of different granularity. The main difference between the approach of Kimball and Inmonis that Kimball’s conformed dimensions are de-normalized, whereas that of the Inmon uses a highly normalized central database model. Inmon's Data Marts store a second copy of the data from the centralized DW tables whereas the dimensions of Kimball’s Data Marts are not copies of the conformed dimensions, but the dimension tables themselves. Kimball *et al*., (1996) refer to the set of conformed dimensions as the DW bus. There is no right or wrong between these two ideas, as they represent different data warehousing philosophies.

Ballard (1998) gave an assessment of the evolution of the concept of data warehousing, as it relates to data modeling for the DW, they defined database warehouse modeling is the process of building a model for the data in order to store in the DW. There are two data modeling techniques that are relevant in a data warehousing environment and they are:

i. Entity Relationship (ER) Modeling: ER modeling produces a data model of the specific area of interest, using two basic concepts: entities and the relationships between those entities. Detailed ER models also contain attributes, which can be properties of either the entities or the relationships. The ER model is an abstraction tool because it can be used to understand and simplify the ambiguous data relationships in the business world and complex systems. ER modeling uses the following concepts: entities, attributes and the relationships between entities. The ER model can be used to understand and simplify the ambiguous data relationships in the business world and complex systems environments. An entity is represented on the diagram as a rectangular box. The name of the entity appears in the top section of the rectangle, such as Instructor and Department in the below Figure 1. Attributes are listed in the lower part of the rectangle that represents the entity which they belong to. A relationship is a line between the boxes that represents the entities that it links. The relationship may be identified in different ways like; one-to-one, many-to many, and one to many.
ii. Dimensional Fact Modeling: Dimensional modeling uses three basic concepts: Measures, facts, and dimensions, Dimensional modeling is powerful in representing the requirements of the business user in the context of database tables. Measures are numeric values that can be added and calculated as can be seen in Figure 2 (Ballard, 1998). Each dimension table is assigned a primary key that are not related to one another. The primary key is unique key for the dimension table, which is replicated in a fact table where it is referred to as a foreign key. A fact table is a table that contains the data (factual history) like Student ID, Course ID, Semester ID, etc. (Ballard, 1998).

Thomas and Carol (2002) derived the way a DW or a Data Mart structure in dimensional modeling into several ways. Flat schema, Terraced Schema, Star Schema, Fact Constellation Schema, Galaxy Schema, Snowflake Schema, Star Cluster Schema, and Star flake Schema. However there are two basic models that are widely used in dimensional modeling: Star and Snowflake models.

i. Star Schema: The Star Schema (in Figure 3) is a relational database schema used to hold measures and dimensions in a Data Mart. The measures are stored in a fact table and the dimensions are stored in dimension tables. For each Data Mart, there is only one measure surrounded by the dimension tables, hence the name star schema. The center of the star is formed by the fact table. The fact table has a column or the measure and the column for each dimension containing the foreign key for a member of that dimensions. The key for this table is formed by concatenate all of the foreign key fields. The primary key for the fact table is usually referred to as composite key. It contain the measures, hence the name "Fact". The dimensions are stored in dimension tables. The dimension table has a column for the unique identifier of a member of the dimension, usually an integer of a short character value. It has another column for a description. In this project to follow the naming convention we are going to name the dimension tables based on the information they contained and prefix with "Dim" (Ballard, 1998).

ii. Snowflake Schema: Snowflake Schema model is derived from the star schema and, as can be seen, looks like a snow flake. The snowflake model is the result of decomposing one or more of the dimensions, which generally have hierarchies between themselves. Many-to-one relationships among members within a dimension table can be defined as a separate dimension table, forming a hierarchy as can be seen in Figure 4.

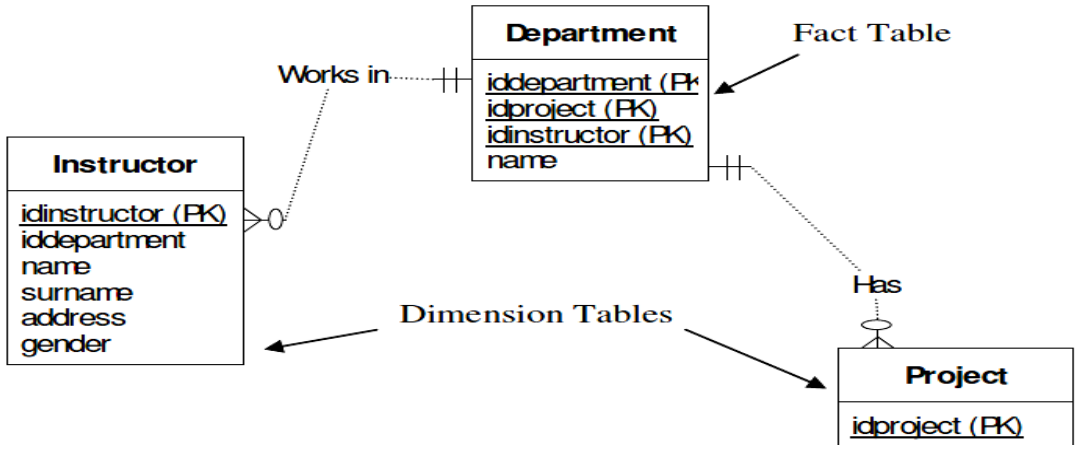

Figure 2: Dimension and Fact Tables with Primary and Foreign Keys (Ballard, 1998).

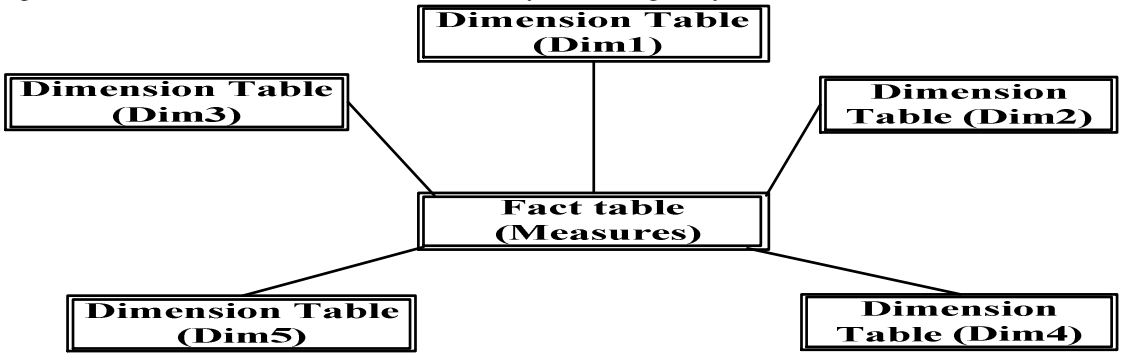

Figure 3: A Star Schema (Ballard, 1998).

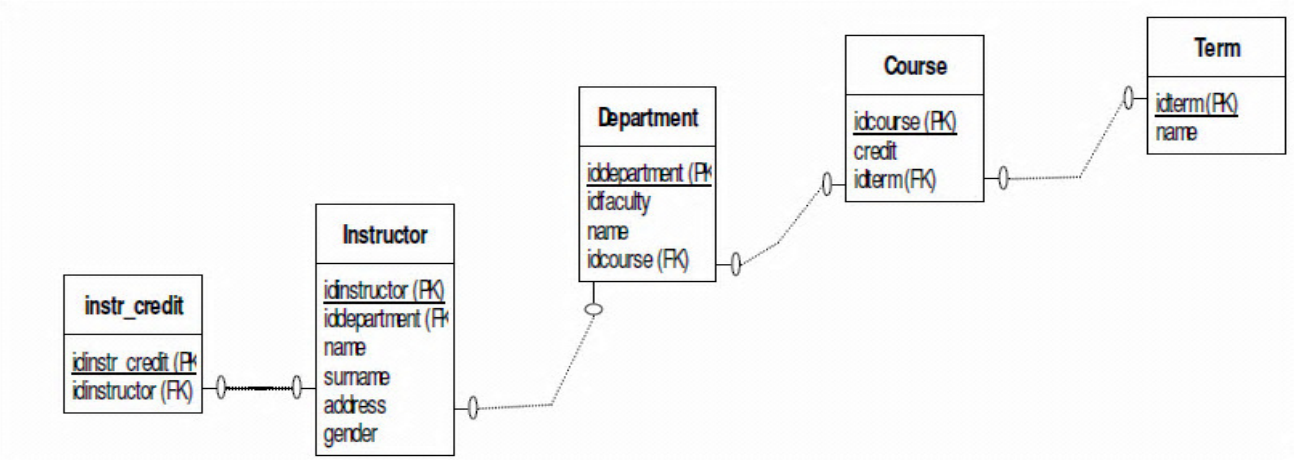

Figure 4: Example of a Snowflake Schema (Ballard, 1998).

There are three levels of data modeling. They are conceptual, logical, and physical. For the purpose of this thesis, we would discuss only the first two. Conceptual design manages concepts that are close to the way users perceive data; logical design deals with concepts related to a certain kind of DBMS; physical design depends on the specific DBMS and describes how data is actually stored. The main goal of conceptual design modeling is developing a formal, complete, abstract design based on the user requirements. DW logical design involves the definition of structures that enable an efficient access to information. The designer builds multidimensional structures considering the conceptual schema representing the information requirements, the source databases, and non-functional (mainly performance) requirements. This phase also includes specifications for data extraction tools, data loading processes, and warehouse access methods. At the end of logical design phase, a working prototype should be created for the end-user (Basaran, 2005).

Galhardas (2001) opines that building a DW from independent data sources is a difficult process. This process involves extracting, converting, cleaning, integration and transformation of the data. In order to do these operations, an ETL (extract, transform, and load) tool is required. The key steps that need to be undertaken to transform raw operational data to a form that can be stored in a DW for analysis are:

i. Extraction the goal of the data extraction step is to bring data from different sources into a database before modification.
ii. Converting the data into a format that is suitable to the DW.
iii. Cleaning of the data, data entry errors and differences in schema formation can cause for example student dimension table to have several corresponding entries for a single student.
iv. Integration of the different datasets to suit the data model of the DW.
v. Transformation, of the data through summarization and creation of new attributes, it is a set of rules and scripts that typically handles the transformation of data from an input schema to the destination schema.

3. Dimensions and Fact Tables of the Proposed DW Design Model

The use of Microsoft Office Visio 2010 tools was adopted to create the Dimension and Fact tables for this paper as can be seen in Figure 5 and 6 respectively. In the Figure 5, the dimensions for each department are stored in dimension tables. The dimension tables have a column for the Primary Keys, (PK Programme, PK Course, PK Student, PK Lecturer, PK Thesis, PK School and PK Department) they are the unique identifiers in the dimension tables. In this study, we adopted a naming convention to name our dimension tables based on the information they contained and prefix with "Dim". For example, the information and prefix "Dim_ Programme IT Dept." is a dimension table from Information Technology department that contained information about programmes. In the Figure 6, the Fact table has a column and measure, the column holds the Primary Keys of dimension tables, the primary keys automatically becomes Foreign Keys (FK) in Fact table. Mark, Grade, Remark and External Exam/Viva date columns are the measures in the Fact tables (measures are numeric values that can be added and calculated). We adopted a naming convention to name our Fact tables based on the information they contained and prefix with "Fact_Table". For example, the information and prefix "Fact Table_Graduate PG Dept. is a Fact table from the PG department containing information about graduates.

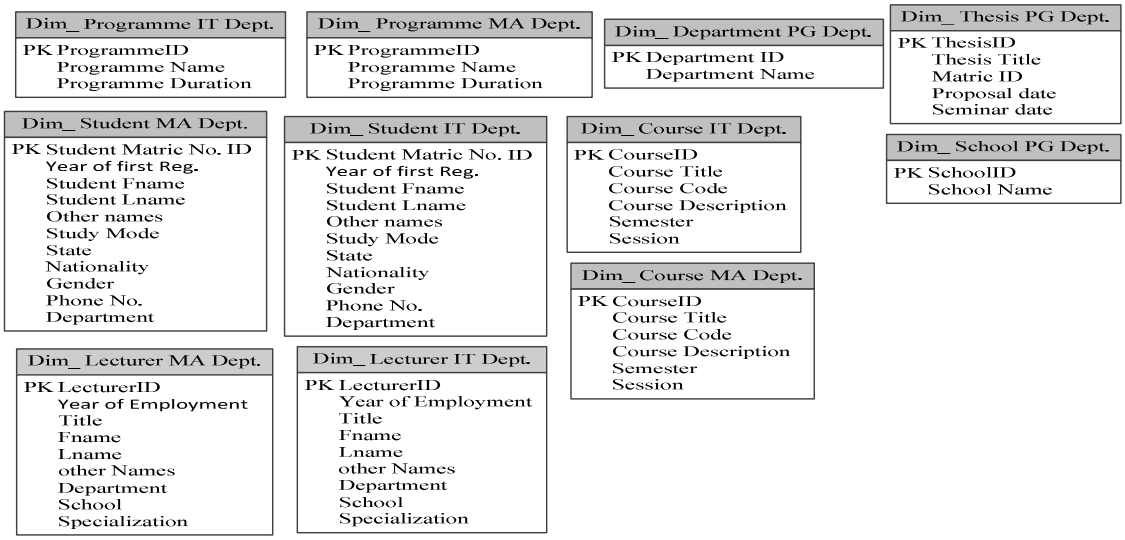

Figure 5: Dimension Tables of the DW Prototype Model

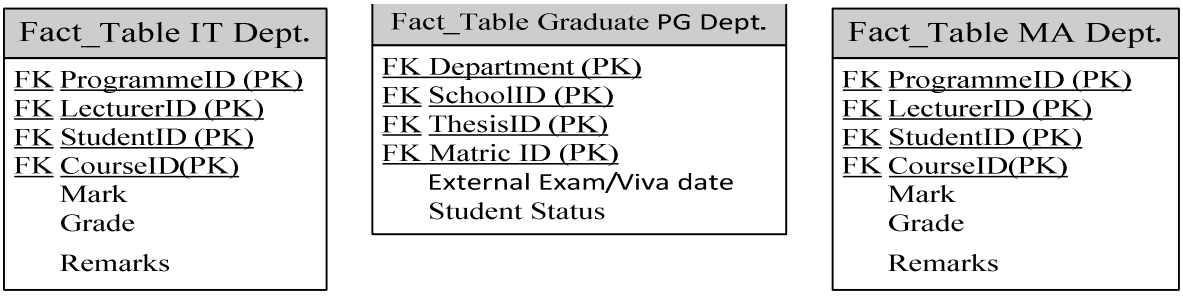

Figure 6: Fact Tables of the DW Prototype Model.

4. The Logical Data Marts Models of the Proposed DW Design

The researchers have started the database design with logical models for all the departmental Data Marts. The logical models are the representation of the data in a way that can be presented as a platform for the physical implementation. The main elements of the logical models are entities, attributes, and relationships; we designed the Data Marts through the Dimension and Fact tables. The Dimensions tables are the foundation of the Dimensional modeling, describing the objects of the database such as Student, Lecturer, Course and other dimension table used in the design of the logical Data Marts models. According to Ralph Kimball (See section II), Data Marts represent a unit or departmental process within an organization. Data Mart is the collection of Fact table and its dimension tables. The Fact table holds the Dimension's table attributes and Measures. The Dimension table attributes (primary keys of the dimension tables) are Foreign Keys or other attributes called degenerate dimension. The Fact table is collection of two types of attributes, (Dimension attributes and measures) and the collection of a Fact and Dimension table formed our Data Mart. We have used the bottom up DW design approach, which is the integration of the departmental Data Mart to form the DW database model by a star schema (Ralph Kimball's idea). The star schema is a relational database schema used to hold measures and dimensions in a Data Mart.

The Figure 7 and Figure 8 are representing the Dimensional Data Mart models of this study; it is the logical independent Data Mart model with entities, attributes, and relationships for the IT and MA departments by a star schema. It also shows the relationships between the Dimension tables (Dim_ programme, Dim_ Lecturer, Dim_ Student and Dim_ Course) and the "Fact_ Table IT Dept." containing the Foreign Keys: FK Programme ID (PK), FK Lecturer ID (PK), FK Student ID (PK) FK Course ID (PK), and their measures (Grade and Remark). The Figure 9 is representing the logical independent Data Mart model with entities, attributes, and relationships for the PG department by a star schema. It also shows the relationships between the Dimension tables (Dim_ School, Dim_ Department and Dim_ Thesis) and the "Fact_ Table PG Dept.", Containing Foreign Keys; FK

Department (PK), FK School ID (PK), FK Thesis ID (PK), and its measures (External Exam/Viva date and Student Status). The DW database logical model is created on an SQL Server 2012, so that the Fact and Dimension tables can be integrated into the DW database. The Figure 10 shows logical model of the DW database by a Star Schema. The DW database model is the one that integrates the Fact and Dimension tables of from different departments.

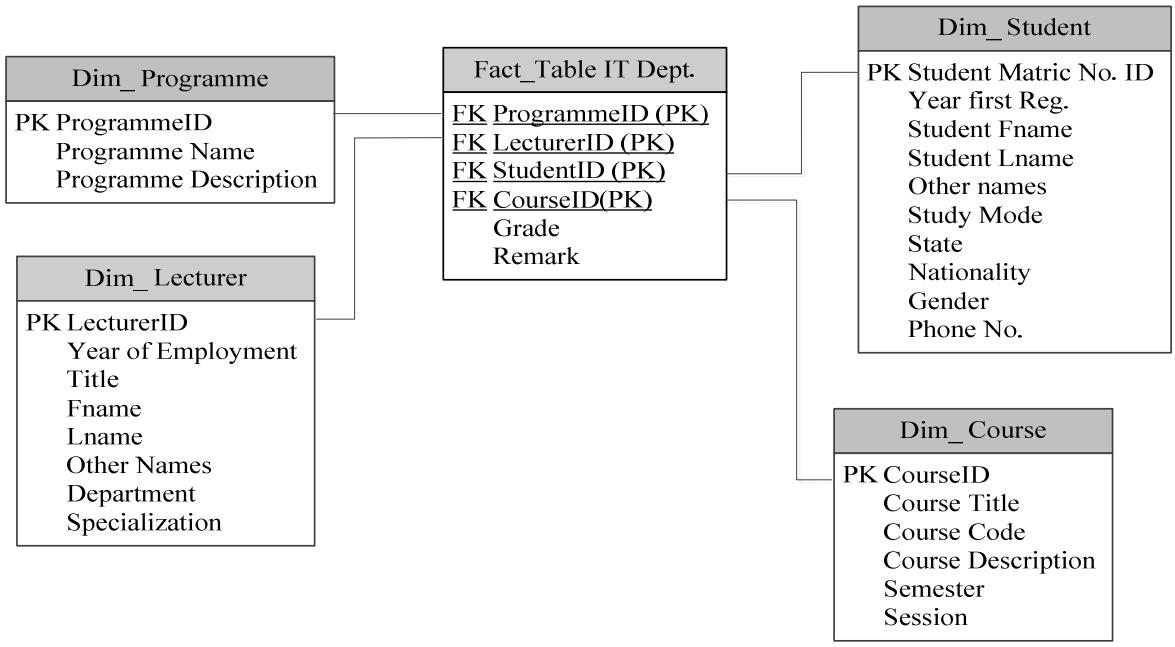

Figure 7: Data Mart Logical Model for Information Technology (IT) Department by Star Schema.

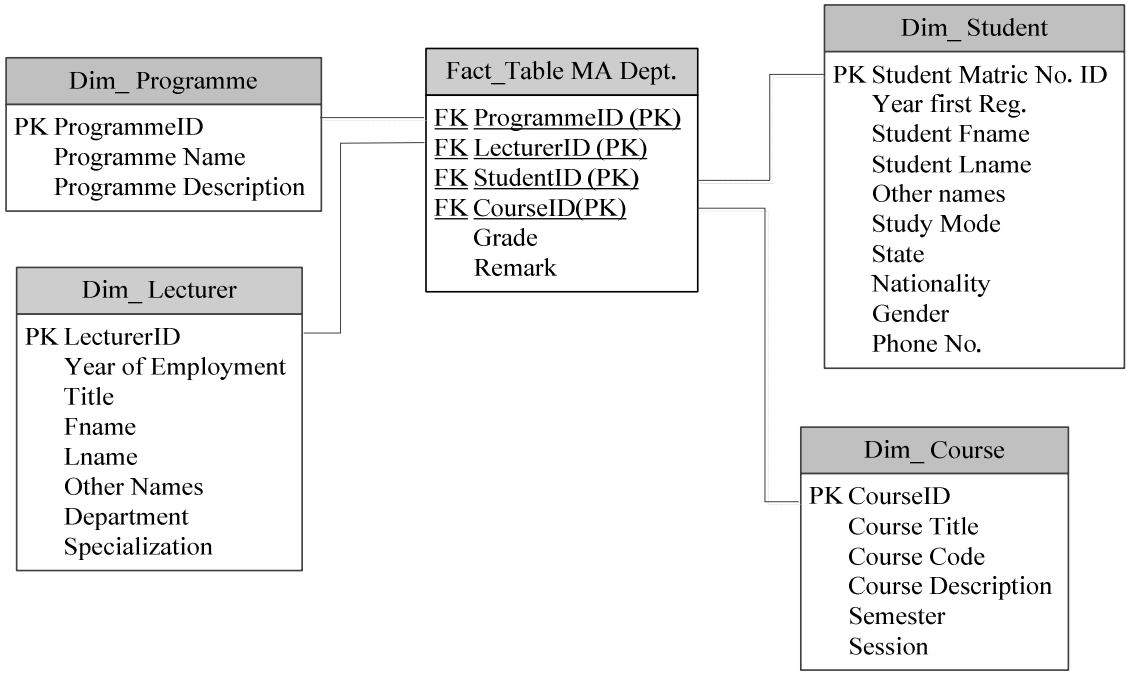

Figure 8: Data Mart Logical Model for Mathematics (MA) Department by Star Schema.

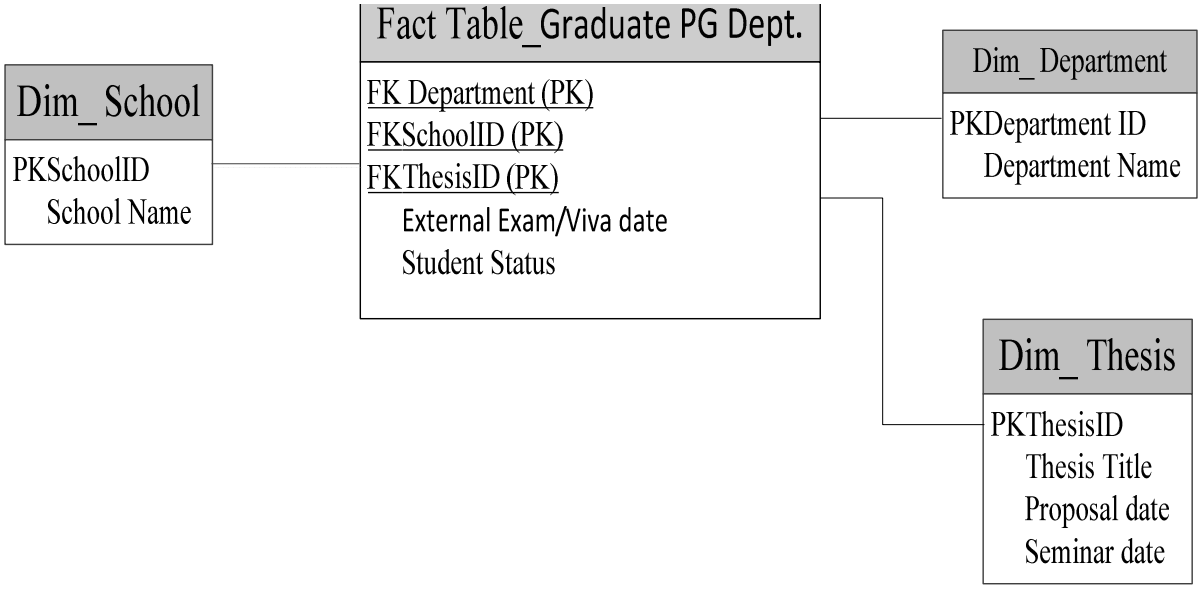

Figure 9: Data Mart Logical Model for the Postgraduate (PG) Department by Star Schema.

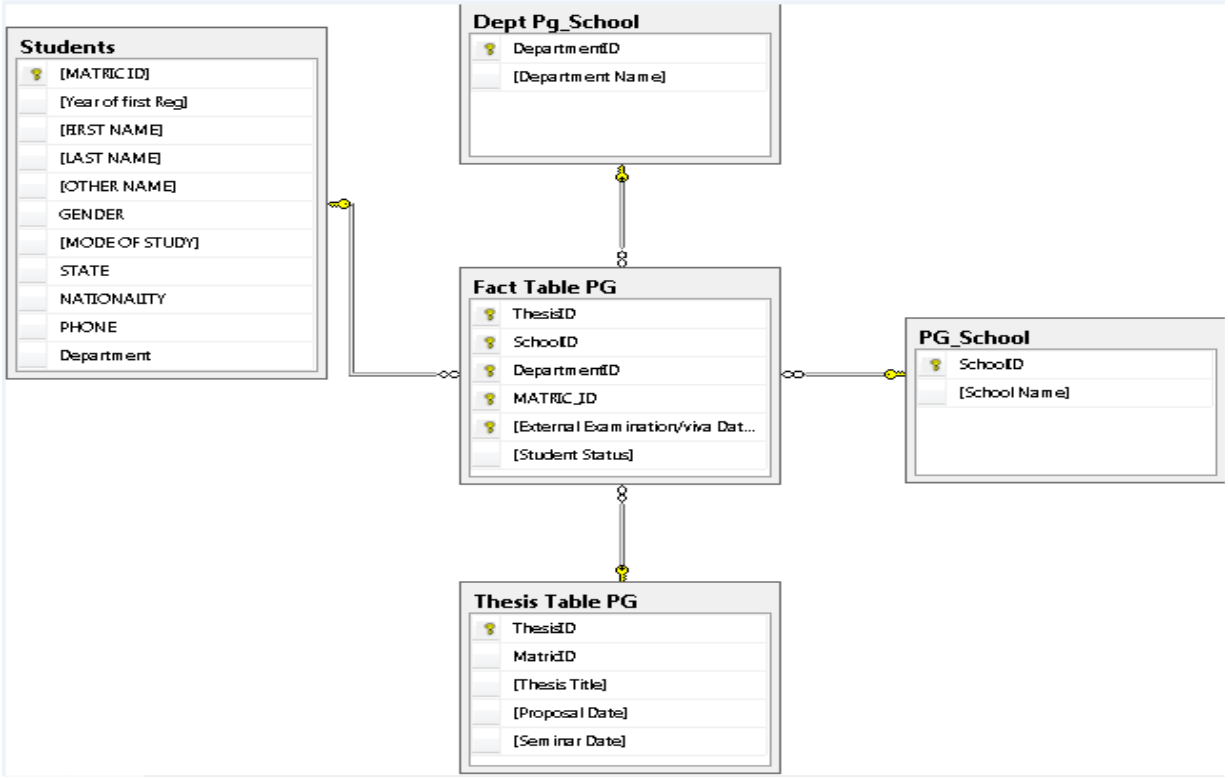

Figure 10: The Proposed DW Database Logical Model by a Star Schema.

5. Conclusion

Relating to multidimensional applications, more than a few well-organized multidimensional data configurations such as summarized cubes (Wang, W., Lu, H., Feng, J. and Yu, J. X., 2002) dwarfs (Sismanis 2003, Sismanis 2002), and QC-Trees (Lakshmanan, Pei and Zha, 2003) have been proposed to manage data cubes. Clearly, without the support of more expressive logical models we cannot expect to achieve a streamlined design process that guarantees quality criteria to be satisfied and seriously takes security considerations into account.

In this paper the researchers have focused on modeling and designing a DW database model for a University decision making. In the course of the study, the researchers have reflected on the thoughts of other scholars in relation to data warehousing modeling and design concepts. Afterwards the researchers proposed a dimensional database modeling technique and demonstrated how fact and dimensional tables is connected to form an

independent departmental (IT, MA and PG departments) data mart by a star schema model. The integration of the independent departmental data mart forms the logical DW database model. The overall task of logical modeling is the transformation of conceptual schemata into logical schemata that can be implemented on a chosen target system. In the DW domain, target database systems are typically either relational or multidimensional. Concerning relational implementations, the so-called star, constellation, and snowflake schemata are widely accepted to manage data cubes and are supported by various vendors (Manifesto of a Dagstuhl Perspectives Seminar, 2004)

References

Ballard, C. (1998). Data Modeling Techniques for Data Warehousing. *IBM International Technical Support Organization.* 36-37.

Başaran, and Beril, P. (2005)**.** A Comparison of Data Warehouse Design Models. The Graduate School of Natural and Applied Sciences Atilim University, Turkey.

Chan S. S., (1999). The Impact of Technology on Users and the Work place. *New Directions for Institutional Research*. 103. 3 – 21.

Cho, V. and Ngai, E.W.T. (2003). Data Mining for Selection of Insurance Sales Agents*. Journal of Expert Systems.* 20, 3-10.

Demarest, M. (2013). Data Warehouse Prototyping: Reducing Risk, Securing Commitment and Improving Project Governance. Retrieved June 28, 2014, from: http// www.wherescape.com.

Galhardas, H.(2001). Declarative Data Cleaning Model, Language and Algorithms. *Proc. VLDB Conf*., Morgan Kaufmann, San Francisco, pp.371-380.

Güzin, T. (2007). Developing a Data Warehouse for a University Decision Support System. The Graduate School of Natural and Applied Sciences, Atilim University.

Inmon, W. H. (1992). *What is a Data Warehouse? Sunnyvale*, CA: PRISM Solutions, Inc. Tech. 1, p. 1.

Inmon, W. H. & Hackathorn, R. D. (1994). *Using the Data Warehouse.* New York. John Wiley & Sons, Inc.

Kimball, R. & Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling.* New York. John Wiley & Sons, Inc.

Kimball, R., Reeves M. & Ross, M. (1996). *The DW Toolkit: The Complete Guide to Dimensional Modeling*. New York. John Wiley & Sons, Inc.

Lakshmanan, L. V. S., Pei, J. and Zhao,Y. (2003). QC-Trees: an Efficient Summary Structure for Semantic OLAP. *ACM SIGMOD Int. Conf. on Management of Data*, pp. 64–75.

Ma, C. (2000). Data Warehousing Technology Assessment and Management. *Journal of Industrial Management and Data Systems.*100, 3, 125 – 135.

Manifesto of a Dagstuhl Perspectives Seminar (2004). Data Warehousing at the Crossroads. Perspectives Workshop Dagstuhl, Germany.

O'Leary, D.E. (1999). *REAL-D*:A Schema for Data Warehouses. *Journal of Information Systems.*13, 49-62.

Thomas, C. & Caroly, B. (2002). *Database Systems*. New York. Addison-Wesley press.

Sismanis,Y., Deligiannakis, A., Roussopoulos, N. and Kotidis, Y. (2002). Dwarf: Shrinking the PetaCube. *ACM SIGMOD Int. Conf. on Management of Data*, pp. 464–475.

Sismanis, Y., Deligiannakis, A., Kotidis Y., and Roussopoulos, N. (2003). Hierarchical Dwarfs for the Rollup Cube. *6th ACM Int. Workshop on Data Warehousing and OLAP (DOLAP)*, pp. 17–24.

Stefano, R., Alberto A., Jens L. and Juan T. (2006). Research in Data Warehouse Modeling and Design: Dead or Alive*? Proceedings of the 9th ACM international workshop on Data warehousing and OLAP* (pp 3-10) ACM New York, NY, USA.

Wang, W., Lu, H., Feng, J. and Yu, J. X. (2002). Condensed Cube: An Efficient Approach to Reducing Data Cube Size. *18th Int. Conf. on Data Engineering (ICDE)*, pp. 155–165.

Youssef, B. (2012). A Data Warehouse Design for a Typical University Information System. *Journal of Computer Science & Research (JCSCR)*. 1, 6, 12-17.

Zeng, Y. (2003). Enterprise Integration with Advanced Information Technologies. ERP and data warehousing. *Journal of Information Management and Computer Security.*2, 6, 23-77.

[1]**Ibrahim Inuwa** receives his Bachelor's Degree in Mathematics with Economics from the Federal University of Technology, Yola – Nigeria in 2010. At the moment he is a Masters student (M. Tech. in IT Management) in the department of Information Technology Faculty of Management and Information Technology at the Modibbo Adama University of Technology, Yola – Nigeria.

[2]**Dr. Oye, N. D**. receives his M.Tech OR (Operations Research) degree from the Federal University of Technology Yola-Nigeria in 2002. He is currently a lecturer in the same University (for the past 15 years). He obtained his PhD in Information System at the Department of Information System, Faculty of Computer Science Univeristi Teknologi Malaysia in 2013.