BUILDING DATA WAREHOUSING AND DATA MINING FROM COURSE MANAGEMENT SYSTEMS: A Case Study of Federal University of Technology (FUTA) Course Management Information Systems

*Akintola K.G., ** Adetunmbi A.O. **Adeola O.S.
*Department of Computer Science, University of Houston-Victoria, Texas 77901 United States of America.
**Department of Computer Science, Federal University of Technology, Akure, Ondo State, Nigeria

ABSTRACT

In recent years, decision support systems otherwise called business Intelligence (BI) have become an integral part of organization's decision making strategy. Organizations nowadays are competing in the global market. In order for a company to gain competitive advantage over the others and also to help make better decisions, Data warehousing cum Data Mining are now playing a significant role in strategic decision making. It helps companies make better decisions, streamline work-flows, and provide better customer services. This paper gives the report about developing data warehouse for business management using the FUTA Student-Course management system as a case study. It describes the process of data warehouse design and development using Microsoft SQL Server Analysis Services. It also outlines the development of a data cube as well as application of Online Analytical processing (OLAP) tools and Data Mining tools in data analysis. It was concluded that the effective use Data-Warehousing and Datamining in Enterprise resource management system will promote the rapid growth of major companies.

Keywords: OLTP, Data warehouse, Data Mining, Dimensional modeling, OLAP

1.0 INTRODUCTION

In simple terms, a data warehouse (DW) is a pool of data produced to support decision making; It is also a repository of current and historical data of potential interest to managers throughout the organization. Data are usually structured to be available in a form ready for analytical processing activities (e.g. online analytical processing (OLAP), data mining, querying, reporting and other decision supporting applications). A data warehouse is a subject-oriented, integrated, time-variant, non-volatile collec-tion of data in support of management's decision-making process. [Efraim T. et al., 2010]. The day-to-day operations of an organization are done by using the OLTP system. This system is good for normal operations and few decision making but the system is inadequate when it comes to strategic decision support. The lack of historical data in OLTP makes it unsuitable to provide a comprehensive information about the operations of the business. DW on the other hand, provides a central repository of historical data which provides an integrated platform for historical analysis of data. With a data warehouse and Online Analytical Processing (OLAP), users can perform better data analysis and gain better knowledge from the repository data.

According to Stephen Brobst and Joe Rarey (2003) Five stages of decision support were identified in data-warehouse:

Stage 1: Reporting

The initial stage of data warehouse deployment typically focuses on reporting from a single source of truth within an organization. The biggest challenge in Stage 1 data warehouse deployment is data inte-gration.

Stage 2: Analyzing

In a Stage 2 data warehouse deployment, decision-makers focus less on what happened and more on why it happened. Analysis activities are concerned with drilling down beneath the numbers on a report to slice and dice data at a detailed level. Performance is also a lot more important in a Stage 2 data warehouse implementation because the information repository is used much more interactively.

Stage 3: Predicting

As an organization becomes well-entrenched in quantitative decision-making techniques and experiences the value proposition for understanding the "whats" and "whys" of its business dynamics, the next step is to leverage information for predictive purposes.

Advanced data mining methods often employ complex mathematical functions such as logarithms, exponentiation, trigono-metric functions and sophisticated statistical functions to obtain the predictive character-istics desired.

Stage 4: Ope-rationalizing

Ope-rationalization in Stage 4 of the evolution starts to bring us into the realm of active data warehousing. Whereas stages 1 to 3 focus on strategic decision-making within an organization, Stage 4 focuses on tactical decision support. Ope-rationalizing typically means providing access to information for immediate decision-making in the field. Two examples are (1) inventory management with just-in-time replenishment and (2) scheduling and routing for package delivery. Many retailers are moving toward vendor managed inventory, with a retail chain and the manufacturers that supply it working as partners. The goal is to reduce inventory costs through more efficient supply chain management.

Stage 5: Active Warehousing

An active data warehouse delivers information and enables decision support throughout an organization rather than being confined to strategic decision-making processes. However, tactical decision support does not replace strategic decision support. Rather, an active data warehouse supports the coexistence of both types of workloads.

The future of data warehousing

We are moving to the stage of a data ware housing applications that can provide information to many decision makers operational, strategic, and tactical and also to the customers as well in an integrated fashion.

Data Mining

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determinethe impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data (*Bill Palace 1996)*

With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

For example, Blockbuster Entertain-ment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures. (***Bill Palace 1996).***

WalMart is pioneering massive data mining to transform its supplier relation-ships. WalMart captures point-of-sale transactions from over 2,900 stores in 6 countries and continuously transmits this data to its massive 7.5 terabyte Teradata data warehouse. WalMart allows more than 3,500 suppliers, to access data on their products and perform data analyses. These suppliers use this data to identify customer buying patterns at the store display level. They use this information to manage local store inventory and identify new merchandising opportunities. In 1995, WalMart computers processed over 1 million complex data queries (***Bill Palace 1996).***

How data mining works

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1). Sometimes called the k-nearest neighbour technique.

- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.
- **Data visualization**: The visual interpretation of complex relationships in multi-dimensional data. Graphics tools are used to illustrate data relationships (Bill Palace 1996).

2.0 THE CASE PROJECT OF STUDENT COURSE MANAGEMENT SYSTEM

The case study used as a model for the project is a student course management system in Federal University of Technology Akure Nigeria. The school registers students every semester and students take courses and exams. These are being managed by an online transaction processing (OLTP) system. A simplified representation of the logical design of the OLTP system is shown in Figure 1.

The Motivation for the Data Warehouse system

The day-to-day operations of the school rely heavily upon the OLTP system. The staffs make use of OLTP system for management information system. The need for strategic decision necessitates the development of the data warehouse. The OLTP can only handle few historical data which are not enough to make strategic decisions. With a data warehouse and OLAP, staff and management can perform roll-up and drill-down operations to student enrollments by year, by semester by course or any combination desired. The system will provide decision support that is flexible and user-friendly. The University would like to use the data-warehouse to answer questions such as: What is the trend of student admission, student's enrollments, Lecturers room schedule, number of annual enrollment etc. These type of questions need a lot of historical data to generate which the current system OLTP system cannot support.

3.0 DESIGNING THE DATA WAREHOUSE

The following discussion outlines the process of the data warehouse design. It involves the logical design, the OLAP design, and Data mining design.

The logical Design

The logical design of data-warehouse is defined by the dimensional data modeling approach. The dimensioning design process followed in this project adheres to the methodology described by Kimball and Ross (2002).

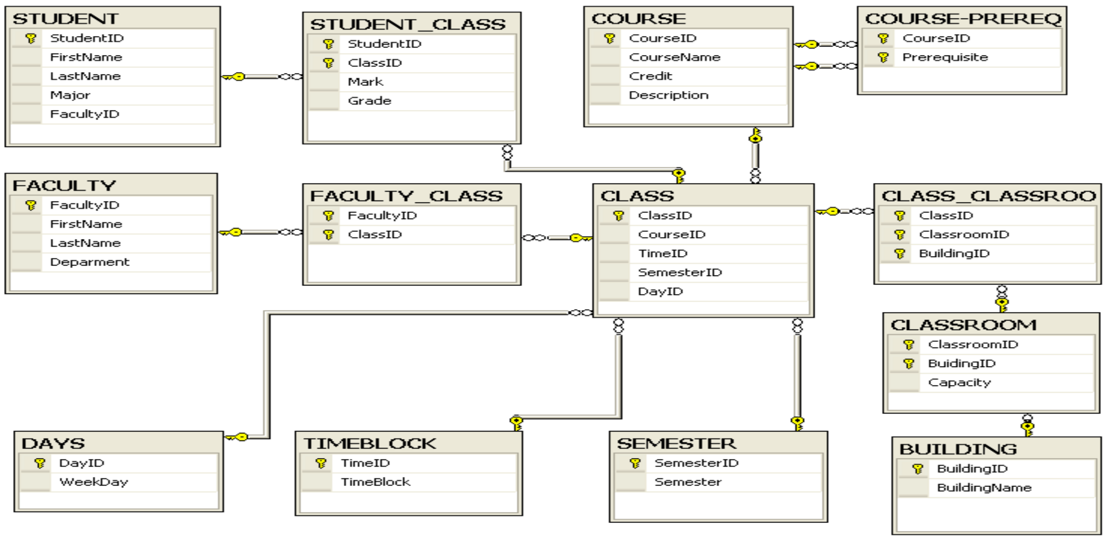

Figure 1. The logical design of OLTP for class registration database

Unlike the Entity Relationship (ER) and Unified Modeling Language (UML) data modeling processes, the logical design of Data Warehouse (DW) is defined by the dimensional data modeling approach. To minimize the join operations which slow down queries, normalization is not the guiding principle in DW design. A schema is a collection of database objects, including tables, views, indexes, and synonyms. There is a variety of ways of arranging schema objects in the schema models designed for data warehousing. The following are the two types of schemas commonly used in dimensional data modeling.

- **Star schema:** The star schema is perhaps the simplest data warehouse schema. It is called a star schema because the entity-relationship diagram of this schema resembles a star, with points radiating from a central table. The center of the star consists of a large fact table and the points of the star are the dimension tables.

  A star schema is characterized by one or more very large **fact** tables that contain the primary information in the data warehouse, and a number of much smaller **dimension** tables (or lookup tables), each of which contains information about the entries for a particular attribute in the fact table.

  Star schema facilitates quick response to queries. The core detailed values are stored in fact table. The dimensional info and hierarchies are kept in dimension tables

- **Snowflake schema:** The snowflake schema is a more complex data warehouse model than a star schema, and is a type of star schema. It is called a snowflake schema because the diagram of the schema resembles a snowflake.

  Snowflake schemas normalize dimensions to eliminate redundancy. That is, the dimension data has been grouped into multiple tables instead of one large table. For example, a product dimension table in a star schema might be normalized into a `products` table, a `product_category` table, and a `product_ manufacturer` table in a snowflake schema. While this saves space, it increases the number of dimension tables and requires more foreign key joins. The result is more complex queries and reduced query performance

Dimensional Data Modeling approach

The dimensional approach is quite different from the normalization approach followed when designing a database for daily operations. Figure 2 shows the five dimension tables used in the project.

Data Hierarchies in Dimensional tables

Each of the dimensions contains at least one hierarchy. The hierarchies allow users to analyze data aggregations using the OLAP. This allows related items to be grouped and summarized for high level analysis while retaining the ability to drill down to more specific product detail.

DimStudent	
PK	***StudentID***
	FisrtName LastName Sex Major Sex

DimFaculty	
PK	**FacultyID**
	FisrtName LastName Department

DimClassroom	
PK	***ClassroomID***
	Classroom BuildingName Capacity

DimSchedule	
PK	***ClassID***
	acadyear Semester WeekDay TimeBlock

DimCourse	
PK	***CourseID***
	CourseName Credit Description

Figure 2: The Dimensional tables showing data hierarchies

for analyzing the data

warehouse. For example we can look for Faculty who taught a particular course.

Fact Table

Fact table contains dimension attributes and measures. Dimension attributes are FKs or other attributes called degenerate dimension (<<dd>>). Measures are the values to be aggregated when queries group rows together. The Fact table is composed of two types of attributes: dimension attributes and measures.

Figure 3 presents the fact table for this research. In the table, the registered field is used as a measure while the other keys are used to link the dimension tables.

The Star Schema:

Data Warehouse are commonly organized with one large central fact table, and many smaller dimensions tables. This

ClassFact Table	
PK	***StudentID***
PK	***CourseID***
PK	***ClassID***
PK	***ClassroomID***
PK	***FacultyID***
	Grade
	Mark
	Registered

Figure 3: The Fact Table

configuration is termed a star schema. A star schema has been adopted in this research. Below is the diagram of the relationship between the fact table and the Dimensional tables.

In the figure 4, the relationships between the fact table and dimension tables are depicted in a star form.

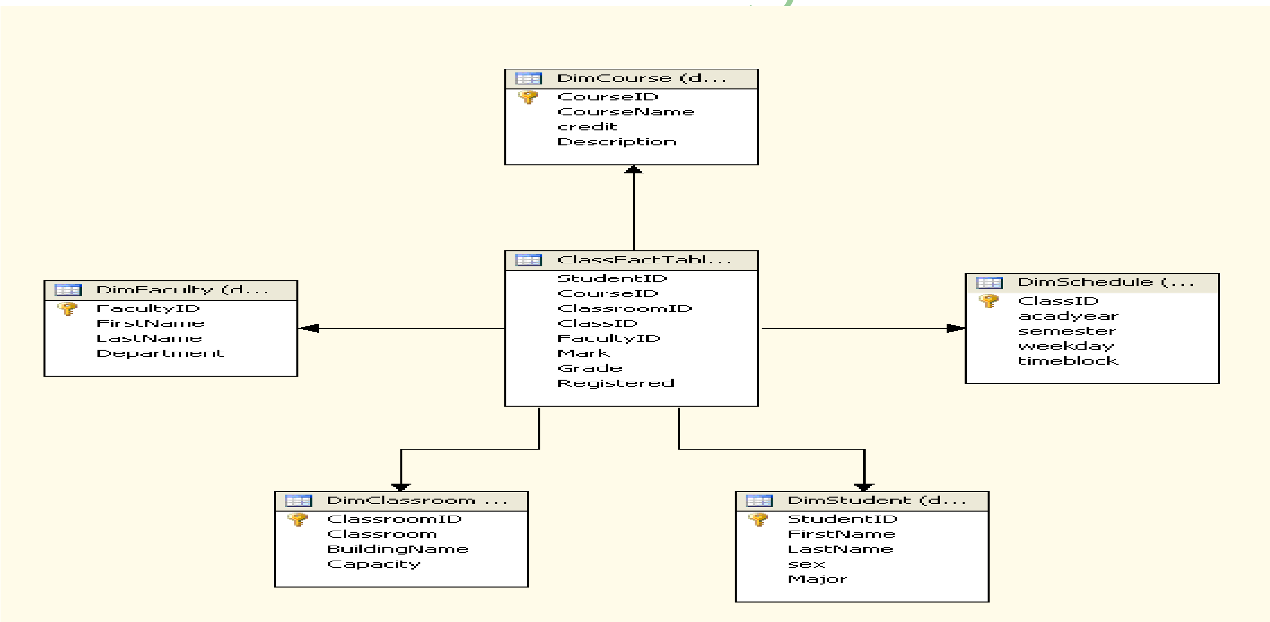

Figure 4. The Star schema

4.0 IMPLEMENTATION

4.1 Data Warehouse

Transporting Data from OLTP Database to Data Warehouse

SQL Server provides the tool SSIS to assist in transporting data in and out of a database. It involves creation of the dimension table's data structure, followed by the creation of integration service project using the Business Intelligence package in the SQL server. After creating the project, a data-source is then created followed by the creation of OLEDB data-source and destination. The following tables show the dimensional tables and fact table created in the course of this project.

a. DimStudent

Table - dbo.DimStudent | Summary

StudentID	FirstName	LastName	sex	Major
1	Ojo	Adewale	M	COMPUTER
2	Aina	Olubukola	F	MATHEMATICS
3	Funmilayo	Olaiya	F	BIOLOGY
4	Fabunmi	Oni	M	CHEMISTRY
5	Kunle	Ajayi	M	Physics
6	Ronke	Dawodu	F	COMPUTER
NULL	NULL	NULL	NULL	NULL

b. **DimFaculty**

Table - dbo.DimFaculty | Table - dbo.DimSchedule | Table - dbo.DimSchedule

FacultyID	FirstName	LastName	Department
1	Kolawole	Olu	Computer Science
2	Ajewole	Ajayi	Computer Science
3	Oni	Dada	Computer science
4	Wale	Ige	Biology
5	Femi	Akinsola	Physis
6	Ojo	Funmilayo	Maths
NULL	NULL	NULL	NULL

c. **DimSchedule**

Table - dbo.DimSchedule | Table - dbo.DimSchedule | Table - dbo.DimStudent | Summary

ClassID	acadyear	semester	weekday	timeblock
1	2009	FISRT	Monday	8-10
2	2009	SECOND	Monday	10-12
3	2009	SECOND	Tuesday	12-2
4	2010	FIRST	Wednesday	2-4
5	2010	SECOND	Wednesday	12-2
6	2010	SECOND	Monday	2-4
7	2011	FIRST	Friday	8-10
8	2011	SECOND	Thursday	10-12
NULL	NULL	NULL	NULL	NULL

d. **DimCourse**

Table - dbo.DimCourse | Table - dbo.DimFaculty | Table - dbo.DimSchedule

CourseID	CourseName	credit	Description
1	COMPUTER ARC...	3	COMPUTER COU...
2	DATABASE MAN...	3	DATABASE
3	SYSTEM ANALYS...	3	ANALYSIS
4	COMPUTER NET...	3	NETWORKS DES...
5	Solid Electronics	3	elctronic
6	Physics	3	physics
7	Solid Electronics	3	elctronic
8	SYSTEM ANALYS...	3	ANALYSIS
NULL	NULL	NULL	NULL

e. DimClassroom

Table - dbo.DimClassroom | Table - dbo.DimCourse | Table - dbo.DimFaculty

ClassroomID	Classroom	BuildingName	Capacity
1	201	ETF	300
2	102	BIG LT OBANLA	300
3	201	ETF	300
4	102	BIG LT OBANLA	300
5	104	ETF	250
6	102	BIG LT OBANLA	300
7	101	SMALL LT OBANLA	200
8	201	ETF	300
NULL	NULL	NULL	NULL

FactTable

Table - dbo.ClassFactTable | Table - dbo.DimStudent | Diagram - FB4...tabaseDiagram | Summary

	StudentID	CourseID	ClassroomID	ClassID	FacultyID	Mark	Grade	Registered
	1	1	1	1	1	91	A	1
	1	2	2	2	2	95	A	1
	1	3	3	3	3	91	A	1
	2	1	1	1	1	88	B	1
	2	2	2	2	2	80	B	1
	2	3	3	3	3	90	A	1
▶	3	1	1	1	1	75	C	1
	3	2	2	2	2	90	A	1
	4	1	1	1	1	90	A	1
	4	2	2	2	2	80	B	1
*	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Fig. 5 Populated Dimensional tables and fact table

The tables 5 a, b, c, d, e and f show the typical contents of the fact and dimension tables after some records have been entered into the tables.

4.2 Online Analytical Process (OLAP)

After Data Warehouse has been populated with data, the next step in the data warehouse development is to provide the users with data analysis tool such as OLAP and Data Mining to analyze the business process for decision making. OLAP is a service that automatically selects a set of summary views (tables), and saves these summary views to disk. OLAP also manage these views and update them when the fact table has new data. To create this, we must first create Analysis Services project using Business Intelligence Development Studio. Then, we define the data source, create the data source view and then create Cube. Figure below shows a screen shot of the OLAP cube created with SQL Server.

Figure 6 shows how a star schema described earlier is transformed into an internal form, which SQL business intelligence can use to fetch appropriate data requested by the users.

4.3 The Data Analysis Reports

The reports blows are the output results obtained as a result of roll-up and drill-down operations of OLAP performed on the data for various hierarchies of the dimensions.

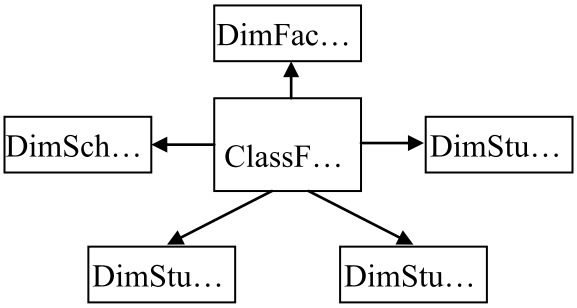

Figure 6: Class Registration

Drop Filter Fields Here

	Course Name			
	COMPUTER ARCHITECTURE	DATABASE MANAGEMENT	SYSTEM ANALYSIS AND DESIGN	Grand Total
Sex	Class Fact Table Count	Class Fact Table Count	Class Fact Table Count	Class Fact Table Count
F	2	2	1	5
M	2	2	1	5
Grand Total	4	4	2	10

Figure 7. Data Analysis based on student gender and courses registered.

Figure 7 shows list of students' sex and by course. It is interactively generated by OLAP based on users' choice.

Registration by academic Year

Drop Filter Fields Here

	First Name			
	Ajewole	Kolawole	Oni	Grand Total
Acadyear	Class Fact Table Count	Class Fact Table Count	Class Fact Table Count	Class Fact Table Count
2009	4	4	2	10
Grand Total	4	4	2	10

Figure 8 Data Analysis based on Academic Year and course registered by students

This figure shows count of students taught by lecturer in academic year 2009. It is interactively generated by OLAP based on users choice.

Lecture Room by Time-block showing Number of students by Courses

Drop Filter Fields Here

		Course Name			
		DATABASE MANAGEMENT	COMPUTER ARCHITECTURE	SYSTEM ANALYSIS AND DESIGN	Grand Total
Building Name	Timeblock	Class Fact Table Count	Class Fact Table Count	Class Fact Table Count	Class Fact Table Count
BIG LT OBANLA	10-12	4			4
	Total	4			4
ETF	12-2			2	2
	8-10		4		4
	Total		4	2	6
Grand Total		4	4	2	10

Figure 9 Data Analysis based on Lecture rooms, time tables versus Courses registered

This figure shows Lecture rooms allocation by course and number of students participating in that course. Such a report can be used for effective lecture theater allocation for course lecturing or examination. It is interactively generated by OLAP based on users' choice.

4.4 The Data Mining

(Discovering Hidden knowledge using Data Mining). Data mining is a process to have the computer search for correlations in the data, and present promising hypothesis to users for consideration. The problems that can be solved by data mining are:

(a) Categorization: Classify a given set cases

(b) Clustering: Find the natural groupings of a given set of cases.

(c) Association rule: Find out what items are frequently processed together.

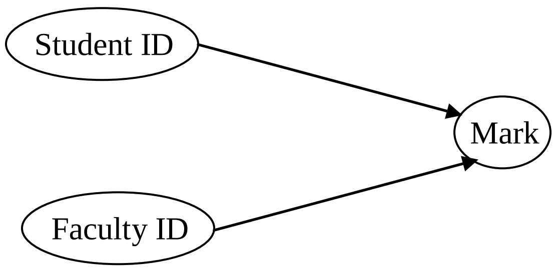

Figure 10: Data Clustering Mining that finds Relationship between Students Marks and Faculties

This figure shows the result obtained from the data-mining model above. It shows class group of Mark 90 and group of mark 75. The class a student favours between mark 90 to 75 and the degree of membership of such a class is shown by the bar. Also shown are faculties and the degree of their membership in awarding marks in the range of 90 and that of group of 75 and the degree of their membership of such a group.

A data-mining that clusters students and Lectures showing the class a student favors' between mark 90 to 75 and the degree of membership of such a class is shown by the bar.

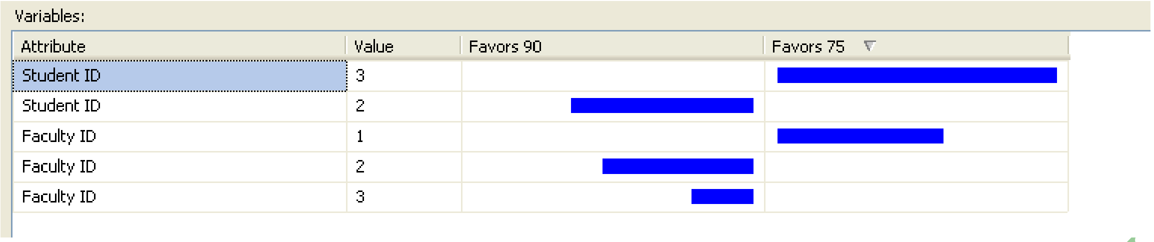

Figure 11: Discriminant mining showing Relationship between Students Marks and Faculties

5.0 CONCLUSION

In this paper, we have been able to demonstrate the process of designing and developing data-warehouse and data mining applications using SQL server business intelligence Development tools using a case study in an academic environment. It is to be noted however that this technique can be applied to any organization wishing to implement business intelligence as part of their strategic decision support operations. The power of Data-Warehousing in data analysis is tremendous and data-mining can discover hidden treasures in the data-warehouse.

Organizations, particularly in Nigeria can begin to implement this project as part of their strategic decision making process tools. With it they can begin to see trends of things historically as their day to day operational data accumulates over the years. They can forecast the future using neural Networks, regression analysis, and other data mining operations incorporated into SQL server Business Intelligence. Data-warehouse is the solution we need at the moment to catapult our business or our organization to the next level.

6.0 REFERENCES

Efraim T., Jay E., Teng-Peng L., Ramesh S, (2010). Decision Support and Business Intelligence Systems (8th ed) Prentice Hall. Infogold Data Warehouse, Data Mart, Data Mining, and Decision Support Resources, http://infogoal.com/dmc/dmcdwh.htm.

BIN (2007) Business Intelligence Network, http://www.beye-network.com/home/

Bill Palace (1996). Data Mining Technology Note prepared for Management 274A Anderson Graduate School of Mana-gement at UCLA http://www.ander-son.ucla.edu/faculty/jason.frand/teacher/technologies/palace/index.htm.

David Heise: An online publication on Data Warehousing in higher Education, http://dheise.andrews.edu/dw/DWData.htm

Fang, R. and Tuladhar, S. 2006. "Teaching Data Warehousing and Data Mining in a Graduate Program in Informa-tion Technology," Journal of Com-puting Sciences in Colleges, Vol. 21, Issue 5, pp. 137-144.

Kimball R. and Ross M. (2002). The Data Warehouse Toolkit, second edition. John Wiley and Sons, Inc., USA.

Michael A. King (2009). A Realistic Data Warehouse Project: An Integration of Microsoft Access and Microsoft Excel Advanced Features and Skills Journal of Information Technology Education Vol. 8, 2009

Innovations in Practice.
Pierce E.M. (1999). "Developing and Deli-vering a Data Warehousing and Data Mining Course," Communications of the AIS, Vol. 2, Article 16, pp. 1-22.
Jacobson R. (2000). Microsoft SQL Server (2000). Analysis Services, Step by Step; Microsoft Press, Redmond, Washington.
Stephen Brobst, Joe Rarey (2003). Five Stages of Data Warehouse Decision Support Evolution http://dssre-sources.com/papers/features/brobst&rarey01062003.html.
Wierschem D., McMillan J. and McBroom R. (2003). "What Academia Can Gain from Building a data Warehouse," Educause Quarterly, Number 1, pp. 41-46.