

CASE PROJECTS IN DATA WAREHOUSING AND DATA MINING

Mohammad A. Rob, University of Houston-Clear Lake, rob@uhcl.edu
Michael E. Ellis, University of Houston-Clear Lake, ellisme@uhcl.edu

ABSTRACT

This paper briefly describes the process of development of several case projects through offering a graduate level course on Data Warehousing and Data Mining. It then outlines a particular case project that describes the process of data extracting, data cleansing, data transfer, data warehouse design and development. It also outlines the development of a data cube as well as application of OLAP tools using Microsoft SQL Server Analysis Services to understand business intelligence. The results can be beneficial to an instructor who wants to develop a practical course or a practitioner venturing into the data warehousing and data mining area.

Key Words: Data warehouse, Business intelligence, BI, OLAP, Data mining, Case, and Course.

INTRODUCTION

In recent years, data warehousing and data mining became a significant part of many organization's IT infrastructure. The historical data in the warehouse play an important role in providing Business Intelligence (BI) that helps companies to streamline workflows, provide better customer services, and target market their products and services. Software development companies are also focused on developing new tools and technologies for data warehousing engines, providing data transfer services from traditional sources to data warehouses, performing analysis for business intelligence, generating reports and ad hoc queries, and executing data mining algorithms. Large companies such as Microsoft, Oracle and IBM have significantly modified their flagship databases such as SQL Server, Oracle and DB2 to accommodate some of these necessities. Other companies having significant market share of the data warehousing and BI tools are SAP, SAS, PeopleSoft, Crystal Reports, Business Objects, Teradata, Informatica, DataMirror, Hyperion, and Cognos [1-2].

Many IT consulting companies help large companies develop and maintain their data warehousing and BI efforts (see for example, www.mindtree.com). Demand for personnel with specific IT skills in the data warehousing and BI technologies has also been

growing. Today, a search for data warehousing, BI, or data mining returns thousands of jobs scattered across the nation [www.careerbuilder.com, www.monster.com, www.dice.com].

The use of data warehousing and BI technology span sectors such as retail, airline, banking, health, government, investment, insurance, manufacturing, telecommunication, transportation, hospitality, pharmaceutical, and entertainment [3]. Due to increasingly stringent budgets, rising operational costs, and competition from online universities, many educational institutions recently adopted data warehousing and BI technology to improve their business processes [4]. Universities are using BI tools in areas such as academics, enrollment, financial aid, alumni, development, finance, and human resources. Wierschem, et al. [5] discuss many important issues upon which universities can focus their data warehousing efforts.

While almost all business sectors, government agencies, and academia moved into adopting data warehousing and BI tools, and there are significant demands for skilled personnel in these areas, the faculty members in computing and MIS programs that are expected to teach the knowledge and skills necessary to prepare their students for the rising job market are lagging behind. Although a systematic research has not been done to find out how many universities offer such a course, a simple search of the web or journal databases reveals very few course offerings or papers in relation to teaching data warehousing and/or data mining [6-8]. Similarly, finding a text book that can be adopted for such a course provides a gloomy picture; few of the books available in the market are of research or tool-specific in nature.

We have developed a graduate course on data warehousing and data mining that provides students with both the theoretical knowledge as well as hands-on practice with data warehousing tools and techniques. In the following, we provide a brief outline of the course as well as the teaching methodology adopted for the course. Since the main focus of the course is for students to apply the knowledge gained in the course to a case project using a specific data warehousing tool, we will

mainly focus on the description of the development of such a case project.

THE COURSE OUTLINE AND TEACHING METHODOLOGY

Course Purpose

The purpose of the course is to acquaint students with fundamental knowledge of data warehouse modeling and design; the tools and techniques of data analysis using OLAP (Online Analytical Processing) techniques; to acquaint students with data mining concepts, techniques and query language; and to prepare students for future careers in data warehouse planning, analysis, design, and implementation. Specifically the course provides an understanding of the concepts and knowledge in the following key areas:

- The difference between a database, a data mart, and a data warehouse; operational versus decision support systems; the difference between Online Transactional Processing (OLTP) and Online Analytical Processing (OLAP).
- The data warehouse architecture and environment; the concepts of facts, dimensions, and hierarchies; the Dimensional Model versus Entity-Relationship Model; data warehouse modeling with Star- and Snow-flake schema.
- The concepts of cube and OLAP processing through roll-up, drill-down, slicing, dicing, and pivot table. The ETL process – the concept of data extraction, transformation and loading; data mining techniques such as classification, clustering, estimation and prediction.
- Designing and developing a data warehouse as well as applying ETL, OLAP, and data mining tools on cubes using Microsoft SQL Server Analysis Services.

Key Components of the Course

The course mainly focuses on three areas: (i) development of data warehousing and data analysis concepts through lecture notes, textbooks, and other online resources; (ii) writing and presenting two research papers; and (iii) development of a case project using the Microsoft SQL Server Analysis Services as well as making three group presentations and submitting a final report on the project. The required textbooks for the course are, “Data Warehousing Fundamentals” by Paulraj Ponniah [9] and “Microsoft SQL Server 2000 Analysis Services

Step by Step” by Reed Jacobson [10]. Other reference books used for the course are found in the references [11-12]. The course is offered once a year and the topics for research papers vary. During the summer of 2006, they were focused on understanding some of the concepts of data mining such as classification, clustering, estimation and prediction. The course website (<http://mis.uhcl.edu/rob/Course/DW/DW.htm>) hosts completed student papers and presentations.

Developing the Case Project

Students are required to work in groups to model, design, and develop a data warehouse. Each group needs to choose a business industry such as airline, education, retail, financial, insurance, hospitality, investment, and healthcare. They will then collect day-to-day business data found in files, databases, spreadsheets, or text documents in various data formats. Next they will develop a strategy to convert these data into a common format as well as in summarized form. They will then design and develop a multi-dimensional data warehouse and transfer the data to the warehouse. Using the data, they will then develop a multidimensional cube using the SQL Server Data Analysis Services. Finally, students will apply OLAP and data mining tools to extract meaningful business intelligence on customers, products, purchases, and so on. Further details of the case project requirements are outlined in the course web site.

Presentations on the Case Project

Each student group is required to make three presentations during the various stages of the project development. The presentations follow a schedule and they cover topics such as: (i) project definition and planning, (ii) logical and physical design of the data warehouse, and (iii) results of the cube implementation through applying OLAP and data mining tools. Further details of the presentation requirements can be found in the course web site. Note, about 50% of the students in the class worked in various industries in and around the city of Houston. This experience allowed almost all student groups to work on real-life cases. All presentations as well as the final report for each case project can be found in the course website. The rest of the discussion is focused on a particular case project.

THE CASE PROJECT OF HVAC/R WHOLESALE COMPANY

Business Scenario

The type of business used as a model for the project is a Heating, Ventilation, Air Conditioning and Refrigeration (HVAC/R) wholesale company. It is modeled on a former employer of one of the team members of a student group. The 50-year-old company has about 260 employees divided between its 215,000 square foot distribution center and 30 branch locations around the state of Texas. The branch locations are setup with a small showroom containing small items on shelves, like an auto parts store. Overstock and larger items are in the warehouse. Customers can self-serve for smaller items, but must be helped for items in the warehouse. All items are invoiced at the counter in the branch location.

The company sells equipment, parts, and supplies to licensed contractors who perform work on heating, air conditioning, and refrigeration systems. It sells approximately 16,000 line items that come from 2,000 stocking manufacturers. These sales are processed by an online transaction processing (OLTP) system. A simplified representation of the existing OLTP system is shown in Figure 1. As invoices are generated, the individual invoice line items are stored in the Sales table.

When more than one item is sold for an invoice, an invoice line item is stored in the Sales table, with the Invoice Number, Date, and Invoice Line Number as a composite primary key. This composite key is also necessary because invoice numbers are recycle every year. Without the Date field as part of the key there would never be any way to tell which invoice you were referring to, invoice 150088 from 1998 or invoice 150088 from 2003.

The other tables in Figure 1, the Products, Customers, and Stores, allow coherent reporting of current sales data through joins between the Sales table and the other three tables as needed to provide descriptive data about products, customers, and branch locations.

Why a Data Warehouse?

The day-to-day operations of the company rely heavily upon the OLTP system. Everyone from the purchasing department to branch managers to accounts receivable rely on it for current information to make normal operational

decisions. For the most part, it does what it is supposed to do – it tells users what is currently happening in the company.

Yet the OLTP system is painfully inadequate when it comes to strategic decision support. These types of information requests from management must be dealt with individually by the information technology (IT) staff. Data aggregations are programmed into reports, but any comparisons across time or products must be done manually. Data history in the OLTP system typically goes back 2 years, even though the company has been generating computerized data for over 20 years.

A data warehouse is the solution. It will provide a central repository for historical data. It will provide an integrated platform for historical analysis of sales data. It will allow the application of online analytical processing (OLAP) techniques by users themselves. With a data warehouse and OLAP, we expect to empower users to perform their own roll-up and drill-down operations to analyze sales across product categories, subcategories, store regions, individual stores, or any combination desired. They will have the flexibility to view data and immediately look at data in another form without sending a request to IT for a new report. They will enjoy a true decision support system that will provide strategic analysis in a user-friendly format.

Expected Challenges

There are some technical challenges that need to be considered. We mentioned previously that the invoice numbers are recycled every year. Each

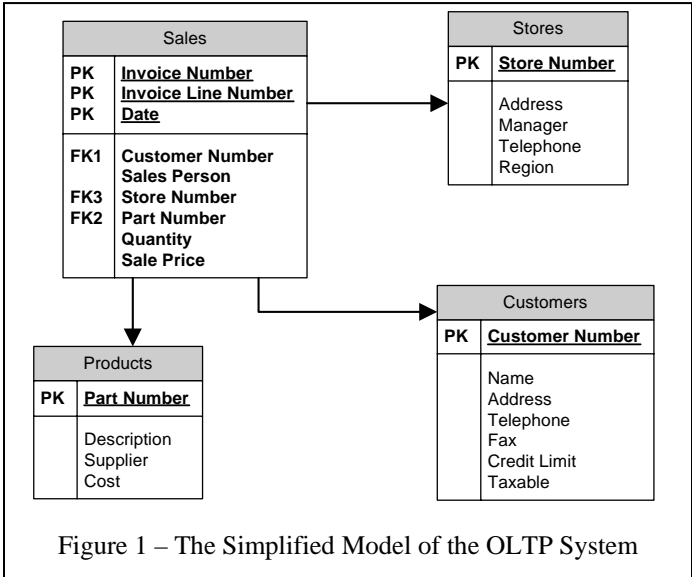


Figure 1 – The Simplified Model of the OLTP System

invoice number also has the store number embedded within it. For example, invoice 150088 was written at store number 15. It is a handy feature if someone is tracking down invoices, but it is a problem for a data warehouse. Product part numbers also contain embedded information. The first three digits of the seven-digit part number are used to identify a category or supplier. For example, part numbers 3411208, 3416442, and 3417880 are all electrical parts. Parts 7602392 and 7601008 are manufactured by Sporlan Valve Company. This must be resolved as we build our data warehouse model.

Designing the Data Warehouse

The following discussion outlines the process of our data warehouse design.

Dimensional Modeling

The tables shown in Figure 2 are the tables that make up the current OLTP system. To create a data warehouse system we must determine how we are going to extract meaningful data and logically group the data. To that end, we prepared the information package shown in Figure 3. The information package

Information Subject: Sales				
Dimensions				
	Time	Product	Customer	Store
Hierarchies/Categories	Year	Category	License type	State
	Quarter	Subcategory	Category	Region
	Season	Product name	Size	City
	Month		Customer Name	Square footage
	Date			Store name
	Day of Month			
	Day of Week			

Facts: Sales quantity, Item dollar amount, Item cost

Figure 2 – Information Package

allows the data warehouse’s designers to layout the requirements for the dimension tables, their hierarchies, and the facts to be modeled.

Dimension Tables

The information package was then used to create the dimension tables shown in Figure 3. The Time, Product, Customer, and Store dimensions are each used to create dimension tables. The dimension hierarchies will be discussed in the next section.

Fact Table

The fact table used for this project was based on sales information. The table contains ten thousand randomly generated sales records for 500 dates beginning on January 1, 2005. Figure 4

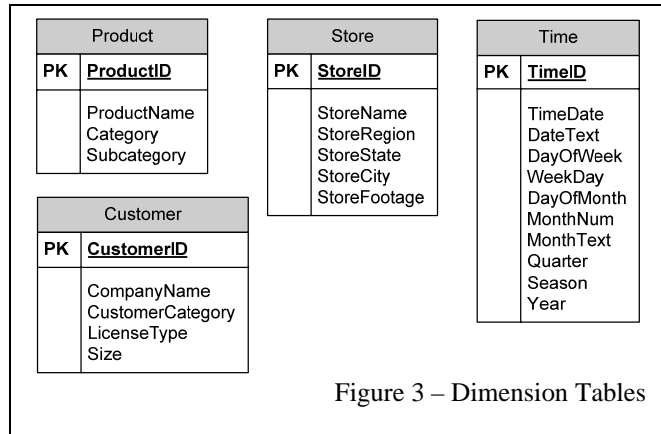


Figure 3 – Dimension Tables

shows the fact table design.

Data Hierarchies

Each of the dimensions contains at least one hierarchy. The hierarchies allow users to analyze data aggregations in a very straightforward manner using the OLAP functions of Analysis Services. The Product hierarchy is shown in Figure 5.

All 16,000 products are grouped into categories and each category if further divided into several subcategories. For example, 2” gray duct tape is within subcategory “Tapes” within category “Supplies.” 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.

The Store hierarchy simply allows individual store data to be rolled up into a regional designation, and then into a state name. Currently all stores are in Texas, but the possibility exists that the company could expand into neighboring states. The state

SalesFact	
PK	ProductID
PK	TimeID
PK	StoreID
PK	CustomerID
	Quantity
	PricePerUnit
	CostPerUnit

Figure 4 – Fact Table

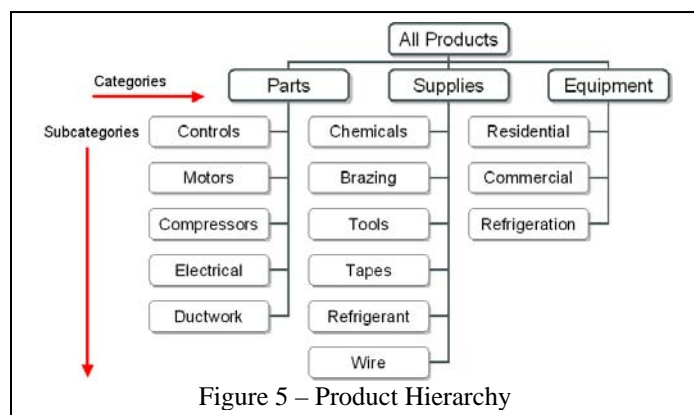


Figure 5 – Product Hierarchy

aggregation level is included now to make any future changes easier to implement.

Customers are included in three hierarchies within the Customer dimension, as shown in Table 1. It is important to note that each customer will be found in each of the three hierarchies. This provides three different ways to look at summarized customer data.

Looking at the Time dimension should be straightforward, until you come to the “Season”

attribute. Because of the nature of the business, seasonal sales differences can be an important analytical topic. A large percentage of sales occur during the summer, but there are also considerable heating sales in the winter and new construction year around. Since this provides two different ways to aggregate according to time periods, there are two hierarchies within the Time dimension, as shown in Table 1.

Table 1: Hierarchies in Customer and Time dimensions				
Customer			Time	
Size	License Type	Category	Calendar	Seasonal
Small (2-3 techs)	“A” license (HVAC and refrigeration)	HVAC	Year	Year
Medium (4-10 techs)	“B” license (HVAC only)	Builder	Quarter	Season
Large (11-20 techs)		Government	Month	Month
Corporate (21+ techs)		Refrigeration	Day	Day of the week
		Maintenance		

Database Schema

Since the dimension tables were not normalized and the size of these dimensions was not too large, the STAR schema was implemented. The STAR schema is shown in Figure 6. The STAR schema was also selected because it provides an intuitive design that can more readily be understood by users.

Database Implementation

While much of the data could be readily transformed into the tables created for the STAR schema implementation, some data cleansing was required.

Data Cleansing

As discussed earlier in the Technical Challenges section, both invoice numbers and part numbers pose a problem in that they contain embedded information. Figure 7 shows an example of the information embedded in a part number. So as part of the data cleansing process, these embedded identifications were replaced by integer keys in the dimension tables and the corresponding records of the fact table. The Product Part Number is then replaced with a new integer value.

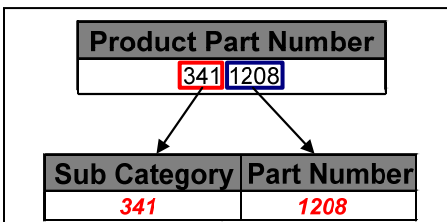


Figure 7 – Embedded information

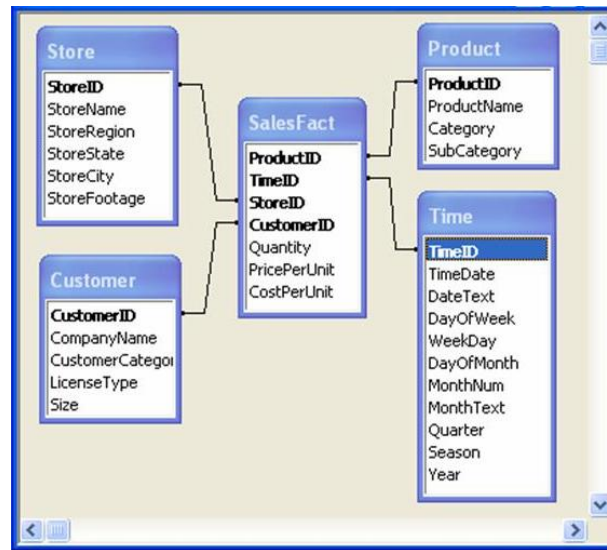


Figure 6 – The STAR Schema

Dimension Implementation

Figures 8 shows the Customer dimension table along with some actual data implemented in Microsoft Access. Other dimensional tables are similarly populated with necessary data.

Fact Implementation

The SalesFact table is implemented according to the design (Figure 6) and sales data were populated according to the discussion before. All four fields of the primary key are shown (Figure 9). Note that TimeID values begin with 3 instead of 1. TimeID = 1 is 1/1/2005, a Saturday, and TimeID = 2 is

1/2/2005, a Sunday. All stores were closed on weekends and there were no sales.

Cube Implementation and OLAP

Figure 10 shows how the dimensions and fact table are implemented in a cube using SQL Server 2000 Analysis Services. Jacobson [10] outlines the process of developing a cube in SQL Server from Microsoft Access database. Figure 11 shows a screen shot of the result of an OLAP operation on the cube, displaying facts for customers in various categories. The final report in the course web site provides results of roll-up and drill-down operations of OLAP performed on the data for various hierarchies of the dimensions.

Three classifications

Customer ID	Company Name	CustomerCategory	LicenseType	Size
1	Keep Kool Air	HVAC	B	Small
2	Jim Thompson Air	HVAC	B	Small
3	Bobcat Heating & Cooling	HVAC	B	Small
4	Carl's Refrigeration	Refrigeration	A	Small
5	Addicks Air	HVAC	B	Medium
6	Cool Breeze Air	HVAC	B	Medium
7	Air Masters	HVAC	A	Medium
8	Becker & Sons Air	HVAC	B	Medium

Figure 8 - Customer dimension

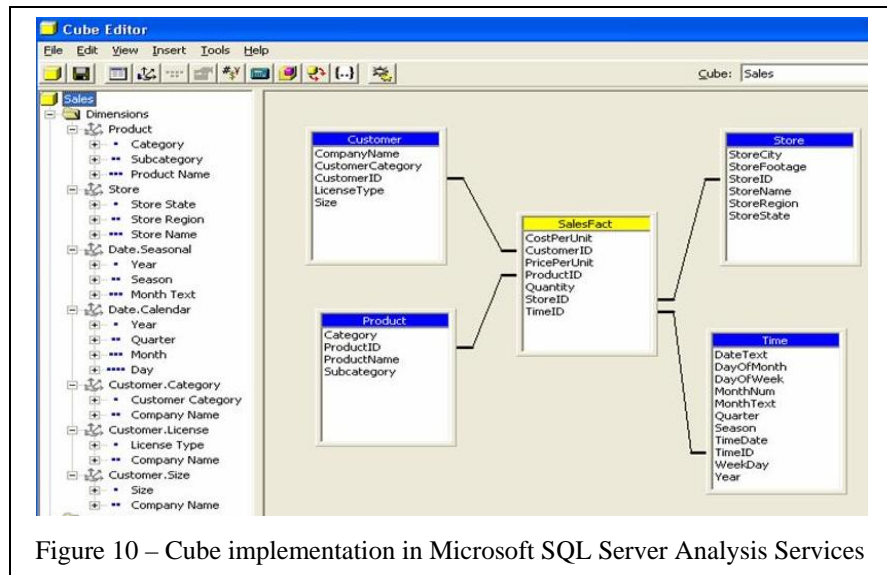
SalesFact : Table

Product ID	TimeID	StoreID	CustomerID	Quantity	Price Per Unit	Cost Per Unit
5	3	5	9	2	20.86	18
43	3	14	3	2	10.38	8.12
57	3	1	1	1	334	272.56
57	3	8	6	1	334	272.56
42	3	11	14	7	22.05	17.42
24	3	16	17	5	23.43	17.1
69	3	15	12	1	327	268

Figure 9 – The SalesFact table implementation

Conclusion and Discussion

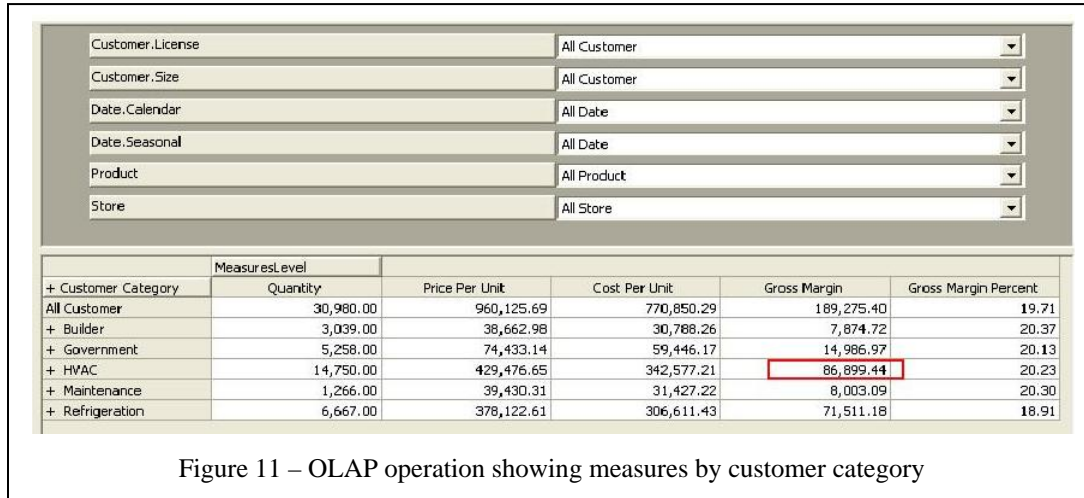
As mentioned earlier, many of the executives within this privately held company are very skeptical towards new technology and resist change. However, even someone with this mindset is impressed when they are given a new set of tools that gives them the ability to make more informed decisions. The power of the OLAP tools alone implemented in this project would make a very persuasive argument for the implementation of a full-scale data warehouse.



With management buy-in also comes new ideas for aggregation levels that can be added or modified to fit the user’s analytical needs. This may be the only way to incorporate views of the data that have been buried in obscure reports or that have possibly never been implemented before due to their complexity. We would include more dimensional attributes and actual data to enable a meaningful data mining effort, which we have tried rather unsuccessfully with the current data.

REFERENCES

1. Data Warehouse, Data Mart, Data Mining, and Decision Support Resources, <http://infogoal.com/dmc/dmcdwh.htm>, downloaded from the web, February 8, 2007.
2. Gray, P. and Israel, C. (1999), “The Data Warehouse Industry,” A research report from the Center for Research on Information Technology and Organization, University of California, Irvine, California.



3. Business Intelligence Network, <http://www.b-eye-network.com/home/>, downloaded from the web on February 2, 2007.
4. Data Warehousing in higher Education, <http://dheise.andrews.edu/dw/DWData.htm>, downloaded from the web on February 4, 2007.
5. Wierschem, D., McMillen, J. and McBroom, R., (2003), "What Academia Can Gain from Building a data Warehouse," *Educause Quarterly*, Number 1, pp. 41-46.
6. Fang, R. and Tuladhar, S. (2006), "Teaching Data Warehousing and Data Mining in a Graduate Program in Information Technology," *Journal of Computing Sciences in Colleges*, Vol. 21, Issue 5, pp. 137-144 .
7. Pierce, E. M. (1999), "Developing and Delivering a Data Warehousing and Data Mining Course," *Communications of the AIS*, Vol. 2, Article 16, pp. 1-22.
8. Slazinski, E. D. (2003), "Teaching Data Warehousing to Undergraduates – Tales from the Warehouse Floor," *CITC'03*, October 16-18, Lafayette, Indiana, pp. 242-248.
9. Ponniah, P. (2001), Data Warehouse Fundamentals, a Comprehensive Guide for IT Professionals; John Wiley & Sons, New York.
10. Jacobson, R. (2000), Microsoft SQL Server 2000 Analysis Services, Step by Step; Microsoft Press, Redmond, Washington.
11. Inmon, W. H. (1996), Building the Data Warehouse, Second Edition; John Wiley & Sons, New York.
12. Berry, M. A. and Linoff, G. S. (2004), Data Mining Techniques, Second Edition; John Wiley & Sons, New York.