

Data Warehousing and Data Mining: A Course in MBA and MSIS Program from Uses Perspective

Shamsul Chowdhury

Joseph O. Chan

Walter E. Heller College of Business Administration

Roosevelt University

schowdhu@Roosevelt.edu

ABSTRACT

Over the past years data warehousing and data mining tools have evolved from research into a unique and popular business application class for decision support and business intelligence. The paper presents and discusses the main components and contents of a course in data warehousing and data mining offered in the graduate business curriculum of an American university in the Midwest. Further, it also discusses how the course was delivered. The course has been designed from lessons learned in research and practice in the field. Key to the development and success of the course is an on-going collaboration between a large company in the retailing branch and the university. Collaboration with the business world has proven invaluable in building bridges between the academic institution and the real world. However, teaching a course in a dynamic field like this will always remain a challenge.

INTRODUCTION

The course in data warehousing (DW) and data mining (DM) presents the necessary fundamentals of DW and DM (methodology, tools, techniques, systems and terminology) to students by putting these concepts into context and comparing expert views in these areas through seminars, discussions, and hands-on-work in computer labs. The prerequisite for the course is a graduate course in Information Resource Management. At least half of the course participants had also taken a graduate course in Database Systems before taking this course and had the skills of ER modeling, normalization, SQL and some other basic DBMS skills. For the required project teamwork, students were split into groups represented by at least a member with fundamental database skills gained from the database system course or gained from real life work experiences. Many of our graduate students are non-traditional and have extensive work experiences in business/information technology or both. Some students also have experiences in data warehousing, ETL (Extract, Transform and Load) and data warehousing project management. The course is delivered in the form of lectures, group discussions, teamwork and seminars where participants are required to actively participate both in presentation and discussions and investigate agreed upon topics.

Besides a number of reference books, articles and web resources, two textbooks are used in the course: "The Enterprise Data Warehouse: The Planning, Building & Implementation" by E. Sperley (1999) and "Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management" by Michael J.A. Berry, Gordon S. Linoff (2004), respectively for DW and DM.

COURSE COMPONENTS AND LEARNING OBJECTIVES

The main purpose of the course is to develop and gain an understanding of the principles, concepts, functions and uses of data warehouses, data modeling and data mining in business. A DW and DM project is usually business-driven and will work to improve the direction of the company by aligning the data warehouse technology with business strategy.

The following areas of interest are addressed in the course:

- DW methodology
- DW architectures
- The DW development processes: Logical and physical DW
- DW data modeling
- ETL, Data access, Data quality
- DM - Query tools
- State-of-the-art in DM tools and technologies
- From research to a mature technology – technological artifact
- DM and Business intelligence – From findings to application
- DW, DM and beyond.

In sections below we will present and discuss the course contents and how it was delivered in some detail.

Data Warehousing Methodology

Data warehousing methodologies, as described by Zen et al. (2005), share a common set of tasks, including business requirements analysis, data design, architectural design, implementation and deployment. The important concept of the separation of conceptual and physical design in systems development is emphasized in the course. Contemporary methodologies are discussed. Typically they contain the framework that consists of requirements analysis, conceptual design, physical design, development and testing, implementation and deployment (Chan 2007). In the following, a description of contents for each phase is presented.

Requirements Analysis: For the topic of requirements analysis, the concept of analytical requirements and their differences from operational requirements are introduced. Key concepts in analytics that deal with forecasting, projection and formation of strategies are explored. The course points out the differences between data and function analysis in DW and DM. Data gathering techniques such as user interviews, joint application design (JAD) are explored and practiced by doing project work in small groups.

Conceptual Design: For the topic of conceptual design, modeling techniques specific to DW are discussed. They include the Entity-Relationship (ER) modeling introduced by Chen (1976) and dimensional modeling including the Star Schema and Snowflake Schema, which utilize fact tables and dimension tables (Todman 2001). Strategies in modeling with regard to the data warehouse ER model, the data warehouse dimensional model and the independent data marts as described by Jukic (2006) are explored and practiced in the classroom setting. The concept of enterprise modeling as described by Chan (2004) as a conceptual framework in building data warehouses is studied. Using the enterprise-modeling framework, the model of the requirements for analytical functions is developed in conjunction with the development of the data model. Data sourcing strategies and the logical mapping of the data schema of sourcing systems to the conceptual model can be developed during the conceptual design phase.

Physical Design: For the topic of physical design, various levels of the physical design for the data warehouse are explored. They consist of three levels: the data level, the application level and the technical infrastructure level. Topics for the data warehouse database design include the general database design principles and specific data warehouse considerations such as de-normalization. Specific topics for database design with regard to the choice of databases such as Oracle, DB2, MS SQL Server, PostGres and MySQL can be included. The data level design also includes the design of data extraction, transformation and loading (ETL). Special topics in the ETL tools can be included. Application level design includes the selection of development tools for custom built analytic applications, packaged software, analytic and reporting tools. Various topics in SQL, statistical modeling, online application processing (OLAP) and data mining are explored. At the technical infrastructure level general concepts of technical architecture for data warehousing include requirements in hardware, software and networking are discussed. The design phase also includes development of user interfaces (UI) and considerations of scalability in terms of both the growth in the number of users and the increase of use by each user.

Development and Testing: For the topic of development, concepts and techniques in the creation of databases and applications in a development environment are introduced. They include the creation and population of databases

and the development of ETL routines, user interface, analytic applications, reports, system and application interfaces. Topics of unit testing, system testing and performance testing are studied. Special topics in development sourcing strategies can be included.

Implementation and Deployment: For the topic of implementation, different deployment strategies are discussed. They include the big-bang approach and various phased approaches. In the big-bang approach the data warehouse is deployed to the entire organization with all functionalities all at once. In the phased approach, the data warehouse can be deployed by phases based on various criteria such as geography, organizational units or data warehouse functions. Other topics for implementation include system installation and conversion in the production environment, user documentation, user training, system support, and performance tuning and system administration.

Data Modeling

For the data modeling part of the course the starting point is to understand the basics of ER modeling and also its limitation for creating an enterprise wide data model for decision-making purposes. The Entity Relationship (ER) diagram is commonly being used to create transaction-oriented relational database system to run the day-to-day operation of the business. The STAR or dimensional modeling is used for creating data warehouses (DW). Idea behind a data warehouse is to centralize company wide information to create and deliver the necessary analytical environment, for example data mining and business intelligence; to meet the business needs.

Three methods are studied for dimensional modeling (Sperley 1999): development by modification; development from template; and, full custom development. In development by modification the existing ER model is modified and reorganized in the form of STAR model by deleting the operational information and adding more of decision making information to accommodate the decision making needs of the business. Development by modification was found to be useful and is becoming a common acceptable (best) practice. Success with this approach depends on factors, such as: the ER models are relatively new, well designed and normalized; the IT staff (for the new DW project) is very familiar with the source system OLTP models; the IT staff also understands the data that is already in the database; and, many DSS (Decision Support System) data warehouse reference and fact tables are already in the OLTP model in a different form.

The use of an accepted methodology provides big advantages in the conversion of the ER model to the STAR model in DW creation. The course addresses the ways and means of the conversion processes providing useful guidelines based on the Ten Commandments of Dimensional Data Modeling. The ten commandments summarizes as to what needs to be done for a successful and useful conversion (Sperley, 1999). The course follows them into practical realization in the form of project works in small groups and has been described elsewhere (Chowdhury, 2007c).

Conversion of the ER diagram to a star schema is a step-by-step process and usually begins with a conversation between end users and IT professionals to discover a list of burning questions. This list consists of a set of analytical problems that decisions makers feel are critical success factors to the future direction of the business. For example, a burning question in a retailing business could be in the form of “who are the top 10% customers and what do they purchase and where do they live”? The retailing business may want to provide the top customers with discounts as an incentive to maintain a long lasting relationship with them. STAR model should be structured to easily find answers to these kinds of analytical questions, which could bring competitive advantages to the business by retaining those customers (Chowdhury, 2007a). In the course we used the ER model of the Northwind Traders Database from Microsoft, implemented in ACCESS relational DBMS. Data modeling is as much art as science; many possibilities exist for a successful conversion. The students are divided into a number of smaller groups and are given the task of studying the Northwind OLTP database and create a STAR model based on each group’s understanding of the data warehouse requirement from business perspective. The groups came up with different STAR models, which are being presented for the whole class. We further studied these alternative STAR models, showed their relative strengths and weaknesses and suggested an optimal model for the business concern to gain and retain competitive advantages. Besides working with ACCESS we have also introduced Oracle Warehouse Builder for data warehousing. We would like to introduce more tools like this in the future.

This group works by the students gave them a better understanding by sharing, comparing and discussing one another's solution as to how an ER model could be converted to a dimensional model or the STAR model. In fact the process acted as a catalyst to enhance student learning by actively doing and participating.

ETL

Decision-making data are extracted from OLTP source and further organized as per fact/s or burning question/s for decision-making purposes. Data transformation is the process that takes the data from the source database and moves it to the DW. Critical pieces of data transformation process addressed in the course are data quality, metadata, source/destination configuration and transformation rules. Data are cleansed, aggregated, transformed and loaded in the DW. The ETL issues were addressed appropriately and the students learned how to extract data from the (NorthWind) OLTP database to the DW (they created).

Data Quality

Data quality is a key issue when an organization implements an enterprise wide data warehouse, for example, in customer relationship management (CRM) or other purposes. Utilizing CRM requires that customer information be of high quality in order to identify, validate and consolidate customers within an organization. Quality of the data will determine the quality of the data warehouse as well as the quality of the decision. In other words data quality is an investment in profitability (Sperley, 1999).

One main purpose for a DW creation is the possibility of having integrated data in one place (DW). It solves the problems with non-integrated data. But it does not really solve the problems with bad or incorrect data in the operational (source) systems. We may still suffer from the syndrome "Garbage in- Garbage out of Control" (Vaas, 2005).

The course examines the aspects of ensuring data quality and recommend data quality improvement program in a data warehouse by utilizing mainly a revised process flow model (Chowdhury, 2007b). The model was originally proposed by Sperley (1999).

1. Where to start with the data quality improvement program?

Data in a data warehouse is gathered from different internal and external operational system sources. The first step in the data quality improvement program would be to discover where data quality problems exist in the source systems. When sources for problems are identified, methods must be developed to improve data quality.

2. When to start with the data quality improvement program?

Data quality improvement should be a continuous process. The sooner we start the better are the chances to have quality data. Data quality improvement program should be started prior to building a data warehouse. We can have a data cleanup program before we enter data in a DW. At the same time we can also initiate a data quality program and try to improve the quality of the data in the source system.

Data Mining

For the data mining portion of the course we thoroughly present and discuss the emergence and need of data mining. It is the sub-field of knowledge discovery in databases (KDD) for finding pieces of unrevealed information and put them together for meaningful decision-making. KDD is a multidisciplinary field of research including, statistics, machine learning, expert systems, database technology and data visualization. Usually statistical methods are being used for data mining purposes (Fayyad, et al 1996, Chowdhury, 1990, 2007b).

Data mining steps include: data preparation, defining a study, reading the data and building a model, understanding your model and prediction. (Ismail, Vargas and Chowdhury, 2004).

Besides, traditional methods, knowledge-based methods that use artificial intelligence techniques are also becoming common for performing advanced database analyses. Data mining tools integrated with new ways of presenting information (for example: OLAP – On-Line Analytical Processing with integrated data mining capabilities) is emerging as a powerful tool for the end-user for better decision-making.

Data mining is one or more analytic techniques (for example: Query tools (SQL), Statistical techniques (k-nearest neighbor), Decision trees, Neural networks and Genetic algorithms) in combination that can be used to turn data into knowledge and knowledge into action. For example, in the case of a customer data warehouse data about customer could be converted to knowledge about customer and appropriate course of action could be initiated (Groth, 1999).

The students are exposed to in depth studies of some of these techniques and use one or the other technique for finding answers to their burning questions from the DW they created from the OLTP database. Further the students are also trained to understand which technique to apply (where and how?). In order to address a business problem with data mining technique we need first to break down the business problem and convert it to a data mining task and then perform step-by-step data mining using suitable tool/s. To be successful in analysis and interpretation of the results we need to understand the data context, statistical/methodological context and the domain context (Chowdhury, 1990). The purpose is to make a substantive (subject matter) interpretation of the result/s obtained by using a data mining methodology. Otherwise the result obtained could be a methodological artifact (a consequence of methods applied and results obtained/produced), which could be compared with the infamous statement “lying with statistics.”

Over the years the focus of computational technology has shifted more from program-centric to data-centric and currently towards customer-centric approach. This shift has enabled business to learn more and more about customers to serve them better - a form of better customer relationship Management (CRM). Other possible uses of data mining are: classification; estimation; prediction; affinity grouping; clustering; and, profiling.

The students complete the (DM) part by writing and presenting a term paper, focusing on uses and applications of data mining tools and technologies like Oracle Data Miner, IBM's Intelligent Miner, Enterprise Miner of SAS, Clementine of SPSS and tools and techniques are being used, presented and discussed (Desai and Chowdhury, 2003).

Data Warehouses, Data Mining And Beyond

We also address that DW and DM are not an end in itself (Chowdhury, 2003). The knowledge gained and business intelligence attained by DW and DM could be retained and reused by using CBR (Case-based Reasoning) technology. It is the technology for solving new problems by adapting the known solutions of previous similar problems (Chan and Chowdhury, 2005). Data mining and CBR can be seen as complementary methodologies for knowledge management in a broader sense (Chowdhury, 2005).

RESULTS AND DISCUSSIONS

The students found the course very practical and useful. The course evaluations were performed in an open-ended qualitative form. The overall student feedbacks from the course offered so far were very encouraging. However, teaching a course in a dynamic field like this will always remain a challenge.

The course allows the creation of a distinctive and rewarding learning experience. This experience is more than just academically rigorous but also practically relevant as the desire is to stretch the students' thinking and actions to real world business problems and solutions. Overall, performing hands-on project work in small groups on real world business problems, discussing the outcome of the work and getting feedback contributed most to student learning. In fact learning by doing, discussing and sharing acted as a catalyst to enhance group as well as individual learning.

Key to the development and success of the course is an on-going collaboration between a large company in the retailing branch (operating one of the largest data warehouse in the country) and the university. Collaboration with the business world has proven invaluable in building bridges between the academic institution and the real world (Lawyer and Chowdhury, 2005). We have also been inviting guest speaker in the course from the industry. Guest

speakers from the industry add value to the course as well. The experience in collaboration between guest speakers and real-world business needs expands student understanding of these concepts.

In the near future we would like to focus on issues like Active Data Warehousing (ADW), Virtual Data Warehouse (VDW), Web-mining, etc. ADW with its Near Real Time (NRT) update will help organizations in both strategic and operational decision making based on (almost) the same data (updated in NRT). In VDW no physical DW environment is required. The topic VDW is not new. However, it has not been a success to date due to performance issues. The topic is again gaining importance due to the advent of faster processing capabilities. Further we would also explore the possibilities of combining DW and DM with CBR technology as a holistic approach for knowledge management.

REFERENCES

- Berry, M.J.A. and Linoff G.S. (2004). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (paperback), 2nd Edition. ISBN: 0-471-47064-3, Wiley.
- Chowdhury, S.I. (1990). *Computer-Based Support for Knowledge Extraction from Clinical Databases*, Linköping Studies in Science & Technology. Dissertations No. 240, Linköping University.
- Chowdhury, S.(2003). *Databases, Data Mining and Beyond*. Published in *The Journal of American Academy of Business, Cambridge*. Vol. 2, No. 2, pp. 576-580.
- Chan, J.O. (2004). *Building Data Warehouses Using The Enterprise Modeling Framework*. *Journal of International Technology and Information Management*, 13(2), 97-110.
- Chan, J and Chowdhury, S. (2005). *Enhancing Customer Services using Case-based Reasoning*. Published in *the communications of the ICISA – the official journal of the International Chinese Information Systems Association*, Vol VII. Number 1, pp 53-61, Summer 2005.
- Chan, J.O. (2007). *Planning For Successful Data Warehousing Deployment*. In *Proceedings of the MBAA/SAIS 43 Annual Conference*, 2007.
- Chen, P.P. (1976). *The Entity-Relationship Model: Toward a Unified View of Data*. *ACM Transactions on Database Systems*, March 1976, 1(1), 9-37.
- Chowdhury, S. (2005). *A Holistic Framework for Knowledge Management*. In *Issues in Information Systems. International Association for Computer Information Systems (IACIS). Volume VI, No. 2, pp 10-16, Fall 2005*.
- Chowdhury S. (2007a). *A Research Proposal on Model Conversion: E-R to Star Model*. Roosevelt University, Spring 2007.
- Chowdhury, S. (2007b). *Methodology for Ensuring Data Quality in Data Warehouses*. In *Proceedings of the MBAA/SAIS 43 Annual Conference*, 2007.
- Chowdhury, S. (2007c). *From E-R to Star Data Modeling – A Methodology*. Paper presented at the 10th International Conference of ASBBS (American Society of Business and Behavioral Sciences) in June, 2007.
- Desai, N. and Chowdhury, S. (2003). *Data Mining for Competitive Business Advantages: Present and Future Implications*. Published in the *Journal of American Society of Business and Behavioral Sciences (ASBBS)*. Vol. 10, NO. 1, pp62-71, Spring 2003.

- Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P., and Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press. ISBN 0-26256-097-6.
- Groth, R. (199). *Data Mining: Building Competitive Advantage*”, ISBN 0130862711, Prentice Hall.
- Ismail, W., Vargas, D. and Chowdhury, S. (2004). *Data Warehousing from a Business Intelligence Perspective. Journal of American Society of Business and Behavioral Sciences (ASBBS). Vol 12, No. 1. pp 90-98, Fall 2004.*
- Lawyer, J and Chowdhury, S. (2005). *Data Warehousing Practices – A success Story. Published in GESTS (Global Engineering, Science and Technology Society) International Transactions on Computer Science and Engineering, Paper field: Computer and its Application. Vol. 21 and No. 1. Nov. 2005.*
- Sperley, Eric. (1999) *The Enterprise Data Warehouse: The Planning, Building & Implementation Volume I: 1st Edition.* ISBN 0139058451, Prentice Hall.
- Todman, C. (2001). *Designing a Data Warehouse Supporting Customer Relationship Management.* Upper Saddle River, NJ: Prentice Hall PTR.
- Vaas, L. (2005). *Garbage In, Garbage Out of Control.* [Www.eweek.com/article2/0,1895,1828456,00.asp](http://www.eweek.com/article2/0,1895,1828456,00.asp)
- Zen, A. and Sinha, A.P. (2005). *A Comparison of Data Warehousing Methodologies. Communications of the ACM, 48(3), 79-84.*
- Jukic, N. (2006). *Modeling Strategies and Alternatives for Data Warehousing Projects. Communications of the ACM, 49(4), 83-88.*

Acknowledgement:

We sincerely thank the unknown reviewers on an earlier version of the manuscript for their comments. We also thank Phil Alonso, a former student in the course for proof reading and providing helpful comments on the work. Their comments have helped to improve the quality of the paper.

