doi:10.2498/cit.1001805

# Bridging the Knowledge Gap between Transactional Databases and Data Warehouses

Nenad Jukic[1] and Boris Jukic[2]

[1]Loyola University Chicago, Illinois, USA
[2]Clarkson University, Potsdam, New York, USA

Data warehouse is widely recognized in the industry as the principal decision support system architecture and an integral part of the corporate information system. However, the majority of academic institutions in the US and world-wide have been slow in developing curriculums that reflect this reality. This paper examines the issues that have contributed to the lag in the coverage of data warehousing topics at universities.

*Keywords:* ICT education, transactional databases, data warehouses, educational resources

## 1. Introduction

Even though data warehouse is widely recognized in the industry as the principal decision support system architecture and an integral part of the corporate information system, the majority of academic institutions in the US and worldwide have been slow in developing curriculum components that are consistent with this development. In this paper we examine the issues that have contributed to the lag in the coverage of data warehousing topics at universities.

Since the advent of data warehousing in the 1990's, industry seminars, vendor specific tutorials, and on-the-job training have been almost the exclusive means of education for the majority of people involved in corporate data warehousing projects. In many cases, the lack of formal academic education related to data warehousing has left these information systems professionals without a clear and meaningful understanding of the overall purpose of the data warehousing process and its various stages. This has been one of the contributing factors to the extraordinary failure rates of data warehousing projects, which are by some estimates higher than 50% [6][9]. Our goal is to bring the attention of the information systems academic community to this issue, by examining of the data warehouse academic-education challenges.

This paper is organized as follows. Section 2 gives an overview of data warehousing and related issues. Section 3 describes contemporary challenges and approaches related to data warehousing teaching and learning. Section 4 points to the existing data warehousing related resources that are available to academic insitutions. The paper is concluded with Section 5 that offers a brief summary.

## 2. Background – Data Warehousing

At the heart of any computer information system is a database and a database management system (DBMS). A database is an organized collection of logically related data tables. DBMS is software through which users interact with a database. Users use DBMS to create database tables, perform updates (insert, modify, and delete data) on database tables, and retrieve data from database tables. DBMS can also be used for creating user-friendly database interfaces and applications (such as forms and reports). A typical organization maintains and utilizes a number of transactional (operational) databases. These transactional databases are used to support the organization's day-to-day operations. A data warehouse is created within an organization as a separate database (using its own DBMS) whose primary purpose is data analysis for the support of management's decision making processes [10]. The data stored in the data warehouse captures many different aspects of the business process such as production, supply-chain management, sales, and marketing. This data reflects strategically important information such as customer behavior patterns, sales trends, outcomes of marketing strategies, and other characteristics. Therefore, this data is of vital importance to the success of the business whose state it captures. That is the reason why companies choose to engage in the relatively expensive and lengthy undertaking of creating and maintaining the data warehouse, often containing multiple terabytes of data. A recent study [7], reports typical cost of $3 million for creating a 1 terabyte data warehouse, with a typical implementation time of 2 years.

Often, the same fact can have both operational and analytical purposes, and subsequently can be stored in both an transactional database and the data warehouse. For example, data describing that product A was bought by customer B in store C can be stored in a transactional data store for business-process support purposes such as financial transaction record keeping or inventory monitoring. That same fact can also be stored in a data warehouse where, combined with vast numbers of similar facts accumulated over a time period, it is used to analyze important trends, such as sales patterns or customer behavior.

Why should the same fact be stored in two separate places? There are two main reasons that necessitate the creation of a data warehouse as a separate analytical data store. The first reason is the performance of queries in different contexts. Operational queries are mostly short and fast, while analytical queries are complex and consume a significant amount of time. The performance of operational queries can be severely diminished if they have to compete with analytical queries for computing resources. The second reason lies in the fact that, even if performance is not an issue, it is often impossible to structure a database which can be used in a straightforward manner for both operational and analytical purposes. Therefore, a data warehouse is created as a separate data store, designed for accommodating analytical queries. A typical data warehouse periodically retrieves selected analytically-useful data from the transactional data sources. The process and the infrastructure that facilitate the retrieval of the data from the transactional databases into the data warehouses is known as ETL, which stands for Extraction, Transformation and Load. Figure 1 illustrates a layout of a typical data warehouse.

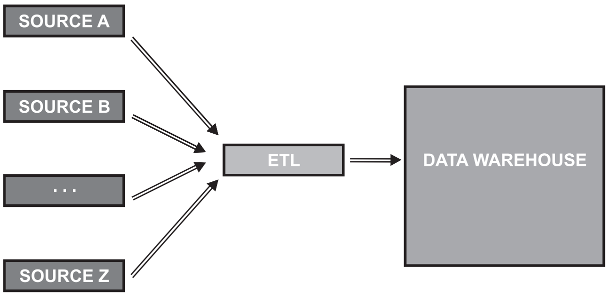

*Figure 1.* Data warehouse – a separate analytical repository.

A data mart is a data store based on the same principles as a data warehouse, but with a more limited scope. Whereas a data warehouse combines data from transactional databases across an entire enterprise, a data mart is smaller and focuses on a particular department or subject.

Typical data warehousing project follows some form of a System Development Life Cycle (SDLC). SDLC is the overall process of developing information systems through a multi-step process including steps such as planning, analysis, design and implementation [4]. One popular data warehouse-focused variation of the SDLC is the Data Warehousing Lifecycle [13]. The steps depicted in Figure 2 are common to any data-warehousing project: *data warehouse requirement collection and definition – data warehouse modeling – ETL design and development – front-end specifications (i.e. front-end requirement collection and definition) – front-end design and development – deployment – use/maintenance/growth.*

The data warehouse requirement stage involves agreeing on and defining the desired capabilities and functionalities of the future data warehouse. The final list of requirements must take into account the availability of information in the transactional data sources.

The data warehouse modeling and development stage uses the defined requirements as a basis

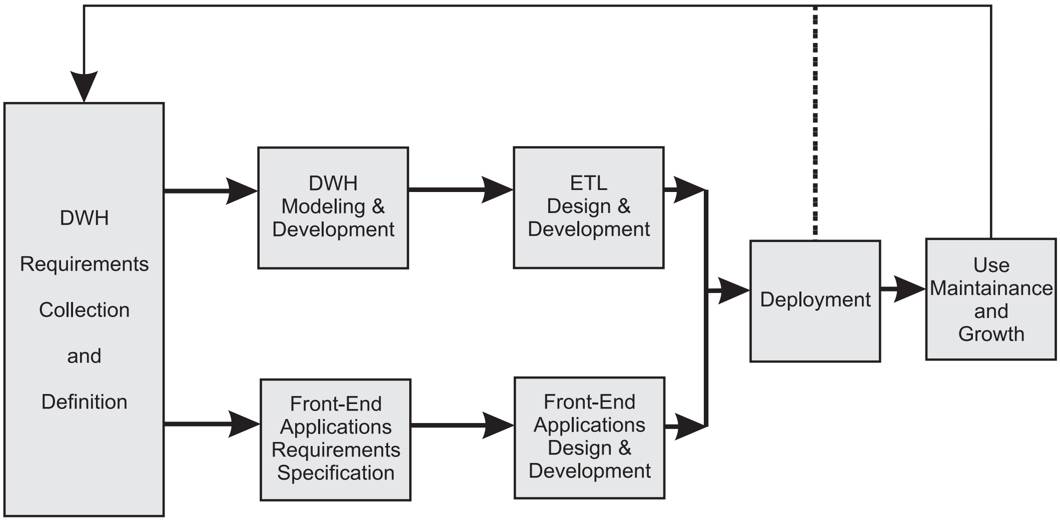
DWH
Requirements
Collection
and
Definition
DWH Modeling & Development
ETL Design & Development
Front-End Applications Requirements Specification
Front-End Applications Design & Development
Deployment
Use Maintainance and Growth

*Figure 2.* Abbreviated data warehouse system development lifecycle.

for creating a target data warehouse model and then using DBMS to implement the model by developing the structure of the actual target data warehouse.

The ETL design and development stage creates infrastructure and procedures for the tasks of retrieving analytically useful data from the transactional sources, transforming such data so that it conforms to the structure of the target data warehouse model, ensuring the quality of the transformed data, and loading the transformed and quality assured data into the target data warehouse. In most real-world data warehousing projects, the ETL stage is the most time and resource intensive. However, it is important to note that the ETL process as the process of moving data from the transactional sources into the data warehouse target, if properly undertaken, is predetermined by the results of the requirements collection and data warehouse modeling stages which specify the sources and the target.

Front-end applications requirements, design, and development stages specify the details of the data warehouse interfaces that are used to present the capabilities and functionalities of the data warehouse that are determined in the initial requirement collection and definition stage of the project to the end-users. Simply put, the role of the front-end applications stages is to provide user-friendly ways of utilizing those capabilities and functionalities.

The deployment of the data warehouse involves activating the data warehousing system composed of its target data warehouse, ETL infrastructure, and the front-end applications. The dashed line in Figure 2 indicates the options of alpha (within the development team) and beta (outside the development team) test releases, prior to the deployment of the actual working-system. These releases are designed to provide for testing and feedback collection, which can result in a modification of the requirements and changes in the system before the actual production deployment takes place. Once the actual production release has taken place, the use of the data warehouse by the analytical community of the organization can commence.

If the deployed corporate data warehouse has been used in a way that increases the efficiencies of the company's processes and opens new business opportunities, the first version of a data warehouse is often followed by the initiative to expand the scope of the project. The subsequent iterations of the data warehouse will have to go through the same stages of the project as the initial version, as is illustrated by the recursive line connecting the left-most and right-most squares in Figure 2.

We conclude this brief overview of basic data warehousing concepts by analyzing a comprehensive definition of data warehouse, originally coined by [10] and paraphrased by the Oracle Corporation in their BI&W glossary of terms, which characterizes data warehouse as an *enterprise-wide integrated structured repository of subject-oriented, time-variant, historical data used for decision support related infor-*

*mation retrieval*. *Enterprise-wide* refers to the fact that a data warehouse provides a company-wide view of the information it contains. *Integrated* refers to the fact the data warehouse integrates data from a number of separate independent transactional data sources within the company, and, in many cases, external sources as well. *Structured repository* refers to the fact that a data warehouse is a structured data repository hosted by a DBMS like any other database. *Historical* refers to the fact that a data warehouse typically contains multiple years worth of data, as opposed to the much shorter time-horizon for data in most traditional transactional databases. *Time-variant* refers to the fact that a data warehouse contains slices of data depicting corporate conditions across different time periods within its time horizon. *Subject-oriented* refers to the fundamental difference in the purpose of a traditional transactional database system and a data warehouse. A typical transactional database system is developed in order to support a specific day-to-day business operation (e.g. company order-entry database). A data warehouse is developed to analyze a specific business subject area (e.g. sales, profit). The data in the data warehouse is organized around major subject areas of an enterprise, and is therefore useful for an enterprise-wide understanding of those subjects [13].

Following this brief overview of basic data warehousing concepts, the next section examines the coverage of data warehousing topics in the academic settings.

## 3. Academic Data Warehouse Challenges and Issues

Practically every contemporary information systems program at universities in the US and worldwide features one or more database management related course. Within the last few decades such courses have become one of the fundamental parts of the management information systems curriculum. There is a plethora of mature database management textbooks, such as [5][8][15], with firm agreement on the topics that form the foundation of a contemporary database management course. Those common topics are centered on the skills and methods that facilitate the main database lifecycle activities. These activities are conceptual database modeling (which includes collecting the requirements and creating an architectural blueprint for a database), database implementation (using DBMS to create actual transactional databases, based on the conceptual design) database utilization (using DBMS to populate and update databases, as well as to retrieve data from the databases) and database administration (security, concurrency control, user rights administration etc). In contemporary database textbooks these processes are typically discussed and demonstrated in the context of transactional (a.k.a. operational) databases supporting day-to-day business processes.

Most contemporary database textbooks include an overview of data warehousing topics, where much of the terminology and fundamental concepts are mentioned, defined, and, in some cases, described by small, typically stand-alone examples. However, in contrast to the detailed and systematic coverage of the topics related to the development and utilization of transactional database systems (such as conceptual data base modeling, normalization, SQL), the amount of hands-on development material and other associated resources is negligible. Therefore, contemporary database textbooks mostly fail to illustrate the details and mission of the data warehousing projects in a comprehensive, integrated and meaningful fashion. Similar abridged summary-style coverage of data warehousing topics and materials can be found in books dealing with encompassing and/or parallel topics such as decision support systems, business intelligence, data mining or data visualization [14][16].

In contrast to the summary-style coverage of data warehousing topics in the mainstream textbooks is the wide-availability of in-depth and detailed practitioner's books dealing with data warehousing such as [1][10][13]. These books provide a wealth of knowledge and are often written by the original inventors of some of the most important concepts in data warehousing. However, these books are almost always intended for usually experienced IS professionals, and not undergraduate or graduate university students. This is demonstrated by the absence of exercises, case studies, integrated running examples, supporting software, and other teaching material that serves to illustrate and enforce the concepts.

In addition to the shortage of comprehensive ready-to-use teaching content, the issue of data warehouse technology complexity and cost has also contributed to the lag in coverage of data warehousing topics at universities. A typical data-warehousing project utilizes a host of software solutions including data modeling tools, ETL tools, OLAP tools, DBMS (in many cases specialized data warehousing DBMS). Many of the professional tools in these categories are expensive, complex to install and maintain, and require significant hardware resources. This fact often serves as a discouraging factor forcing educators into token coverage of data warehousing topics.

In order to provide a strong argument about the unsatisfactory state of academic data warehouse coverage, an analysis was performed on the course descriptions of 103 representative undergraduate information systems programs in the US. We considered schools that can reasonably be viewed as either mainstream or trendsetting institutions when it comes to the information systems curriculum and implementation. The undergraduate programs we analyzed either appear in the Business Week top 50 undergraduate programs ranking or the top 100 ranking by the US News and World report (or both, of course) and have an undergraduate program in information system. The remainder of the sample consists of the schools whose graduate information systems programs were ranked, and have an undergraduate program in information systems. Results of our analysis reveal that the students in standard database management classes acquire a great deal of knowledge and skills that prepare them for developing, using and managing transactional database systems. On the other hand, most programs are typically less focused and consistent when it comes to the teaching of the skills that are necessary for the design and use of decision-supporting data warehouses. The typical course sequence in an undergraduate information systems program includes one or two database design and management courses followed by, or taken concurrently with, a decision support systems course. In few instances there is a whole decision support track, with two, three or even more courses offered. The purpose and content of the data design and management courses is relatively uniform, exposing students to the fundamentals of relational database design, ER modeling, and normalization as well as writing of SQL queries of varying degrees of complexity. The content of the decision support courses varies widely across institutions, often reflecting individual institution's understanding of which of the many different decision support activities and processes merits emphasis. In most surveyed institutions, data warehousing is not covered in either database or decision support courses. In only 34% of surveyed institutions there is some mention of data warehousing within any of the course descriptions in their entire information systems-related curriculum. Moreover, less than half of those institutions (15% overall) have courses that offer any meaningful coverage of data warehouse design and modeling. It is reasonable to assume that in the broader academic environment, beyond this sample of highly ranked schools, these numbers are even lower.

This situation reflects a problematic disconnection between the industry and academia. The current job market for IS majors requires competence in both transactional database systems and data warehouses, while academic programs for the most part focus strictly on the former. This is not to say that other decision support processes and activities should be de-emphasized at the expense of extending the curriculum to cover data warehousing issues, but rather that the schools seem not to be aware of the paramount importance of data warehousing in all the decision support activities within large (and increasingly in small to medium) enterprises.

The lack of appropriate amount of attention and focus on data warehousing in most academic programs and curriculums is strongly related to the unavailability of appropriate educational resources. In the next section we will describe some of the available free resources that can enable educators to deal with the issues and challenges described above.

## 4. Free Academic Resources

We developed two easy–to-use completely free software packages in order to support the modeling-centered approach to teaching data warehousing. The software, which is free of charge to all academic institutions, takes into account both widely-recognized data warehouse modeling methods. Before we describe the tools, we provide a short overview of the two data modeling techniques for modeling large-scale, enterprise-wide data warehouses [11]: dimensional modeling and ER modeling.

Dimensional modeling [13] groups data attributes into two types of tables: facts and dimensions. A fact table contains one or more measures (usually numerical) of a subject that is being modeled for analysis. Dimension tables contain various descriptive attributes (usually textual) that are related to the subject depicted by the fact table. The data model that is produced by the dimensional modeling method is known as a star-schema [2], or a star-schema extension such as snowflake or constellation. The intent of the dimensional model is to represent relevant questions whose answers enable appropriate decision-making in a specific business area [3]. Dimensional data warehousing modeling method, championed by Kimball [13], views a data warehouse as a collection of dimensionally modeled data marts. Another option for modeling data warehouses, first proposed by Inmon [10], envisions a data warehouse as an integrated database modeled by using the traditional database modeling technique (ER modeling). After such a data warehouse is created, it then serves as a source of data for dimensionally modeled data marts.

One tool, named FatFreeERD, is a standard ER modeling tool, an alternative to commercial data modeling tools such as CA ERwin Data Modeler. Unlike commercial tools, FatFreeERD is completely free of charge for academic institutions and requires no licensing or tool-specific training. The other tool, named FatFreeStar, is a dimensional modeling tool for creating star schemas, complete with fact and dimension tables. Like FatFreeERD, FatFreeStar is also completely free of charge for academic institutions and requires no licensing or tool-specific training. The diagrams created by these tools can be easily embedded into a variety of documents of different formats, such as MS Word or MS PowerPoint. In addition to FatFreeERD and FatFreeStar tools, we have created a number of data warehouse modeling-assignments, and we have made them available to the wider academic community. These tools and case-assignments are available for free download and use by any academic institution at the Teradata University Network (TUN) website www.teradatauniversitynetwork.com. Teradata University Network is a no-catch free educational portal that uses a zero cost-zero installation-zero maintenance model to provide tools, data-sets, and other materials for teaching and learning about database management, data warehousing, decision support, and other contemporary data management and analysis topics.

In addition to the resources we just described, other free data warehouse education-related content offered on TUN includes data warehousing DBMS software, OLAP software, large industry data-sets, articles, case studies, tests, exercises, lectures, syllabi, lesson plans, tutorials, demos, and podcasts. Detailed descriptions of the TUN resources can be found in [12]. The DBMS and OLAP software is hosted by TUN and is used via the internet.

## 5. Summary and Conclusion

The focus of this paper was on examining the issues that encumber the pervasive, effective, and meaningful inclusion of data warehousing topic into information systems curriculums worldwide and describing resources that can help dealing with these issues.

Many of the professional data warehousing software tools are expensive, complex to install and maintain, and require significant hardware resources, which hinders widespread meaningful data warehouse education in academic settings.

With the resources we described here, the cost, maintenance, installation, hardware, and support personnel factors are eliminated from the data warehouse teaching equation, enabling any faculty member with internet access to provide a meaningful data warehouse learning experience to their students.

## References

[1] S. Adelman, L. T. Moss, *Data Warehouse Project Management*, Addison-Wesley, 2000.

[2] A. Chaudhuri U. Dayal, An Overview of Data Warehousing and OLAP Technology. *ACM SIGMOD Record 26*, 1, pp. 65–74, 1997.

[3] T. Chenoweth, D. Schuff, and R. St. Louis, Method for developing dimensional data darts. *Communications of the ACM*, 46, 12, 93–98, 2003.

[4] A. Dennis, B. Wixom and R. Roth, *System Analysis & Design*, 3rd Edition, John Wiley and Sons, 2006.

[5] R. Elmasri and S. Navathe, *Fundamentals of Database Systems*, Addison Wesley, Boston, MA, 2007.

[6] M. N. Frolick, K. Lindsey, Critical Factors for Data Warehouse Failure. *Business Intelligence Journal*, 8, 1, Winter 2003.

[7] P. Gray, *Manager's Guide to Making Decisions about Information Systems*, Wiley, 2006.

[8] J. Hoffer, M. Prescott, and H. Topi, *Modern Database Management*. Prentice Hall, Upper Saddle River, NJ. 2008.

[9] M. I. Hwang, H. Xu, The Effect of Implementation Factors on Data Warehousing Success: An Exploratory Study. *Journal of Information, Information Technology, and Organizations*, 2, pp. 1–14, 2007.

[10] W. Inmon, *Building the Data Warehouse*, 4th Edition, Wiley, 2005.

[11] N. Jukic, Data Modeling Strategies and Alternatives for Data Warehousing Projects. *Communications of the ACM*, 49, 4, pp. 83–88, 2006.

[12] N. Jukic and P. Gray, Using Real Data to Invigorate Student Learning. *INROADS – the ACM SIGCSE Bulletin*, Vol. 40, No. 2, 2008, pp. 6–10.

[13] R. Kimball, M. Ross, W. Thornthwaite, J. Mundy, B. Becker, *The Data Warehouse Lifecycle Toolkit*. Wiley, 2007.

[14] G. M. Marakas, *Modern Data Warehousing, Mining, and Visualization – Core Concepts*. Prentice Hall, 2003.

[15] P. Rob and C. Coronel, *Database Systems: Design, Implementation and Management*. Cengage Learning, Florence, KY, 2008.

[16] E. Turban, R. Sharda, J. E. Aronson, D. King, *Business Intelligence – A Managerial Apporach*. Prentice Hall, 2008.

*Received*: June, 2009
*Accepted*: April, 2010

*Contact addresses:*

Nenad Jukic
Loyola University Chicago
1 E Pearson, Chicago IL 60611, USA
e-mail: njukic@luc.edu

Boris Jukic
Clarkson University
372 Bertrand H. Snell Hall
PO Box 5790, Potsdam NY 13699-5790, USA
e-mail: bjukic@clarkson.edu

Nenad Jukic, Associate Professor of Information Systems and the Director of Data Warehousing and Business Intelligence Graduate Program at Loyola University Chicago (US) first attended and presented a paper at ITI in 1998, when it was held in Pula. Since then Nenad has attended seven more ITI conferences. He has been a vice chair for the Business Intelligence, Databases and Information Systems track at ITI for more than five years. Nenad has brought a number of distinguished speakers to ITI over the years and continues to be involved with the organization of the conference.

Boris Jukic, Associate Professor of Operations & Information Systems and the Director of Graduate Business Programs at Clarkson University, Potsdam, New York (US) has been a contributor and a tireless reviewer for the ITI conference for more than seven years. Boris is happy to be an ambassador for the ITI among business information systems academics in the US and world-wide.