Design of a Data Warehouse for Academic Planning at Private College: A Case Study at Cosmopoint College Metro Campus Sungai Petani

Sharmila Mat Yusof

Faculty of Information Technology
06010, Sintok, Universiti Utara Malaysia
sharmila_yusof@yahoo.com

Abstract

A data warehouse is a repository of integrated information used to perform query and analysis on vast amount of operational data. Since data warehouse project requires extensive resources to be completed on time, many organizations chose to incrementally develop data marts instead, gradually merge to be the organization's data warehouse. However, due to the non-profit advantage, data warehouse has been rarely implemented in higher education institution although majority of the institutions have some kind of information system. As a result, the reporting and analysis of the student record system is often ad hoc and time-consuming tasks. This research was intended to analyze and design a data warehouse that will enable high improvement data analysis for the Academic planning of the college. During the requirement analysis, the fact-finding techniques of observations and interviews have been conducted with the key users of the Online Analytical System (OLTP) to gain deep understanding and verify the facts collected. From the user requirement analysis, the logical model of Entity Relationship (ER) Diagram was developed and has iteratively been used to verify the user requirements. Then, the dimensional model was developed using the Kimball's 'Nine-Step Methodology'. Finally, the design was validated using the Microsoft SQL server 7.0.

Keywords: data warehouse, data mart, academic planning, college

1.0 INTRODUCTION

The original concept of a data warehouse was devised by IBM as the 'information warehouse and presented as a solution for accessing data held in non-relational systems. The information warehouse was proposed to allow organizations to use their data archives to help them gain a business advantage. However, due to the sheer complexity and performance problems associated with the implementation of such solutions, the early attempts at creating an information warehouse were mostly rejected. Since then, the concept of data warehousing has been raised several times but it is only in recent years that the potential of data warehousing is now seen as a valuable and viable solution. The latest and most successful advocate for data warehousing is Bill Inmon, who has earned the title of 'father of data warehousing' due to his active promotion of the concept [1].

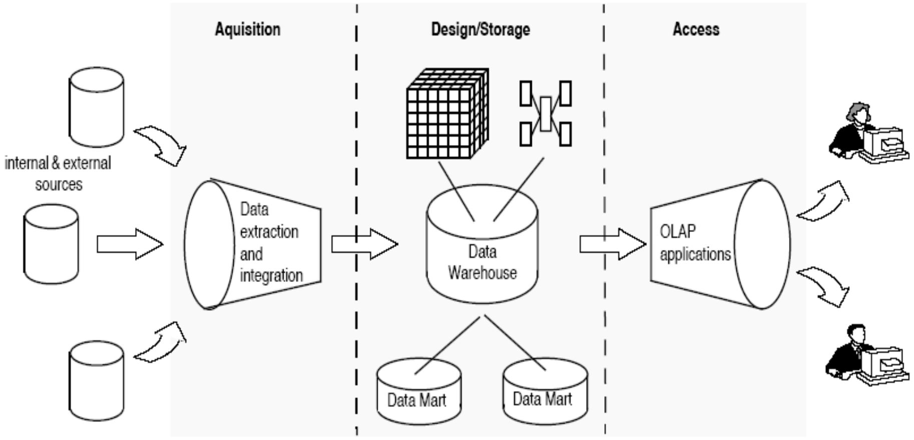
Since the 1970s, organizations have mostly focused their investment in new computer systems that automate business processes that is an Online Transaction Processing (OLTP) system. In this way, organizations gained competitive advantage through systems that offered more efficient and cost-effective services to the customer. Throughout this period, organizations accumulated growing amounts of data stored in their OLTP databases. However, in recent times, where such systems are commonplace, organizations are focusing on ways to use operational data to support decision-making, as a means of regaining competitive advantage.

OLTP systems were never designed to support such business activities and so using these systems for decision-making may never be an easy solution. The legacy is that a typical organization may have numerous operational systems with overlapping and sometimes contradictory definitions, such as data types. The challenge for an organization is to turn its archives of data into a source of knowledge, so that a single consolidated view of the organization's data is presented to the user. The concept of a data warehouse was deemed the solution to meet the requirements of a system capable of supporting decision-making, capable of receiving data from multiple operational or OLTP data sources.

As defined by Inmon (1996), data warehousing is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision-making process. Regardless various definitions, the ultimate goal of data warehousing is to integrate organization-wide corporate data into a single repository from which users can easily run queries, produce reports, and do performance analysis [15].

Accompanying the rapid emergence of data warehouses is the related concept of data marts. A data mart holds a subset of the data in a data warehouse normally in the form of summary data relating to a particular department or business function. The data mart can be standalone or linked centrally to the corporate data warehouse. The typical architecture for a data warehouse system and associated data mart is shown in Figure 1. The characteristics that differentiate data marts and data warehouse include:

- A data mart focuses on only the requirements of users associated with one department or business function
- Data marts do not normally contain detailed operational data, unlike data warehouses
- As data marts contain less data compared with data warehouses, data marts are more easily understood and navigated



(Source: [5])

The Cosmopoint College as one of the higher educational institution, there is an OLTP system namely a College Administration System 2000 (CAS 2000) system that caters for the daily operation of each branch of campuses throughout Malaysia. The system comprises of five main modules, which are Admission Management, Class Management, Attendance Management, Exam and Financial. This system generates data that are vital to the college decision makers. Currently, the CAS 2000 system mostly in each branch resides on the DELL Server PowerEdge 1500SC using Microsoft SQL Server 7.0 database and Windows 2000 Server operating system.

This research involves the analysis of user requirement and designs a data warehouse specifically for Academic Planning of the Cosmopoint Metro Campus Sungai Petani.

1.1 PROBLEM STATEMENT

To date, very few efforts have been made on designing a data warehouse model for higher education information systems although majority of institutions of higher education have some kind of information system, i.e. have some way of gathering and accumulating data [3]. This is probably due to the fact that this area is not as commercially attractive as other money related businesses such as accounting, banking etc. As a result, the management and maintenance of student record system is often ad hoc, and tends to be more resource and attention intensive than accounting systems [3].

Cosmopoint College was established in 1991. With the initial service as an IT training center, the center expanded tremendously over the last decade. Cosmopoint specializes in education, corporate training, multimedia development, software development and IT consulting. Its education unit offers courses in the area of Information Technology, Multimedia and Computer Graphic Design. To date, Cosmopoint has spawned over 5,500 graduates and has 14 branches nationwide. In each branch except headquarters in KL, there are 3 main departments, which are Academic department, Operation department and Marketing department. The most important department is the Operation Department, which responsible for providing Academic data and information for the management's decision.

At Cosmopoint College, due to its scattered branches with no dedicated network between the branches, the operational reporting including academic reporting has become the responsibilities of each branch. Due to the standalone CAS 2000 system in each branch, each branch would need to submit pre-defined monthly reports and ad-hoc reports to management in headquarter for performance evaluation and decision-making. The person responsible at each branch for submitting the reports is Assistant Manager Academic (AMA). As an education institution, the college every day creates data about students, supporting programs, fees etc, of which are important in supporting the daily works of the Operation Department of the college especially for Academic planning, but for the most part, this data is lockup up in a myriad of manual and computer systems and is exceedingly difficult for the AMA to get at.

This research is intended to analyze and design a data warehouse that will enable high improvement data access for the Operation Department of the college.

According to Michael Haisten (2002), the most powerful justifications for opting Data Warehouse investment are:

- Quality goals, since its typical objective are improving information access.
- Bringing the user in touch with their data.
- Enhancing the quality of their decisions.

The result obtained will then be useful for future development of successful Data Warehouse of the Cosmopoint College Metro Campus Sungai Petani.

1.2 OBJECTIVE

The general purpose of the research is to study and analyze the data warehouse architecture, design Data Warehouse for Operation Department data mart specific to Academic planning.

The proposed study objectives, derived from this general purpose are:

- To study the architecture of the data warehouse.
- To develop the logical model of the data warehouse that is the ER Diagram and Dimensional Model.
- To validate the design of the data mart for Academic Planning.

1.3 SCOPE

The scope of this research covers the analysis and design of a data mart for Academic planning in Operation Department within the context of Relational Online Analytical Processing (ROLAP).

2.0 LITERATURE REVIEW

Data warehouse is a computer based information systems that are home for "secondhand" data that originated from either another application or from an external system or source [2]. Data warehouse system consists of a collection of decision support technologies, aimed at enabling the decision makers to make better and faster decisions. Data warehousing technologies have been successfully deployed in many industries such as in manufacturing, retail, financial services, transportation, telecommunications, utilities, and healthcare [4]. Nevertheless, there is less effort to deploy a data warehouse technologies at higher educational institution [3].

Typically, a data warehouse is maintained separately from the organization’s operational databases. There are many reasons for doing this. The data warehouse supports on-line analytical processing (OLAP) technology, the functional and performance requirements of which are quite different from those of the on-line transaction processing (OLTP) applications (operational systems) traditionally supported by the operational databases. OLAP is a term that describes a technology that uses a multi-dimensional view of aggregated data to provide quick access to strategic information for the purpose of advanced analysis [1]. OLAP enables users to gain deeper understanding and knowledge about various aspects of the institution data through fast, consistent, interactive access to a wide variety of possible views of the data.

OLTP applications typically automate transactional processing tasks such as order entry and banking transactions that are the core daily operations of an organization. These tasks are structured and repetitive, and consist of short, atomic, isolated transactions. The transactions require detailed, up-to-date data, and read or update a few (tens of) records accessed typically on their primary keys. Operational databases tend to be hundreds of megabytes to gigabytes in size. Consistency and recoverability of the database are critical, and maximizing transaction throughput is the key performance metric. Consequently, the database is designed to reflect the operational semantics of known applications, and, in particular, to minimize concurrency conflicts.

Data warehouses, on the other hand, are targeted for decision support. Historical, summarized and consolidated data is more important than detailed, individual records. Since data warehouses contain consolidated data, perhaps from several operational databases, over potentially long periods of time, they tend to be orders of magnitude larger than operational databases; enterprise data warehouses are projected to be hundreds of gigabytes to terabytes in size. The workloads are query intensive with mostly ad hoc, complex queries that can access millions of records and perform a lot of scans, joins, and aggregates.

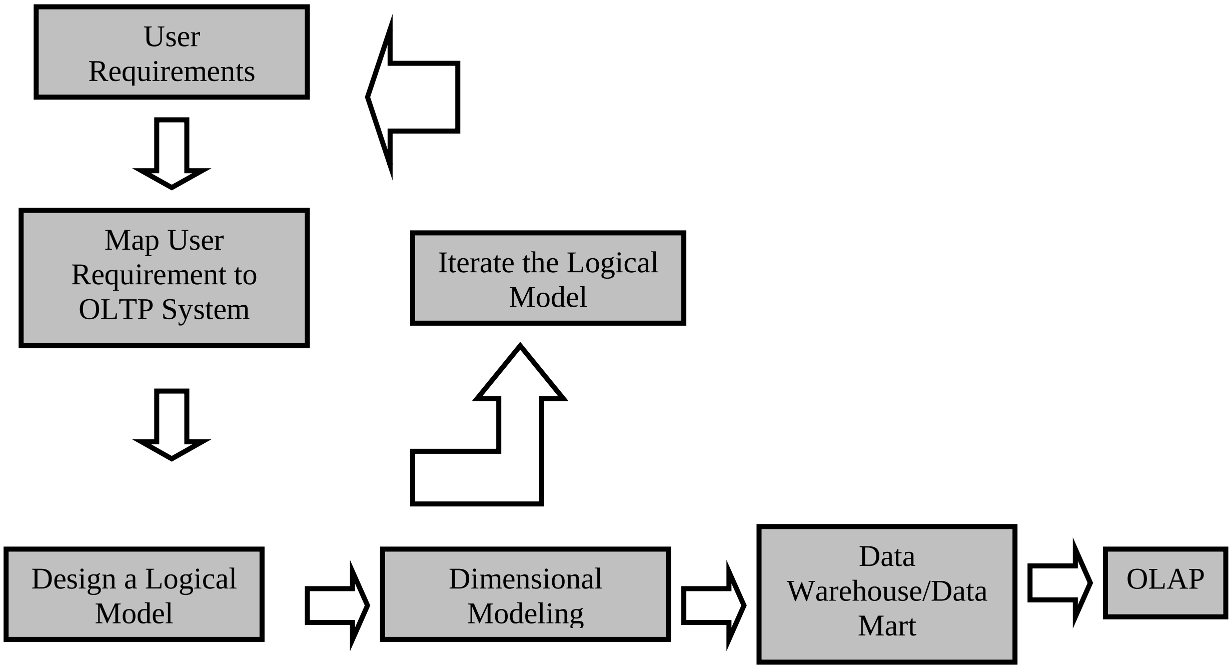
To facilitate complex analyses and visualization of the data in a warehouse, the data is typically modeled in *multidimensional* using OLAP technology. Typical OLAP operations include *rollup* (increasing the level of aggregation) and *drill-down* (decreasing the level of aggregation or increasing detail) along one or more dimension hierarchies, *slice_and_dice* (selection and projection), and *pivot* (re-orienting the multidimensional view of data). OLAP system not only it can answer ‘who?’ and ‘what?’ questions, it also able to answer more complex questions such as ‘what is?’ and ‘why?’ type of questions. Thus, OLAP enables decision-making about future actions.

Data warehouses might be implemented on standard or extended relational DBMSs, called Relational OLAP (ROLAP) servers. These servers assume that data is stored in relational databases, and they support extensions to SQL and special access and implementation methods to efficiently implement the multidimensional data model and operations. In contrast, multidimensional OLAP (MOLAP) servers are servers that directly store multidimensional data in special data structures (e.g., arrays) and implement the OLAP operations over these special data structures.

In addition to selecting an OLAP server and defining a schema and some complex queries for the data warehouse, there is more to building and maintaining it. For that reason, different architectural alternatives exist to suite a particular organizational requirement. Many organizations want to implement an integrated enterprise warehouse that collects information about all subjects (e.g., customers, products, sales, assets, personnel) spanning the whole organization. However, building an enterprise warehouse is a long and complex process, requiring extensive business modeling, and may take many years to succeed. Some organizations are settling for *data marts* instead, which are departmental subsets focused on selected subjects (e.g., a marketing data mart may include customer, product, and sales information)[4]. These data marts enable faster roll out, since they do not require institution-wide consensus…. [4].

3.0 METHODOLOGY

This project followed five steps in accomplishing the project objectives. Steps used in the project methodology are depicted below.



3.1 User Requirements – Fact Finding Techniques

3.1.1 Observation

Observation is one of the most effective fact-finding techniques for understanding a system. With this technique, it is possible to either participate in, or watch, a person perform activities to learn about the system. The persons observed were the CAS 2000 users - Admin Clerk, Assistant Manager (Academic) and lecturers. The aim of the observation was to conduct a detailed notation of behaviors, events and the contexts surrounding the daily operation of the OLTP system in order to understand the current flow of the system.

3.1.2 Documents Examination

Examining documentation can be useful when we are trying to gain some insight as to how the need for the data warehouse arose. By examining documents, forms, reports and files associated with the OLTP system, we can quickly gain some understanding of the system. The types of documents that were examined were memos, emails, monthly and ad-hoc reports produced weekly, monthly and yearly with regards to Academic planning.

3.1.3 Interviews

At this stage, interviews were conducted to collect information from respective persons in order to identify requirements, gathering ideas and opinions. An interview has been held with the Assistant Manager (Operation), Assistant Manager (Academic), IT Executive and clerks of the CAS 2000 system to enable the identification of a prioritized set of requirements for the college that the data warehouse must meet in the area of Academic planning.

3.2 Map User Requirements to OLTP System

During this stage, again the interviews with the Assistant Manager (Operation) and Assistant Manager (Academic) have been held to verify and clarify facts gathered during the fact-findings techniques conducted earlier.

3.3 Design a Logical Model (ER Diagram)

According to Ralph Kimball (2002), ER modeling is a powerful technique for designing transaction-processing systems in relational environments. The value of the entity relationship model is achieved in modeling institutional data. Gaining consensus within the business as to the name, definition and business rules about the data within the institution is step one for anything we choose to do with the data to determine whether the data is appropriate for a data warehouse system. From the fact-finding techniques, the Entity- Relationship (ER) Diagram has been produced and iteratively been used to verify the user requirement of the data mart.

3.4 Design a Dimensional Model

There are many approaches that offer alternative routes to the creation of a data warehouse. A typical approach is to decompose the design of the data warehouse into manageable parts that are data marts. At a later stage, the integration of the smaller data marts leads to the creation of the institution-wide data warehouse. The methodology specifies the steps required for the design of a data mart, however, the methodology also ties together separate data marts so that over time they merge together into a coherent overall data warehouse. The Kimball's '***Nine-Step Methodology***' will be used to design a data warehouse [1]. The steps are:

1. Choosing the process

   The process (function) refers to the subject matter of a particular data mart. The first data mart to be built should be the one that is most likely to be delivered on time, within budget, and to answer the most commercially important business questions.

2. Choosing the grain

   The grain is referred to the level of detail available in a star schema. The grain of the star schema is the finest level of detail implied by joining of the fact and dimension tables. For example, the grain of a student record star scheme with dimensions of time (academic year, term), student bio (one record per student) term (one record per student per term) and student matriculation (one record per student per course of study undertaken) would be "student per term per course of study."

3. Identifying and conforming the dimensions

Dimensions set the context for asking questions about the facts table. A well-built set of dimensions makes the data mart understandable and easy to use. A poorly presented or incomplete set of dimensions will reduce the usefulness of a data mart to an institution. If any dimension occurs in two data marts, they must be exactly the same dimension, or one must be a mathematical subset of the other. Only in this way can two data marts share one or more dimensions in the same application. When a dimension is used in more than one data mart, the dimension is referred to as being conformed.

4. Choosing the facts

   The grain of the fact table determines which facts can be used in the data mart. All the facts must be expressed at the level implied by the grain. Additional facts can be added to a fact table at any time provided they are consistent with the grain of the table.

5. Storing pre-calculations in the fact table

   Once the facts have been selected each should be re-examined to determine whether there are opportunities to use pre-calculations. A common example of the need to store pre-calculation occurs when the facts comprise a profit and loss statement.

6. Rounding out the dimension tables

   In this step, we return to the dimension tables and add as many text descriptions to the dimensions as possible. The text descriptions should be as intuitive and understandable to the users as possible. The usefulness of a data mart is determined by the scope and nature of the attributes of the dimension tables.

7. Choosing the duration of the database

   The duration measures how far back in time the fact table goes. In many enterprises, there is a requirement to look at the same time period a year or two earlier while others, there may be a legal requirement to retain data extending back five or more years such as in insurance companies.

8. Tracking slowly changing dimensions

   The slowly changing dimension problem means, for example, that the proper description of the old attributes i.e. student bio must be used with the old transaction history. There are three basic types of slowly changing dimensions:

   Type 1 – where a changed dimension attribute is overwritten.

   Type 2 – where a changed dimension attribute causes a new dimension record to be created.

   Type 3 – where a changed dimension attribute causes an alternate attribute to be created so that both the old and new values of the attribute are simultaneously accessible in the same dimension record.

9. Deciding the query priorities and the query modes

   In this step we consider physical design issues. The most critical physical design issues affecting the end-user's perception of the data mart are the physical sort order of the fact table on disk and the presence of pre-stored summaries or aggregations.

At the end of this methodology, we have a design for a data mart that supports the requirements of a particular business process and also allows the easy integration with other related data marts to ultimately form the institution-wide data warehouse.

4.0 FINDINGS

4.1 CAS 2000 System Overview

Due to the scattered branches across Malaysia and there is no dedicated network between theses branches, the operational reporting has become the responsibility of each branches. Currently, the CAS 2000 system resides on the DELL server using Microsoft SQL Server 7 database. Every month and year, there are various reports that need to be submitted to headquarters. The reports are basically on students' progress and collection. Based on this, all the branches would be ranked for monthly and yearly performance, which contribute to the branch performance evaluation.

From the observations and interviews with CAS 2000 users, the CAS 2000 systems tend to have 5 basic components that serve four basic constituencies.

1. **Admission Management**

   The Admission module is for receiving application information; registration of newly admitted students and application for sponsorship for those eligible.

2. **Class Management**

   The Class Management module is for the creation of the classes offered every semester. It also includes creation of the additional and replacement classes. The Class Management also serves a function as subject registration of continuing students, tracking student's academic progress and providing information to facilitate advising.

3. **Class Attendance Management**

   The Class Attendance module is for lecturers to record the attendance of the students for each class they taught every semester.

4. **Financial**

   The financial module is for receiving and recording main course fees and additional fees such as referral exam fees, fines, Convocation fees, lab fee s and etc.. It also receives funds from Sponsorship, applies those credits to the balances owned, and sends bills for the reminder due.

5. **Examination**

   The Examination module is for lecturer to enter student marks for coursework, mid-term examination and final examination.

4.2 ER Diagram

Refer to Attachment ***A*** *and Attachment* ***B.***

4.3 Dimensional Model

Refer to Attachment ***C*** *to* ***J.***

5.0 SIGNIFICANCE OF STUDY

The result obtained from this study can be used to develop data warehouse for the whole Department of the college. Also the documentation of the research will be used as reference for any other study on the topic of data warehouse especially for the Cosmopoint College.

This study will also encourage many organizations especially in higher educational institution to opt for data warehouse investment in order to improve information access within their organizations, bringing the user of their information in touch with their data, and providing cross-function integration of operation systems within the organization. Data warehouse for the Operation Department will enable the decision makers to access data, understood data and manipulate them while making decisions for the Cosmopoint College.

6.0 CONCLUSION

In this paper, the requirement analysis and the design of the data mart prototype was presented. The data mart development specifically for the Academic Planning can reduce the burden of the Assistant Manager Academic in coping with the reporting and student's data analysis requirement from the headquarters. The successful implementation of the students' data mart would lead to the whole creation of the data warehouse for Cosmopoint Metro Campus Sungai Petani.

7.0 REFERENCES

[1] Thomas M. Connolly, Carolyn E. Begg, Anne D. Strachan, Database systems: a practical approach to design, implementation and management, Addison-Wesley. 2005.

[2] Data Warehouse Description. Minnesota State Archives; 2002. Retrieved on 10th November, 2005 from
http://www.mnhs.org/preserve/records/dwintro.html

[3] Allan RG. Data Model for a Registrar's Data Mart. Georgetown University; 2000. Retrieved on n10th November, 2005 from
http://www.educause.edu/ir/library/pdf/CMR0033.pdf

[4] S. Chaudhuri and U. Dayal. An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD Record, 26(1):65-74, March 1997.

[5] Gatziu, S., & Vavouras A. (1999). Data Warehousing: Concepts and Mechanisms. Retrieved on 25th November, 2005 from
http://www.ifi.unizh.ch/groups/dbtg/Staff/Vavouras/InformSIJan99.pdf

[6] Sen, A., & Sinha, A. P. A Comparison of Data Warehousing Methodologies. Communications of The ACM March 2005/Vol. 48, No. 3. Retrieved on 25th November, 2005 from http://www.ifs.tuwien.ac.at/~bruckner/pubs/dexa2002_dwh_development.pdf.

[7] Baranovic, M., Madunic, M. & Mekterovic, I. (2003). Data Warehouse as a Part of the Higher Education Information System in Croatia.

[8] Lechtenborger, J. (n.d.). Data warehouse Schema Design. Retrieved on 1st October, 2005 from
http://doesen0.informatik.uni-leipzig.de/proceedings/paper/diss2.pdf

[9] Data Warehousing Case Study: Image Craft. (2005). Retrieved on 7th November, 2005 from http://www.interprisesoftware.com/data-warehouse.html

[10] Brown, B. (n.d.). Data Warehouse Implementation with the SAS System. Retrieved on 10th November, 2005 from
(http://www2.sas.com/proceedings/sugi22/DATAWARE/PAPER132.PDF

[11] Lin, M. C. (2001) University Data Warehouse Design Issues: A Case Study. Proceedings of the 2001 American Society for Engineering Education Annual Conference & Exposition. Retrieved on 4th November, 2005 from http://www.ecet.ipfw.edu/~linm/publications/ASEE_DataWareHouse01063_2001.pdf

[12] Allan RG. Snowflakes and Grain: Snowflaking to Meet A Design Constraint. Retrieved on 10th November, 2005 from
http://www.educause.edu/ir/library/pdf/CMR0033.pdf

[13] Relational Online Analytical Processing. (2005). Retrieved on 25th December, 2005 from http://searchoracle.techtarget.com/sDefinition/0,,sid41_gci214582,00.html

[14] Kimball, R. and Ross, M. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 2nd Edition, Wiley, New York, 2002.

[15] Inmon, W. H. Building the Data Warehouse, 2nd Edition. New York: John Wiley and Sons, 1996.

[16] Haisten, M. (2002) Data Warehousing: What's Next?. DM Direct. Retrieved on 25th December, 2005 from http://www.dmreview.com/article_sub.cfm?articleId=5176