Investigating the Status of Data Mining in Practice

Teh Ying Wah and Zaitun Abu Bakar
University of Malaya, Kuala Lumpur, Malaysia

tehyw@um.edu.my zab@um.edu.my

Abstract

This paper is based on a survey carried out in the Malaysian environment. The paper starts with a definition of data warehouses, data mining and this is followed by a description of its current status in the Malaysian data mining environment. This is followed by a discussion on why data mining is a great challenge for an implementation in the Malaysian environment. Based on the feedback obtained from the respondents, a conclusion is drawn on the appropriateness of the data mining techniques in the Malaysian environment

Keywords: Data Mining, Survey, Data Warehouses

Introduction

In the mid-1980s, the Malaysian government made serious efforts to get the banks to merge as a result of the economic downturn and the problems besetting small banks. The mergers were to be carried out in a staggered manner, as shown in Table 1 (***Bank Negara explains rationale for bank mergers***, 1999). The Governor of Bank Negara (National Bank), Tan Sri Ali Abul Hassan Sulaiman, claimed that Malaysia's domestic banking institutions would be able to face the pressure and challenges arising from an increasingly competitive global environment (***Bank Negara explains rationale for bank merger***, 1999). As a result of the consolidation that took place in the Malaysian banking sector, twenty-one domestic commercial banks were merged into ten anchor banks by 31 December 2001 (*Bank Negara*, 2002).

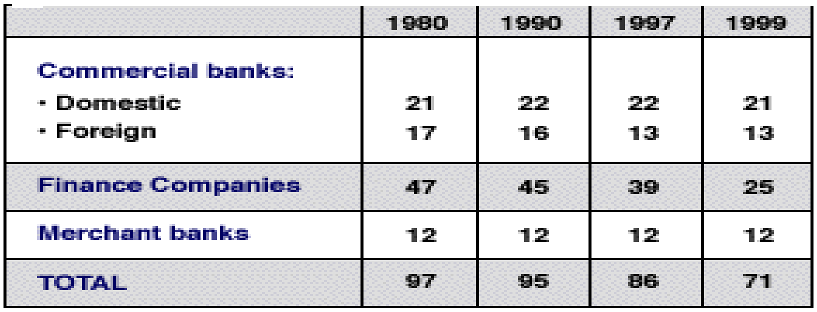
Banks mergers have played an important role in the development of data warehousing. In fact, the banking industry has been the leader in the use of data warehouses (Gupta, 1997). For example, when a banking unit uses different operational systems in the different branches, the top management still needs to view the consolidated business and manage the associated risks accordingly.

Data warehouses mean different things to different people. According to the original definition of Bill Inmon (1996), the father of data warehouses, a data warehouse is *a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process.*

As such, data warehouses are the foundation of the business IT infrastructures that collect data from several dispersed information sources and are designed to allow decision makers have prompt access to information for purpose of reporting.

Data mining is *a variety of techniques such as neural networks, decision trees or standard statistical techniques to identify nugget of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in areas such as decision support, prediction, forecasting, and estimation* (Dan et al., 2001).

Given a query, there might be many irrelevant attributes, irrelevant tuples, or both irrelevant attributes and tuples. Data warehouses use redundant data structures (e.g. indexes and materialised views) to efficiently process complex queries. Determining the most right set of redundant data structures is a complex physical design problem (Gupta, 2002). Incorporating data mining techniques in the physical design of a data warehouse helps us select an appropriate set of redundant data structures.



	1980	1990	1997	1999
Commercial banks:				
• Domestic	21	22	22	21
• Foreign	17	16	13	13
Finance Companies	47	45	39	25
Merchant banks	12	12	12	12
TOTAL	97	95	86	71

Table 1: Merger of Malaysian Banks

This paper describes and discusses the findings of a survey on the status of using data mining tools in decision-support systems and in the physical design of data warehouse. This survey constitutes the quantitative part of the research. The study elicits information and relevant data on the growth rate of data warehouse capacity, data warehouse platform, data warehouse performance tuning process, existing tools to help in improving the response time of a query and classification of users in the decision-support system. The data collected will direct the study towards building an overview on the use of data mining in a decision-support system, the physical design of data warehouse, and identifying related problems and solutions.

The Survey

The questionnaire survey method was used to investigate the status of using data mining tools in the physical design of data warehouse. This method is appropriate because data can be gathered from a large number of organisations with a wide variety of experiences and practices.

The survey consisted of two parts. The first part of the survey gathered current information on the systems in the organisation. It elicits information on awareness of data warehouse capacity, understanding of data warehouse performance tuning process, understanding of existing of data mining tools, and the type of users with access to the decision-support system. The second part of the survey collected data on the application of data mining in the physical design of a data warehouse.

Methodology

The organisations selected for the survey utilise computers in their daily operations. They were assured that all information given would be treated with the strictest confidence. The questionnaires were posted or sent by email to 100 organisations, and they were given a period of one month to respond. Some of the responses were collected via face-to-face interview with the company operation/management staff of companies. The returned questionnaires were checked for consistency of the answers and for completeness. The data were coded and analysed using the statistical software package, SPSS version 11.

Figure 1 represents the conceptual model used in this study. The model for this study tests certain factors that apply data mining tools to improve the response times of queries. In this model, factors that affect the implementation of data mining tools include end-users (data warehouse administrators or decision makers) involvement and non-end-users involvement. To implement data mining successfully, factors that hinder this need to be resolved by the data warehouse administrators or decision makers.

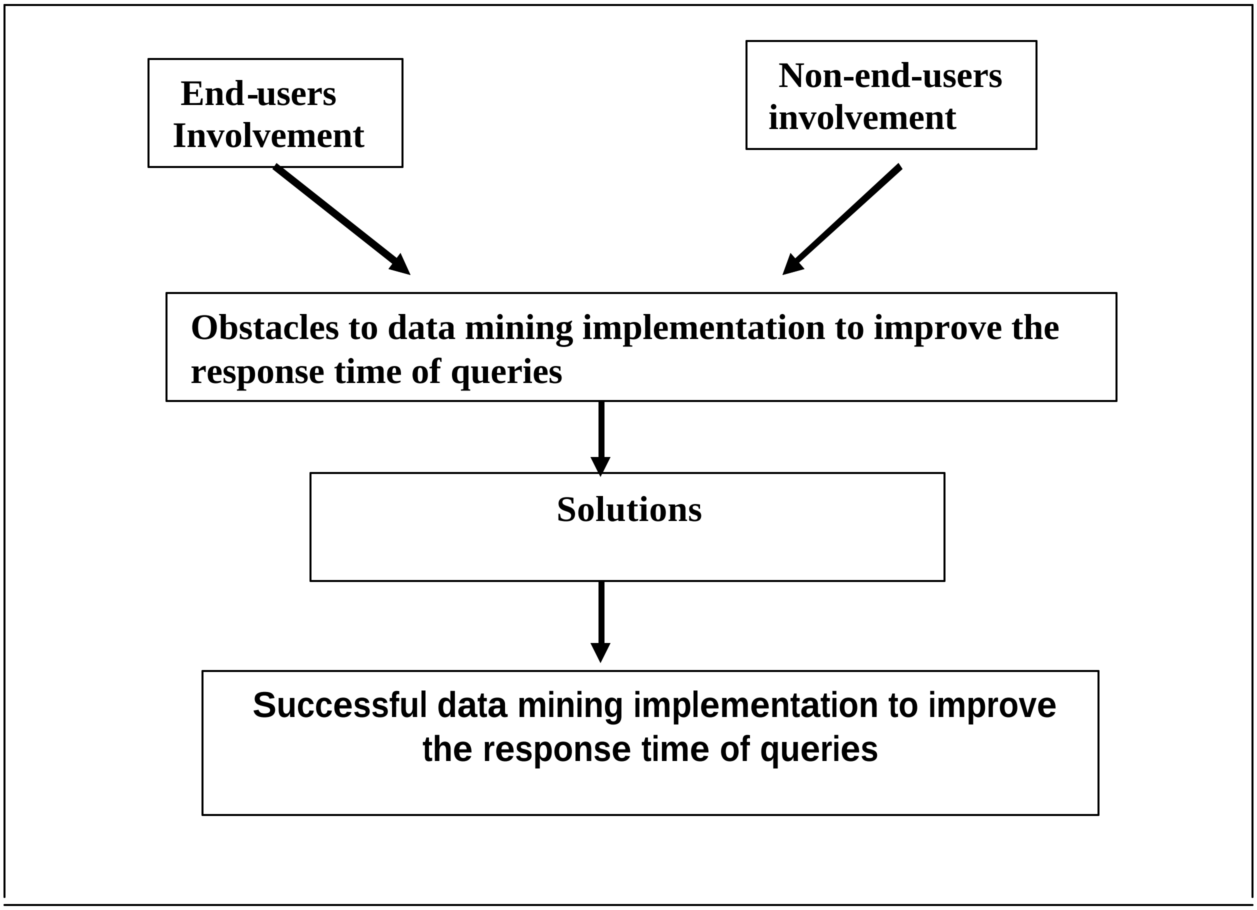

End-users Involvement
Non-end-users involvement
Obstacles to data mining implementation to improve the response time of queries
Solutions
Successful data mining implementation to improve the response time of queries

Figure. 1: The Conceptual Model

End-users (data warehouse administrators or decision makers) involvement plays an important role in the successful implementation of information technology project. Data mining implementation is no exception. End-users who have been given adequate training on data mining can contribute to its successful implementation. Developing countries, such as Malaysia, rely heavily on data mining experts to give training to end-users. With effective technology transfer and systematic training, data mining can be successfully implemented in the country.

Apart from end-user involvement in data mining implementation, there are other matters that relate to software cost, user-interface and database (DB) issues that affect implementation. To determine the actual reasons to why data mining is not implementation to improve the response time of queries, the following hypothesis is formulated:

H1. Data mining is not implemented to improve the response time of queries is due more to the end-users related rather non-end users related problems.

Rick (2002) states that within the next five year many businesses will be managing a petabyte (10^{15}) of data, which is equivalent to 250 billion pages of text, or enough to fill 20 million four-drawer filing cabinets. Business transactions in electronic format are continuously growing. The efficient processing of information is crucial for the business to function well and profitably. SQL processing usually accounts for 60 to 90 percent of the computer resources of a relational database server. Normally, over 60 percent of performance problems in database applications are caused by poorly performing SQL statements, and the performance of at least 30 percent of all SQL statements can be extensively improved. Most data

Most data warehouse administrators, when asked, would say that 90 percent or more of data warehouse application performance problems are due to poor SQL performance (Frank & Richard, 2001).

Data warehouse performance improvement through SQL tuning, can result in vast savings by delaying expensive hardware upgrades, avoiding time-consuming (therefore costly) data warehouse redesign or reconfiguration, and improving business productivity (Frank & Richard, 2001).

This research investigated the current tools that can be applied to address the problems relating to the response time of queries. The objective is to find out whether the decision makers or data warehouse administrators know how some of the existing tools help to improve the response time of a query. New and existing data mining techniques were integrated to select relevant attributes, relevant tuples, or both relevant attributes and tuples to form redundant data structure as a complete solution to improve the response time of queries. The next question is, what are the critical success factors? Hence, the following hypothesis is formulated:

H2. Factors that contribute to the success of data mining implementation to improve the response time of queries are more the end-users related rather than non- end-users involvement.

In the survey, the questionnaire developed used questions, which are open ended and having multiple-choice answers and using Likert Scale answers. Data collection has been constituted by the response to the questionnaire and mostly sent by e-mail. Some of the responses were collected via face-to-face interviewing with company operation and management staff.

Discussion of Results

The questionnaires were sent to 100 Malaysian business organisations, randomly chosen. Forty-two questionnaires were returned – a response rate of 42%. Eight business organisations, among those that responded, do not have any digital data transmission, but, they utilise computers in their daily operation

All the 42 business organisations, a great percentage (80.95%) use electronic data transmission for business purpose. This is a positive sign indicating that the business community is keen to adopt electronic communication as a major business medium. It has been reported that *e-business revenues will increase from $61 billion in 2001 to $148 billion in 2005* (James, 2001). Sixteen (38.1 %) business organisation indicated their current data warehouse capacity for their decision-support systems to be more than or equal to 1 GB. In the next five years, fifteen (35.71 %) organisations indicated their data warehouse capacity for their decision-support systems to be more than or equal to 1 GB. Based on a recent survey conducted by META group (Dave, 2002) *over 90 percent of Global 2000 companies reported that having less than 10 GB of data.* After 12 months of implementing a data warehouse, 43% of the companies project that the size of the data warehouse will be between 10 GB to 250 GB (Dave, 2002).

Data warehouse performance tuning is an important part of the management and administration of a decision-support system. It helps data warehouse administrators or decision makers to improve the response time of a query. Table 2 shows that most responding companies (71.43%) understand the processes of data warehouse performance tuning.

Table 2: Companies and Their Understanding of Data Warehouse Performance Tuning Process

	Yes	**No**
Understand data warehouse performance tuning process	30 (71.43%)	12 (28.57%)

Table 3, indicates that 59.52% of responding companies, do not use any tool to improve the response time of queries. The data warehouse server takes care of most of the tuning work (Auto-configuring, self-tuning). Most of the responding companies (64.29%) suggested having a better tool for the data warehouse administrators or decision makers to optimise the performances of their data warehouse. Most of the responding companies (85.71%) are not sure and do not use any data mining tools in their decision-support system. Only a few companies (14.29%) use any data mining tools in their decision-support system. Most of responding companies (95.24%) are not sure and do not use any data mining tools to improve the response time of queries. One of responding company states that they use SQL Server Data Transformation Services (DTS) to improve the response time of queries. As is known, the function of SQL Server DTS are data-manipulation utility services in SQL Server 7.0, and provide import, export, and data-manipulating services between OLE DB, ODBC, and ASCII data stores. DTS is not a data mining tool. This highlights the need to have data mining tools to help the data warehouse administrators or decision makers.

Table 3 Understanding Data Mining Tools

The responding companies do not use any tools to improve the response time of queries	25 (59.52%)
The responding companies use any tools to improve the response time of queries	17 (40.48%)
The responding companies suggest having a better tool for data warehouse administrators or decision makers to optimise the performance of their database	27 (64.29%)
The responding companies think that the current tools good are enough for data warehouse administrators or decision makers to optimise the performance of the database	15 (35.71%)
The responding companies have never used data mining tools in decision-support systems	21 (50%)
The responding companies are not sure whether any data mining tools have been used in decision-support systems	15 (35.71%)
The responding companies never used data mining tools in decision-support systems	6 (14.29%)
The responding companies have never used data mining tools to improve the response time of queries	28 (66.67%)
The responding companies are not sure whether any data mining tools have been used to improve the response time of queries	12 (28.57%)
The responding companies have used data mining tools to improve the response time of queries	2 (4.76%)

Reasons for Not Implementing Data Mining to Improve the Response Time of Queries

Table 4 lists the reason for not implementing data mining to improve the response time of queries. Three reasons have mean values higher than 3: *Lack of required expertise, high software cost* and *lack of knowledge about data mining*. Other reasons are not considered significant.

Reasons	Mean
1. Lack of required expertise	3.44
2. High software cost	3.31
3. Lack of knowledge about data mining	3.06
4. High training cost	2.88
5. It is difficult to use	2.69
6. Difficult to improve the response time of queries	2.56
7. Data Mining only handles the logical level of a data warehouse rather than physical level of a data warehouse	2.56
8. Data warehouse self-tuning takes care of improving the response time of queries	2.31
9. Small data warehouse capacity	2.31
10. Lack of enthusiasm	0.31
11. Advancement in DB technology	0.31

Table 4: Reasons for Not Implementing Data Mining to improve the response time of queries

Obstacles to Data Mining Implementation to Improve the Response Time of Queries

The respondents were asked to express in their opinions concerning the difficulty in implementing data mining to improve the response time of queries. Table 5 summarises the data collected.

There are 3 reasons with mean value above 3 and the important reasons given in order of significance are:

1. *Lack of required expertise*
2. *High software cost*
3. *High training cost*

Reasons	Mean
1. Lack of required expertise	3.81
2. High software cost	3.31
3. High training cost	3.31
4. Data warehouse's self tuning takes care of improving the response time of queries	2.88
5. Lack of knowledge about data mining	2.75
6. Not user friendly interface	2.5
7. Lack of enthusiasm	0.31
8. Advancement in technology in DB	0.31

Table 5: Obstacles to data mining implementation to improve the response time of queries

Factors that Contribute Towards the Success of Data Mining Implementation to Improve the Response Time of Queries

Table 6 lists 9 factors that contribute towards the success of data mining implementation to improve the response time of queries.

Factors	Mean
1. Sufficient knowledge about data mining	3.94
2. Availability of the required expertise	3.69
3. Support from top level management	3.56
4. Low software cost	3.43
5. Low training cost	3.31
6. User-friendly interface	3.25
7. Data mining is able to interact with multiple database platforms	3.25
8. Lack of enthusiasm	0.31
9. Advancement in technology in DB	0.31

Table 6: Factors that contribute towards the success of data mining implementation to improve the response time of queries

Seven (7) factors have a mean value of higher than 3. The following top five factors in the order of significance:

1. Ensure sufficient knowledge about data mining
2. The availability of required expertise
3. Support from top level management
4. Low software cost
5. Low training cost

Conclusion of the Survey

To test hypothesis H1, the reasons were classified into end-users and non-end-users related reasons. The statistical package SPSS version 11 was then used to compute the mean for each group.

<u>The end-users related reasons are:</u>

1. *High training cost*
2. *Lack of required expertise*
3. *Lack of knowledge about data mining*
4. *Lack of enthusiasm*

<u>The non-end-users related reasons are:</u>

1. *High software cost*
2. *Data warehouse self tuning takes care of improving the response time of queries*
3. *No user-friendly interface*
4. *Advancement in DB technology*

In Table 7 shows the mean ranking given to end-users related reasons significantly higher than the mean ranking of the non-end-users related reasons. This infers that the reasons why *data mining is not carried out to improve the response time of queries are due more to end-users involvement rather than non-end-users involvement.*

Pair	**Total Mean**	**Mean of Group**	**Mean Difference** End-users related reasons – Non-end-users related reasons
End-users related reasons	10.19	2.55	0.3
Non-end-users related reasons	9	2.25	

Table 7: Paired sample test between end-users and non-end-users related reasons to why data mining is not carried out to improve the response time of queries

The results of the survey showed that there are 9 factors considered important enough to contribute to the success of data mining implementation to improve the response time of queries (Table 6). To test the hypothesis H2, the success factors were classified into end-users related reasons and non-end-users related reasons. SPSS was used to compute the mean for each group. The results are summarised in Table 8.

The end-user related reasons are:

1. Ensure sufficient knowledge about data mining
2. The availability of required expertise
3. Support from top-level management
4. Low training cost
5. Lack of enthusiasm

The non-end-user related reasons are:

1. Data mining is able to interact with multiple database platforms
2. Low software cost
3. User-friendly interface
4. Advancement in DB technology

Pair	**Total Mean**	**Mean of Group**	**Mean Difference** End-users related reasons – Non-end-users related reasons
End-users related reasons	14.81	2.96	0.4
Non-end-users related reasons	10.25	2.56	

Table 8: Paired sample test between end-users related reasons and non-end-users related reasons success factors of data mining implementation to improve the response time of queries

On the average, end-user related reasons received higher ranking than the non-end-user related reasons. This infers that success factors of data mining to improve the response time of queries are more end-users related reasons rather than non-end-users related reasons. The conceptual model can now be refined and filled in with the answers for the research questions. The result is depicted in Figure 2.

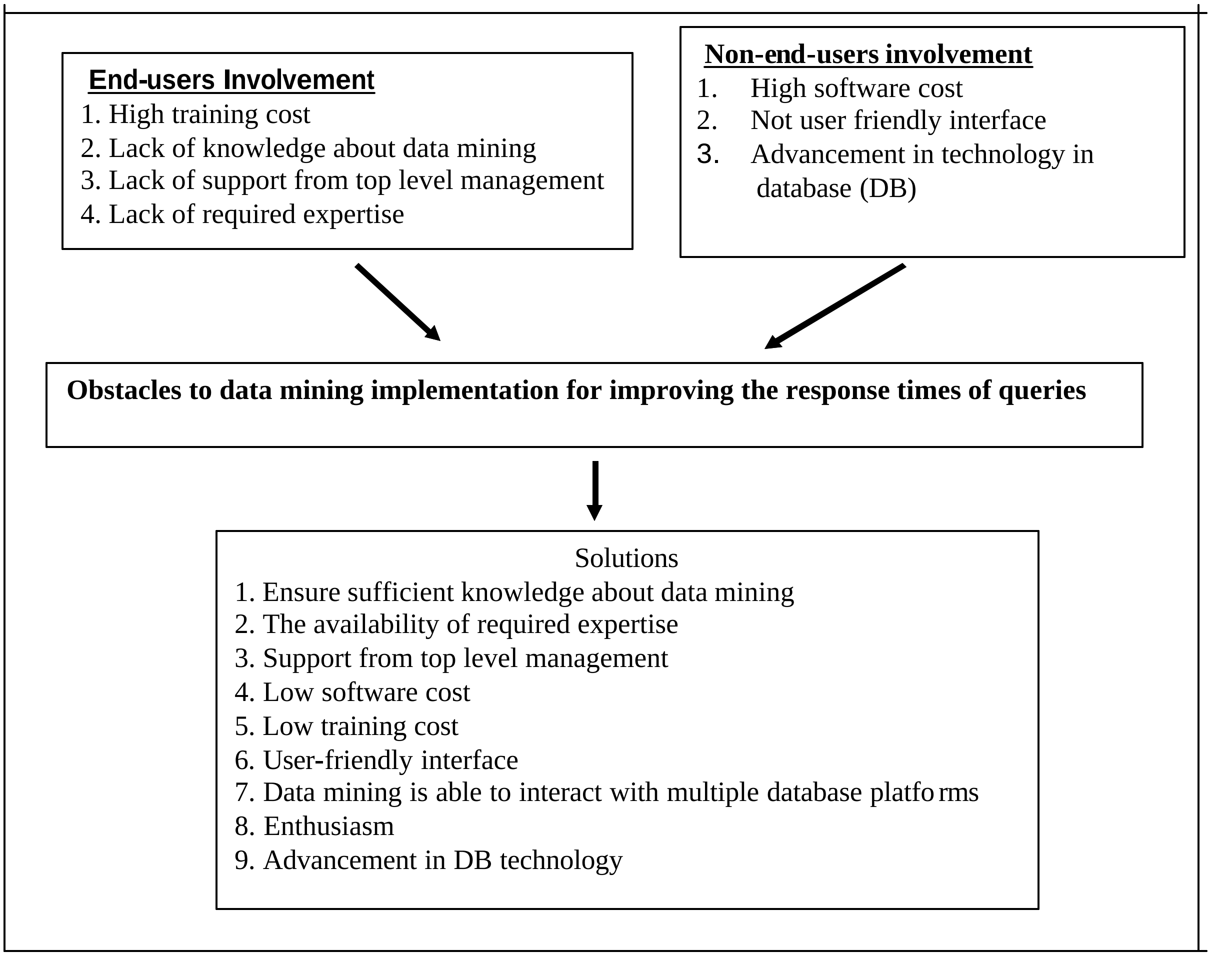

End-users Involvement
1. High training cost
2. Lack of knowledge about data mining
3. Lack of support from top level management
4. Lack of required expertise
Non-end-users involvement
1. High software cost
2. Not user friendly interface
3. Advancement in technology in database (DB)
Obstacles to data mining implementation for improving the response times of queries
Solutions
1. Ensure sufficient knowledge about data mining
2. The availability of required expertise
3. Support from top level management
4. Low software cost
5. Low training cost
6. User-friendly interface
7. Data mining is able to interact with multiple database platforms
8. Enthusiasm
9. Advancement in DB technology

Figure 2. The Refined Conceptual Model

References

Bank Negara explains rationale for bank mergers. (1999, August 10). *The Star*.

Bank Negara Malaysia. (2002). Consolidation of the Banking Sector. Available from World Wide Web: http://www.bnm.gov.my/en/News/releases.asp?yr=2002&sid=0128a Last modified : 22 March, 2002.

Chaudhuri, S., Dayal U., Ganti, Venkatesh. (2001). Database Technology for Decision Support Systems. *Computer*. IEEE Computer Society.

Frank I. and Richard T. (2001, August). *SQL Optimization for the Data Warehousing Environment.* Lecco Technology. Available from World Wide Web: http://www.hkcs.org.hk/dbwp1805.doc

GeneCards: Knowledge Discovery In Biology and Medicine. (2002). Available from World Wide Web: http://bioinfo.weizmann.ac.il/cards/knowledge.html Last modified: February 26 2002.

Gupta, V. R. (2002). An Introduction to Data Warehousing. Available from World Wide Web: http://system-services.com/dwintro.asp Last modified : 22 March, 2002.

Inmon, W. (1996). *Building the Data Warehouse*. John Wiley & Sons, Inc. 2nd Edition, 1996.

Rick W. (2002). Tower of power. *Information Week*, February 11, 2002, Available from World Wide Web: http://www.informationweek.com/story/IWK20020208S0009