2006:38

MASTER'S THESIS

Data Warehouse Maintenance

Improving Data Warehouse Perfomance through Efficient Maintenance

Asad Javed
Sardar Saad Rafique

Luleå University of Technology

Master Thesis, Continuation Courses
Computer and Systems Science
Department of Business Administration and Social Sciences
Division of Information Systems Sciences

2006:38 - ISSN: 1653-0187 - ISRN: LTU-PB-EX--06/38--SE

To Our
Beloved Parents

ABSTRACT

Data warehousing has been a buzz word in the IT industry since mid 90's. Researchers are constantly involved in finding new and improved ways for the design and development of data warehouses. But unlike traditional operational information systems used for running the day to day business of an organization, data warehouses require a lot more maintenance and support. The real work of taking output from the data warehouse depends largely on how it is managed. Although a lot of research is going on to enhance the design and development of data warehouses, very little effort has been spent on the maintenance side. Without proper maintenance data warehouse is not going to give the desired output which is expected of it. As data warehousing projects are very expensive, it is extremely desirable that it gives the desired results and functions smoothly. In this research study we have tried to figure out the currently available maintenance methods being used by the industry today to enhance the data warehouse performance. First we have gathered data from the books, journals, articles and the internet to see which maintenance mechanisms are available. Than we have gathered data related to data warehouse maintenance from a company using data warehouse and finally we have compared the theoretical findings with the real world findings and gave our opinion on the best possible strategies to improve data warehouse performance through efficient maintenance.

ACKNOWLEDGEMENT

We are really thankful to God Almighty (Allah) enabling us to achieve another goal in our lives.

We are really thankful for guidance and motivation given by our supervisor Jörgen Nilsson to complete our work professionally. During the time spent with him in meetings and discussions during the thesis we found him really cooperative and helpful and he tried his level best to guide us in the right direction. We would specially like to thank Mr. Lars Furberg of Information Systems Department for giving us the starting guideline in data warehouses. Additionally we would like to thank Mr. Junaid Rafique, data warehouse consultant at NCR Corporation, Pakistan for all his support and help during the entire process, arranging interviews with personnel from Telenor Pakistan and giving us a lot of help in the technical aspects of thesis.

I, Sardar Saad Rafique would like to thank my parents, brothers Junaid and Ahmed and sister Rahima for supporting me throughout my study time in Sweden. It would have been impossible for me to complete the studies and thesis without their love, affection and encouragement.

I, Asad Javed would like to thank my parents and my brother for their encouragement and support during my study period. Also I am thankful to every member of the Information Systems Department for bearing us and supporting during the entire program.

---------- ----------

Sardar Saad Rafique Asad Javed

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

1.1 BACKGROUND 1
1.2 DATA WAREHOUSE MARKET 3
1.3 DATA WAREHOUSE 4
1.4 GOALS OF DATA WAREHOUSE 5
1.5 PROBLEMS IN DATA WAREHOUSING 6
1.6 PROBLEMS AFTER DEPLOYMENT OF DATA WAREHOUSE 6
1.7 IMPORTANT TERMS 7
1.7.1 Data Warehouse Architecture 7
1.7.2 Schema 11
1.7.3 View 11
1.7.4 Views, OLAP and Warehousing 12
1.7.5 Corporate Information factory (CIF) 13
1.8 AIM OF STUDY 14
1.9 RESEARCH QUESTION 15
1.10 DELIMITATIONS 16
1.11 DISPOSITION OF THESIS 16

CHAPTER 2: METHODOLOGY 18

2.1 RESEARCH APPROACH 18
2.2 RESEARCH STRATEGY 20
2.3 DATA COLLECTION 21
2.4 DATA ANALYSIS 22
2.5 VALIDITY AND RELIABILITY 23

CHAPTER 3: LITERATURE REVIEW 25

3.1 DATA WAREHOUSE PERFORMANCE MANAGEMENT 26
3.2 DATA WAREHOUSE MAINTENANCE 27
3.3 PERFORMANCE TUNING MECHANISMS 27
3.3.1 Communication and Training 28
3.3.2 Help Desk and Problem Management 29
3.3.3 Network Management 29
3.3.4 Capacity Planning 29
3.3.5 Data Loading Performance 30
3.3.6 Query Management 31
3.4 COMMUNICATION PROCESS 32
3.4.1 Communication Process Implementation 32
3.5 TRAINING PROGRAM 33
3.5.1 What Should Be Taught 34
3.5.2 The Training Program Implementation 35
3.6 ROLE OF HELP DESK 36
3.6.1 Help Desk Services 37
3.6.2 Developing a Help Desk 38
3.7 THE PROBLEM MANAGEMENT PROCESS 39
3.7.1 Problem Management Process Development 39
3.8 NETWORK MANAGEMENT 40
3.9 SOFTWARE AND HARDWARE ISSUES 43
3.9.1 Implementation Strategies 43
3.9.2 The Certification Testing Process 45
3.9.3 Certification Program 46
3.9.4 Certification Business Requirements 46
3.9.5 Certification Test Results 47
3.10 EXTRACT, TRANSFORM AND LOAD (ETL) 47
3.10.1 A Typical ETL Process 49
3.11 VIEW MATERIALIZATION 52
3.11.1 Maintenance of Materialized Views 55

3.11.2 View Maintenance Policies *57*

CHAPTER 4: EMPIRICAL FINDINGS **58**

4.1 INTRODUCTION 58
4.2 COMMUNICATION & TRAINING 61
4.3 HELP DESK & PROBLEM MANAGEMENT 62
4.4 NETWORK MANAGEMENT 63
4.5 SOFTWARE & HARDWARE ISSUES 64
4.6 EXTRACT, TRANSFORM AND LOAD (ETL) 65
4.7 VIEW MAINTENANCE 66

CHAPTER 5: ANALYSIS AND DISCUSSIONS **67**

5.1 COMMUNICATION & TRAINING 67
5.2 HELP DESK & PROBLEM MANAGEMENT 68
5.3 NETWORK MANAGEMENT 69
5.4 SOFTWARE & HARDWARE ISSUES 69
5.5 EXTRACT, TRANSFORM & LOAD (ETL) 70
5.6 VIEW MAINTENANCE 70

CHAPTER 6: CONCLUSION AND DISCUSSIONS **72**

6.1 CONCLUSION AND DISCUSSION 72
6.2 FURTHER RESEARCH 75

REFERENCES **76**

APPENDIX A **80**

LIST OF FIGURES

1.1: Data warehouse in an Organization

1.2: Data Warehouse Architecture

1.3: Example of a Schema

1.4: CIF Architecture

2.1: Types of Research

2.2: Components of Qualitative Data Analysis

3.1: ETL function

3.2: ETL in Corporate Information Factory

3.3: Materialized views within a client server DBMS

LIST OF TABLES

3.1: Who should be trained in what areas

4.1: Three levels of education and training

CHAPTER 1: INTRODUCTION

This chapter will explain the background of the research study under subject. Starting with introduction to concept of data warehousing and data warehouse market it will provide the basic concepts associated with data warehouses. Additionally we have formulated the research question and aim of study in this chapter. Finally the disposition of the thesis is given.

1.1 Background

In the current scenario of changing business conditions organization's management needs to have access to more and better information. Most organizations are now days operating using information technology as the backbone of their operations but the fact is that despite having a large number of powerful desktop and notebook computers and a fast and reliable network, access to information that is already available within the organization is very difficult or otherwise not possible (KO00). All organizations whether large or small using Information Technology for the operations produce large amount of data about their business including data about sales, customers, products, services and people. But in most cases this data remains in the operational systems and can't be used by the organization. This phenomenon is called 'data in jail'. Experts say that only a small portion of this data that is entered, processed and stored is actually available to decision makers and management of the enterprise. The unavailability of this data can cause significant reduction in sales and profits of organizations and vice versa. (ibid)

In the 1990's as large scale organizations began to need more timely data about their business, they found that traditional information systems technology was simply too slow and complex to provide relevant data efficiently and quickly(WI06). Completing reporting requests could take days or weeks using traditional reporting tools that were designed more or less to 'execute' the business rather than 'run' the business. As a cure for this problem the concept of data warehouse started as a place where relevant data could be held for completing strategic reports for management. The key here is the word 'strategic' as most executives were less concerned with the day to day operations than they were with a more overall look at the model and business functions. (ibid)

In the latter half of the 20th century, there existed a large number and types of databases (WI06). Many large businesses found themselves with data scattered across multiple platforms and variations of technology, making it almost impossible for any one individual to use data from multiple sources. A key idea within data warehousing is to take data from multiple platforms/technologies (As varied as spreadsheets, DB2 databases, IDMS records, and VSAM files) and place them in a common location that uses a common querying tool. In this way operational databases could be held on whatever system was most efficient for the operational business, while the reporting / strategic information could be held in a common location using a common language. Data Warehouses take this even a step farther by giving the data itself commonality by defining what each term means and keeping it standard. An example of this would be gender which can be referred to in many ways, but should be

standardized on a data warehouse with one common way of referring to each sex. The purpose behind all these developments was to make decision support more readily available and without affecting day to day operations. One aspect of a data warehouse that should be stressed is that it is NOT a location for ALL of a businesses data, but rather a location for data that is 'interesting' and 'important'. Data that is interesting will assist decision makers in making strategic decisions relative to the organization's overall mission (ibid).

Significant users of this technology include retail giants such as Wal-Mart, credit card companies such as Visa and American Express, and major banks and transportation companies which include Bank of America, Royal Bank of Canada, Allied Irish Bank, United Airlines, Continental Airlines and many more (NCR05). Planning the design and construction of these huge and complex information repositories have led to the development of data warehousing. Its major role remains crucial in understanding, planning, scoping, and delivering knowledge capital back to the enterprise in a timely and cost effective fashion (ibid).

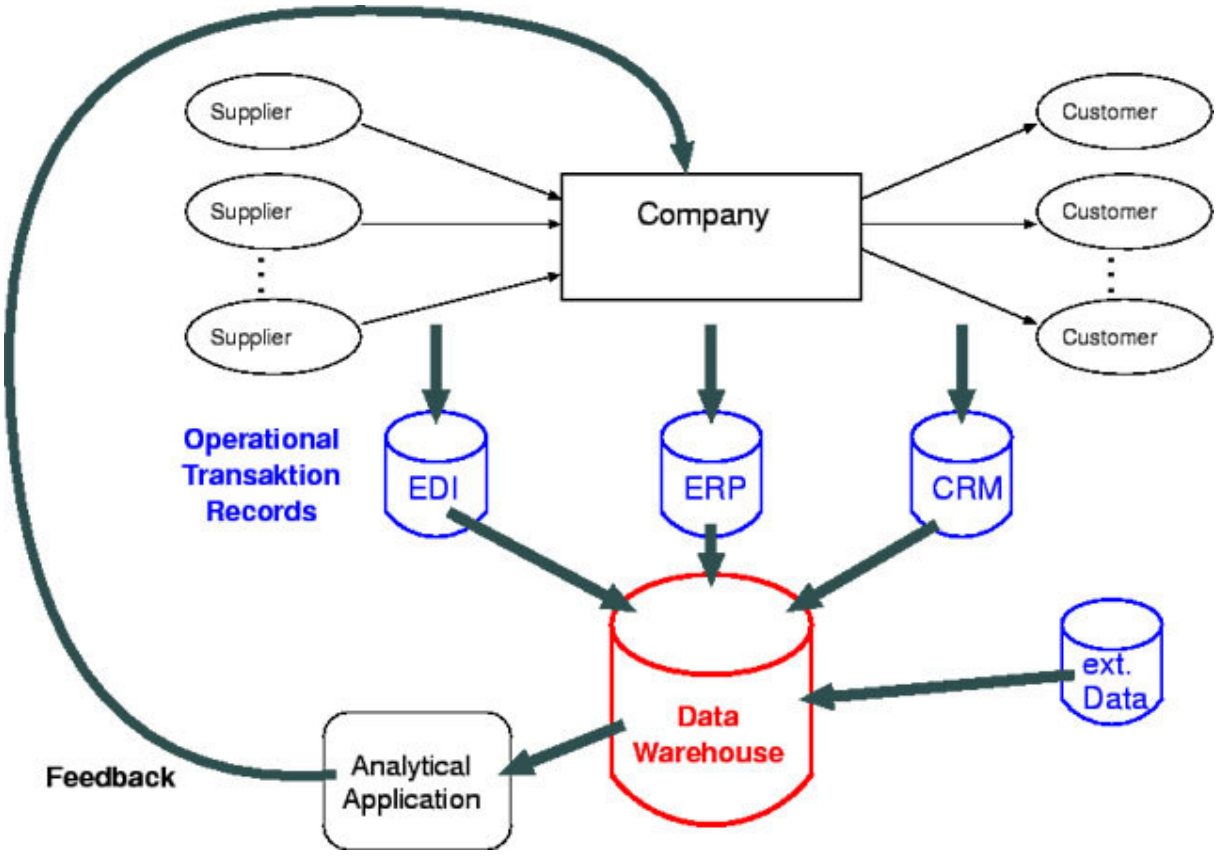

Fig 1.1: Data Warehouse in an Organization
http://wwwai.wu-wien.ac.at/~hahsler/research/datawarehouse_webster2001/talk/node4.html

Another important concept that has come out of the data warehousing concept is the recognition that there are two different types of information systems in all organizations namely operational systems and information systems (KO00). Operational systems are used to

perform the day to day operations of the organization. They function as a backbone to any enterprise. For e.g. order entry, inventory control, payroll, accounting etc are all operational systems. Because of their importance the operational systems are always the first to be computerized in an enterprise. In fact most of the organizations around couldn't operate without these operational systems. On the other hand there are other functions within the organization which have to work with planning, forecasting and management. These functions are quite different from operational functions. For e.g. resource planning, financial analysis and strategy planning etc. These functions require a lot of support from operational systems but these are actually different from operational systems. These are knowledge based systems called informational systems (ibid).

1.2 Data Warehouse Market

The concept of data warehousing was in the industry since the early 1980's but during the early 90's it's real importance was recognized. Since than virtually every global 2000 company has acquired some form of data warehousing technology and is using it in some form for decision support (CG04). During the early stages of data warehouse evolution most industry professionals were thinking that this technology will develop at a very rapid pace but the reality is not the same. Not very much has been accomplished market wise since its evolution. The users of data warehouse still complain about the problems of data quality, metadata management and warehouse maintenance. Users still complain that they can't get the required results from the data warehouse.

Enterprise data warehousing (EDW) is a submarket within the overall data warehousing/business intelligence market. Companies considering investing in an EDW solution have matured to a point where data marts (small data warehouses intended for a particular function or department of an enterprise) can no longer satisfy the organization's increased need for higher-quality and more timely business analytics. These organizations seek a platform solution that can handle the demands of multiple subject areas and larger numbers of concurrent users, all while providing end users with the freedom to ask any question. Indeed, the warehousing market has reached a point where need, opportunity, and capability have merged. Meta group believe that these forces will drive double-digit growth of the EDW market through 2008.

By the end of the 1990s, most global 2000 companies had finished the major task of implementing an ERP software infrastructure. Internet has totally changed the business scenario and most of the companies have changed their existing transactional infrastructure rapidly into the web model. Missing in all this was an organization's ability to create meaning from all the transactional and sub-transactional (e.g., usage, traffic levels) data being captured. Indeed, META Group research indicates that 77% of organizations plan to capture even more detailed business data in 2004 than was captured in 2003.

There are a number of companies competing in the data warehouse industry but still not a single vendor can address all the needs of one customer. The data warehouse market is evolving continuously and rapidly. Each vendor that joins the battle is hoping to address the concerns of at least a slice of what is estimated to be a $4Billion market currently, and which will grow to an estimated $9.9Billion by 2008. Through 2006, Meta Group expects to see

vendors increase their sales, marketing, and development focus on this market as the transaction processing market recedes in emphasis. This will mean larger services organizations for some vendors, while others will consider solidifying or expanding relationships with third-party value-added resellers. Some consolidation of end-user business intelligence tools (e.g., Business Objects, Cognos), extract/transform/load vendors (e.g. Informatica, Ab Initio), or boutique services firms that specialize in data warehousing is also likely to be seen through 2007. For larger organizations, this will mean investing in data mart consolidation projects. For companies of virtually any size, it will mean increased development and planning to achieve analytic maturity (ibid).

Currently there are a lot of companies working in the data warehouse sector among which Teradata (a subsidiary of NCR), IBM and Oracle are the most prominent with their specialized data warehousing products.

1.3 Data Warehouse

According to (WR94) a data warehouse is a subject oriented (high level entities of enterprise for e.g. customer, product), integrated (consistent naming convention, consistent variables, consistent attributes of data), time variant (data obtained over a long period of time), and nonvolatile (arrival of new data doesn't updates previous data) collection of data to support the management's decision making. The data that enters the data warehouse comes from the traditional operational systems working in the enterprise (ibid).

A data warehouse is a copy of transaction data specifically structured for querying and reporting (RM04). It is a huge (sometimes terabytes of disk storage) database, which stores volumes of historical data for the company. The database can be of any form. It can be a relational database, multi dimensional database, flat file, or hierarchical database, etc. A data warehouse organized for a single department or for a specific group of people is termed as a data mart. (ibid)

The concept of a data warehouse came into existence as a result of two different sets of requirements (RM04). First, the end users need to view and understand the company wide view of information and second, the information system (IS) department's need to manage the data for technological and economic reasons. Although, the two requirements may seem completely different from each other, the data warehouse gurus confirm that addressing any one of these requirements will make it easier to meet the other requirements and vice versa. (ibid)

In reality the ultimate purpose of a data warehouse is to integrate enterprise wide corporate data into a single repository from which users can easily extract data according to the organizational needs for the purposes of producing reports, performance analysis and long term decision making (RM04).

In short data warehouse is a collection of decision support technologies which enable a manager, analyst or any worker in an organization to take better and effective decisions (SU96). In the recent years a tremendous increase in the number of products and the services offered relative to data warehousing in the industry has been seen.

The concept of data warehousing consists of tools, technologies, and methodologies that allow for the construction, usage, management, and maintenance of the hardware and software used for a data warehouse, as well as the actual data itself (MR05).

1.4 Goals of Data Warehouse

A major problem faced by today's organizations is the access to useful data (RK96). Although a lot of data is available to them, the data required for planning of future policy, future planning and market capturing etc is not easily available. The solution to this problem is the data warehouse. It's the place where people can access their data. All the data within the enterprise is scattered as different operational systems are placed at different locations, while in a data warehouse all the data of the enterprise is integrated and stored in a common data base. Ralph Kimball has identified 6 major goals of a data warehouse, which are described briefly:

1. It provides access to corporate and organizational data. By using a data warehouse, managers and analysts of an organization can get data easily on their personal computers or laptops. The tools available to managers and analysts are user friendly and easy to operate.

2. Data in data warehouse is consistent. When user of a data warehouse request data from it, it will generate the same data for the same query. For e.g. if a manager and analyst both want to get the sales figure for the month of January, they both will get the same data.

3. Data can be combined and separated. As rows and columns can be joined in a relational database, same is the case with a data warehouse where data can be combined or separated from other data.

4. Data warehouses have made data querying pretty easy provided you have the proper tools. Data warehouse is not just a data it also has a set of tools and techniques to query, analyze and present information.

5. It is a place to publish used data. A data warehouse is not only a place to store raw data. Rather in a data warehouse data is carefully assembled from a variety of information sources around the organization, cleaned up, quality assured, and then released only if it is fit for use.

6. The quality of data can be a guiding force for business reengineering. The data warehouse cannot fix poor quality data. The publishing of this poor quality data can lead to pressure arising within the organization when people see how valuable the data could be if only it was of better quality. In this way a data warehouse can play a key role in the business reengineering efforts in an organization (ibid).

1.5 Data Warehouse Maintenance Problem

As data warehousing is an emerging area, a lot of problems are found in the system. One of the major problems faced by the industry today is data warehouse maintenance. As data warehouses are huge systems. Lot of time is spent on data extraction, cleansing and loading process. Experts say usually 80% of the time building a data warehouse is consumed by these tasks. As the users of the data warehouse experience the capabilities of the data warehouse their demands will increase gradually. The developers of data warehouse often find problems in the operational systems from where data must be captured. It's a tough decision for the developers whether to fix the problem in the operational system or fix it in the warehouse. Sometimes data captured from operational systems needed to be validated before it can be stored in the warehouse. Typically once data are in warehouse many inconsistencies are found with fields containing 'descriptive' information. For example, many times no controls are put on customer names. Therefore, you could have 'DEC', 'Digital' and, 'Digital Equipment' in your database. This is going to cause problems for a warehouse user who expects to perform an ad hoc query selecting on customer name. The warehouse developer, again, may have to modify the transaction processing systems or develop (or buy) some data cleaning technology. Large scale data warehouses may require very large amount of disk space for storage purposes.

Data warehouses require a very high level of maintenance. Any reorganization of the business processes and the source systems may require the data warehouse to change. Updates are often needed in these situations. Keeping in view this problem we have decided to do some research on maintenance of data warehouses. Experts say that more resources are required for maintenance of a data warehouse rather than its development (LG05).

According to (SL00) there are many ways for a data warehouse project to fail. The project can be over budget, the schedule may slip, critical functions may not be implemented, the users could be unhappy and the performance may be unacceptable. The system may not be available when the users expect it, the system may not be able to expand function or users, the data and the reports coming from the data may be of poor quality, the interface may be too complicated for the users, the project may not be cost justified and management might not recognize the benefits of the data warehouse. The most important task for a data warehouse project manager is picking the right people who have the competence to manage a large enterprise data warehouse. (ibid)

1.6 Problems After Deployment Of Data Warehouse

Some of the common problems faced after the deployment of a data warehouse include (LG05):

1. Some times after the deployment of data warehouse it may be needed to delete some useless data. Someone has to make a decision which data to delete and which one to keep. The usual cause for this problem is the storage cost.

2. In a data warehouse queries to retrieve information from data warehouse need to be written. Someone has to decide which queries should be user written and which should be written by the information system.

3. The users of the system will see "holes" in the data they store in the data warehouse. Mainly for the sake of completeness, they will be tempted to add this data. Unfortunately, when they have added this data several times, they will find, the size and complexity of the data warehouse has increased substantially without proper consideration of whether the incremental size and complexity had business worth.

4. After the deployment of a data warehouse, the users will find a lot of loop holes where there are opportunities to fine tune the data warehouse.

5. The users of the data warehouse need to know which data is going where. They are uncertain in determining which reports should be generated from operational systems and which one from the warehouse.

6. The users will find problems in feeding the warehouse from source systems (operational systems). In that updates have to be applied to keep data warehouse in working order.

7. Maintaining data warehouse architecture is more difficult than establishing the warehouse architecture. The architecture here refers to the consistent use of dimensions, definitions of derived data, attribute names, and data sources for specific information.

8. Security policies may need to be changed depending on the user interaction with the system. Security should not be a hindrance in accessing useful information for the user of warehouse.

1.7 Important Terms

In order to go deep into the understanding of data warehouse maintenance, it is better to first understand the basic architecture of a data warehouse and some of the related concepts.

1.7.1 Data Warehouse Architecture

A typical data warehouse is shown in figure 1.2. Here the components of the data warehouse architecture are described in some detail (TC05).

Operational data is the data residing in the operational systems of the company. This is the data required for the day to day operations of the company. For e.g. accounts data from accounting system, reserves data from inventory system etc (TC05).

An operational data store is a repository of current and integrated operational data used for analysis (TC05). Its main purpose is to collect data from the operational systems and move that data into the warehouse. It is often structured and supplied with data in the same way as the data warehouse but may in fact perform simply as a staging area for data to be moved into the data warehouse. Operational data store (ODS) is commonly created when legacy operational systems can not perform the reporting functionality efficiently. While using ODS users have the comfort of relational database while remaining distant from the decision support functionality of a data warehouse. Building an ODS can be a helpful step towards the development of a data warehouse because an ODS can supply data that has been already extracted from the source systems and cleaned making remaining work of integrating and restructuring the data for data warehouse easy and simple. (ibid)

Load manager has the responsibility to manage and perform all the operations required for the extraction and loading of data into the data warehouse (TC05). The data can be extracted directly from the operational data systems or more commonly from the operational data store. The operations performed by the load manager may include simple transformations of data to prepare the data for entry into the warehouse. The size and complexity of this component will vary between data warehouses and may be constructed using a combination of vendor data loading tools and custom built programs. (ibid)

The warehouse manager performs all the operations associated with the management of data in the warehouse (TC05). This component is constructed using vendor data management tools and custom built programs. Operations of the warehouse manager include:

1. Analysis of data to ensure consistency

2. Transformation and merging of source data from temporary storage into data warehouse tables.

3. Creation of indexes and views on base tables.

4. Generation of de-normalizations (if necessary).

5. Generation of aggregation (if necessary).

6. Backing-up and archiving data.

Sometimes warehouse manager also generates query profiles to determine which indexes and aggregations are appropriate. A query profile can be generated for each user, group of users, or the data warehouse and is based on information that describes the characteristics of the queries such as frequency, target tables and size of result sets. (ibid)

Query manager manages all the work related to management of user queries (TC05). It is typically constructed using vendor end-user data tools, data warehouse monitoring tools, database facilities and custom built programs. Responsibilities of a query manager include directing queries to the appropriate tables and scheduling the execution of queries. Sometimes the query manager also generates query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate. (ibid)

Detailed data is the actual data stored in the database schema. In most cases the detailed data is not stored online but is made available by aggregating the data to the next level of detail. (TC05)

Lightly and highly summarized data is generated by the warehouse manager. The basic purpose of summary information is to speed up the performance of queries. The summary data is updated on a regular basis as new data enters the data warehouse. (TC05)

Archive / backup data is all the detailed and summarized data, stored for the purpose of archiving and backup. This data is usually stored on storage media such as magnetic tapes or optical disks. (TC05)

Metadata is the data about data (TC05). It provides information about data format, data mapping between source and destination systems etc. For e.g. in a university data warehouse the statement 'a good student is a student with a CGPA of 3.5' is metadata. Metadata is helpful in a lot of processes including:

1. Extraction and loading processes – metadata is used to map data sources to a common location of data within the data warehouse.

2. Warehouse management process – metadata helps in automating the production of summary tables.

3. Query management process – used to direct a query to the most appropriate data source. (ibid)

According to (RH97) metadata describes data structure, syntax, and semantics are typical key elements of metadata. Other system related data that is present in a dictionary, directory or a repository can also be regarded as metadata. The responsibility for managing metadata lies with warehouse staff, but its ownership is distributed and external (ibid).

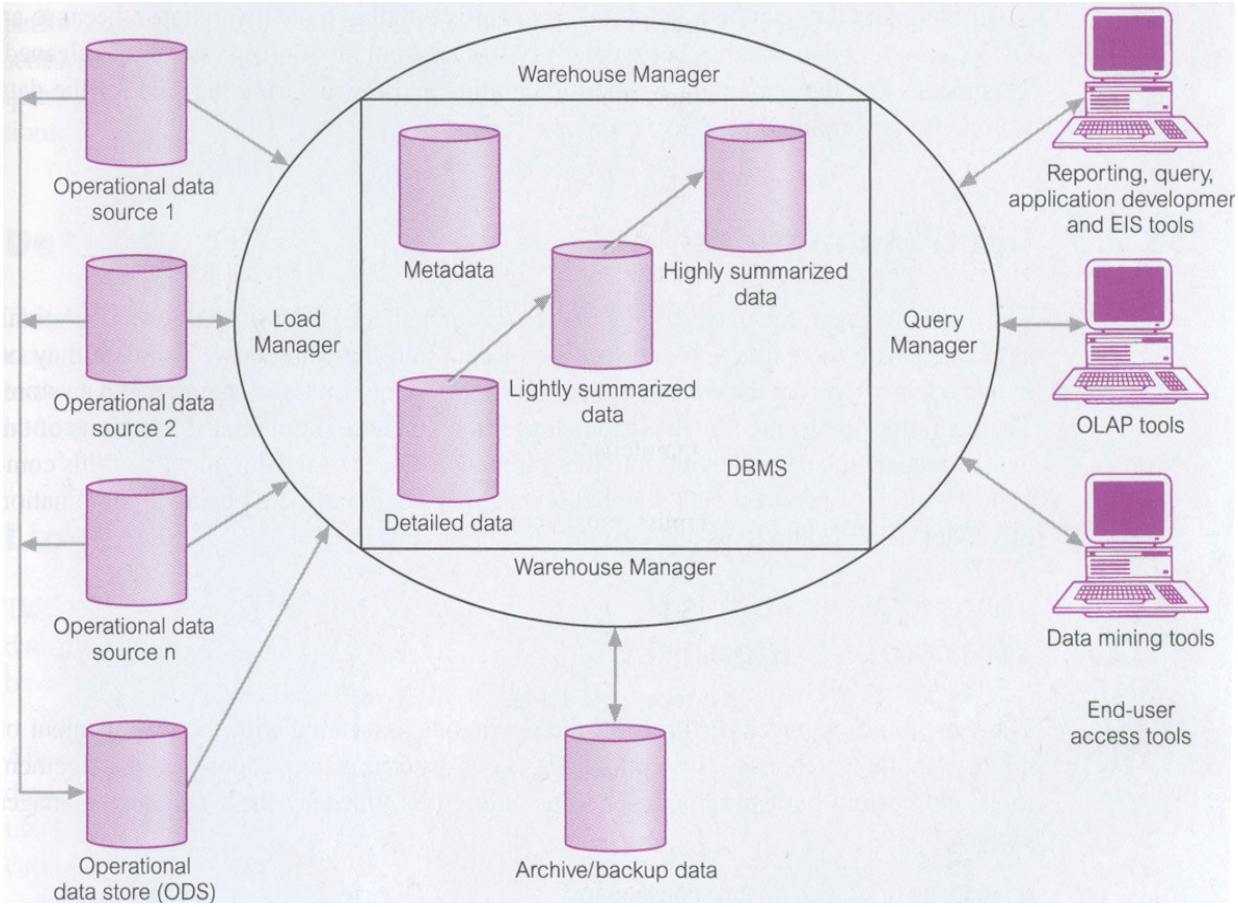

Fig 1.2: Data Warehouse Architecture (TC05, Page # 1157)

End user access tools enable the business user to extract the required data from data warehouse (TC05). The main purpose of data warehousing is to provide information to business users for strategic decision making while interacting with the warehouse using end user access tools. The principal advantage of data warehouse is efficient support of ad hoc and routine analysis and it could be achieved by pre-planning the requirement for joins, summations, and periodic reports by end-users. (ibid)

Keeping in view the context of (BS 97) user access tools can be further sub-divided into five major groups namely:

1. Reporting and query tools
2. Application development tools
3. Executives information system (EIS) tools
4. Online analytical processing (OLAP) tools
5. Data mining tools (ibid)

A description of all these tools is outside the scope of our thesis so we are not discussing these terms here. Those interested can refer to the book (TC05).

1.7.2 Schema

A schema is the definition of an entire database. It defines the structure and the type of contents that each data element within the structure can contain. Schemas are often designed with visual modeling tools (Erwin, Rational Rose) that automatically create the SQL code necessary to define the table structures. In figure 1.3 there are five tables. Each table relates to the others through a foreign key.

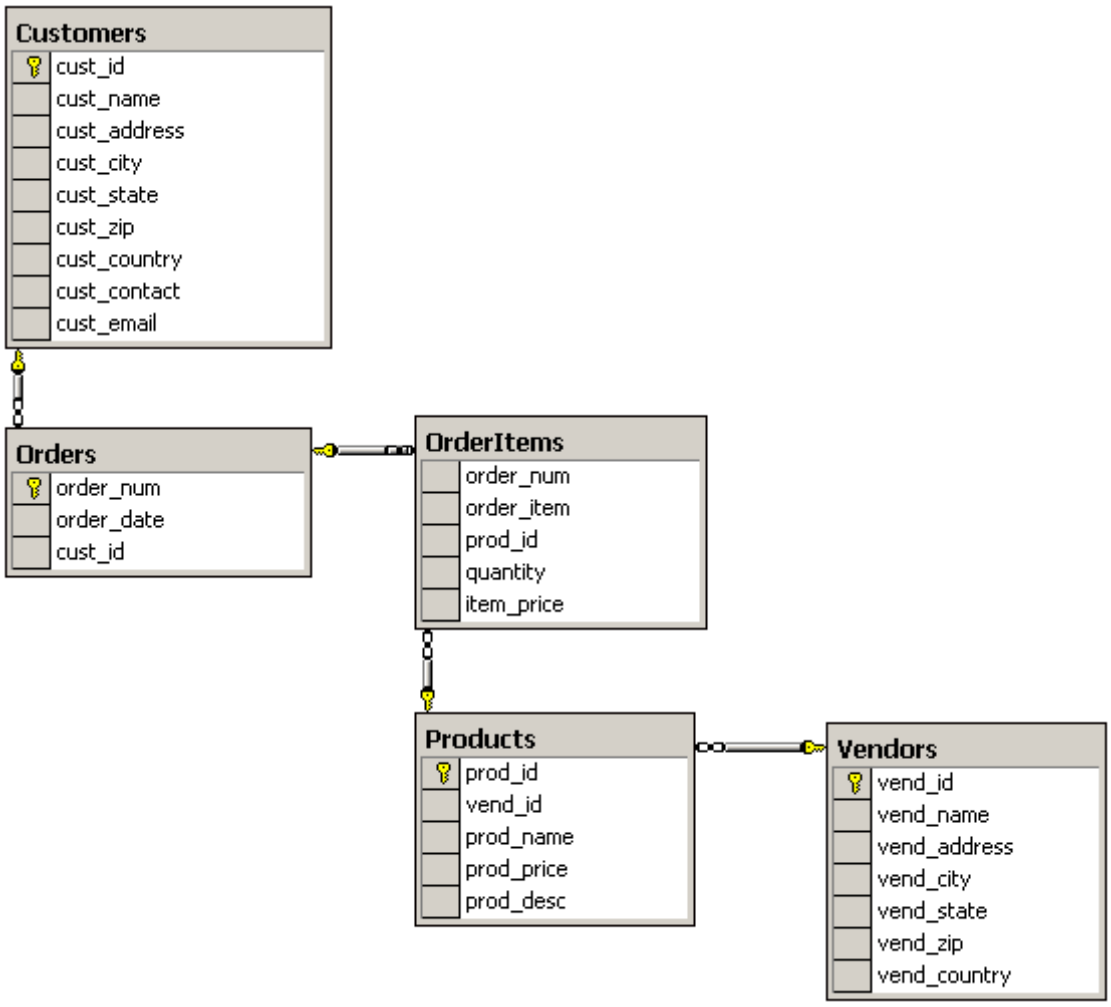

Fig 1.3: Example of a Schema
(http://www.sql-tutor.com/sql_tutor/students/lessons/lesson1.asp)

1.7.3 View

Views are widely used in decision support applications (RG03). An organization commonly has more than one analyst or groups of analysts which are typically concerned with different aspects of the business and it is convenient to define views that give each group insight into the business details that are useful for it. Once a view is defined queries can be written on new

view definitions that use it. Evaluating queries defined against views is very important for decision support applications. (ibid)

Originally, in database theory, a view is a read only virtual or logical table composed of the result set of a query (RG03). Unlike ordinary tables in a relational database, a view is not part of the physical schema; it is a dynamic, virtual table computed or collated from data in the database. Changing the data in a table alters the data shown in the view.

Views have the following advantages over tables:

1. They can subset the data contained in a table.

2. They can join and simplify multiple tables into a single virtual view.

3. Views can act as aggregated tables, where aggregated data (sum, average etc.) are calculated and presented as part of the data.

4. Views can hide the complexity of data, for example a view could appear as Sales2000 or Sales2001, transparently partitioning the actual underlying table.

5. Views do not incur any extra storage overhead.

6. Depending on the SQL engine used, views can provide extra security. (ibid)

Various database management systems have extended the views from read-only subsets of data (RG03). The Oracle database introduced the concept of materialized views, which are pre-executed, non-virtual views commonly used in data warehousing. They are a static snapshot of the data and may include data from remote sources. The accuracy of a materialized view depends on the frequency or trigger mechanisms behind its updates. The equivalent of this in Microsoft SQL Server, introduced in the 2000 version, is an indexed view. (ibid)

1.7.4 Views, OLAP and Warehousing

Views are very closely related to OLAP and data warehousing (RG03). OLAP queries are typically aggregate queries. Analysts want fast answers to these queries over very large data sets and it is natural to consider precomputing views. For e.g. the CUBE operator in SQL gives rise to several aggregate queries that are closely related to each other. The relationships that exist between the many aggregate queries that arise from a single CUBE operation can be exploited to develop very effective precomputation strategies. The idea here is to choose a subset of the aggregate queries for materialization in such a way that typical CUBE queries can be quickly answered by using the materialized views and doing some additional computation. The choice of views to materialize is influenced by how many queries they can potentially speed up and by the amount of space required to store the materialized view because it is needed to take into account cost of storage as well. (ibid)

A data warehouse is in fact a collection of replicated tables and periodically synchronized views (RG03). A warehouse is characterized by its size, the number of tables involved, and

the fact that most of the underlying tables are from external databases of OLTP systems. In reality the basic problem in warehouse maintenance is asynchronous maintenance of replicated tables and materialized views. (ibid)

Some people consider data warehouses as an extension to database views (EN04). Views however provide only a subset of the functionality and capabilities of the data warehouse. Views and data warehouse are similar in the sense that both have read only snapshots of data from OLTP systems and subject orientation. However data warehouses have quite a few differences as well with views including:

1. Data warehouses are multidimensional while views are relational.

2. Data warehouse can be indexed while views cannot.

3. Data warehouses provide large amount of data generally more than is contained in one database whereas views are extracts of a database (ibid).

1.7.5 Corporate Information factory (CIF)

According to (CI99) Corporate Information Factory (CIF) is a logical architecture whose purpose is to deliver business intelligence and business management capabilities by using data provided from business operations/operational information systems. The CIF has proven to be a stable and enduring technical architecture for any size enterprise desiring to build strategic and tactical decision support systems (DSSs). The CIF consists of producers of data and consumers of information (ibid). Figure 1.4 shows all the components found within the Corporate Information Factory architecture.

As CIF is beyond the scope of our thesis we'll not discuss its detailed architecture here. For those interested they can reach it at the following link:
http://www.dmreview.com/article_sub.cfm?articleId=1667

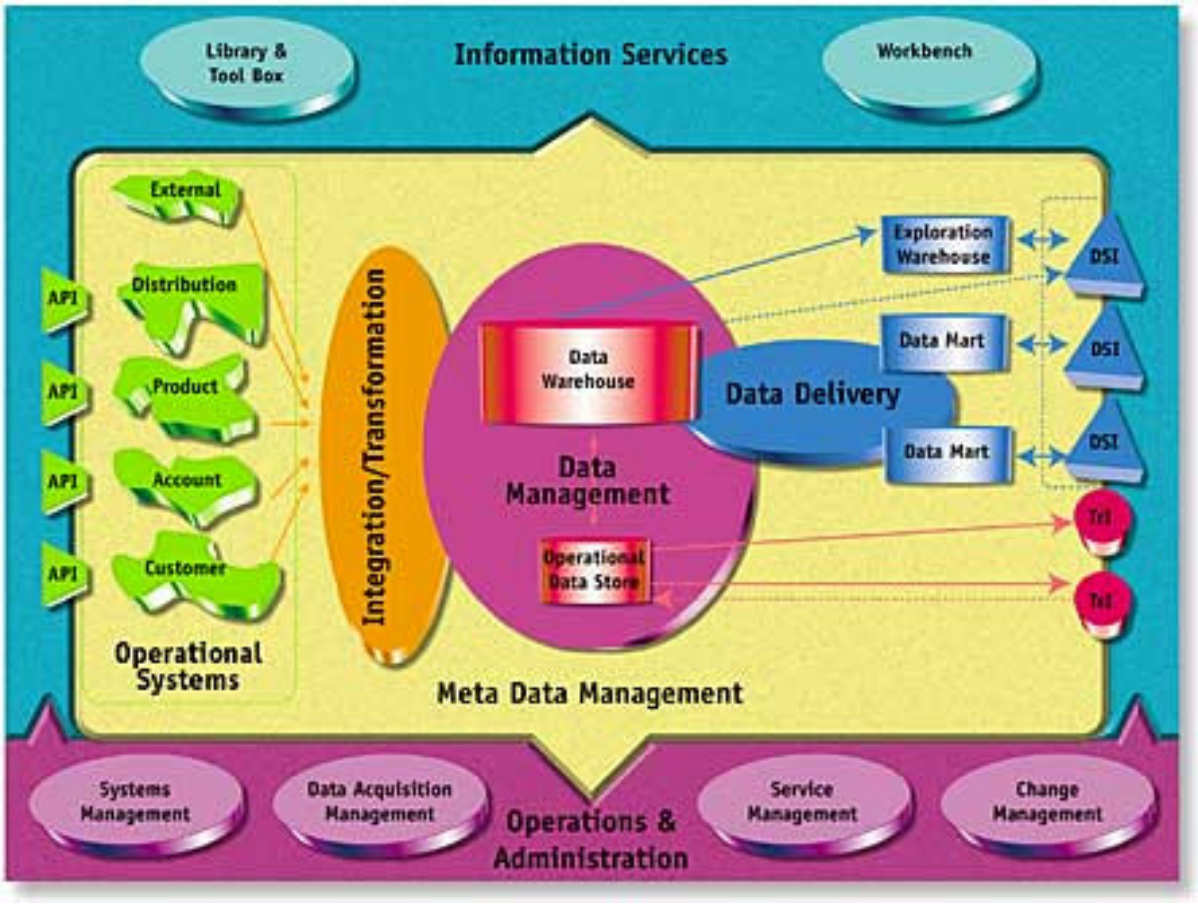

Fig 1.4: CIF Architecture
http://www.dmreview.com/article_sub.cfm?articleId=1667

1.8 Aim of study

The aim of our study is to investigate the maintenance strategies available for data warehouse performance management. Data warehousing projects are very expensive projects as compared to typical operational information systems, therefore it is highly desirable that they produce the expected results and help in making decision making an easy process.

Studies have shown that maintaining a data warehouse is more difficult and cumbersome than its development (LG05). Reorganizations, product introductions, new pricing schemes, new customers, changes in production systems, etc. are going to affect the warehouse. If the warehouse is going to stay 'current' (and being current will be a big selling point of the warehouse), changes to the warehouse have to be made fast. (ibid)

The question arises that why we have selected data warehouse maintenance? As data warehousing is an emerging area, a lot of effort and research is spent on its architecture, design and development phases. But not enough is done for the maintenance issues. We still can't find any book giving a reasonable description of all the issues involved in data warehouse maintenance and their solution. However there are articles and research papers

putting light on one or more of these maintenance issues. We'll try to gather this information from a lot of articles, research papers and books so that the issues related to data warehouse maintenance can be addressed.

According to (JT97) the focus of most writings about data warehousing is on planning, designing, and building them. But no one is considering what will happen after their development. Is that the end? Certainly not. The real work of taking output from the data warehouse starts from here. Suppose you have successfully implemented a data warehouse for an organization so it is worthwhile to setup solid procedures for managing this project. You will than be able to test their effectiveness on a smaller scale and improve on them before a much larger number of users have come to depend on the system. (ibid)

While you may have some solid systems administration procedures installed at your company, they may need modification or extension for the data warehouse environment (JT97). It is possible that they may need to be enhanced to serve the new applications and new users who will be accessing the system, as these people have less experience with technology, may have different expectations of service, or may be viewing more massive amounts of data than your other users. (ibid)

Where transaction rate, high availability and sub-second response time are the key concerns of operational systems management, other concerns may take their place in a data warehouse environment (JT97). The success of this application may be measured by factors such as ability to manage massive amounts of data, the speed with which users can execute their analysis, the timeliness and quality of the data, the ability to manage within a budget for hardware and software and user satisfaction. (ibid)

1.9 Research Question

Data warehousing is becoming an increasingly important technology for information integration and data analysis (BSE02). Given the dynamic nature of modern distributed environments, both source data and schema changes are likely to occur autonomously and even concurrently in different sources. In this scenario it is needed to have a comprehensive data warehouse management program to carry out all the routine work and take maximum output from the data warehouse. (ibid)

Keeping in view this phenomenon our research question is:

1. How can a data warehouse be successfully managed from performance perspective after its deployment?

The term successful here refers to the efficient and optimized performance by the data warehouse. It should provide the information that it was meant to deliver to make future planning and decision making an easy task.

After the deployment of data warehouse, users must now address related issues that augment the production environment i.e. establishing and maintaining ongoing communication and training, providing help support services, and managing technical infrastructure updates

concerning new releases of hardware, software and services. These services are often not discussed as part of the data warehouse project development life cycle. (RK00)

After the deployment of the first iteration or the complete data warehouse, the team managing it should be able to take care of the tasks and methods used to maintain a data warehouse. In this regard we will figure out the tasks and procedures used to maintain a data warehouse and will compare our finding with some empirical data obtained from a company already working with data warehouses.

1.10 Delimitations

Due to the reason that data warehouse stores enormous amount of data and data comes from a lot of sources data warehouses are highly maintainable systems. It also happens mostly that the systems feeding the data warehouse are on a different platform, use a different architecture or are located at different physical locations. All these problems require lot of resources and maintenance for the data warehouse to function properly and give the desired results. Keeping in view the high maintenance needs of data warehouse it is not possible to discuss all the maintenance issues here as it will be a long list. We will not discuss any issue related to design of a data warehouse. For e.g. data backup, data purging, account maintenance, role maintenance etc. Although these functions are also the responsibility of data warehouse maintenance team but are more related to the design of the data warehouse, therefore we are not going to discuss any such function.

Like ordinary DBMS's data warehouses could also be distributed or centralized. There is another form of data warehouses called data marts which are intended for a particular department or a function of the organization. We are not limiting our research to centralized or distributed data warehouses neither to the data marts. Our research addresses all the types of data warehouses.

1.11 Disposition of Thesis

We have divided our research work into six chapters as described below:

Chapter 1: This chapter will explain the background of the research study under subject. Starting with introduction to concept of data warehousing and data warehouse market it will provide the basic concepts associated with data warehouses. Additionally we have formulated the research question and aim of study in this chapter. Finally the disposition of the thesis is given.

Chapter 2: In this chapter we have discussed our research approach, research strategy, data collection strategies, data analysis methods and the methods to validate the research findings. This chapter will serve us as a guide throughout the remaining part of the thesis.

Chapter 3: In this chapter we have collected theory related to data warehouse maintenance and performance found in literature. We have collected theory from books, articles, research papers, white papers and the internet

Chapter 4: In this chapter we have presented our findings from the case study that we have done. We have presented how data warehouse maintenance is actually carried out in an organization.

Chapter 5: In this chapter we will compare the theoretical findings (Chapter 3) with the empirical findings (Chapter 4) and will discuss which ways are better than others and why?

Chapter 6: In this chapter we have presented our findings from the thesis. Keeping an eye on all the previous work here we will present our conclusions. Additionally we will present some areas where future research could be done in the area of data warehouse maintenance.

CHAPTER 2: METHODOLOGY

In this chapter we have discussed our research approach, research strategy, data collection strategies, data analysis methods and the methods to validate the research findings. This chapter will serve us as a guide throughout the remaining part of the thesis.

2.1 Research Approach

According to (RJ96) research can be classified into three perspectives:

1. The application of the research study
2. The objectives in undertaking the research
3. The type of information sought

These three classifications are not mutually exclusive i.e. a research study classified from the view point of 'application' can also be classified from the view point of 'objectives' and 'type of information sought'. For e.g. a research project may be classified as pure or applied research (from the perspective of application), as descriptive, co relational, explanatory or exploratory (from the perspective of objectives and as qualitative or quantitative (from the perspective of the type of information sought).

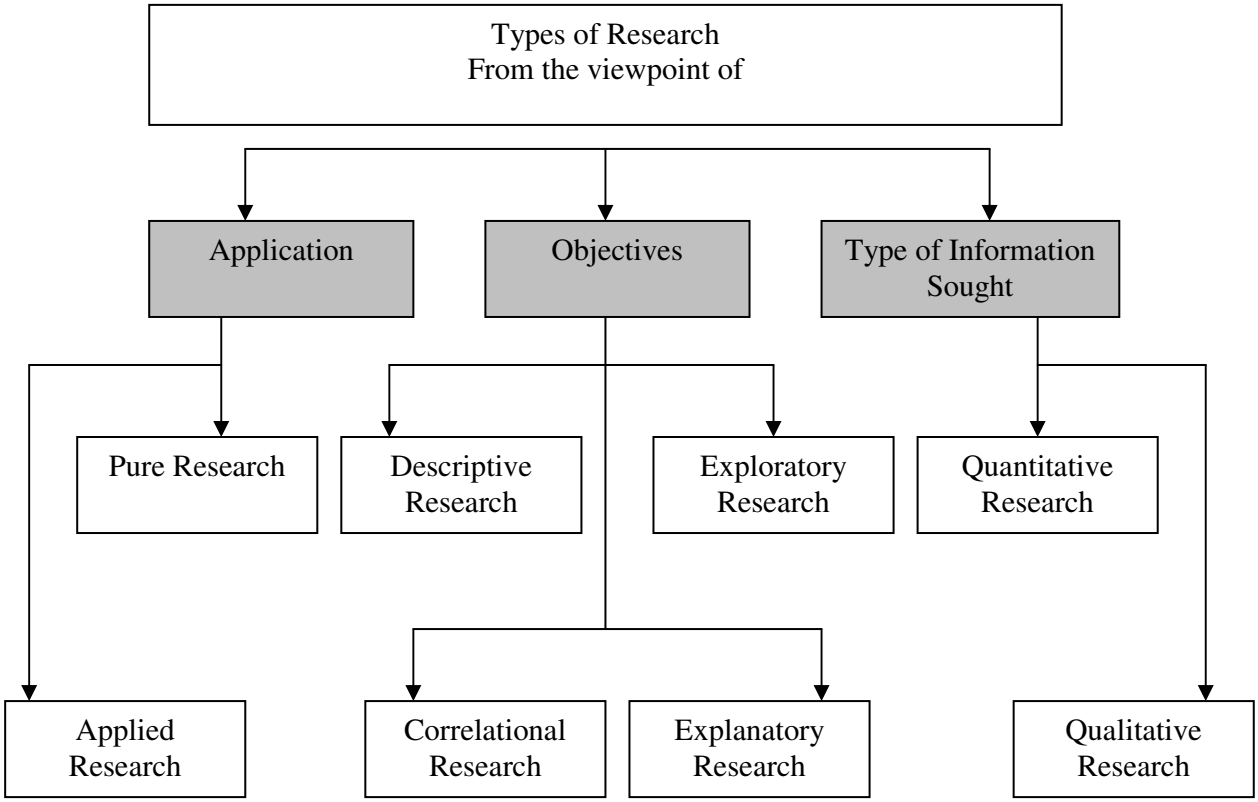

Fig 2.1: Types of Research (Reproduced from RJ96, Page # 8)

Keeping in view all these perspectives we found out that the type of information sought is the best option for us because we need to find out some information related to data warehouse maintenance.

Now keeping the type of information sought as the final perspective, we came to know that research can be classified as quantitative or qualitative. The quantitative or qualitative classification is dependent on three criteria:

1. The purpose of the study
2. How the variable are measured
3. How the information is analyzed

The study is classified as qualitative if: the purpose of the study is primarily to describe a situation, phenomenon, problem or event; the information is gathered through the use of variables measured on nominal or ordinal scales (qualitative measurement scales); and if analysis is done to establish the variation in the situation, phenomenon or problem without quantifying it. The description of an observed situation, the historical enumeration of events, an account of the different opinions people have about an issue, and a description of the living conditions of a community are examples of qualitative research.

On the other hand, if you want to quantify the variation in a phenomenon, situation, problem or issue, if information is gathered using predominantly quantitative variables, and if the analysis is gathered to ascertain the magnitude of the variation, the study is classified as a quantitative study. Example of quantities aspects of a research study are: How many people have a particular problem? How many people hold a particular attitude? (ibid)

In the social sciences, qualitative research is a wide term that describes research that focuses on how individuals and groups view and realize the world and create meaning out of their experiences (WI06). Qualitative research methods are sometimes used together with quantitative research methods to gain deeper understanding of the causes of social phenomena, or to help generate questions for further research. Unlike quantitative methods, qualitative research methods place little importance on developing statistically valid samples, or on searching for conclusive proof of hypotheses. (ibid)

Based on our research purpose and research question, the qualitative method has been chosen. The qualitative research seeks a better understanding of the complex situations. Further the qualitative method can describe a situation or problem easily and using this method the information gathering becomes an easy task as well.

We consider the qualitative method will be the best way due to the fact that we want to find out how we can deal with maintenance issues that need to be addressed after the development and deployment of the data warehouse. Using qualitative research approach we will gain a deep insight into the concepts and methods used for data warehouse maintenance.

Based on the research question, our research is based on deductive studies. The reason being that we'll take the ideas relating to data warehouse maintenance from the theory and will test them against some real world situation. We believe that we can gain a deeper and better understanding of under what circumstances data warehouses need maintenance. Are there any mechanisms within the data warehouse for its maintenance or some external resources are required to keep it up to date. We will also study the existing maintenance mechanisms and will try to figure out which techniques for maintenance are suitable and which are not.

2.2 Research Strategy

According to (RD02) case study is one of several ways of doing social science research. Other ways include experiments, surveys, histories, and the analysis of archival information. Each strategy has peculiar advantages and disadvantages, depending on three conditions:

1. The type of research question

2. The control a researcher has over actual behavioral patterns

3. The focus on contemporary as opposed to historical phenomenon

In general, case studies are an obvious choice when question of type 'how' or 'why' are raised. As a research strategy, the case study is used in many situations to contribute to our knowledge of individual, group, organizational, social, political, and related phenomena. In

short the case study method allows investigators to retain the holistic and meaningful characteristics of real life event such as individual life cycles, organizational and managerial processes, neighborhood change, international relations and the maturation of industries (ibid).

We will use the case study method to gather in-depth data relative to data warehousing, including the books written, research papers, and the articles written for the purpose of learning more about data warehouses itself and its maintenance which is a poorly understood situation up till now.

2.3 Data Collection

After deciding the most suited research strategy we had to decide how to collect data for our purpose. Because our research needs an in-depth data, and also considering the nature of qualitative research, we will use documentation as our main source of data collection. In addition the use of interviews will help us to gather valid and reliable data that are relevant to our research questions and objectives. We will try to arrange interviews with professionals working in companies which are already working with data warehouses. We will mainly use semi-structured or unstructured interviews in this regard because we don't have any formal interview requirements. All the correspondence will be through email or on phone as the companies are not situated in the city. We will ask them questions like how they are carrying out maintenance of their data warehouses, what they think about data warehouse maintenance and how it can be improved. To complement the interviews and gain as much information as possible, we will also use the documents from the company's website if available.

In addition we will gather data from the internet including articles, research papers etc. The data from internet will also help us a lot because we can find the latest developments in this field on the internet.

For the purpose of case study we have selected Telenor Pakistan Ltd. Telenor has been using the data warehouse for more than two years and according to their staff their data warehouse is working properly and giving them the desired results. The data warehouse has became a very popular information source among all the divisions of Telenor and nearly all the business users are using the data warehouse in some way or the other, which shows that the data warehouse is properly maintained and given due support to function properly.

We'll try to arrange interviews with data warehouse project manager, data warehouse staff and the data warehouse users at Telenor involved in the day to day operations and the maintenance of data warehouse. Interviews with these technical staff will enable us to rightly analyze the best maintenance policies in practice from a performance perspective. Keeping in view the location of the company and the cost of telephonic interviews we have decided to use some person as an intermediate interviewee. For this purpose we need a person who can communicate easily with us and the case site. We have identified the Data Warehouse Consultant at Teradata, Islamabad as the most suitable person. We'll send our interview questions to him and he'll send us the interview response afterwards.

2.4 Data Analysis

Qualitative modes of data analysis provide ways of discerning, examining, comparing and contrasting, and interpreting meaningful patterns or themes (MH94). Meaningfulness is determined by the particular goals and objectives of the project at hand: the same data can be analyzed and synthesized from multiple angles depending on the particular research or evaluation questions being addressed. The varieties of approaches including ethnography, narrative analysis, discourse analysis, and textual analysis correspond to different types of data, disciplinary traditions, objectives, and philosophical orientations. However, all share several common characteristics that distinguish them from quantitative analytic approaches. In quantitative analysis, numbers and what they stand for are the material of analysis. By contrast, qualitative analysis deals in words and is guided by fewer universal rules and standardized procedures than statistical analysis. (ibid)

They further say that data analysis consists of three concurrent flows of activities (MH94). These are reducing data, displaying data and drawing conclusion and verifying the conclusion (figure 2.2). Some times it is assumed that data reduction is not part of data analysis but its not true. Data reduction helps in making data sharp, sorted, focused, discarded, and organized in order to be able to draw and verify conclusion. (ibid)

Miles and Huberman conclude by saying ‘We have few agreed-on canons for qualitative data analysis, in the sense of shared ground rules for drawing conclusions and verifying their sturdiness (MH94).

As we have selected qualitative research so we attempt to gather data from several sources to aid in the validation of the data collection The analysis of qualitative data is not nearly as straightforward as quantitative data and requires a great deal more thought and effort to do well. Qualitative data analysis is not as easy as statistical data analysis.

We will use inductive reasoning for this purpose. First we will try to make some observations on data warehouse maintenance on a small scale and than we will draw inferences about data warehouse maintenance on a large scale.

We will also use comparisons to compare the theoretical findings with the empirical findings. This will help in understanding which things are common and which are not between the two and why. This will help us in validating our research findings as well.

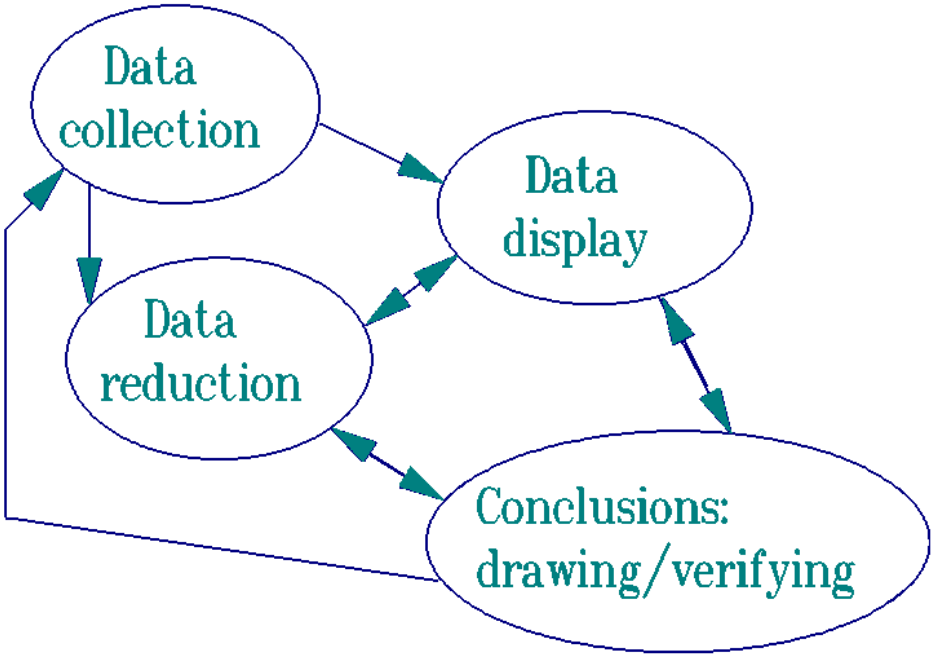

Fig 2.2: Components of Qualitative Data Analysis (MH 94, Page # 12)

2.5 Validity and Reliability

The concept of validity refers to quality and can be applied to any aspect of the research process (RJ96). With respect to measurement procedures it relates to whether a research instrument is measuring what it is used to measure. There are two approaches used to establish the validity of a research: the establishment of a logical link between the objectives of a study and the questions used in an instrument, and the use of statistical analysis to demonstrate this link. There are three types of validity which are: face and content, concurrent and predictive, and construct validity. (ibid)

The reliability of a research method refers to its ability to produce consistent measurements each time. When we perform any research under same or similar condition to the same or similar population and obtain similar results we say that the research is reliable. The more similar the results, the greater the reliability.

We will get validity and reliability in our research by consulting with the documentation as well as with the industry professionals working with data warehouses. We will then compare the results found in the documents to those what the professionals found while working with data warehouses. In this way we could say that we have achieved some form of validity and reliability in our research work.

To further validate the research findings we will collect data from reliable sources, such as data warehouse administrators who are in charge of data warehouse maintenance and interview questions were made based on literature review to ensure validity of research. If

required we may be able to also use triangulation where we will gather information from multiple resources and will use that information to support our own findings.

CHAPTER 3: LITERATURE REVIEW

In this chapter we have collected theory related to data warehouse maintenance and performance found in literature. The theory presented here has been collected from books, articles, research papers, white papers and the internet.

The information in an organization can be categorized in three ways according to need of managerial level i.e. strategic information used by top management e.g. information relating to long term planning, managerial information used by middle management e.g. sales analysis and operational information used by e.g. current stock levels etc.(EA99). EA99 suggested three different computer based systems to cater for the needs of each management level i.e. Decision support system (DSS), Management information system (MIS) and Data processing system (DPS) respectively. (ibid) As we know from general discussion in chapter one that data warehouse can cater for need of all three management levels in an efficient and effective manner.

All information system go through four stages namely system planning and selection, system analysis, system design and system implementation and operation (EA99, VGH01). As our topic fall in the last stage of system development life cycle so we will discuss accordingly. VGH01 listed seven steps for system implementation and operation, i.e. Coding, Testing, Installation, Documentation, Training, Support and Maintenance. These stages can further be grouped as, activities that lead to system going into operation –coding, testing and installation, activities that are necessary for successful system operation – documenting the system and training and supporting the users, activities that are ongoing and needed to keep the system working and up-to-date – maintenance (ibid).

Coding is an intensive activity starts when design team is final with the physical design specification to turn the system into working computer code. Though testing is parallel to coding but it also needs proper planning to achieve the designed objectives of the system. Installation of system is the replacing the existing system with the new one and includes conversion of existing data, software, documentation and work procedure according to new system. (VGH01) The outcomes of these activities are code, program documentation, test plan, test data, test results, user guides, user training plan, installation and conversion plan, hardware and software installation schedule, data conversion plan and site and facility remodeling plan. (ibid)

The process of documentation is carried through out life cycle of the system but on this stage all the information about the system are properly and fully documented both for users and maintainers of the system. In corporations there may be specialized staff to provide training to maintenance workers and users and in small organization some users can be trained and rest can learn from them. The outcomes of these activates are system documentation, user documentation, user training classes, user training tutorials, computer-based training aids, help desk, online help, bulletin boards and other support mechanisms. (VGH01)

The process of maintaining an information system is actually returning to the beginning of SDLC and repeating development steps. Four major activities occur within maintenance, obtaining maintenance requests, transforming request into changes, designing changes and implementing changes. The outcome of maintenance activity is new version of system along with update in all kind of documentation. In a system development life cycle this is the final

stage and again leads to beginning of system development so this is an critically important issue to deal with intensive care. As fundamental assumptions and properties of data, login or process module does not change in maintenance activity so wrongly identified maintenance will transform the system into a mess. (VGH01)

According to VGH01 maintenance is changes made to system to fix or enhance its functionality. They listed four types of maintenance namely corrective, adaptive, perfective and preventive. Corrective maintenance is changes made to a system to repair flaws in its design, coding or implementation. Adaptive maintenance is changes made to a system to evolve its functionality to changing business needs or technology. Perfective maintenance is changes made to a system to add new features or to improve performance and lastly preventive maintenance which is changes made to system to avoid possible future problems.(ibid)

Now onward we will discuss maintenance issues relating to data warehouse to improve the performance of data warehouse through efficient maintenance.

3.1 Data Warehouse Performance Management

The process of data warehouse performance management is similar to that of the design of a data warehouse (AH96). It is similar in that like the design and analysis phases, the procedures utilized are very different from the processes adopted in a conventional OLTP type system life cycle. In a conventional system life cycle there exists usually numerous levels of analysis and planning. In the data warehouse environment the system builders are seldom given this luxury and are required to assemble a data warehouse in a rapid manner with little time for performance analysis and capacity planning. This makes the data warehouse performance management process extremely difficult as the work loads very often cannot be predicted until finally the system is built for the first time and the data is in a production status. As a system goes into production for the first time only then may a system administrator discover there are performance problems. (ibid)

If there are too many performance problems in the running of the data warehouse the viability of the project becomes marginal or questionable (AH96). It is important to remember the success of a data warehouse can only be measured once the data is loaded and users are able to make business level decisions by extracting data from the data warehouse. (ibid)

The workload on a data warehouse hardly ever remains fixed (IB06). New users carry different kinds of demands, existing users change their focus and often the depth of their studies, the business cycle presents its own kinds of peaks and valleys, and in most cases the data warehouse expands as it stores data to cover longer periods of time. As the demand on a data warehouse changes, a lot of changes needed to be carried out to keep the performance graph in a positive direction. Some training courses needed to be introduced, some changed are needed for help desk, some indexes become obsolete and others need to be created, some aggregates are no longer referenced and others need to be evaluated, and the limits on parallel processing must be assessed and adjusted to fit the current demand. These and other tuning tasks should be carried out periodically to keep data warehouse performance smooth and constant. (ibid)

3.2 Data Warehouse Maintenance

Data warehousing is becoming an increasingly important technology for information integration and data analysis (BL02). Given the dynamic nature of modern distributed environments, both source data updates and schema changes are likely to occur autonomously and even concurrently in different data sources. (ibid)

The data warehouse after its deployment needs to be treated as a production system, complete with service level agreements (RM02). Technical support for the data warehouse should constantly monitor the performance and system capacity trends and take measures to get maximum output from the system. (ibid)

Six factors needed to be taken care of when dealing with ongoing data warehouse performance monitoring (RK00):

1. The data warehouse grows exponentially over time in terms of size and processing requirements.

2. Capacity management estimates, even based on the most precise calculations, are most likely to be still too conservative, requiring you to consider data warehouse expansion sooner than planned.

3. Data staging is a continuing challenge, especially if source systems are constantly in a state of change due to problems or changes. Some may be due for replacement under an ERP initiative, causing enormous changes to how data are sourced in the future.

4. Advances in technology in terms of network, hardware and software require more rapid release changes to be applied.

5. Ad hoc query access grows over time and must be carefully monitored as new and inexperienced users continue to run requests against base tables rather than summary or aggregate tables to produce totals.

6. An ongoing training program for business analysts, executives and decision support tool programmers keeps everyone informed as how to use the current version of the data warehouse or mart and find the information they need. (ibid)

3.3 Performance Tuning Mechanisms

While the implementation of a specific phase of the data warehouse may be completed, but the data warehouse program needs to be continued (RM02). Progress monitoring needed to be continued against the agreed-on success criteria. The data warehouse team must ensure that

the existing implementations remain on track and continue to address the needs of business. (ibid)

Performance issues in data warehousing are centralized around access performance for running queries and incremental loading of snapshot changes from the source systems (RK00, RH97, JT97). The following six concepts can be considered for a better performance:

1. Communication and Training
2. Help Desk & Problem Management
3. Network Management
4. Capacity Planning
5. Data Loading Performance
6. Query Management

3.3.1 Communication and Training

After the deployment of data warehouse one needs to address the issues of establishing and maintaining ongoing communication and training, providing help support services, and managing technical infrastructure updates concerning updated versions of hardware, software and services (RK00). Communication and training are two interrelated activities. A communication and training program is helpful in keeping the business community and IT community within the organization informed on the current and proposed future developments in the data warehousing environment. A communication process also offers the data warehouse team to measure progress and identify and resolve issues before they become a serious problem. The communications program provides the business components with increased capabilities and functions. Training expands the communication process by maintaining a level competence in both the business and IT community as to the tools and mechanisms of the data warehouses. (ibid)

As data warehousing starts to penetrate large numbers of enterprises, and it is currently doing so, it becomes critical that there be ample training resources to meet the demand generated by its widespread adoption and dissemination (RH97). Further organizations and individuals in need of instruction should have some level of assurance that the training they get is the right training and they can benefit from the warehouse after the training. (ibid)

Training of data warehouse users is significant and provides the desired output (JJ95). In most computing projects, management identifies the need for training, but does not always fund training. This is true for Arizona State University's data warehouse. With every new database there is a need for another training course, complete with reference materials. Every enhancement or change to the warehouse must be documented and communicated to warehouse users. At Arizona State University, the data administration department assumed responsibility for training and documentation of the data warehouse. While training users is

essential, it distracts from future warehouse development unless new resources are allocated. (ibid)

3.3.2 Help Desk and Problem Management

While training reduces the number of data warehouse questions, a support infrastructure is the key to handling other support needs. (JJ95)

In order to ensure success one needs to develop a support structure and plan an approach. When people are using the system, the questions will flow (VP96). If there are no questions than it is likely that no one is using the system. The question asked could be about the validity of the data itself, how to use the data, what calculations make sense and what levels of aggregation are valid, how to pick a report, how report measures were calculated, how to use the applications, how to change an application, and how to build an own application etc. (ibid)

User support is crucial immediately following the deployment in order to ensure that the business community gets hooked (RM02). For the first several weeks following user education, the support team should be working proactively with the users. It can't sit back and assume that no news from the community is good news. If there is nothing heard from the business users it means that no one is using the data warehouse. In such a case the support professionals should turn up to the business community so that the users of the data warehouse have easy access to support resources. If problems with the data or applications are uncovered, immediately try to rectify the problems. Again if the warehouse deliverables are not of high quality, the unanticipated support demands for data reconciliation and application rework can be devastating (ibid).

3.3.3 Network Management

If there is a heterogeneous group of platforms for the data warehouse implementation, network management is going to be one of the most demanding tasks (JT97). Not only are users coming constantly on-line, but users and equipment are invariably moving to new locations. The networking hardware is proliferating with LANs, WANs, hubs, routers, switches and multiplexers. Leaving behind all this is the next stage – users wanting to access internet based data sources along with the corporate data, requiring even greater bandwidth and network management resources. Managing this environment is one big challenge, capacity planning for the future is another. If the data warehouse team is not quite good in networking technology than there should be at least one person in the organization who understands technology. (ibid)

3.3.4 Capacity Planning

Capacity planning refers to determining the required future configuration of hardware and software for a network, datacenter or web site (AN06). There are numerous capacity planning tools in the market used to monitor and analyze the performance of the current hardware and

software. However, capacity planning also requires insightful forecasting: what if traffic triples overnight; what if a company merger occurs, etc. As a result of all the analyses and forecasts, systems can be upgraded to allow for the projected traffic or be enhanced so that they can be ready for a quick changeover when required. (ibid)

According to (WI06) capacity planning enables the determination of sufficient resources so that user satisfaction can be maximized through timely, efficient and accurate responses. (ibid)

Capacity planning is important when starting a new organization, extending the operations of an existing business, considering additions or modifications to product lines, and introducing new techniques, equipment and materials (WI06). In business, capacity is the maximum rate of output for a process. This means that capacity is the work that the system is capable of doing in a given period of time. The goal of capacity planning is to meet current and future demand with a minimal amount of waste. (ibid)

3.3.5 Data Loading Performance

To load a data warehouse, regular loading or propagation of data from operational systems is needed (JT96). A schedule for summarizing, loading, and making the information available to the user community needs to be developed and it should be presented to the user community. For e.g. daily summary data may be available by 7 AM the next morning and weekly summary data by 8 AM Monday morning. The users should also know if and when the data was loaded. (ibid)

It is also necessary to develop procedures for managing the results of a bad load (JT97). For e.g. there should be some defined procedures if the data loaded is corrupted due to some operational mistakes and the problem with data is discovered after some time. In that case the data needs to be reloaded. One needs to consider what are the impacts of data reloading, how it will be reloaded, what will happen if data is reloaded during peak hours (any effects on operational systems)? User notification regarding corrupted data should be part of the data loading procedure. (ibid)

Factors affecting data loading performance include (RK00):

1. The decreasing batch window (time available for an intensive batch processing operation such as a disk backup) to load data from more and more sources, coupled with increased usage of the data warehouse.

2. The frequency and size of these loads.

3. The changing natures of these loads as source systems change or are replaced.

4. The increasing demand for more metadata regarding the data to be loaded.

5. If load performance remains an issue, consider maintaining a synchronous replica of the full database to source data and to the warehouse.

All these problems mean that the data warehouse team must plan for and examine performance on an ongoing basis. Creating a performance management system that logs all relevant data or selecting a product that provides this functionality will arm you with a critical tool in the fight to keep on top of the growing data management problem. (ibid)

How data is loaded is more important than what data is loaded (RK00). Do not consider loading data based on real time source to target mapping. Data quality is hard to monitor with this method, and critical performance problems do occur that affect the source systems and the data warehouse. Data can be loaded efficiently by dropping all indexes and rebuilding them after the load completes. Turning off row level locking if possible can help in achieving better performance. DBMS's bulk loading facilities can be used to load data as well if available. Try to turn off DBMS journaling (tracking data changes). Data load personnel should ensure that the data staging tables map directly to the data warehouse data mart tables. These personnel should also try to calculate and prepare derived data for loading into data warehouse. (ibid)

Cleaning and transforming the data in the data staging environment prior to loading can be a great help in improved data loading performance (RK00). Data warehouse maintenance staff can plan the division or segmenting of the big dimensions of warehouse, such as customer and product, by subtype, defining each to separate physical tables prior to mapping the data across the disk drives. The parallel loading features of the processors can be used to further enhance the data loading performance. Scaling up the bandwidth of the network to accommodate more traffic for e.g. assuming 15 GB of change data per day at

1. 1 MB per second will take 41 hours to load

2. 10 MB per second will take 25 minutes to load

3. 100 MB per second will take 2.5 minutes to load (ibid)

3.3.6 Query Management

In a data warehousing environment users queries need to be very efficiently and carefully written as some tables of the data warehouse are very huge and queries posted against these tables could days or weeks to complete.

To have an efficient query management system most of the predefined and ad hoc queries should access summary data instead of detailed data (RK00). One needs to employ query navigators to redirect base table queries to the aggregate and summary table level and examine which tables and columns were accessed and the number of rows retrieved. Check response time for these queries and any effects these have for e.g. paging or locking. Break down the predefined queries into smaller queries for processing. Consider doing most of your resource intensive processing away from the current level of detail and try to run large queries during off-peak hours. This will give more processor time to smaller queries and will aid in getting quick results. Try to push query processing up from the client to the application server level. As part of end user workstation design, consider the employment of thin clients, forcing query processing and scheduling up to the server level. In a data warehouse data should be distributed in a hierarchical manner where the most common information has the least amount of distribution and the least common information has the highest level distribution. Minimize

the amount of cross network data retrieval and combination of data from different locations. Allow information access users to follow a hierarchical path when searching for data (ibid).

3.4 Communication Process

The communication program is intended to improve the understanding, attitude, and commitment of the various players who are involved in the data warehousing process (RK00). Communication processes for data warehouse project and program support deals with issues like corresponding with internal and external groups of the data warehouse project, scheduling deliverable reviews and conducting meetings to get status of the project and the issues arising in the project, helping quality assurance representatives to manage quality and notifying the data warehouse group and others in the organization when project milestones are reached. (ibid)

The scope of the selected communication program identifies the people who should be contacted, the main messages to pass, and the type of communication and its frequency (RK00). This can be achieved by giving a detailed and scheduled program of education and training for developing and supporting vision clarity for the data warehousing environment. Educating the knowledge workers about the purpose and benefits of a real and complete data warehouse and feedback and suggestions from data warehouse users can also help in improving data warehouse performance. By providing a place or group to contact outside the helpdesk to address concerns or to act as a contact point for marketing data warehousing services to new business units and involving other interested parties in data warehouse planning, analysis and design sessions, users understanding of data warehouses can be greatly increased. (ibid)

3.4.1 Communication Process Implementation

The activities listed below describe the definition and implementation of a communication program for data warehousing (RK00). A communication plan is an absolute necessity for medium or large scale data warehousing projects. These projects are delivered in multiple iterations; therefore an ongoing communication process is critical for success. Data mart projects that cross lines of business or where the focus is on corporate or executive information analysis should also consider implementing a communication program as a required infrastructure process. (ibid)

To establish an efficient and result oriented communication program one needs to establish a communication team and develop a financial plan for budgeting (RK00). The members of the team should establish a list of recommended technology components and standards that should be used. The team further reviews and approves the proposed approach, functional architecture and technology components and standards. (ibid)

The next responsibility of the team is to design the data warehouse communication program structure (RK00). The team analyzes current communication channels and mechanisms, defines vendor to internal IT involvement and responsibility in this process, determines which communication components to buy and which to develop in-house, conduct the vendor selection evaluation process, defines the timing and frequency of the proposed communication process, determines the required supporting organizational structure. After going through all these tasks the team finally compiles a proposed implementation and support plan. (ibid)

The next step is the implementation of the communication program (RK00). This includes confirming completeness of all designed components, processes, human resources and facilities and preparing implementation of the communication program in support of data warehousing efforts, selecting and train delivery or user team members, setting up facilities, reporting, and feedback procedures and the supporting desktop environment for the communication program. (ibid)

The communication process produces a number of documents, procedures, facilities, and systems capabilities to deliver the data warehouse communication program (RK00). These include a communication program charter or mandate containing program goals and objectives, budget, organizational structure for communication team, standards to be followed and the program procedures for a better communication channel between the users of the data warehouse. Documents related to communication program facility locations and staffing are also produced along with data warehouse program office requirements and data warehouse marketing plan. (ibid)

The effectiveness of the communication program will be in direct proportion to the amount of time, effort and visibility placed on this essential infrastructure process (RK00). Good and frequent communication between development team members, users and project managers and sponsors will assist through the many peaks and valleys to come as you prepare to deliver and maintain the data warehouse. (ibid)

3.5 Training Program

A continuous education and training program is always required for the data warehouse (RM02). The curriculum should include formal refresher and advanced courses, as well as repeated introductory courses. More informal education should be offered to the developers and power users to encourage the interchange of ideas. (ibid)

According to (VP96) users are undoubtedly much more comfortable receiving reports, even if they have to go to five different reports for their information, than learning a whole new information system. Learning to think in a multidimensional, heuristic mode is a skill set that is learned and improved as it is used. It is quite different from a flat file mindset. (ibid)

Most users have no idea and cannot visualize the breadth of functionality these sophisticated decision support tools offer. Front end data access tools are not simple to use. In fact they can be relatively difficult. (VP96)

The data warehouse is not an operational system (VP96). In many cases users don't have to use it, they can choose to use it or not to do their jobs. If the data warehouse has to give some output it must be used and it will be used only when the users know how to use it (ibid).

3.5.1 What Should Be Taught

The most successful data warehouse implementations create ongoing, well-designed and implemented training programs for their users (RK00). Training needs to be focused on data warehouse concepts and terminologies, introduction to the organizational data, where is that located in the warehouse and how it is related to the reports or systems user already is using, the mechanics of using the tool. It is important for people to understand basic navigation within the tool. The type of analysis that can be performed and how use the tool against the data. What starter set of reports has been developed, how to use them and how they are organized. (ibid)

Within data warehousing there are a number of major topics and an even larger number of secondary topics (RH97). In addition there are a series of related subject matters that can be covered from the perspective of data warehousing such as object-oriented technologies, client/server technologies and the internet. (ibid)

The users of the data warehouse could be divided into five categories namely (RH97):

1. Data warehouse staff
2. Data warehouse managers
3. Business users
4. Company executives
5. General audience (Independent of their job description, with no prior knowledge of data warehousing)

Table 3.1 shows which categories need education in what types of data warehousing fields? It should be noted that in each category there may be some individuals who already have some level of data warehousing knowledge and related concepts and thus may start at a different point in the curriculum. (ibid)

	Awareness	Plan	Design	Implement	Manage	Use	Maintain
DWH staff	X	X	X	X			X
Business users	X	X				X	
DWH managers	X	X	X	X	X	X	X
Executives	X				X	X	
General	X						

Table 3.1: Who should be trained in what areas (RH97)

The optimal learning environment may be a customized class created either internally or by a vendor, using a subset of the company's own data (VP96). The advantages this approach has are listed below:

1. Uses data that users know and can identify with (company's own data).

2. Provides custom training material and manuals.

3. Provides formal exercises and materials to assist in training new personnel. (ibid)

It should be noted that one day long training at the vendor site is not enough for an average user of the data warehouse (VP96). The tools used to extract information from a data warehouse are extremely sophisticated. Often users get confused by the overload of information or forget the information before having a chance to use it. It is imperative that procedures and programs be implemented that can provide continuous help and assistance on the data warehouse and the front end tool. (ibid)

Training is an essential component of most industrial pursuits (RH97). It is the process whereby an individual acquires the necessary skills to perform a specific job or carry out specific tasks (ibid).

In his landmark book on data warehousing, the data warehouse toolkit; Ralph Kimball insists to follow the following points for an effective education program (RM02):

1. Understand your target audience. Don't overwhelm.

2. Don't train the business community early prior to the availability of data and analytic applications.

3. Postpone the education if the data warehouse is not ready to be released.

4. Gain the sponsor's commitment to 'no education, no access' policy (ibid).

3.5.2 The Training Program Implementation

Without proper training the intended users would not be able to take full advantage of the capabilities of the data warehouse (RK00). A training program should be started for medium or large scale data warehousing projects. These types of projects are delivered in multiple iterations therefore the program is essential for the success of the data warehousing project. Data mart projects that overlap the business or where the focus is on corporate or executive information should also consider implementing a training program as a required infrastructure process. (ibid)

To establish an efficient and result oriented training program one needs to establish a training team and develop a financial plan for budgeting (RK00). The members of the team should establish a list of recommended technology components and standards that should be taught.

The team further reviews and approves the proposed approach, functional architecture and technology components and standards. (ibid)

The next responsibility of the team is to design the data warehouse training program structure (RK00). The team analyzes current training channels and mechanisms, defines vendor to internal IT involvement and responsibility in this process, determines which training components to buy and which to develop in-house, conduct the vendor selection evaluation process, defines the timing and frequency of the proposed training process, determines the required supporting organizational structure and conduct training developer team training, designs and tests the required organizational structure, roles and responsibilities, deliverables, and procedures for the training program and required interfaces to the data warehouse project development life cycle, help desk function and program office. After going through all these tasks the team finally compiles a proposed implementation and support plan. (ibid)

The next step is the implementation of the data warehouse training program (RK00). This includes confirming completeness of all designed components, processes, human resources and facilities and preparing implementation of the training program in support of data warehousing efforts, selecting and training delivery of user team members, setting up facilities, reporting, and feedback procedures and the supporting desktop environment for the training program. The training program is reviewed from time to time for betterment. (ibid)

As a result of the efforts of the training team members, a lot of documents, procedures, facilities, and systems capabilities are developed to deliver the data warehouse training program. These include a training program charter or mandate containing program goals and objectives, budget, organizational structure for training team, standards to be followed and the program procedures to deliver training. Additionally documents pertaining to training program facility locations and staffing are also produced. (ibid)

Communication and training are two essential change management processes that enable both the technology and user communities to more rapidly accept and deploy this new business technology (RK00). By focusing on these two critical human resource development functions the company can go a long way towards assuring the success of their data warehousing initiative (ibid).

3.6 Role of Help Desk

A help desk acts as an extension to the user and as an assistant to the data warehouse support group (RK00). For the end user the help desk addresses a number of issues that include:

1. Security and sign on (from the client network, or remote)

2. Access to detail or aggregated data stored in a warehouse, operational data store or data mart

3. Data staging or data quality management problems

4. Ad hoc query processing

5. Predefined and scheduled reports or automated routine processing

6. System response time

For the data warehouse IT support and monitoring team, the help desk quickly points out challenges and issues including installation of new functionality, certification of new hardware or software, facility or network capacity thresholds, and scheduled job performance and environment backups etc. (ibid)

The help desk based on its authority and expertise of its staff members, resolves, redirects or escalates a problem event that has occurred (RK00). Help desk services are provided in a number of ways including phone, intranet or internet based help, in person or through the automated help facilities available in the data warehouse. (ibid)

The first and most obvious method of support is to create and expand the help desk (VP96). This gives the users one place to call when help is needed. The people at the help desk needed to be able enough to solve a technical problem themselves, but also need to have an understanding of the business, the data that is in the warehouse and how it can/should be used. This complete skill set may not reside with one single person, instead a support team may be needed that can control the situation. (ibid)

Another role of the data warehouse support and maintenance group is the problem resolution when some problem is encountered in the data warehouse (RK00). A help desk acts a coordinating body for not only collecting and logging problems with the data warehouse environment but also determining where future requirements may lay. In some organizations the data warehouse help service is seen as an extension to an existing OLTP help desk based program while in others a separate organization is struck to deal with the distinctive nature of this environment. Within this context, the problem management process is the vehicle for recording and resolving issues as they arise, again pointing the way towards future improvements or the need for new functionality in the data warehouse. Problems can be maintenance, enhancement based, or they can point the way towards new development. The problem management process much like the project change management process, acts as a vehicle for initiating and determining the nature of work for our maintenance, enhancement, or new development data warehousing projects. (ibid)

3.6.1 Help Desk Services

Help desk services are provided from a central call center or they are distributed to the various business locations (RK00). If more hands on, or personal support is required especially during the early months of a major data warehouse project, this support is usually phased out of the data warehouse group or centralized over time. Other options include ‘train the trainer’ approach where local representatives are given more extensive product and application based training than the average end user. Other options for providing help desk services include identifying local ‘power users’, those more in tune with the technology or using a much broader band of the available services. These people can be tapped in a backup or support role to the help desk. To keep them interested, these talented staff can be offered ‘first look’ and more proactive involvement in future software selections or service enhancements. (ibid)

3.6.2 Developing a Help Desk

The stages of growth for help desk are much like that of the data warehouse (RK00). Initially these services should be provided by the data warehouse implementation and monitoring team for a period up to three to six week months, or until the new data warehouse increment is established and relatively problem free. For enhancements or maintenance efforts, this level of support can be reduced to two months and one month respectively. After establishing a help desk function as part of the role out process, responsibility for this service can be passed to IT. With the migration of this service away from the core group of expertise, training and support of help desk personnel must become part of the overall data warehouse training program. Due to the more volatile nature of this technology, support of help desk personnel is critical to their ability to provide good service. Unlike the OLTP environment, help desk personnel are faced with a large degree of the unknown, based on the ad hoc and unpredictable nature of the query environment they support. Additional challenges these staff members will face include:

1. Understanding the functionality of changing DSS, EIS, and modeling software

2. The potential impact of rapid growth causing changes to processors, network bandwidth, and disk storage capacity

3. The critical importance of error free data staging and what can happen if problems related to incremental data refresh runs are not quickly addressed (ibid)

The process of help desk development starts by establishing a help desk team that develops a budget plan, establishes IT and business sponsorship for the process by conducting awareness sessions, develops an initial position statement or approach for steering committee and/or sponsorship review (RK00). The team further defines a functional framework for the components and procedures to be developed and establishes a list of recommended technology components and standards. (ibid)

The next step is design of help desk which includes analysis of current help facilities and mechanisms, definition of IT involvement and responsibility in this process, determining which help desk components to buy and which to develop in-house, conducting the vendor selection evaluation process (RK00).

The help desk team is than trained to perform their role efficiently. After going through all the process the help desk support team gets the approval for the program (RK00).

Help desk and support process results in a number of documents, procedures, facilities, and systems capabilities to deliver the data warehouse help desk program (RK00). These include the help desk program charter or mandate containing program goals and objectives, budget, organizational structure for help desk team, standards to be followed and the program procedures to perform help desk functions. Additionally documents pertaining to training program facility locations and staffing are also produced. The help desk team also develops training procedures for data warehouse tools for e.g. education of the features and functions of a new decision support system. (ibid)

3.7 The Problem Management Process

The problem management process is the glue that holds the help desk together (RK00). This process specifies how to collect, document, answer and/or escalate calls, requests, and queries related to issues with the data warehousing environment. Problem documentation can be completed either by the help desk representative and/or in conjunction with a form completed by the end user or IT support person requesting a service or action. All inquiries, no matter how trivial should be logged, especially during the start of a new data warehouse or mart. These bits of information can form clues to taking proactive action to bigger problems before they emerge. Having a production ready data warehouse means support must be expedited in an efficient, responsive, and businesslike manner. At stake is the ability of the business to stay competitive if the business information the warehouse contains is not current, accurate, timely and available when needed. This thought must be kept in mind by all help desk personnel as they strive to answer those nagging questions: why queries don't run the way or as fast as they expect. (ibid)

Ongoing checkpoint reviews are a key tool to assess and identify opportunities for improvement with prior deliverables (RM02). Data warehouses most often fall off track when they lose their focus on serving the information needs of the business users. (ibid)

3.7.1 Problem Management Process Development

The problem management process is developed entirely from scratch or it is implemented using an available IT procedure if available. Problem management routines are defined using vendor services in conjunction with their data warehouse monitoring software (RK00). It is always best to refine what is available, rather than having to build a new procedure from scratch, especially if this is the first problem management process created by the business. A good problem management process, integrated within the overall functionality of the helpdesk, is critical for ongoing success in data warehousing. Unresponsive or inefficient procedures inevitably lead to data warehouse performance, usability, and availability issues in near future. Major challenges faced by some early installations don't often come to the attention of management in time. Good documentation, much like responsive, timely support is critical to success of data warehouse. (ibid)

The following procedure highlights a problem management process (RK00):

1. When an issue or problem is identified and raised by an end user, help desk support person or IT/data warehouse support or monitoring team member concerning the performance of data warehouse records that event.

2. If the problem cannot be resolved by immediate action, then an initial entry in a problem record is created. The problem is given the status of 'open' and is assigned a follow on action owner. Whoever raised the problem is designated as problem reporter.

3. As action is taken it is documented as an entry against the problem. Every individual taking an action on a problem makes an entry in the problem log and decides who should be notified next.

4. Numerous actions may be required before a problem is resolved. Unresolved problems are escalated through the data warehouse project's escalation mechanism on a regular basis.

5. When a problem is resolved, the follow on action owner is specified as the original reporter. The original help desk reporter should be the only individual who can close a problem unless it was reassigned during problem resolution to another person.

6. When the reporter receives notification of problem resolution, he or she determines if it can be closed. If the person or group determines that the problem cannot be closed, then it is returned to the help desk for reassignment. A decision to reopen the problem is then documented as an action entry in the problem management log.

All problems no matter how trivial should be documented, especially if related to new development or enhancement data warehouse project. In the event that a problem remains open and requires authority beyond that of the help desk, than the problem should be escalated immediately to the data warehouse program or project leader. (ibid)

Out of this process the following documents, procedures, facilities, and systems capabilities are developed to deliver the data warehouse problem management process (RK00):

1. Problem management log

2. Problem report form

3. Problem action form

4. Problem resolution form (ibid)

3.8 Network Management

If the data warehouse is implemented using a heterogeneous group of platforms, network management will be one of the most difficult and tough tasks (RH97). New users will continuously come online and users along with equipment are invariably moving to new locations as well. The networking hardware is always increasing in numbers with LANs, WANs, hubs, switches, routers and multiplexers. Users always want to access internet based data sources along with the corporate data, requiring even more bandwidth and network

management resources. There should be some knowledgeable person in the organization who could handle these issues. (ibid)

Some integrated tools are required to assist data warehouse team or the network management team in monitoring the network performance (JT97). Fortunately there are several such tools now available, and enhancements are being made with each new release to improve their functionality. Because simple network management protocol (SNMP) is the common standard in this area; most vendors concentrate on SNMP based distributed network management features. (ibid)

Listed below are some of the features to look for in these tools (JT97):

1. Distributed console support: The ability to access the tool from several different consoles.

2. Heterogeneous relational database support: The ability to work with many different database management systems, in a mixed environment.

3. Scalability: The tool should be able to work with an increasing number of servers and platforms without any loss in capability.

4. Interfaces for alerts (error messages requiring action): Should be able to support a variety of operating systems.

5. Security features: Should have features such as user ID authentication and auditing of attempted invalid accesses.

6. Tracking of network utilization in real time and in summary reports.

7. Measurement of network response time.

8. Graphical display of key measurements.

9. Troubleshooting assistance: The ability to provide diagnostics when problems occur.

10. Identification of under-utilized and over-utilized resources to permit load balancing and tuning.

11. Optimization tool to improve overall performance. (ibid)

Even with these sophisticated tools, many warehousing systems find that their staff lacks the expertise to use them to full extent (JT97). This is a complex area and if the staff members do not utilize the tools and associated methods frequently enough, they do not build up enough experience to become experts. So some companies find it cost effective to use outside service providers who specialize in this area to help them identify their best options and sometimes, implement the recommendations. Such firms can supply network planning, design, implementation, management and monitoring services, either remotely or on site. (ibid)

By proactively monitoring the network and resolving any bottleneck issues, you can analyze performance and put a plan in place to be able to support the critical applications (RH97).

There is no doubt that the new tools and services can provide much better insight into network performance than their predecessors. But it is still the job of network administrator to maintain network performance and meet the company's objectives (ibid).

In developing network strategy consider the limitations of the current environment, as if it were not structured for the type of use you are now intending to place on it (RK00). Using mechanisms like ODBC to move small amounts of data in batch for periodic updates is one thing while moving large amounts of data for both loading and querying on an ongoing basis requires careful planning. The available network capacity has a huge impact on the data warehouse data management plan (data topology). High volume data access and data loading over slow pipes results in unacceptable performance. It would be an ideal situation if the database management layer is defined before identifying the required network infrastructure. If a network infrastructure already exists, however, capacity planning must be completed before the data topology design for the data warehouse is finalized. Consider the protocols available with the extraction and transformation software, database and information access software and check whether they are compatible or extendable. With the closing of the gap between technologies such as Asynchronous Transfer Mode (ATM) and Fiber Distributed Data Interface (FDDI), the distinction between the two greatly decreases. This increased compatibility and integration eases the planning of parallel based architectures with respect to data distribution and access, ranging over today's parallel server architecture, local and wide area networks. (ibid)

These parallel architectures are transparent to the end user (RK00). The knowledge worker accessing the data warehouse decision support system views the virtual parallel system through his or her personal computer system not caring if the access is local or distributed. His or her main consideration is and will always remain based on performance, how much and how fast our architecture can provide the answers to a growing list of more complex and extensive business questions. (ibid)

In developing a strategy for data staging and data replication use the DBMS's own data movement facilities to move data asynchronously between different database systems (RK00). Try to use application program controlled asynchronous data movement when data from multiple sources are required to insert aggregate data for better performance. Data movement to multidimensional databases should rely on the data extract and movement facilities supplied by the multidimensional software vendor. Along the same lines, unstructured data movement should be managed by the products vendor. (ibid)

What role will data replication play in data warehouse database management (RK00)? Data replication should be considered only when either

1. Data mirroring is required

2. Data distribution is required

3. Data movement of a subset of a data warehouse is required to update one or more dimensions of one or more marts that require this information.

4. Data replication should never be considered as a replacement for the data staging process by moving data directly from source systems to either an operational data store, data warehouse, and/or data mart. (ibid)

3.9 Software and Hardware Issues

According to (JJ95) client/server technology is less reliable, secure, and timely than its mainframe predecessor. Data access tools are just beginning to mature. Networks add new layers of complexity, and monitoring performance and tuning of servers is imperfect. The results are gaps in available technology and software, leaving users frustrated and their needs unmet. To overcome these problems warehouses needed to get their software and hardware updated in a timely manner to avoid any shortcomings in performance. (ibid)

Updates to the data warehouse are inevitable; so too will be changes to package software, hardware servers, and the supporting network infrastructure (RK00). Three strategies are available to make changes to this technical layer depending upon the scope, timeframe and criticality of the data warehouse environment. These strategies include:

1. Installing new software releases, patches, hardware components or upgrades, and network connections (logical and physical) directly in the production environment.

2. Installing new software versions, hardware upgrades, and network improvement tasks in a temporary test environment and migrates or reconnects to production once certification testing has concluded.

3. Installing technical infrastructure changes into a permanent test or maintenance environment and migrate the production environment once certification testing has concluded. (ibid)

3.9.1 Implementation Strategies

There are three implementation strategies to carry out these changes which are:

1. Peer to peer

2. Master to slave

3. Hybrid

Trade Transition Approach (Peer to Peer)

In this approach each environment acts as a distinct entity (RK00). This entails a large first data warehouse iteration, to be followed by smaller migrations of critical components over time because we are moving between distinct environment, this approach is the most time

consuming and costly to adopt as a standard practice on an ongoing basis. This approach, however, safely isolates certification testing from the production environment. (ibid)

One Master Environment (Master to Slave)

The master slave approach utilizes one environment for integrating changes and upgrades (RK00). The benefit of this approach is that new functionality and components are integrated and tested in one place as various new features and capabilities are introduced. After the certification testing is complete the master releases control to the slave. (ibid)

A Hybrid Approach

In some cases, a combination, or hybrid, approach is adopted based on the type and nature of the certification requirement (RK00). For example, a traditional approach may be adequate for migrations of software releases but not so for upgrades to the network topology. (ibid)

These three aspects of software and hardware installation are further complicated by the differing requirements between software, hardware, and networking components, since the features and functions of each product should align (e.g. a DBMS with strong parallelism features should be deployed on data servers with same properties). For implementation of new software, hardware, and network devices our considerations should include (RK00):

1. For software releases, the potential impact of changes to code and data structures should be taken into consideration. Even a simple unload/reload can become quite time consuming, especially if there is a VLDB implementation. In such cases, the users loose access to their data for a period of time unless some type of data mirroring is employed.

2. For installation of updated hardware such as servers and disks, one needs to be careful that the end users do not suffer. During certification and testing process users need to have access to their data all the time for enhanced performance. The options available largely depend upon the processor technology used since Cluster, AMP, MPP, and NUMA architectures have differing requirements for how the client and desktop technology are deployed.

3. The challenges for the network/communications environment are even more important especially if there is a distributed database operating across geographic regions and global boundaries. Enhancing the bandwidth and the ability of the network components and servers to handle large traffic volumes, as well as managing the differing types of protocols, often results in significant and risky challenges if not planned out well in advance. (ibid)

3.9.2 The Certification Testing Process

Given below is a certification procedure which can be followed to effectively perform certification testing for software, hardware and the network (RK00). It describes overall management, training, procedure and script definition, and project control mechanisms that are required to provide due diligence to a testing process.

1. Initiate product certification: Includes collecting data warehouse product certification requirements and reviewing certification products requested for conformance to existing or future architecture. A core team member is selected who defines a certification strategy and leads it. The proposed certification budget is also presented in this phase.

2. Establish certification team environments: The working place and environment is defined here. The certification team obtains copies of products(s) for certification. Certification procedures and techniques are outlined and developed along with the training of certification team.

3. Define certification test environment: A test environment is established for product certification. Data warehouse release stress testing data and procedures for test verification are obtained. In the end certification test procedures, scripts and test data are developed / refined.

4. Undertake certification testing: During this phase the new products are installed or existing products are upgraded. Certification testing for all databases, decision support services, data staging, information access and supporting technology equipment is performed. Additional activities include product installation and testing, database testing, application components testing running in the same environment etc.

5. Conclude certification testing and prepare for production turnover: During this phase the capacity management is updated which after updating includes the new features, functionality and capabilities of the new environment. Update to the current data warehouse production environment is scheduled. Lists of certified products are also updated and certification results are prepared for review and approval.

6. Implement certified products: Once the approval for certification testing is received it is the time for implementing certified products. Current production components are backed up if required and new product releases are activated. The monitoring of the new product versions is handed over to the data warehouse support team. In the same way product problem management and support is also transferred to the data warehouse support group and the certification cycle is closed. (ibid)

In undertaking a certification process, it is necessary to document what occurred and why and what was in or out of scope through the certification process (RK00). The generic deliverables described here discuss the types of information to be collected, which includes

1. Certification program

2. Certification business requirements

3. Certification test results (ibid)

3.9.3 Certification Program

A certification program is used to define the procedures, schedules, and facilities required to conduct certification testing (RK00). It describes the technical version or view of certification testing to the information services owner of the proposed data warehouse infrastructure improvement. To undertake the technical aspect of the certification program the following information is required.

1. A certification plan containing a list of steps, dates and resources presented as a GANTT chart or spreadsheet showing the certification testing scope and schedule or references to where this information can be found in electronic format.

2. A description of the test cycle(s) and each test cycle, test case, and the order in which they are exercised to verify the functionality of the hardware, software, or network improvement. Each test case consists of expected performance, inputs and outputs, before and after expected results, and a description of hardware, software or network product benchmark(s) to be met and a scoring method for testing.

3. A description of the automated and manual tools and techniques, if different from the development environment , for testing tools (software and hardware support facilities), test data generation method (generators, data entry, etc) and test bench marking approach and related techniques.

4. A description of the proposed /available certification environment in terms of hardware, software, communications and user access profiles. (ibid)

3.9.4 Certification Business Requirements

Certification business requirements describes the scope of certification testing in no technical terms and any conclusion or recommendations appropriate to the generation of any future enhancements to this or similar such efforts (RK00). It is also used to convey this understanding to the business stake holders and owners of the data warehouse. This deliverable contains the certification program definition explaining how the certification program was developed? Whether it was bought from some vendor or build in house? Is it customized from the current environment and describes the staffing model used to staff the certification program in terms of internal staff or external consulting? (ibid)

A scope of testing explanation, which describes the scope in terms of duration, impact, risk, and cost of certification testing as detailed in the certification test program and includes information on test schedule, an end user introductory seminar schedule, list of facilities, allocated technology components like printers etc, and a list of technicians and business area end users for participation in the certification process. (RK00)

It also describes any risks associated with the certification program and certification program recommendations, which identify any final productivity improvements. (RK00)

3.9.5 Certification Test Results

As the name implies these are the results gathered after going through the certification testing process (RK00). Certification test results provide an evaluation on the "state of readiness" of the data warehouse infrastructure improvement, its associated interfaces, and all data or software conversion procedures. It identifies what was tested and what was not tested what problem or issues remain outstanding and any procedural impacts, work around, or risks. (ibid)

Finally it provides recommendations whether to proceed or not proceed to implementation with the hardware, software or network improvement (RK00) .This report should contain the following information:

1. A statement of purpose describing the intention behind certification testing and the approach used.

2. Certification program results describing the success achieved and an assessment of the state of readiness of the various hardware, software, and network components.

3. A summary of the certification program, which details the evaluation method. The evaluation methods may include interviewing, walk throughs, kit review, feedback assessment and scoring results review.

4. Questions used to conduct certification testing and evaluation.

5. Any final recommendation to proceed or not proceed further based on possible impact on production process performance or capacity management limitations.

6. An appendix containing all detailed certification test program material compiled during the test process. (ibid)

3.10 Extract, Transform and Load (ETL)

ETL is a data integration function that involves extracting data from outside sources (operational systems), transforming it to fit business needs, and ultimately loading it into a data warehouse (TD04).

Companies know they have valuable data lying around throughout their networks that needs to be moved from one place to another such as from one business application to another or to a data warehouse for analysis (CM04). The only problem is that the data lies in all sorts of heterogeneous systems, and therefore in all sorts of formats. For instance, a CRM system may

define a customer in one way, while a back-end accounting system may define the same customer differently. (ibid)

To solve the problem, companies use extract, transform and load (ETL) technology, which includes reading data from its source, cleaning it up and formatting it uniformly, and then writing it to the target repository to be exploited (figure 3.1). The data used in ETL processes can come from any source: a mainframe application, an ERP application, a CRM tool, a flat file or an Excel spreadsheet. (CM04)

According to Mike Schiff an analyst at Current Analysis Inc data can be extracted using Java Database Connectivity, Microsoft Corp's Open Database Connectivity technology, proprietary code or by creating flat files. (CM04)

After extraction, the data is transformed, or modified, depending on the specific business logic involved so that it can be sent to the target data base (CM04). There are a variety of ways to perform the transformation, and the work involved varies. The data may require reformatting only, but most ETL operations also involve cleansing the data to remove duplicates and enforce consistency. Part of what the software does is examine individual data fields and apply rules to consistently convert the contents to the form required by the target repository or application, says Schiff. For example, the category "male" might be represented in three different systems as M, male and 0/1. The ETL software would recognize that these entries mean the same thing and convert them to the target format. (ibid)

In addition, the ETL process could involve standardizing name and address fields, verifying telephone numbers or expanding records with additional fields containing demographic information or data from other systems. (CM04)

Harriet Fryman, group director of product marketing at data warehousing vendor Informatica Corp. in Redwood City, Calif, offers an example (CM04). Say, for instance, that a customer runs Oracle financials, PeopleSoft human resources software and SAP manufacturing applications and needs to access the data in each of these systems to complete an order-to-cash process. This will require the company's ETL software to extract data from the originating systems, which isn't as easy as it sounds in some instances, for example, pulling data from the SAP manufacturing application would require the generation of SAP proprietary ABAP code to extract the shipping and open purchase-order information. The transformation occurs when the data from each source is mapped, cleansed and reconciled so it all can be tied together, with receivables tied to invoices and so on. (ibid)

Fryman says customers are using ETL not only for data warehousing and business intelligence activities, they're also moving data from one operational system to another, such as from an ERP system to a CRM application. (CM04)

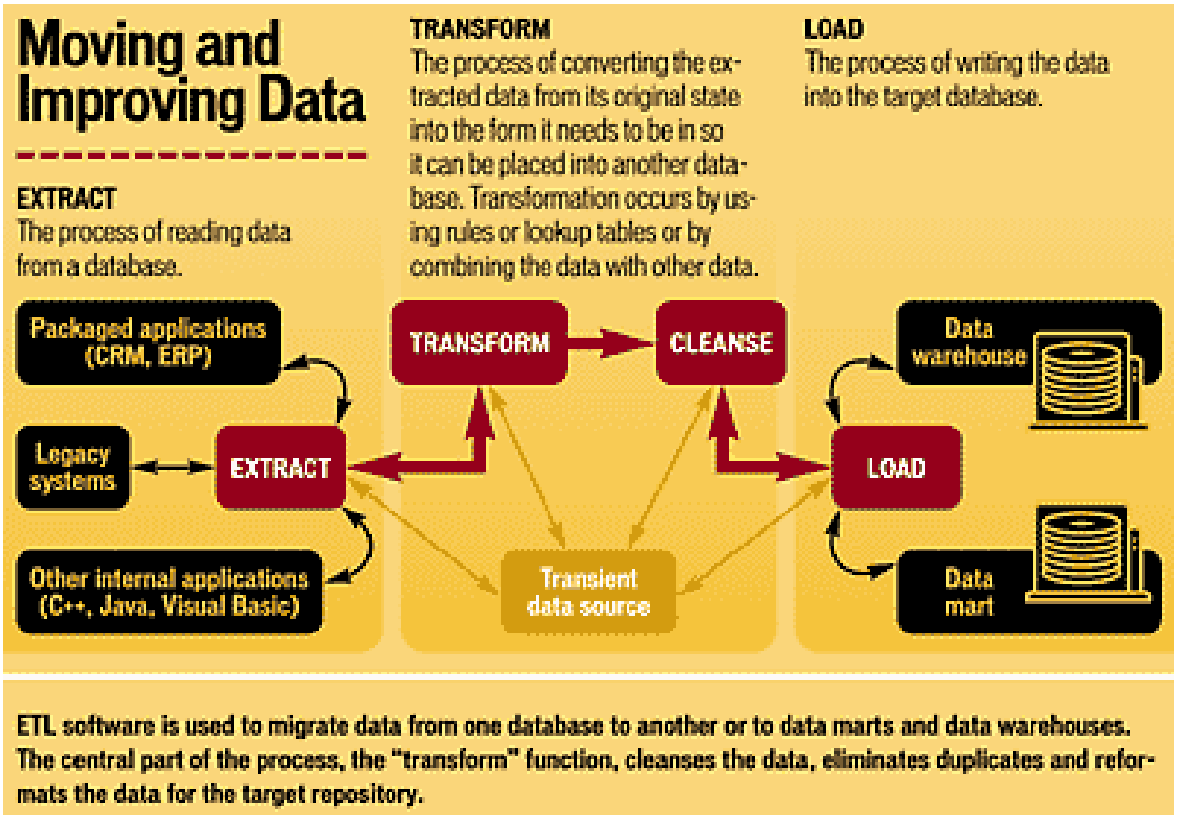

Fig 3.1: ETL function
http://www.computerworld.com/databasetopics/businessintelligence/datawarehouse/story/0,10801,89534,00.html

3.10.1 A Typical ETL Process

According to (CT03) the function of data acquisition in corporate information factory in which the data warehouse and operational data store are populated from operational sources is the most technically challenging and difficult part of any data warehousing environment. According to some industry experts approximately 60-80 percent of a data warehousing project effort is spent on this process alone. In today's high volume, client/server environment data acquisition techniques have to coordinate staging operations, filtering, data hygiene routines, data transformation and data load techniques in addition to cooperating with network technology to populate the data warehouse and operational data stores. (ibid)

The CIF architecture illustrates where the various ETL processes take place (figure 3.2) (CT03). It's more than just in the data acquisition process. While data acquisition is the predominant process using the ETL tools, the data delivery process and movement of data from the analytical functions to the ODS or operational systems use ETL processing as well. The full-blown set of ETL operations must combine into a cohesive, integrated system. A system that ensures each process will fit into the overall effort efficiently, determines how the tool will be used for each component and synchronizes all ETL events. (ibid)

There should be ETL expert in the organization who ensures that the ETL processes have strength and endurance (CT03). This requires an overarching view and control over the entire

environment and is the job of an ETL architect. The ETL architect ensures program efficiency by creating a cohesive ETL architecture to ensure that the various ETL functions form one cohesive system. (ibid)

Taking the time to properly architect a highly integrated set of processes and procedures up front is the fastest way to achieve a smoothly running system that is maintainable and sustainable over the long haul. To accomplish an efficient, scalable and maintainable process, the ETL architect must have the following roles and responsibilities (CT03):

1. The ETL architect should have a close eye on the needs and requirements of the organization. He/she must understand the overall operational environment and strategic performance requirements of the proposed system. The architect must interact with the source system operational and technical staff, the project database administrator (DBA) and the technical infrastructure architects (if different from the ETL architect) to develop the most efficient method to extract source data, identify the proper set of indexes for the sources, architect the staging platform, design intermediate databases needed for efficient data transformation and produce the programming infrastructure for a successful ETL operation.

2. An ETL programmer should not only see his or her single-threaded set of programs. The architect must see the entire system of programs, how they are interconnected, how they will influence and affect each other and, ultimately, how the software coding tools must interact with the technical infrastructure to create a seamless environment. He/she must ensure the technical team understands the target database design and its usage so that the transformations which convert the source data into the target data structures are clearly documented and understood. The ETL architect oversees each and every one of the ETL components and their subcomponents.

3. The ETL process is much more than code written to move data. The ETL architect also serves as the central point for understanding the various technical standards that need to be developed if they don't already exist. These might include limits on file size when transmitting data over the company intranet, requirements for passing data through firewalls that exist between internal and external environments, data design standards, standards for usage of logical and physical design tools and configuration management of source code, executables and documentation. The ETL architect must also ensure that the ETL design process is repeatable, documented and put under proper change control.

4. A key consideration for the ETL architect is to recognize the significant differences that the design and implementation methods for a business intelligence system have from an online transaction processing (OLTP) system approach. An OLTP system only changes in design when the operational process it manages changes, while BI systems must constantly adapt as business users discover new and different ways of analyzing their businesses. BI systems must be scalable from a volume perspective but must also adapt to changing business processes and technologies without requiring a complete redesign and conversion. For example, the current Web-based business environment demands that BI not only address strategic needs but integrate on a tactical level with day-to-day processing. This requires a forward-thinking architect who recognizes that the ETL processes must integrate with the needs of a real-time warehousing effort. (Note: This is indicated in the figure ‘ETL in CIF’ by the lines

linking the warehouse and the data delivery layer back to the ODS and then back through the user transaction interface to the Web site.)

5. The role of the ETL architect also extends to that of consultant to the programming effort. The architect works closely with the programmers to answer questions and plays a key role in problem resolution. Depending on the size of the programming effort and the project organization, the ETL architect may also supervise the development of the programming specifications. In any case, the ETL architect plays a key role as a reviewer and approver during the peer review process.

6. One last role for the ETL architect must be to ensure that the various software tools needed to perform the different types of data processing are properly selected. The yellow boxes in the figure 'ETL in CIF' show each point in the CIF architecture requiring some kind of extract, data transformation or data load operation. Each of these ETL functions has a different purpose and, as such, may not necessarily require the same functions within the software tool. Under the guidance of the ETL architect, a well planned and documented ETL architecture, at least at a high level, will define these purposes and functions as input into the tool selection process. (ibid)

ETL is one of the most important sets of processes for the sustenance and maintenance of Business Intelligence architecture and strategy (CT03). Time and thought are required to ensure the best architecture for its various components as well as for the selection of appropriate software tools and procedures within each component. Ongoing Business Intelligence development demands a flexible, scalable and easily maintainable environment that can only come from an architected approach. (ibid)

This type of architecture must be driven from a central focal point led by the ETL architect (CT03). The need for an ETL architect should be obvious for large systems growing exponentially. The ETL architect is equally important in smaller setups as well. Small systems have a way of growing, and the smart development team will be ready, willing and able from the start to take on the growth with a well-architected environment. (ibid)

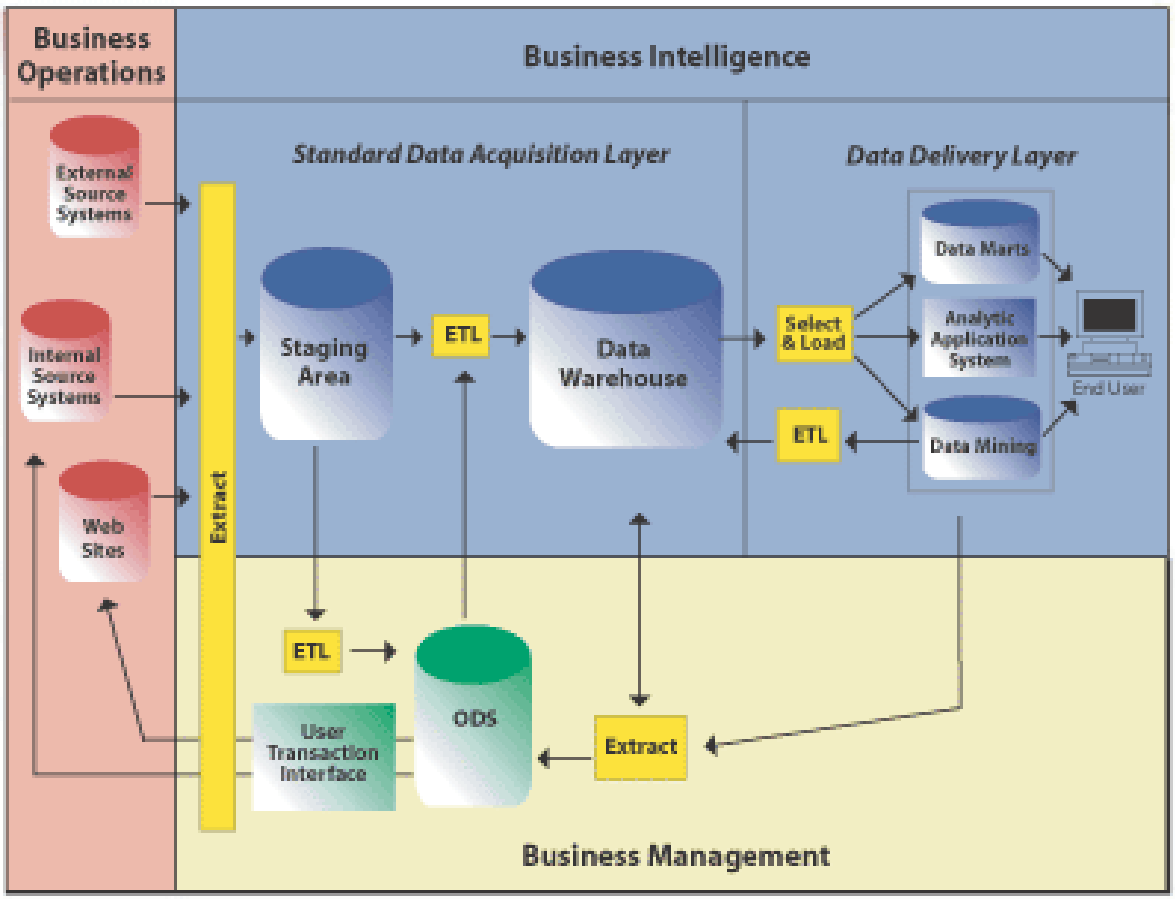

Fig 3.2: ETL in Corporate Information Factory

3.11 View Materialization

There are two approaches towards providing integrated access to multiple, distributed, and heterogeneous databases:

1. Lazy or on-demand approach to data integration, which often uses virtual view(s) techniques.

2. Data warehousing approach, where the repository serves as a warehouse storing the data of interest (JKQ97).

One of the techniques this approach uses is materialized view(s) (figure 3.3) (JKQ97). The virtual view approach may be better if the information sources are changing frequently. On the other hand, the materialized approach would be superior if the information sources change infrequently and very fast query response time is needed. The virtual and materialized view approaches represent two ends of vast spectrum of possibilities. In view of (JKQ97) it may be more efficient not to materialize all the views, but rather to materialize certain “shared” portions of the base data, from which the warehouse views can be achieved. (ibid)

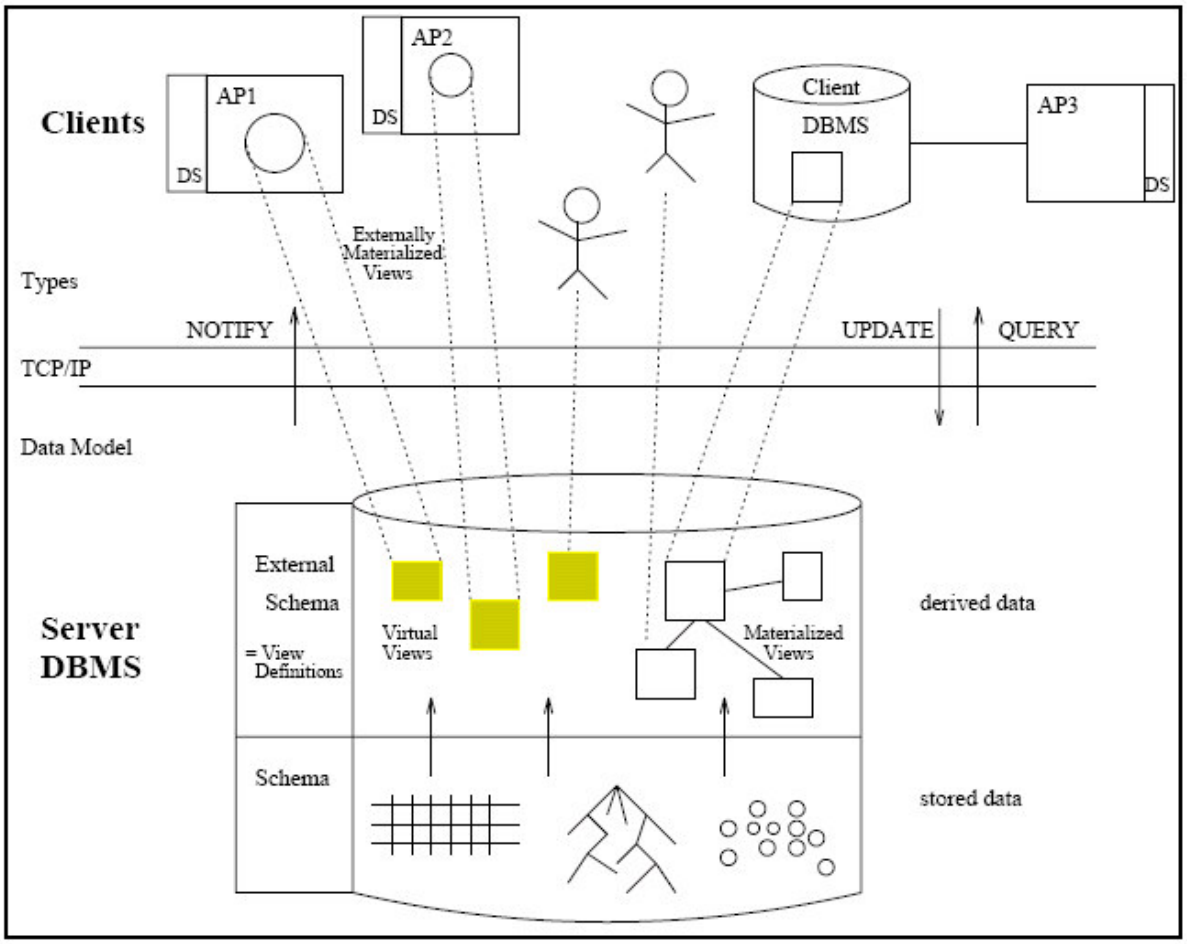

Fig 3.3: Materialized views within a client server DBMS (SJ96)

Queries against complex view definitions must be answered very fast because users engaged in decision support activities require fast and quick answers to their questions (RG03). Even with sophisticated optimization and evaluation techniques, there is a limit to how fast one can answer such queries. Also if the underlying tables are in a remote database, the query modification may not even be feasible because of issues like connectivity and availability. (ibid)

View materialization is another approach used in place of query modification under such circumstances (RG03). Generally in view materialization we precompute the view definition and store the result. When a query is posed on the view, the unmodified query is executed directly on the precomputed result. This approach is much faster than query modification approach because the complex view is not computed again when the query is computed. Materialized views can be used during query processing in the same way as regular relations, for e.g. we can create indexes on materialized views to further speed up query processing. But to have best results we have to maintain the consistency of the precomputed or materialized view whenever the underlying tables are updated. (ibid)

Three questions needed to be considered when thinking of view materialization (RG03):

1. What views should we be materialized and what indices should be built on the materialized views?

2. Given a query on a view and a set of materialized views, can the materialized views be exploited to answer the query?

3. How should one synchronize materialized views with changes to the underlying tables? The choice of synchronization technique depends on several factors such as whether the underlying tables are in a remote database or not etc. (ibid)

The answers to the first two questions are related (RG03). The choice of views to materialize and indices to build is dependent on the expected workload. The choice of views to materialize is more complex than just choosing indexing on a set of database tables, however, because the range of alternative views to materialize is wider, the goal is to materialize a small carefully chosen set of views that can be utilized to quickly answer most of the important queries. Conversely once we have chosen a set of views to materialize, one has to consider how they can be used to answer a given query. (ibid)

View materialization is fast and efficient way for accessing views just like a cache (RG03). A view can be materialized by storing the tuples of the view in the database. Index structures can be built on the materialized view. Consequently, database accesses to the materialized view can be much faster than recomputing the view. (ibid)

A large number of data warehouse queries require summary data, and, therefore, use aggregates (SU96). Hence, in addition to indices, materializing summary data can help to accelerate many common queries. For example, in a production environment, majority of the queries may be based on the production results of the most recent quarter and the current fiscal year. Having summary data on these parameters can significantly speed up query processing. These days the most common strategy used is based on materializing views that have a relatively simple structure. Such views consist of joins of the fact table with a subset of dimension tables (possibly after some selections on those dimensions), with the aggregation of one or more measures grouped by a set of attributes from the dimension tables. (ibid)

Points to consider while selecting views to materialize include workload characteristics, the costs for incremental updates, and storage requirements. (RG03)

A simple and very useful technique to use a materialized view is to use selection on the materialized view by grouping and aggregating on additional columns (RG03). For e.g. assume that a materialized view contains the total sales for a quarter for every product the company produces. This materialized view can be used to answer a query that requests the total sales of a particular product for a year by first applying the selection and than rolling up from quarter to year. (ibid)

The main objective of a materialized view is to improve query performance (KJM01). However, when a warehouse is updated especially due to the changes of remote information sources, the materialized views must also be updated. While queries calling for up-to-date information are growing and the amount of data reflected to data warehouses has been increasing, the time window available for making the warehouse up-to-date has been shrinking. Hence, an efficient view maintenance strategy is one of the outstanding issues in the data warehouse environment. This can improve the performance of query processing by

minimizing OLAP queries down time and interference. There can be roughly two different methods in reflecting data changes to materialized views: recomputation and incremental maintenance. Incrementally maintaining a materialized view includes computing and propagating only its changes. Compared to the sizes of base relations and views, their changes are generally very small. Hence, it is cheaper to compute only the changes of a view rather than to recompute it from scratch. (ibid)

3.11.1 Maintenance of Materialized Views

Data warehouses usually contain a very large amount of data (SU96, HGB01). In this scenario it is very important to answer queries efficiently therefore we need to use highly efficient access methods and query processing techniques. It is an important physical design decision to decide which indices to build and which views to materialize. The next major issue to deal with is how to use these indices and materialized views efficiently for maximum output. Optimization of complex queries is another important problem. We also need to take advantage of parallel query processing to reduce query response time.

1. How a view when an underlying table is modified? Two issues of particular interest are how to maintain views incrementally, that is, without recomputing from scratch when there is a change to an underlying table, and how to maintain views in a distributed environment?

2. When should we refresh a view in response to a change to an underlying table? (ibid)

A data warehouse contains data from autonomous sources (SU96, HGB01). When data in sources are updated there is a need to maintain the warehouse views in order to keep them up-to-date. This propagation of changes is commonly referred to as view maintenance and several potential policies for this have been suggested in the literature. As an example, the warehouse can be maintained immediately when a change is reported from a source, or it can be maintained periodically, for example, once per night. Also there may be a choice to maintain the warehouse incrementally, or to reload all data from scratch. Given a set of potential policies, a data warehouse designer has to decide which policy to use for a specific situation. Choosing an optimum policy is not a trivial task. Impact on system overhead and quality of service needs to be considered, and the level of impact is, amongst other things, dependent on the services provided by sources. (ibid)

A data warehouse stores integrated information from multiple data sources in the form of materialized views over the source data (SMVY99). The data sources may be heterogeneous, distributed and autonomous. When the data in any source changes, the materialized views at the data warehouse need to be updated accordingly. The process of updating a materialized view in response to the changes in the underlying source data is called view maintenance. The view maintenance problem has evoked great interest in the past few years. This view maintenance in such a distributed environment gives rise to inconsistencies since there is a finite unpredictable amount of time required for propagating changes from the data sources to the data warehouse and computing view updates in response to these changes. (ibid)

Data consistency can be maintained at the data warehouse by performing the following steps (SMVY99):

1. Propagate changes from the data sources (ST1 - current state of the data sources at the time of propagation of these changes) to the data warehouse to ensure that each view reflects a consistent state of the base data.

2. Compute view updates in response to these changes using the state ST1 of the data sources.

3. Install the view updates at the data warehouse in the same order as the changes have occurred at the data sources. (ibid)

The inconsistencies at the data warehouse occur since the changes taking place at the data sources are random and dynamic (SMVY99). Before the data warehouse is able to compute the view update for the old changes, the new changes change the state of the data sources from ST1 to ST2. This violates the consistency criterion. Making the materialized views at the data warehouse self-maintainable decimates the problem of inconsistencies by eliminating the finite unpredictable time required to query the data source for computing the view updates. (ibid)

Data warehousing is used for reducing the load of on-line transactional systems by extracting and storing the data needed for analytical purposes (e.g., decision support, data mining) (AASY97). A materialized view of the system is kept at a site called the data warehouse, and user queries are processed using this view. The view has to be maintained to reflect the updates done against the base relations stored at the various data sources. Efficient view maintenance of materialized views is very important because the update efficiency of the warehouse view is counter balanced by the query overhead at the data sources. Several approaches have focused on the problems associated with incremental view maintenance. Such problems include dealing with the anomalies resulting from the order of processing events, the levels of consistency in reflecting the source states in the view and the current validity of the view. (ibid)

A very simple approach to refreshing a view is simply to recompute the view when an underlying table is modified (AI99, RG03). In most cases it is wasteful to maintain a view by recomputing it from scratch. Often it is cheaper when only parts of the view changes in response to changes in the base relations and thus compute only the changes in the view to update its materialization. We stress that the above is only a heuristic. For example, if an entire base relation is deleted, it may be cheaper to recompute a view that depends on the deleted relation (if the new view will quickly evaluate to an empty relation) than to compute the changes to the view. Incremental view maintenance could be reasonable strategy if the underlying tables are in a remote database. In this case the views can be periodically recomputed and sent to the data warehouse where the view is materialized. This has the advantage that the underlying tables would not be replicated at the warehouse. (ibid)

Algorithms that compute changes to a view in response to changes to the base relations are called incremental view maintenance algorithms (RG03). In incremental view maintenance algorithms; the cost for refreshing a view is proportional to the extent of the change rather than the cost of recomputing the view. (ibid)

To understand the intuition behind incremental view maintenance algorithms, observe that a given row in the materialized view can appear several times, depending on how often it was

derived (RG03). The main idea behind incremental maintenance algorithms is to efficiently compute changes to the rows of the view, either new rows or changes to the count associated with a row, if the count of a row becomes 0; the row is deleted from the view (ibid).

3.11.2 View Maintenance Policies

A view maintenance policy is a decision about when a view is refreshed, independent of whether the refresh is incremental or not (RG03). The two common strategies used are:

Immediate View Maintenance

A view can also be refreshed within the same transaction that updates the underlying tables (RG03). This view maintenance technique is called immediate view maintenance. In this scenario the update transaction is slowed by the refresh step, and the impact of refresh increases with the number of materialized views that depend on the updated table. (ibid)

Deferred View Maintenance

As an alternative to immediate view maintenance in this technique updates to the base tables are captured in a log and applied at a later stage to the materialized views (RG03). There are further three techniques in deferred view maintenance which are:

1. Lazy: The materialized view is updated when a query accesses the view.

2. Periodic**:** The materialized view is updated after a certain period of time for e.g. once in a day or during the nights etc.

3. Forced**:** The materialized view is refreshed after a certain number of changes have been made to the underlying tables. (ibid)

It's the responsibility of the data warehouse team to make decision regarding the view maintenance strategies (RG03). Generally the data warehouse designers provide capabilities in the warehouse to select which policies to use. The data warehouse management team needs to decide which policies to use keeping in view:

1. Data warehouse usage during peak hours

2. Number of users accessing the data warehouse

3. Number of accesses to a particular view or table

4. Network capacity

5. Base tables update frequency (ibid)

CHAPTER 4: EMPIRICAL FINDINGS

In this chapter we have presented our findings from the case study that we have done. We have presented how data warehouse maintenance is actually carried out in an organization.

Case Study: Telenor Pakistan Ltd

4.1 Introduction

Telenor is a leading provider of communication services and one of the largest mobile operators worldwide. Telenor's market value as of 31 March 2005 was NOK 100 billion making it the third largest company listed on Oslo Stock Exchange. At the end of 2004, the Group had 21,750 employees, 9,750 of whom were employed outside Norway.

Telenor acquired the license for providing GSM services in Pakistan in April 2004, and has launched its services commercially in Islamabad, Rawalpindi and Karachi on March 15, 2005. On March 23, 2005 Telenor started its services in Lahore, Faisalabad and Hyderabad. Telenor is currently planning to launch its services in other cities of the country as per the roll out plan. Telenor has its corporate headquarters in Islamabad, with regional offices in Karachi and Lahore.

The license terms stipulates that by year 4, Telenor will cover 70% of Pakistan's 297 administrative headquarters (cities). Telenor will fulfil the license requirements and provide superior quality coverage. The company has covered several milestones over the past eleven months and has grown in a number of directions.

Telenor has successfully signed interconnect agreements with all five incumbents during December 2004, allowing its subscribers to exchange voice and data with subscribers on all active mobile networks including Paktel, Instaphone, Ufone, Mobilink and Al-Warid telecom.

In addition to recruiting hundreds of people, Telenor established its call center on January 28 in Lahore. Telenor Pakistan has currently employed more than 1000 employees all over Pakistan to execute its business and will employ more and more people as its business will expand.

Telenor’s primary aim is to offer top quality mobile services and promote healthy competition in the mobile market. Also Telenor aims to create value for shareholders through the serving of customers, employees, partners and the general public interest. In a long-term perspective, a strong market and customer focus, as well as a strong commitment to their employees and to society, will provide the best platform for creating incremental value in their business.

To achieve this goal Telenor Pakistan has been totally computerized for its day to day operations and is completely dependant on its information systems for the running of daily business. There are 5 major operational systems namely:

1. SIEBEL CRM (customer relationship management)

2. GENEVA (billing and postpaid traffic)

3. MEDIATION
    i. Postpaid: For the postpaid subscribers of Telenor the traffic is rated using Mediation postpaid.
    ii. Prepaid: For the prepaid subscribers of Telenor the traffic is rated using Mediation prepaid.

4. VOMS (Voucher Management System): All the prepaid scratch cards, electronic credit transfers and easy loads are managed through VOMS.

5. MSC: The raw traffic source for Telenor is MSC source system. In this all the traffic including prepaid, postpaid, inter-connect and transit is managed.

In addition to these five major operational systems, Telenor, since the day it started its operations in Pakistan has been using an Oracle based data warehouse to strengthen its decision making process. However since April 2005 it has switched to Teradata warehousing solutions. The data warehouse is located at its central head office in Islamabad. This is a centralized data warehouse having the main database at one location but it is in the process of conversion to a distributed data warehouse and professionals from Teradata (a subsidiary of NCR) are working on it to finish this job.

According to the data warehouse project manager at Telenor, data that is useful and helpful from all the source systems is stored in the data warehouse. This helps in the consolidation of data at a central repository. When this data is projected over a period of time a trend can easily be detected in the projections. For e.g. it can be detected which are the cell sites that are used more often and which are not. In this way a capacity planning for the cell sites can be conducted by the BSS department. Similarly activations for each area can be projected over a certain period of time and it can be made out what are the places where the growth is maximum and the brand is popular. Similarly the age group in which the brand is famous is another valuable demographic which can be put to use by Marketing and Brand Management department.

There are nearly 50 employees currently accessing data warehouse including personnel from data warehouse department, operations department, business intelligence department and business analysts group. The data warehouse department is responsible for performing tasks related to maintenance of data warehouse. All the operational systems of Telenor listed above are centralized. Data from all the operational systems is consolidated at a certain central location called COB (close of business). This COB is actually the ODS (operational data store) as described in data warehouse architecture (section 1.7.1). Telenor is storing data from several operational systems in the data warehouse. This includes all the data related to finance, call history, and subscriber's database.

Then the data warehouse is fed using the push or pull scheme depending on the operational system. Mostly the traffic or call history related data is pushed through ftp to a location which is then parsed, moved to staging, transformed and then loaded into the data warehouse. The entire subscriber's related data is extracted from the current source systems and then it is moved to staging and then transformed and loaded into the production system of the data warehouse.
According to the data warehouse project manager, Telenor has nearly 50 employees using the data warehouse. Out of these 50 employees, 9 are member of the data warehouse core team and about 4 people in the support department. Following are the major positions in the enterprise data warehouse team.

1. Project Manager (PM): This person is the over-all in-charge of the project. He is answerable for the data warehouse at the highest level. He reports to the higher management and is responsible of delegation of powers and trainings of the core personnel.

2. Technical Team Lead (TTL): TTL is the person looking after the technical side of the project. He leads the separate ETL (Extraction Transformation and Loading) team, BI (Business Intelligence) team, BA (Business Analysts) team and Managed Services (MS) Team. He is the key personnel who is accountable to the PM for all the workings of the data warehouse. He should have a strong LDM (Logical Data Model) knowledge and should have clear understanding of the business rules related to the specific industry he is working for.

3. Extract, Transform & Load Lead (ETL Lead): ETL Lead is the person who is responsible for mapping the source system data into the data warehouse. He carries out all the extractions from the source system, fast loads them into the raw format of the data warehouse and then applies transformations and loading into production. He has a team of techies who carry out the coding related tasks.

4. Business Intelligence Lead (BI Lead): The BI team is lead by the person who builds aggregates on the production data and transforms into the form which is easier and faster for the reporting and ah-hoc querying. BI team is the team constantly in touch with the business users and this team is the point of contact between the business users and the technical team.

5. Business Analysts Group (BA group): The Business Analysts group consists of the persons responsible for suggesting value additions to the business. This group analyses the trends in the data projections and then provides suggestions to the data warehouse team to bring about a change in the reporting and data transformation strategy for an effective and improved reporting system. Moreover they also suggests to business to make decisions on certain reports.

6. Support and Backup Team: The support and backup team performs functions related to the help desk, problem management and data backup. Members of the support and backup team are not permanent and they perform other duties such as related to ETL and training etc.

Now we'll present an overview of the tasks performed by Telenor to maintain its data warehouse for optimal performance.

4.2 Communication & Training

According to the Project Manager, Telenor has started its communication program by publishing a booklet related to the data warehouse. The booklet contains complete information about the purpose of the data warehouse, the scope of the project, the aim behind the implementation of data warehouse, what input is given to the system, what output can be taken from the system, and who are the responsible persons, those should be contacted to get any further information and can help in any matter regarding data warehouse. The members of the data warehouse core team are in constant contact with the business users, and they keep them informed about any development or shortcomings in the projects.

Project manager further elaborated that the most difficult part in implementing a data warehouse in the company is the training of business users. Most of the business users are technology shy. They just want the reports and analysis on paper and everything printed. Making them comfortable with the use of computer and other multimedia facilities is a very tough task. Once the business users have dipped their hands in this area they realize how important and how beneficial it is for their better understanding of the use of data warehouse. The rich reporting which is possible with the modern BI tools such as Cognos, Business Objects (BO) and Microstrategy can never be imagined with paper and pencil. Business users cannot be pushed to start using data warehouse. They have to be informed about what can be achieved from the data warehouse and how easily.

Telenor has started a comprehensive education and training program for its employees. For this purpose they have set aside a good amount in the overall data warehouse budget. The training is conducted for each and every person having any kind of interaction with the data warehouse. First of all the technical team as a whole is briefed about the over-all scope of the data warehouse, then they are delegated different jobs and then accordingly each one of them is trained. The persons related to the ETL and Design and Infrastructure are trained accordingly. They are made to take up some LDM related trainings and industry level courses. Similarly the BI team enrolls in some courses related to the business tools being used in the organization. Same goes for the managed services people who are trained on the lines of the DBA. Managed services personnel are trained locally by the vendor, while for the logical data model training the personnel are sent to the state of the art training centers of the vendors situated in Germany.

If we see the overall structure of the program, it is divided into three levels of detail depending on the competence of employees. The employees with little or no knowledge of computers are given level 1. Employees with intermediate competence are placed in level 2. Here intermediate competence means that they can operate a computer system, have basic knowledge of databases and networking and can make reports using the computer systems. These are the employees which are already working on any other operational system and are just operating that. They have no information about the working or the functioning of the operational system but they can operate the system and can take output from that. If there is some problem with the operational system they are helpless. Employees with advanced skills are placed in group 3. These are professional computer experts and are responsible for the management, maintenance and development of operational systems used by Telenor.

According to the employee competence, education is also divided into three levels namely basic, intermediate and advanced. The basic education is intended for level 1 employee with

little or no computer literacy. The basic education consists of lessons about working of computer, what are data bases, what are data warehouses, why they are used and how they can help in day to day matter of the company and other topics similar to these. A level 1 employee who gets the basic education is promoted to level 2 and he becomes eligible for intermediate level education.

A level two employee is given intermediate level education consisting of how to use the tools to extract data from data warehouse, trouble shooting any problem (problem management) and some other like thing. A level 2 employee who gets the basic education is promoted to level 3 and he becomes eligible for advanced level education.

A level 3 employee is given advanced level education consisting of issues related to data warehouse design and implementation, query maintenance, network management and future planning (capacity planning). It is the management's decision to decide which employees can get advanced level education.

It is the responsibility of level 3 employees to provide education to level 1 and 2 employees, while level 3 employees are trained by the professionals from the industry or from the data warehouse vendor.

	Basic	Intermediate	Advanced
Level 1	X		
Level 2	X	X	
Level 3	X	X	X

Table 4.1: Three levels of education and training

According to a Business User of Telenor, the training and education program has helped him a lot in understanding the capabilities and functionalities of data warehouse and he can now easily look for historical data that can greatly help in estimating future network and technological requirements for the business.

4.3 Help Desk & Problem Management

According to the data warehouse Project Manager, there is no formal help desk created in Telenor for providing support to the data warehouse users. As the data warehouse project is in the beginning the data warehouse core team is still thinking of how to implement the help desk services. For the time being they have created a support and backup team for providing the 24x7 hours support to the data warehouse users. The team members provide answers to user's queries and help them if there is any technical or other problem with data warehouse usage. The problems encountered by the users can be of any type including report generation, usage of tools for data mining, query problems, and user management etc. The members of this team are knowledgeable in the field of data warehousing and have the ability to solve any problem related to data warehouse on their own. If the problem is beyond their control they report the problem to the Technical Team Lead. It is the responsibility of the support and backup team to report the most commonly and most frequently found problems to the Project

Manager so that the root cause of the problem could be solved, whether in the form of any update to the system or development of a new module.

They are also given training for this purpose from time to time. The support and backup team members can be contacted either by phone, email or in person at any time. The support team logs all the questions and their responses in a web based system where other users can also check and learn with their experiences. The support and backup team can also be contacted through a special interface on the company's local web based portal, where the users can send a memo or a note to the backup and support team related to any functionality problem, or any enhancements for the data warehouse.

According to a member of the support and backup team 'they receive a lot of queries each day regarding what data can be taken out from the system, how that data can help the business users, how to get the data, how to generate a report and how to use a front end tool etc'. They respond to all these questions quickly, so as the business user will not feel any discomfort with the data warehouse.

There are special features provided by the vendor when there are any problems with the system. For e.g. the members of support and backup team are automatically intimated when there is a problem with the data warehouse. For instance during an ETL window when the loading fails a SMS is generated for the specific area and is sent to the mobiles of the support team members. They are also responsible for taking weekly back-ups and archiving the data warehouse. Incase of a hardware failure this team is responsible for rectifying the defect or replacing the hardware.

The problem management team comprises of the Project Manager, Technical Team Lead and the Support Team Lead. Project Manager is the person who decides how to cater the problem. The Technical Team Lead highlights its implications on the data warehouse and its daily loading. The lead of the backup and support team solves the problem.

4.4 Network Management

According to the data warehouse project manager the data warehouse needs constant high speed connectivity with the network. It needs high speed connection with the source systems preferably with the fiber optic link to execute the daily push and pull operation in which the files are extracted from the source or the files are thrown onto the landing server to be picked up by data warehouse through the FTP. Moreover the repository of the business application is placed on a separate server. A SAN (storage area network) is needed for storing any valuable data from the data warehouse. The FTP is configured on a landing server. The files are transferred on a landing server which is then loaded into the data warehouse. All the scripts for transformation and loading are also present on the landing server.

Telenor currently has a 100 MBPS network, which is operating fine for its data warehousing needs. Due to the fact that Telenor's data warehouse is not too old, its usage is not too much as well and also the data warehouse is centralized it doesn't need high maintenance of the network. Once all the links are configured and made secure the process normally runs smoothly. A team of network engineers is responsible for the maintenance of the network

related tasks. This team is also responsible for the overall network of the company. The team utilizes some network monitoring tools to ensure the smooth and reliable operation of the network. Once in a week the team checks as if all the links are working properly and require any maintenance or tuning.

After the up gradation of the data warehouse from centralized to a distributed there will definitely be a need to upgrade the network. In that case Telenor is planning to implement a fiber optic network.

4.5 Software & Hardware Issues

From the interview with Project Manager we found out that the minor hardware problems are looked after by the backup and support team. Moreover incase of problems relating to the major hardware changes are looked after by the hardware vendors like Teradata (a subsidiary of NCR Corporation).

The hardware certification is done by the vendor as they are responsible for its installation, maintenance and up gradation. Telenor has adopted the policy to use hardware from the same vendor if available because of compatibility and performance issues. As far as data warehouse is concerned Telenor is using hardware from the same vendor so they are not having too much problems related to hardware. The same is the case with software. The most common thing that needed to be updated frequently is the storage media. Technical Team Lead is the person responsible for keeping an eye on the hardware resources present and the hardware resources required. If there is some shortcoming in performance due to hardware issues it is reported to the Project Manager and a decision for hardware up gradation is taken in consultation with the vendor.

From the interviews it was observed that as far as the software side is concerned, the professional services side of the vendor trains all the data warehouse resources for keeping them up to date. The business software problems are being looked after by the 3rd party whose software Telenor is using. Telenor has entered into an agreement with the data warehouse vendor for the software up-gradation. Whenever a bug or some problem is found in the system, it is reported to the vendor. Afterwards it is the responsibility of the vendor to solve the problem. This is usually done by providing a software update for the product or by doing some troubleshooting. After every three months professionals from the vendor side also visit the data warehouse site and check its performance and determine if there are any problems or any update is required or not. The areas where problems are frequently found include data loading mechanisms and query management. The vendor also notifies if there is any need for hardware up-gradation or not.

For the logical data model training, the resources are sent to the state of the art training centers of the vendors situated in Germany. Managed services personnel are trained locally by the vendor.

4.6 Extract, Transform and Load (ETL)

According to the data warehouse Project Manager at Telenor, ETL is one the most important and most time consuming process. There are five systems feeding the data warehouse. All these applications have a different platform than that of the data warehouse. Also there are different formats used by the operational systems and by the data warehouse. For e.g. there may be a situation where the operational systems allow the use of null values while the data warehouse does not. To cop with these types of problems and ensure a consistent and reliable ETL operation Telenor has employed a specialist ETL person and given the title ETL Lead.

Professional services for the ETL function are provided by the vendors. Once all the services are in place, the data warehouse team for Telenor takes over and they are the owners of the data warehouse. The ETL portion comes under the authority of the Technical Team Lead which eventually follows it up with the ETL Lead and his resources (personnel). So at present the data warehouse ETL team is carrying out all the ETL daily jobs. ETL lead is responsible for managing all the ETL related tasks. ETL lead is a specialized person in the field of databases having complete command over database structures, database connectivity, data extraction, data transmission, and programming. It is his responsibility to assign duties to members of the ETL team. Any problem that is out of control for ETL team members are reported to ETL lead who then tries to solve it himself otherwise reports it to the Technical Team Lead.

Technical Team Lead and ETL Team Lead are helped by the industry consultants for communications in this regard. Although the role of the industry consultant is limited but during the UAT (user acceptance testing) and SIT (system integration testing) the presence of the industry consultant is a must. He is the person who understands the business as well as the technical side of the product. So he is the best person to ask for help incase a problem arises in the ETL architecture.

On the other hand there is very little to do once the process is in place. It is thoroughly automated with the help of SLJM (simple load job monitor).

ETL Lead works in close cooperation with the Business Intelligence Lead. This helps both the parties in determining what level of data detail are required and which levels of aggregation are best suited for the needs of Telenor.

The problems found during ETL operation can be summarized as follows:

1. As told earlier that Telenor migrated from an Oracle based data warehouse to a Teradata platform. Although the baseline queries for data extraction remain the same but the methodology is totally different.

2. While switching from the legacy systems to the new data warehouse the most difficult part is the extract specification. There are fields which are in compatible, changing them to the new specifications. Moreover when there is a change in the extracts specification the source system doesn't intimate the data warehouse team.

3. There are times when the data type which is used for a certain attribute is wrong. A field which has to be aggregated should be some number like decimal or integer.

4. Mapping issues are also very critical. Once the mapping is in place and there is a change in the mapping by the source system the whole scripts need to be re-visited and there are times when the table definitions needed to be changed.

4.7 View Maintenance

According to data warehouse Project Manager at Telenor views are made according to the business user needs for the aggregates and for the base tables. They allow limiting the user access needs on the production tables. So once there are changes in base tables, views needed to be updated as well. Services for view maintenance are provided in the data warehouse by the vendor. It is the decision of the user to decide which views to materialize and how they will be updated. They just have to use certain commands to update a view in response to changes to the underlying tables.

Telenor is using incremental view maintenance for refreshing views on tables. Instead of loading all the data from scratch the views are updated incrementally whenever the data in source systems or base tables is updated.

The policy used for view maintenance depends on the type of data and the frequency of its usage. For e.g. for data that is important (a minor change can effect the results badly) the views are immediately updated while for views containing data that is not important (changes in data do not affect the results too much) deferred view maintenance is used. Functionalities for all these operations related to view maintenance are provided by the vendor in the form of data warehouse tools.

If there is some problem found in the view maintenance it is the job of the backup and support team to intimate the ETL team regarding any changes in the structure of the tables. Afterwards ETL team takes care of the problem.

CHAPTER 5: ANALYSIS AND DISCUSSIONS

In this chapter we will compare the theoretical findings (Chapter 3) with the empirical findings (Chapter 4) and will discuss which approaches are better than others and why?

Now we will analyze our theoretical findings and the empirical findings by comparing with each other. We will try to find out what are the similarities and differences between the theory and the real world system. Here again we will follow our previous structure for data warehouse maintenance by using communication and training, help desk and problem management, network management, software and hardware issues, ETL operations and view maintenance as the main concepts for analysis and comparisons.

5.1 Communication & Training

According to the data warehouse Project Manager at Telenor Pakistan LTD, communication between the users of data warehouse and training of the data warehouse users are the cornerstone for the success of any data warehousing project. Without proper communication and training users will not know what the system is meant for and how to take output from it. In this case the expensive data warehouse project will fail and all the investment will be lost.

Telenor has nearly 50 data warehouse users in the organization, majority being the business users having no or little knowledge about the data warehouse. In the starting of the data warehousing project most of the users have no idea about how data warehouse should be operated and how it can help in decision making? Exchange of views, experiences and ideas between the business users and the technical people has greatly helped business users in having a better understanding of the system and they can use the system according to their needs and requirements. The data warehouse experts within Telenor pass on their knowledge of the system to the business users by exercising the concept of communication and training.

Project Manger further elaborated that it is worth mentioning that Telenor has not implemented any communication program which is contradicting to the guidelines found in the theory written by RK00 (section 3.4), who is of the view that there should be a separate communication program having a communication team, its own budget, standards, and resources such as facility location and digital equipment. Instead they have merged it with the education and training program. They have followed a very simple path for the communication process by just distributing a booklet amongst the company employees containing all the information that is useful for any ordinary employee. Secondly the members of the core team are in constant contact with the non-technical employees keeping them informed about the capabilities and functionalities of the data warehouse.

The training program implemented by Telenor is very effective and diverse. It has provided an opportunity to nearly every employee in the organization to take benefit and learn the tools

and techniques to gather data according to his needs. Employees who even don't know what are databases or information systems can take up the beginner's level course and can hugely increase their skills in general computer usage. Employees with some understanding of computers and information systems can further increase their skills and knowledge by taking the intermediate level courses. After these courses they can operate the data warehouse front end tools (OLAP tools) and take the desired output from the system. Similarly advanced level courses are meant for data warehouse administrators and support staffs so they can better manage the data warehouse and provide the help services at any time.

The training program implemented at Telenor is more or less similar to the findings of the theoretical part (section 3.5). In the theory (section 3.5.1) we find out that there are five categories of data warehouse staff and seven areas of education while Telenor has divided the staff and the education into three levels each. The reason for this division being the company size, employee expertise in computers and financial constraints due to the fact that company has recently started its business and is trying to establish itself in the market right now.

As a result of the communication and training program, a lot of business users are using the data warehouse for information retrieval and they have found it very interesting and important. This program has enabled them to use the data warehouse and take output from it, which without training would have been impossible.

5.2 Help Desk & Problem Management

Telenor has not implemented any help desk as was found in the theoretical part (section 3.6). One of the reasons is that currently the data warehouse is not very huge one, and only 50 employees are using it. As more and more users start using it and the warehouse size will increase they will setup a help desk. The data warehouse Project Manager at Telenor has assigned the responsibility for providing help desk services to the members of the backup & support team. Members of the support team are trained at advanced level and can solve any general query regarding the front end tools, data collection and report generation etc. If they can't solve any problem, it is reported to the Technical Team Lead, who decides how to act in the situation.

The methods used to reach support team members are quite similar to the theoretical findings (section 3.6.2). They can be contacted using direct phone, email, or in person. The users can also record their responses regarding data warehouse through the local web based portal of the company. They support team is functioning in a similar way found in theory (section 3.6.2), so as to help the business users, provide 24X7 support, record and solve any problem etc. They are also logging the problems and their solutions for future references. The most common and frequent problems are reported to the Project Manager for enhancements to the data warehouse.

In the theory (section 3.7) it is said that there should be some problem management team that should solve any problem in a combined way by helping each other. Same is the case in Telenor where there is problem management team comprising of Project Manager, Technical Team Lead and the Support Team Lead.

Regarding role of help desk, it was found in the theoretical part (section 3.6) that help desk solves a number of issues including, security and sign in, access levels for data, data quality management problems etc. The backup and support team in Telenor is performing a similar function as well by solving all issues related to user logging on to the warehouse, data access to sensitive data, query processing etc.

Also Telenor is using 'train the trainer' approach as was said in the theory (section 3.6.1), by giving advanced training to its employees who can in return help other colleagues around them.

5.3 Network Management

As the major business of Telenor is telecommunications, it hasn't found any problems in managing the network. In the theory (section 3.8) we found out that if the data warehouse is composed of components from different platforms, network management will be a difficult part but at Telenor, as far as data warehouse is concerned all the hardware and software are from the same vendor so network management is not a difficult task but the operational systems are not from the same vendor, therefore Telenor has implemented a state of the art latest high speed network to connect the data warehouse with the source systems and the users. This high speed link greatly reduces the time required for data loading from the source systems into the data warehouse. Secondly users get quick response from the data warehouse for their queries.

In the theoretical part (section 3.8) it was found that there should be some person having good knowledge about networks and technology in the organization. At Telenor there is a team of network engineers that is responsible for managing the whole LAN of Telenor head office. This team also takes care of the data warehouse link. The team is using state of the art network monitoring and management tools as was found in theory (section 3.8). Members of the network support team are qualified engineers and they can operate the tools to full extent for network monitoring.

5.4 Software & Hardware Issues

Telenor has adopted the policy to keep its systems up to date, whether it be data warehouse, whether it be the network, whether it be the hardware or any thing else. Telenor always tries to keep itself inline with the latest technology in confirmation with the findings of the theoretical part (section 3.9).

Regarding hardware issues Telenor has an agreement with the vendor who is responsible for all the hardware updates. There is a growing trend in organizations where a deal is struck for hardware or software certification with the vendors. Organizations are no longer keeping the hardware or software certification processes within the organization saving them from recruiting employees and spending money.

Telenor has signed an agreement with its data warehouse vendor to provide necessary updates for its software as well. In case there is any problem in the system it is reported to the vendor. The vendor after checking the problem area provides its solution.

5.5 Extract, Transform & Load (ETL)

ETL is one of the most important maintenance functions related to data warehouse maintenance. ETL is the process which determines what data should be extracted from the source systems, how it should be transformed into a format that is acceptable for a data warehouse, cleaning the data that is not useful and finally the loading of data in the data warehouse.

In the theoretical part (section 3.10.1) it was found that there should an ETL architect in the organization, responsible for ensuring strength, endurance and reliability in an ETL process. The ETL architect should understand all the technicalities of the operational systems and the data warehouse. Same is the case in Telenor. Although the title for ETL architect in Telenor is ETL Team Lead, but the responsibilities and duties are the same. ETL Team Lead is responsible for creating a smooth and flawless ETL architecture to ensure that various ETL functions perform as one unit.

ETL Team Lead at Telenor is fulfilling all the duties of the ETL architect described in the literature with the help of his support staff including identifying data to be extracted, writing code for data extraction and transmission, storing data in a temporary database, transforming data, cleaning data, and loading data in the warehouse. Although a lot of ETL capabilities are provided by the data warehouse but still there is a need for ETL architect to carry out these very sensitive tasks.

5.6 View Maintenance

From the interview with Project Manager (section 4.7) at Telenor we found out that there is nothing much a user can do about view maintenance. Most of the functionalities related to view maintenance are provided in the system by the vendor. Users just need to have knowledge about when to use those functionalities and when to create and refresh views. In a data warehousing environment users queries need to be very efficiently and carefully written as some tables of the data warehouse are very huge and queries posted against these tables could days or weeks to complete.

During the communication & training process users are given advice on how to handle issues related to views. Users should not access base tables directly; instead there should be views for those tables, which needed to be accessed. In this way users can get quick response to their queries from the data warehouse. Views are commonly built on tables that are accessed frequently and have large data sets.

The commonly used strategy for view maintenance by the industry is incremental where a view is updated incrementally instead of refreshing the view from scratch and re-running the view query.

The use of view maintenance policy depends on the type of data and its effectiveness on the overall query result.

CHAPTER 6: CONCLUSION AND DISCUSSIONS

In this chapter we have presented our findings from the thesis. Keeping an eye on all the previous work here we will present with our conclusions. Additionally we will present some areas where future research could be done in the area of data warehouse maintenance.

6.1 Conclusion and Discussion

Data warehousing is the leading and most reliable technology used today by companies for planning, forecasting, and management for e.g. resource planning, financial forecasting and control etc. After the evolution of the concept of data warehousing during the early 90's it was thought that this technology will grow at a very rapid pace but unfortunately it's not the reality. A lot has been done in this field regarding design and development of data warehouses and a lot still needs to be done but one area which needs special attention from research community is data warehouse maintenance.

A major reason for data warehouse project failures is poor maintenance. Without proper maintenance desired results are nearly impossible to attain from a data warehouse. Unlike operational systems data warehouses need a lot more maintenance and a support team of qualified professionals is needed to take care of the issues that arise after its deployment including data extraction, data loading, network management, training and communication, query management and some other related tasks. To carry out all these functions and processes a qualified team of full time skilled professionals is required who can efficiently and constantly take care of all the data warehouse maintenance issues in a timely manner. Keeping in view the data warehouse maintenance requirements and performance requirements we proposed our research question as:

1. How can a data warehouse be successfully managed from performance perspective after its deployment?

During our empirical study we have validated six major areas which have to be managed in a timely manner for a better and improved performance. All these concepts are closely related to the theoretical findings and the empirical findings with very little differences. These are:

1. Communication and Training
2. Help Desk and Problem Management
3. Network Management
4. Software and Hardware Issues
5. Extract, Transform and Load Process (ETL)
6. Query Management

Our case study shows that the first and the most important part of a data warehouse maintenance program is the training of its users. The case study shows that most business users are reluctant to adopt technology to carry out their work, therefore pursuing a business user to use data warehouse is inevitable. To pursue business users in using data warehouse the communication and training program is a must. The training program gives the users of data warehouse an insight into the qualities and capabilities of a data warehouse and teaches them the methods to benefit from it. Often the data warehouse projects fail because the users don't know how to use it according to the business needs. No one is going to use the data warehouse until they know how to use it, especially the business users who are more comfortable in receiving reports in a paper form instead of using computers for this purpose. The communication process also continues along with the training program. The communication process keeps the business users and IT users in contact with each other to have exchange of views, suggestions and any guidance towards enhanced performance of a data warehouse.

We further concluded from the case study that the education program should be implemented according to the capabilities and skills of the users. Different levels of training should be carried out for different groups of users. The business users can be trained in-house provided qualified professionals are present in the data warehouse support team which is a must while advanced technical training for support team can be carried out by the vendor or some outside party. The advanced training includes lectures on logical data modeling and ETL functions. We found out that the communication program doesn't needs any formal implementations. Data warehouse support team members can continue the communication program by keeping themselves in close contact with the business users and exchanging views on the present and proposed performance of data warehouse etc. Some informal parties or get-togethers could be arranged to bridge the gap between business community and the IT community. Information about the data warehouse in the form of documents, emails or presentations should be regularly sent to the data warehouse users.

Our case study showed that services of help desk and problem management play an important role in taking valuable output from the data warehouse. Support is always required in any information system, same is the case with a data warehouse but here the support is needed 24 hours a day. Some of the processes like ETL are carried out during the night, which require presence of support staff to rectify any problem. There are mechanisms implemented where the support staff is notified of any problem by the help of a SMS to his/her mobile phone. The support staff after receiving the message takes necessary steps to rectify the problem. Support staff can also be contacted by data warehouse users if they had some problem or difficulty in using the system. There should be some sort of formal or informal help desk created solely for this purpose depending on the size of organization and the data warehouse. The help desk personnel should be available at all times through any means such as by phone, by email or in person. The support team logs all the problems found and reported by users for later references and user community can also benefit from previous experiences. Users can contact support team in any problem whether related to security and sign in, data access to detailed or aggregated data, query processing, report generation, front end tools usage etc. The support team also points out if there is any loop hole or problem area within the data warehouse that should be addressed. Apart from help desk each data warehouse support team develops its own problem management process. The process defines necessary routines and instructions to counter any problem found in the warehouse. If the problems found in the data warehouse are not addressed at the right time, this leads to performance shortfalls, and usability and

availability issues in near future. Thus help desk and problem management play a key role in improving data warehouse performance and getting the desired out put from it.

Network management also plays its part in improving data warehouse performance. From the case study we concluded that by having a fast and reliable network user queries get a much shorter response time especially in a distributed data warehouse. There should be some specialized person(s) responsible for managing the network in the organization. As more and more users start using the data warehouse the load on the network also increases and response time becomes longer and longer. The network monitoring personnel(s) can use some specialized tools for monitoring network performance. The features these tools should provide are listed in the theoretical part (section 3.8). Another approach for network management could be using the services of some specialized company in this field. Such firms can apply network planning, design, implementation, management and monitoring services either remotely or on site.

Future planning for hardware and software resources for the data warehouse is compulsory for taking maximum output from it. In the beginning as the data warehouse usage is less, fewer resources are required, but as time passes and data warehouse starts becoming popular within the users, more hardware and software resources are required for a smooth running. It is the responsibility of the data warehouse support team to keep an eye on the data warehouse trends within the organization and develop a strategy for calculating the required hardware and software resources for the future.

Having up to date versions of software and hardware is another issue that the data warehouse support team needs to handle. For this purpose the companies can sign agreements with vendors or they could implement it in house if they have the required competence and resources within the organization. Software updates are usually provided by the vendors. The data warehouse support team should be in contact with the vendors and needed to provide feedback on data warehouse performance to them. Hardware updates are inevitable when the existing hardware can't withstand the computing load and the response time increases. In such cases one needs to install/update new hardware. The updating could be in the form of ram upgrading or disk upgrading, alternatively new and latest hardware could be installed.

The case study showed that ETL functions needed to be carried out by a competent and trained ETL team. The ETL team is headed by an ETL expert. It's the responsibility of ETL architect to devise a comprehensive and effective ETL process to load the data warehouse. The ETL architect/expert ensures that the ETL processes have strength and endurance. The ETL architect works in close coordination with the business users and identifies which data and at what level of detail is required. The ETL architect has a clear understanding of the company's business and knows what type of data is exactly required. The data warehouse will not address reporting requirements, until and unless it has the data that is useful. If the data is of no use for the business, there is no need storing it in the warehouse. Thus ETL architect has this very sensitive responsibility of selecting the right and useful data for loading into the data warehouse. After identifying data that needs to be extracted from the operational systems ETL team transforms it into a format acceptable for the data warehouse and cleanses it as well by eliminating useless data. The data is finally loaded into the warehouse. During all these tasks it is the responsibility of ETL team to make sure that all the operational systems and the data warehouse are available to users.

View materialization is a strategy used to provide fast answers to user queries. But it is important to have updated views whenever the base tables upon which views are built are updated. It is the responsibility of data warehouse support team to devise a flexible and optimal strategy for maintenance of materialized views. Although it is more of a design decision to provide capabilities for view maintenance but it is the user's decision to decide which views to materialize and when to refresh them.

6.2 Further Research

During the time spent on writing this thesis we have identified a lot of other areas within the data warehousing field that need some attention from the research community.

A lot of companies are now thinking of outsourcing their data warehouse maintenance because of high costs. This is a relatively new area and further research could be done on the pros and cons of offshore data warehouse maintenance.

Companies are still finding it difficult to find an efficient view maintenance mechanism. ETL functions of extract, transform and load can also be considered for further research.

It is often found that personnel in an organization are reluctant to share their data with others. This concept is known as data ownership. This area can also be considered for further research. Another area of research is data warehouse politics where certain users consider them more powerful than other because of having access to more and confidential data.

REFERENCES

AASY97: Efficient View Maintenance at Data Warehouses. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek.

AH96: Data Warehouse Performance Management Techniques. Andrew Holdworth of Oracle Corporation. 1996

AI99: Maintenance of Materialized Views: Problems, Techniques, and Applications, Ashish Gupta and Inderpal Singh Mumick. ACM digital library

AN06: Information from Answers.com, http://www.answers.com/capacity%20planning

BL02: Optimization strategies for data warehouse maintenance in distributed environments. Master's thesis by Bin Liu of Worcester Polytechnic Institute

BS97: Data warehousing, data mining & olap authors: Alex Berson and Stephen J. Smith Publisher: Mcgraw-Hill

BSE02: A Transactional Approach to Parallel Data Warehouse Maintenance by Bin Liu, Songting Chen, and Elke A. Rundensteiner. Worcester Polytechnicque institute.2002

CG04: The evolving data warehouse market: Part1. Charlie Garry copyright 2004 Meta Delta.

CG04(2): The evolving data warehouse market: Part2. Charlie Garry copyright 2004 Meta Delta.

CI99: The corporate information factory. Claudia Imhoff. 1999 DMReview magazine

CM04: The computer world magazine
http://www.computerworld.com/databasetopics/businessintelligence/datawarehouse/story/0,10801,89534,00.html

CT03: The ETL in a box. Claudia Imhoff and Tom Kerr. 2003 DMReview Magazine

EA99: System Analysis and Design. 2nd Edition. 1999. Elias M. Awad

EN04: Fundamentals of database systems. 4th Edition. Persons international and Addison Wesley. Ramez Elmasri and Shamkant B. Navathe

HGB01: A Benchmark Comparison of Maintenance Policies in a Data Warehouse Environment, Henrik Engström, Gionata Gelati, Brian Lings.

IB06: Information from IBM website
http://publib.boulder.ibm.com/infocenter/rb63help/index.jsp?topic=/com.ibm.redbrick.doc6.3/perf/perf11.htm

JJ95: Lessons from a successful data warehouse implementation. Dr. John. D Porter and john. J Rome. Arizona State university.

JKQ97: Algorithms for Materialized View Design in Data Warehousing Environment, Jian Yang, Kamalakar Karlapalem. Qing Li. Proceedings of the 23rd VLDB conference, Athens, Greece, 1997

JT97: Data warehouse, Practical advice from the experts. 1997. Prentice hall by Joyce Bischoff & Ted Alexander

KJM01: Efficient incremental view maintenance in data warehouses. Ki Yong Lee, Jin Hyun Son, Myoung Ho Kim. Korea Advanced Institute of Science and Technology

KO00: Data Warehousing Technology. Ken Orr. A white paper. Revised edition, 2000 by Ken Orr Institute.

LG05: Data Warehouse Information Center. www.dwinfocenter.org/gotchas.html Copyright 2005 Larry Greenfield

MH94: Qualitative data analysis: A sourcebook of new methods. Miles M.M. & Huberman A. M.1994, Sage publications

MJ96: Research Design Explained 3rd edition. Mark Mitchell and Janina Jolley (1996).

MR05: A definition of data warehousing by Michael Read of Technology Evaluation. http://www.intranetjournal.com/features/datawarehousing.html

NCR06: NCR customers success stories, http://www.teradata.com/t/go.aspx/?id=92886

RD02: Case Study Research: Design and Methods by Robert K Yin, Donald T Campbell. Sage Publications

RH97: Building, using, and managing the data warehouse. Ramon Barquin, George Zagelow, Katherine hammer, Mark sweiger, George Burch, Dennis Berg, Christopher Heagele, Katherine Glassey-Edholm, David Menninger, Paul Barth, J.D. Welch, Narsim ganti, Herb Edelstein, Bernard Boar, Robert Small. Data warehousing institute series from Prentice Hall.

RJ96: Research Methodology: A step by step approach for beginners. Ranjit Kumar. 1996 Sage Publications Inc.

RM02: The data warehouse toolkit. 2nd edition. Ralph Kimball, Margy Ross. 2002 Wiley computer publishing.

RK96: The Data Warehouse Toolkit by Ralph Kimball. 1996 John Wiley and Sons.

RK00: Data Warehouse Management Handbook by Richard Kachur. 2000 Prentice Hall

RM04: Simple strategies to improve data warehouse performance. Masters thesis by Reena Mathews of North Carolina State University, 2004.

SJ96: Incremental Maintenance of externally materialized views. M. Staudt, M. Jarke. In Proc. of the 22nd VLDB Conference, Mumbai, India, 1996.

SL00: Risks in Data Warehouse Project Management, Sid Adelman and Larissa Moss, Addison Wesley Longman, 2000

SMVY99: Recent advances and research problems in data warehousing. Sunil Samtani, Mukesh Mohania, Vijay Kumar, and Yahiko Kambayashi

SR05: Eleven steps to success in data warehousing. White paper by Sanjay Raizada of Syntel corporation. 2005

SU96: An Overview of Data Warehousing and OLAP Technology by Surajit Chaudhry and Umeshwar Dayal

TC05: Database Systems, a Practical Approach to Design, Implementation and Management by Thomas Connely and Carolyn Begg. 4th edition, 2005. Addison Wesley

TD04: The concise technical dictionary
http://www.thetechdictionary.com/term/etl_%28data_integration%29

VP96: Building a data warehouse for decision support. 1996 Prentice Hall. By Vidette Poe with contributions from Laura L. Reeves.

VGH01: Essentials of System Analysis & Design. 2001 Prentice Hall. By Joseph S. Valacich, Joey F. George, Jeffrey A. Hoffer.

WI06: Wikipedia, the web's free encyclopedia http://en.wikipedia.org/wiki/Data_warehouse

WR94: Using the Data Warehouse by W.H. Inmon and R.D. Hackathorn. 1994 John Wiley and Sons.

APPENDIX A

Interview Questions

1. Give an overview of your organization's business?
2. How many employees are there in your organization?
3. How your organization is managing its business by the help of IT, means which operational systems are you using and in which field. (Inventory, orders, customer relations etc)?
4. How long your organization has been using data warehouse?
5. Is it a centralized data warehouse or distributed data warehouse?
6. How many user access the data warehouse?
7. Are the operational systems feeding the data warehouse located in the same place or in different places(cities or offices)?
8. Which data you are storing in the data warehouse and from which operational systems?
9. How data warehouse is helping you in decision support?
10. How many members you have in the data warehouse team (give role/s of each person as well)?
11. How you have trained your business users to use data warehouse?
12. How you have applied the communication process between the data warehouse users in the organization? Give some detail.
13. How you have carried the training program? (Give schedule in some detail)
14. Are there any helpdesk services for the data warehouse? If so what are their functions and how it is helping data warehouse users?
15. Is there any problem management team specifically formed for data warehouse. If so what is its role in solving the problems?
16. How the data warehouse is affecting your network? (include network attributes)
17. How you are managing the network?
18. How are you managing any software and hardware problems?

19. Are there any software and hardware certification processes in your organization?

20. Who is carrying out the ETL function?

21. What problems you have faced in ETL function and how you have managed them? (Briefly)

22. Is there any ETL architect. If so what are his job responsibilities?

23. How are you managing the view maintenance problem?

24. Which strategies are you using for view synchronization and why?