CHAPTER - 1
INTRODUCTION

1.1 OVERVIEW

Data warehousing encompasses architectures, algorithms, and tools for bringing together selected data from multiple databases or other information sources into a single repository, called a data warehouse suitable for direct querying or analysis. Data warehousing is the important and most reliable technology used today by companies for development, forecasting, and management. After the evolution of the concept of data warehousing during last two decade it was thought that this technology will grow up at a very rapid speed but unfortunately it’s not the truth.

Many research works has been done in this pasture regarding design and development of data warehouses and many still needs to be done but one area which needs special attention from research community is data warehouse maintenance. In this research, I have made an effort to take out the present available maintenance technique used by the corporatist to enhance the data warehouse performance.

Finally I conclude that without proper technique of searching records, expected result are unfeasible to achieve from a data warehouse. Unlike operational systems data warehouses need a lot more maintenance and a help and support team of qualified expert is needed to take care of the issues that arise after its implementation including communication, coordination, data cleansing , data loading, education, training and materialized view and some other related tasks.

In the present circumstances business conditions is altering then the management of the organization require to use new and best information for making timely decisions and respond to changing business conditions. Many companies are now a day have changed their business focus towards customer orientation to remain competitive using information technology as the backbone of their operations but the fact is that in spite of having a great number of influential desktop and notebook PC and a speedy and consistent network, access to information that is previously available within the organization is very hard or otherwise not possible.

All companies use Information Technology for the operations to create large amount of data about their business but in many cases this data remains in the operational systems and can’t be used by the organization. Experts state that only a small portion of this data that is entered, processed and stored is actually accurate data. Data that is entered, processed and stored is actually available to decision makers and management of the enterprise.

The mixture of correct and incorrect data can cause wrong decision and significant reduction in sales and profits of companies and vice versa. Because the primary goal of data is use the information so that decision makers and business analyst can make queries, analysis and planning regardless of the data changes in operational database.

It is accepted that right data is a very powerful feature that can provide significant benefits to any organization. The most fundamental aspect in a particular organization is the critical decision making capacity of the management, which influence the successful running of business operations. For such decisions, the information must be reliable, accurate, real-time and easy-to-access. For such information, all the enterprise-related data should be appropriately analyzed from a multi-dimensional point of view and presented at one place.

In the 1990's companies start to require accurate data that can store and retrieve on time, they found that traditional information Technology systems was simply too time-consuming and difficult to provide such facility. These work take a lot of time using traditional applications 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 important data could be held for carrying out strategic reports for companies 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.

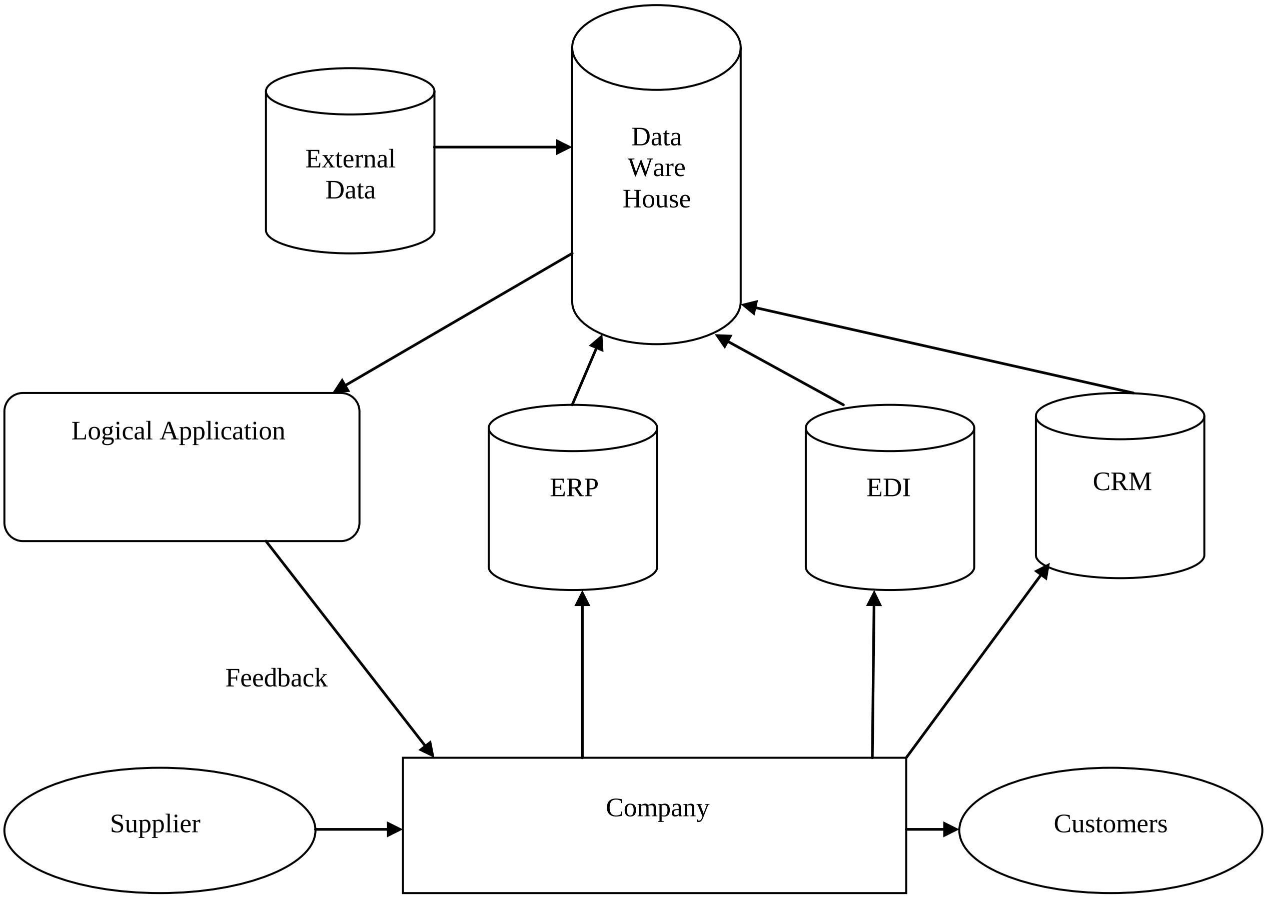


Fig 1.1: Data Warehouse in a Company

In the latter half of the 20th century, there survive a large number

nd types of databases .Many large businesses found themselves with data spread 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 (Spreadsheet, SQL Server, Oracle, DB2 Databases, IDMS Records and VSMS Files) and place them in a common location that uses common querying tool.

In this way operational databases could be held on whatever system was most capable for the operational business while the reporting information could be held in a common location using a common language.

One aspect of a data warehouse that should be worried is that it is NOT a location for ALL of a business's data, but rather a location for data that is 'interesting' and 'important'. Data that is interesting assist decision creator in making considered decisions relative to the organization's overall mission. Its major role remains fundamental in understanding, planning and delivering knowledge to the enterprise in a timely and cost effective fashion.

Another main concept that has come out of the data warehousing concept is the identification the two different types of information systems in all companies namely operational systems and information systems. Operational systems are used to achieve the day to day process of the companies. They function as a backbone to any project. In fact most of the companies around couldn't operate without these operational systems. On the other hand there are other functions within the companies which have to work with planning, forecasting and management. These functions are quite different from operational functions. 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.

Data warehouses are large, special-purpose databases that contain data integrated from a number of independent sources, supporting clients who wish to analyze the data for trends and anomalies. The process of analysis is usually performed with queries that aggregate, filter, and group the data in a variety of ways. Because the queries are often complex and the warehouse database is often very large, processing the queries quickly is a critical issue in the data warehousing environment.[1,2,11,12,14]

Data ware house are computer based data information system that optimized database query and reporting tool because of their ability to analyze data often from disparate database and in interesting ways. They are a way for managers and decision makers to extract information quickly and easily in order to answer question about their business.[3,6,8]

1.2 Market trend in recent years in corporate warehousing

The idea of data warehousing was in the companies since the early 1980's but during the early 90's all universal company have obtained some form of data warehousing technology and using it in some form for decision support . During the early phase of data warehouse development many company professionals were belief that this technology expand at a very rapid pace but the reality is not the same. [87] 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. [21]

Enterprise Data Warehouse is a centralized warehouse which provides service for the entire enterprise. It is a specialized data warehouse which may have several interpretations. [112,114] Companies considering investing in an EDW solution have matured to a level where data marts can no longer assure the companies increased requirement for superior quality and more timely business analytics[104,106].

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.[122]

In 2008 well-known corporate companies consider that these services (pressure to conduct these) drive double digit growth of the EDW market. According to my survey, in the end of 1990, almost universal 2500 companies had finished the major task of implementing an Enterprises Resource Planning software infrastructure. [129] Internet has totally changed the business scenario and most of the companies have changed their existing transactional infrastructure rapidly into the web model. 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 $3.9Billion market currently, and which grow to an estimated $ 10 Billion by 2008.

Through 2006, Meta Group expect to see vendors increase their sales, marketing, and development focus on this market as the transaction Missing in all this was an association’s ability to generate meaning from all the transactional and sub transactional (e.g., usage, traffic levels) data being detain.

Indeed, META Group research indicates that 77% of organizations plan to detain even more detailed business data in 2009 than was detaining in 2008.

There are a number of companies rival in the data warehouse industry but still not a single purveyor can address all the requirements of one customer. The data warehouse market is evolving endlessly and swiftly. [133] Each purveyor that joins the battle is hoping to address the apprehension of at slightest a slice of what is estimated to be a $5Billion market at this time, and which grow to an probable $ 9.9Billion by 2008. Through 2006, Meta Group expects to see purveyor boost their sales, marketing, and development focus on this market as the transaction processing market go back in stress. [135]

This have mean larger services association for some purveyor, while others consider coagulating or growing relationships with third party value added resellers. Some joining of end user business intelligence tools (e.g., Business Objects), extract/transform/load purveyors , or boutique services firms that specialize in data warehousing is also likely to be seen through 2007.[84]

For bigger associations, this mean investing in data mart join projects. For companies of nearly any size, it means increased development and planning to achieve investigative ripeness. at this time there are a lot of corporation working in the data warehouse sector among which Tera data , IBM and Oracle, Microsoft are the most famous with their specialized data warehousing products.

1.3 Data Warehouse

Data warehouse are actually build new system environment. This new environment is kept separate from the system environment helping in the operations. The primary goal of data warehouse is use the information so that decision makers and business analyst can make queries, analysis and planning regardless of the data changes in operational database.[68,73,71] The data ware house essentially holds the business intelligence for the enterprise to enable strategic decision making.[75]

A data ware house is designed for analysis and query rather than for a transaction processing. It separate analysis from transaction and permit an organization to combine data from several sources. A data warehouse is the copy of transaction data especially structured for querying, reporting and analysis purpose. The data warehouse contains copy of transaction which cannot be updated or altered by the transaction system.[59,62,64]

Data warehouse is the source of stable and integrated data designed to support decision makers and business analysts. [43] Data are acquired from various operational data stores across the enterprise. When a database management system (DBMS) parses a query it decides the best strategy to execute it based on statistics it retains about DB structure, indexes, and number of distinct values.[44]

A data warehouse is a system that retrieves and consolidates data periodically from the source systems into a dimensional or normalized data store. It usually keeps years of history and is queried for business intelligence or other analytical activities. It is typically updated in batches, not every time a transaction happens in the source system. [132]

According to Bill Inmon,

“a data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management’s decisions”.

According to Ralph Kimball, [71]

“A data warehouse is a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making.”

Both of them agree that a data warehouse integrates data from various operational source systems. In Inmon’s approach, the data warehouse is physically implemented as a normalized data store. In Kimball’s approach, the data warehouse is physically implemented in a dimensional data store.[132,171]

Another interesting definition is from Alan Simon:[72]

“The coordinated, architected, and periodic copying of data from various sources into an environment optimized for analytical and informational processing.”

According to Bill Inmon a data warehouse is a subject, integrated , time variant and nonvolatile 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. Data Warehouse can provide more data/information than is available from the OLTP system by incorporating additional business rules into the EXTRACT, TRANSFORM, AND LOAD processes or Semantic Layer.

“A well designed DW can provide much more information to users than can be found in the source system”

Data warehouse is the design and implementation of process, tools to manage and deliver entire, timely accurate and logical information for decision making. It is accepted that information is a very powerful feature that can provide significant benefits to any organization.[1] Data ware house are computer based data information system that optimized database query and reporting tool because of their ability to analyze data often from disparate database and in interesting ways.

They are a way for managers and decision makers to extract information quickly and easily in order to answer question about their business. [2]Data warehouse are actually build new system environment. This new environment is kept separate from the system environment helping in the operations. The primary goal of data warehouse is use the information so that decision makers and business analyst can make queries, analysis and planning regardless of the data changes in operational database.[21,24] The data ware house essentially holds the business intelligence for the enterprise to enable strategic decision making.

1.4 Data Warehouse Characteristics

1.4.1 Subject-Oriented Data

A data ware house is organized around major subjects, such as customer, supplier, product and sales. Rather than concentrating on the day today operations and transaction processing of an organization, a data ware house focuses on the modeling and analysis of data for decision makers[29,30]

1.4.2 Integrated Data

A data warehouse is constructed by integrating data from varied, heterogeneous databases and information systems such as relational databases and flat files. As an example the representation of 'male' and 'female' could be coded by 'M' and 'F', '0' and '1', or by 'true' and 'false'. Frequently inconsistencies are more complex and subtle, but by definition, data in a data warehouse is always maintained in a consistent fashion.[33,37]

1.4.3 Time Variant Data

As opposed to operational databases historical data is of high importance in the DWH world. Data might be available in daily, weekly, monthly, quarterly and/or yearly aggregates. Therefore time variance calls for storage of multiple copies of the underlying detail of differing periodicity and/or time frames. The time variant strategy is essential, not only for performance but also for maintaining the consistency of reported summaries across organizational units and over time.[31,32]

1.4.4 Non volatile Data

Non volatility, the final primary aspect of data warehouses, means that after data are loaded into the warehouse, changes, inserts, or deletes are no longer performed. The data warehouse is subsequently reloaded or, more likely, appended on a periodic basis [usually nightly, weekly, or monthly] with new, transformed or summarized data. Apart from this loading process, the information contained in the data warehouse remains static. [38,40,41]

This is required to represent data as of any point in time. If a data row were updated, information would be destroyed. “Maintaining ‘institutional memory’” [84] is one of the most important features of DWHs, besides the property of non volatility also permits a data warehouse to be heavily optimized for query processing. Besides these four primary principals that went into Inmon’s definition of a data warehouse he laid out several others.

“Some of these principles were initially controversial but are commonly accepted now. Others are still in dispute or have fallen into disfavour”

By Michael Haisten[84].

However besides these principles there is also a point raised concerning the functionality of a DWH directly in Inmon’s definition, the support of management decisions.

Data warehouse can be a treasure of information for companies, a place where employees can access data as well as the relevant, accurate, and timely information about consumers, products, and market and technology developments. It is major challenge for companies to creating and maintaining a valuable data warehouse.[87,88] If there’s too much data in the warehouse but, if the data is not kept relevant, accurate, and, if the tools used to access the data warehouse are not well incorporated, the value of the warehouse is inadequate.

1.5 The Architecture of a data warehouse

A data warehouse system has two main architectures: the data flow architecture and the system architecture. The data flow architecture is about how the data stores are arranged within a data warehouse and how the data flows from the source systems to the users through these data stores.

The system architecture is about the physical configuration of the servers, network, software, storage, and clients. In data warehousing, the data flow architecture is a configuration of data stores within a data warehouse system, along with the arrangement of how the data flows from the source systems through these data stores to the applications used by the end users [147,150]. This includes how the data flows are controlled, logged, and monitored, as well as the mechanism to ensure the quality of the data in the data stores.

The data flow architecture is different from data architecture. Data architecture is about how the data is arranged in each data store and how a data store is designed to reflect the business processes. The activity to produce data architecture is known as data modeling. Data stores are important components of data flow architecture.[126] A data store is one or more databases or files containing data warehouse data, arranged in a particular format and involved in data warehouse processes. Based on the user accessibility, it can classify data warehouse data stores into three types:

- A user-facing data store is a data store that is available to end users and is queried by the end users and end-user applications.

- An internal data store is a data store that is used internally by data warehouse components for the purpose of integrating, cleansing, logging, and preparing data and it is not open for query by the end users and end-user applications.
- A hybrid data store is used for both internal data warehouse mechanisms and for query by the end users and end-user applications.

A master data store is a user-facing or hybrid data store containing a complete set of data in a data warehouse, including all versions and all historical data. Based on the data format, it can classify data warehouse data stores into four types:

- A stage is an internal data store used for transforming and preparing the data obtained from the source systems, before the data is loaded to other data stores in a data warehouse.
- A normalized data store (NDS) is an internal master data store in the form of one or more normalized relational databases for the purpose of integrating data from various source systems captured in a stage, before the data is loaded to a user-facing data store.
- An operational data store (ODS) is a hybrid data store in the form of one or more normalized relational databases, containing the transaction data and the most recent version of master data, for the purpose of supporting operational applications.
- A dimensional data store is a user-facing data store, in the form of one or more relational databases, where the data is arranged in dimensional format for the purpose of supporting analytical queries.

Operational data is the data live in the operational systems of the company. This is the data required for the day to day operations of the company. An operational data store is a repository of current and integrated operational data used for analysis [37]. Its main purpose is to collect data from the operational systems and transfer that data into the warehouse.

Load manager has to manage and perform the entire task required for the extraction and loading of data into the data warehouse [37]. The entire task performed by the load manager may include alteration of data to prepare the data for entry into the data warehouse. The size and complexity of this component vary between data warehouses and may be constructed using a combination of vendor data loading tools and custom built programs.

The warehouse manager executes all the process related with the management of data in the warehouse [37]. A warehouse manager directs, coordinates and plans the warehouse storage and distribution of the products and materials within a company. Others perform these activities for organizations that store products and go ODS for many different organizations. A warehouse manager supervises the activities of employees, engages in shipping, receiving, storing and testing materials and products.

They are responsible for safety development, along with implementing security and safety programs. They issue job assignments and review work orders, invoices and confirm reports and monitor the distribution of materials or products. Warehouse manager also creates query profiles to verify which indexes and aggregations are proper.

A query profile can be create 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. The main works of a query manager is to translate the logical database design into a physically implementable design.

The query manager responsible for The Data Warehouse Database Administrator is responsible for the synchronization, or replication, process. Query manager include queries to the appropriate tables and scheduling the execution of queries. Query manager Assuring transformation rules keep the data meaningful and consistent. He also create query to allow the warehouse manager to determine which indexes and aggregations are suitable. He is responsible party for the quality of the data.

Some applications require the data to be in the form of a multidimensional database (MDB) rather than a relational database. An MDB is a form of database where the data is stored in cells and the position of each cell is defined by a number of variables called dimensions. Each cell represents a business event, and the value of the dimensions shows when and where this event happened. MDB is populated from DIMENSIONAL DATA STORE. Extract, transform, and load (EXTRACT, TRANSFORM, AND LOAD) is a system that has the capability to read the data from one data store, transform the data, and load it into another data store. The data store where the EXTRACT, TRANSFORM, AND LOAD reads the data from is called a source, and the data store that the EXTRACT, TRANSFORM, AND LOAD the data into is called a target. Figure 1-4 shows data flow architecture with four data stores: stage, ODS, DIMENSIONAL DATA STORE, and MDB with the control system, the metadata, and the components of data quality processes.

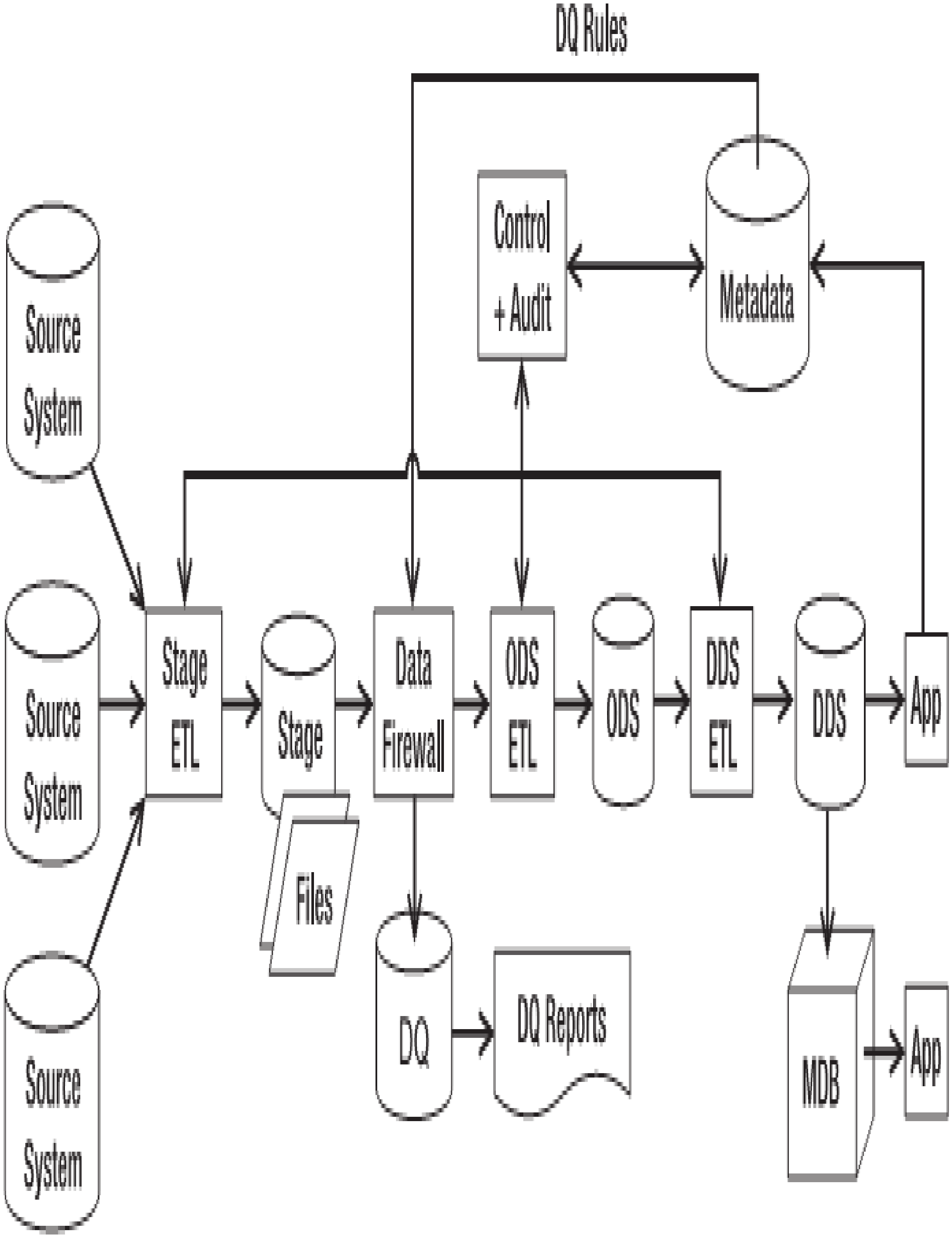


Figure 1.2 – Flow of data in data ware house

The arrows in figure 1-2 show the flows of data. The data flows from the source systems to a stage, to ODS, to dimensional data store, and then to the applications used by the users. There are three extract, transform, and load packages between the four data stores. A stage extract, transform, and load retrieve data from the source systems and load it into a stage. ODS extract, transform, and load retrieves data from a stage and loads it into ODS. Dimensional data store extract, transform, and load retrieves data from ODS and loads it into dimensional data store.

An extract, transform, and load package consists of several extract, transform, and load processes.[127] An extract, transform, and load process is a program that is part of an extract, transform, and load package that retrieves data from one or several sources and populates one target table.

An extract, transform, and load process consists of several steps. A step is a component of an extract, transform, and load process that does a specific task. An example of a step is extracting particular data from a source data store or performing certain data transformations. The extract, transform, and load packages in the data warehouse are managed by a control system, which is a system that manages the time each extract, transform, and load package runs, governs the sequence of execution of processes within an extract, transform, and load package, and provides the capability to restart the extract, transform, and load packages from the point of failure.[92]

The mechanism to log the result of each step of an extract, transform, and load process is called extract, transform, and load audit. Examples of the results logged by extract, transform, and load audits are how many records are transformed or loaded in that step, the time the step started and finished, and the step identifier so you can trace it down when debugging or for auditing purposes.

The description of each extract, transform, and load process is stored in metadata. This includes the source it extracts the data from, the target it loads the data into, the transformation applied, the parent process, and the schedule each extract, transform, and load process is supposed to run. In data warehousing, meta-data is a data store containing the description of the structure, data, and processes within the data warehouse.

This includes the data definitions and mapping, the data structure of each data store, the data structure of the source systems, the descriptions of each extract, transform, and load process, the a data flow architecture is one of the first things you need to decide when building a data warehouse system because the data flow architecture determines what components need to be built and therefore affects the project plan and costs. The data flow architecture shows how the data flows through the data stores within a data warehouse.

The data flow architecture is designed based on the data requirements from the applications, including the data quality requirements. Data warehouse applications require data in different formats. These formats dictate the data stores you need to have. If the applications require dimensional format, then you need to build a dimensional data store. If the applications require a normalized format for operational purposes, then you need to build an ODS.

If the application requires multidimensional format, then you need to build an MDB. Once you determine the data stores you need to build, you can design the extract, transform, and load to populate those data stores. Then you build a data quality mechanism to make sure the data in the data ware house is correct and complete.

The main advantage of this architecture is that you can accommodate existing data warehouses, and therefore the development time is shorter.

The system architecture is about the physical configuration of the servers, network, software, storage, and clients. In data warehousing, the data flow architecture is a configuration of data stores within a data warehouse system, along with the arrangement of how the data flows from the source systems through these data stores to the applications used by the end users.

The three level data ware house architecture allow a clear separation of the conceptual view (materialized view) from the user view and from the storage view layout. A data ware house is able to separate the three views of data is likely to be flexible and adaptable and this flexibility and adaptability is data independence. There are three views in data ware house.

1. User View

The user view is the top view of three level data ware house architecture. In this view, only restricted portion of the data ware house is available to end user or application programmer. As data ware house is a shared resource, so each user has a view of the real world represented in the form that is familiar to him.

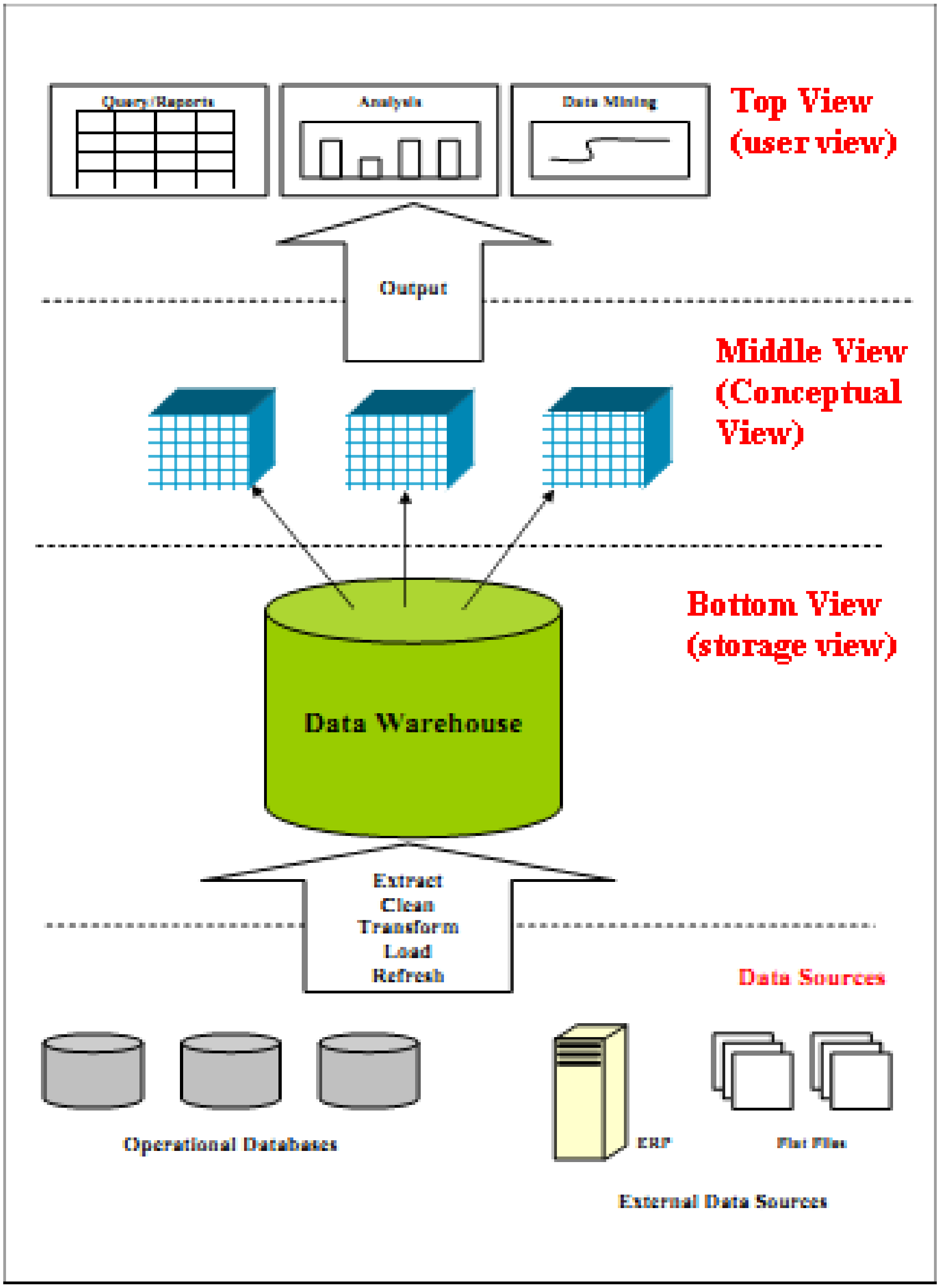


Figure 1.3 – Three level architecture of data ware house

The top view is a user view, which contains query-and-reporting tools, multidimensional analysis tools, statistical analysis tools, and/or data mining tools. Creating separate view of the data base for the different users help is ensuring the data ware house security. When different users have access only to the data that is needed and the access is limited to subset of entire data ware house, the chances of security breakdown are minimized.

2. Conceptual view

   It is materialized view which stores the aggregated and summarized data. It is represented by middle view in three levels architecture. It did not contain any storage level details. It is independent of both hardware and software. The software independent means that the view is not dependent on the data ware house used to implement the database. The hardware independent means that the view does not depend on hardware used in data ware house. Thus change in either hardware and software will not affect the data ware house at conceptual view.

3. Storage View

   It is bottom view of three levels architecture. The storage view is the view of the actual physical storage of data. It defines how data is stored physically in data ware house. It also tell the physical implementation of the data ware house so as to achieve optimal runtime performance and storage space utilization which is required for the efficient data ware house.

1.6 Data Warehouse components

A Data Warehouse consists of six components, these components are; data sources, data extraction and transformation tools, data modeling tools, central repository, target DB, and front-end tools. The following sections provide some details of these components.

1.6.1 Data source

These are the ODS' databases, external, personal or archival data sources. Different data source formats can be used as sources of data, for example VSAM, IMS, RMS, DB2, relational, flat files and other formats [44].

1.6.2 Data extraction and transformation tools

These are used to extract data from the data source files, clean and transform the data, and to ensure that all the relevant data required by the users is available in the DW. The extraction can be done using a standard RDBMS (e.g. ORACLE, SYBASE, INFORMIX, DB2, SQL, and ACCESS).

1.6.3 Data modeling tools

These tools are used to prepare the DW structure from both the data source and the target data warehouse database;

1.6.4 Central repository

This component is used to store the metadata (data about data). Metadata describes the transformation between the source and the target database.

1.6.5 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. In a relational database, the schema defines the tables, the fields in each table, and the relationships between fields and tables. Schemas are generally stored in a data dictionary. Although a schema is defined in text database language, the term is often used to refer to a graphical depiction of the database structure Schemas are often designed with visual modeling tools (Erwin, Rational Rose) that automatically create the SQL code necessary to define the table structures.

1.6.6 View

Views are widely used in decision support applications. A view is a virtual table. A Database View is a subset of the database sorted and displayed in a particular way. A company 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. Originally, in database theory, a view is a read only virtual or logical table composed of the result set of a query. 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.

1.6.7 Views, OLAP and Warehousing

Views are very closely related to OLAP and data warehousing .OLAP queries are typically aggregate queries. Analysts want fast answers to these queries over very large data sets and it is natural to consider pre computing 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 pre computation 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.

A data warehouse is in fact a collection of replicated tables and periodically synchronized views. 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.

Some people consider data warehouses as an extension to database views [41. 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

1.7 Important outcomes of Data Ware house

The following tangible benefits have been reported [20].

1. Product inventory turnover is improved;

2. More cost-effective decision making process by separating the query processing from the operational databases [57];

3. Enhancing asset and liability management by providing the overall picture of the enterprise purchasing and inventory transactions;

4. Supporting the corporate strategy that positions the clients at the center of all operations. This client-centered strategy could not be achieved without a DW [13];

5. To record the past history accurately;

6. Supporting the Reengineering of decisional processes [32];

7. Improving productivity by keeping all the required data in a single location

[45];

8. Reduces redundant processing, support, and software to enhance DSS applications;

9. Enhancing the work process, that also affects the success of business process reengineering.

Generally, a DW is expected to deliver competitive advantage to organizations through increasing the productivity and the effectiveness of decision making within them [15].

1.8 Goals of Data Warehouse

The Data Warehouse contains intelligent data collections which are modeled to support the reporting and analysis needs of DSS users. Consequently, not ALL institutional data is available from the Data Warehouse.). 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.

For those who develop or maintain institutional data systems, who want to make their data available to a wider audience of users, the Data Warehouse offers a central location where staff, faculty and even students can access information for all their reporting needs.[19]

In many cases data ware house assignment are viewed as a stopgap evaluate to get users off to backs or to provide something for nothing .But data ware house require alert management. A data ware house is good investment only if end users actually can get important information quicker and cheaper than they can use modern technology. [22] So management has to take decision for very high level of maintenance. Because previous researchers have focused on methodological, data related, operational, educational and technical issue of data warehouse implementation.

There is require explore studies in which discuss organizational, project-related and environmental dimensions regarding implementation of data warehouse technology in general and in companies' then we came to know that current scenario require the data warehouse to change. Keeping in view this problem we have decided to do some research on management of data warehouses.

It is clear that the data warehouse design and management is extremely critical for the successful functioning of the data warehouse system. There survive much theoretical explanation on how to improve performance of data warehouse but many companies still face problem when working in the data warehouse environment. They face various issues which hamper the smooth functioning. Therefore, there is a strong need to understand the issues affecting the data warehouse efficiency. The goal of this project is to identify these issues and cross verify if these issues are valid in reality by using interview, case studies and online survey.

Goal 1

This research first try to understand what are the performance tuning strategies after implementation of a data warehouse environment.

Goal 2

The project further tries to understand what are the current major problems faced by businesses through case studies and survey.

Goal 3

The topic affecting the data warehouse performance in theoretically and empirical study are discussed.

Goal 4

Finally, a few ways to overcome or minimize these issues are discussed. The overall goal of the project is to understand what the performance tuning strategies of data warehouse design are and how they should be taken into consideration while implementing a data warehouse environment.

1.9 Cardinality

Definition of cardinality in set theory refers to the number of members in the set. On database theory, the cardinality of a table refers to the number of rows contained in a particular table. In terms of OLAP system, cardinality refers to the number of rows in a table. On the other hand, on a data warehousing point of view, cardinality usually refers to the number of distinct values in a column. Generally, there are four levels of cardinalities (as following items); Low, Normal, High and Very high cardinality (also known as Full Cardinality).[66]

Low-cardinality

It refers to columns which have a very few unique values. Low-cardinality column values are typically Boolean values such as gender or a check-box. For instance, the Product table with a column named Active-Bt is a column with low-cardinality. This column contains only 2 distinct values: 1 or 0, denoting whether the product is available. Because there are just 2 possible values in this column, its cardinality level would be called as low-cardinality.

Normal-cardinality

It refers to columns which have sporadic unique values. Examples of such columns with normal cardinality are addresses or product types. For instance, column named Name-Bit in Order table contains the name of the customers. There may be some customers with the general name, such as John, while others have dissimilar names. While there are many possible values in this column, its cardinality level would be called as normal-cardinality.

High-cardinality

It is related to columns which has a large number of distinct values containing very unique values. From the DW point of view, since the grouping of characteristics that are not related in one dimension; high cardinality can be called the number of unique combination of values in a dimension which are very high.

Full-cardinality

It is related to columns which has a very large number of distinct values. Full cardinality values are generally like identification number or e-mail addresses. As an example, in the USER table, an auto-generated number is assigned to each user to uniquely identify them. Recently, full-cardinality is also known as Very High Cardinality in the database community.[94]

1.10 Indexing

In today's scenario, each and every company needs strategical information for the purpose of business analysis to look at the prevailing information crisis. This information needs to be gathered as quickly as possible for the business analysis. Hence, to speedup the query processing in data warehousing environment, indexing is used.[54]

Indexes are basically database objects, which are used to speed up the data retrieval from huge databases. Indexing is an old technique and it exists in relational database from past so many years but it’s not efficient on large number of data records as present in data warehouse, which is used for strategic analysis.[155]

Indexing the data warehouse can reduce the amount of time it takes to see query results. Indexing is the process of creating indexes for record collections. Having indexes allows researchers to more quickly find records for specific individuals; without them, researchers might have to look through hundreds or thousands of records to locate an individual record. Indexing is a way of sorting a number of records on multiple fields.[173] Creating an index on a field in a table creates another data structure which holds the field value, and pointer to the record it relates to. This index structure is then sorted, allowing Binary Searches to be performed on it.

When data is stored on disk based storage devices, it is stored as blocks of data. These blocks are accessed in their entirety, making them the atomic disk access operation.[178] Disk blocks are structured in much the same way as linked lists; both contain a section for data, a pointer to the location of the next node (or block), and both need not be stored contiguously.

Due to the fact that a number of records can only be sorted on one field, we can state that searching on a field that isn’t sorted requires a Linear Search which requires N/2 block accesses, where N is the number of blocks that the table spans. If that field is a non-key field (i.e. doesn’t contain unique entries) then the entire table space must be searched at N block accesses.

Whereas with a sorted field, a Binary Search may be used, this has log2 N block accesses. Also since the data is sorted given a non-key field, the rest of the table doesn’t need to be searched for duplicate values, once a higher value is found. [186]Thus the performance increase is substantial.

To overcome this problem of slow query processing and to improve response time, the concept of indexing came. Data warehouse reports needs quick response. [182] Hence, many indexing techniques have been created to accomplish this target. So indexing enhances the ability to extract data to answer complex and ad hoc queries speedily.

A database index is a data structure that improves the speed of data retrieval operations on a database table at the cost of slower writes and increased storage space. Indexes can be created using one or more columns of a database table, providing the basis for both rapid random lookups and efficient access of ordered records

Characteristics of indexing:

1. Index size should be less.
2. Index should support with indexes of another data ware-house indexes.
3. Index should take less memory for processing.
4. Creation time of index should be least.
5. Indexes should be able to work with joins & ad-hoc queries.

There are existing techniques of indexing for searching the data that is Full Text index, Clustered index and Non Clustered index etc.

Types of indexing

Recently, data warehouse system is becoming more and more important for decision-makers. Most of the queries against a large data warehouse are complex and iterative. The ability to answer these queries efficiently is a critical issue in the data warehouse environment.

Indexing in data warehouse is very much essential these days, as importance of strategic information is increasing day by day. Companies need to analyze their data frequently and this can only be analyzed via querying on giant data warehouse.[106] Querying on such a giant data warehouse, takes a long time to produce results of these queries. To solve this problem of long response, indexing is used.

Indexing basically increases the speed of searching data depending upon different criteria's. Depending upon the size of data warehouse, cost, space requirement, types of queries to be run on etc., indexing technique is chosen.[99] There are different indexing techniques possible, but according to the requirement of company, data warehouse design and expected queries to be run on (i.e. expected outcomes from the data warehouse) indexing technique is being chosen.

Indexing techniques, based upon some specific domains; depending upon which one can choose which technique they can apply to their data warehouse. Indexing type should to chosen such that it is cheap, uses less space, is more efficient in terms of query processing speed and most importantly produces correct results. [197] Depending upon the type of queries to be run onto the data warehouse, indexing technique performance also changes.

If the right index structures are built on columns, the performance of queries, especially ad hoc queries will be greatly enhanced. Although the indexing has been around since the early days of computers, there have been great advances in indexing technology over the years. Indexing technology is the most effective way to reduce the disk I/O required to query, analyze, summarize and retrieve data. Following advanced indexes deliver dramatic performance improvements without major investments in hardware. Different types of indexing techniques explained in this paper are listed below:

- **Full text index**
- **Cluster index**
- **Non Cluster Index**
- **Full Text Index**

It is hard to overstate how much computers have changed the way people approach information. In the days before the Internet and good search engines like Google, looking up information was hard work and people did not bother if they did not have a real need for the information.

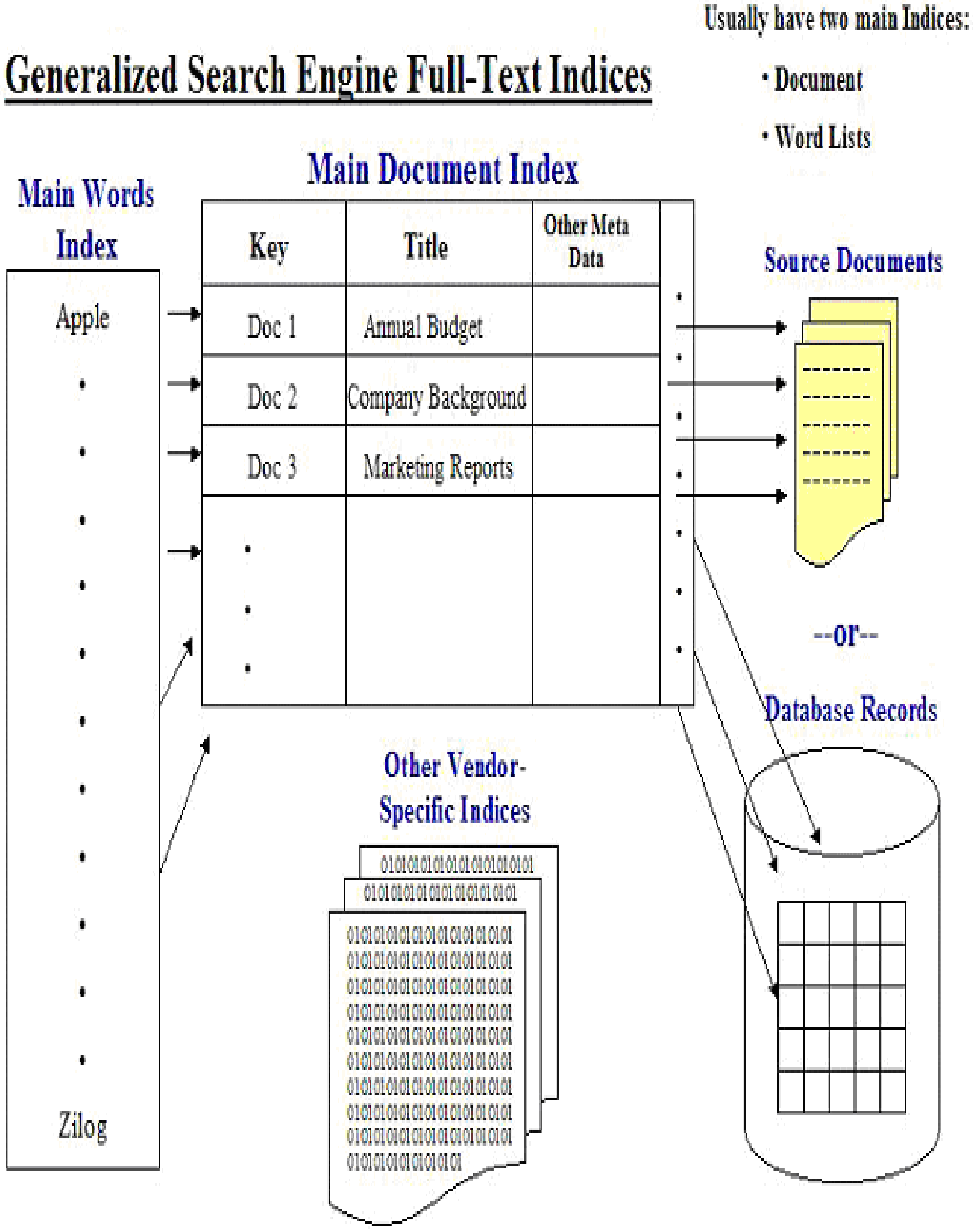
Now, many types of information can be found by typing in a few relevant words to an Internet search engine, which then somehow manages to find the most relevant pages from among the billions of pages available. It does all this in under a second. The above has become such a common part of our daily lives that we are no longer amazed by it.

But searching through the full text of millions of documents, and ranking the results such that the most relevant results are returned first, needs specialized search technology. Such technologies are the subject of this thesis. Before computers, full text search was not possible, so information had to be categorized in various ways so that people could find it.

The information in full-text indexes is used by the Full-Text Engine to compile full-text queries that can quickly search a table for particular words or combinations of words. A full-text index stores information about significant words and their location within one or more columns of a database table. A full-text index is a special type of token-based functional index that is built and maintained by the Full-Text Engine for SQL Server. The process of building a full-text index differs from building other types of indexes.

Instead of constructing a B-tree structure based on a value stored in a particular row, the Full-Text Engine builds an inverted, stacked, compressed index structure based on individual tokens from the text being indexed. In SQL Server 2008, the size of a full-text index is limited only by the available memory resources of the computer on which the instance of SQL Server is running.

Full text indexing provides the feature of full text queries for character based data where there is LOB (large object) kind of columns is present like varchar which gives better efficiency because it gives accurate result due to division of characters strings into tokens and each token searching is fast like searching in documents. [154] When there are fewer documents then full text searching scans tokens in documents called serial scanning. If the large document is present for searching keywords then full text has to follow two steps:



Usually have two main Indices:

- Document
- Word Lists

Generalized Search Engine Full-Text Indices

Main Words Index

Apple

Zilog

Main Document Index

Key	Title	Other Meta Data
Doc 1	Annual Budget	
Doc 2	Company Background	
Doc 3	Marketing Reports	

Source Documents

--or--

Database Records

Other Vendor-Specific Indices

Figure 1.4 : Full text Search engine

a) Indexing

b) Searching

Indexing means scanning the text from documents often called index of the text searched and searching phase only search references rather than large documents which makes full text indexing more efficient than other indexing techniques.

Disadvantages

1. Fails to identify repeated words which consume additional time in processing similar words.
2. Sequence of searching is not supported by full text index.
3. It gives the results of queries as large collection of data.

Cluster Index –

It is a type of index in which the data is arranged in distinct order (in sequence) which means clustered index determines the physical order of data in table. It is beneficial when there is need to access the records sequentially or in the reverse order [157,158]. When the order of the records in a data file is in the same order as or similar to the order of the entries in the index file, the index file is said to be clustered. There may be at most one clustered index for a given data file.

A data file that is clustered on an index is not necessarily maintained as a sorted file, even though it may begin life that way, since the cost to maintain the sort is expensive in the face of frequent inserts and deletes.[162] The benefit of a clustered index is evident when performing range queries since the index entries point to records that are distributed across the smallest number of pages.

A table can only have one **clustered** index, because the clustered index sorts the rows in the table itself. *Every table in the database should have a well-chosen clustered index* to aid data retrieval and modification.

It is also possible to create a database file which is clustered on an attribute of the relation without creating an index for the attribute. Furthermore, some DBMS allow attributes from different files to be clustered (inter-file clustering), which is most useful when the attributes are frequently retrieved in the same query.

A clustered index determines the physical order of data in a table. A clustered index is analogous to a telephone directory, which arranges data by last name.[172] Because the clustered index dictates the physical storage order of the data in the table, a table can contain only one clustered index. However, the index can comprise multiple columns (a composite index), like the way a telephone directory is organized by last name and first name.

A clustered index is particularly efficient on columns that are often searched for ranges of values. After the row with the first value is found using the clustered index, rows with subsequent indexed values are guaranteed to be physically adjacent.[170] For example, if an application frequently executes a query to retrieve records between a range of dates, a clustered index can quickly locate the row containing the beginning date, and then retrieve all adjacent rows in the table until the last date is reached. This can help increase the performance of this type of query. Also, if there is a column(s) that is used frequently to sort the data retrieved from a table, it can be advantageous to cluster (physically sort) the table on that column(s) to save the cost of a sort each time the column(s) is queried.

Clustered indexes are also efficient for finding a specific row when the indexed value is unique.[174] For example, the fastest way to find a particular employee using the unique employee ID column emp_id is to create a clustered index or PRIMARY KEY constraint on the emp_id column.

There can only be one clustered index per table, because the data rows themselves can only be sorted in one order. There are row locators which is clustered index key on the row. The only time the data rows in a table are stored in sorted order is when the table contains a clustered index. If a table has no clustered index, its data rows are stored in a heap

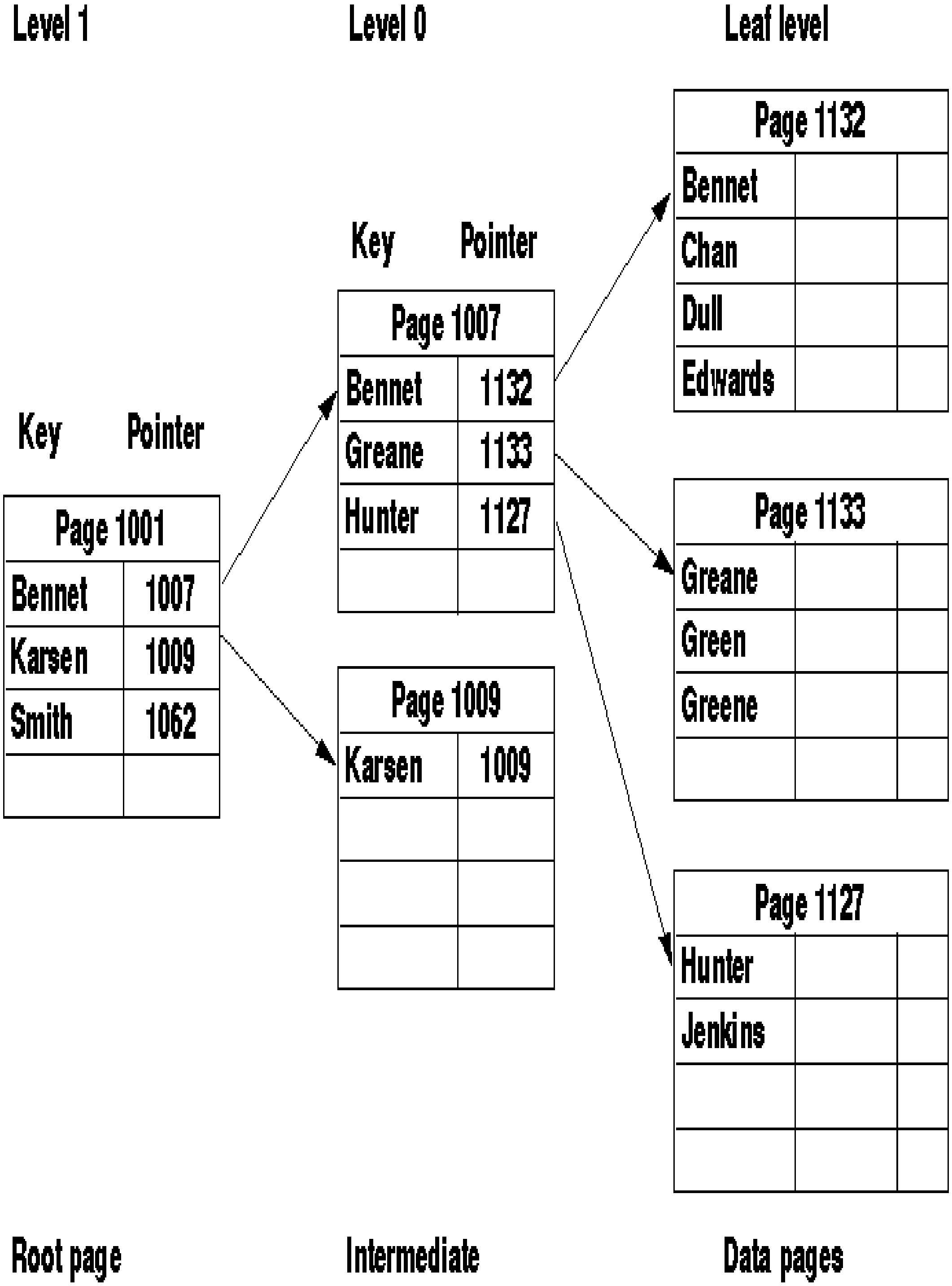


Figure1.5 – Architecture of cluster indexing

Disadvantages

1. Inserts and updates take longer time with clustered index
2. Cluster index is avoided when there are concurrent inserts on almost same clustering index value.

Non Clustered Index

The data of records are present in random order but the logical ordering is created by index. The index partition is one but when there are more partitions, then every partition is kind of B-tree structure which contains the indexes. The physical order of the rows is not the same as the index order.

A **non clustered** index is a separate physical structure from the underlying table. It contains the values for the included columns – called index keys – along with pointers back to the corresponding table row. On a table that has a clustered index, each non clustered index's pointer is the clustered index key. Note that a clustered index is ordered, but it does not alter the order of the rows in the table [159].

A non clustered index **is analogous to an index in a textbook**. **The data is stored in one place, the index in another**, with pointers to the storage location of the data. The items in the index are stored in the order of the index key values, but the information in the table is stored in a different order (which can be dictated by a clustered index). If no clustered index is created on the table, the rows are not guaranteed to be in any particular order.[170]

These indexes are mostly created on column where queries like JOIN, WHERE, and ORDER BY clauses are used and better for those tables whose values are modified frequently.

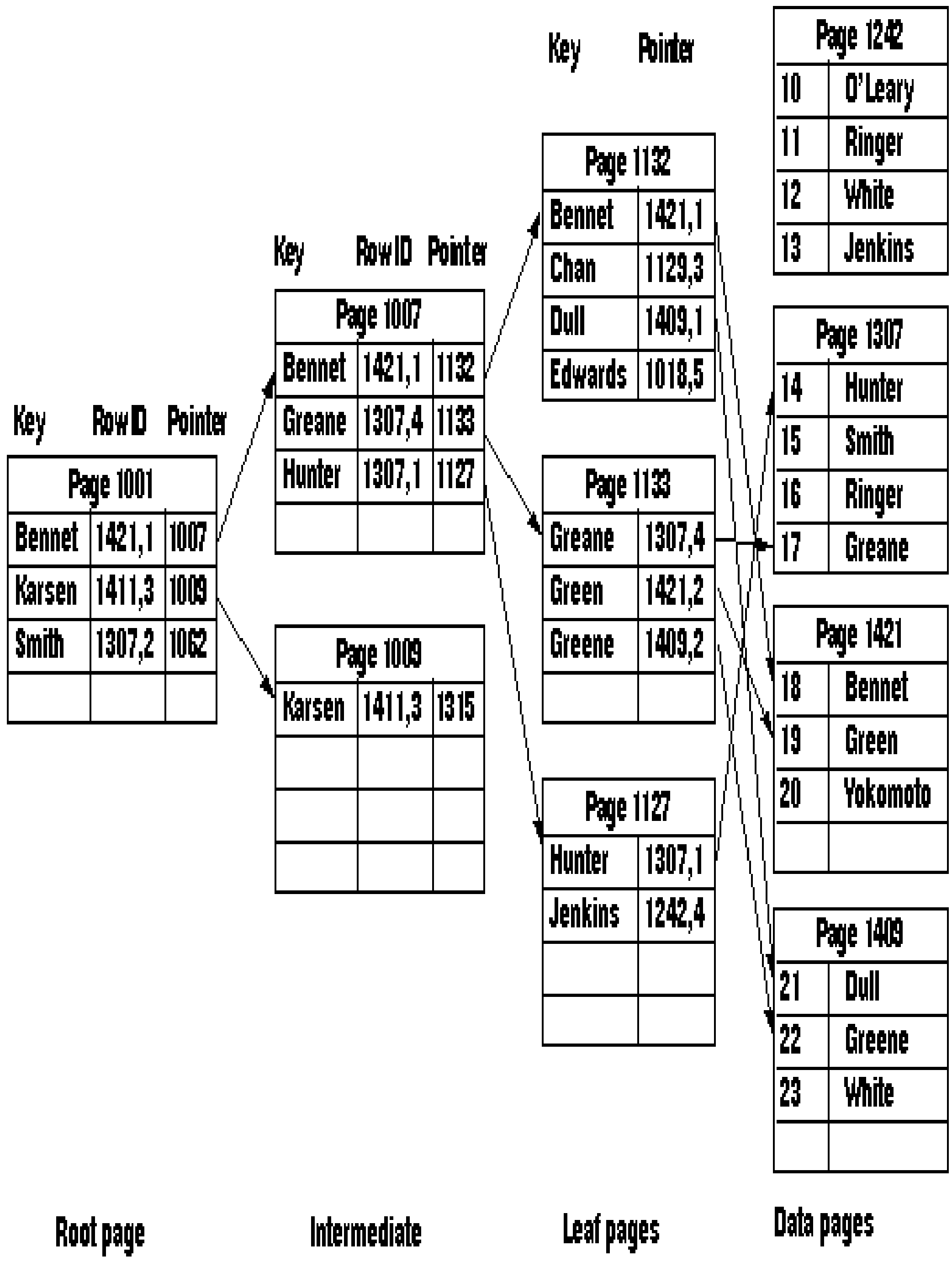


Figure1.6 – Architecture of non cluster indexing

Similar to the way you use an index in a book, Microsoft® SQL Server™ 2000 searches for a data value by searching the non clustered index to find the location of the data value in the table and then retrieves the data directly from that location[172].

This makes non clustered indexes the optimal choice for exact match queries because the index contains entries describing the exact location in the table of the data values being searched for in the queries. If the underlying table is sorted using a clustered index, the location is the clustering key value; otherwise, the location is the row ID (RID) comprised of the file number, page number, and slot number of the row[174].

Considerations

Consider using non clustered indexes for:

- Columns that contain a large number of distinct values, such as a combination of last name and first name (if a clustered index is used for other columns). If there are very few distinct values, such as only 1 and 0, most queries will not use the index because a table scan is usually more efficient.
- Queries that do not return large result sets.
- Columns frequently involved in search conditions of a query (WHERE clause) that return exact matches.
- Decision-support-system applications for which joins and grouping are frequently required. Create multiple non clustered indexes on columns involved in join and grouping operations, and a clustered index on any foreign key columns.
- Covering all columns from one table in a given query. This eliminates

accessing the table or clustered index altogether.

Disadvantages

1) Each index takes up disk space and drag on data modification resulting in serious issues regarding storage requirements.

2) By using non clustered index the number of rows return are less.

3) When using non cluster index there is no saving from one row to other as successive entries of non clustered index reference rows on disk pages that are likely to be far apart.

1.8 Objective

Data ware house is not magic –they require a great deal of very hard work. In many cases data ware house assignment are viewed as a stopgap evaluate to get users off to backs or to provide something for nothing .But data ware house require alert management [3].A data ware house is good investment only if end users actually can get important information quicker and cheaper than they can use modern technology.

Management has to think seriously about how they want their ware house to perform and how they are going to get the word out to the end user community and management has to recognize that the maintenance of any other mission –critical application. So management has to take decision for very high level of maintenance.

Because previous researchers have focused on methodological, data related, operational, educational and technical issue of data warehouse implementation. There is require explore studies in which discuss organizational, project-related and environmental dimensions regarding implementation of data warehouse technology in general and in companies' then we came to know that current scenario require the data warehouse to change[5]. Keeping in view this problem we have decided to do some research on management of data warehouses. This research is aimed at achieving the following objectives:

- Evaluate the current techniques being used by Industries for data ware house.
- Introduce the data warehouse concepts and its applications;
- Suggest new techniques for the data warehouse applications.

The purpose of this research is to analysis and evaluates the problem of data warehouse and provides techniques that help to speed up the DW application performance.

1.9 Problem Definition

`As data warehousing is a rising area, many problems are found in the system. One of the main problems in company is management of data warehouse. As data warehouses is titanic systems for a company. Lot of time is spent on data extraction, cleansing and loading process. Experts say usually 80% of the time building a data warehouse is taken by these tasks. As the knowledge and capability of data ware house their demands also 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.

Data warehouses management should by a very high level. Any reimplementation of the business processes and the source systems may want the data warehouse to change. Updates are often requiring in these condition. 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.

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.

At the beginning stage of data warehouse development, many projects feel difficulty because Managers and designers failed to identify the size, possibility, and difficulty of the challenge they make. The following are some of the challenges that may arise when implementing a data warehouse project. DW applications have some problem which is faced when data ware house is managed.

1. Some times after the implementation of data warehouse it may be required to remove some useless data. Someone has to take a decision which data to remove and which one to continue. The usual cause for this problem is the storage cost.

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

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

4. After the implementation of a data warehouse, the users find a lot of loop breaks where there are chances to fine tune the data warehouse.

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

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

7. Establishing the warehouse architecture is easy than maintaining warehouse architecture. Here architecture mean to the regular 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 barrier in accessing useful information for the user of warehouse.

9. If the data warehouse has inaccurate data or values that cannot be altered properly, it is important for the data warehouse alteration process to use intelligent default values for the missing or corrupt data. It is also important to devise a mechanism for users of the data warehouse to be aware of these default values.

For the prelisted points; these are considered the cornerstones of DW applications and their performance issues. One of the problems with data warehousing is that companies are rushing to build it without regard to how that impact existing systems architectures, or how it be integrated with other applications. Critics also contend that warehouses ignore process and function, essentially sterilizing the data and removing the application context by isolating it in one or more relational database engine.

The lack of the following criteria is the common causes of data warehouse failures: pre-launch clear purpose, insider presence on data warehouse projects team, user demand for difficult data analysis, and initial involvement of business managers. Another potential difficulty involves increasing user demand and unrealistic expectations.

1.10 Research Importance

As data warehousing is an emerging area, a lot of effort and researcher's time is spent on its architecture, design and development phases. But it is not enough for the management issues. Now I am trying to putting some rays on these issues. The glory of this research stems is starts from the fact that DW applications access huge amounts of data and the end users actually get vital information faster and cheaper by using this technology. Therefore it is highly desirable that they produce the expected results which help in decision making, an easy process. the focus of most of writings about data warehousing is on planning, designing, and building them.

But no one is considering what happen after the implementation, Communication, Training, Networking, loading and Query processing are key element in improve performance of data ware house. So my Research is putting stress on above burning issues of corporate.

The genuine work of taking output from the data warehouse starts from here. Suppose successfully implemented a data warehouse for an organization. so it is useful to setup solid procedures for managing this project. Designing data warehouses is very different from designing traditional operational systems.

Designing a data warehouse often involves thinking in terms of much wider, and more difficult to define, business concepts than does designing an operational system. So data ware housing is quite close to Business Process Reengineering.

Finally, the ideal performance strategy is communication and training, help and support, 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.

1.11 Thesis Organization

This thesis comes on six distinct chapters as follows: I have divided our research work into five chapters as described below:

Chapter 2: This chapter have explain the literature survey of the research study under subject. Starting with performance strategies of data warehousing which affect the data ware house performance.

Chapter 3: In this chapter I have discussed my research material, objective and research problem, data collection strategies, data analysis methods and the methods to validate the research findings. This chapter serves us as a guide throughout the remaining part of the thesis.

Chapter 4: In this chapter I have present my findings from the case study that I have done. I have present how data warehouse maintenance is actually carried out in an organization. In this chapter I have compare the theoretical findings (Chapter 2) with the empirical findings (Chapter 4) and have discuss which ways are better than others and why?

Chapter 5: In this chapter I have presented our findings from the thesis. Keeping an eye on all the previous work here I have present our conclusions. Additionally I have presented some areas where future research could be done in the area of data warehouse maintenance.