DATA WAREHOUSING

REEMA THAREJA
Assistant Professor
Department of Computer Science
Shyama Prasad Mukherjee College for Women
University of Delhi

OXFORD
UNIVERSITY PRESS

CONTENTS

Preface *v*

PART I

Chapter 1: Introduction to Data Warehousing **3**
1.1 A Short Historical Note *4*
1.2 Increasing Demand for Strategic Information *5*
1.3 Data Warehouse Defined *10*
1.4 Data Warehouse Users *14*
1.5 Benefits of Data Warehousing *17*
1.6 Concerns in Data Warehousing *19*

Chapter 2: Data Warehouse: Defining Features **23**
2.1 Introduction *24*
2.2 Features of a Data Warehouse *24*
2.3 Data Granularity *31*
2.4 The Information Flow Mechanism *33*
2.5 Metadata *47*
2.6 Two Classes of Data *52*
2.7 The Lifecycle of Data *53*
2.8 Data Flow from Warehouse to Operational Systems *56*

Chapter 3: Architecture of a Data Warehouse **61**
3.1 Introduction *62*
3.2 Characteristics of Data Warehouse Architecture *62*
3.3 Data Warehouse Architecture Goals *65*
3.4 Data Warehouse Architecture *65*
3.5 Data Warehouse and Data Mart *73*
3.6 Issues in Building Data Marts *74*
3.7 Building Data Marts *79*
3.8 Other Data Mart Issues *81*
3.9 Increased Popularity of Data Marts *85*
3.10 Can Data Warehouse and Data Mart Co-exist? *85*
3.11 Pushing and Pulling Data *86*

PART II

Chapter 4: Gathering the Business Requirements **91**
4.1 Introduction *92*
4.2 Determining the End-user Requirements *93*
4.3 Requirements Gathering Methods *95*
4.4 Requirements Analysis *102*
4.5 Dimensional Analysis *103*
4.6 Information Package Diagrams (IPD) *107*

Chapter 5: Planning and Project Management **114**
5.1 Project Management Principles *115*
5.2 Data Warehouse Readiness Assessment *117*
5.3 Data Warehouse Project Team *120*
5.4 Planning for the Data Warehouse *121*
5.5 Data Warehouse Project Plan *124*
5.6 Economic Feasibility Analysis *126*
5.7 Planning for the Data Warehouse Server *130*
5.8 Capacity Planning *135*
5.9 Selecting the Operating System *140*
5.10 Selecting the Database Software *142*
5.11 Selecting the Tools *144*

Chapter 6: Data Warehouse Schema **155**
6.1 Introduction *155*
6.2 Dimensional Modelling *156*
6.3 The Star Schema *158*
6.4 The Snowflake Schema *164*
6.5 Aggregate Tables *169*
6.6 Fact Constellation Schema *173*
6.7 The Strengths of Dimensional Modelling *177*
6.8 Data Warehouse and the Data Model *177*

Chapter 7: Dimensional Modelling **185**
7.1 Characteristics of a Dimension Table *186*
7.2 Characteristics of a Fact Table *187*
7.3 The Factless Fact Table *189*
7.4 Updates to the Dimension Tables *189*
7.5 Cyclicity of Data—The Wrinkle of Time *195*
7.6 Other Types of Dimension Tables *196*
7.7 Keys in the Data Warehouse (Star) Schema *198*
7.8 Enhancing the Data Warehouse Performance *200*
7.9 Technology Requirements *214*

Chapter 8: The ETL Process **219**
8.1 Introduction *220*
8.2 Data Extraction *221*
8.3 Data Transformation *229*
8.4 Data Loading *232*
8.5 Data Quality *237*

Chapter 9: Testing, Growth, and Maintenance **246**
9.1 Data Warehouse Design Review *247*
9.2 Developing the Data Warehouse Iteratively *250*
9.3 Testing *250*
9.4 Monitoring the Data Warehouse *256*
9.5 Tuning the Data Warehouse *259*
9.6 The Feedback Loop *262*

Chapter 10: OLAP in the Data Warehouse **266**
10.1 Need For OLAP *267*
10.2 OLAP *269*
10.3 OLAP and Multidimensional Analysis *272*
10.4 OLAP Functions *276*
10.5 OLAP Applications *284*
10.6 OLAP Models *285*
10.7 OLAP Design Considerations *293*
10.8 OLAP Tools and Products *293*
10.9 Existing OLAP Tools *295*
10.10 Data Design *297*
10.11 Administration and Performance *298*
10.12 OLAP Platforms *299*

PART III

Chapter 11: Building a Data Warehouse **305**
11.1 Introduction *306*
11.2 Problem Definition *306*
11.3 Critical Success Factors *307*
11.4 Requirement Analysis *309*
11.5 Planning for the Data Warehouse *310*
11.6 The Data Warehouse Design Stage *313*
11.7 Building and Implementing Data Marts *317*
11.8 Building Data Warehouses *317*
11.9 Backup and Recovery *323*
11.10 Establish the Data Quality Framework *325*
11.12 Operating the Warehouse *327*
11.13 Recipe for a Successful Warehouse *336*
11.14 Data Warehouse Pitfalls *336*

Chapter 12: Data Mining Basics **340**
12.1 Introduction *341*
12.2 Architecture of a Data Mining System *348*
12.3 The Knowledge Discovery Process *349*
12.4 Integrating Data Mining with Data Warehouse *352*
12.5 Related Areas of Data Mining *356*
12.6 Data Mining Techniques *357*

Chapter 13: Moving into Data Mining **379**
13.1 Introduction *380*
13.2 How Do We Categorize Data Mining Systems? *381*
13.3 Interesting and Useful Data *382*
13.4 Applications of Data Mining *383*
13.5 Web Mining *391*
13.6 Text Mining *395*
13.7 Temporal Data Mining *396*
13.8 Sequence Mining *399*
13.9 Time-series Analysis *400*
13.10 Spatial Data Mining *402*
13.11 Issues and Challenges in Data Mining *404*
13.12 Current Trends in Data Mining *405*

Chapter 14: Trends in Data Warehousing **410**
14.1 Introduction *411*
14.2 Data Warehouse Solutions *411*
14.3 Web-enabled Data Warehouse *416*
14.4 Distributed Data Warehouse *427*
14.5 Virtual Data Warehouse *430*
14.6 Operational Data Store *434*
14.7 Integration with Other Technologies *438*
14.8 Trends in Data Warehousing *447*
14.9 Data Warehouse Futures *452*

Glossary *457*
Bibliography *466*
Index *470*

1

INTRODUCTION TO DATA WAREHOUSING

Learning Objectives

This chapter aims to provide an overview of the fundamental concepts of data warehousing. It endeavours to answer questions regarding the need for a data warehouse, its evolution, characteristics, and applications.

Case Study

We shall introduce a case study of a company that requires a data warehouse to ease the running of its operations. Through the decisions taken by the managers, we shall study construction, operation and management of a data warehouse. This case study shall run throughout the book.

Pallav Raj is the CEO of a large garments retail chain called JRTs. He asks one of his employees to provide him with a status report on the business as he wishes to know if the company was making an overall profit or loss. JRTs has approximately 100 stores spread throughout the country. Although this is not a difficult question to answer, the problem lies in collecting the relevant data that is spread across 100 stores.

With great difficulty, the employee contacts each and every store and asks the store managers to give a summarized figure describing whether the store is running at a profit or loss. After obtaining 100 such figures, he calculates the cumulative result and gives it to Pallav Raj.

The problem does not end here for the employee! Pallav Raj now wants a detailed product report of the previous year as he wishes to know which products sold well and those that did not even have a marginal sale. Again the employee contacts each and every store and thus the entire process is repeated.

Such situations prevail in a non-data warehouse environment. This is where the concept of data warehouses comes into picture. In a data warehouse environment, the entire data of all the stores is stored at one place, that is, on one single computer system at the main office. In such a situation, the employee's work would have been very easy. Or rather, he would not have been required as the CEO could have himself gained access to all the data while sitting on his chair.

1.1 A SHORT HISTORICAL NOTE

Computers came into existence in 1914 and since then, the field of computer science has witnessed a tremendous growth in hardware as well as software technologies. Computers that were one day meant only for scientific applications became so widespread, that today hardly any business runs without a computer.

Information technology professionals work on computer applications as analysts, programmers, designers, developers, database administrators or project managers. Depending upon the industries in which they work, they are usually involved in applications such as order processing, general ledger inventory, inpatient billing, checking accounts, insurance claims, and so on. These applications are the lifelines of the business as they are used to run the businesses.

These applications process orders, maintain inventory, keep the accounting books, service the clients, receive payments, and process claims. The situation today is that without these computer systems, no modern business can survive. Companies started using these kinds of applications in the 1960s and now they cannot even think of running their businesses without a computer.

However, in the 1990s, businesses grew more complex with corporations spreading globally. As the competition in the market became fiercer, business executives became desperate for information to stay competitive and improve the bottom lines of their business. The operational computer systems were meant to provide information to run the day-to-day business operations, but the business executives and managers needed different kinds of information that could be readily used to make strategic decisions. For example, business executives need to know where to build the next warehouse, which product lines to expand, and which markets should be strengthened. The operational systems could not provide this strategic information to its users. Businesses, therefore, were compelled to turn to new ways of getting strategic information. This new way is called *data warehousing*.

Data warehousing is thus a new paradigm that provides strategic information to its users. In the 1990s, organizations began to achieve competitive advantages by moving into this technology. Basically, data warehousing is a comprehensive term which indicates the various activities involved in the construction, maintenance, and use of the *information oriented architecture.*

Business organizations can achieve considerable competitive advantages by analysing their historical data. This analysis can reveal certain unusual trends in sales that in turn can indicate opportunities for new business. Moreover, the analysis of past customer demands can help to forecast production needs. A data warehouse is thus an integrated collection of *enterprise-wide data*, oriented to decision making that is built to support this activity.

Data warehousing systems facilitate business executives and managers to acquire and integrate information from heterogeneous sources and to query very large databases efficiently. Building and implementing a data warehouse

calls for adoption of design and implementation techniques that are strikingly different from those applied in underlying operational information systems.

1.2 INCREASING DEMAND FOR STRATEGIC INFORMATION

Before we look into the need for strategic information, let us first clarify what is meant by strategic information. Strategic information is not required for running day-to-day operations of the business and neither is it required to produce an invoice, make a shipment, settle a claim, or post a withdrawal from a bank account. Yet it is critical for the survival of the corporation in a highly competitive world as critical business decisions depend on the availability of proper strategic information in an enterprise.

Strategic information in an enterprise is meant for the executives and managers who are responsible for keeping the enterprise competitive. They need this information to make the right decisions at the right time, to formulate business strategies, establish goals, set objectives, and monitor results. Here are some examples of business objectives:

- Retain the current customers of the business.
- Add to the customer base by at least 10% over the next 3 years.
- Enhance the market share by 15% in the next 2 years.
- Launch new and better products in the market by the next year.
- Improve product quality of top five selling products.
- Increase sales in the north west region.

For making decisions about business objectives, executives and managers need different kinds of information to:

- Get detailed knowledge of the company's operations.
- Analyse how key business factors affect each other.
- Monitor how the business factors change with time.
- Compare the company's performance with that of their competitor's.

For making effective decisions the executives and managers need to focus on customer's needs and preferences, emerging technologies, sales and marketing results, and quality levels of products and services. The information needed for formulating and executing business strategies and objectives is meant for the entire organization. Such type of information is referred to as *strategic information*. Table 1.1 lists the characteristics of strategic information.

Table 1.1 Characteristics of strategic information

▪ **Integrated**	▪ Must have an overall enterprise-wide view
▪ **Data integrity**	▪ Data in all the tables must be accurate
▪ **Accessible**	▪ Easily accessible by the users with intuitive access paths
▪ **Timely**	▪ Information must be available within the stipulated time

1.2.1 The Information Crisis

The various computer applications in an organization produce a huge amount of data which keeps building up and getting stored over a period of several years. The organizations are thus faced with two astonishing facts:

- Organizations have a huge amount of data.
- Information systems that they have are ineffective at turning this into useful strategic information.

Since the past several years, companies have been accumulating tons and tons of data from their day-to-day operations. Thus, colossal amounts of data already exist and this information is said to double every 18 months. As a result, these companies have witnessed an information crises not because of lack of sufficient data, but due to the unavailability of data that is readily useful for strategic decision making.

The large quantities of data that exist within an organization are very useful for running the business operations but hardly amenable for use in making decisions about business strategies and objectives. Hence, information crises persist because of two main reasons:

- The data in a corporation resides in various disparate systems, multiple platforms, and diverse structures. But, for proper decision making on overall corporate strategies and objectives, we need information integrated from all systems.
- Data needed for making strategic decisions must be available in a format that enables executives and managers to analyse trends in order to lead their companies in the right direction. For this they need to review the data from different business viewpoints. The tremendous amounts of operational systems data cannot be readily used to spot trends.

Operational data is event driven, that is, you record the details of each and every transaction that happens. This data cannot be used to state the prevailing trend in the market. For this purpose you need to provide data from different viewpoints to the managers and executives. For example, they must be able to review sales quantities by product, salesperson, region, and customer demographics. Of course, operational data cannot be directly used for reviewing data from different angles.

1.2.2 Inability of Past Decision-support Systems

To start with the topic, let us first analyse a real time scenario that existed when the concept of a data warehouse was not present. The marketing department in a company has been concerned about the performance of a particular region as the sales numbers from the monthly report of that month are drastically low. The marketing manager wants to get some report from the IT department to analyse the performance over the past two years, product by product and compared to monthly targets. He wants to take quick strategic decisions to

rectify the situation. Now, there may not be any regular reports to give to the marketing department on what they want. The IT department has to gather the data from multiple applications and start forming the report from scratch.

Sometimes, they have to get the information required for such ad hoc reports from the databases of not one but several applications, perhaps running on different platforms. What happens next? The marketing department likes the report but now they may like the report to be produced in a different form, containing some more information as illustrated in Fig. 1.1.

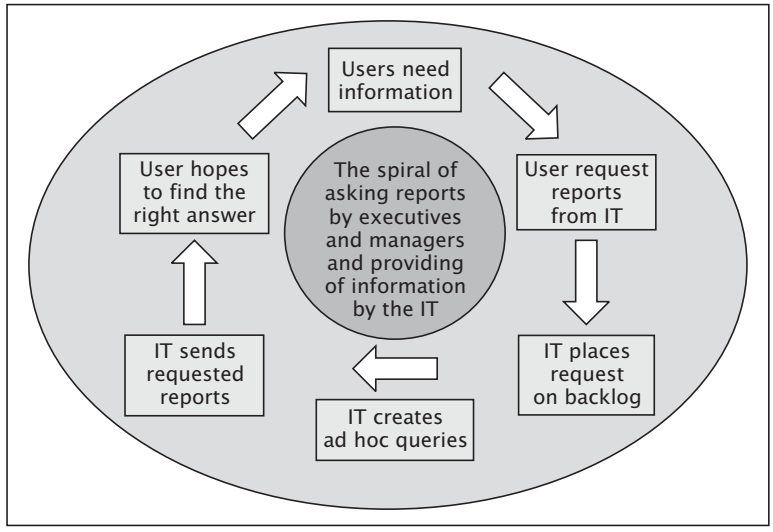

Figure 1.1 Inadequate attempts by IT to provide strategic information

Most of these attempts by IT in the past ended in failure as the users could not clearly define what they wanted in the first place. After seeing the first set of reports, they wanted more data in different formats. The chain continued. The mess was clearly due to the very nature of the process of making strategic decisions.

Information needed for making strategic decisions must be available in an interactive manner so that the users can query online, get results, and query further. The information must be in a format suitable for analysis. Hence, some factors that were responsible for the inability to provide strategic information in the past prior to data warehousing are as follows:

- IT received too many ad hoc requests for a variety of reports. But with limited resources, IT was not able to generate all the reports in the requested manner and within the assigned timeframe.
- Requests were not only numerous, but also kept changing over time with users wanting more reports subsequently to expand and understand earlier reports.

- The users indulged themselves into the spiral of asking for more and more supplementary reports thereby increasing the IT load even further.
- The users depended on IT to provide the information as they could not access the information directly in an interactive manner.
- As a result, IT was unable to provide an environment for flexible and conducive analysis to the managers and executives for making strategic decisions.

1.2.3 Presence of Better Technology

The field of information technology has witnessed the breathtaking changes that have taken place in the last decade wherein the IT infrastructure has changed rapidly and its capabilities have constantly increased as evidenced by the following:

- The power of microprocessors has been doubling every two years.
- The processing speed of the microprocessors has increased while its cost has declined.
- The price of digital storage has been coming down for several years.
- There has been a constant increase in the network bandwidth and a decrease in its cost to access that bandwidth.
- The workplace has now become heterogeneous in terms of hardware and software.
- Legacy systems are now integrated with new applications.

Also hardware economics and miniaturization allow a workstation on every desk and provide increasing power at reducing costs. New software provides user friendly systems. Improved connectivity, networking, and the internet open up interaction with an enormous number of systems and databases. All of these improvements in technology are meritorious as they have made computing faster, cheaper, and widely available.

To provide strategic information a large collection of corporate data stored in suitable formats is required. Technology advances in data storage and reduction in storage costs fulfils the data storage needs for strategic decision-support systems. Executives, managers, and business analysts use strategic information to analyse data and spot prevailing trends. The user does the analysis in an interactive manner by asking a question and getting the results then asking another question, looking at the results to ask yet another question. Tremendous advances in interface software make such *interactive analysis* possible.

Processing huge volumes of data and providing interactive analysis demands massive computing power. The tremendous increase in the computing power and its lower costs has made strategic information feasible and thus we can think of a new system that provides the users with this type of information.

1.2.4 Expectations from the Decision-support System

We need a different type of decision-support system to provide strategic information that is different from available operational systems. We need a new type of system environment for the purpose of providing strategic information for analysis, discerning trends, and monitoring performance. The advantages of this type of system environment designed for strategic information are:

- Database designed for analysing large volumes of data.
- Data extracted from multiple applications.
- User friendliness.
- Intuitive to use for long interactive sessions by users.
- Containing read intensive data usage that is stable in nature.
- Enables easy usage of the system by the users without assistance from IT professionals.
- Periodically updating of data content.
- Contain current as well as historical data.
- Ability for users to formulate and execute queries and get results online.

1.2.5 Operational vs. Decision-support System

Table 1.2 given below summarizes the differences between the two systems–the current operational system and the needed decision-support system.

Table 1.2 Operational versus decision-support system

Attributes	Operational Systems	Decision-support System
Data content	Current values	Archived, summarized, derived
Data structure	Optimized for transactions	Optimized for complex queries
Access frequency	High	Medium to low
Access type	Read, update, delete	Read
Usage	Predictive, repetitive	Ad hoc, random
Response time	Sub-seconds	Several seconds to minutes
User number	Large numbers	Relatively small number
Characteristic	Operational processing	Informational processing
Orientation	Transaction	Analysis
Users	Clerk, DBA, database professional	Executives, managers, business executives
Function	Day-to-day operations	Long-term informational requirements, decision support
Database design	ER based, application oriented	Star/snowflake, subject oriented
Summarization	Primitive, highly detailed	Summarized, consolidated
View	Detailed, flat relational	Summarized, multidimensional
Unit of work	Short, simple transaction	Complex query
Records accessed	Tens	Millions

Table 1.2 Continued

Table 1.2 Continued

Attributes	Operational Systems	Decision-support System
Database size	100 MB to GB	100 GB to TB
Priority	High performance, high availability	High flexibility, end-user autonomy
Indexes	Few	Many
Joins	Many	Some
Duplicated data	Normalized DBMS	Denomalized DBMS
Derived data and aggragates	Rare	Common

In the data warehouse model, operational systems are not accessed directly to perform information processing. But still they play a key role by acting as the source of data for the data warehouse. The users will use this data warehouse which is the information repository and point of access for information processing.

1.3 DATA WAREHOUSE DEFINED

Data warehouses were developed to meet the growing demand for information analysis that could not be met by operational systems for a range of reasons:

- The processing load of reporting affected the response time of the operational systems.
- The database designs of operational systems were not optimized for information analysis and strategic decision making.
- Generally all big organizations had a number of operational systems so enterprise-wide reporting could not be supported from a single system.

As a result, separate databases were built that were specifically designed to support management information and analysis purposes. Data warehouses collected data from a range of different data sources, such as mainframe computers, minicomputers, as well as personal computers and office automation software such as spreadsheet, and integrate this information in a single place. The bottomline of success of such type of databases is its capability, coupled with user-friendly reporting tools and freedom from operational impacts.

The data warehouse enables the organization to make use of an enterprise-wide data store to link information from diverse sources and make the information accessible to the users for strategic analysis. Here the term strategic analysis is a comprehensive term that includes trend analysis, forecasting, competitive analysis, and targeted market research.

The data warehouse is thus an informational environment which does the following:

- provides an integrated view of the enterprise.
- renders the enterprise's current as well as historical data readily available for making strategic decisions.
- makes decision making possible without hindering operational systems.
- makes the organization's information consistent and easily accessible.
- provides a flexible, conducive and interactive source of strategic information.

1.3.1 What Can a Data Warehouse Do?

In this section, we will study about the capabilities of a data warehouse. So let us have a look at what a data warehouse can do.

Immediate information delivery Data warehouses reduces the time period lapsed between the request for information and the actual delivery of information to the users. For example, the sales report was formed once in every month, usually in the first week of every month. But with data warehouses the same report can be formulated on a daily basis thereby enabling the business analysts to exploit opportunities that could otherwise have been raised.

Integration of data from within and outside the organization Data warehouses combine data from multiple sources. The data is collected from different departments like sales, marketing, finance, and accounting. Besides this, data is also taken from external sources like business magazines, news reports, survey's etc.

Provides an insight into the future Data warehouses store large amounts of historical information that enables the decision makers to analyse the prevailing trends in the market and produce goods according to the customers demands.

Enables users to look at the same data in different ways A data warehouse provides its users with tools for analysing and manipulating data in many different ways. It facilitates the users to drill down into detailed data with the click of a mouse that could have otherwise taken a few days with the traditional approach.

Provides freedom from the dependency on IT With data warehouses, the users have to no longer depend on the availability of IT professionals to answer their queries. Now, if the manager needs an ad hoc report, he can himself form it without the assistance of any computer guru.

Table 1.3 illustrates how a data warehouse can help its users to analyse sales.

Table 1.3 Sales analysis by a data warehouse

Sales Analysis
▪ Determine sales to take vital decisions regarding price and distribution. ▪ Determine the success and failure attributes by studying the historical sales data. ▪ Determine successful products and learn about their key success factors. ▪ Understand the profits as well as revenue implications of a decision. ▪ Identify the most promising customers. ▪ Identify customers who are no longer loyal to the organization. ▪ Identify the salesperson that have performed extremely well and those who could not work to fulfill their expectations.

1.3.2 What Can a Data Warehouse Not Do?

A data warehouse is not a magical box; it does have some limitations. It acts as an information repository that collects and reports data that already exists. It cannot create additional data on its own. For example, if a manager wants to analyse the sales of a product based on customer's income level, and if the income of the customer is not captured by the source systems, then the data warehouse will not be able to help the users in any way until and unless a mechanism is devised to gather the income data.

Apart from this, if an organization has dirty data in the source systems, the data warehouse will not be able to correct results until and unless the data is first cleaned. In this context, a data warehouse will only be able to identify where the problem exists, but corrections will have to be made in the source systems that capture that data.

1.3.3 Data Warehouse—An Environment or a Product?

A data warehouse is a user centric environment that enables its users to use the data stored in it directly for making strategic decisions. To consider it as a part of either a software or hardware product that can be purchased from the market to provide strategic information is not correct. Rather, it is an overall strategy, or process, for building decision support systems, a knowledge-based applications architecture and an environment that supports long-term decision making.

It is an architectural construct of information systems that provides users with current and historical information to support strategic decisions that are otherwise hard to access or present in traditional operational data stores. In fact, the data warehouse is a cornerstone of the organization's ability to do effective information processing that thereby enables the discovery and analysis of trends and dependencies that otherwise would have gone unnoticed.

In principle, data warehouses are designed to satisfy the informational needs of managers and executives. A data warehouse once deployed is meant to provide strategic business opportunities by allowing all its users to access the corporate data without violating the security measures. The characteristics of this new computing environment called the data warehouse are:

- An ideal environment for data analysis and decision making.
- Flexible, intuitive, and interactive.
- User friendly.
- Conducive and responsive to formulate and execute interactive queries.
- Enables the users to discover answers to complex queries.

1.3.4 A Blend of Many Technologies

The key reason for the implementation of a data warehouse is to bring together information from disparate sources and put the information into a format that is conducive for making business decisions. This calls for a set of activities that are far more complex than just collecting data and reporting against it (Refer Fig. 1.2). Data warehousing requires both business and technical expertise and involves the following activities:

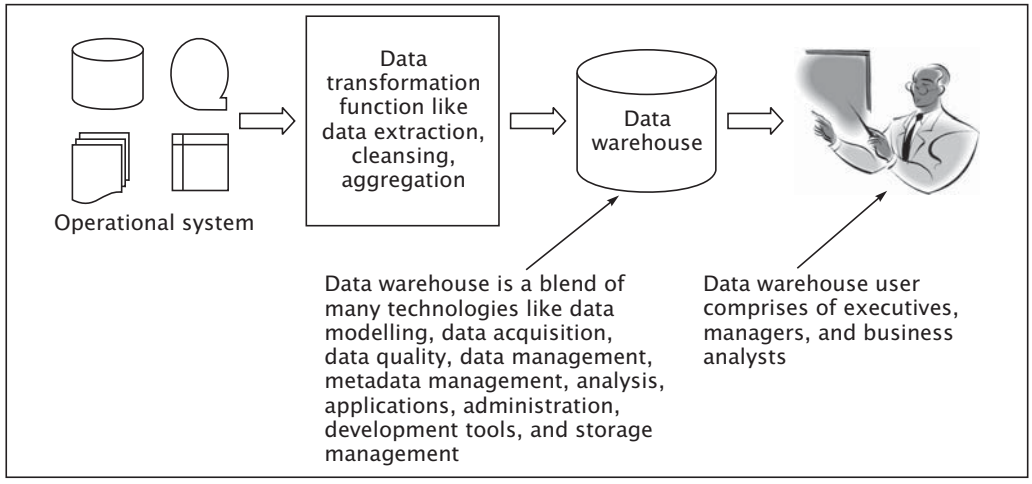

Figure 1.2 Data warehousing is a blend of many technologies

- Accurate identification of business information that must be stored in the warehouse.
- Identification and prioritization of subject areas to be included in it.
- Defining the scope of each subject area.
- Development of a scaleable architecture.
- Selection of the hardware/software/middleware components needed.

- Extracting, cleansing, aggregating, transforming, and validating the data to ensure accuracy and consistency.
- Providing user-friendly, powerful tools to the users with which they can gain access to the data warehouse.
- Giving adequate training to the users.
- Establishing a data warehouse helpdesk to support the users in their day-to-day tasks.
- Establishing procedures for maintenance and enhancement of the data warehouse.

Thus, the basic operations of data warehousing are to:

- Extract data from the operational systems.
- Include relevant data from outside sources like magazines, journals, reports of other organizations in the same industry.
- Remove inconsistencies and transform and clean the data.
- Store the data in such a way so that it is for easy access for decision making.

Although data warehousing seems to be a simple concept, it however involves different functions like data extraction, data loading, transforming, storing the data, and providing user interfaces. Table 1.4 shows various uses of a data warehouse system.

Table 1.4 Applications of a data warehouse system

Industry	Applications
Retail	Customer loyalty, targeted marketing
Financial and banking	Risk management, fraud detection
Airlines	Route profitability, promotional schemes
Manufacturing	Cost reduction, resource management
Government	Manpower planning, development, and cost control
Other application areas include: insurance companies, utilities providers, health care providers, financial services companies, telecommunications service providers, travel, transport and tourism companies, security agencies, logistic, inventory, and purchasing.	

1.4 DATA WAREHOUSE USERS

A data warehouse is primarily designed to support executives, senior managers, and business analysts in making complex business decisions. They provide the business users with access to accurate, consolidated information from various internal and external sources.

The ideal data warehouse users are the people who can be described by characteristics as given here.

- People whose job involves analysing data to draw meaningful conclusions and make decisions based on large masses of data without the need to organize the data for this purpose.
- Those who are not supposed to access the database in a highly technical fashion to find the desired information.
- Those whose decisions have enough value in terms of enhanced productivity, increased sales, better quality of products, targeted advertising, etc. to the organization to justify the data warehousing effort.

A data warehouse is a high level solution for making strategic decisions and is not meant to be used by every user in the organization. That is, a data warehouse is not the universal solution to all types of business's information needs. The users of the data warehouse include executives, senior managers, CEOs, business analysts, and some high level computer professionals. Now after having a look at the people who need to access the data warehouse, let us also study which are the people who do not need to have a data warehouse.

- People whose job involves dealing with individual data records (daily transactions).
- Anyone whose job includes updating the organizational database, not just looking at what data is already stored in it. These users may need an operational database for the purpose and not the data warehouse.

Therefore, we see that, anyone who needs strategic information is expected to be a part of the group of users of the data warehouse. The user group includes business analysts, business planners, managers, and senior executives. Every group of users has specific business needs for which they expect to get answers from the data warehouse. It is always better to classify the user groups depending on what information they expect from the warehouse. Every user is supposed to perform a particular business function and needs information to support that function.

In order to make the information delivery mechanism best suited for the data warehouse environment, you need to have a good understanding of the classes of users (Refer Fig. 1.3). We will classify the users based upon two perspectives—their computing proficiency and their job function.

Casual or novice user Uses the data warehouse occasionally and needs a very intuitive information interface.

Regular user Uses the data warehouse almost daily. These users are comfortable with computing options but cannot create reports and queries on their own and thus make use of query templates and predefined reports.

Power user These users are well versed and highly proficient with the technology. They are capable of creating reports and executing queries on their own and can even write macros and scripts for their applications.

Users can also be classified based upon their job functions as below:

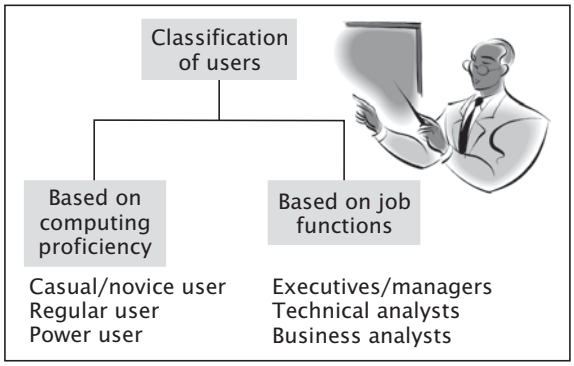

Figure 1.3 Data warehouse users classification

Executives and managers They need information for making high level strategic decisions. They prefer customized and personalized reports.

Technical analysts They perform complex analysis and statistical analysis, perform drill-down, roll-up, slice and dice operations on the data.

Business analysts Although these users are comfortable with the technology, they may not be able to write queries and create reports from scratch. So, they rely on predefined queries and reports to satisfy their information needs.

1.4.1 Why Do They Want Information?

Any organization that has implemented data warehousing can exploit the extensive data stored in it for activities like planning, execution, and assessment of results. The data warehouse can assist in planning for a market expansion and in the assessment of the results of the execution of marketing campaigns for that purpose. We will go through a few general areas of the enterprise where the data warehouse can be useful in the planning and assessment phases of the management loop.

Profitability growth To increase profits of the business, the managers and executives need to understand the underlying relationship that exists between profit and product categories, markets and services. For this managers must know which products and markets produce greater profits. The information from the data warehouse provides support to plan for profitability, growth, and to assess the results when plans are executed.

Strategic marketing Strategic marketing guides the process of business growth. The data warehouse offers great information potential for strategic marketing by providing the users with information regarding up-selling and cross-selling to its existing customers and for expanding the customer base.

Customer relationship management The data warehouse contains all the information about the customers that is extracted from various disparate source systems, transformed and integrated. This provides an opportunity to the executives and managers to learn their customers individually from the information available in the data warehouse and thus create better relations with them.

Corporate purchasing The data warehouse contains corporate-wide information about the different vendors' and the customers' purchasing patterns. So, the data warehouse can easily empower the corporate management to plan for streamlining purchasing process.

For these purposes, the end-users try to find information in two distinct modes.

Verification mode In this mode, the data warehouse user proposes a hypothesis and then asks a series of questions to either confirm or repudiate it. Consider, that the marketing department has launched a new product in the market. Now the marketing department wants to assess the sales of the product. So, the marketing department goes to the data warehouse with the hypothesis that the product has a tremendous sale in the market. The information from the data warehouse will help confirm the validity of this hypothesis.

In this mode, the users can retrieve historical as well as current data and perform statistical analysis using query and reporting tools. The results may either be in the form of reports or charts. The users may also use complex operations like roll-up, drill-down slice, dice, and pivot. We will learn about all these operations in the coming chapters.

Discovery mode In this mode, the user does not use a predefined hypothesis. Rather, he attempts to discover patterns of customer behaviour and relationships among the products that sell together. We will learn more about this mode in Chapter 12. Basically, in this mode the user does not have any preconceived notions of the result sets.

To cater to the needs of different users, the user-information interface must have the following features.

- Be easy to use, intuitive, and enticing to the users thereby being very user friendly.
- Support the ability to express business needs clearly in the form of rules.
- Be linked with metadata (data about data, similar to data dictionary).
- Be capable of formatting and structuring output in a variety of ways, both textual and graphical.

1.5 BENEFITS OF DATA WAREHOUSING

There are many benefits in using a data warehouse, some of which are:

- Data warehouses enable end-users to access a wide variety of data.
- Business analysts and decision makers can analyse the current trends in the market to predict future trends for example the analyst can analyse the products sales to find the item with maximum sales in a particular area for the last two years. This may be helpful for future investments in a particular item.

- Data warehouse provides consistent data.
- It helps to increase productivity and decrease computing costs.
- Data warehouses contain data that has been integrated from a number of different sources.
- The results obtained can be presented in a variety of formats in the form of reports, graphs, etc.
- Data warehouse users can obtain trend reports, for example the products that had maximum sales in the northern region within the last two years and exception reports that show actual performance versus goals.
- Data warehouses enhance the value of operational business applications, notably customer relationship management (CRM) systems.

Data warehouse architecture not only enhances the availability of business intelligence data but also improves the effectiveness and timeliness of the startegic decisions. The benefits of the data warehouse can be further subdivided into two categories–tangible benefits and intangible benefits. They are explained as follows.

1.5.1 Tangible Benefits

Successfully implemented data warehouse can realize some significant tangible benefits. For example, assuming an improvement in out-of-stock conditions in the retailing business that results in a 1% increase in sales can mean a sizable cost benefit to the business. As even for a small retail business with $200 million in annual sales, a 1% improvement in sales can yield additional annual revenue of $2 million. However, this benefit is in addition to retaining customers who might not have stayed loyal to the organization if, because of out-of-stock problems, they had to do business with other retailers. Other examples of tangible benefits of implementing a successful data warehouse includes the following:

- Cost of product introduction comes down with targeted marketing campaigns.
- Better decisions in terms of cost and quality are taken by separating query processing from running on operational systems.
- Data warehouses have lead to enhanced asset and liability management since it provides a clear picture of enterprise wide purchasing and inventory patterns thereby indicating otherwise unseen credit exposure and opportunities for cost savings.

1.5.2 Intangible Benefits

Apart from the tangible benfits, data warehouses also provide a number of intangible benefits. Although difficult to quantify, they must also be considered when planning for the data warehouse.

Examples of intangible benefits are:

- Improved productivity that is achieved by keeping all the data in a single location.
- Enhanced customer relations through improved knowledge of individual customer's requirements and trends in the market.
- The information extracted from the data warehouse enables better customer relationship management by tailored product offerings and improved customization.
- Data warehouses enable reengineering of business processes by providing useful insights into the work processes.

1.6 CONCERNS IN DATA WAREHOUSING

- Extracting, cleaning, and loading data are complex, time consuming activities. But tools available in the market can be used to make them easier.
- It is not uncommon for data warehouse projects to go beyond their scope.
- There can be problems of compatibility with the existing systems like the operational systems.
- Providing training to end-users, who may not otherwise use the warehouse at all.
- Security could be a serious bottleneck especially if the data warehouse is web accessible.
- Data warehouse operating and maintenance costs are very high.
- Data warehouses get outdated very quickly, hence there is a risk of delivering suboptimal information to the organization.

1.6.1 Nothing is for Free

No doubt, data warehousing provides a vast range of benefits to its users, but all this comes at a cost. The cost of designing the data model, implementing the data warehouse, addition of extra hardware and ongoing costs that stem from daily data transfer, cleansing and storage of new data entering the warehouse environment can be substantial. Table 1.5 lists the various cost factors involved in moving into this technolology.

Storing large volumes of data has a severe impact on the following:

Cost As the amount of data that has to be stored in the data warehouse goes up so does the cost of storage media. Initially, the data warehouse starts with a small budget but with the increase in the size of the data, the budget allocated for the data warehouse also grows. This cost is required for having a disk with higher storage capacity, disk controller, communication lines, robust operating systems, business intelligence software, etc.

Table 1.5 Costs incurred in deploying a data warehouse

	Recurring Costs	One-time Costs
Capital expenditures	▪ Hardware maintenance ▪ Software maintenance ▪ Middleware technology	▪ Hard Disk ▪ CPU ▪ Network hardware and software ▪ DBMS software ▪ Middleware software
Operational expenditures	▪ Ongoing data refreshing ▪ Integration of data ▪ Data transformation activities ▪ Maintenance of data model ▪ Data archival	▪ Integration of data ▪ Data transformation ▪ Database design ▪ Data model definition ▪ Netwrok related issues ▪ Data dictionaries

Usefulness Initially when organizations start with, say 50 GB data, the probability of all the data being used is quite high. But as the data grows up in size, the percentage of data that is actually used goes down.

Data management When a data warehouse is recently deployed, it has small amounts of data, so data management is not a complexity. But as the data grows in size, the data management activities become more and more complex and take much more time to accomplish. For example, refreshing the data with new values might have taken only an hour when there was a meagre 50 GB of data in the database but now when the size of database has grown to 50 TB, the same activity may take several hours to complete.

Recapitulation

The operational computer systems provide information to run the day-to-day operations, but they cannot be readily used to make strategic decisions.

Data warehousing is a new paradigm specifically intended to provide strategic information.

Data warehouses support decision making and presents flexible, conducive, and interactive source of strategic information to the managers and executives.

A data warehouse is not a single software or hardware product. Rather it is a computing environment where users are put directly in touch with the data they need to make better decisions. It is a user-centric environment.

A data warehouse is a blend of many technologies as it takes data from different operational systems and from outside sources like magazines, journals, reports of other organizations in the same industry; removes inconsistencies, transforms the data, and finally stores it in formats suitable for easy access for decision making.

Data warehouses are meant to be used by executives, mangers, and other people at higher managerial levels who may not have much technical expertise in handling the databases.

Advantages of data warehouses include better decisions, increased productivity, lower operational costs, enhanced asset and liability management, and better CRM.

While implementing a data warehouse in your organization, you need to be careful about extracting, cleaning, and loading of data; checking its compatibility with systems already in place; providing training to end-users and paying special attention to the security of the data.

The data warehouse is used in two basic modes. In the verification mode, the user proposes a hypothesis and asks a series of questions to either confirm or repudiate it. In the discovery mode, the user desires to discover patterns of customer behaviour and relationships among the products that sell together.

Objective Questions

1. Choose the right statements

(a) Operational systems are meant to provide information to run the day-to-day business.

(b) Operational systems are used to make strategic decisions.

(c) Data warehouse stores historical as well as current data.

(d) Historical data is used to study unusual trends in sales.

(e) Data warehouse contains integrated information from heterogeneous sources.

(f) Strategic information is required to run day-to-day operations.

(g) Strategic information is needed for the survival of the corporation in a highly competitive world.

(h) Operational staff needs strategic information.

2. Fill in the blanks

(a) A _____ is a user centric environment.

(b) _____ provides the users with access to accurate, consolidated information from various internal and external sources.

(c) The users of the data warehouse include _____, _____, _____ and _____.

(d) Data warehouse provides _____ data.

3. Multiple choice questions

(a) Reasons for moving into data warehousing include

(i) Processing huge volumes of data
(ii) Providing interactive analysis
(iii) Increase in the computing power
(iv) Lower costs
(v) None of these
(vi) All of these

(b) Characteristics of a data warehouse include

(i) Stores only current data
(ii) Facilitates analyses of large volumes of data
(iii) Data extracted from only a single application
(iv) User friendliness
(v) Contains read intensive data
(vi) Can be updated

(c) Choose the characteristics of an operational system.

(i) Current data
(ii) Optimized for complex queries
(iii) Predictive usage
(iv) 100 MB – 1 GB database size
(v) High access frequency

4. Match the following

(a) Integrated	1. Data in all the tables must be accurate
(b) Data integrity	2. Information must be available within the stipulated time
(c) Accessible	3. Must have an overall enterprise-wide view
(d) Timely	4. Easily accessible by the users with intuitive access paths

5. Categorize as Tangible or Intangible Benefit

(a) Better customer relationship management

(b) Reengineering of business process

(c) Reduction in cost of product introduction

(d) Better decisions in terms of cost and quality

Review Questions

1. What do you understand by strategic information? Give suitable examples. Also write down some of the characteristics of strategic information. For a commercial bank, name five types of strategic objectives.
2. Explain the term Information Crisis.
3. As you have seen, a retail store collects huge amounts of data through its operational systems. Name any four types of transaction data that are likely to be collected by the retail store through its daily operations.
4. Differentiate between operational systems and informational systems.
5. Give reasons why operational systems are not useful for making strategic decisions.
6. Explain the factors which lead to the growth and usage of data warehouses.
7. Data warehouse is an environment, not a product. Comment.
8. Write a short note on benefits of data warehousing.
9. How can you say that data warehousing is a blend of many technologies?
10. Data warehousing is the only viable means to resolve the information crisis and to provide strategic information. Justify the statement.