V Skills
Skills for a secure future

Certified Data Mining and Warehousing Professional Sample Material

V-Skills Certifications

A Government of India
&
Government of NCT Delhi Initiative

V-Skills

Skills for a secure future

1. DATA WAREHOUSING INTRODUCTION

1.1. Introduction

Traditionally, business organizations create billions o f bytes of data about all aspects of business everyday, which contain millions of individual facts about their customers, products, operations, and people. However, this data is locked up and is extremely difficult to get at. Only a small fraction of the data that is captured, processed, and stored in the enterprise is actually available to executives and decision makers.

Recently, new concepts and tools have evolved into a new technology that make it possible to provide all the key people within the enterprise with access to whatever level of information needed for the enterprise to survive and prosper in an increasingly competitive world. The term that is used for this new technology is "data warehousing". In this unit I will be discussing about the basic concept and terminology relating to Data Warehousing.

The Lotus was your first test of "What if "processing on the Desktop. This is what a data warehouse is all about using information your business has gathered to help it react better, smarter, quicker and more efficiently.

1.2. Meaning of Data Warehousing

Data warehouse potential can be magnify if the appropriate data has been collected and stored in a data warehouse. A data warehouse is a relational database management system (RDBMS) designed specifically to meet the needs of transaction processing system. It can be loosely defined as any centralized data repository, which can be queried for business benefit, but this will be more clearly defined letter. Data warehouse is a new powerful technique making. It possible to extract archived operational data and overcome inconsistencies between different legacy data formats, as well as integrating data through-out an enterprise, regardless of location, format, or communication requirements it is possible to incorporate additional or expert information it is.

The logical link between what the managers see in their decision Support EIS application and the company's operational activities Johan McIntyre of SAS institute Inc.

In other words the data warehouse provides warehouse provides data that is already transformed and summarized, therefore making it an appropriate environment for the more efficient DSS and EIS applications.

A data warehouse is a collection of corporate information, derived directly from operational system and some external data sources.

Its specific purpose is to support business decisions, not business ask "What if?" questions. The answer to these questions will ensure your business is proactive, instead of reactive, a necessity in today's information ago.

The industry trend today is moving towards more powerful hardware and software configuration, we now have the ability to process vast volumes o f information analytically, which would have

been unheard of tenor even five years ago. A business today must we able to use this emerging technology or run the risk if being information under loaded. As you read that correctly under loaded - the opposite of over loaded. Over loaded means you are so determine what is important. If you are under loaded, you are information deficient. You cannot cope with decision making expectation because you do not know where you stand. You are missing critical pieces of information required to make informed decisions.

To illustrate the danger o f being information under loaded, consider the children's story of the country mouse is unable to cope with and environment its does not understand.

What is a cat? Is it friend or foe?
Why is the chess in the middle of the floor on the top of a platform with a spring mechanism?

Sensory deprivation and information overload set in. The picture set country mouse cowering in the corner. If is stays there, it will shrivel up and die. The same fate awaits the business that does not respond to or understand the environment around it. The competition will moves in like cultures and exploit all like weaknesses.

In today's world, you do not want to be the country mouse. In today's world, full o f vast amounts o f unfiltered information, a business that does not effectively use technology to shift through that information will not survive the information age Access to, and the understating of, information is power. This power equate to a competitive advantage and survival. This unit
will discuss building own data warehouse-a repository for storing information your business needs to use if it hopes to survive and thrive in the information age. We will help you

understand what a data warehouse is and what it is not. You will learn what human resources are required, as well as the roles and responsibilities of each player. You will be given an overview of good project management techniques to help ensure the data warehouse initiative does not fail due the poor project management. You will learn how to physically implements a data warehouse with some new tools currently available to help you mine those vast amounts of information stored within the warehouse. Without fine running this ability to mine the warehouse, even the most complete warehouse, would be useless.

1.3. History of Data Warehousing

Let us first review the historical management schemes of the analysis data and the factors that have led to the evolution of the data warehousing application class.

Traditional Approaches to Historical Data

Throughout the history of systems development, the primary emphasis had been given to the operational systems and the data they process. It was not practical to keep data in the operational systems indefinitely; and only as an afterthought was a structure designed for archiving the data that the operational system has processed. The fundamental requirements of the operational and analysis systems are different: the operational systems need performance, whereas the analysis systems need flexibility and broad scope.

Data from Legacy Systems

Different platforms have been developed with the development of the computer systems over past three decades. In the 1970's, business system development was done on the IBM mainframe computers using tools such as Cobol, CICS, IMS, DB2, etc. With the advent of 1980's computer platforms such as AS/400 and VAX/VMS were developed. In late eighties and early nineties UNIX had become a popular server platform introducing the client/server architecture which remains popular till date. Despite all the changes in the platforms, architectures, tools, and technologies, a large number of business applications continue to run in the mainframe environment o f the 1970's. The most important reason is that over the years these systems have captured the business knowledge and rules that are incredibly difficult to carry to a new platform or application. These systems are, generically called legacy systems. The data stored in such systems ultimately becomes remote and becomes difficult to get at.

Extracted Information on the Desktop

During the past decade the personal computer has become very popular for business analysis. Business Analysts now have many of the tools required to use spreadsheets for analysis and graphic representation. Advanced users will frequently use desktop database programs to store and work with the information extracted from the legacy sources. The disadvantage of the above is that it leaves the data fragmented and oriented towards very specific needs. Each individual user has obtained only the information that she/he requires. The extracts are unable to address the requirements of multiple users and uses. The time and cost involved in addressing the requirements of only one user are large. Due to the disadvantages faced it led to the development of the new application called **Data Warehousing**

Factors, which Lead To Data Warehousing

Many factors have influenced the quick evolution of the data warehousing discipline. The most important factor has been the advancement in the hardware and software technologies. Hardware and Software prices: Software and hardware prices have fallen to a great extent. Higher capacity memory chips are available at very low prices.

- ✓ **Powerful Preprocessors:** Today's preprocessor are many times powerful than yesterday's mainframes: e.g. Pentium III and Alpha processors
- ✓ **Inexpensive disks:** The hard disks of today can store hundreds of gigabytes with their prices falling. The amount of information that can be stored on just a single one-inch high disk drive would have required a roomful of disk drives in 1970's and early eighties.
- ✓ **Desktop powerful for analysis tools:** Easy to use GUI interfaces, client/server architecture or multi-tier computing can be done on the desktops as opposed to the mainframe
- ✓ computers of yesterday.
- ✓ **Server software:** Server software is inexpensive, powerful, and easy to maintain as compared to that of the past. Example of this is Windows NT that have made setup of powerful systems very easy as well as reduced the cost. The skyrocketing power of hardware and software, along with the availability of affordable and easy-to-use reporting and analysis tools have played the most important role in evolution of data warehouses.

Emergence of Standard Business Applications

New vendors provide to end-users with popular business application suites. German software vendor SAP AG, Baan, PeopleSoft, and Oracle have come out with suites of software that provide

different strengths but have comparable functionality. These application suites provide standard applications that can replace the existing custom developed legacy applications. This has led to the increase in popularity of such applications. Also, the data acquisition from these applications is much simpler than the mainframes.

End-user more Technology Oriented

One of the most important results of the massive investment in technology and movement towards the powerful personal computer has been the evolution of a technology-oriented business analyst. Even though the technology-oriented end users are not always beneficial to all projects, this trend certainly has produced a crop of technology-leading business analysts that are becoming essential to today's business. These technology-oriented end users have frequently played an important role in the development and deployment of data warehouses. They have become the core users that are first to demonstrate the initial benefits of data warehouses. These end users are also critical to the development of the data warehouse model: as they become experts with the data warehousing system, they train other users.