LESSON 1
INTRODUCTION TO
DATA WAREHOUSING

Structure

- Objective
- Introduction
- Meaning of Data warehousing
- History of Data warehousing
- Traditional Approaches To Historical Data
- Data from legacy systems
- Extracted information on the Desktop
- Factors, which Lead To Data Warehousing

Objective

The main objective of this lesson is to introduce you with the basic concept and terminology relating to Data Warehousing. By the end of this lesson you will be able to understand:

- Meaning of a Data warehouse
- Evolution of Data warehouse

Introduction

Traditionally, business organizations create billions of 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.

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 over come inconsistencies between different legacy data formats, as well as integrating data throughout 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 of 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 of 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 of vast amounts of 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 dose 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 with in the warehouse. Without fine running this ability to mine the warehouse, even the most complete warehouse, would be useless.

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 of 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.

LESSON 2
MEANING AND CHARACTERISTICS OF DATA WAREHOUSING

Structure

- Objective
- Introduction
- Data warehousing
- Operational vs. Informational Systems
- Characteristics of Data warehousing
- Subject oriented
- Integrated
- Time variant
- Non-volatiles

Objective

The objective of this lesson is to explain you the significance and difference between Operational systems and Informational systems. This lesson also includes various characteristics of a Data warehouse.

Introduction

In the previous section, we have discussed about the need of data warehousing and the factors that lead to it. In this section I will explore the technical concepts relating to data warehousing to you.

A company can have data items that are unrelated to each other. Data warehousing is the process of collecting together such data items within a company and placing it in an integrated data store. This integration is over time, geographies, and application platforms. By adding access methods (on-line querying, reporting), this converts a 'dead' data store into a 'dynamic' source of information. In other words, turning a liability into an asset. Some of the definitions of data warehousing are:

"A data warehouse is a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use in a business context." (Devlin 1997)

"Data warehousing is a process, not a product, for assembling and managing data from various sources for the purpose of gaining a single, detailed view of part or all of the business." (Gardner 1998)

A Data Warehouse is a capability that provides comprehensive and high integrity data in forms suitable for decision support to end users and decision makers throughout the organization. A data warehouse is managed data situated after and outside the operational systems. A complete definition requires discussion of many key attributes of a data warehouse system Data Warehousing has been the result of the repeated attempts of various researchers and organizations to provide their organizations flexible, effective and efficient means of getting at the valuable sets of data.

Data warehousing evolved with the integration of a number of different technologies and experiences over the last two decades, which have led to the identification of key problems.

Data Warehousing

Because data warehouses have been developed in numerous organizations to meet partic-ular needs, there is no single, canonical definition of the term data warehouse.[1] Profes-sional magazine articles and books in the popular press have elaborated on the meaning in a variety of ways. Vendors have capitalized on the popularity of the term to help mar-ket a variety of related products, and consultants have provided a large variety of services, all under the data-warehousing banner. However, data warehouses are quite distinct from traditional databases in their structure, functioning, performance, and purpose.

Operational vs. Informational Systems

Perhaps the most important concepts that has come out of the Data Warehouse movement is the recognition that there are two fundamentally different types of information systems in all organizations: operational systems and informational systems.

"Operational systems" are just what their name implies, they are the systems that help us run the enterprise operate day-to-day. These are the backbone systems of any enterprise, our "order entry', "inventory", "manufacturing", "payroll" and "accounting" systems. Because of their importance to the organization, operational systems were almost always the first parts of the enterprise to be computerized. Over the years, these operational systems have been extended and rewritten, enhanced and maintained to the point that they are completely integrated into the organization. Indeed, most large organizations around the world today couldn't operate without their operational systems and that data that these systems maintain.

On the other hand, there are other functions that go on within the enterprise that have to do with planning, forecasting and managing the organization. These functions are also critical to the survival of the organization, especially in our current fast paced world. Functions like "marketing planning", "engineering planning" and "financial analysis" also require information systems to support them. But these functions are different from operational ones, and the types of systems and information required are also different. The knowledge-based functions are informational systems.

"Informational systems" have to do with analyzing data and making decisions, often major decisions about how the enterprise will operate, now and in the future. And not only do informational systems have a different focus from operational ones, they often have a different scope. Where operational data needs are normally focused upon a single area, informational data needs often span a number of different areas and need large amounts of related operational data.

In the last few years, Data Warehousing has grown rapidly from a set of related ideas into architecture for data delivery for enterprise end user computing.

They support high-performance demands on an organization's data and information. Several types of applications-OLAP, DSS, and data mining applications-are supported. OLAP (on-line analytical processing) is a term used to describe the analysis of complex data from the data warehouse. In the hands of skilled knowledge workers. OLAP tools use distributed computing capabilities for analyses that require more storage and processing power than can be economically and efficiently located on an individual desktop. DSS (Decision-Support Systems) also known as EIS (Executive Information Systems, not to be confused with enterprise integration systems) support an organization's leading deci-sion makers with higher-level data for complex and important decisions. Data mining is used for knowledge discovery, the pro-cess of searching data for unanticipated new knowledge.

Traditional databases support On-Line Transaction Processing (OLTP), which includes insertions, updates, and deletions, while also supporting information query requirements. Traditional relational databases are optimized to process queries that may touch a small part of the database and transactions that deal with insertions or updates of a few tuples per relation to process. Thus, they cannot be optimized for OLAP, DSS, or data mining. By contrast, data warehouses are designed precisely to support efficient extraction, process-ing, and presentation for analytic and decision-making purposes. In comparison to traditional databases, data warehouses generally contain very large amounts of data from multiple sources that may include databases from different data models and sometimes lies acquired from independent systems and platforms.

A database is a collection of related data and a database system is a database and database software together. A data warehouse is also a collection of information as well as supporting system. However, a clear distinction exists, Traditional databases are transactional: relational, object-oriented, network, or hierarchical. Data warehouses have the distinguishing characteristic that they are mainly intended for decision-support applications. They are optimized for data retrieval, not routine transaction processing.

Characteristics of Data Warehousing

As per W. H. Inmon, author of building the data warehouse and the guru who is ready widely considered to be the originator of the data warehousing concept, there are generally four character that describe a data warehouse:

W. H. Inmon characterized a data warehouse as "a subject-oriented, integrated, nonvola-tile, time-variant collection of data in support of management's decisions." Data ware-houses provide access to data for complex analysis, knowledge discovery, and decision-making.

Subject Oriented

Data are organized according to subject instead of application e.g. an insurance company using a data warehouse would organize their data by costumer, premium, and claim, instead of by different products (auto. Life etc.). The data organized by subject contain only the information necessary for decision support processing.

Integrated

When data resides in money separate applications in the operational environment, encoding of data is often inconsistent. For instance in one application, gender might be coded as "m" and "f" in another by o and l. When data are moved from the operational environment in to the data warehouse, when data are moved from the operational environment in to the data warehouse, they assume a consistent coding convention e.g. gender data is transformed to "m" and "f".

Time variant

The data warehouse contains a place for storing data that are five to ten years old, or older, to be used for comparisons, trends, and forecasting. These data are not up dated.

Non-volatile

Data are not update or changed in any way once they enter the data warehouse, but are only loaded and accessed.

Data warehouses have the following distinctive characteristics.

- Multidimensional conceptual view.
- Generic dimensionality.
- Unlimited dimensions and aggregation levels.
- Unrestricted cross-dimensional operations.
- Dynamic sparse matrix handling.
- Client-server architecture.
- Multi-user support.
- Accessibility.
- Transparency.
- Intuitive data manipulation.
- Consistent reporting performance.
- Flexible reporting

Because they encompass large volumes of data, data warehouses are generally an order of magnitude (sometimes two orders of magnitude) larger than the source databases. The sheer volume of data (likely to be in terabytes) is an issue that has been dealt with through enterprise-wide data warehouses, virtual data warehouses, and data marts:

- Enterprise-wide data warehouses are huge projects requiring massive investment of time and resources.
- Virtual data warehouses provide views of operational databases that are materialized for efficient access.
- Data marts generally are targeted to a subset of the organization, such as a dependent, and are more tightly focused.

To summarize the above, here are some important points to remember about various characteristics of a Data warehouse:

- **Subject-oriented**
  - Organized around major subjects, such as customer, product, sales.

LESSON 3
ONLINE TRANSACTION PROCESSING

Structure

- Objective
- Introduction
- Data warehousing and OLTP systems
- Similarities and Differences in OLTP and Data Warehousing Processes in Data Warehousing OLTP
- What is OLAP?
- Who uses OLAP and WHY?
- Multi-Dimensional Views
- Benefits of OLAP

Objective

The main objective of this lesson is to introduce you with Online Transaction Processing. You will learn about the importance and advantages of an OLTP system.

Introduction

Relational databases are used in the areas of operations and control with emphasis on transaction processing. Recently relational databases are used for building data warehouses, which stores tactical information (<1year into the future) that answers who and what questions. In contrast OLAP uses MD views of aggregate data to provide access strategic information. OLAP enables users to gain insight to a wide variety of possible views of information and transforms raw data to reflect the enterprise as understood by the user e.g., Analysts, managers and executives.

Data Warehousing and OLTP Systems

A data base which in built for on line transaction processing, OLTP, is generally regarded as inappropriate for warehousing as they have been designed with a different set of need in mind i.e., maximizing transaction capacity and typically having hundreds of table in order not to look out user etc. Data warehouse are interested in query processing as opposed to transaction processing.

OLTP systems cannot be receptacle stored of repositories of facts and historical data for business analysis. They cannot be quickly answer adhoc queries is rapid retrieval is almost impossible. The data is inconsistent and changing, duplicate entries exist, entries can be missing and there is an absence of historical data, which is necessary to analyses trends. Basically OLTP offers large amounts of raw data, which is not easily understood. The data warehouse offers the potential to retrieve and analysis information quickly and easily. Data warehouse do have similarities with OLTP as shown in the table below.

Similarities and Differences in OLTP and Data Warehousing

	OLTP	Data Warehouse
Purpose	Run day-to-day operation	Information retrieval and analysis
Structure	RDBMS	RDBMS
Data Model	Normalized	Multi-dimensional
Access	SQL	SQL plus data analysis extensions
Type of Data	Data that run the business	Data that analyses the business
Condition of Data	Changing incomplete	Historical descriptive

The data warehouse server a different purpose from that of OLTP systems by allowing business analysis queries to be answered as opposed to "simple aggregation" such as 'what is the current account balance for this customer?' Typical data warehouse queries include such things as 'which product line sells best in middle America and how dose this correlate to demographic data?

Processes in Data Warehousing OLTP

The first step in data warehousing is to "insulate" your current operational information, i.e. to preserve the security and integrity of mission- critical OLTP applications, while giving you access to the broadest possible base of data. The resulting database or data warehouse may consume hundred of gigabytes-or even terabytes of disk space, what is required than are capable efficient techniques for storing and retrieving massive amounts of information. Increasingly, large organizations have found that only parallel processing systems offer sufficient bandwidth.

The data warehouse thus retrieves data from a varsity of heterogeneous operational database. The data is than transformed and delivered to the data warehouse/ store based in a selected modal (or mapping definition). The data transformation and movement processes are completed whenever an update to the warehouse data is required so there should some from of automation to manage and carry out these functions. The information that describes the modal metadata is the means by which the end user finds and understands the data in the warehouse and is an important part of the warehouse. The metadata should at the very least contain:

- Structure of the data;
- Algorithm used for summarization;

- Mapping from the operational environment to the data warehouse.

Data cleansing is an important viewpoint of creating an efficient data warehouse of creating an efficient data warehouse in that is the removal of creation aspects Operational data such as low level transaction information which sloe down the query times. The cleansing stage has to be as dynamic as possible to accommodate all types of queries even those, which may require low-level information. Data should be extracted from production sources at regular interval and pooled centrally but the cleansing process has to remove duplication and reconcile differences between various styles of data collection.

Once the data has been cleaned it is than transfer to the data warehouse, which typically is a large database on a high performance box, either SMP Symmetric Multi- Processing or MPP, Massively parallel Processing Number crunching power is another importance aspect of data warehousing because of the complexity involved in processing adhoc queries and because of the vast quantities of data that the organization want to use in the warehouse. A data warehouse can be used in different ways, for example it can be a central store against which the queries are run of it can be used like a data mart, data mart which are small warehouses can be established to provide subsets of the main store and summarized information depending on the requirements of a specific group/ department. The central stores approach generally uses every simple data structures with very little assumptions about the relationships between data where as marts often uses multidimensional data base which can speed up query processing as they can have data structures which reflect the most likely questions.

Many vendors have products that provide on the more of the above data warehouse functions. However, it can take a significant amount of work and specialized programming to provide the interoperability needed between products form. Multiple vendors to enable them to perform the required data warehouse processes a typical implementation usually involves a mixture of procedure forma verity of suppliers.

Another approach to data warehousing is the Parsaye Sandwich paradigm put forward by Dr. Kamran Parsaye , CED of information discovery, Hermosa beach. This paradigm or philosophy encourages acceptance of the probability that the first iteration of data warehousing effort will require considerable revision. The Sandwich paradigm advocates the following approach.

- Pre-mine the data to determine what formats and data are needed to support a data- mining application;
- Build a prototype mini- data warehouse i.e. the, the meat of sandwich most of features envisaged for the and product;
- Revise the strategies as necessary;
- Build the final warehouse.

What is OLAP?

- Relational databases are used in the areas of operations and control with emphasis on transaction processing.
- Recently relational databases are used for building data warehouses, which stores tactical information (< 1 year into the future) that answers who and what questions.
- In contrast OLAP uses Multi-Dimensional (MD) views of aggregate data to provide access strategic information.
- OLAP enables users to gain insight to a wide variety of possible views of information and transforms raw data to reflect the enterprise as understood by the user e.g. Analysts, managers and executives.
- In addition to answering who and what questions OLAPs can answer "what if " and "why".
- Thus OLAP enables strategic decision-making.
- OLAP calculations are more complex than simply summing data.
- However, OLAP and Data Warehouses are complementary
- The data warehouse stores and manages data while the OLAP transforms this data into strategic information.

Who uses OLAP and WHY?

- OLAP applications are used by a variety of the functions of an organisation.
- Finance and accounting:

  Budgeting

  Activity-based costing

  Financial performance analysis

  And financial modelling
- Sales and Marketing

  Sales analysis and forecasting

  Market research analysis

  Promotion analysis

  Customer analysis

  Market and customer segmentation
- Production

  Production planning

  Defect analysis

Thus, OLAP must provide managers with the information they need for effective decision-making. The KPI (key performance indicator) of an OLAP application is to provide just-in-time (JIT) information for effective decision-making. JIT information reflects complex data relationships and is calculated on the fly. Such an approach is only practical if the response times are always short The data model must be flexible and respond to changing business requirements as needed for effective decision making.

In order to achieve this in widely divergent functional areas OLAP applications all require:

MD views of data

Complex calculation capabilities

Time intelligence

Multi-Dimensional Views

- MD views inherently represent actual business models, which normally have more than three dimensions e.g., Sales data is looked at by product, geography, channel and time.

LESSON 5
ARCHITECTURE AND PRINCIPLES OF DATA WAREHOUSING

Structure

- Objective
- Introduction
- Structure of a Data warehouse
- Data Warehouse Physical Architectures
- Generic Two-Level
- Expanded Three-Level
- Enterprise data warehouse (EDW)
- Data marts
- Principles of a Data warehousing

Objective

The objective of this lesson is to let you know the basic structure of a Data warehouse. You will also learn about Data warehouse physical architecture and various principles of a Data warehousing.

Introduction

Let me start the lesson with an example, which illustrates the importance and need of a data warehouse. Until several years ago Coca Cola had no idea how many bottles of Coke it produced each day because production data were stored on 24 different computer systems. Then, it began a technique called Data warehousing. One airline spent and wasted over $100 million each year on inefficient mass media advertising campaigns to reach frequent flyers…then it began data warehousing. Several years ago, the rail industry needed 8 working days to deliver a freight quote to a customer. The trucking industry, by contrast, could deliver a freight quote to a customer on the phone instantly, because unlike the rail industry, truckers were using…data warehousing.

A data warehouse is a data base that collects current information, transforms it to ways it can be used by the warehouse owner, transforms that information for clients, and offers portals of access to members of your firm to help them make decisions and future plans.

Data warehousing is the technology trend most often associated with enterprise computing today. The term conjures up images of vast data banks fed from systems all over the globe, with legions of corporate analysts mining them for golden nuggets of information that will make their companies more profitable.

All of the developments in database technology over the past 20 years have culminated in the data warehouse. Entity-relationship modeling, heuristic searches, mass data storage, neural networks, multiprocessing, and natural-language interfaces have all found their niches in the data warehouse. But aside from being a database engineer's dream, what practical benefits does a data warehouse offer the enterprise?

When asked, corporate executives often say that having a data warehouse gives them a competitive advantage, because it gives them a better understanding of their data and a better understanding of their business in relation to their competitors, and it lets them provide better customer service.

So, what exactly is a data warehouse? Should your company have one, and if so, what should it look like?

Structure of a Data Warehouse

Essentially, a data warehouse provides historical data for decision-support applications. Such applications include reporting, online analytical processing (OLAP), executive information systems (EIS), and data mining.

According to W. H. Inmon, the man who originally came up with the term, a data warehouse is a centralized, integrated repository of information. Here, integrated means cleaned up, merged, and redesigned. This may be more or less complicated depending on how many systems feed into a warehouse and how widely they differ in handling similar information.

But most companies already have repositories of information in their production systems and many of them are centralized. Aren't these data warehouses? Not really.

Data warehouses differ from production databases, or online transaction-processing (OLTP) systems, in their purpose and design. An OLTP system is designed and optimized for data entry and updates, whereas a data warehouse is optimized for data retrieval and reporting, and it is usually a read-only system. An OLTP system contains data needed for running the day-to-day operations of a business but a data warehouse contains data used for analyzing the business. The data in an OLPT system is current and highly volatile, which data elements that may be incomplete or unknown at the time of entry. A warehouse contains historical, nonvolatile data that has been adjusted for transactions errors. Finally, since their purposes are so different, OLPT systems and data warehouses use different data-modeling strategies. Redundancy is almost nonexistent in OLTP systems, since redundant data complicates updates. So OLPT systems are highly normalized and are usually based on a relational model. But redundancy is desirable in a data warehouse, since it simplifies user access and enhances performance by minimizing the number of tables that have to be joined. Some data warehouses don't use a relational model at all, preferring a multidimensional design instead.

To discuss data warehouses and distinguish them from transactional databases calls for an appropriate data model. The multidimensional data model is a good fit for OLAP and decision-support technologies. In contrast to multi-databases, which provide access to disjoint and usually heterogeneous databases, a data warehouse is frequently a store of integrated data from multiple sources, processed for storage in a multidimensional model. Unlike most transactional databases, data warehouses typically support time-series and trend analysis, both of which requires more historical data than are generally

LESSON 6
DATA WAREHOUSING AND OPERATIONAL SYSTEMS

Structure

- Objective
- Introduction
- Operational Systems
- Warehousing" data outside the operational systems
- Integrating data from more than one operational system
- Differences between transaction and analysis processes
- Data is mostly non-volatile
- Data saved for longer periods than in transaction systems
- Logical transformation of operational data
- Structured extensible data model
- Data warehouse model aligns with the business structure
- Transformation of the operational state information

Objective

The aim of this lesson is to explain you the need and importance of an Operational Systems in a Data warehouse.

Introduction

Data warehouse, a collection of data designed to support management decision-making. Data warehouses contain a wide variety of data that present a coherent picture of business conditions at a single point in time.

Development of a data warehouse includes development of systems to extract data from operating systems plus installation of a warehouse database system that provides managers flexible access to the data.

The term data warehousing generally refers to the combination of many different databases across an entire enterprise.

Operational Systems

Up to now, the early database system of the primary purpose was to, meet the needs of operational systems, which are typically transactional in nature. Classic examples of operational systems include

- General Ledgers
- Accounts Payable
- Financial Management
- Order Processing
- Order Entry
- Inventory

Operational systems by nature are primarily concerned with the handling of a single transaction. Look at a banking system, when you, the customer, make a deposit to your checking account, the banking operational system is responsible for recording the transaction to ensure the corresponding debit appears in your account record.

A typical operational system deals with one order, one account, one inventory item. An operational system typically dea1s with predefined events and, due to the nature of these events, requires fast access. Each transaction usually deals with small amounts of data.

Most of the time, the business needs of an operational system do not change much. The application that records the transaction, as well as the application that controls access to the information, that is, there porting side of the- banking business does not change much over time. In this type of system, the information required, when a customer initiates a transaction must be current. Before a bank will allow a withdrawal, it must first be certain of your current balance.

"Warehousing" Data outside the Operational Systems

The primary concept of data warehousing is that the data stored for business analysis can most effectively be accessed by separating it from the data in the operational systems. Many of the reasons for this separation have evolved over the years. In the past, legacy systems archived data onto tapes as it became inactive and many analysis reports ran from these tapes or mirror data sources to minimize the performance impact on the operational systems.

These reasons to separate the operational data from analysis data have not significantly changed with the evolution of the data warehousing systems, except that now they are considered more formally during the data warehouse building process. Advances in technology and changes in the nature of business have made many of the business analysis processes much more complex and sophisticated. In addition to producing standard reports, today's data warehousing systems support very sophisticated online analysis including multi-dimensional analysis.

Integrating Data from more than one Operational System

Data warehousing systems are most successful when data can be combined from more than one operational system. When the data needs to be brought together from more than one source application, it is natural that this integration be done at a place independent of the source applications. Before the evolution of structured data warehouses, analysts in many instances would combine data extracted from more than one operational system into a single spreadsheet or a database. The data warehouse may very effectively combine data from multiple source applications such as sales, marketing, finance, and production. Many large data warehouse architectures allow for the source applications to be integrated into the data warehouse incrementally.

The primary reason for combining data from multiple source applications is the ability to cross-reference data from these

applications. Nearly all data in a typical data warehouse is built around the time dimension. Time is the primary filtering criterion for a very large percentage of all activity against the data warehouse. An analyst may generate queries for a given week, month, quarter, or a year. Another popular query in many data warehousing applications is the review of year-on-year activity. For example, one may compare sales for the first quarter of this year with the sales for first quarter of the prior years. The time dimension in the data warehouse also serves as a fundamental cross-referencing attribute. For example, an analyst may attempt to access the impact of a new marketing campaign run during selected months by reviewing the sales during the same periods. The ability to establish and understand the correlation between activities of different organizational groups within a company is often cited as the single biggest advanced feature of the data warehousing systems.

The data warehouse system can serve not only as an effective platform to merge data from multiple current applications; it can also integrate multiple versions of the same application. For example, an organization may have migrated to a new standard business application that replaces an old mainframe-based, custom-developed legacy application. The data warehouse system can serve as a very powerful and much needed platform to combine the data from the old and the new applications. Designed properly, the data warehouse can allow for year-on-year analysis even though the base operational application has changed.

Differences between Transaction and Analysis Processes

The most important reason for separating data for business analysis from the operational data has always been the potential performance degradation on the operational system that can result from the analysis processes. High performance and quick response time is almost universally critical for operational systems. The loss of efficiency and the costs incurred with slower responses on the predefined transactions are usually easy to calculate and measure. For example, a loss of five seconds of processing time is perhaps negligible in and of itself; but it compounds out to considerably more time and high costs once all the other operations it impacts are brought into the picture. On the other hand, business analysis processes in a data warehouse are difficult to predefine and they rarely need to have rigid response time requirements.

Operational systems are designed for acceptable performance for pre-defined transactions. For an operational system, it is typically possible to identify the mix of business transaction types in a given time frame including the peak loads. It also relatively easy to specify the maximum acceptable response time given a specific load on the system. The cost of a long response time can then be computed by considering factors such as the cost of operators, telecommunication costs, and the cost of any lost business. For example, an order processing system might specify the number of active order takers and the average number of orders for each operational hour. Even the query and reporting transactions against the operational system are most likely to be predefined with predictable volume.

Even though many of the queries and reports that are run against a data warehouse are predefined, it is nearly impossible to accurately predict the activity against a data warehouse. The process of data exploration in a data warehouse takes a business analyst through previously undefined paths. It is also common to have runaway queries in a data warehouse that are triggered by unexpected results or by users' lack of understanding of the data model. Further, many of the analysis processes tend to be all encompassing whereas the operational processes are well segmented. A user may decide to explore detail data while reviewing the results of a report from the summary tables. After finding some interesting sales activity in a particular month, the user may join the activity for this month with the marketing programs that were run during that particular month to further understand the sales. Of course, there would be instances where a user attempts to run a query that will try to build a temporary table that is a Cartesian product of two tables containing a million rows each! While an activity like this would unacceptably degrade an operational system's performance, it is expected and planned for in a data warehousing system.

Data is mostly Non-volatile

Another key attribute of the data in a data warehouse system is that the data is brought to the warehouse after it has become mostly non-volatile. This means that after the data is in the data warehouse, there are no modifications to be made to this information. For example, the order status does not change, the inventory snapshot does not change, and the marketing promotion details do not change. This attribute of the data warehouse has many very important implications for the kind of data that is brought to the data warehouse and the timing of the data transfer.

Let us further review what it means for the data to be non-volatile. In an operational system the data entities go through many attribute changes. For example, an order may go through many statuses before it is completed. Or, a product moving through the assembly line has many processes applied to it. Generally speaking, the data from an operational system is triggered to go to the data warehouse when most of the activity on these business entity data has been completed. This may mean completion of an order or final assembly of an accepted product. Once an order is completed and shipped, it is unlikely to go back to backorder status. Or, once a product is built and accepted, it is unlikely to go back to the first assembly station. Another important example can be the constantly changing data that is transferred to the data warehouse one snapshot at a time. The inventory module in an operational system may change with nearly every transaction; it is impossible to carry all of these changes to the data warehouse. You may determine that a snapshot of inventory carried once every week to the data warehouse is adequate for all analysis. Such snapshot data naturally is non-volatile.

It is important to realize that once data is brought to the data warehouse, it should be modified only on rare occasions. It is very difficult, if not impossible, to maintain dynamic data in the data warehouse. Many data warehousing projects have failed miserably when they attempted to synchronize volatile data between the operational and data warehousing systems.