BIT 324 DATA WAREHOUSING AND MINING

ACADEMIC YEAR: 2017/2018 **SEMESTER:** II

LECTURER: ROSELIDA MAROKO ONGARE

CONTACT: **Mobile:** 0721597710 **Email Address:** rongare@kibu.ac.ke

DAY: TUESDAY **TIME:** **ROOM:**

AIM/PURPOSE: The purpose of this course is to introduce the core concepts of data warehousing and mining, their techniques, implementation, benefits, and outcome expectations from this new technology

PRE-REQUISITE: Database Systems

COURSE HOURS PER WEEK: 3

EXPECTED LEARNING OUTCOMES

Upon completion of this course the student should be able to:

i. Understand the fundamental concepts of data warehousing, OLAP and data mining
ii. Demonstrate understanding of data warehouse, data mining algorithms, methods, and tools
iii. Be able to efficiently design and manage data storages using data warehousing
iv. Select and apply appropriate data mining techniques for different real life applications.

COURSE CONTENT

Week	Activity	Assessment
1.	Data Warehousing Concepts • Need for data warehousing • Basic elements of data warehousing • Trends in data warehousing	Review Questions
2.	Architecture and Infrastructure • Architectural components • Infrastructure and metadata	Review Questions
3.	Data Design and Data Representation • Principles of dimensional modelling • Dimensional modelling advanced topics • Data extraction • Transformation and loading • Data quality	Review Questions
4.	Information Access and Delivery • Matching information to classes of users • OLAP in data warehouse • Data warehousing and the web	Review Questions

5.	Data mining
	• Data mining basics
	  ○ The Knowledge Discovery Process
	  ○ OLAP Versus Data Mining
		Review Questions

Week	Activity	Assessment
	○ Data Mining and the Data Warehouse	
6.	CAT 1	
7.	• Data mining techniques ○ Cluster Detection ○ Decision Trees ○ Memory-Based Reasoning ○ Link Analysis ○ Neural Networks ○ Genetic Algorithms ○ Moving into Data Mining ○ Data mining primitives, languages and systems • Discovery and analysis of patterns ○ Trends and deviations • Data mining applications ○ Benefits of Data Mining ○ Applications in Retail Industry ○ Applications in Telecommunications Industry ○ Applications in Banking and Finance	Review Questions
8.	Descriptive data mining • Characterization and comparison • Association analysis • Classification and prediction • Cluster analysis ○ Clustering ○ Enabling data mining through data warehouse. • Knowledge Discovery • KDD Process	Review Questions
9.	Implementation and Maintenance • Physical design process • Data warehouse deployment • Growth and maintenance	Review Questions
10.	Web Mining • Web Content Mining • Web Structure Mining • Web Usage mining	Review Questions
11.	CAT 2	Review Questions
12.	Data marts • Multidimensional databases • Mining complex types of data • Applications and trends in data mining	Review Questions
13.	Revision	Sample Exam Questions
14.	Revision	Sample Exam Questions

MODE OF DELIVERY
Lectures, tutorials, group discussions, research and presentation

INSTRUCTIONAL MATERIALS AND/OR EQUIPMENT
Audio visual equipment, whiteboard

COURSE ASSESSMENT

Exam	70%
CATs	30%
TOTAL	**100%**

CORE REFERENCES

Berson, A., Smith, S. J. Smith, (1997), Data Warehousing, Data Mining, and OLAP, McGraw-Hill, Inc. New York, NY, USA

Kimball, R., Ross, M., (2013), The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modelling 3rd Edition

Ponniah, P., (2001), Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals. John Wiley & Sons, Inc.

Approval

----------	**----------**
Lecturer/Instructor	**Date**
----------	**----------**
COD Information Technology	**Date**

Concepts of Data Warehouse

Why Data Warehousing?

- In today's competitive business environment, large companies have operations in many places within their home country and even other parts of the world. Each of their branch offices may generate huge volumes of data on a daily basis, and corporate decision makers require access to those data sources. This is where data warehouses come in. The data warehouse is one of the most important business intelligence tools a business needs to have. Data warehousing turns the massive amount of data generated from multiple sources into a format that is easy to understand.
- Data Warehousing & OLAP are important to businesses because they help the decision makers discover information **hidden** within the organization's data.
  - See data from different angles: **product**, **client**, **time**, **area**
  - Get adequate **statistics** to get your point of argumentation across
  - Get a glimpse of the future

Data Warehouses

- The term Data Warehouse was first invented by Bill Inmom in 1990. A Data Warehouse is always kept separate from an operational Database.
- The data warehouse is an informational environment that:
  - Provides an integrated and total view of the enterprise
  - Makes the enterprise's current and historical information easily available for decision making
  - Makes decision-support transactions possible without hindering operational systems
  - Renders the organization's information consistent
  - Presents a flexible and interactive source of strategic information

Definitions of Data Warehousing & Data Warehouse

Data Warehousing

Data warehousing is a technology that aggregates structured data from one or more sources so that it can be compared and analyzed for greater business intelligence.

Data Warehouse

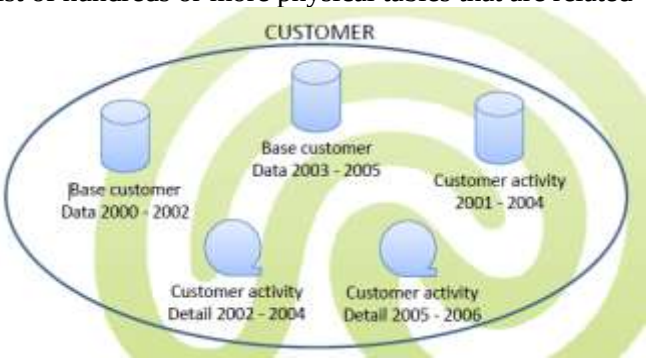
- A Data Warehouse consists of data from **multiple heterogeneous data sources** and is used for analytical reporting and decision making. Many authors have given different definitions of data warehouse
  - Data Warehouse is a central place where data is stored from different data sources and applications.
  - A data warehouse is a centralized store of all data generated by the departments of a large organization. It is specially designed for data analysis, generating reports, and for other ad-hoc queries. A data warehouse extracts the huge streams of data from a company's operational and external databases and turns them into meaningful data, so business decisions can be made based on this information.
  - A data warehouse is a centralized repository that stores data from multiple information sources and transforms them into a common, multidimensional data model for efficient querying and analysis.
  - A data warehouse is a large repository of some organization's electronically stored data specifically designed to facilitate reporting and analysis

- A data warehouse is a subject-oriented, integrated, time-varying, non-volatile collection of data in support of the management's decision-making process.

Subject Oriented

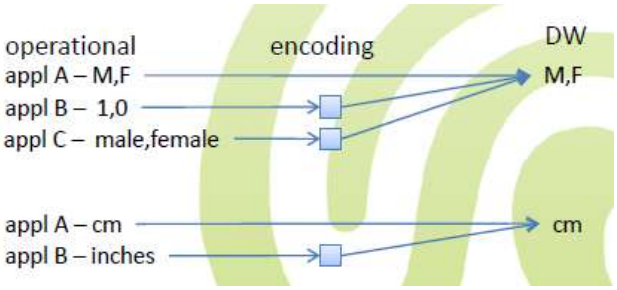
The data in the DW is organized in such a way that all the data elements relating to the same real-world event or object are linked together.

- Typical subject areas in DWs are Customer, Product, Order, Claim, Account etc
- Example: customer as central subject in some DW
    - The complete DW is organized by customer
    - It may consist of hundreds or more physical tables that are related



Integrated

The DW contains data from most or all the organization's operational systems and this data is made consistent. E.g. gender, measurement, conflicting keys.



Non-volatile

Data in the DW is hardly ever over-written or deleted - once committed, the data is static, read-only, and retained for future reporting

- Data is loaded, but not updated
- When subsequent changes occur, a new version or snapshot record is written

Time-varying

The changes to the data in the DW are tracked and recorded so that reports show changes over time. Different environments have different time horizons associated. While for operational systems a 60-to-90 day time horizon is normal, DWs have a 5-to-10 year horizon

The concept of a data warehouse to take all the data you already have in the organization, clean and transform it, and then provide useful strategic information. It is not to generate fresh data, but to make use of the large volumes of existing data and to transform it into forms suitable for providing strategic information.

A Data Warehouse is an Environment Not a Product

A data warehouse is not a single software or hardware product you purchase to provide strategic information. It is a computing environment where users can find strategic information, an 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 Blend of Many Technologies

- Data warehousing is a blend of technologies. It involves different functions: data extraction, the function of loading the data, transforming the data, storing the data, and providing user interfaces. This includes:
  - Taking all the data from the operational systems
  - Where necessary, including relevant data from outside, such as industry benchmark indicators
  - Integrating all the data from the various sources
  - Removing inconsistencies and transforming the data
  - Storing the data in formats suitable for easy access for decision making

Accessing data in a data warehouse

- Data in data warehouse is accessed by BI (Business Intelligence) users for analytical reporting, data mining and analysis. This is used for decision making by business users, sales manager, and analysts to define future strategy.
- BI is a technology infrastructure for gaining maximum information from available data for the purpose of improving business processes. Typical BI infrastructure components are as follows: software solution for gathering, cleansing, integrating, analyzing and sharing data.
- Business Intelligence produces analysis and provides believable information to help making effective and high quality business decisions.
- The most common kinds of Business Intelligence systems are:
  - **EIS** - Executive Information Systems
  - **DSS** - Decision Support Systems
  - **MIS** - Management Information Systems
  - **GIS** - Geographic Information Systems
  - **OLAP** - Online Analytical Processing and multidimensional analysis
  - **CRM** - Customer Relationship Management
- Business Intelligence systems are based on Data Warehouse technology. A Data Warehouse (DW) gathers information from a wide range of company's operational systems. Data loaded to DW is usually good integrated and cleaned that allows producing credible information which reflected so called 'one version of the true'.

Differences between Data Warehouses and Operational Databases

The purpose of an operational database is to record and store current data from users. A database is suitable for the traditional type of data storage method. For instance, a bank ATM uses a database to record their customers' money transactions in real-time. A data warehouse, on the other hand, is a type of database but specifically designed for data analysis. It is used to store and summarize large volumes of historical data.

Data Warehouse Vs Operational Database

- An **Operational database** is designed for known workloads and transactions like updating a user record, searching a record, etc. However, Data Warehouse transactions are more complex and present a general form of data.
- An **Operational database** contains the current data of an organization and Data warehouse normally contains the historical data.
- An **Operational Database** supports parallel processing of multiple transactions. Concurrency control and recovery mechanisms are required to maintain consistency of the database.
- An **Operational Database** query allows to read and modify operations (insert, delete and Update) while an OLAP query needs only read-only access of stored data (Select statement).

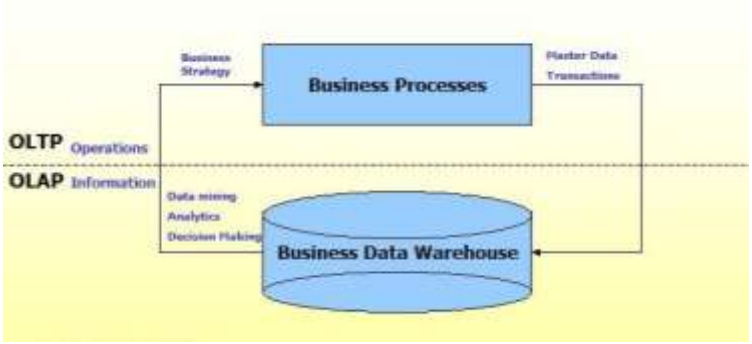
Benefits of Data Warehouses

A data warehouse is essential for any business that wants to profit from sound business decisions. A goal common to all businesses is to make better business decisions than their competitors. Once a data warehouse is implemented into your business intelligence plans, your company can benefit from it in many ways.

- **Better decision-making –** Corporate decision makers will no longer have to make important business decisions based on limited data and hunches. Data warehouses store credible facts and statistics, and decision makers will be able to retrieve that information from the data warehouse based on their personal needs. In addition to making strategic decisions, a data warehouse can also assist in marketing segmentation, inventory management, financial management, and sales.
- **Quick and easy access to data –** Speed is an important factor that sets you above your competitors. Business users can quickly access data from multiple sources from a data warehouse, meaning that precious time won't be wasted on retrieving data from multiple sources. This allows you to make quick and accurate decisions, with little or no support from your IT department.
- **Data quality and consistency –** Since data warehouses gather information from different sources and convert it into a single and widely used format, departments will produce results that are in line and consistent with each other. When data is standardized, you can have confidence in its accuracy, and accurate data is what makes for strong business decisions.

Online Transaction Processing (OLTP) vs Online Analytical Processing (OLAP)

Information Systems can be divided into transactional (OLTP) and analytical (OLAP). OLTP systems provide source data to data warehouses, whereas OLAP systems help to analyze it.



- **OLTP (On-line Transaction Processing)** is characterized by a large number of short on-line transactions (INSERT, UPDATE, and DELETE). The main emphasis for OLTP systems is put on very fast query processing, maintaining data integrity in multi-access environments and an effectiveness measured by number of transactions per second. In OLTP database there is detailed and current data, and schema used to store transactional databases is the entity model (usually 3NF).

  For Example

  A Day-to-Day transaction system in a retail store, where the customer records are inserted, updated and deleted on a daily basis. It provides faster query processing. OLTP databases contain detailed and current data. The schema used to store OLTP database is the Entity model.

- **OLAP (On-line Analytical Processing)** is characterized by relatively low volume of transactions. Queries are often very complex and involve aggregations. For OLAP systems a response time is an effectiveness measure. OLAP applications are widely used by Data Mining techniques. In OLAP database there is aggregated, historical data, stored in multi-dimensional schemas (usually star schema).

The following table summarizes the major differences between OLTP and OLAP systems.

	OLTP System **Online Transaction Processing** **(Operational System)**	**OLAP System** **Online Analytical Processing** **(Data Warehouse)**
Source of data	Operational data; OLTPs are the original source of the data.	Consolidation data; OLAP data comes from the various OLTP Databases
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support
What the data	Reveals a snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and Updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Relatively standardized and simple queries returning relatively few records	Often complex queries involving aggregations
Processing Speed	Typically very fast	Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes

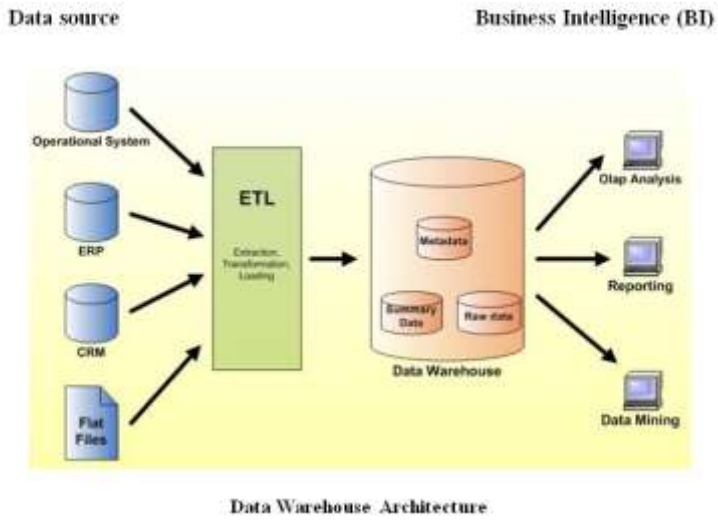
	OLTP System Online Transaction Processing (Operational System)	OLAP System Online Analytical Processing (Data Warehouse)
Space Requirements	Can be relatively small if historical data is archived	Larger due to the existence of aggregation structures and history data; requires more indexes than OLTP
Database Design	Highly normalized with many tables	Typically de-normalized with fewer tables; use of star and/or snowflake schemas
Backup and Recovery	Backup religiously; operational data is critical to run the business, data loss is likely to entail significant monetary loss and legal liability	Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method

Need for a Data Warehouse

A fundamental concept of a data warehouse is the distinction between **data** and **information**. **Data** is composed of raw observable and recordable facts that are often found in operational or transactional systems. At Kibabii University these systems include the registrar's data on students, human resource and payroll databases, course scheduling data, and data on financial aid. In a data warehouse environment, **data** only comes to have value to end-users when it is organized and presented as **information**. **Information** is an integrated collection of facts and is used as the basis for decision making.

For example, an academic unit needs to have diachronic information about its extent of instructional output of its different faculty members to gauge if it is becoming more or less reliant on part-time faculty. For a company to be successful in the future, they must make good decisions. And to make good decisions requires all relevant data to be taking into consideration. And the best source for that data is a well-designed data warehouse.

- A data warehouse exists to answer questions users have about the business, the performance of the various operations, the business trends, and about what can be done to improve the business.
- The data warehouse exists to provide business users with direct access to data, to provide a single unified version of the performance indicators, to record the past accurately, and to provide the ability to view the data from many different perspectives.
- A Data Warehouse is used for reporting and analyzing of information and stores both historical and current data. The data in DW system is used for Analytical reporting, which is later used by Business Analysts, Sales Managers or Knowledge workers for decision-making.
- The data in a DW system is loaded from operational transaction systems such as Sales, Marketing, and Human Resource.
- From the figure below, you can see that the data is coming from multiple heterogeneous data sources to a Data Warehouse. Common data sources for a data warehouse includes Operational databases, Flat Files (xls, csv, txt files)



Data Warehouse Architecture

As an information technology professional, you will work on computer applications as an analyst, programmer, designer, developer, database administrator, or project manager. You will be been involved in the design, implementation, and maintenance of systems that support day-to-day business operations. Depending on the industries you will work in, you will be involved in applications such as order processing, general ledger, inventory, in-patient billing, checking accounts, insurance claims, and so on.

These applications are important systems that run businesses. They process orders, maintain inventory, keep the accounting books, service the clients, receive payments, and process claims. Without these computer systems, no modern business can survive. These applications are effective in what they are designed to do. They gather, store, and process all the data needed to successfully perform the daily operations. They provide online information and produce a variety of reports to monitor and run the business.

In the 1990s, as businesses grew more complex, corporations spread globally, and competition became fiercer, business executives became desperate for information to stay competitive and improve the bottom line. The operational systems did provide information to run the day-to-day operations, but what the executives needed were different kinds of information that could be readily used to make strategic decisions. They wanted to know where to build the next warehouse, which product lines to expand, and which markets they should strengthen. The operational systems, important as they were, could not provide strategic information. Businesses, therefore, were compelled to turn to new ways of getting strategic information.

Data warehousing is a new paradigm specifically intended to provide vital strategic information. In the 1990s, organizations began to achieve competitive advantage by building data warehouse systems.

Escalating Need for Strategic Information

Enterprises need for strategic information to make strategic decisions but the prevailing information crisis holds them back. However, data warehouse enables enterprises to provide strategic information.

The executives and managers who are responsible for keeping the enterprise competitive need information to make proper decisions. They need information to formulate the business strategies, establish goals, set objectives, and monitor results.

Here are some examples of business objectives:

- Retain the present customer base
- Increase the customer base by 15% over the next 5 years
- Gain market share by 10% in the next 3 years
- Improve product quality levels in the top five product groups
- Enhance customer service level in shipments
- Bring three new products to market in 2 years
- Increase sales by 15% in the North East Division

For making decisions about these objectives, executives and managers need information for the following purposes:

- to get in-depth knowledge of their company's operations
- learn about the key business factors and how these affect one another
- monitor how the business factors change over time
- compare their company's performance relative to the competition and to industry benchmarks

Executives and managers need to focus their attention on customers' needs and preferences, emerging technologies, sales and marketing results, and quality levels of products and services. The types of information needed to make decisions in the formulation and execution of business strategies and objectives are broad-based and encompass the entire organization. All these types of essential information is called **strategic information.**

Strategic information is not for running the day-to-day operations of the business. It is not intended to produce an invoice, make a shipment, settle a claim, or post a withdrawal from a bank account. Strategic information is far more important for the continued health and survival of the corporation. Critical business decisions depend on the availability of proper strategic information in an enterprise. Figure below lists the desired characteristics of strategic information.

INTEGRATED	Must have a single, enterprise-wide view.
DATA INTEGRITY	Information must be accurate and must conform to business rules.
ACCESSIBLE	Easily accessible with intuitive access paths, and responsive for analysis.
CREDIBLE	Every business factor must have one and only one value.
TIMELY	Information must be available within the stipulated time frame.

The Information Crisis

Whatever the size of a company may be, think of all the various computer applications in the company. Think of all the databases and the quantities of data that support the operations of a company;

- How many years' worth of customer data is saved and available?
- How many years' worth of financial data is kept in storage?
- Where is all this data? (On one platform, in legacy systems or in client/server applications)

We are faced with two facts:

(1) organizations have lots of data
(2) information technology resources and systems are not effective at turning all that data into useful strategic information

Most companies are faced with an information crisis not because of lack of sufficient data, but because the available data is not readily usable for strategic decision making. These large quantities of data are very useful and good for running the business operations, but hardly amenable for use in making decisions about business strategies and objectives because;

- The data of an enterprise is spread across many types of incompatible structures and systems. Your order processing system might have been developed 20 years ago and is running on an old computer. Some of the data may still be in old file formats. Your later credit assignment and verification system might be on a client/server platform and the data for this application might be in relational tables.
- The data in a corporation resides in various disparate systems, multiple platforms, and diverse structures. The more technology a company has used in the past, the more disparate the data of a company will be.

For proper decision making on overall corporate strategies and objectives, there is need information integrated from all systems. Data needed for strategic decision making must be in a format suitable for analyzing trends. Executives and managers need to look at trends over time and steer their companies

in the proper direction. You get snapshots of transactions that happen at specific times. You have data about units of sale of a single product in a specific order on a given date to a certain customer. In the operational systems, you do not readily have the trends of a single product over the period of a month, a quarter, or a year.

For strategic decision making, executives and managers must be able to review data from different business viewpoints. For example, they must be able to review sales quantities by product, salesperson, district, region, and customer groups. Operational data is not directly suitable for review from different viewpoints.

Technology Trends

The entire spectrum of computing has undergone tremendous changes. The computing focus itself has changed over the years. Old practices could not meet new needs. Screens and preformatted reports are no longer adequate to meet user requirements. Over the years, the price of computers is continuing to decline, digital storage is costing less and less, and network bandwidth is increasing as its price decreases. We have seen explosive changes in these critical areas:

- Computing technology
- Human/machine interface
- Processing options

Hardware economics and miniaturization allow a workstation on every desk and provide increasing power at reducing costs. New software provides easy-to-use systems. Open systems architecture creates cooperation and enables usage of multivendor software. Improved connectivity, networking, and the Internet open up interaction with an enormous number of systems and databases.

All of these improvements in technology are commendable. These have made computing faster, cheaper, and widely available. The current state of the technology is conducive to providing strategic information because strategic information requires collection of large volumes of corporate data and storing it in suitable formats. Technology advances in data storage and reduction in storage costs readily accommodate data storage needs for strategic decision-support systems.

Analysts, executives, and managers use strategic information interactively to analyze and spot business trends. The user will ask a question and get the results, then ask another question, look at the results, and ask yet another question. This interactive process continues. Tremendous advances in interface software make such interactive analysis possible.

Processing large volumes of data and providing interactive analysis requires extra computing power. The explosive increase in computing power and its lower costs make provision of strategic information feasible. What we could not accomplish a few years earlier for providing strategic information is now possible with the current advanced stage of information technology.

Operational Systems (Online Transaction Systems)

Operational systems such as order processing, inventory control, claims processing, outpatient billing, and so on are not designed or intended to provide strategic information. If we need the ability to provide strategic information, we must get the information from altogether different types of systems.

Only specially designed decision support systems or informational systems can provide strategic information.

Basic Elements of Data Warehousing

The desired features of data warehousing are:

- Database designed for analytical tasks
- Data from multiple applications
- Easy to use and conducive to long interactive sessions by users
- Read-intensive data usage
- Direct interaction with the system by the users without IT assistance
- Content updated periodically and stable
- Content to include current and historical data
- Ability for users to run queries and get results online
- Ability for users to initiate reports

Processing Requirements in a Data Warehouse

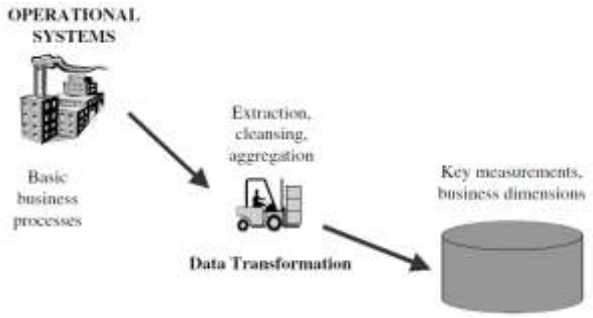
Most of the processing in data warehouse environment for strategic information will have to be analytical. There are four levels of analytical processing requirements:

1. Running of simple queries and reports against current and historical data
2. Ability to perform "what if " analysis in many different ways
3. Ability to query, step back, analyze, and then continue the process to any desired length
4. Spot historical trends and apply them for future results

Business Intelligence (BI) at the Data Warehouse

A goal of every business is to make better business decisions than their competitors. BI turns the massive amount of data from operational systems into a format that is easy to understand, current, and correct so decisions can be made on the data. You can then analyze current and long-term trends, be instantly alerted to opportunities and problems, and receive continuous feedback on the effectiveness of your decisions. The data warehouse holds the business intelligence for the enterprise to enable strategic decision making.

At a high level of interpretation, the data warehouse contains critical measurements of the business processes stored along business dimensions. For example, a data warehouse might contain units of sales, by product, day, customer group, sales district, sales region, and promotion. Here the business dimensions are product, day, customer group, sales district, sales region, and promotion. The data in the warehouse is derived from the operational systems that support the basic business processes of the organization. In between the operational systems and the data warehouse, there is a data staging area. In this staging area, the operational data is cleansed and transformed into a form suitable for placement in the data warehouse for easy retrieval.



Business Intelligence at the data warehouse

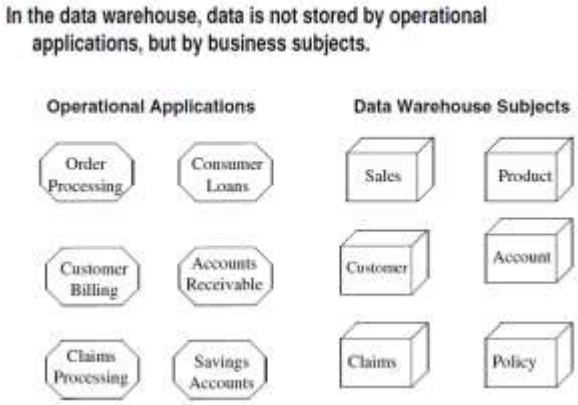
Features of a data warehouse

The data in the data warehouse is:

- Separate
- Available
- Integrated
- Time stamped
- Subject oriented
- Non-volatile
- Accessible

Subject-Oriented Data

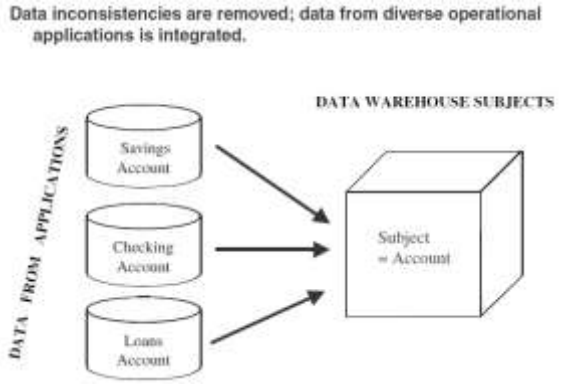
A data warehouse is subject oriented, thus information is presented according to specific subjects or areas of interest, not simply as computer files. Data is manipulated to provide information about a particular subject. For example, a University data warehouse is not simply made accessible to end-users, but is provided structure and organized according to the specific needs.



In a data warehouse, there is no application flavour. The data in a data warehouse cut across applications.

Integrated Data

For proper decision making, you need to pull together all the relevant data from the various applications. The data in the data warehouse comes from several operational systems. Source data are in different databases, files, and data segments. These are disparate applications, so the operational platforms and operating systems could be different. The file layouts, character code representations, and field naming conventions all could be different. In addition to data from internal operational systems, for many enterprises, data from outside sources is likely to be very important. A data warehouse provides one-stop shopping and contains information about a variety of subjects. Thus a University data warehouse has information on students, faculty and staff, instructional workload, and student results.

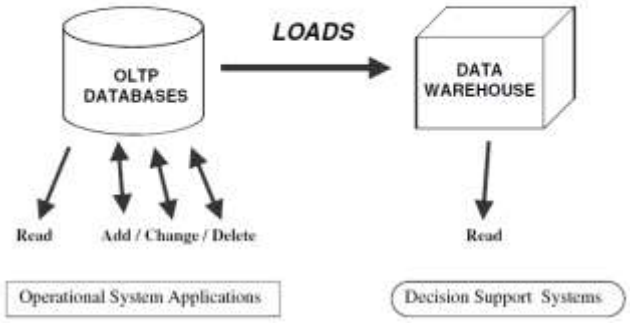


Non-Volatile Data

Data extracted from the various operational systems and pertinent data obtained from outside sources are transformed, integrated, and stored in the data warehouse. The data in the data warehouse is not intended to run the day-to-day business. When you want to process the next order received from a customer, you do not look into the data warehouse to find the current stock status. The operational order entry application is meant for that purpose. In the data warehouse, you keep the extracted stock status data as snapshots over time. You do not update the data warehouse every time you process a single order.

Data from the operational systems are moved into the data warehouse at specific intervals. Depending on the requirements of the business, these data movements take place twice a day, once a day, once a week, or once in two weeks. In fact, in a typical data warehouse, data movements to different data sets may take place at different frequencies. The changes to the attributes of the products may be moved once a week. Any revisions to geographical setup may be moved once a month. The units of sales may be moved once a day. **A data warehouse** holds stable information that doesn't change each time an operational process is executed. Information is consistent regardless of when the warehouse is accessed.

Usually the data in the data warehouse is not updated or deleted.



Time-Variant Data

A data warehouse contains a history of the subject, as well as current information. Historical information is an important component of a data warehouse. A data warehouse, because of the very nature of its purpose, has to contain historical data, not just current values. Data is stored as snapshots over past and current periods. Every data structure in the data warehouse contains the time element. You will find historical snapshots of the operational data in the data warehouse.

The time-variant nature of the data in a data warehouse

- Allows for analysis of the past
- Relates information to the present
- Enables forecasts for the future

Accessible

The primary purpose of a data warehouse is to provide readily accessible information to end users.

Process-Oriented

It is important to view data warehousing as a process for delivery of information. The maintenance of a data warehouse is ongoing and iterative in nature.

Data Warehouses and Data Marts

The single most important issue facing IT managers is whether to build the data warehouse first or the data mart first.

Before deciding to build a data warehouse for your organization, you need to ask the following basic and fundamental questions and address the relevant issues:

- Top-down or bottom-up approach: Should you look at the big picture of your organization, take a top-down approach, and build a huge data warehouse?
- Enterprise-wide or departmental: should you adopt a bottom-up approach, look at the individual local and departmental requirements, and build bite-size departmental data marts?
- Which first—data warehouse or data mart: Should you build a large data warehouse and then let that repository feed data into local, departmental data marts?

- Build pilot or go with a full-fledged implementation: should you build individual local data marts, and combine them to form your overall data warehouse? Should these local data marts be independent of one another?
- Dependent or independent data marts: should they be dependent on the overall data warehouse for data feed?

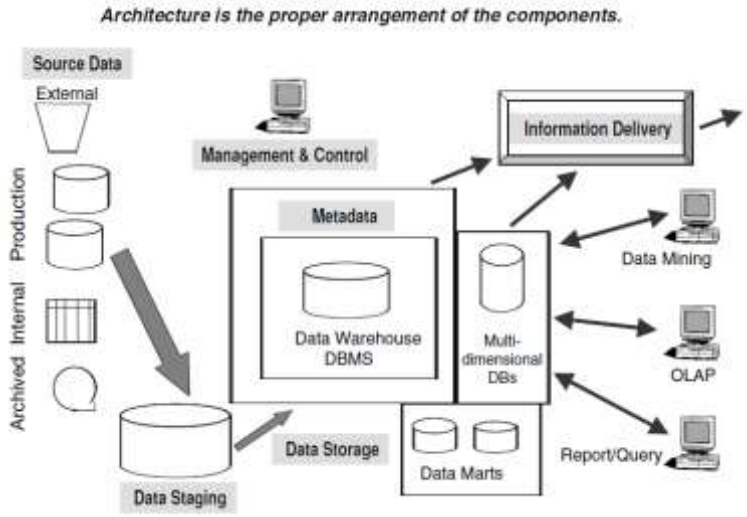
DATA WAREHOUSE	DATA MART
Corporate/Enterprise-wide	Departmental
Union of all data marts	A single business process
Data received from staging area	Star-join (facts & dimensions)
Queries on presentation resource	Technology optimal for data access and analysis
Structure for corporate view of data	Structure to suit the departmental view of data
Organized on E-R model	

A data mart is a logical subset of the complete data warehouse a sort of pie-wedge of the whole data warehouse. A data warehouse, therefore, is a conformed union of all data marts. Individual data marts are targeted to particular business groups in the enterprise, but the collection of all the data marts form an integrated whole, called the enterprise data warehouse.

Review Questions

1. Define the following terms
   (a) Data warehousing
   (b) Data warehouse
   (c) Business intelligence
2. Describe at least six characteristics or features of a data warehouse.
3. A data warehouse is an environment, not a product. Discuss
4. A Data Warehouse is Blend of Many Technologies. Discuss
5. Explain the differences between Data Warehouses and Operational Databases
6. Discuss the benefits of Data Warehouses
7. Explain the differences between Online Transaction Processing (OLTP) and Online Analytical Processing (OLAP)
8. Discuss the need for a Data Warehouse
9. Why is data integration required in a data warehouse, more so there than in an operational application?
10. A data warehouse is subject-oriented. What would be the major critical business subjects for the following companies?
    a. an international manufacturing company
    b. a local community bank
    c. a domestic hotel chain
11. Every data structure in the data warehouse contains the time element. Why?
12. Explain data granularity and how it is applicable to the data warehouse.
13. How are the top-down and bottom-up approaches for building a data warehouse different? Discuss the merits and disadvantages of each approach.

Architectural Components of a Data Warehouse



Source Data Component

Source data coming into the data warehouse may be grouped into four broad categories:

1. ***Production Data.*** This category of data comes from the various operational systems of the enterprise. The significant and disturbing characteristic of production data is disparity. The great challenge is to standardize and transform the disparate data from the various production systems, convert the data, and integrate the pieces into useful data for storage in the data warehouse.
2. ***Internal Data.*** In every organization, users keep their "private" spreadsheets, documents, customer profiles, and sometimes even departmental databases. This is the internal data, parts of which could be useful in a data warehouse.
3. ***Archived Data.*** Operational systems are primarily intended to run the current business. In every operational system, you periodically take the old data and store it in archived files. The circumstances in an organization dictate how often and which portions of the operational databases are archived for storage. Some data is archived after a year. Sometimes data is left in the operational system databases for as long as five years.
4. ***External Data.*** Most executives depend on data from external sources for a high percentage of the information they use.
    - They use statistics relating to their industry produced by external agencies.
    - They use market share data of competitors.
    - They use standard values of financial indicators for their business to check on their performance.
    - For example, the data warehouse of a car rental company contains data on the current production schedules of the leading automobile manufacturers. This external data in the data warehouse helps the car rental company plan for their fleet management.
    - The purposes served by such external data sources cannot be fulfilled by the data available within your organization itself. The insights gleaned from your production data and your archived data are somewhat limited. They give you a picture based on what you are doing or have done in the past. In order to spot industry trends and compare performance against other organizations, you need data from external sources. Usually, data from outside sources do not conform to your formats. You have to devise conversions of data into your internal formats and data types. You have to organize the data transmissions from the external sources.

Data Staging Component

After data is extracted data from various operational systems and from external sources, it has to be prepared for storage in the data warehouse. The extracted data coming from several disparate sources needs to be changed, converted, and made ready in a format that is suitable to be stored for querying and analysis.

Three major functions need to be performed for getting the data ready;

- extract the data,
- transform the data, and
- load the data into the data warehouse storage

These three major functions of extraction, transformation, and preparation for loading take place in a staging area. This area consists of a workbench for these functions. It provides a place and an area with a set of functions to clean, change, combine, convert, de-duplicate, and prepare source data for storage and use in the data warehouse.

Why do we need a separate place or component to perform the data preparation?

- Remember that data in a data warehouse is subject-oriented and cuts across operational applications.
- A separate staging area, therefore, is a necessity for preparing data for the data warehouse.

Data Extraction

This function deals with numerous data sources. Appropriate technique needs to be employed for each data source. Source data may be from different source machines in diverse data formats. Part of the source data may be in relational database systems. Some data may be on other network and hierarchical data models. Many data sources may still be in flat files.

Data extraction may become quite complex. Tools are available in the market for data extraction. Consider using outside tools suitable for certain data sources. For the other data sources, a company may want to develop in-house programs to do the data extraction. Purchasing outside tools may entail high initial costs. In-house programs, on the other hand, may mean ongoing costs for development and maintenance.

Data warehouse implementation teams extract the source into a separate physical environment from which moving the data into the data warehouse would be easier. In the separate environment, source data needs to be extracted into a group of flat files, or a data-staging relational database, or a combination of both.

Data Transformation

- Data transformation involves many forms of combining pieces of data from the different sources. Data is combined from single source record or related data elements from many source records.
- Data transformation also involves purging source data that is not useful and separating out source records into new combinations.
- Sorting and merging of data takes place on a large scale in the data staging area.

Data Transformation Tasks

A number of individual tasks form part of data transformation.

1. Clean the data extracted from each source. Cleaning may be correction of misspellings or may deal with providing default values for missing data elements, or elimination of duplicates when you bring in the same data from multiple source systems.
2. Standardize the data types and field lengths for same data elements retrieved from the various sources.
3. Resolve synonyms and homonyms. When two or more terms from different source systems mean the same thing, it must be resolved. When a single term means many different things in different source systems, it must also be resolved.
4. The assignment of surrogate keys derived from the source system primary keys.

In many cases, the keys chosen for the operational systems are field values with built-in meanings. For example, the product key value may be a combination of characters indicating the product category, the code of the warehouse where the product is stored, and some code to show the production batch. Primary keys in the data warehouse cannot have built-in meanings.

A grocery chain point-of-sale operational system keeps the unit sales and revenue amounts by individual transactions at the check-out counter at each store. But in the data warehouse, it may not be necessary to keep the data at this detailed level. Data may be summarized by product totals at each store for a given day and keep the summary totals of the sale units and revenue in the data warehouse storage. In such cases, the data transformation function would include appropriate summarization.

The end of data transformation function result into a collection of integrated data that is cleaned, standardized, and summarized ready to load into each data set in the data warehouse.

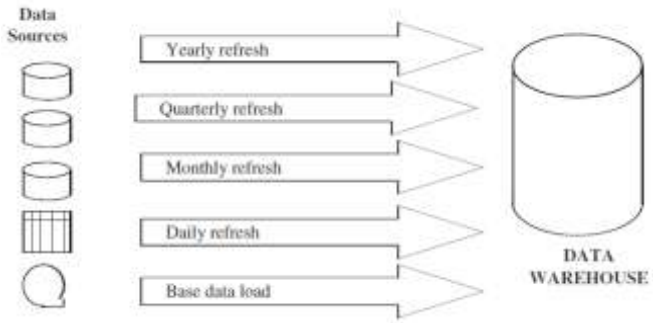
Data Loading

- Two distinct groups of tasks form the data loading function. When the design and construction of the data warehouse is complete and go live for the first time, the initial loading of the data into the data warehouse storage takes place.
- The initial load moves large volumes of data using up substantial amounts of time.
- As the data warehouse starts functioning, there is need to continue extracting the changes to the source data, transforming the data revisions, and feeding the incremental data revisions on an ongoing basis.

Data Storage Component

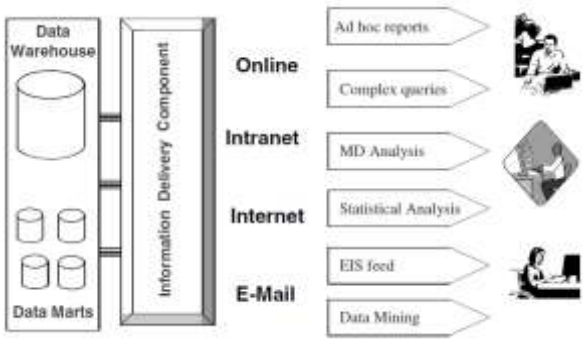
- The data storage for the data warehouse is a separate repository.
- The operational systems (online transaction processing applications) of an enterprise support the day-to-day operations. The data repositories for the operational systems typically contain only the current data. These data repositories contain the data structured in highly normalized formats for fast and efficient processing.
- In contrast, in the data repository for a data warehouse, large volumes of historical data are kept for analysis.
- Further, the data kept in the data warehouse need to be in structures suitable for analysis, and not for quick retrieval of individual pieces of information.
- Therefore, the data storage for the data warehouse is kept separate from the data storage for operational systems.

- In databases supporting operational systems, the updates to data happen as transactions occur. These transactions hit the databases in a random fashion. How and when the transactions change the data in the databases is not completely within the users' control.
    - The data in the operational databases could change from moment to moment.
- When analysts use the data in the data warehouse for analysis, they need to know that the data is stable and that it represents snapshots at specified periods. As they work with the data, the data storage must not be in a state of continual updating. For this reason, the data warehouses are "read-only" data repositories.
- The database in a data warehouse must be open. Depending on your requirements, one can use tools from multiple vendors. The data warehouse must be open to different tools.
    - Most of the data warehouses employ relational database management systems.
    - Many of the data warehouses also employ multidimensional database management systems.
- Data extracted from the data warehouse storage is aggregated in many ways and the summary data is kept in the multidimensional databases (MDDBs). Such multidimensional database systems are usually proprietary products.



Information Delivery Component

- Many users need information from the data warehouse. The range is fairly comprehensive.
    - The novice user comes to the data warehouse with no training and, therefore, needs prefabricated reports and preset queries.
    - The casual user needs information once in a while, not regularly. This type of user also needs pre-packaged information.
    - The business analyst looks for ability to do complex analysis using the information in the data warehouse.
    - The power user wants to be able to navigate throughout the data warehouse, pick up interesting data, format his/her own queries, drill through the data layers, and create custom reports and ad hoc queries.
- In order to provide information to the wide community of data warehouse users, the information delivery component includes different methods of information delivery.
- Ad hoc reports are predefined reports primarily meant for novice and casual users.
- Provision for complex queries, multidimensional (MD) analysis, and statistical analysis cater to the needs of the business analysts and power users.
- Information fed into Executive Information Systems (EIS) is meant for senior executives and high-level managers.



- Some data warehouses also provide data to data-mining applications. Data-mining applications are knowledge discovery systems where the mining algorithms help in discovering trends and patterns from the usage of your data.
- A data warehouse may include several information delivery mechanisms such as **online queries** and **reports**. The users will enter their requests online and will receive the results online. One may set up delivery of scheduled reports through e-mail or make adequate use of the organization's intranet for information delivery. Information delivery over the Internet has been gaining ground.

Metadata Component

- Metadata in a data warehouse is similar to the data dictionary or the data catalogue in a database management system. In the data dictionary, the information about the logical data structures, the information about the files and addresses, the information about the indexes, etc is kept. The data dictionary contains data about the data in the database.
- The metadata component is the data about the data in the data warehouse.

Management and Control Component

This component of the data warehouse architecture sits on top of all the other components. It interacts with the metadata component to perform the management and control functions. As the metadata component contains information about the data warehouse itself, the metadata is the source of information for the management module.

Functions of management and control component

1. It coordinates the services and activities within the data warehouse.
2. It controls the data transformation and the data transfer into the data warehouse storage.
3. It moderates the information delivery to the users.
4. It works with the database management systems and enables data to be properly stored in the repositories.
5. It monitors the movement of data into the staging area and from there into the data warehouse storage itself.

Summary

- Defining features of the data warehouse are: separate, subject-oriented, integrated, time-variant, and non-volatile.
- You may use a top-down approach and build a large, comprehensive, enterprise data warehouse; or, you may use a bottom-up approach and build small, independent, departmental data marts. In spite of some advantages, both approaches have serious shortcomings.
- A viable practical approach is to build conformed data marts, which together form the corporate data warehouse.

- Data warehouse building blocks or components are: source data, data staging, data storage, information delivery, metadata, and management and control.
- In a data warehouse, metadata is especially significant because it acts as the glue holding all the components together and serves as a roadmap for the end-users.

Review Questions

1. What are the various data sources for the data warehouse?
2. Why do you need a separate data staging component?
3. Under data transformation, list five different functions you can think of.
4. Name any six different methods for information delivery.
5. You are the data analyst on the project team building a data warehouse for an insurance company. List the possible data sources from which you will bring the data into your data warehouse. State your assumptions.
6. For an airlines company, identify three operational applications that would feed into the data warehouse. What would be the data load and refresh cycles?
7. Prepare a table showing all the potential users and information delivery methods for a data warehouse supporting a large national grocery chain.

The Significant Role of Metadata

Objectives

- Explain why metadata is so important
- Understand who needs metadata and what types they need
- Review metadata types by the three functional areas
- Discuss business metadata and technical metadata in detail
- Examine all the requirements metadata must satisfy
- Describe the challenges for metadata management
- Describe options for providing metadata

Importance of Metadata

Metadata in a data warehouse contains the answers to questions about the data in the data warehouse. The answers are kept in a place called the metadata repository. Definitions of Meta data

- Data about the data
- Table of contents for the data
- Catalogue for the data
- Data warehouse atlas
- Data warehouse roadmap
- Data warehouse directory
- Glue that holds the data warehouse contents together
- The nerve centre

Meta data describes all the pertinent aspects of the data in the data warehouse fully and precisely.

A Critical Need in the Data Warehouse

For Using the Data Warehouse

- There is one big difference between a data warehouse and any operational. The difference is in the information access.
    - In operational systems, users get information through the relevant screens.
    - They get information from specific reports.
- In contrast, users themselves retrieve information from the data warehouse.
    - Users themselves create ad hoc queries and run these against the data warehouse.
    - They format their own reports.
- Because of this major difference, before they can create and run their queries, users need to know about the data in the data warehouse.
- In operational systems:
    - Users do not have any easy and flexible methods for knowing the nature of the contents of the database.
    - There is no need for user-friendly interfaces to the database contents.
    - The data dictionary or catalogue is meant for IT uses only.
- In data warehouse, users need to receive maximum value from a data warehouse.
    - They need sophisticated methods for browsing and examining the contents of the data warehouse.
    - They need to know the meanings of the data items.
    - Users have to be prevented from drawing wrong conclusions from their analysis through their ignorance about the exact meanings.

For Building the Data Warehouse

- In order to apply expertise, one must know the source systems and their data structures.
  - You need to know the structures and the data content in the data warehouse.
  - You need to determine the mappings and the data transformations.
- To perform tasks in building the data extraction and data transformation component of the data warehouse, one needs metadata about the source systems, source-to-target mappings, and data transformation rules.
- Metadata is an overall compelling necessity and a very significant component in a data warehouse. Metadata is absolutely essential for building a data warehouse.

For Administering the Data Warehouse

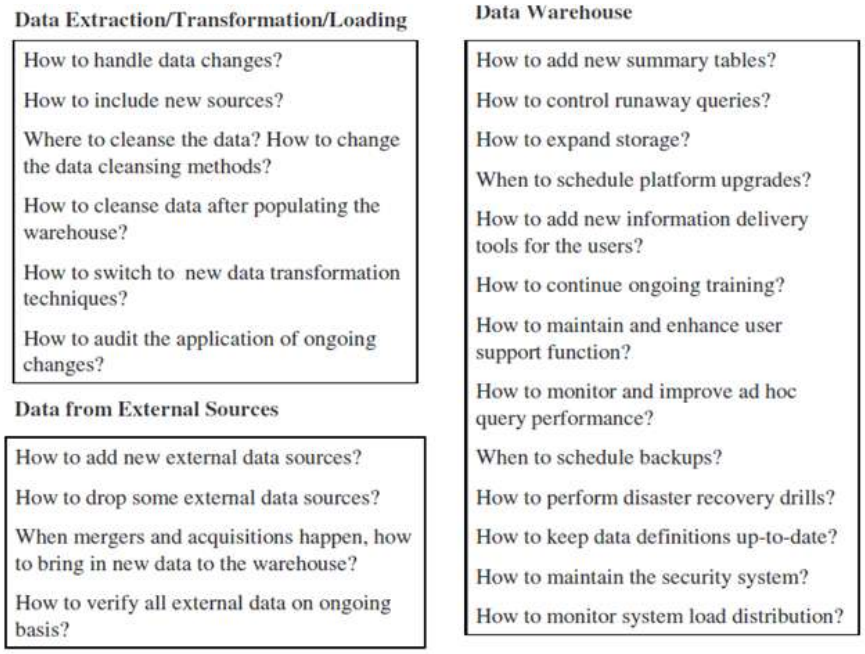
Because of the complexities and enormous sizes of modern data warehouses, it is impossible to administer the data warehouse without substantial metadata.

Responsibilities of Data warehouse DBA

- Responsible for the physical design of the database and for doing the initial loading.
- Responsible for periodic incremental loads.

In order to perform the tasks of physical design and loading, a DBA need metadata about a number of things:

- The layouts in the staging area.
- The logical structure of the data warehouse database.
- The data refresh and load cycles.



Data warehouse administration: questions and issues

Metadata is like a Nerve Centre

- Various processes during the building and administering of the data warehouse generate parts of the data warehouse metadata.
- Parts of metadata generated by one process are used by another.
- In the data warehouse, metadata assumes a key position and enables communication among various processes.
- It acts like a nerve centre in the data warehouse.

Metadata acts as a nerve centre

Who Needs Metadata

- Imagine a filing cabinet stuffed with documents without any folders and labels
- Without metadata, a data warehouse is like such a filing cabinet. It is probably filled with information very useful for users and for IT developers and administrators. But without any easy means to know what is there, the data warehouse is of very limited value.

Metadata is needed by two groups of people:

(1) end-users
(2) IT (developers and administrators)
In the next two subsections, we will review why metadata is critical for each of these two groups

Importance of Metadata to End-Users

The following would be a typical use of a data warehouse by a key user, e.g. a business analyst. The Marketing manager of a company has asked this business analyst to do a thorough analysis of a problem that recently surfaced. Because of the enormous sales potential in the Western and Rift Valley regions, a company has opened five new stores in each region.

Although overall countrywide sales increased nicely for two months following the opening of the stores, after that the sales went back to the prior levels and remained flat. The Marketing manager wants to know why, so that he can take appropriate action.

As a user, the business analyst expects to find answers from the new data warehouse, but he does not know the details about the data in the data warehouse. Specifically, he does not know the answers to the following questions:

- Are the sale units and shillings stored by individual transactions or as summary totals, by product, for each day in each store?
- Can sales be analyzed by product, promotion, store, and month?
- Can current month sales be compared to sales in the same month last year?
- Can sales be compared to targets?
- How is profit margin calculated?
- What are the business rules?
- What is the definition of a sales region?
- Which counties are included in each of the two regions being analyzed?
- Where did the sales come from? From which source systems?
- How old are the sales numbers? How often do these numbers get updated?

If the analyst is not sure of the nature of the data, he is likely to interpret the results of the analysis incorrectly. It is possible that the new stores are cannibalizing sales from their own existing stores and that is why the overall sales remain flat. But the analyst may not find the right reasons because of misinterpretation of the results.

- The analysis will be more effective if he can access adequate metadata to help as a powerful roadmap of the data.
- If there is sufficient and proper metadata, the analyst does not have to get assistance from IT every time he needs to run an analysis.
- Easily accessible metadata is crucial for end-users.

Importance of Metadata to the IT Staff

- Development and deployment of a data warehouse is a joint effort between the IT staff and user representatives.
- Because of the technical issues, IT staff is primarily responsible for the design and ongoing administration of the data warehouse.
- For performing the responsibilities for design and administration, IT staff must have access to proper metadata.
- Throughout the entire development process, metadata is essential for IT staff.
    - Beginning with the data extraction and ending with information delivery.
    - As the development process moves through data extraction, data transformation, data integration, data cleansing, data staging, data storage, query and report design, design for OLAP, and other front-end systems, metadata is critical for IT staff to perform their development activities.
- List of processes in which metadata is significant for IT staffs:
    - Data extraction from sources
    - Data transformation
    - Data cleaning
    - Data aggregation and summarization
    - Data staging
    - Data refreshment
    - Database design
    - Query and report design

Automation of Warehousing Tasks

- Traditionally, metadata has been created and maintained as documentation about the data for each process. Now metadata is assuming a new active role.
- Tools perform major functions in a data warehouse environment. For example, tools enable the extraction of data from designated sources.
- When providing the mapping algorithms, data transformation tools transform data elements to suit the target data structures.
- At the front end, tools empower the users to browse the data content and gain access to the data warehouse.
- These tools generally fall into two categories:
  - development tools for IT professionals
  - information access tools for end-users
- Tool for design and development lets designers to create and record a part of the data warehouse metadata.
- When one uses another tool to perform another process in the design and development, this tool uses the metadata created by the first tool.
- When end-user uses a query tool for information access at the front end, that query tool uses metadata created by some of the back-end tools.

Metadata assumes an active role

- Metadata is no longer passive documentation.
- Metadata takes part in the process. It aids in the automation of data warehouse processes.
- Considering the back-end processes beginning with the defining of the data sources, as the data movement takes place from the data sources to the data warehouse database through the data staging area, several processes occur. In a typical data warehouse, appropriate tools assist in these processes. Each tool records its own metadata as data movement takes place. The metadata recorded by one tool drives one or more processes that follow. This is how metadata assumes an active role and assists in the automation of data warehouse processes.

List of back-end processes in the order in which they occur:

1. Source data structure definition
2. Data extraction
3. Initial reformatting/merging
4. Preliminary data cleansing
5. Data transformation and consolidation
6. Validation and quality check
7. Data warehouse structure definition
8. Load image creation

Metadata is important in a data warehouse because it drives the processes. The metadata recorded by each tool may reside on the platform where the corresponding process runs.

Establishing the Context of Information

Imagine this scenario. One user wants to run a query to retrieve sales data for three products during the first seven days of April in the Western Region. This user composes the query as follows:

Product = Samsung-1 or Sumsung-2 or Sumsung-3

Region = ‘WESTERN’

Period = 04-10-2017 to 04-17-2017

The result comes back:

Sale Units Amount

Sumsung-1— 25,355 253,550

Sumsung-2— 16,978 254,670

Sumsung-3— 7,994 271,796

How does a user find out what exactly each data element in the query is and what the result set means?

Metadata gives a user the meaning of each data element. Metadata establishes the context for the data elements.

Metadata Types by Functional Areas

Different authors and data warehouse practitioners classify and group metadata in various ways: some by usage, and some by who uses it.

Metadata classification

In each line of the list shown below are the different methods for classification of metadata:

- Administrative/End-user/Optimization
- Development/Usage
- In the data mart/At the workstation
- Building/Maintaining/Managing/Using
- Technical/Business
- Back room/Front room
- Internal/External

Metadata Types

- Data warehouse environment is functionally divided into the three areas of:
  - Data Acquisition
  - Data Storage, and
  - Information Delivery
- All data warehouse processes occur in these three functional areas. Processes are designed in each of the three functional areas. Each of the tools used for these processes creates and records metadata. These tools may also use and be driven by the metadata recorded by other tools.

- Metadata types are grouped by these three functional areas (Data acquisition, Data storage, Information delivery) because every data warehouse process occurs in one of these three areas. All the processes happening in each functional area are taken into account and then put together all the processes in all the three functional areas.

Data Acquisition

In this area, the data warehouse processes relate to the following functions:

- Data extraction
- Data transformation
- Data cleansing
- Data integration
- Data staging

The tools record the metadata:

- The tools record the metadata elements during the development phases.
- As the processes take place, the appropriate tools record the metadata elements relating to the processes.
- The tools record while the data warehouse is in operation after deployment.

How IT professionals use metadata relating to data acquisition

- Some other tools used for other processes either in this area or in some other area may use the metadata recorded by other tools in this area. For example, when one uses a query tool to create standard queries, he/she will be using metadata recorded by processes in the data acquisition area. The query tool is meant for a process in a different area, namely, the information delivery area.
- They use metadata recorded by processes in the data acquisition area for administering and monitoring the ongoing functions of the data warehouse after deployment. Metadata from this area are used to monitor ongoing data extraction and transformation.

How users use metadata relating to data acquisition

- When a user wants to find the data sources for the data elements in his or her query, he or she will look up the metadata from the data acquisition area.
- When the user wants to know how the profit margin has been calculated and stored in the data warehouse, he or she will look up the derivation rules in the metadata recorded in the data acquisition area.

Data Storage

- In this area, the data warehouse processes relate to the following functions:
  - Data loading
  - Data archiving
  - Data management
- The tools record metadata:
  - elements during the development phases

    - as processes take place in the data storage functional area, the appropriate tools record the metadata elements relating to the processes
    - while the data warehouse is in operation after deployment
- Similar to metadata recorded by processes in the data acquisition area, metadata recorded by processes in the data storage area is used for development, administration, and by the users.
    - Designers use the metadata from this area for designing the full data refreshes and the incremental data loads.
    - The DBA use metadata for the processes of backup, recovery, and tuning the database.
    - Data warehouse administration use metadata for purging the data warehouse and for periodic archiving of data.
    - Users use metadata from the data storage functional area to create queries and reports. For example, if one of the users wants to create a query breaking the total quarterly sales down by sale counties. Before the user runs the query, he or she would like to know when the last time the data on county description was loaded was. The user gets the information about load dates of the county description from metadata recorded by the data loading process in the data storage functional area.

Information Delivery

- In this area, the data warehouse processes relate to the following functions:
    - Report generation
    - Query processing
    - Complex analysis
- Mostly, the processes in this area are meant for end-users.
- While using the processes, end-users use metadata recorded in processes of the other two areas of data acquisition and data storage.
- When a user creates a query with the aid of a query processing tool, he or she can refer back to metadata recorded in the data acquisition and data storage areas and can look up the source data configurations, data structures, and data transformations from the metadata recorded in the data acquisition area.
- In the same way, from metadata recorded in the data storage area, the user can find the date of the last full refresh and the incremental loads for various tables in the data warehouse database.
- Metadata recorded in the information delivery functional area relate to predefined queries, predefined reports, and input parameter definitions for queries and reports.
- Metadata recorded in this functional area also include information for OLAP.

Business Metadata and Technical Metadata Classification

Metadata types may also be classified as business metadata and technical metadata.

Business Metadata

- Business metadata connects business users to a data warehouse.
- Business users need to know what is available in the data warehouse from a perspective different from that of IT professionals.
- Business metadata is like a roadmap or an easy-to-use information directory showing the contents and how to get there. It is like a tour guide for executives and a route map for managers and business analysts.
- Business metadata must describe the contents in plain language giving information in business terms. For example, the names of the data tables or individual data elements must not be cryptic but be meaningful terms that business users are familiar with. For example, the data item name *calc_pr_sle* is not acceptable. You need to rename this as *calculatedprior-month-sale.*

- Business metadata is much less structured than technical metadata.
- A substantial portion of business metadata originates from textual documents, spreadsheets, and even business rules and policies not written down completely.
- Even though much of business metadata is from informal sources, it is as important as metadata from formal sources such as data dictionary entries.
- All of the informal metadata must be captured, put in a standard form, and stored as business metadata in the data warehouse.
- A large segment of business users do not have enough technical expertise to create their own queries or format their own reports. Therefore:
  - they need to know what predefined queries are available and what preformatted reports can be produced
  - they must be able to identify the tables and columns in the data warehouse by referring to them by business names
  - business metadata should, therefore, express all of this information in plain language
- Business metadata focuses on providing support for the end-user at the workstation.
- It must make it easy for the end-users to understand what data is available in the data ware house and how they can use it.
- Business metadata portrays the data warehouse purely from the perspective of the end users.
- It is like an external view of the data warehouse designed and composed in simple business terms that users can easily understand.

Examples of Business Metadata

- Connectivity procedures
- Security and access privileges
- The overall structure of data in business terms
- Source systems
- Source-to-target mappings
- Data transformation business rules
- Summarization and derivations
- Table names and business definitions
- Attribute names and business definitions
- Data ownership
- Query and reporting tools
- Predefined queries
- Predefined reports
- Report distribution information
- Common information access routes
- Rules for analysis using OLAP
- Currency of OLAP data
- Data warehouse refresh schedule

The list below provides sample questions business metadata can answer for the end-users.

- How can I sign onto and connect with the data warehouse?
- Which parts of the data warehouse can I access?
- Can I see all the attributes from a specific table?
- What are the definitions of the attributes I need in my query?
- Are there any queries and reports already predefined to give the results I need?
- Which source system did the data I want come from?
- What default values were used for the data items retrieved by my query?

- What types of aggregations are available for the metrics needed?
- How is the value I need in the data item derived from other data items?
- When was the last update for the data items in my query?
- On which data items can I perform drill down analysis?
- How old is the OLAP data? Should I wait for the next update?

Beneficiaries

Beneficiaries of business metadata include:

- Managers
- Business analysts
- Power users
- Regular users
- Casual users
- Senior managers/junior executives

Technical Metadata

- Technical metadata is meant for the IT staff responsible for the development and administration of the data warehouse.
- The technical personnel need information to design each process. These are processes in every functional area of the data warehouse.
- Different members on the project team need different kinds of information from technical metadata.
- If business metadata is like a roadmap for the users to use the data warehouse, technical metadata is like a support guide for the IT professionals to build, maintain, and administer the data warehouse.
- IT staff working on the data warehouse project need technical metadata for different purposes.
- Data acquisition expert needs metadata that is different from that of the information access developer on the team.
- The technical staffs on the project need to understand the data extraction, data transformation, and data cleansing processes.
- They have to know the output layouts from every extraction routine and must understand the data transformation rules.

IT staff require technical metadata for three distinct purposes.

- IT personnel need technical metadata for the initial development of the data warehouse
- Technical metadata is absolutely essential for ongoing growth and maintenance of the data warehouse.
- Technical metadata is also critical for the continuous administration of the production data warehouse.
    - Administrators have to monitor the ongoing data extractions.
    - Administrators have to ensure that the incremental loads are completed correctly and on time.
    - Administrators are responsible for database backups and archiving of old data.
- Data warehouse administration is almost impossible without technical metadata.
- Technical metadata concentrates on support for the IT staff responsible for development, maintenance, and administration.
- Technical metadata is more structured than business metadata.

- Technical metadata is like an internal view of the data warehouse showing the inner details in technical terms.

Examples of Technical Metadata

- Data models of source systems
- Record layouts of outside sources
- Source-to-staging area mappings
- Staging area-to-data warehouse mappings
- Data extraction rules and schedules
- Data transformation rules and versioning
- Data aggregation rules
- Data cleansing rules
- Summarization and derivations
- Data loading and refresh schedules and controls
- Job dependencies
- Program names and descriptions
- Data warehouse data model
- Database names
- Table/view names
- Column names and descriptions
- Key attributes
- Business rules for entities and relationships
- Mapping between logical and physical models
- Network/server information
- Connectivity data
- Data movement audit controls
- Data purge and archival rules
- Authority/access privileges
- Data usage/timings
- Query and report access patterns
- Query and reporting tools

Questions technical metadata can answer for developers and administrators

- What databases and tables exist?
- What are the columns for each table?
- What are the keys and indexes?
- What are the physical files?
- Do the business descriptions correspond to the technical ones?
- When was the last successful update?
- What are the source systems and their data structures?
- What are the data extraction rules for each data source?
- What is source-to-target mapping for each data item in the data warehouse?
- What are the data transformation rules?
- What default values were used for the data items while cleaning up missing data?
- What types of aggregations are available?
- What are the derived fields and their rules for derivation?
- When was the last update for the data items in my query?
- What are the load and refresh schedules?

- How often data is purged or archived? Which data items?
- What is schedule for creating data for OLAP?
- What query and report tools are available?

Beneficiaries of technical metadata

- Project manager
- Data warehouse administrator
- Database administrator
- Metadata manager
- Data warehouse architect
- Data acquisition developer
- Data quality analyst
- Business analyst
- System administrator
- Infrastructure specialist
- Data modeller
- Security architect

How to Provide Metadata

- As a data warehouse is being designed and built, metadata needs to be collected and recorded.
- Metadata describes a data warehouse from various points of view.
- One looks into the data warehouse through the metadata to:
  - find the data sources
  - understand the data extractions and transformations
  - determine how to navigate through the contents
  - retrieve information
- Most of the data warehouse processes are performed with the aid of software tools.
- Metadata management presents great challenges. The challenges are not in the capturing of metadata through the use of the tools during data warehouse processes but lie in the integration of the metadata from the various tools that create and maintain their own metadata.

Metadata Requirements

Metadata must serve as a roadmap to the data warehouse for users and support IT in the development and administration of the data warehouse.

Capturing and Storing Data

- The data dictionary in an operational system stores the structure and business rules as they are at the current time. For operational systems, it is not necessary to keep the history of the data dictionary entries. However, the history of the data in a data warehouse spans several years, typically five to ten in most data warehouses. During this time, changes do occur in the source systems, data extraction methods, data transformation algorithms, and in the structure and content of the data warehouse database itself.
- Metadata in a data warehouse environment must, therefore, keep track of the revisions.
- Metadata management must provide means for capturing and storing metadata with proper versioning to indicate its time-variant feature.

Variety of Metadata Sources

- Metadata for a data warehouse never comes from a single source. CASE tools, the source operational systems, data extraction tools, data transformation tools, the data dictionary definitions, and other sources all contribute to the data warehouse metadata.
- Metadata management, therefore, must be open enough to capture metadata from a large variety of sources.

Metadata Integration

- All these elements must be integrated and merged in a unified manner for them to be meaningful to end-users.
- Metadata from the data models of the source systems must be integrated with metadata from the data models of the data warehouse databases. The integration must continue further to the front-end tools used by the end-users.

All these are difficult propositions and very challenging.

Metadata Standardization

If data extraction tool and the data transformation tool represent data structures, then both tools must record the metadata about the data structures in the same standard way. The same metadata in different metadata stores of different tools must be represented in the same manner.

Rippling Through of Revisions

Revisions occur in metadata as data or business rules change. As the metadata revisions are tracked in one data warehouse process, the revisions must ripple throughout the data warehouse to the other processes.

Keeping Metadata Synchronized

Metadata about data structures, data elements, events, rules, and so on must be kept synchronized at all times throughout the data warehouse.

Metadata Exchange

While end-users are using the front-end tools for information access, they must be able to view the metadata recorded by back-end tools like the data transformation tool. Free and easy exchange of metadata from one tool to another must be possible

Support for End-Users

Metadata management must provide simple graphical and tabular presentations to end-users, making it easy for them to browse through the metadata and understand the data in the data warehouse purely from a business perspective.

Integration and standardization of metadata are great challenges.

Sources of Metadata

As tools are used for the various data warehouse processes, metadata gets recorded as a by product. For example, when a data transformation tool is used, the metadata on the source-to-target mappings get recorded as a by product of the process carried out with that tool.

Source Systems

- Data models of operational systems (manual or with CASE tools)
- Definitions of data elements from system documentation
- Physical file layouts and field definitions
- Program specifications
- File layouts and field definitions for data from outside sources
- Other sources such as spreadsheets and manual lists

Data Extraction

- Data on source platforms and connectivity
- Layouts and definitions of selected data sources
- Definitions of fields selected for extraction
- Criteria for merging into initial extract files on each platform
- Rules for standardizing field types and lengths
- Data extraction schedules
- Extraction methods for incremental changes
- Data extraction job streams

Data Transformation and Cleansing

- Specifications for mapping extracted files to data staging files
- Conversion rules for individual files
- Default values for fields with missing values
- Business rules for validity checking
- Sorting and re-sequencing arrangements
- Audit trail for the movement from data extraction to data staging

Data Loading

- Specifications for mapping data staging files to load images
- Rules for assigning keys for each file
- Audit trail for the movement from data staging to load images
- Schedules for full refreshes
- Schedules for incremental loads
- Data loading job streams

Data Storage

- Data models for centralized data warehouse and dependent data marts
- Subject area groupings of tables
- Data models for conformed data marts
- Physical files
- Table and column definitions
- Business rules for validity checking

Information Delivery

- List of query and report tools
- List of predefined queries and reports
- Data model for special databases for OLAP
- Schedules for retrieving data for OLAP

Challenges for Metadata Management

- Although metadata is so vital in a data warehouse environment, flawlessly integrating all the parts of metadata is a difficult task.
- Industry-wide standardization is far from being a reality. Metadata created by a process at one end cannot be viewed through a tool used at another end without going through convoluted transformations.
- These challenges force many data warehouse developers to abandon the requirements for proper metadata management.

Major challenges to be addressed while providing metadata include:

- Each software tool has its own propriety metadata. If you are using several tools in a data warehouse, how can you reconcile the formats?
- No industry-wide accepted standards exist for metadata formats.
- There are conflicting claims on the advantages of a centralized metadata repository as opposed to a collection of fragmented metadata stores.
- There are no easy and accepted methods of passing metadata along the processes as data moves from the source systems to the staging area and thereafter to the data warehouse storage.
- Preserving version control of metadata uniformly throughout the data warehouse is tedious and difficult.
- In a large data warehouse with numerous source systems, unifying the metadata relating to the data sources can be an enormous task.
- You have to deal with conflicting standards, formats, data naming conventions, data definitions, attributes, values, business rules, and units of measure.
- You have to resolve indiscriminate use of aliases and compensate for inadequate data validation rules.

Metadata Repository

- Metadata repository is a general-purpose information directory or cataloguing device to classify, store, and manage metadata.
- Business metadata and technical metadata serve different purposes.

- The end-users need the business metadata, data warehouse developers and administrators require the technical metadata.
- The structures of these two categories of metadata also vary.
- The metadata repository can be thought of as two distinct information directories, one to store business metadata and the other to store technical metadata. This division may also be logical within a single physical repository.

Summary

- Metadata is a critical need for using, building, and administering the data warehouse.
- For end-users, metadata is like a roadmap to the data warehouse contents.
- For IT professionals, metadata supports development and administration functions.
- Metadata has an active role in the data warehouse and assists in the automation of the processes.
- Metadata types may be classified by the three functional areas of the data warehouse, namely, data acquisition, data storage, and information delivery. The types are linked to the processes that take places in these three areas.
- Business metadata connects the business users to the data warehouse. Technical metadata is meant for the IT staff responsible for development and administration.
- Effective metadata must meet a number of requirements. Metadata management is difficult; many challenges need to be faced.
- Universal metadata standardization is still an elusive goal. Lack of standardization inhibits seamless passing of metadata from one tool to another.
- A metadata repository is like a general-purpose information directory that includes several enhancing functions.

Review Questions

1. Explain the importance of metadata a data warehouse environment.
2. Explain how metadata is critical for data warehouse development and administration.
3. Metadata is like a nerve centre. Describe how the concept applies to the data warehouse environment.
4. What are the three major types of metadata in a data warehouse? Briefly mention the purpose of each type.
5. Describe three major reasons why metadata is vital for end-users.
6. Explain why metadata essential for IT.
7. (a) List six processes in which metadata is significant for IT and explain why.
   (b) Identify three processes in which metadata assists in the automation of these processes. Show how metadata plays an active role in these processes.
8. (a) What is the meaning of ‘establishing the context of information’?
   (b) Briefly explain with an example how metadata establishes the context of information in a data warehouse.
9. Identify four metadata types used in each of the three areas of data acquisition, data storage, and information delivery.
10. List any ten examples of business metadata.
11. (a) Identify four major requirements that metadata must satisfy.
    (b) Describe each of these four requirements.
12. As the project manager for the development of the data warehouse for a domestic soft drinks manufacturer, your assignment is to write a proposal for providing metadata. Consider the options and come up with what you think is needed and how you plan to implement a metadata strategy.
13. As the data warehouse administrator, describe all the types of metadata you would need for performing your job. Explain how these types would assist you.

14. You are responsible for training the data warehouse end-users. Write a short procedure for your casual end-users to use the business metadata and run queries. Describe the procedure in user terms without using the word metadata.
15. As the data acquisition specialist, what types of metadata can help you? Choose one of the data acquisition processes and explain the role of metadata in that process.