1010043363 – Data Mining & Business Intelligence

Unit-1

Overview and concepts Data Warehousing and Business Intelligence

Prof. Deepa R

deepa.it@socet.edu.in

Outline

- Why Reporting & Analyzing Data?
- Introduction to Business Intelligence
- Introduction to Data Warehousing
- Features of Data Warehousing
- Introduction to Data marts
- Types of Data Marts
- Meta Data

Why Reporting & Analyzing Data?

- The amount of data stored in databases is growing exponentially & databases are now measured in gigabytes(GBs) and terabytes(TBs).
- However row data does not provide useful information.
- In today's highly competitive business environment, companies need to turn these terabytes of raw data into some **useful information**.
- The general methods of analysis/reporting can be broadly classified into two categories: **non-parametric** analysis & **parametric** analysis
- **Example**
  - Managers will generally be more interested in **actual data** and non-parametric analysis results, while engineers will be more concerned with parametric analysis.


Large amounts of
RAW DATA
EXTRACT
TRANSFORM
LOAD
DATA WAREHOUSE
FORECASTING
$f(x) = \frac{1}{\sigma 2\pi} e^{-(x-\mu)^2/(2\sigma^2)}$
Predicting FUTURE customer behaviour
bell curve
BUSINESS INTELLIGENCE
Useful
Improve STRATEGY
DATA MINING
Searching for hidden patterns ...
012012
012012
012012
Decision-making support
This way
That way
QUERY & REPORTING
Graphical ANALYSIS
Spot NEW OPPORTUNITIES

What is Business Intelligence?

- BI technologies provide **historical, current and predictive views** of business operations.
- Common functions of business intelligence technologies include reporting, online analytical processing, analytics, data mining, process mining, business performance management, text mining, predictive analytics and prescriptive analytics.
- BI technologies can handle **large amounts of structured** and sometimes **unstructured data** to **help business & also identify, develop** new strategic **business opportunities**.
- Identifying new opportunities and implementing an effective strategy based on insights can provide businesses with a **competitive market advantage** and **long-term stability**.

Business Intelligence (Cont..)

- Business intelligence (BI) make up the **strategies and technologies used by enterprises for the data analysis** of business information.
- BI tools access and analyze data sets and present analytical findings in **reports, summaries, dashboards, graphs, charts and maps** to provide users with **detailed intelligence** about the **state of the business.**
- Typical BI infrastructure components are as follows:
  - Software solution for gathering, cleansing, integrating, analyzing and sharing data.
- It produces analysis and provides **believable information** to help **making effective and high quality business decisions**.

Business Intelligence (Cont..)

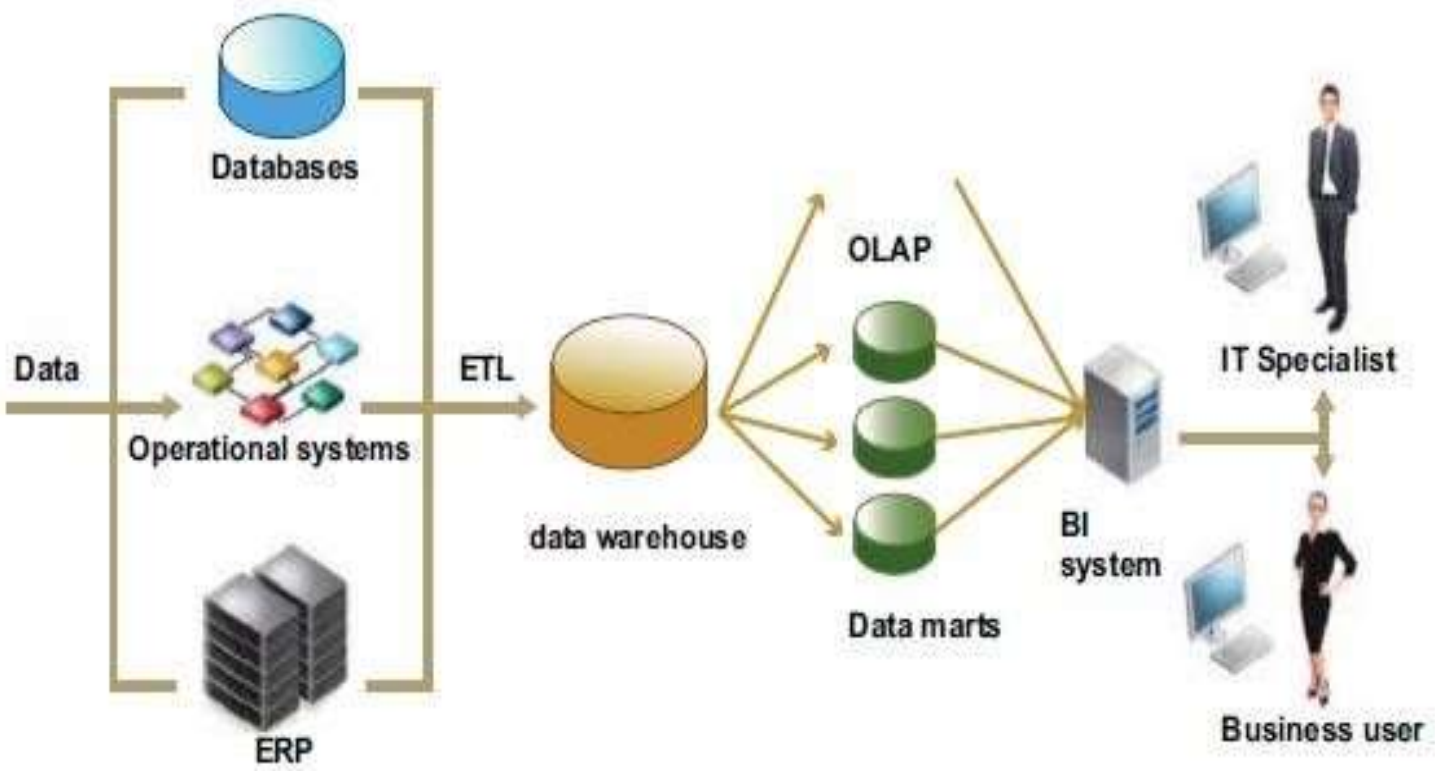
- The most common kinds of **business intelligence systems** are:
  - **MIS** - Management Information Systems
  - **CRM** - Customer Relationship Management
  - **EIS** - Executive Information Systems
  - **DSS** - Decision Support Systems
  - **GIS** - Geographic Information Systems
  - **OLAP** - Online Analytical Processing

BI and DW in today’s perspective

BI and DW in today's perspective :

- While there are varying definitions for *BI*, Forrester defines it broadly as a "set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information that allows business users to make informed business decisions with real-time data that can put a company ahead of its competitors".
- In other words, the high-level goal of BI is to help a business user turn business-related data into actionable knowledge.

BI and DW in today's perspective :



BI and DW in today's perspective :

- BI traditionally focused on reports, dashboards, and answering predefined questions
- Today BI also includes a focus on deeper, exploratory, and interactive analyses of the data using *Business Analytics* such as data mining, predictive analytics, statistical analysis, and natural language processing solutions.
- BI systems evolved by adding layers of data staging to increase the accessibility of the business data to business users.
- Data from the operational systems and ERP were extracted, transformed into a more consumable form (e.g., column names labeled for human rather than computer consumption, errors corrected, duplication eliminated).
- Data from a warehouse were then loaded into OLAP cubes, as well as data marts stored in data warehouses.
- OLAP cubes facilitated the analysis of data over several dimensions.
- Data marts present a subset of the data in the warehouse, tailored to a specific line of business.
- Using Business Intelligence, the business user, with the help of an IT specialist who had set up the system for her, could now more easily access and analyze the data through a BI system.

Introduction to Data Warehouse

- Collections of databases that work together are called **data warehouses.**
- This makes it possible to **integrate data from multiple databases** & it is used to help individuals and organizations **make better decisions.**
- A database consists of one or more files that need to be stored on a computer.
- In large organizations, databases are typically not stored on the individual computers of

Data Warehouse (Cont..)

- A **server** is a computer system that provides a **service over a network**. The server is often located in a specific place with controlled access, so only authorized person can get physical access for it.
- In a typical setting, the database files reside on the server, but it can be accessed from many different computers in the organization.
- As the number and complexity of databases grows, we start referring to them together as a data warehouse.
- **The ultimate goal of a database is not just to store data, but to help businesses make decisions based on that data.**
- A data warehouse supports this goal by providing an architecture and tools to systematically organize and understand data from multiple databases.

Data Warehouse (Cont..)

- According to William H. Inmon, a leading architect in the construction of data warehouse systems, “A data warehouse is a **subject-oriented**, **integrated**, **time-variant**, and **nonvolatile** collection of data in support of management’s decision making process”.
- **Features of Data Warehousing**
  - Subject-oriented
  - Integrated
  - Time-variant
  - Nonvolatile

Features of Data Warehouse

- **Subject-oriented:**
  - A data warehouse is organized around major subjects, such as **customer, supplier, product, and sales.**
  - Rather than concentrating on the day-to-day operations and transaction processing of an organization, a **data warehouse focuses on the modeling and analysis of data for decision makers.**
  - Data warehouses **typically provide a simple and concise view around particular subject** issues by excluding data that are not useful in the decision support process.

Features of Data Warehouse (Cont..)

- **Integrated:**
  - A data warehouse is usually constructed by **integrating multiple heterogeneous sources**, such as **relational databases, flat files, and on-line transaction records.**
  - Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.
- **Time-variant:**
  - Data are stored to provide information from a historical perspective (e.g., the past 5–10 years).
  - Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.

Features of Data Warehouse (Cont..)

- **Nonvolatile:**
  - A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment.
  - Due to this separation, a **data warehouse does not require transaction processing, recovery, and concurrency control mechanisms**.
  - It usually requires only two operations in data accessing: **initial loading of data and access of data**.

Data Warehouse Design Process

- A data warehouse can be built using a **top-down** approach, a **bottom-up** approach, or a **combination of both**.
- **Top Down Approach**
  - The top-down approach starts with the overall design and planning.
  - It is useful in cases where the technology is mature and well known, and where the business problems that must be solved are clear and well understood.
- **Bottom up Approach**
  - The bottom-up approach starts with experiments and prototypes.
  - This is useful in the early stage of business modeling and technology development.
  - It allows an organization to move forward at considerably less expense and to evaluate the benefits of the technology before making significant commitments.
- **Combined Approach**
  - In the combined approach, an organization can exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom-up approach.

Need for data warehousing

Benefits of Data Warehouse

- Understand business trends and make better forecasting decisions.
- Data Warehouses are designed to perform well enormous amounts of data.
- The structure of data warehouses is more accessible for end-users to navigate, understand, and query.
- Queries that would be complex in many normalized databases could be easier to build and maintain in data warehouses.
- Data warehousing is an efficient method to manage demand for lots of information from lots of users.
- Data warehousing provide the capabilities to analyze a large amount of historical data.

Types of Data Warehouse

- The three main types of data warehouses are:
  - **Enterprise Data Warehouse**
  - **Operational Data Store**
  - **Data Mart**

Data Warehouse Types (Cont..)

- **Enterprise Data Warehouse**:
  - Enterprise Data Warehouse is a **centralized warehouse**, which provides **decision support service across the enterprise**.
  - It offers a **unified approach to organizing and representing data**.
  - It also provides the **ability to classify data according to the subject** and give access according to those divisions.
- **Operational Data Store**:
  - Operational Data Store, also called ODS, is data store required when neither data warehouse nor OLTP systems support organizations reporting needs.
  - It is widely preferred for **routine activities like storing records.**.
  - In ODS, Data warehouse is refreshed in real time.
- **Data Mart**:
  - A Data Mart is a subset of the data warehouse.
  - It specially designed for specific segments like sales, finance, sales, or finance.
  - In an independent data mart, data can collect directly from sources.

Introduction to Data Marts

- A data mart is a **simple form of a data warehouse** that is focused on a single subject (or functional area), such as **Sales or Finance or Marketing**.
- Data marts are often built and controlled by a **single department within an organization,** given their single-subject focus, data marts usually draw data from only a few sources.
- The sources could be internal operational systems, a central data warehouse, or external data.

Introduction to Data Marts (Cont..)

- A data mart is a repository of data that is **designed to serve a particular community of knowledge workers**.
- The difference between a data warehouse and a data mart can be confusing because the two terms are sometimes used incorrectly as synonyms.
- A data warehouse is a **central repository for all an organization's data**.
- The goal of a data mart, however, is to meet the particular demands of a specific group of users within the organization, such as human resource management (HRM).
- Generally, an organization's data marts are subsets of the organization's data warehouse.

Reasons for Creating a Data Marts

- Easy access to frequently needed data
- Creates collective view by a group of users
- Improves end-user response time
- Ease of creation
- Lower cost than implementing a full data warehouse
- Potential users are more clearly defined than in a full data warehouse
- Contains only business essential data and is less cluttered

Data Warehouse v/s Data Mart

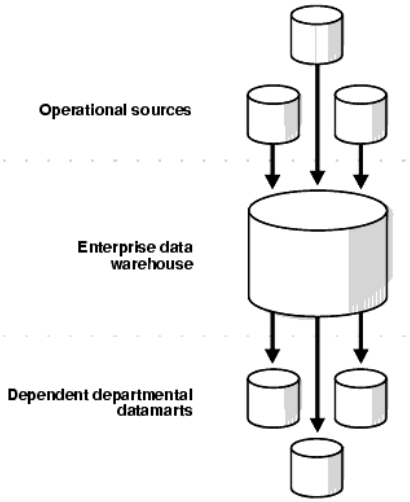
- **Data warehouse**:
  - Holds multiple subject areas
  - Holds very detailed information
  - Works to integrate all data sources
  - Size (typical) 100 GB-TB+
  - Implementation Time : Months to Years
- **Data mart**:
  - Often holds only one subject area- for example, Finance, or Sales
  - May hold more summarized data
  - Concentrates on integrating information from a given subject area or set of source systems
  - Size (typical) < 100GB
  - Implementation Time : Months

Types of Data Marts

- There are three kinds of Data-Marts (DMs), which are as follows:
  1) **Dependent DM**: Created from a data warehouse to a separate physical data-store. (build over data warehouse physically)
  2) **Independent DM**: Created from operational systems and have separate physical data-store.
  3) **Logical or Hybrid DM**: Exists as a subset of data warehouse. (build over data warehouse logically)

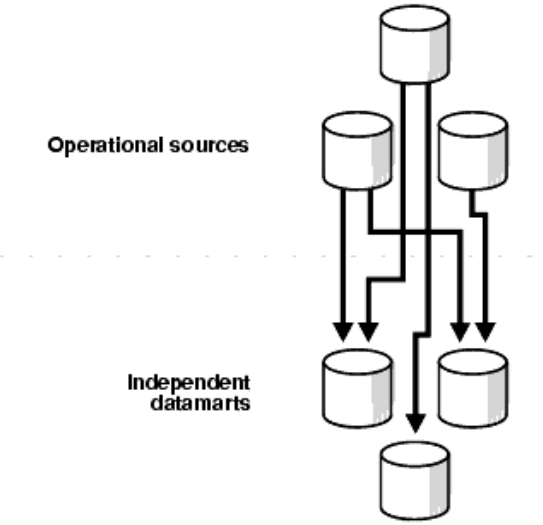
1) Dependent Data Marts

- A dependent data mart allows you to unite your organization's data in one data warehouse.
- This gives you the usual advantages of centralization.



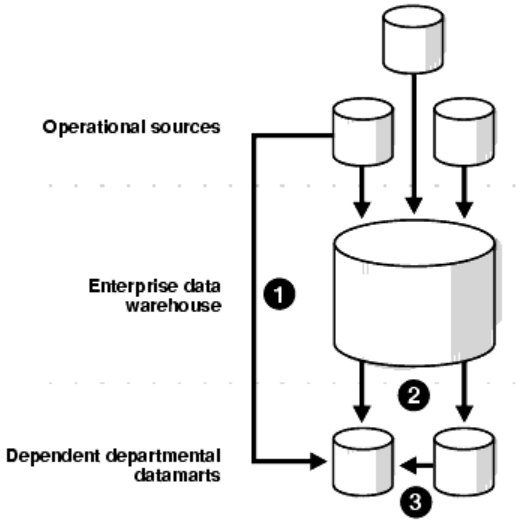
2) Independent Data Marts

- An independent data mart is created without the use of a central data warehouse.
- This could be desirable for smaller groups within an organization.



3) Hybrid Data Mart

- A hybrid data mart allows you to combine input from sources other than a data warehouse.
- This could be useful for many situations, especially when you need ad hoc integration, such as after a new group or product is added to the organization.



Meta data

- Metadata are **data about data**.
- When meta data is used in a data warehouse, that defines warehouse objects.
- Metadata are created for the data names and definitions of the given warehouse.
- Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

Metadata – Example

- **To Describe Meta Data of a Book Store:**
  - Name of Book
  - Summary of the Book
  - The Date of publication
  - High level description of what it contains
  - How you can find the book
  - Author of the book
  - Whether the book is available OR not
- **The information helps you to:**
  - Search for the book
  - Access the book
  - Understand the book before you access OR buy it.

Meta Data Repository

Meta Data Repository :

- Metadata are data about data. When used in a data warehouse, metadata are the data that define warehouse objects.
- Metadata are created for the data names and definitions of the given warehouse.
- Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes.

- **A metadata repository should contain the following:**

  - A description of the structure of the data warehouse, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents.

  - Operational metadata, which include data lineage (history of migrated data and the sequence of transformations applied to it), currency of data (active, archived, or purged), and monitoring information (warehouse usage statistics, error reports, and audit trails).

Meta Data Repository :

- The algorithms used for summarization, which include measure and dimension definition algorithms, data on granularity, partitions, subject areas, aggregation, summarization and predefined queries and reports.

- The mapping from the operational environment to the data warehouse, which includes source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security (user authorization and access control).

- Data related to system performance, which include indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles.

- Business metadata, which include business terms and definitions, data ownership information, and charging policies.

Thank you!