# Data warehousing
# Components

# What Is Data Warehousing?

- Data Warehousing is an architectural construct of information systems that provides users with current and historical decision support information that is hard to access or present in traditional operational data stores.
- It is a blend of technologies and components aimed at effective integration of operational databases into environment that enables strategic use of data.
- Technologies includes : relational/multidimensional DBMS, client/server architecture, meta data modelling and repositories, GUI etc.,

SSN

# The need for data warehousing

- Business perspective
  - In order to survive and succeed in today's highly competitive global environment:
    - Decisions need to be made quickly and correctly
    - The amount of data doubles every 18 months, which affects response time and the sheer ability to comprehend its content
    - Rapid changes

# Business Problem Definition

- The business problems solved by data warehousing and complementary technologies provide organizations with a sustainable competitive advantage.
- An decision support of business applications helps to take decisions about all aspects of their business which includes
  - Customer retention
  - Sales and customer service
  - Marketing
  - Risk assessment and fraud detection

# Business Problem Definition

- Data warehousing classify the business problems into
  - **Retrospective analysis**
  - **Predictive Analysis**
- **Retrospective analysis** :Focuses on the present and past events.
  - Example: Analysis of the performance of the sales organization for the last 2 years across different geographic regions, demographics, and types of products
- **Predictive analysis**: Focuses on predicting certain events or behaviour based on historical information.
  - Example: predictive model which describes the attrition rates of their customers and define steps that reduce it
- This technique further classified as classification, clustering and segmentation, Associations and sequencing

# Operational Data Vs Informational Data

These differences between the informational and operational databases are summarized in the following table.

| | Operational data | Informational data |
|---|---|---|
| Data content | Current values | Summarized, archived, derived |
| Data organization | By application | By subject |
| Data stability | Dynamic | Static until refreshed |
| Data structure | Optimized for transactions | Optimized for complex queries |
| Access frequency | High | Medium to low |
| Access type | Read/update/delete<br>Field-by-field | Read/aggregate<br>Added to |
| Usage | Predictable<br>Repetitive | Ad hoc, unstructured<br>Heuristic |
| Response time | Subsecond (<1 s) to 2–3 s | Several seconds to minutes |

SSN

# Data Warehouse Characteristics

- A data warehouse can be viewed as an information system with the following attributes:
- –It is a database designed for analytical tasks
- –It's content is periodically updated
- –It contains current and historical data to provide a historical perspective of information
- **"A data warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management decisions"**

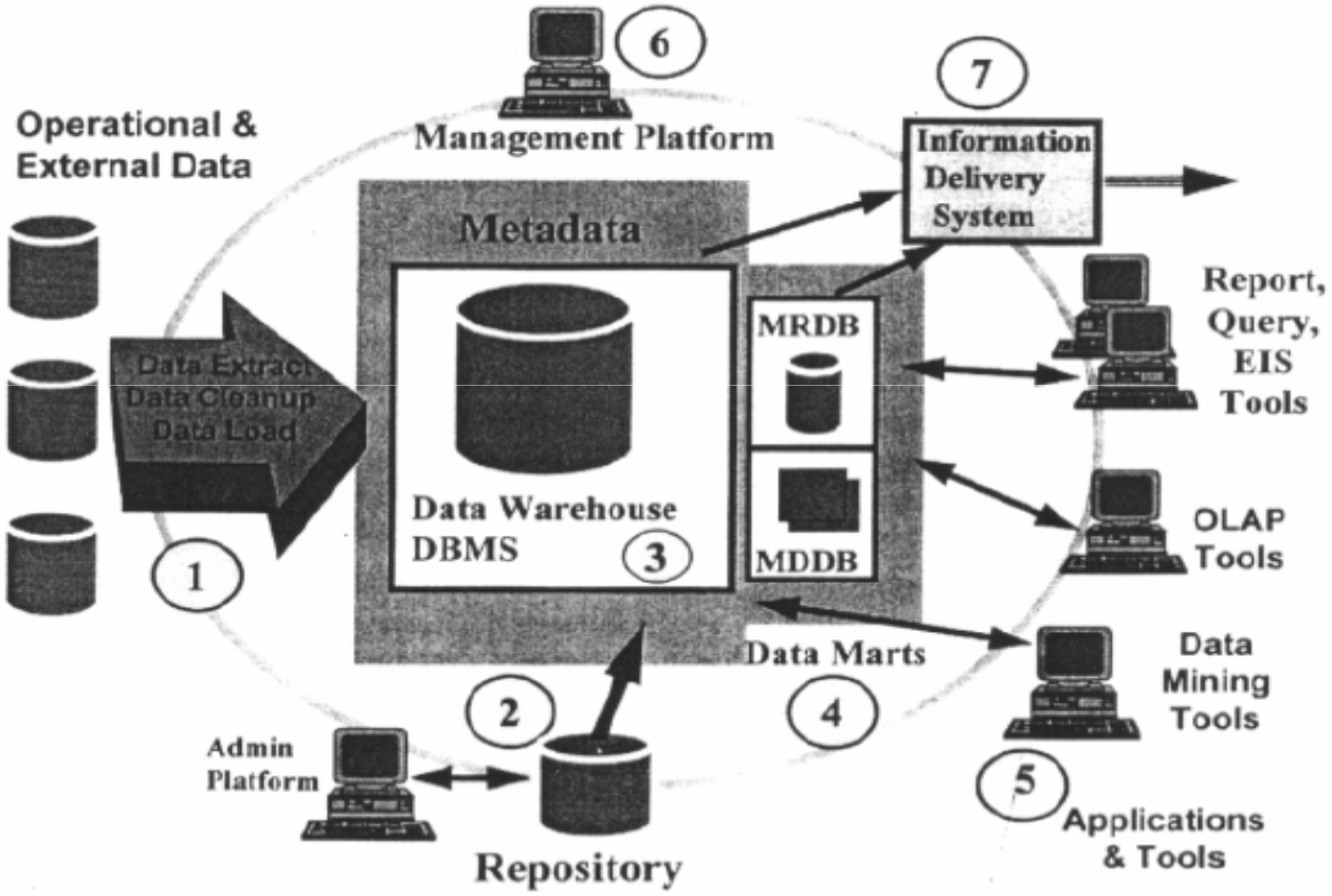
# Data Warehouse Characteristics

- **Data warehouse is subject-oriented:** Organized around major subjects, such as customer, product, sales.
- **Data warehouse is integrated:** It is constructed by integrating multiple, heterogeneous data sources such as relational databases, flat files, on-line transaction records. Data cleaning and data integration techniques have to be applied
- **Data warehouse is time-variant: The time horizon for the data warehouse is** significantly longer than that of operational systems. They provide information from a historical perspective (e.g., past 5-10 years)
- **Data warehouse is non-volatile: DW is a physically separate store of data**

# Architecture of Data warehouse

- The data warehouse architecture is based on relational database management system server that functions as central repository for informational data.
- Source data for the warehouse comes from operational applications
- Data has to be transformed into integrated structure and format.
- Transformation process involves conversation, summarization, filtering and condensations of data.

# Architecture of Data warehouse



# Architecture of Data warehouse

Seven data warehouse components

- **Data sourcing, cleanup, transformation, and migration tools**
- **Metadata repository**
- **Warehouse/database technology**
- **Data marts**
- **Data query, reporting, analysis, and mining tools**
- **Data warehouse administration and management**
- **Information delivery system**

SSN

# Data Warehouse Database

- It is an important concept in the Warehouse environment.
- In additional to transaction operation such as ad hoc query processing, and the need for flexible user view creation including aggregation, multiple joins, and drill-down.
- Parallel relational database designs that require a parallel computing platform.
- Using new index structures to speed up a traditional RDBMS.
- Multidimensional database (MDDBS) that are based on proprietary database technology or implemented using already familiar RDBMS.

# Data sourcing, cleanup, transformation, and migration tools

- This step is used to perform summarizations, key changes, structural changes and condensations needed to transform disparate data into information.
- The functionalities include:
    - Removing unwanted data from operational database
    - Calculating summaries and derived data.
    - Establishing defaults for missing data.
    - Accommodating source data definition changes.

# Data sourcing, cleanup, transformation, and migration tools

- The issues in data sourcing, clean-up and transformation tools are :
- **Database heterogeneity:** DBMS are very different in data model, data access language, data navigation, operation, concurrency, integrity, recovery etc.
- **Data heterogeneity:** This is the difference in the way data is defined and used in different models, different attributes for the same entity and different ways of modelling the same effect.

# Metadata

- Metadata is data about data that describes the warehouse.
- It is used for building, maintaining, and using the data warehouse.
- Metadata provides interactive access to users to help understand content and find data.
- It also includes the information directory. – helps technical and business users to exploit the power of data warehousing.
- This should be accessible by any Web Browser and should run on all major platforms, including MS Windows, Windows NT and UNIX.

# Metadata

- Metadata management is provided via metadata repository and accompanying software.
- Metadata is classified into – Technical metadata and business metadata

# Technical metadata

- Contains information about warehouse data for use **by warehouse designers and administrators** when carrying out warehouse development and management tasks, Documents includes:
- Information about data sources
- Transformation, descriptions, i.e., the mapping methods from operational databases into the warehouse and algorithms used to convert, enhance or transform data.
- Warehouse objects and data structure definitions for data targets.
- The rules used to perform data cleanup and data enhancement.
- Data mapping operations when capturing data from source systems and applying to the target warehouse database.
- Access authorization, backup history, archive history, information delivery history, data acquisition history, data access etc.

# Business metadata

Gives perspective of the information stored in the data warehouse

- Subject areas and information object type, including queries, reports, images, video, and / or audio clips.
- Other information to support all data warehouse components.
- Data warehouse operational information e.g., data history, ownership, extract, audit trail, usage data.

# Metadata

- Metadata management is provided via a metadata repository and accompanying software.
- Metadata software can be used to
  - map the source data to the target database
  - generate code for data transformations
  - integrate and transform the data
  - control moving the data to the warehouse
- The important functional components of the metadata repository is the information directory.
- This directory helps integrate, maintain, and view the contents of the data warehousing system

# Access Tools

- The principal purpose of data warehousing is to provide information to business users for strategic decision making.
- These users interact with the data warehouse using the front-end tools and end user tools.
- **Adhoc requests, regular reports and custom applications** are the primary delivery vehicles for the analysis.
- **Alerts** are exceptional reports allows user to know about certain event has occurred.

# Access Tools

The tools divided into five main groups.

- Data query and reporting tools
- Application development tools
- Executive information system (EIS) tools
- On-line analytical processing tools
- Data mining tools

# Query and Reporting tools

This category can be further divided into two groups.

- **Reporting tools**
- **Managed query tools**
- Reporting tools can be divided into **production reporting tools and desktop report writers.**
- **Production reporting tools** allow companies generate regular operational reports or support high-volume batch jobs.
- **Report writers,** on the other hand, are inexpensive desktop tools designed for end users.
- **Managed query tools** shield end users from the complexities of SQL and database structures by inserting a metalayer between users and the database

# Access Tools

- **Application development tools – Due to the complex set of queries and sophisticated** data models, many data experts rely on application development using graphical data access environments designed primarily for client server environments
- **OLAP tools** – These tools are based on the concept of multi-dimensional databases.
- **Data mining tools** – Data mining is defined as the process of discovering meaningful new correlations, patterns and trends by mining into large amounts of data stored in data warehouse using Artificial Intelligence, statistical and mathematical techniques.

# Access Tools

- **Data mining tools** is used to Discover knowledge: This includes the following functionalities:
  - Segmentation
  - Classification
  - Association
  - Preferencing
- Visualize Data - with a goal to –humanize the mass of data
- Correct Data- Data mining techniques can help identify and correct problems in the most consistent way possible.

# Access Tools

- **Data visualization tools – It is a method of presenting the output of all the previously** mentioned tools.
- Data visualization techniques experiment with various colors, shapes, 3D imaging, sound and virtual reality.

# Data Marts

- Data marts are presented as an inexpensive alternative to a data warehouse.
- It is defined as a **data store that is subsidiary to data warehouse.**
- It is directed at a partition of data (often called subject area) that is created for use of a dedicated group of users.
- Data mart is a physically separate store of data and is normally resident on separate database server.

# Data Marts

- The two types of data marts are – dependent and independent data marts.
- **Dependent data marts**
  - Data is sourced from the data warehouse.
  - Gives advantages of centralization.
- **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.

# Data warehouse Administration and Management

- Managing a data warehouse includes:
  - Data quality checks
  - Data warehouse storage management
  - Auditing and reporting data warehouse usage and status
  - Monitoring updates from multiple sources
  - Managing and updating metadata
  - Backup and recover
  - Security and priority management

# Information Delivery System

- Information delivery Component is used to enable the process of subscribing the DW information and having it delivered to one or more destinations according to some scheduling algorithm.
- IDS distributes warehouse-stored data and other information objects to other data warehouses and end-user products such as spreadsheets and local databases.
- Delivery of info may be based on time of day or on completion of an external event.