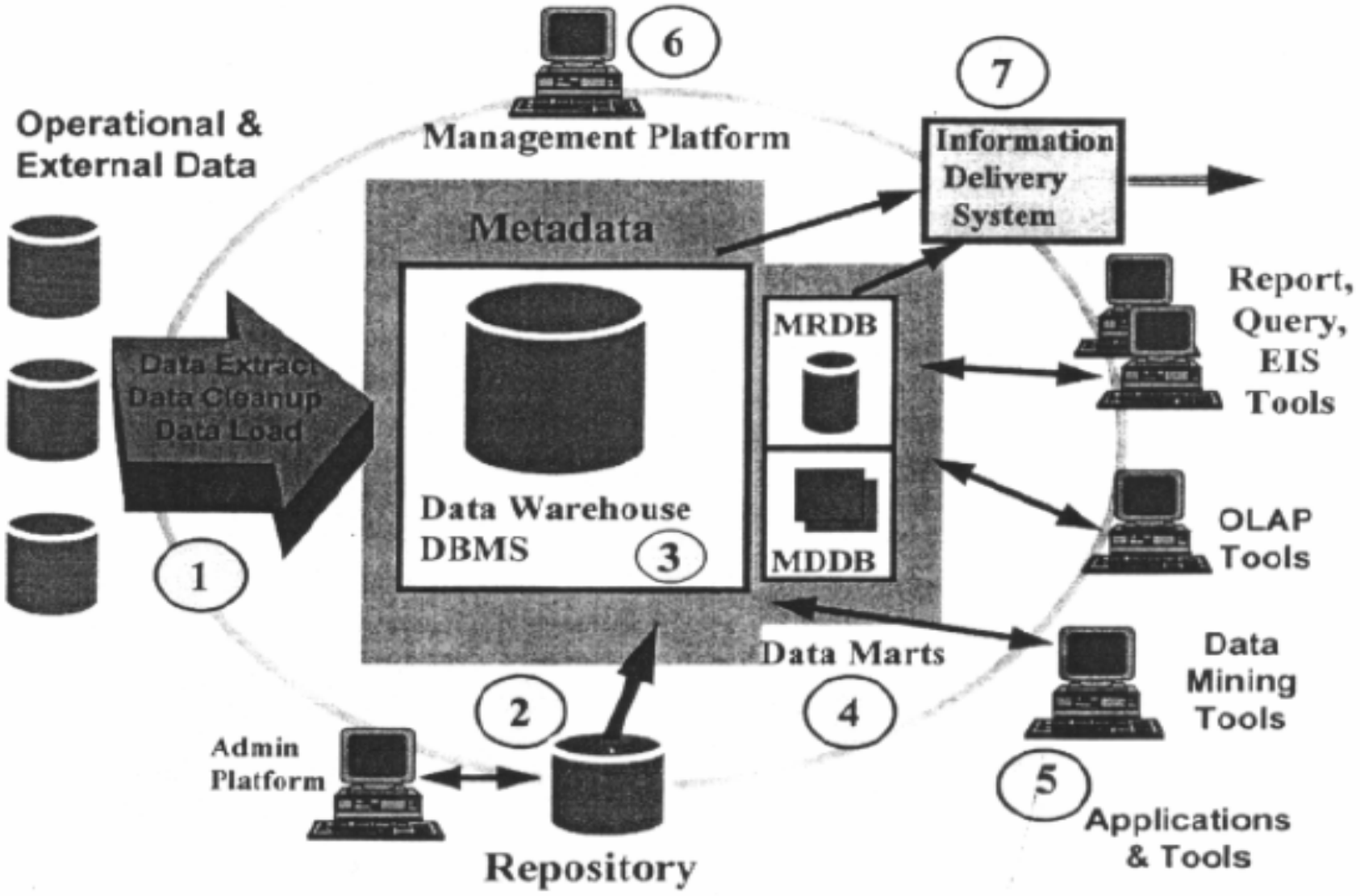
Data Warehouse Architecture

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

Data Warehouse Environment



Data Warehousing Components

- Data warehouse is an environment, not a product which is based on relational database management system that functions as the central repository for informational data.
- The central repository information is surrounded by number of key components designed to make the environment as functional, manageable and accessible.
- The data source for data warehouse is coming from operational applications.
- The data entered into the data warehouse transformed into an integrated structure and format.
- The transformation process involves conversion, summarization, filtering and condensation.
- The data warehouse must be capable of holding and managing large volumes of data as well as different data structures over time.

Operational Data store (ODS)

- ODS can be used for decision support activities
  - Against operational data
  - Act as staging area for data aquisition into Dw
  - ODS is subject oriented
  - ODS is integrated
- ODS VS DW
  - ODS is volatile while DW is non-volatile
  - Contains current data while DW contains current and historical data
  - Contains detailed data only
  - Updata intensive

Data Warehouse Database :

- Central datawarehouse database is a corner stone of the datawarehousing environment
- Database are implemented in RDBMS technology optimized for transactional database processing
- 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-downs.
- Parallel relational database designs that require a parallel computing platform such as symmetric multiprocessors, massively parallel processors clusters of uni or multiprocessors
- Approach for 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.

Sourcing, Acquisition, Cleaning, and Transformation tools

To perform all of the conversations, summarizations, key changes, structural changes, and condensations needed to transform disparate data into information.

- It produces the programs and control statements and maintains the metadata
- The data transformation is required so that the information can by used by decision support tools.
- **The functionality of the tools are listed below:**
  - Removing unwanted data from operational database
  - Converting to common data names and definitions
  - Calculating summarizes and derived data.
  - Establishing default for missing data.
  - Accommodating source data definition changes.

Issues to be considered:

- **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.

Metadata

Metadata is data about data that describes the DW.

- Used for building, maintaining, and using the data warehouse
- Classified into : Technical and Business metadata

Technical metadata :

- About warehouse data for use by warehouse designers and administrators when carrying out warehouse development and management tasks
- 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.

Metadata

- 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

- Users easy to understand and 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.
    - It helps to map source data to the target database
    - Generate code for data transformations
    - Integrate and transform the data
    - Control moving 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
    - Act as a gateway to the dW
    - Should be searchable by business oriented keywords
    - Should support distribution of the query results
    - Should support sharing of information objects such as queries, reports , data collections and subscriptions between users

Access Tools

- Principal purpose of DW is to provide information to business users for strategic decision making
- Users interact with DW using front-end tools
- Front-end tools, ad hoc request, regular reports, and custom applications are the primary delivery of the analysis.
- Exceptional reporting Alerts, which let a user know when a certain event has occurred.
- End user tools uses metadata definitions to obtain access to data stored in the DW
- Some tools act as specialized data stores for a given end-user tool covering a specific subject data mart

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

- Query and reporting tools are used to generate query and report
- 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 will let 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 act as a metalayer between users and database

Application development tools: This is a graphical data access environment which integrates OLAP tools with data warehouse and can be used to access all db systems.

OLAP:

- used to analyze the data in multi dimensional and complex views.
- To enable multidimensional properties it uses MDDB and MRDB
- Multidimensional model supported by multidimensional database
- Relational database designed to enable multidimensional properties

Data mining

To discovery meaningful new correlations, patterns, and trends by digging into (mining) large amount of data stored in warehouse using artificial-intelligence (AI) and statistical and mathematical techniques

Data visualization

- Presenting the output of all the previously mentioned tools
- Displaying complex relationships and patterns on a 2D monitor.
- Colors, shapes, 3-D images, sound, and virtual reality

Data Marts

Data store that is subsidiary to data warehouse.

- It is partition of data that is created for the use of dedicated group of users focus on dedicated subjects.
- They are used for rapid delivery of enhanced decision support functionality to end users.
- In most instance, the data mart is physically separate store of data and is normally resident on separate database server
- Sometimes it may be placed on the DW database than separate store of data

Data Warehouse administration and Management

Managing data warehouse includes

- Security and priority management
- Monitoring updates form multiple sources
- Data quality checks
- Managing and updating metadata
- Auditing and reporting data warehouse usage and status
- Replicating, sub setting, and distributing data
- Backup and recover
- Data warehouse storage management

Information Delivery System

Information delivery component is used to enable the process of subscribing for data warehouse info.

- It distributes warehouse-stored data and other information objects to other data warehouses and end-users