innovate achieve lead

SS ZG515 - Data Warehousing

BITS Pilani
Pilani Campus

Dr. Yashvardhan Sharma
CSIS Dept., BITS-Pilani

The goals of a Data Warehouse

- The data warehouse must make an organization's information easily accessible.
- The data warehouse must present the organization's information consistently.
- The data warehouse must be adaptive and resilient to change.
- The data warehouse must be a secure bastion that protects our information assets.
- The data warehouse must serve as the foundation for improved decision making.
- The business community must accept the data warehouse if it is to be deemed successful.

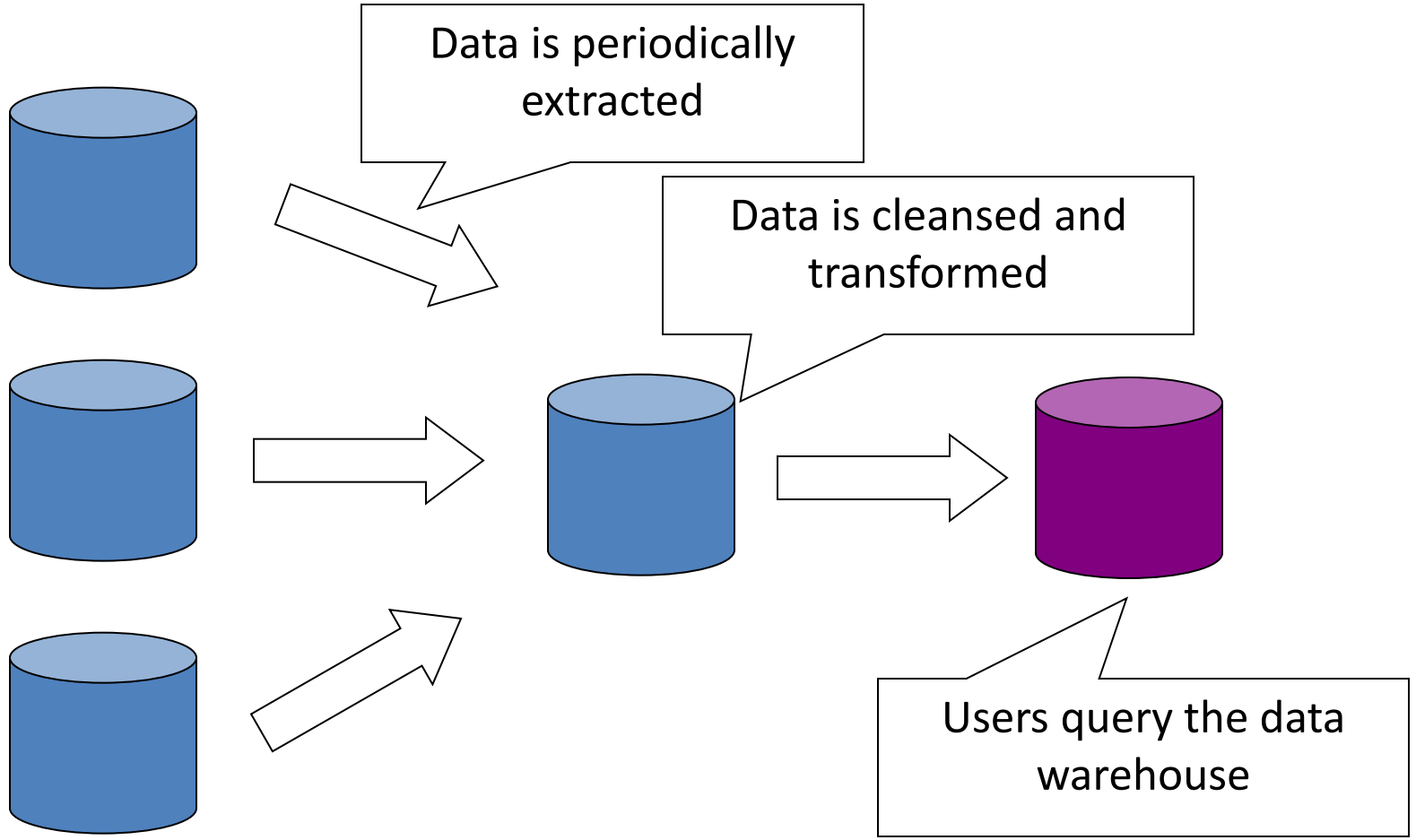
Data Warehouse Usage

- Three kinds of data warehouse applications
  - Information processing
    - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
  - Analytical processing
    - multidimensional analysis of data warehouse data
    - supports basic OLAP operations, slice-dice, drilling, pivoting
  - Data mining
    - knowledge discovery from hidden patterns
    - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools

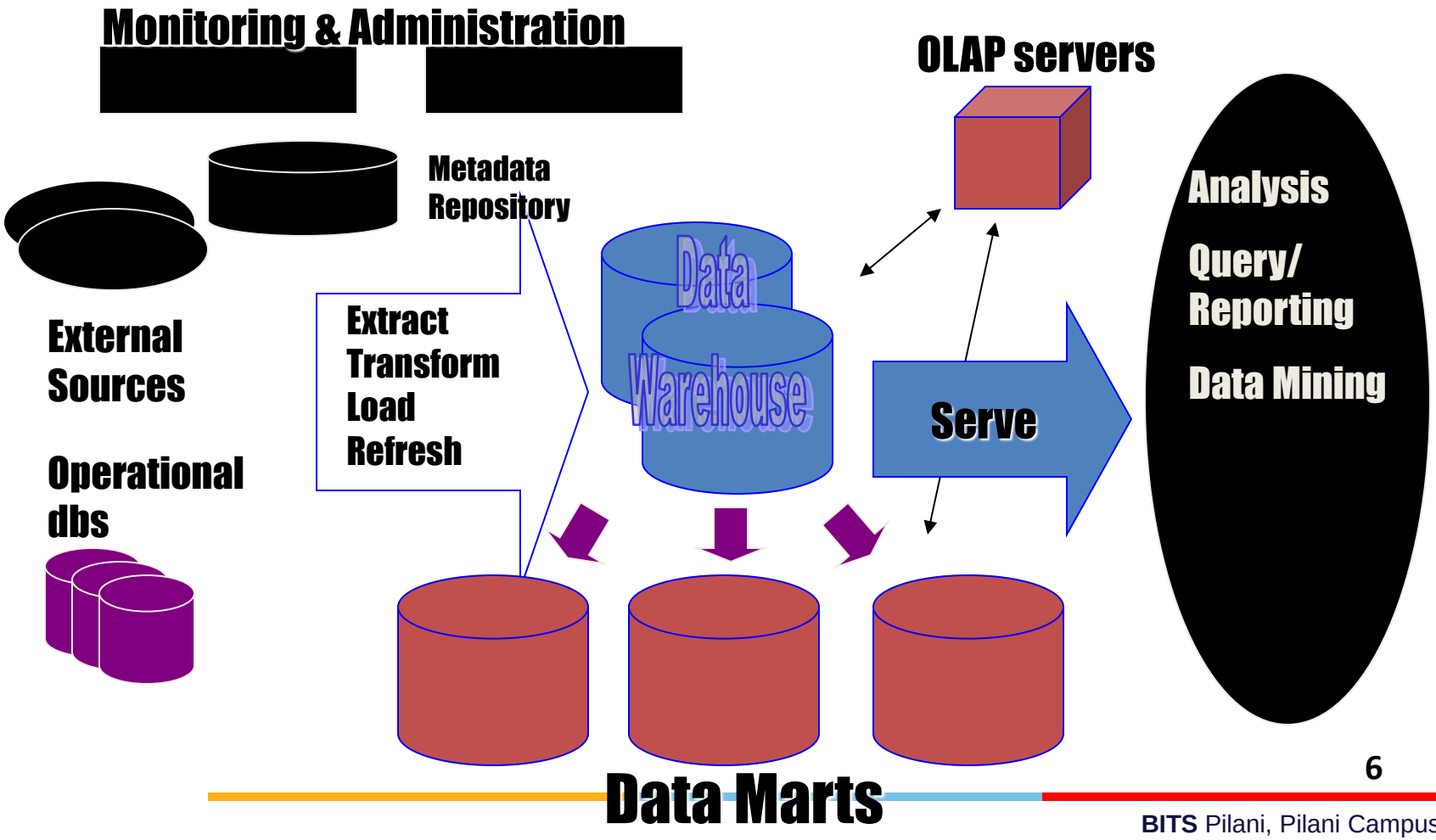
Problems with Data Warehousing

- Underestimation of resources for data loading
- Hidden problems with source systems
- Required data not captured
- Increased end-user demands
- Data homogenization
- High demand for resources
- Data ownership
- High maintenance
- Long-duration projects
- Complexity of integration

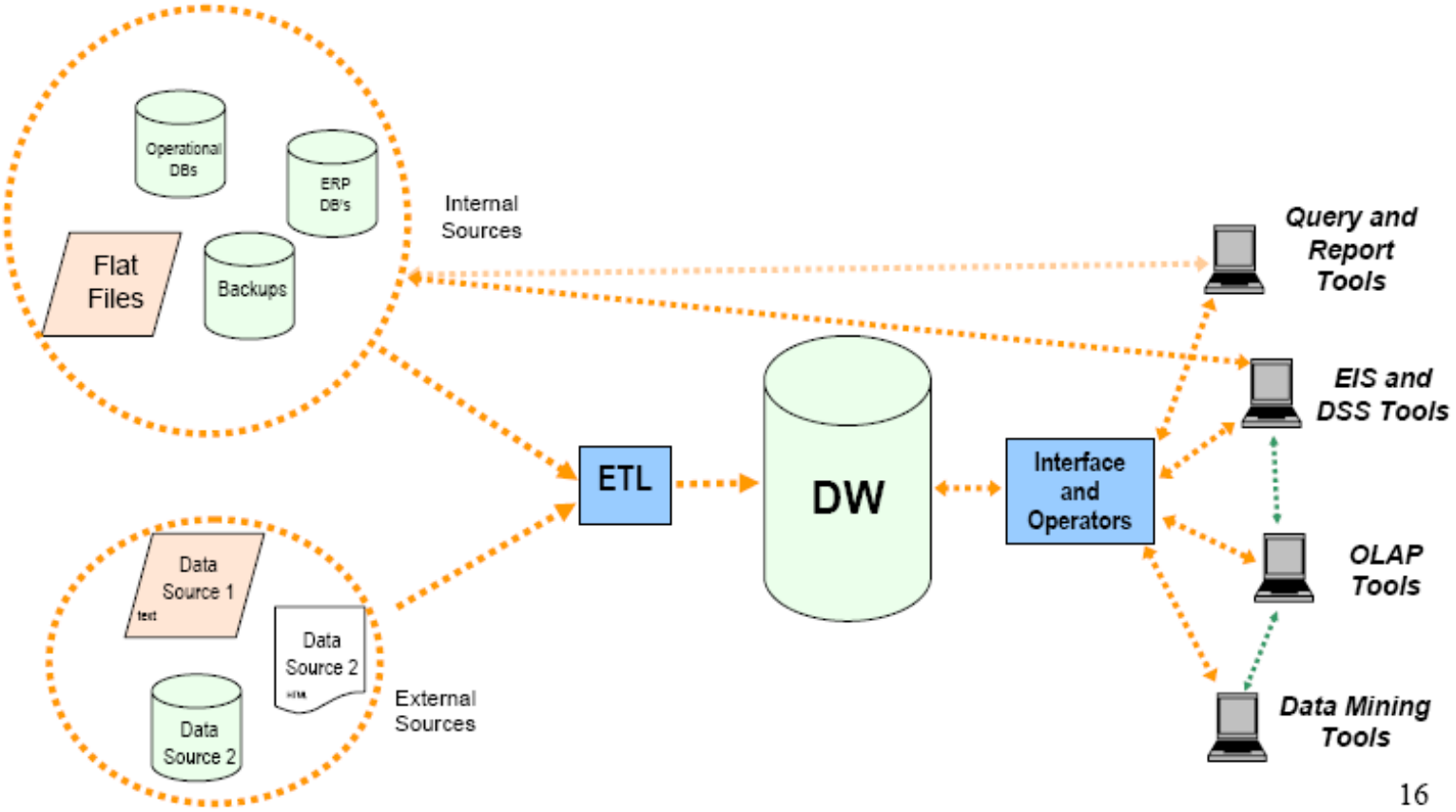
Loading the Data Warehouse



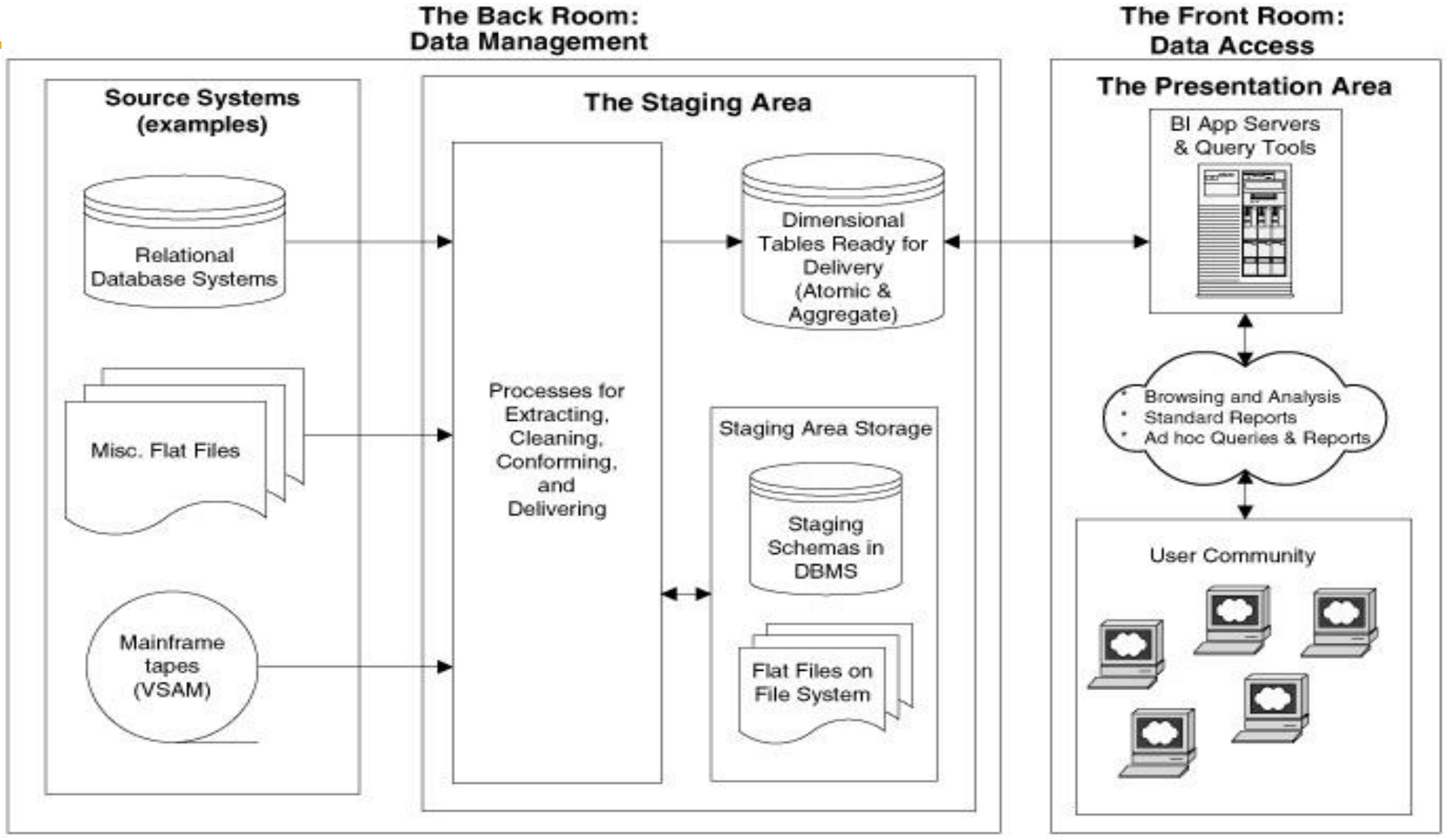
Data Warehousing Architecture



Data Warehousing Architecture



Data Warehouse Architecture



Data Warehouse COMPONENTS

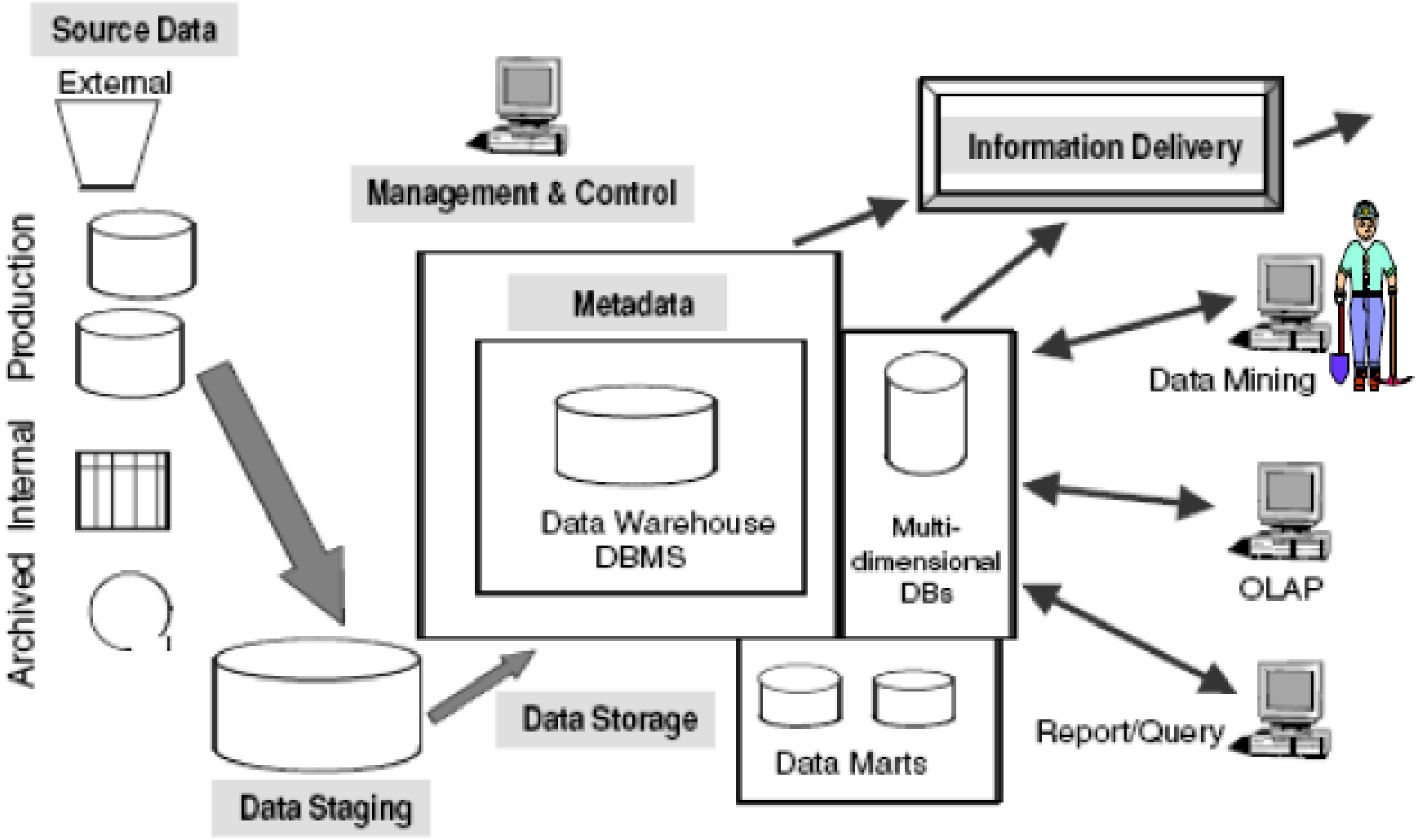


Figure 2-6 Data warehouse: building blocks or components.

Data Warehouse COMPONENTS

- **Source Data Component**
  - ***Production Data.***
  - ***Internal Data.***
  - ***Archived Data.***
  - ***External Data.***
- **Data Staging Component**
  - ***Data Extraction***
  - ***Data Transformation.***
  - ***Data Loading.***

Data Loading

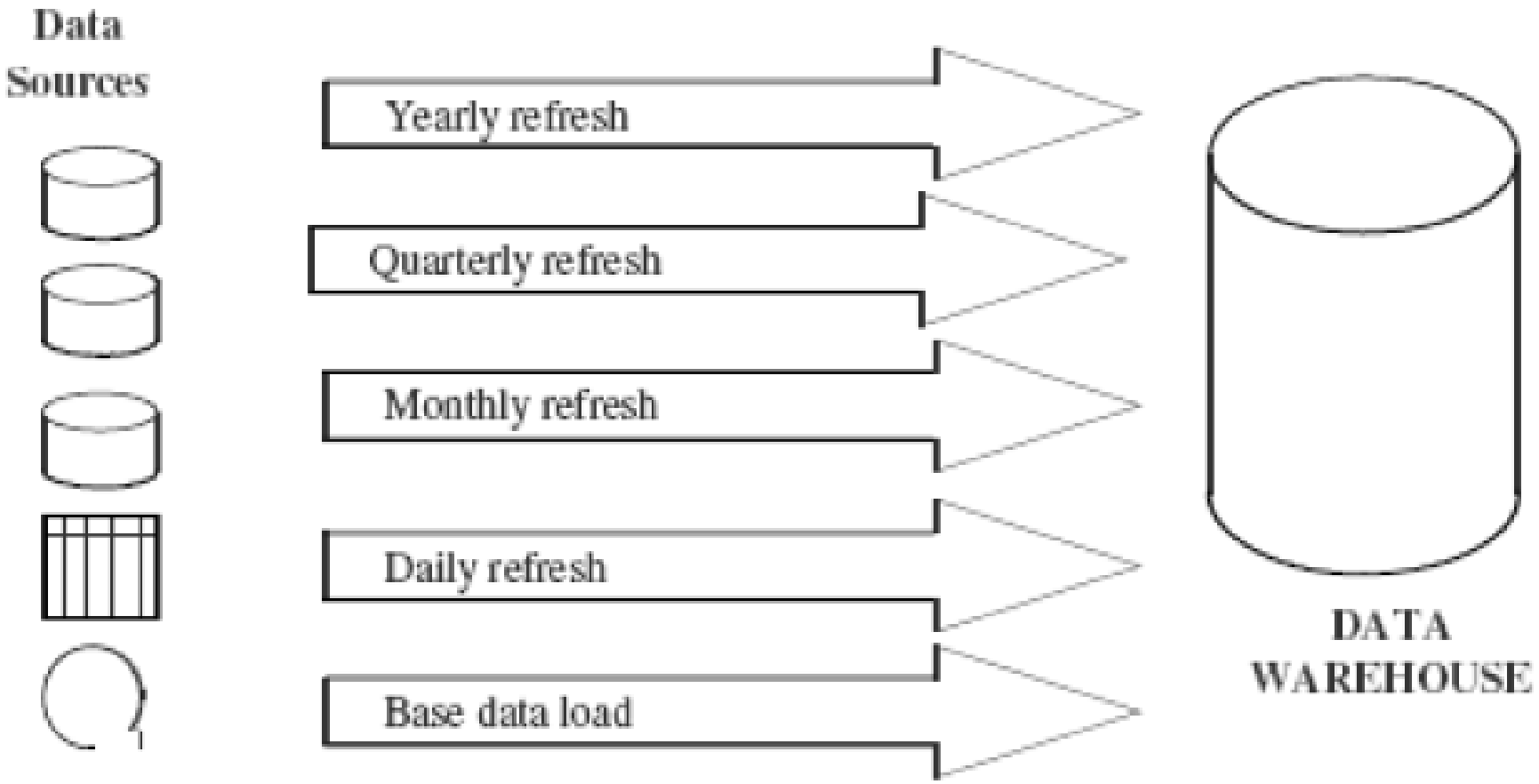


Figure 2-7 Data movements to the data warehouse.

Data Storage Component

- 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

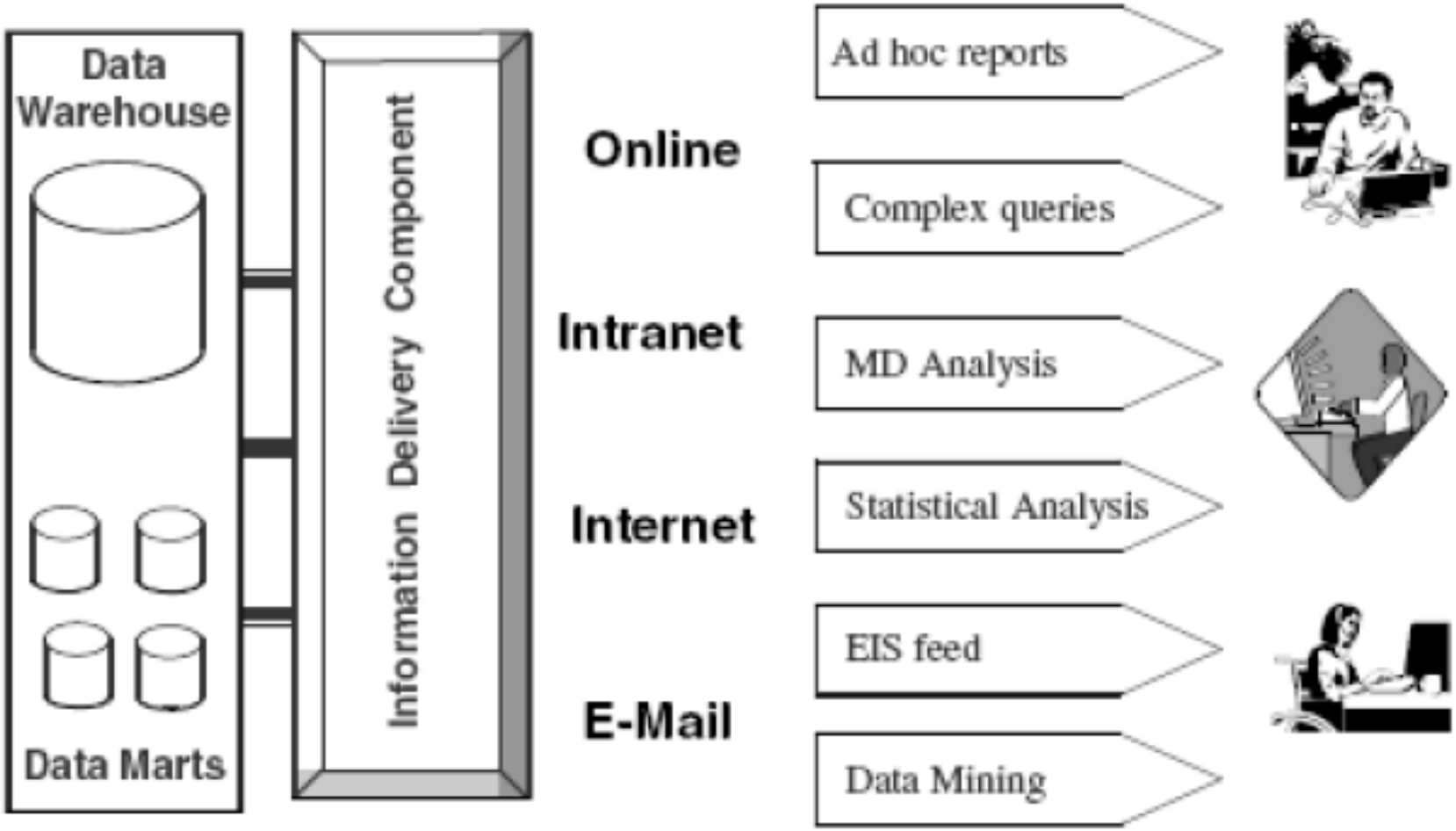


Figure 2-8 Information delivery component.

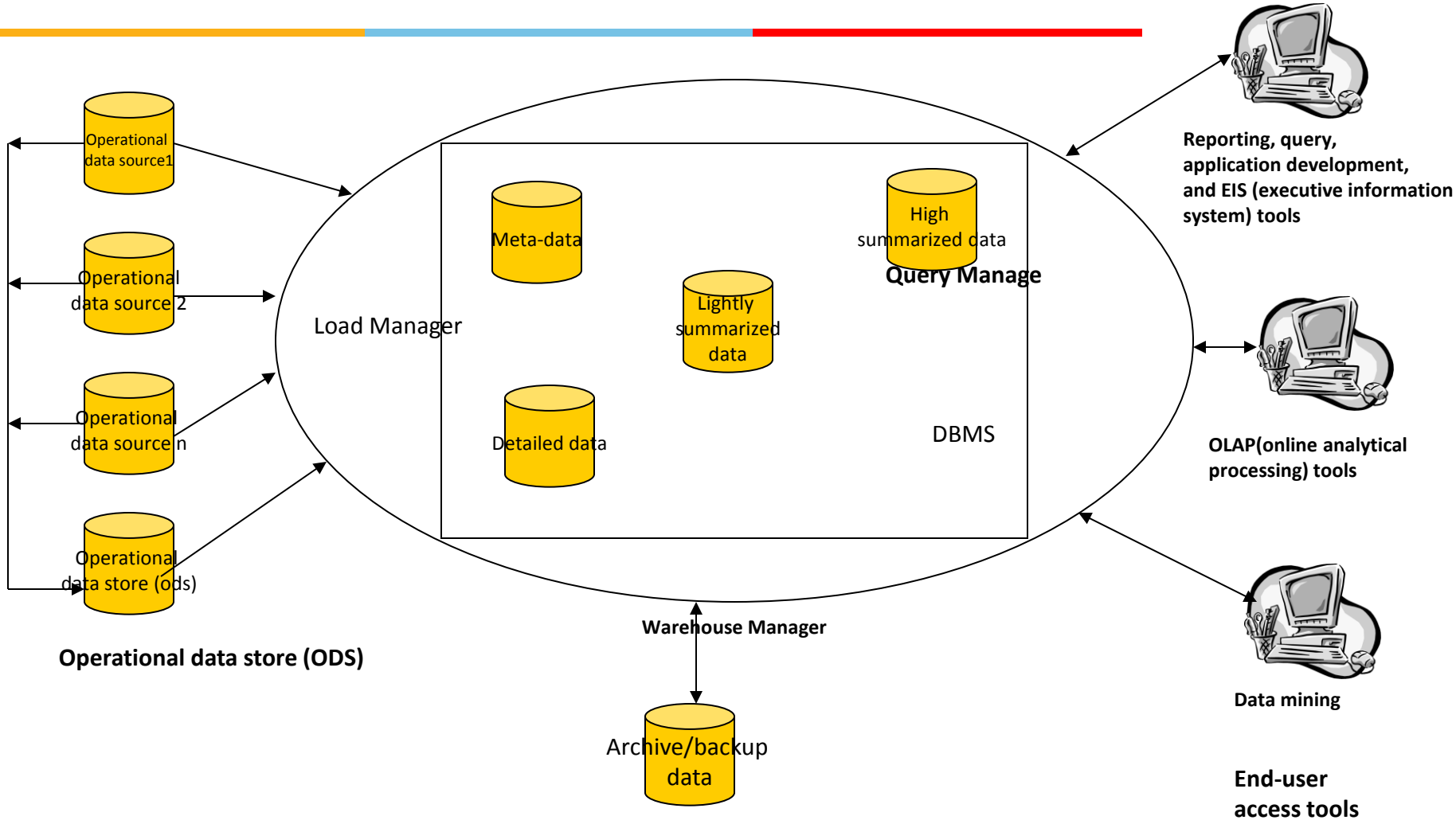
Metadata Component

- Metadata in a data warehouse is similar to a data dictionary, but much more than a data dictionary.
- **Types of Metadata**
  - Operational Metadata
  - Extraction and Transformation Metadata
  - End-User Metadata
- More Details in Chapter 9.

Why Meta Data: Special Significance

- First, it acts as the glue that connects all parts of the data warehouse.
- Next, it provides information about the contents and structures to the developers.
- Finally, it opens the door to the end-users and makes the contents recognizable in their own terms.

The architecture



Typical architecture of a data warehouse

The main components

- Operational data sources → for the DW is supplied from mainframe operational data held in first generation hierarchical and network databases, departmental data held in proprietary file systems, private data held on workstaions and private serves and external systems such as the Internet, commercially available DB, or DB assoicated with and organization's suppliers or customers
- Operational datastore (ODS) → is a repository of current and integrated operational data used for analysis. It is often structured and supplied with data in the same way as the data warehouse, but may in fact simply act as a staging area for data to be moved into the warehouse

The main components

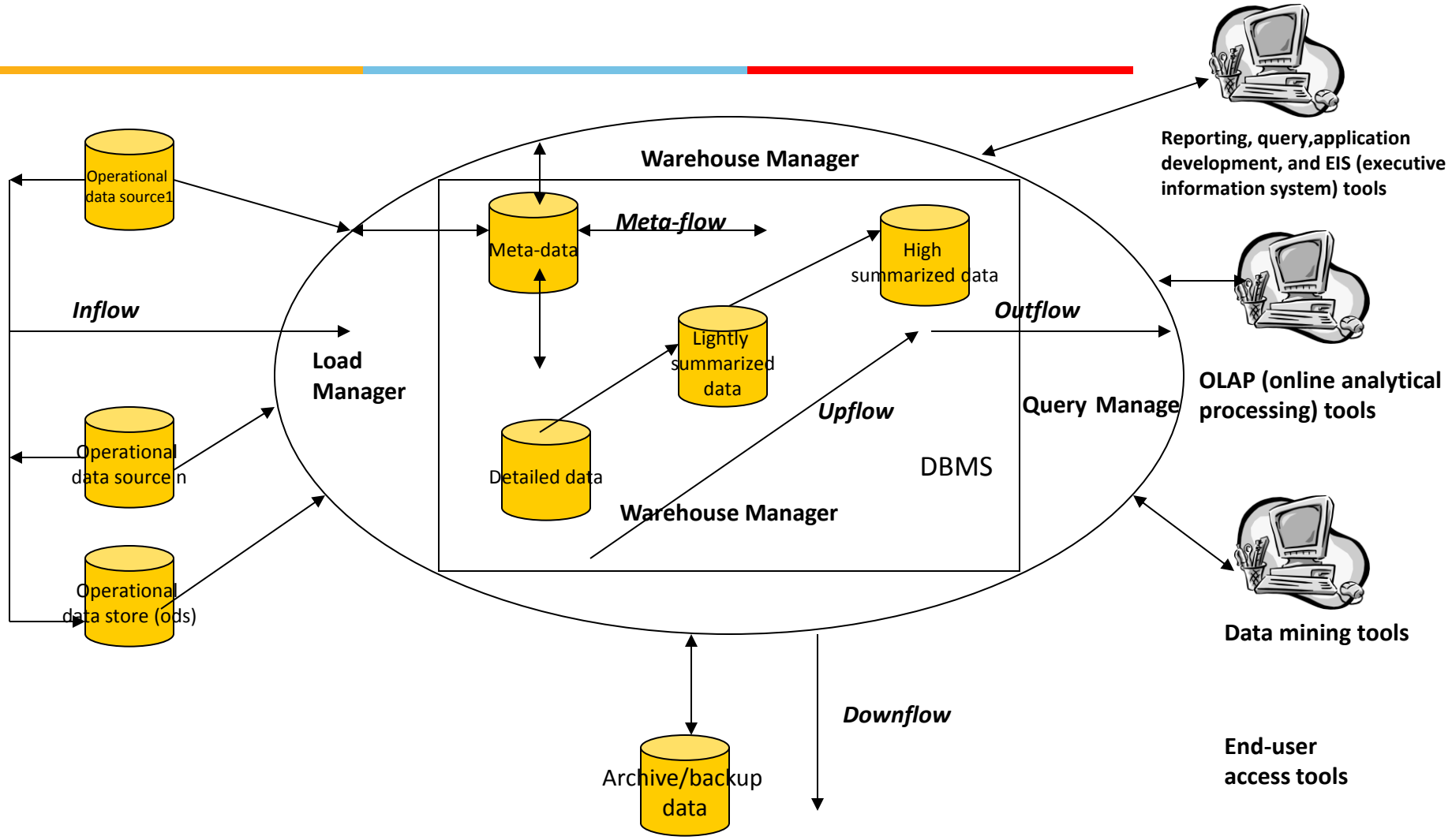
- load manager → also called the *frontend* component, it performance all the operations associated with the extraction and loading of data into the warehouse. These operations include simple transformations of the data to prepare the data for entry into the warehouse
- warehouse manager → performs all the operations associated with the management of the data in the warehouse. The operations performed by this component include analysis of data to ensure consistency, transformation and merging of source data, creation of indexes and views, generation of denormalizations and aggregations, and archiving and backing-up data

The main components

- query manager → also called backend component, it performs all the operations associated with the management of user queries. The operations performed by this component include directing queries to the appropriate tables and scheduling the execution of queries
- detailed, lightly and lightly summarized data, archive/backup data
- meta-data
- end-user access tools → can be categorized into five main groups: data reporting and query tools, application development tools, executive information system (EIS) tools, online analytical processing (OLAP) tools, and data mining tools

Data flows

- Inflow- The processes associated with the extraction, cleansing, and loading of the data from the source systems into the data warehouse.
- upflow- The process associated with adding value to the data in the warehouse through summarizing, packaging, packaging, and distribution of the data
- downflow- The processes associated with archiving and backing-up of data in the warehouse
- outflow- The process associated with making the data availabe to the end-users
- Meta-flow- The processes associated with the management of the meta-data



Information flows of a data warehouse

Tools and Technologies

- The critical steps in the construction of a data warehouse:

  a. Extraction

  b. Cleansing

  c. Transformation

- after the critical steps, loading the results into target system can be carried out either by separate products, or by a single, categories:

- code generators
- database data replication tools
- dynamic transformation engines

Data Cleaning

- Why?
  - Data warehouse contains data that is analyzed for business decisions
  - More data and multiple sources could mean more errors in the data and harder to trace such errors
  - Results in incorrect analysis
- Detecting data anomalies and rectifying them early has huge payoffs
- Long Term Solution
  - Change business practices and data entry tools
  - Repository for meta-data

Soundex Algorithms

- Misspelled terms
- For example NAMES
- Phonetic algorithms – can find similar sounding names
- Based on the six phonetic classifications of human speech sounds

Data Warehouse Design

- OLTP Systems are Data Capture Systems
- "DATA IN" systems
- DW are "DATA OUT" systems

Analyzing the DATA

- Active Analysis – User Queries
  - User-guided data analysis
  - Show me how X varies with Y
  - OLAP
- Automated Analysis – Data Mining
  - What's in there?
  - Set the computer FREE on your data
  - Supervised Learning (classification)
  - Unsupervised Learning (clustering)

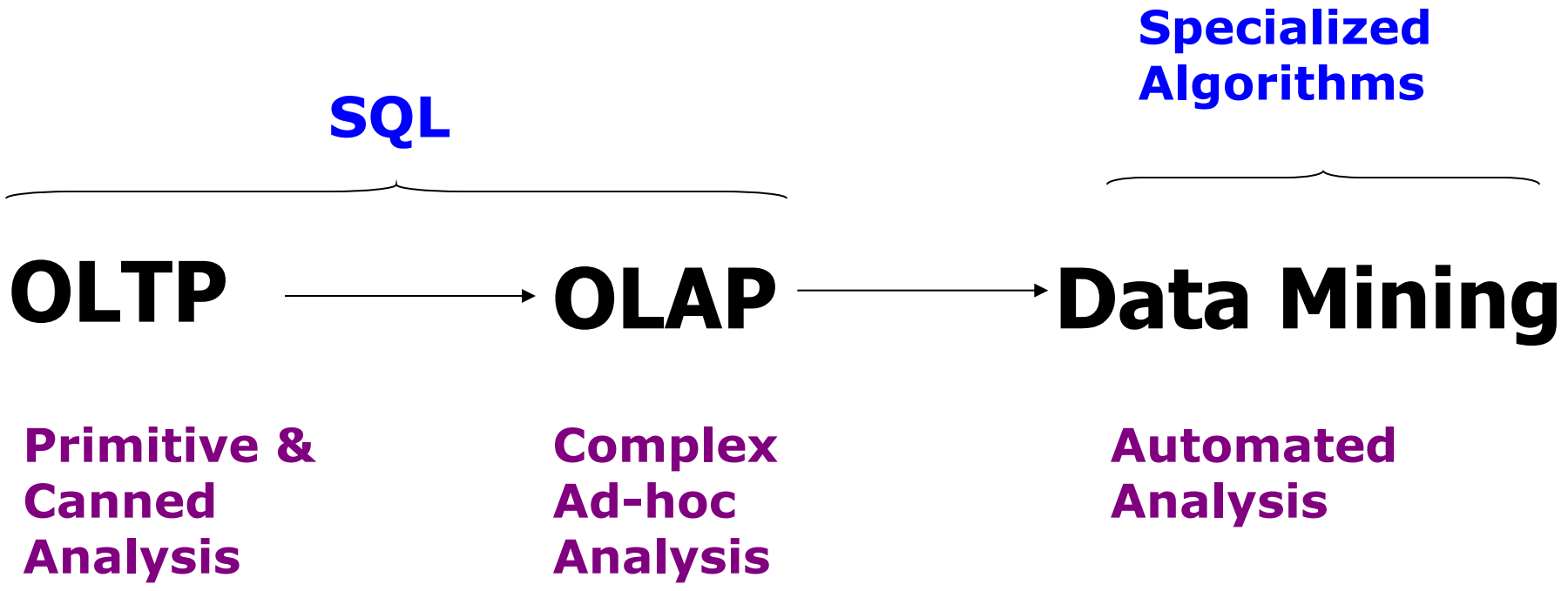
OLAP Queries

- How much of product P1 was sold in 2009 state wise?
- Top 5 selling products in 2010
- Total Sales in Q1 of FY 2008-09?
- Color wise sales figure of cars from 2008 to 2010
- Model wise sales of cars for the month of Jan from 2006 to 2010

Data Mining Investigations

- Which type of customers are more likely to spend most with us in the coming year?
- What additional products are most likely to be sold to customers who buy sportswear?
- In which area should we open a new store in the next year?
- What are the characteristics of customers most likely to default on their loans before the year is out?

Continuum of Analysis



Net Resources

- Online Resources
  - The Data Warehousing Institute www.tdwi.org
  - Data Warehousing on www www.datawarehousing.org www.datawarehousing.com

- Online Magazines & Periodicals
  - www.intelligententerprise.com
  - www.dmreview.com
  - www.cio.com
  - www.daniel-lemire.com/OLAP/index.html