Armazéns de Dados

Departamento de Engenharia Informática (DEI/ISEP)
Paulo Oliveira
PJO@isep.ipp.pt

Adaptado do Original de:
Fátima Rodrigues (DEI/ISEP)

Introduction to Data Warehouses

Summary

- Definition of Data Warehouse
  - Based on four keywords
- Data Warehouse Characteristics
- A Data Warehouse Architecture
- Definition of Data Mart

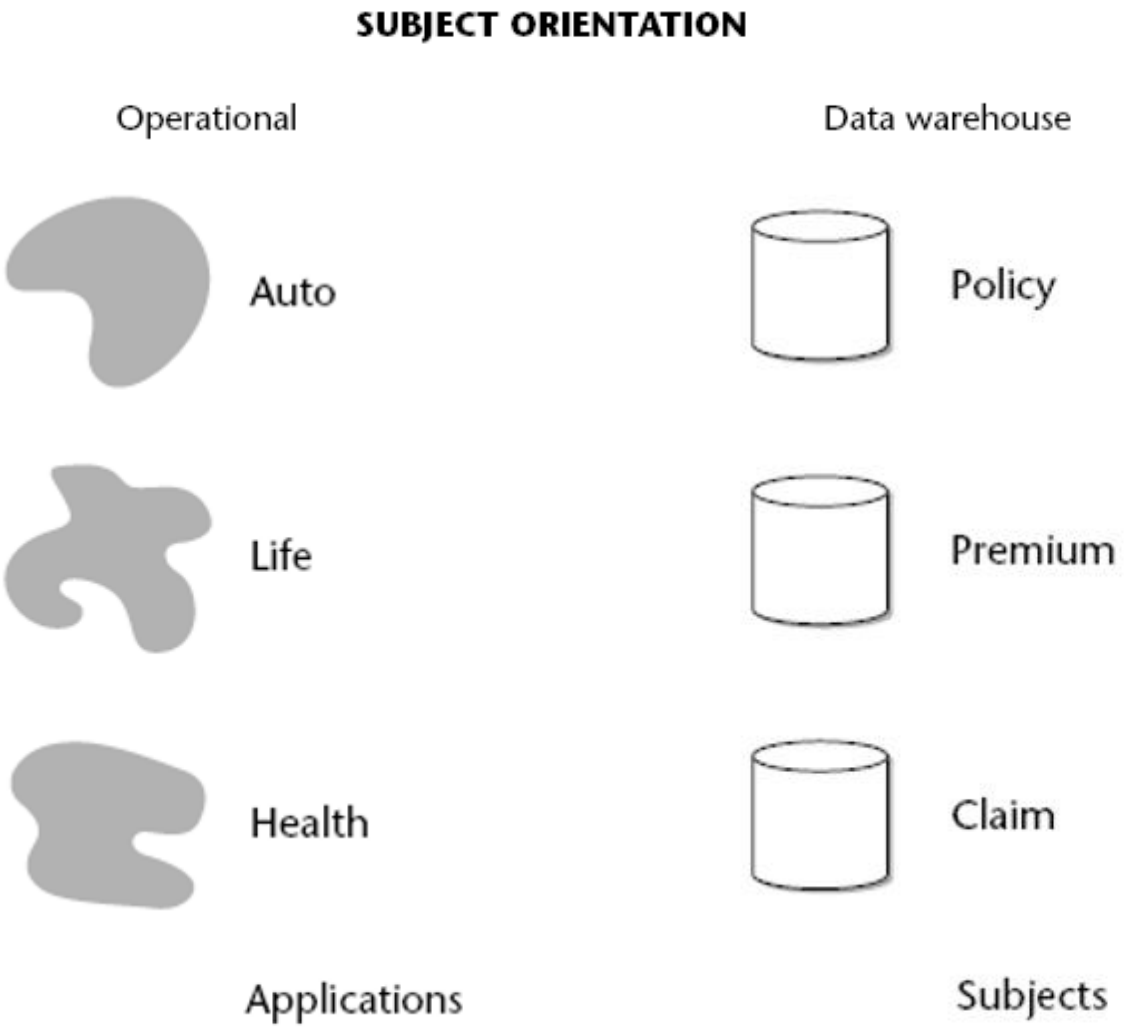
Definition of Data Warehouse (DW)

- **Data warehouse is a subject-oriented, integrated, time-variant, non-updatable collection of data used in support of the decision-making process and business intelligence** [Inmon and Hackathorn, 1994]

Data Warehouse - Subject-Oriented

- Organized around key subjects of the enterprise, such as **sales, purchases, inscriptions**
- Focused on the **modeling and analysis of data for decision makers**
  - not on daily operations or transaction processing
- Provide **a simple and concise** view around particular subject issues, **excluding data not useful in the decision support process**

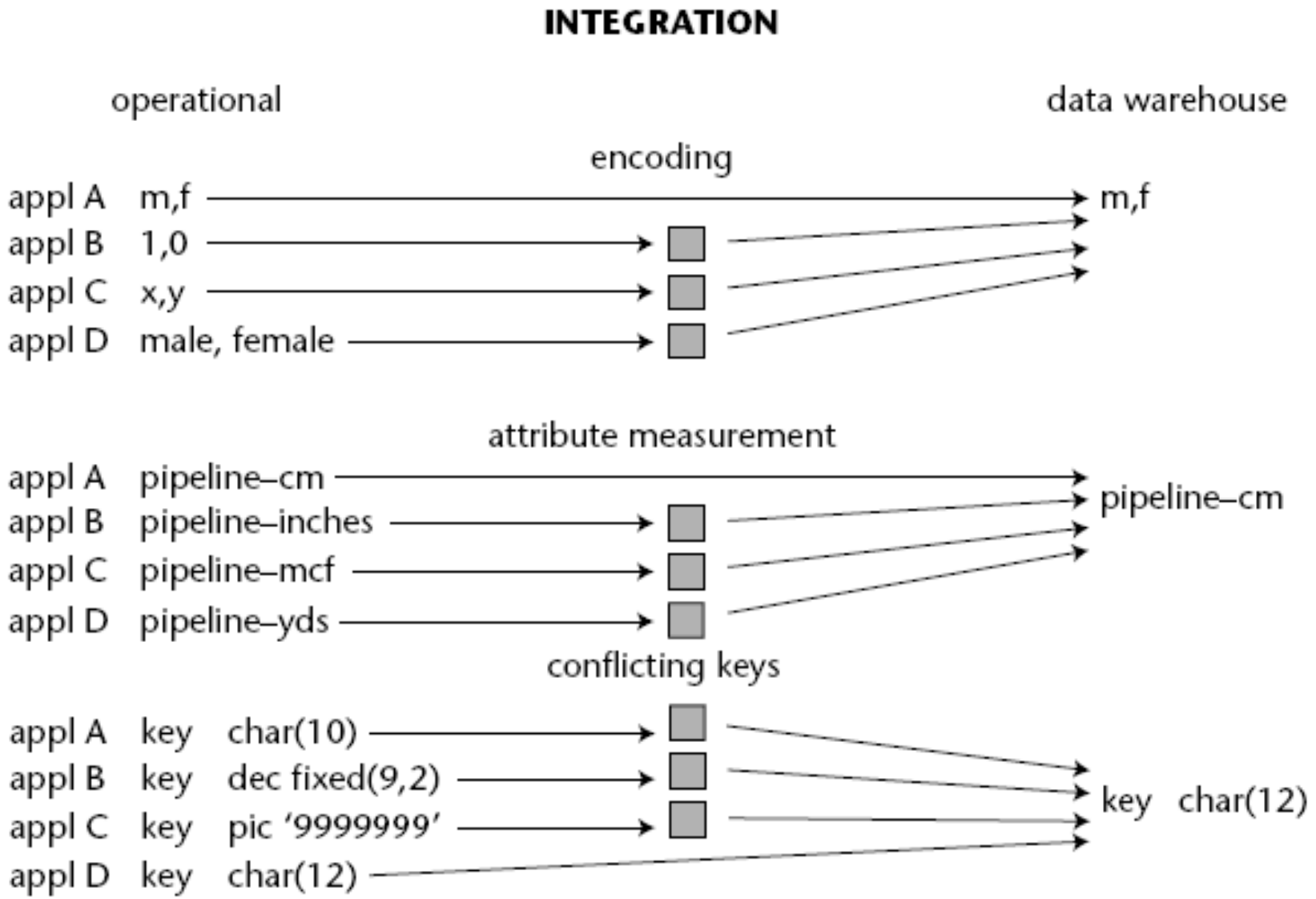
Data Warehouse - Subject-Oriented



Data Warehouse - Integrated

- Constructed by integrating **multiple, heterogeneous data sources**
  - relational databases, text files, binary files, …
- **Data cleaning** and **data transformation** techniques are applied
  - consistent naming conventions, formats, encoding structures, attribute measures, … among different data sources (internal and external)
- DW holds the **version of "the truth"**

Data Warehouse - Integrated



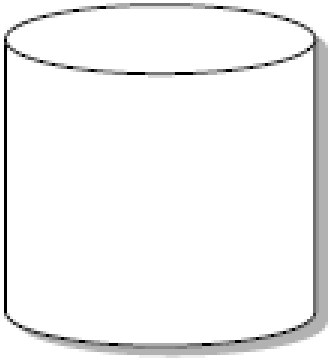
Data Warehouse - Time Variant

- Time horizon is longer than that of On-Line Transaction Processing (OLTP) systems
  - Operational database: current data
  - Data warehouse: provide **information from a historical perspective** (e.g., past 5-10 years)
  - Snapshots of OLTP systems are moved to the DW as a series of data layers – much like **geologic layers**
- Some form of **time marking** to show the moment in time during which **the record is/was accurate**

Data Warehouse - Time Variant

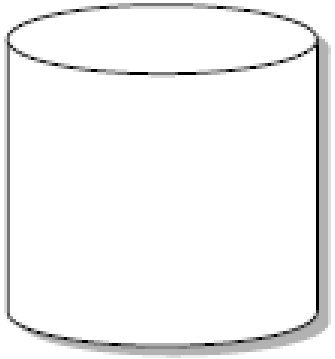
TIME VARIANCY

Operational



- Time horizon – current
- Update of records
- Schema may or may not contain an element of time

Data warehouse

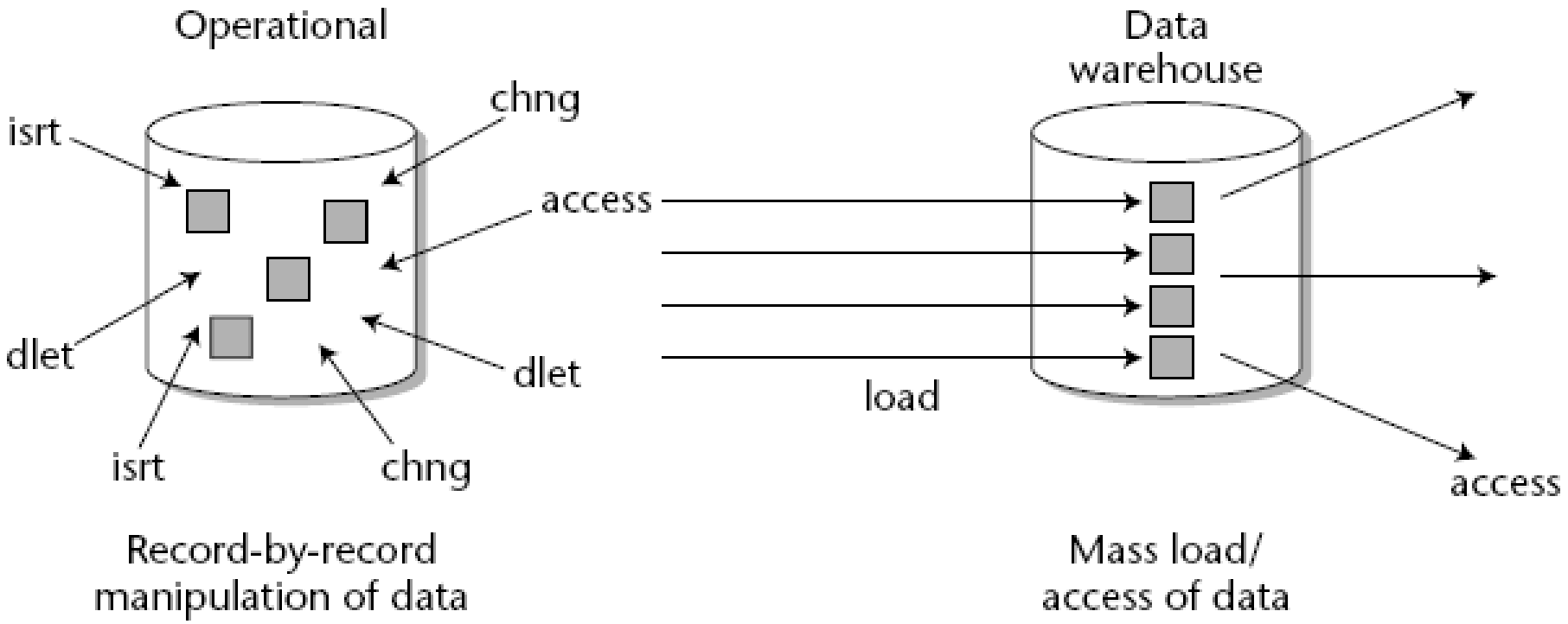


- Time horizon – 5–10 years
- Sophisticated snapshots of data
- Schema contains an element of time

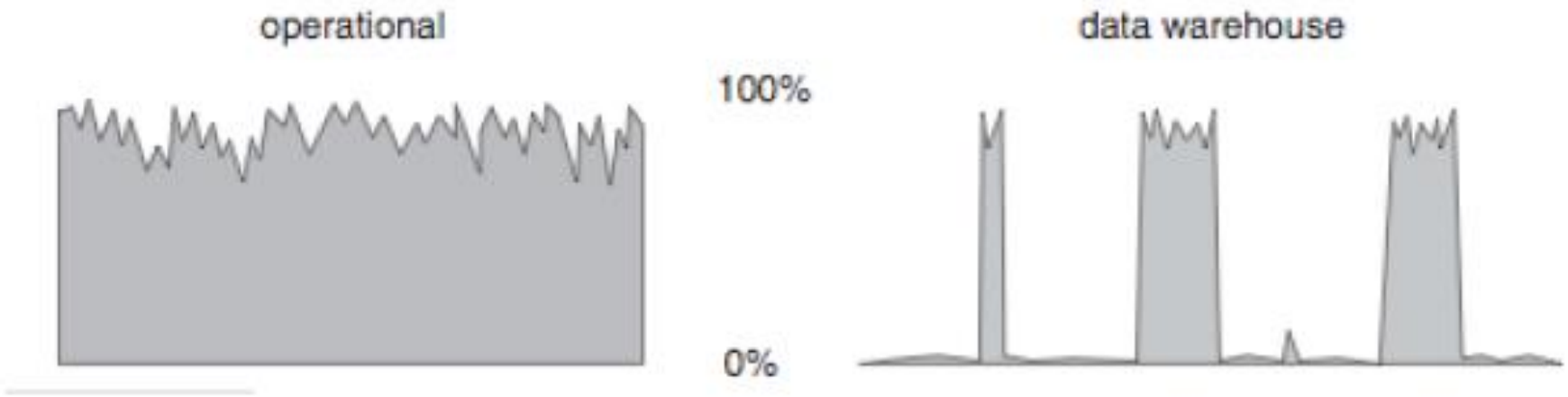
Data Warehouse - Non-Updatable

- **Physically separated store** of data transformed from the operational environment
- **Update of data does not occur** in the DW environment
  - Does not require **recovery** and **concurrency** control mechanisms
  - Requires only two data operations: ***load*** and ***access***
  - **Cannot be updated** by end-users

Data Warehouse - Non-Updatable



Utilization Patterns for OLTP and DSS



Data Warehouse Characteristics

- **Collection of technologies** for the *knowledge worker* make **better** and **faster** decisions
- Targeted for **decision support**
- Contain **consolidated data**, sometimes from several operational databases, over **long periods of time**
- Tend to be **orders of magnitude larger than operational databases**
- Projected to be **hundreds of gigabytes to terabytes** in size

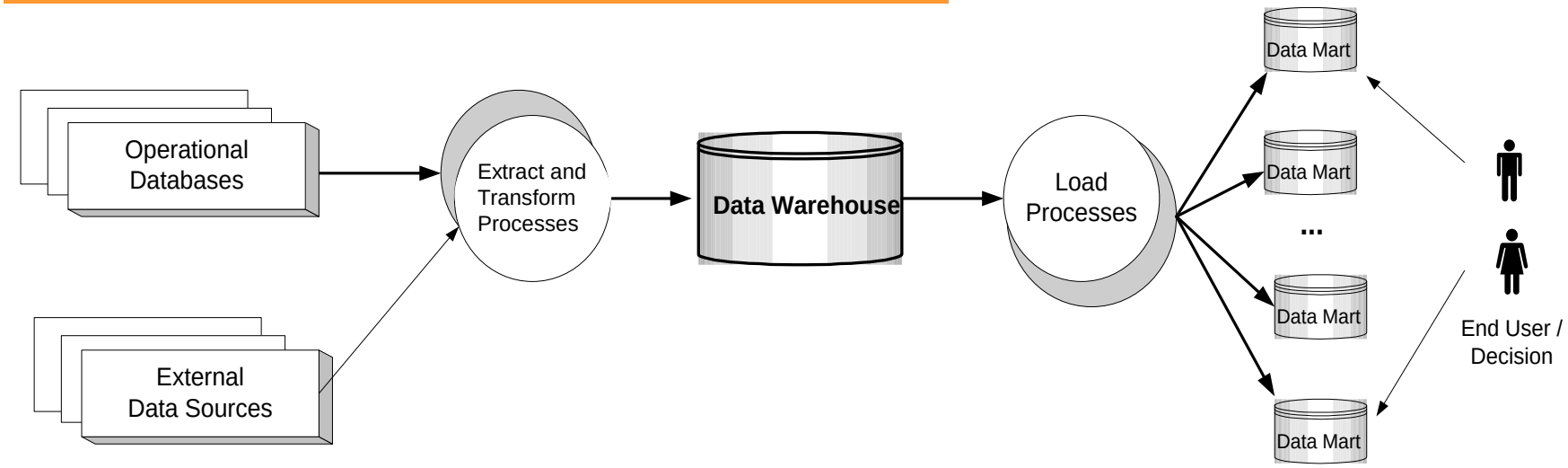
Advantages of a Data Warehouse

- Query performance
- Queries not visible outside the data warehouse
- Local processing at sources unaffected
- Can operate when sources unavailable
- Can query data not stored in the OLTP systems
- Extra information at the DW
  - Summarized (aggregated)
  - Historical information
- Consolidated view of organizational data

Disadvantages of a Data Warehouse

- Initial cost
- Time consuming development
- Maintenance cost
- Difficulty with heterogeneity of hardware and software

A Data Warehouse Architecture



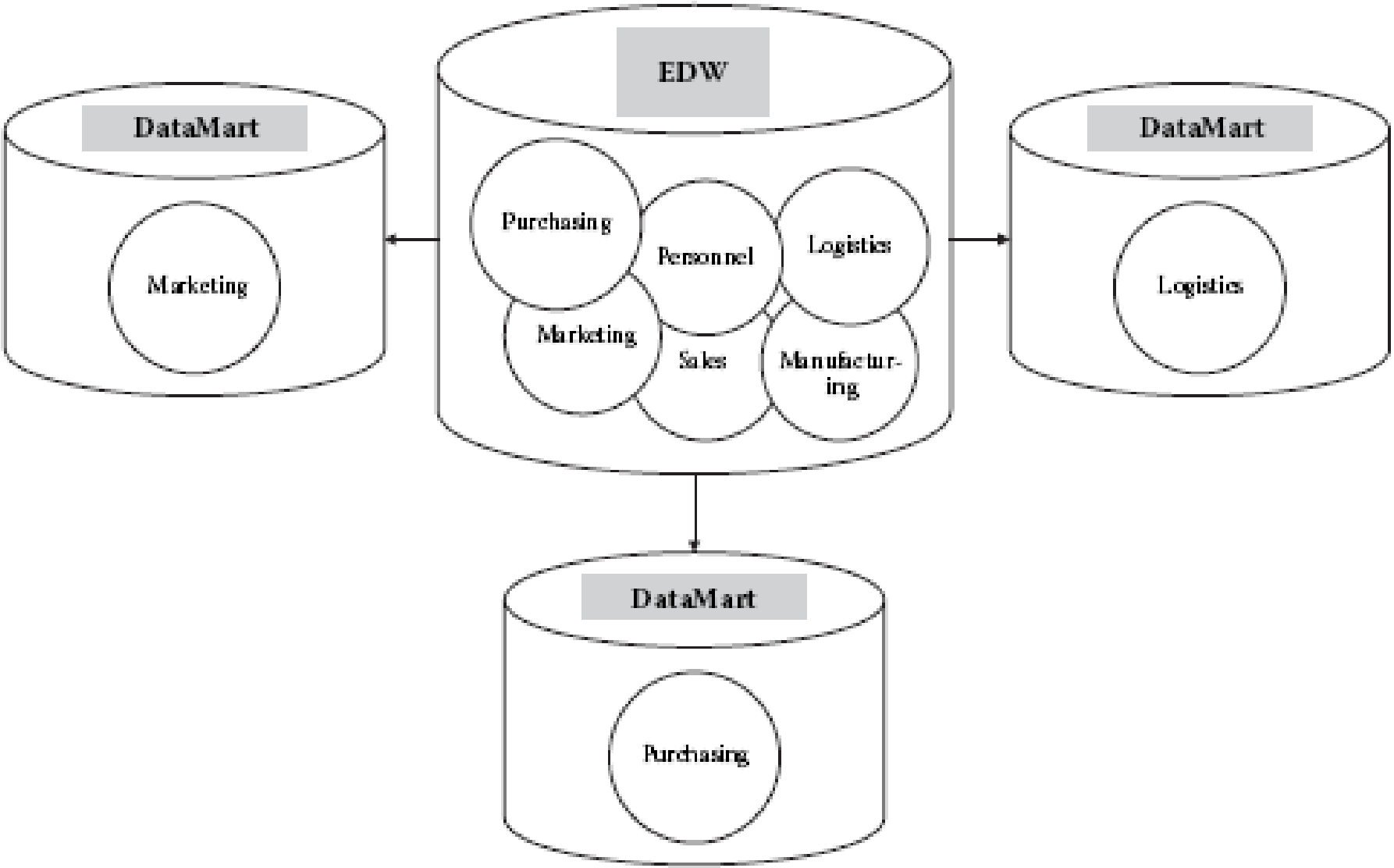
Data warehouse is based on a "supply chain" metaphor:

1. "Data raw material" is obtained from "data suppliers"
2. Data is stored in a central "data warehouse"
3. Data is delivered via "data marts" to "data consumers"

Data Mart

- **Small DW** which contains only a subset of the Enterprise-wide Data Warehouse (EDW)
- DW **limited in scope**
- **Specific to a department or group of users** containing only the data which is relevant
  - Example: marketing data mart has only information about customers, products and sales

Data Mart


EDW
DataMart
Marketing
Purchasing
Personnel
Logistics
Marketing
Sales
Manufactur-
ing
DataMart
Logistics
DataMart
Purchasing

Data Warehouse vs. Data Mart

- Data mart enables fast response to queries
- Data in the data mart is usually more aggregated and less voluminous than in a DW
- Data in a DW is detailed, voluminous (data from various periods of time) and lightly aggregated or not