# Armazéns de Dados

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

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

1

# Introduction to Data Warehouses

2

## Summary

- Data Warehouse Definition

- Data Warehouse Characteristics

- A Data Warehouse Architecture

- Data Mart Definition

## Data Warehouse (DW) Definition

- **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**

5

5

# Data Warehouse - Subject-Oriented

**SUBJECT ORIENTATION**

Operational

Data warehouse

Auto

Policy

Life

Premium

Health

Claim

Applications

Subjects
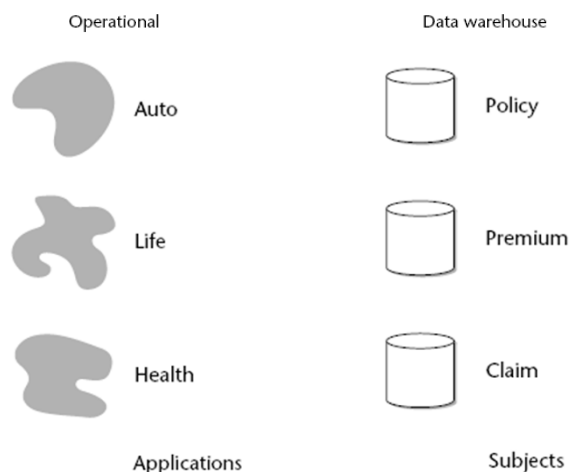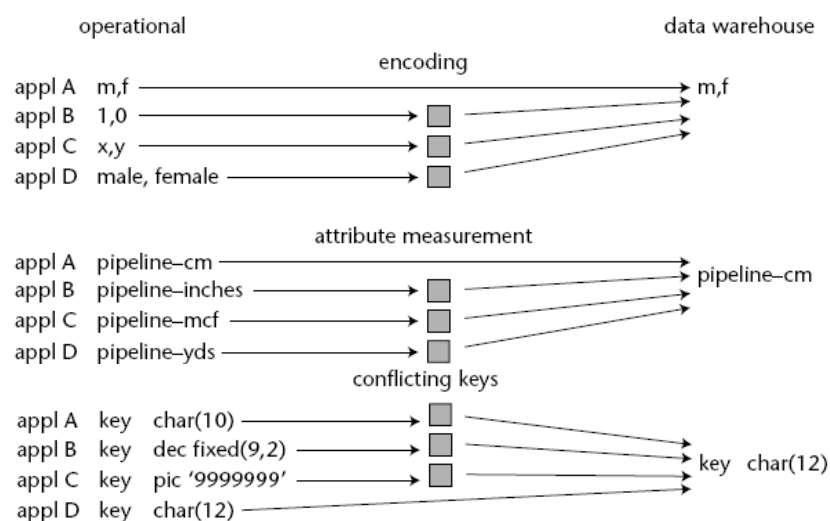
6

6

# Data Warehouse - Integrated

- Constructed by integrating **multiple, heterogeneous data sources**
  - relational databases, text files, binary files, …

- **Data transformation** and **data cleaning** 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

**INTEGRATION**

operational                                         data warehouse

encoding

appl A   m,f ——————————————————————————→ m,f
appl B   1,0 ———————————→ ▢
appl C   x,y ———————————→ ▢
appl D   male, female ——→ ▢

attribute measurement

appl A   pipeline–cm ——————————————————→ pipeline–cm
appl B   pipeline–inches ——→ ▢
appl C   pipeline–mcf ———→ ▢
appl D   pipeline–yds ———→ ▢

conflicting keys

appl A   key   char(10) ——→ ▢
appl B   key   dec fixed(9,2) ——→ ▢       key   char(12)
appl C   key   pic '9999999' ——→ ▢
appl D   key   char(12) ——————

# 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**

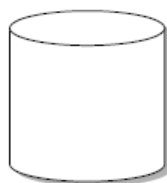- Some form of **time marking** to show the moment in time during which **the record is/was accurate**
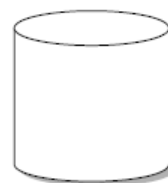
# Data Warehouse - Time Variant

**TIME VARIANCY**

Operational                                Data warehouse

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

- 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*
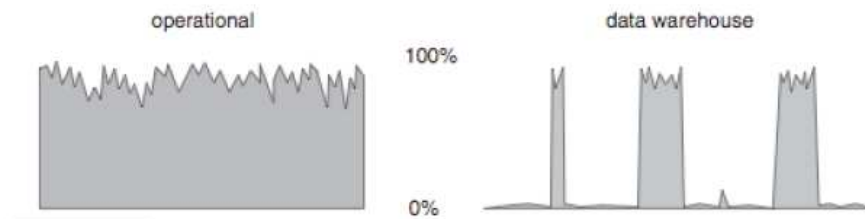    - **Cannot be updated** by end-users

# Data Warehouse - Non-Updatable
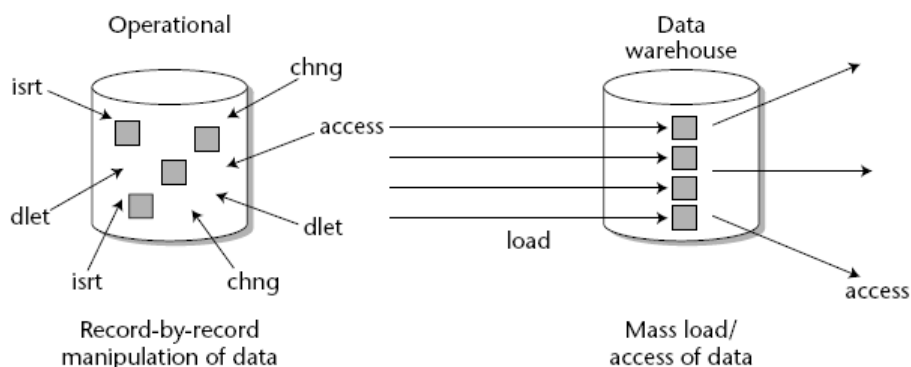
# Utilization Patterns for OLTP and DW



operational                 data warehouse

100%

0%

13

13

# 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

14

14

# 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
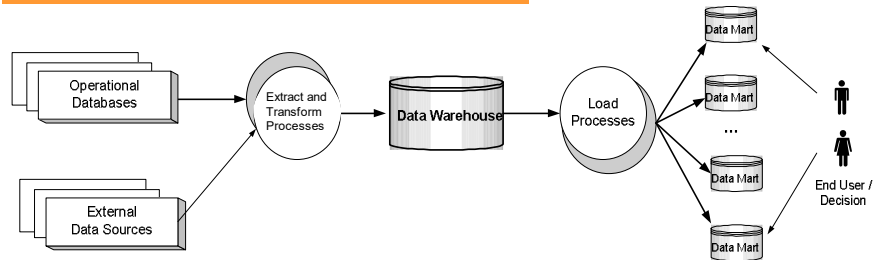
15

15

# Disadvantages of a Data Warehouse

- Time consuming development

- Initial cost

- Maintenance cost

- Data updates

- Heterogeneity of software and hardware

16

16

# 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"
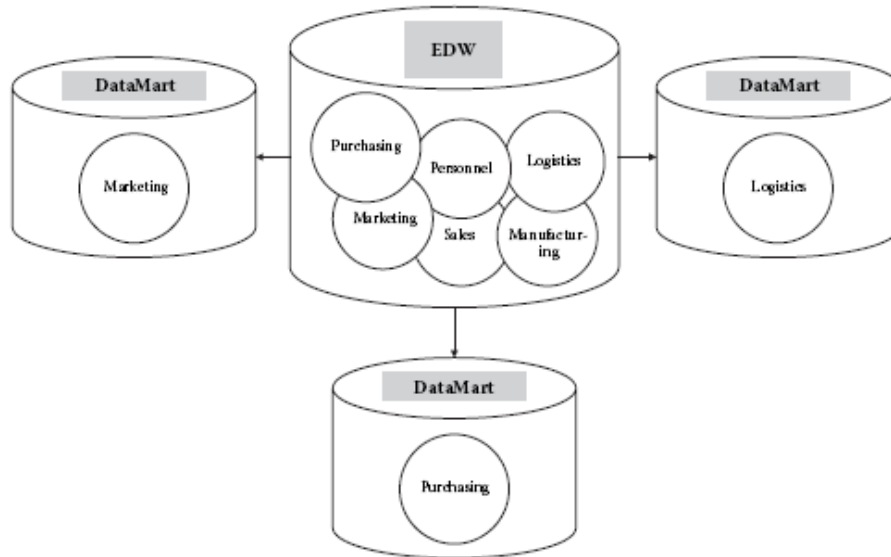
17

17

# Data Mart Definition

- **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

18

18

## Data Mart

19

## Data Warehouse vs. Data Mart

- Data mart enables **fast response to queries**

- Data in the data mart can be **more aggregated** and is **less voluminous** than in a DW

- Data in a DW is **detailed**, **voluminous** (data from various periods of time) and **lightly aggregated or not**

20