Data Warehousing

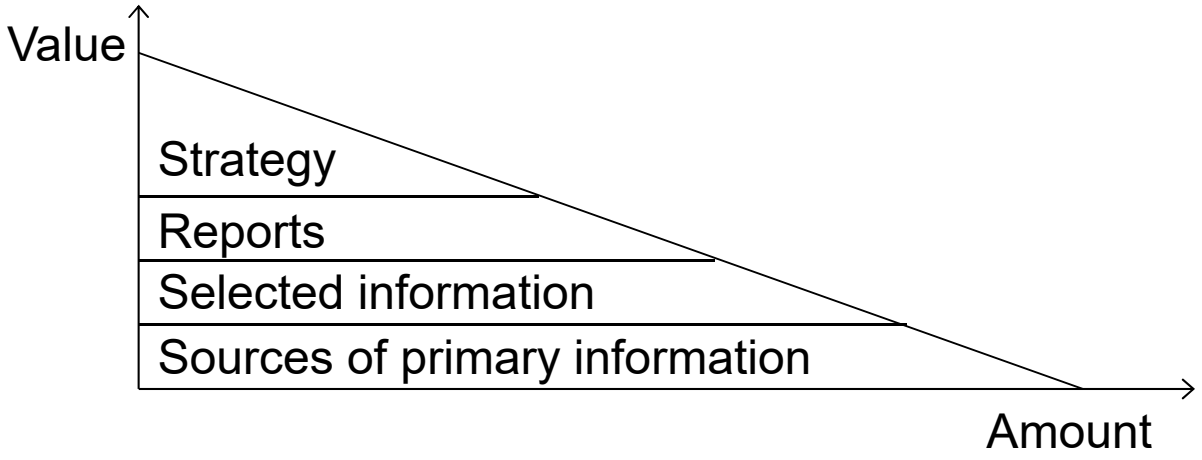
Knowledge Objectives

1. Explain the different requirements, characteristics, kinds of users and tools of a decisional DB, compared with an operational one
2. Give the definition and four characteristics of a "data warehouse"
3. Distinguish "Data Warehouse" from "Data Mart" and "Operational Data Store" in a three layer architecture
4. Explain the importance of metadata in a decisional environment
5. Enumerate different kinds of metadata

OPERATIONAL AND ANALYTICAL DATA

Operational vs analytical

- Operational data
- Analytical data



OLTP vs DW

- Evolution (time management)
- Data volumes
- Aggregation levels
- Actualization
- Response time
- Users
- Functionality

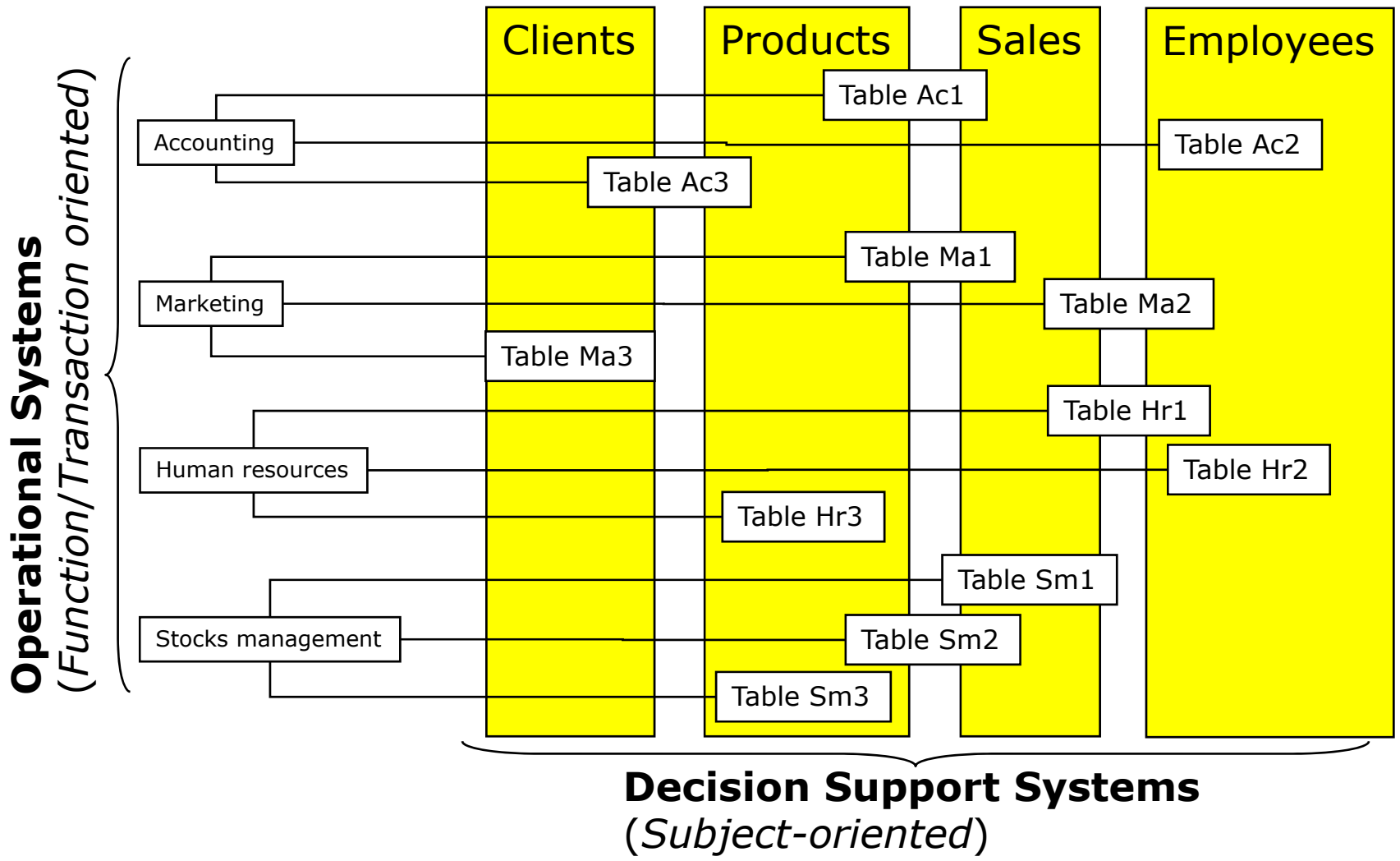
DATA WAREHOUSING

Definition

"A Data Warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision-making process."

W. Inmon, 1992

Subject-oriented



Time variant and nonvolatile

Operational

Name	Salary
Jordi	1200E

Time variant and nonvolatile

Time variant
(Valid Time)

Name	Salary	VT
Jordi	1000E	Jan
Jordi	1100E	Mar
Jordi	1200E	Jul

Operational

Name	Salary
Jordi	1200E

Time variant and nonvolatile

Time variant (Valid Time)

Name	Salary	VT
Jordi	1000E	Jan
Jordi	1100E	Mar
Jordi	1200E	Jul

Operational

Name	Salary
Jordi	1200E

Nonvolatile (Transaction Time)

Name	Salary	TT
Jordi	1000E	Jan
Jordi	900E	Mar
Jordi	1100E	Apr
Jordi	1200E	Sep

Time variant and nonvolatile

Time variant (Valid Time)

Name	Salary	VT
Jordi	1000E	Jan
Jordi	1100E	Mar
Jordi	1200E	Jul

Operational

Name	Salary
Jordi	1200E

Nonvolatile (Transaction Time)

Name	Salary	TT
Jordi	1000E	Jan
Jordi	900E	Mar
Jordi	1100E	Apr
Jordi	1200E	Sep

Data warehouse

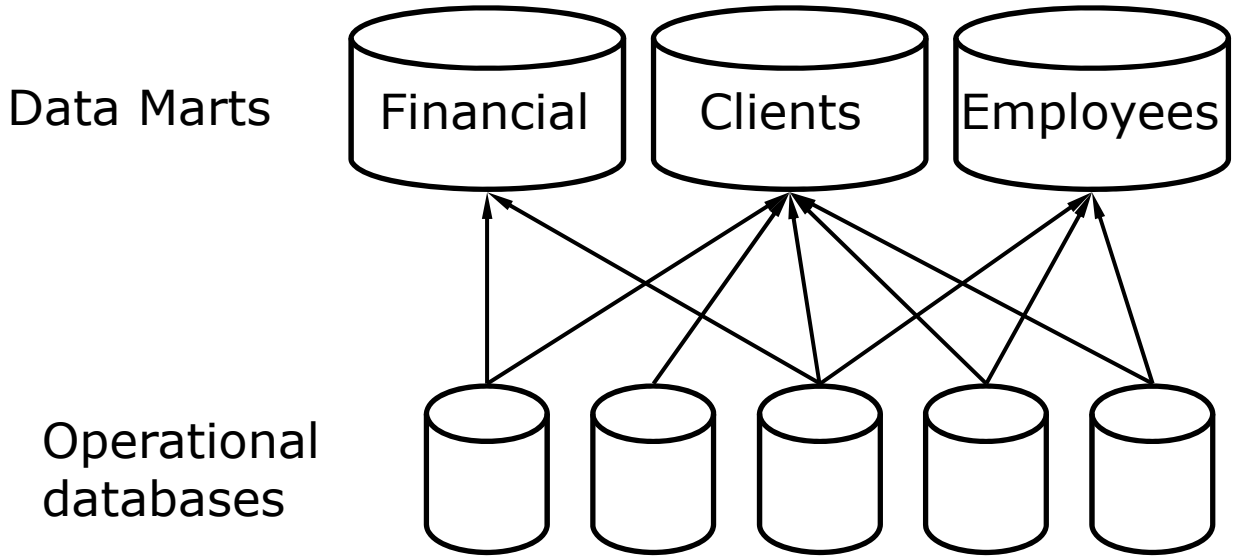
Name	Salary	TT	VT
Jordi	1000E	Jan	Jan
Jordi	1000E	Mar	Jan
	900E		Mar
Jordi	1000E	Apr	Jan
	1100E		Mar
Jordi	1000E	Sep	Jan
	1100E		Mar
	1200E		Jul

Data warehousing

"*Data Warehous**ing** is a process, not a product, for assembling and managing data from various sources for the purpose of gaining a single, detailed view of part or all of a business*."

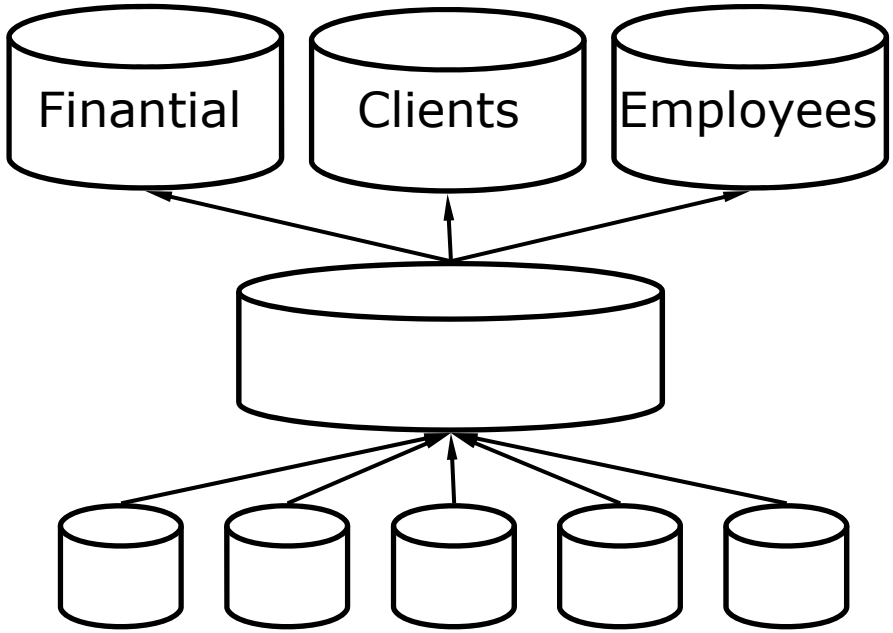
S. Gardner, 1998

Data Marts



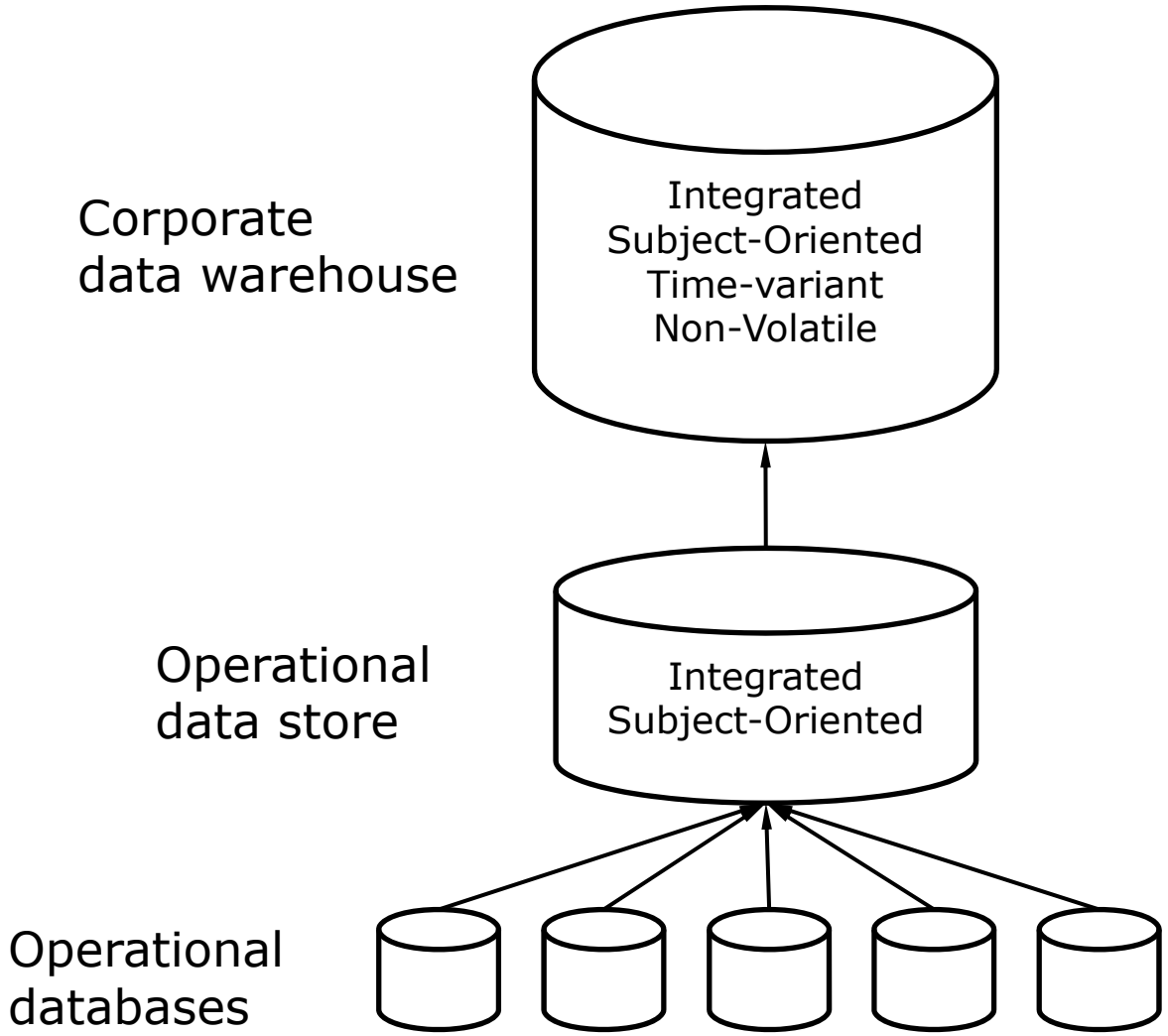
- Analysis oriented
- **Usually** multidimensional
- Only required data:
  - Partial history
  - Only some data sources
  - Not the finest granularity
- Allow cost reduction

Data Warehouse

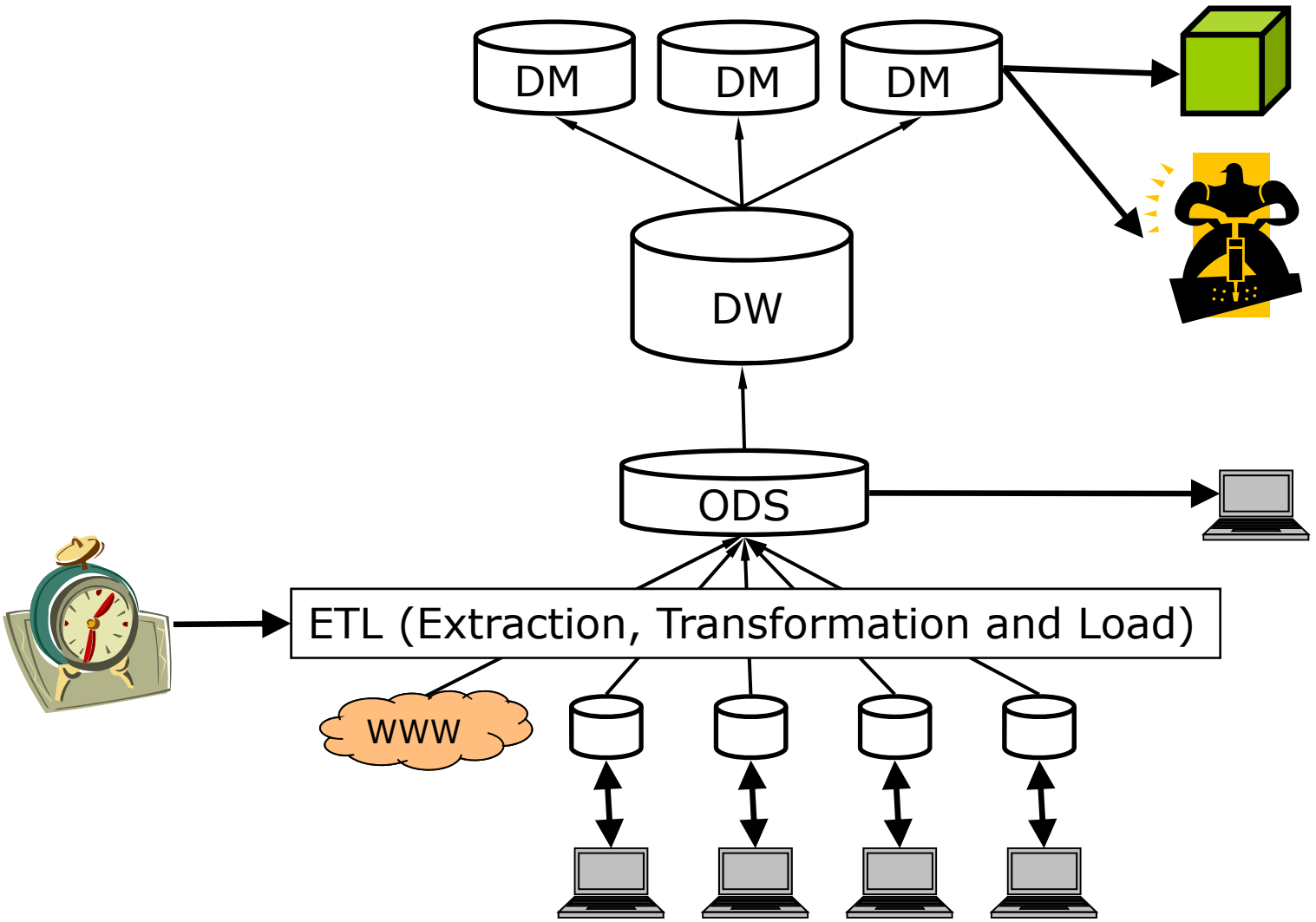


Characteristics	Data warehouse	
	Departamental	Corporate
Subjects	Specific	Generic
Data sources	Few (some)	Many (all)
Size	Gigabytes	Terabytes
Development time	Months	Years
Data model	Multidimensional	Relational (file system)

Operational Data Store



Corporate Information Factory



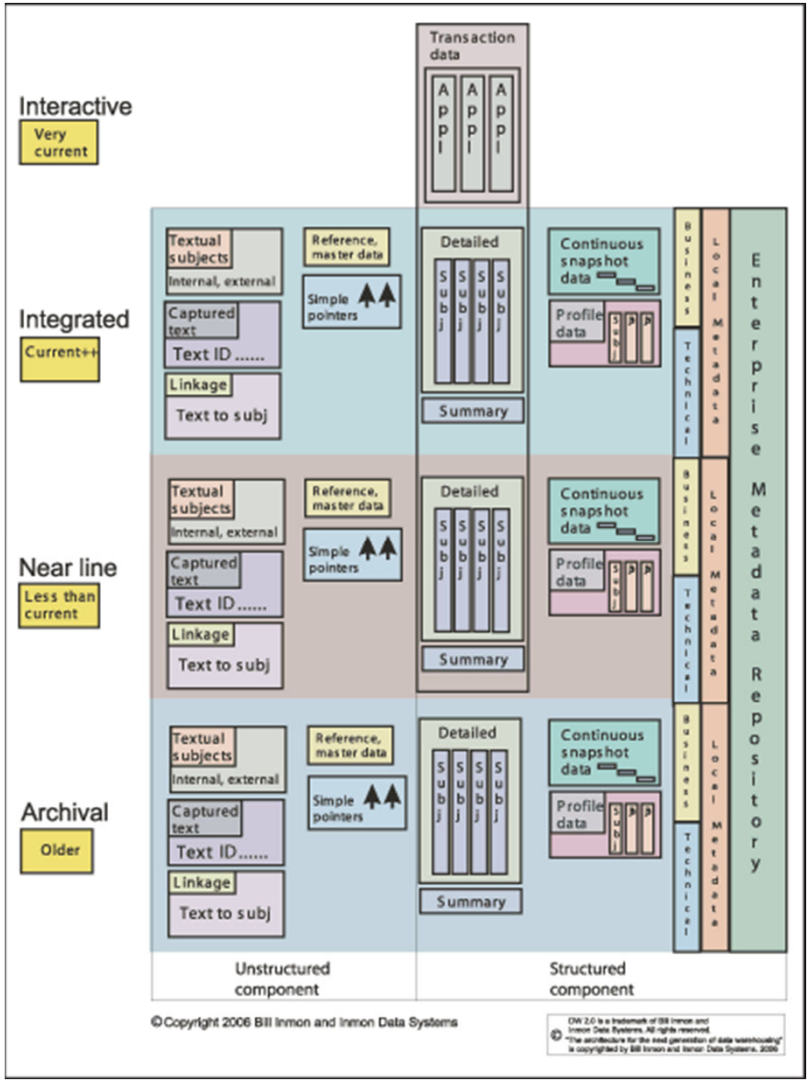
Data sources

Easy and common

- Own operational systems
  - Known
  - Control over changes
- Partners' systems
  - Formats have to be negotiated
  - Undesirable changes are possible
- World Wide Web
  - There is no control
  - Changes are common
- Non computerized sources (OCR or by hand)

Hard and rare

New challenges: Data Warehousing 2.0



W. Inmon

METADATA

Prefix "meta-"

- In science domain means "change". Eg:
  - Metamorphosis
  - Metabolism
- In philosophy domain means "more abstract". Eg:
  - Metarule (Eg: Transitivity)
  - Metaheuristic
  - Metalanguage
  - Metaknowledge (Eg: *Modus ponens*)
  - Metamodel (Eg: Structures, operations and constraints)

Metadata (data about data)

"**Data** is a representation of facts, concepts and instructions, [illegible] in a formalized manner, useful for communication, interpretation and process, by human beings as well as automated means."

ISO 010101

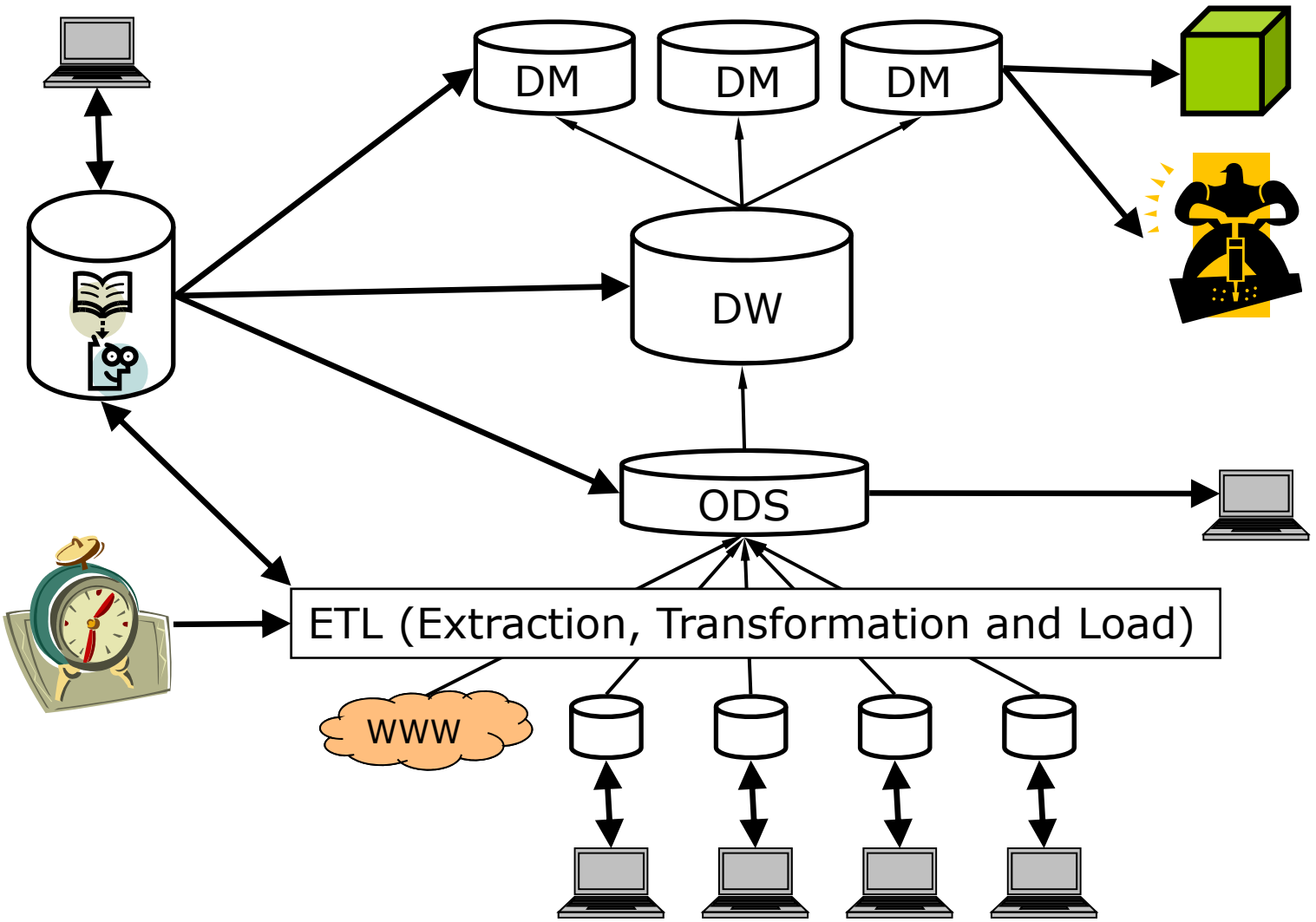
"**Information**, in the processing of data and office machines, is the meaning given to data from the conventional rules used in their representation."

ISO 010102

Derived, summarized and aggregated

- Are obtained from other data
- They may be physically stored
  - Not really necessary
  - Improves query time
  - Space consuming
  - Refresh frequency must be considered
  - Algorithms may also change

Metadata repository



Contents in the metadata repository

- Location map
- Relationships between technical and business metadata
- Extraction mechanisms
- Business rules (integrity constraints)
- Authorization and access information
- Schemas (and their versions)
- Contents and structure of stored data
- Data sources
- Integration information
- Actualization information
- Contents statistics
- Access structures
- General documentation
- Derivation and aggregation algorithms

CLOSING

Differences between OLTP and DW

- Decisional and not operational
  - Subject oriented
  - Huge amount of information
    - Integrate several data sources
    - Contain several versions (of both: data and schemas)
  - Can be composed by several storage systems
- Operational and not decisional
  - Redundancy is not allowed
  - Constant data actualization
  - Transactions are used (concurrency control)

Summary

- Comparison between decisional and operational environments
- Definitions
  - Data Warehouse
  - Data Marts
  - Operational Data Store
  - Corporate Information Factory
- Kinds of data
  - Metadata (data about data)

Bibliography

- W. H. Inmon, C. Imhoff and R. Sousa. *Corporate Information Factory*. John Wiley & Sons, 1998
- W. H. Inmon, D. Strauss and G. Neushloss. *DW2.0*. Morgan Kaufmann, 2008
- M. Golfarelli and S. Rizzi. *Data Warehouse Design*. McGrau-Hill, 2009
- A. Vaisman and E. Zimányi. Data Warehouse Systems. Springer, 2014