MSc Computing- Enterprise Software Systems

Organisation Memory I

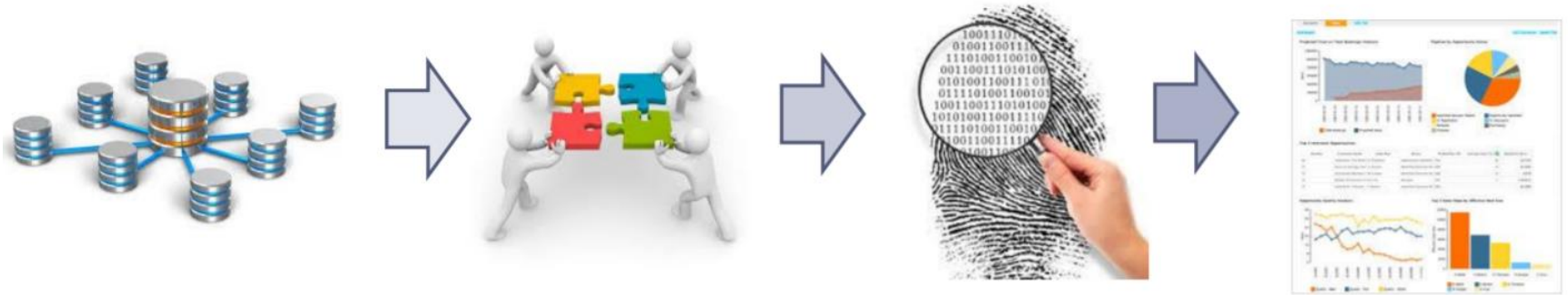
Lecturer: Ruth Barry
Department: Computing and Mathematics
Email: rbarry@wit.ie
Website: www.wit.ie

Waterford Institute *of* Technology
INSTITIÚID TEICNEOLAÍOCHTA PHORT LÁIRGE

Learning Objectives:

- Explain organisation memory
- Understand the basic definitions and characteristics of data warehousing
- Describe the processes used in developing and managing data warehouses
- Understand data warehousing architectures
- Explain data warehousing operations
- Explain the role of data warehouses in decision support
- Explain data integration and the extraction, transformation, and load (ETL) processes

BI Capabilities



Organisational Memory

Information Integration

Information Insights

Information Presentation

Organisational Memory

- At any point in time, organisations have large amounts of accumulated data and information which comprise an important component of org memory.
- Also termed as corporate memory, or institutional memory
- There are large amounts of data and information existing in transactional systems, Enterprise Resource Planning (ERP) and databases and Data Warehouses.
- Key to current data skills required for business intelligence is a knowledge of data warehousing

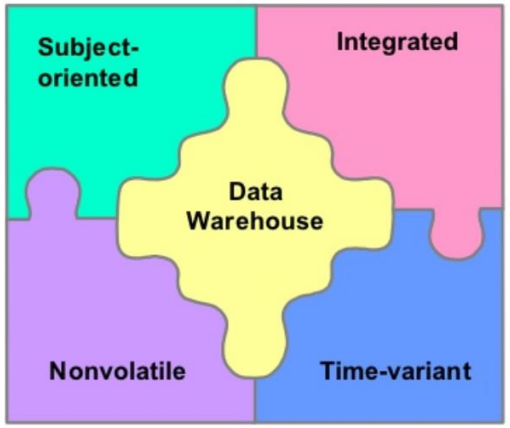
What is a Data Warehouse?

- A physical repository where relational data are specially organized to provide enterprise-wide, cleansed data in a standardized format.
- Holds a copy of transactional data, structured for querying and reporting.

Characteristics of a Data Warehouse

"The data warehouse is a collection of integrated, subject-oriented databases designed to support DSS functions, where each unit of data is non-volatile and relevant to some moment in time"

- Subject-oriented- data is organized around major subjects of the enterprise, such as sales, rather than individual transactions and is oriented to decision making
- Integrated – the same piece of information collected from various systems is referred to in only one way. Example: Gender: M, F; Male, Female; Sex: 0,1
- Nonvolatile: Data is loaded into a data warehouse on a scheduled basis.
- Time-variant: Historical data to support time-series and trend analysis



Some additional characteristics:

Summarized

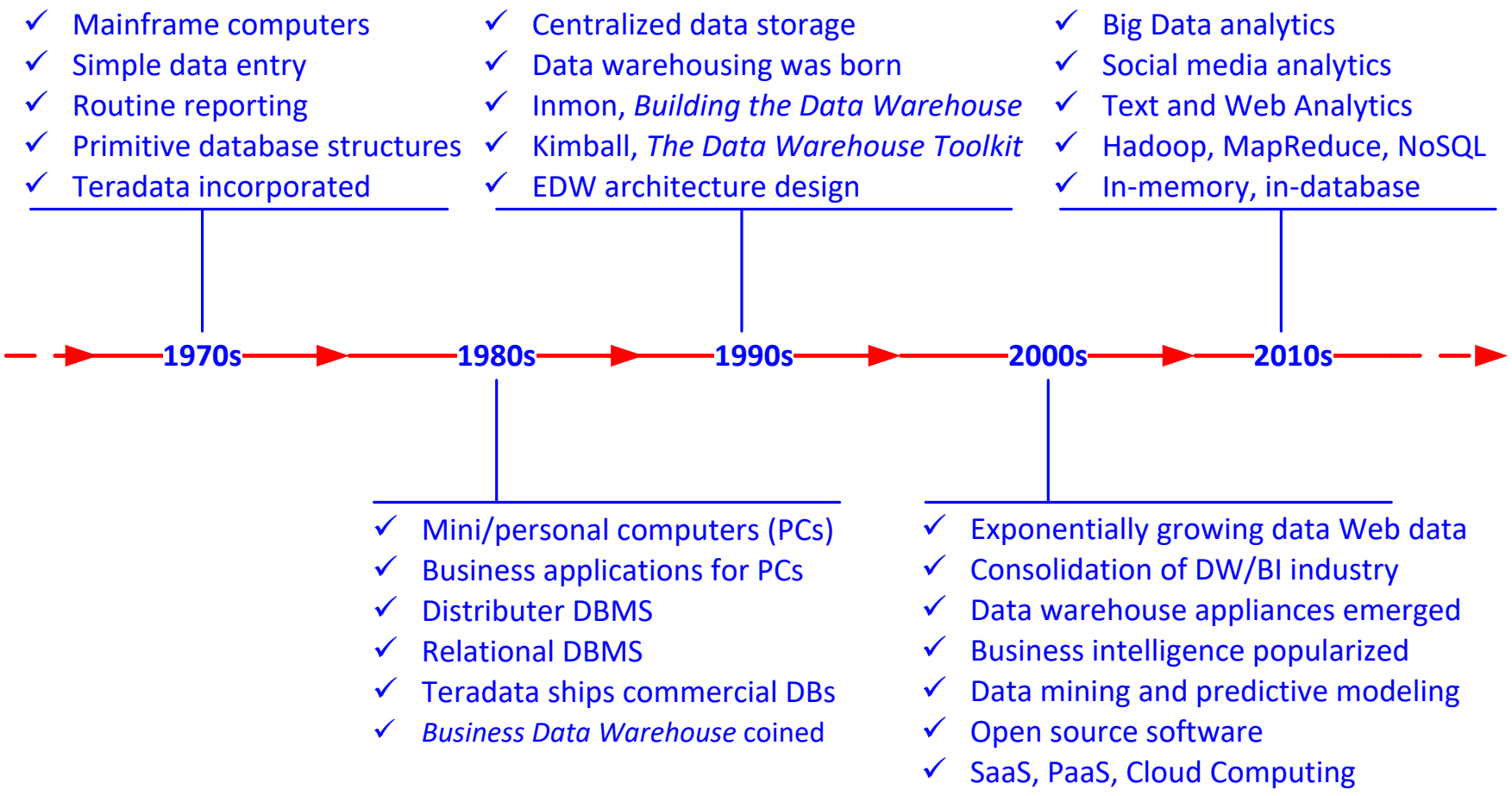
Not normalized

Metadata

Web based, relational/multi-dimensional

Client/server, real-time/right-time/active...

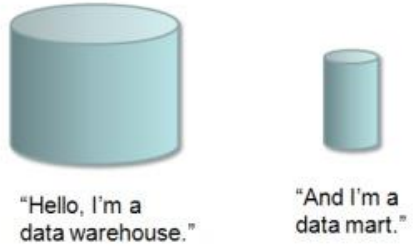
A Historical Perspective to Data Warehousing



DW definitions I

Data Mart: A departmental data warehouse that stores only relevant data:

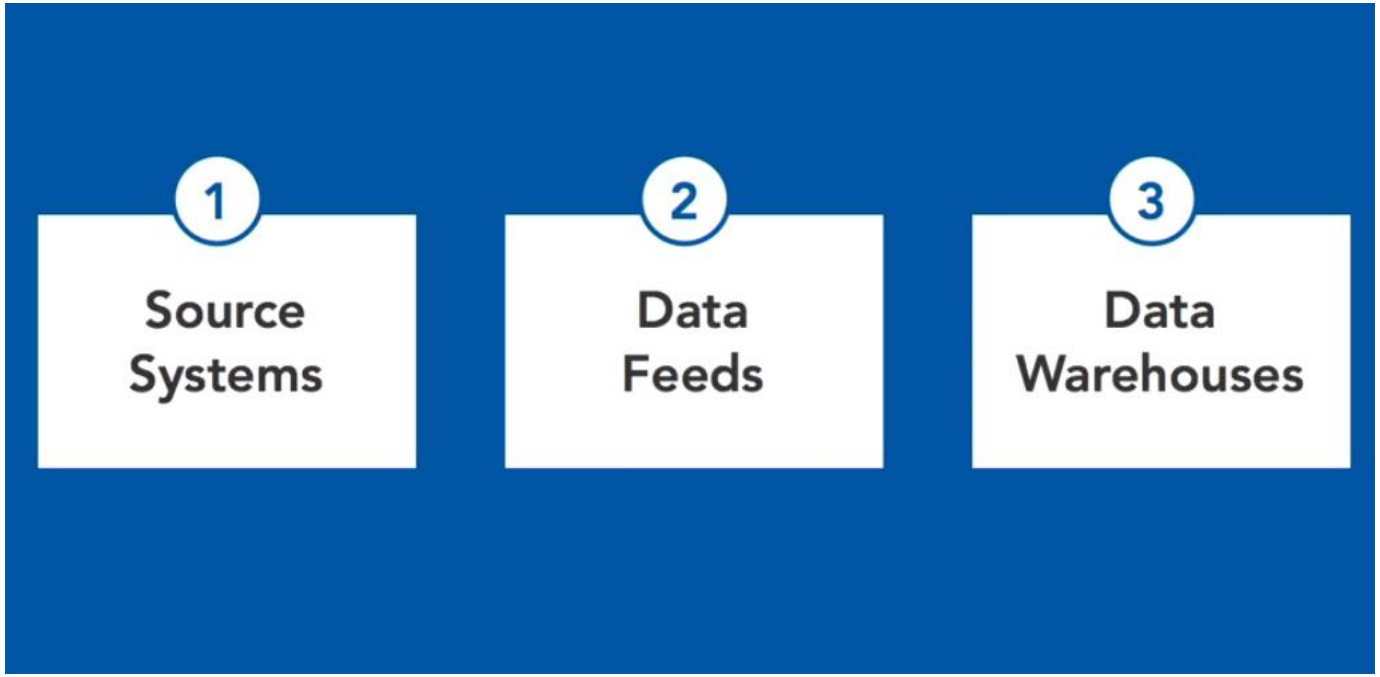
- Dependent data mart: a subset that is created directly from a data warehouse
- Independent data mart: a small data warehouse designed for a strategic business unit or a department



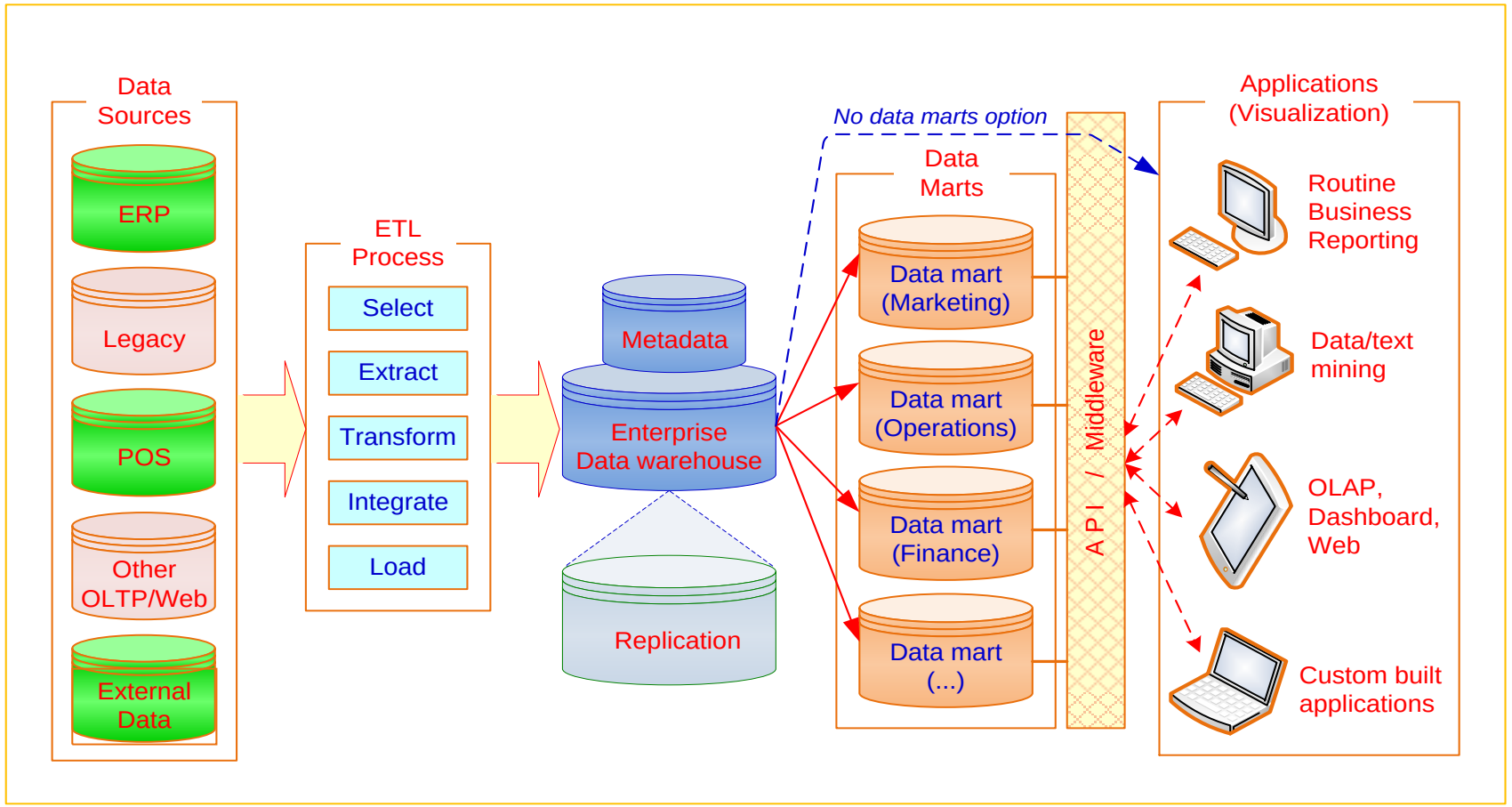
DW Definitions II

- Operational data stores (ODS): a type of database often used as an interim area for a data warehouse
- Enterprise data warehouse (EDW): a data warehouse for the enterprise
- Metadata: data about data. In a data warehouse, metadata describes the contents of a data warehouse and the manner of its acquisition and use.

Three Main Phases of DW Architecture



A Generic DW Framework



DW Process:

Data sources

Data extraction and transformation

Data loading

Comprehensive database – EDW

Metadata

Middleware tools

Data Warehouse Development

Data warehouse development approaches:

- **Inmon Model:** EDW approach (top-down)
- **Kimball Model:** Data mart approach (bottom-up)

  - Which model is best?

Comparing EDW and Data Mart (1 of 2)

Effort	DM Approach	EDW Approach
Scope	One subject area	Several subject areas
Development time	Months	Years
Development cost	$10,000 to $100,000+	$1,000,000+
Development difficulty	Low to medium	High
Data prerequisite for sharing	Common (within business area)	Common (across enterprise)
Sources	Only some operational and external systems	Many operational and external systems
Size	Megabytes to several gigabytes	Gigabytes to petabytes
Time horizon	Near-current and historical data	Historical data
Data transformations	Low to medium	High

Comparing EDW and Data Mart (2 of 2)

Effort	DM Approach	EDW Approach
Update frequency	Hourly, daily, weekly	Weekly, monthly
Hardware	Workstations and departmental Servers	Enterprise servers and mainframe computers
Operating system	Windows and Linux	Unix, Z/OS, OS/390
Databases	Workgroup or standard database servers	Enterprise database servers
Number of simultaneous Users	10s	100s to 1,000s
User types	Business area analysts and Managers	Enterprise analysts and senior executives
Business spotlight	Optimizing activities within the business area	Cross-functional optimization and decision making

Development Choices

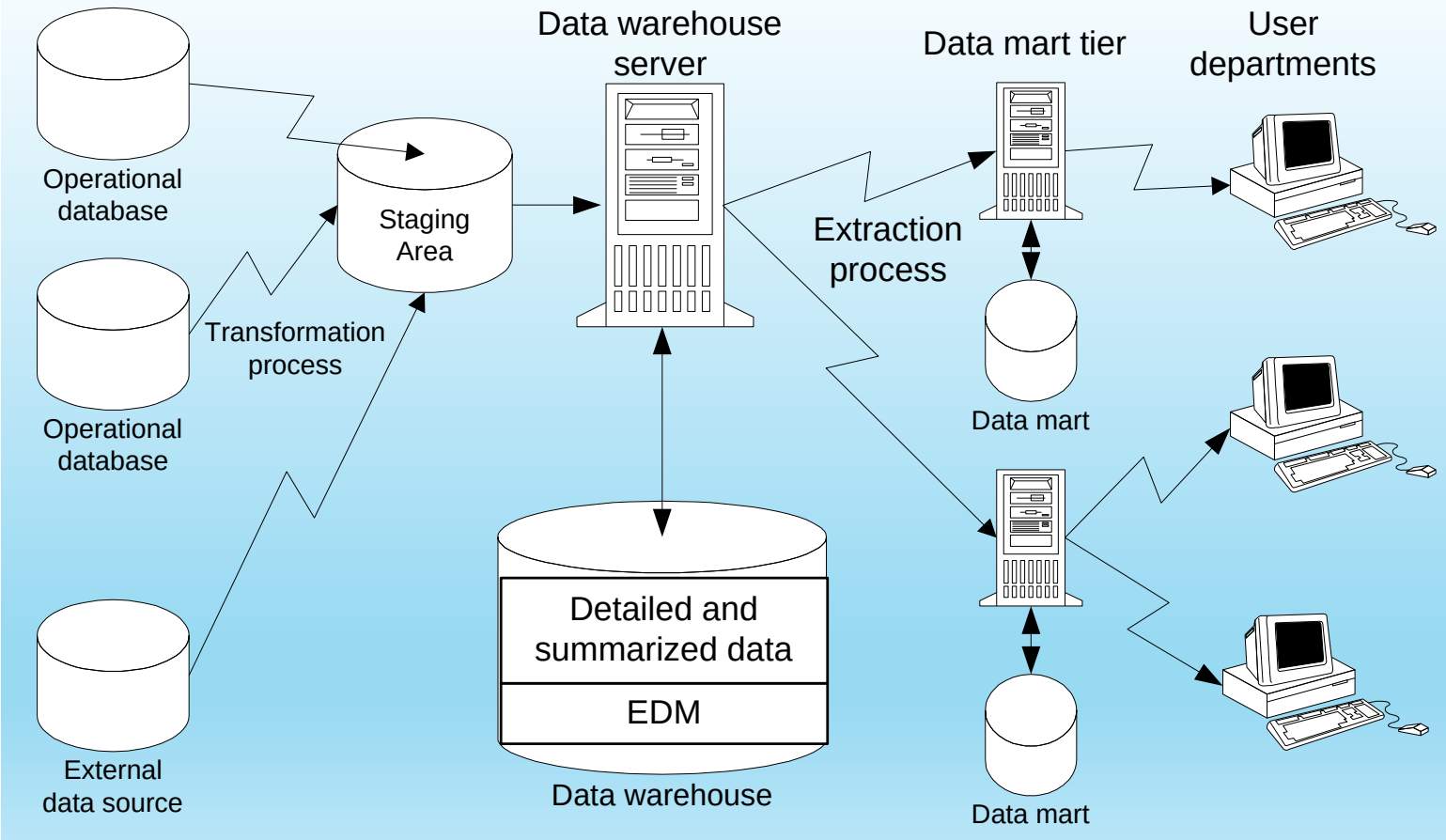
Top Down

- Enterprise data warehouse
- Higher integration levels
- Centralized
- Larger project scope

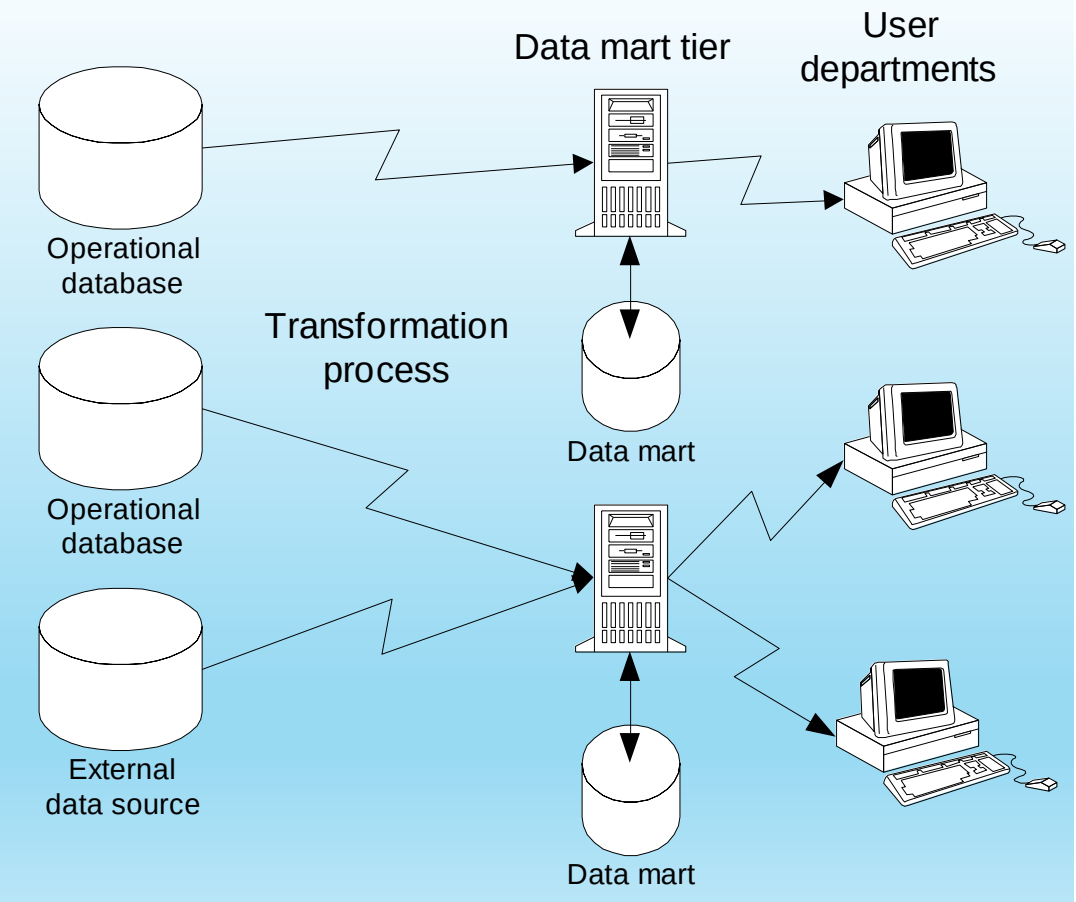
Bottom Up

- Independent data marts
- Lower integration levels
- Decentralized
- Smaller project scope

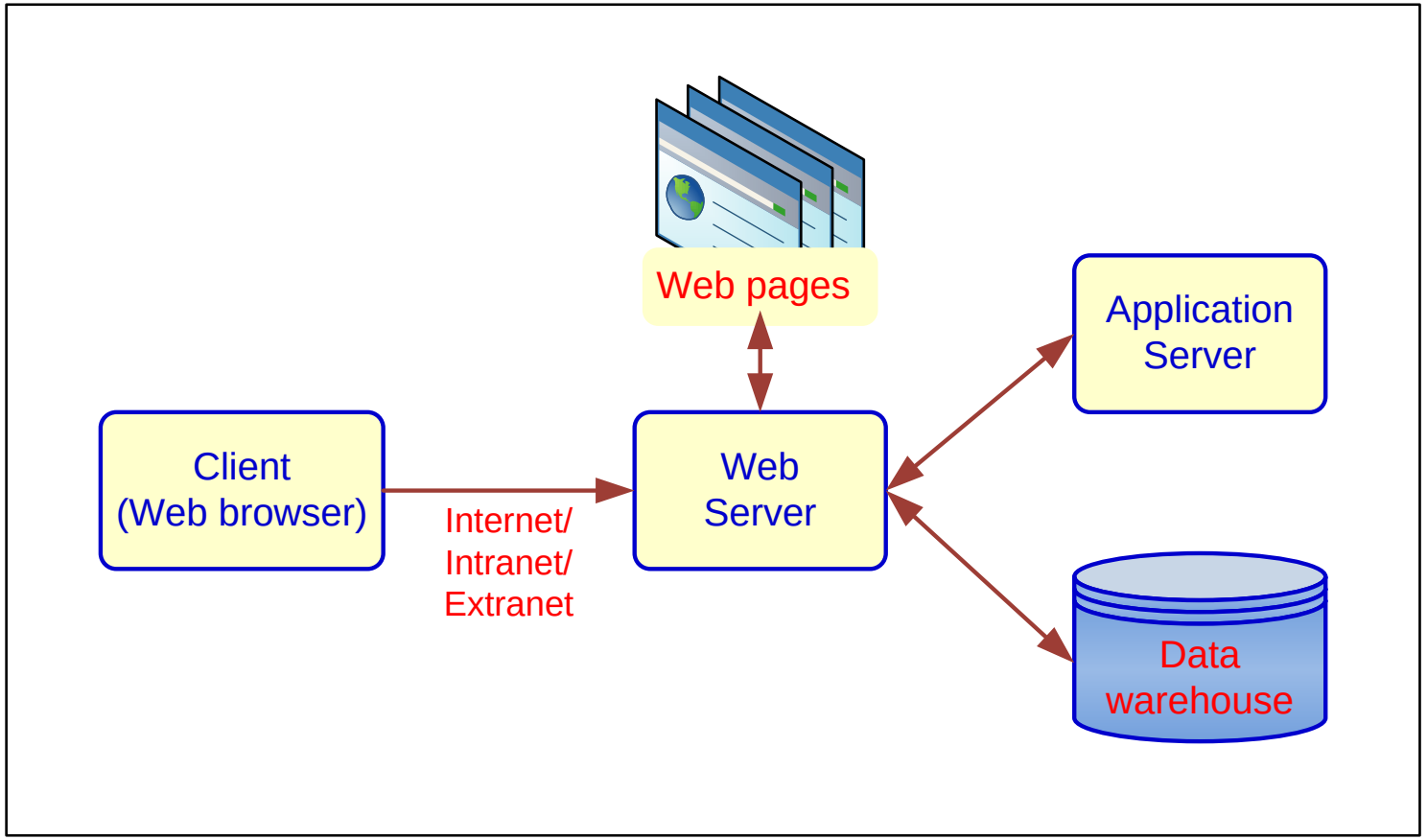
Top-Down


Data warehouse server
Data mart tier
User departments
Operational database
Staging Area
Extraction process
Operational database
Transformation process
Data mart
Detailed and summarized data
EDM
External data source
Data warehouse
Data mart

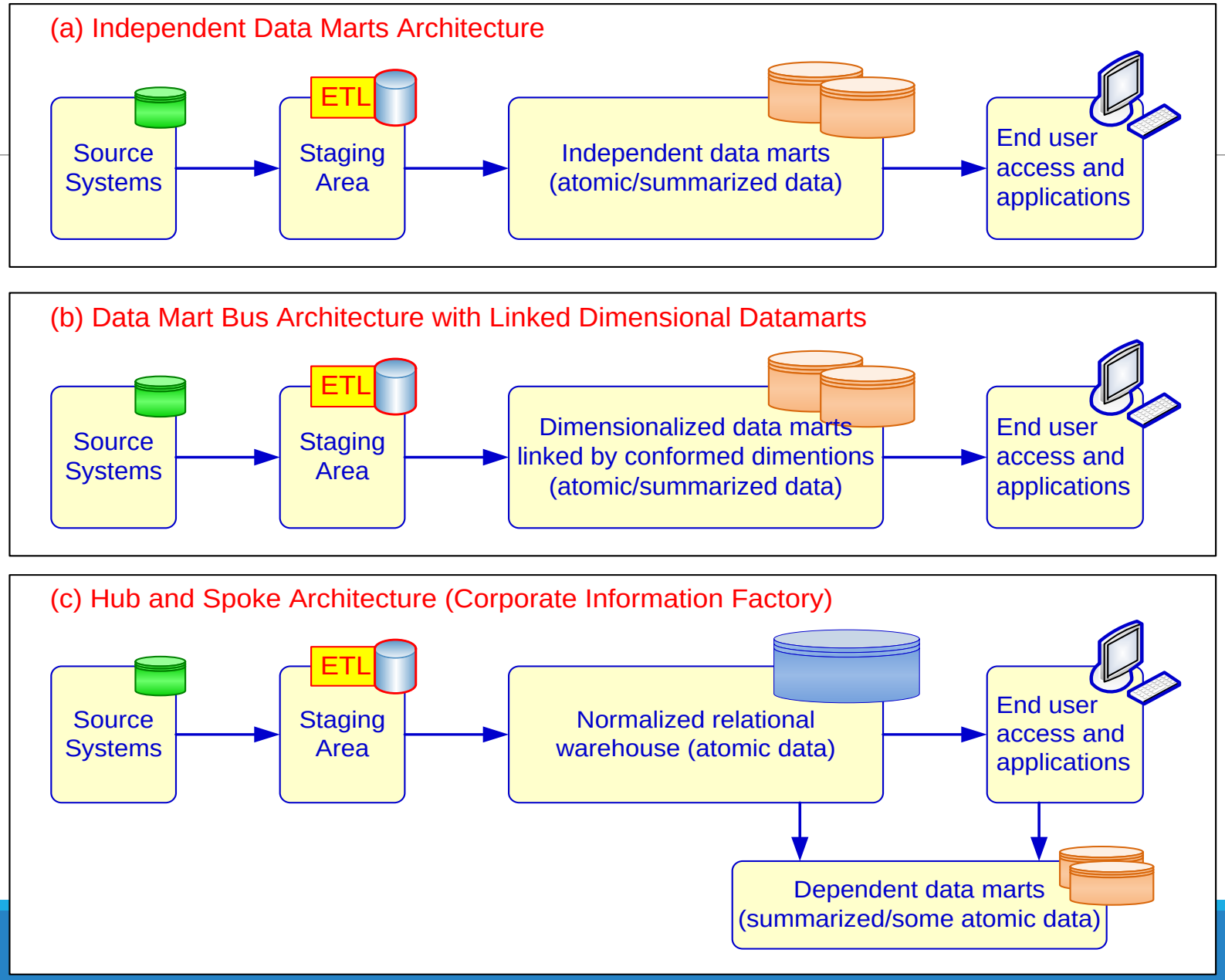
Bottom-up


Data mart tier
User departments
Operational database
Transformation process
Data mart
Operational database
External data source
Data mart

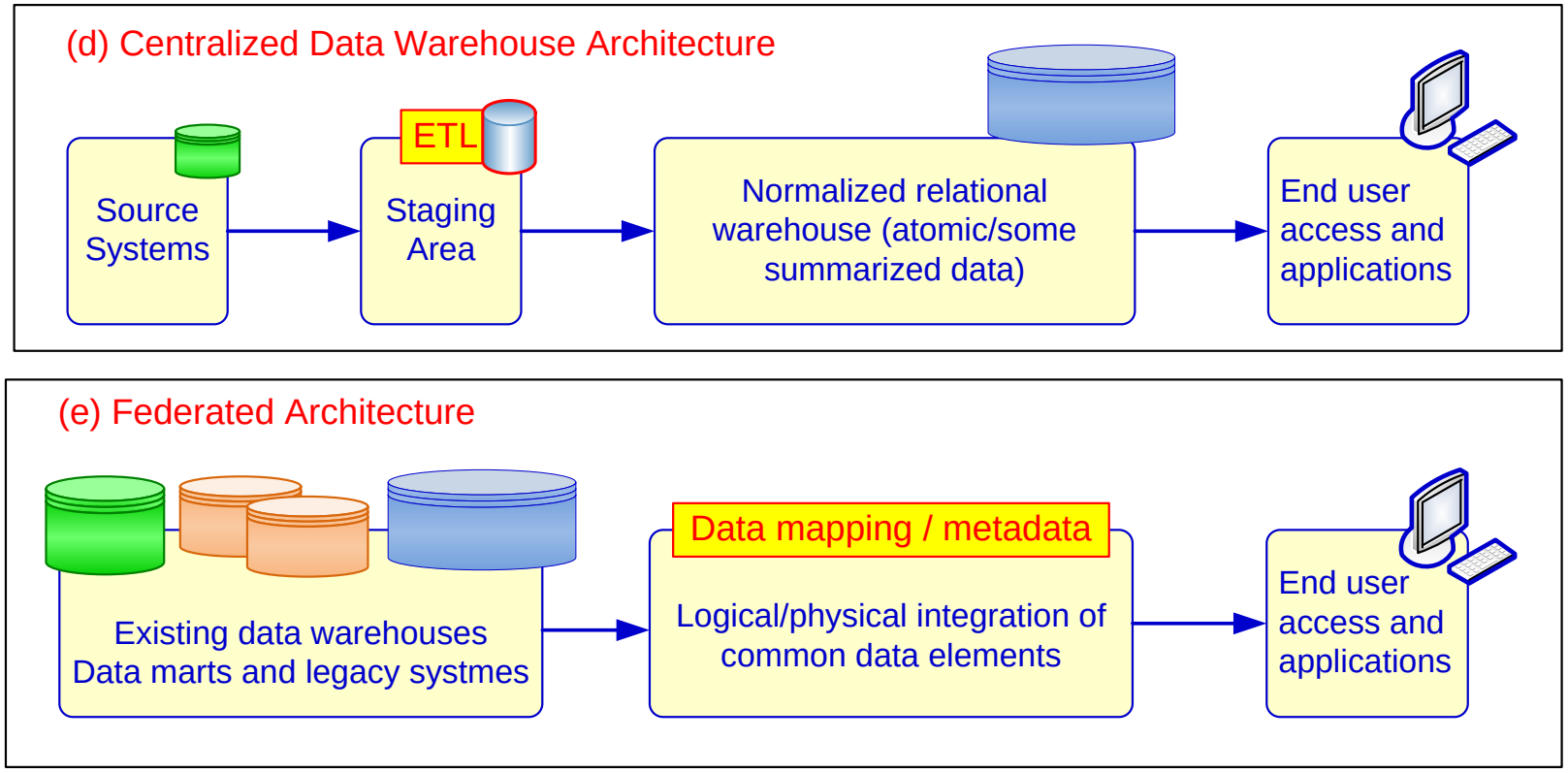
A Web-based DW Architecture



Alternative DW Architectures



DW Architectures



DW Architectures

1. Independent Data Marts
2. Data Mart Bus Architecture
3. Hub-and-Spoke Architecture
4. Centralized Data Warehouse
5. Federated Data Warehouse

Each has pros and cons!

Ten Factors that Potentially Affect the Architecture Selection Decision

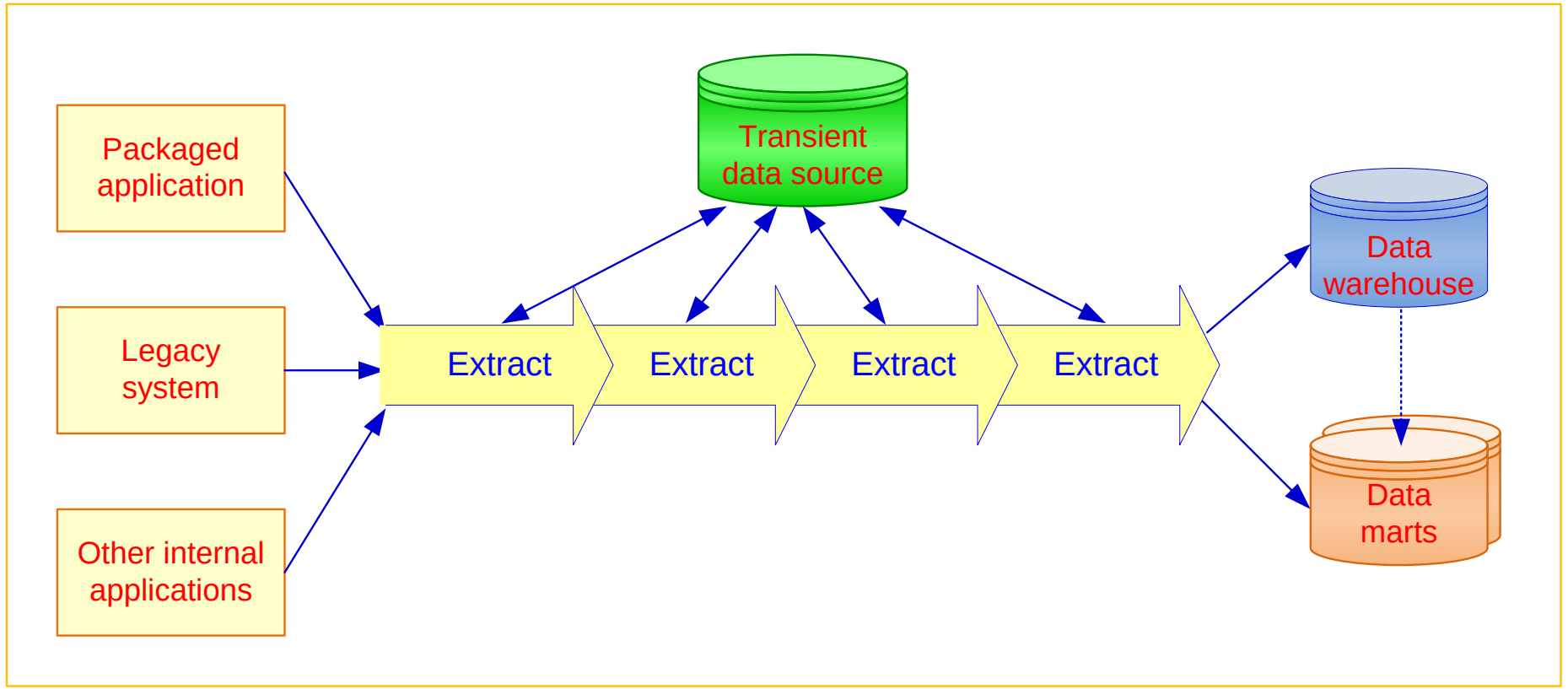
1. Information interdependence between organizational units
2. Upper management's information needs
3. Urgency of need for a data warehouse
4. Nature of end-user tasks
5. Constraints on resources
6. Strategic view of the data warehouse prior to implementation
7. Compatibility with existing systems
8. Perceived ability of the in-house IT staff
9. Technical issues
10. Social/political factors

Data Feeds – Integration of Data in DW

This process is known as Extraction Transformation Loading (ETL)

- Help ensure that only clean data is fed into the data warehouse.
- By tradition, its batch oriented
- Different architecture required if real-time feeds
- ETL is heavily driven by business rules.
- Performance is difficult to manage (as DW expands)

Data Integration and the Extraction, Transformation, and Load Process



ETL (Extract, Transform, Load)

Issues affecting the purchase of an ETL tool

- Data transformation tools are expensive
- Data transformation tools may have a long learning curve

Important criteria in selecting an ETL tool

- Ability to read from and write to an unlimited number of data sources/architectures
- Automatic capturing and delivery of metadata
- A history of conforming to open standards
- An easy-to-use interface for the developer and the functional user

Data Warehouses Challenges

Significant coordination across organisational units

Uncertain data quality in data sources

Difficult to scale data warehouse

- Enhancing a DW is time consuming

They are built slowly

DW and BI have been dominated by insights into what happened in the past.

Data latency - Operational BI requires insights to what's happening currently