CHANDIGARH UNIVERSITY
Discover. Learn. Empower.

APEX INSTITUTE OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Introduction to Data Science (21CST-292)

Faculty: Dr. Jitender Kaushal (E14621)

Associate Professor

Lecture - **Introduction to data warehousing and data mart**

Introduction to Data Science: Course Objectives

COURSE OBJECTIVES

The Course aims to:

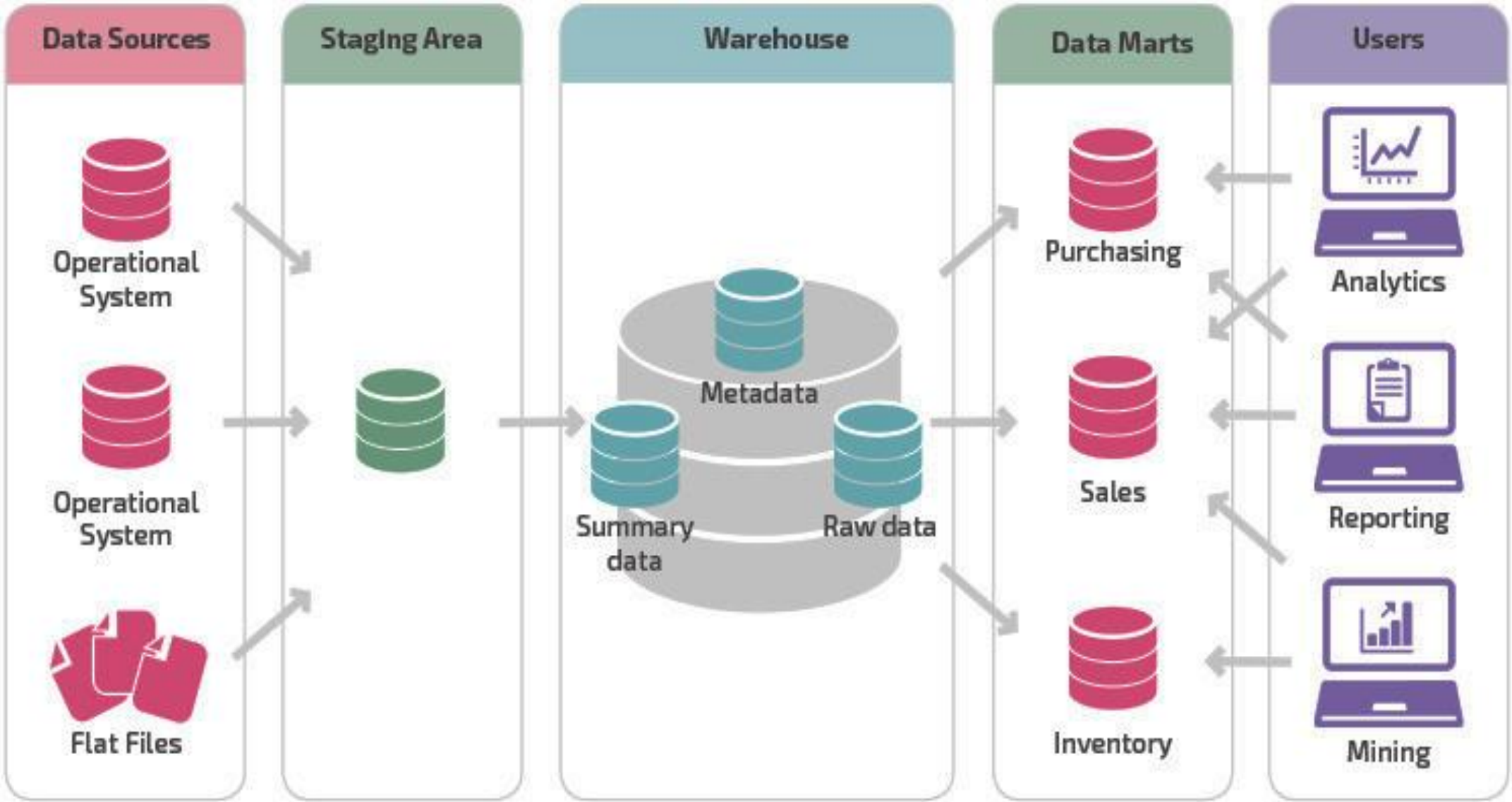
- This course brings together several key big data problems and solutions.
- To recognize the key concepts of Extraction, Transformation and Loading
- To prepare a sample project in Hadoop Environment

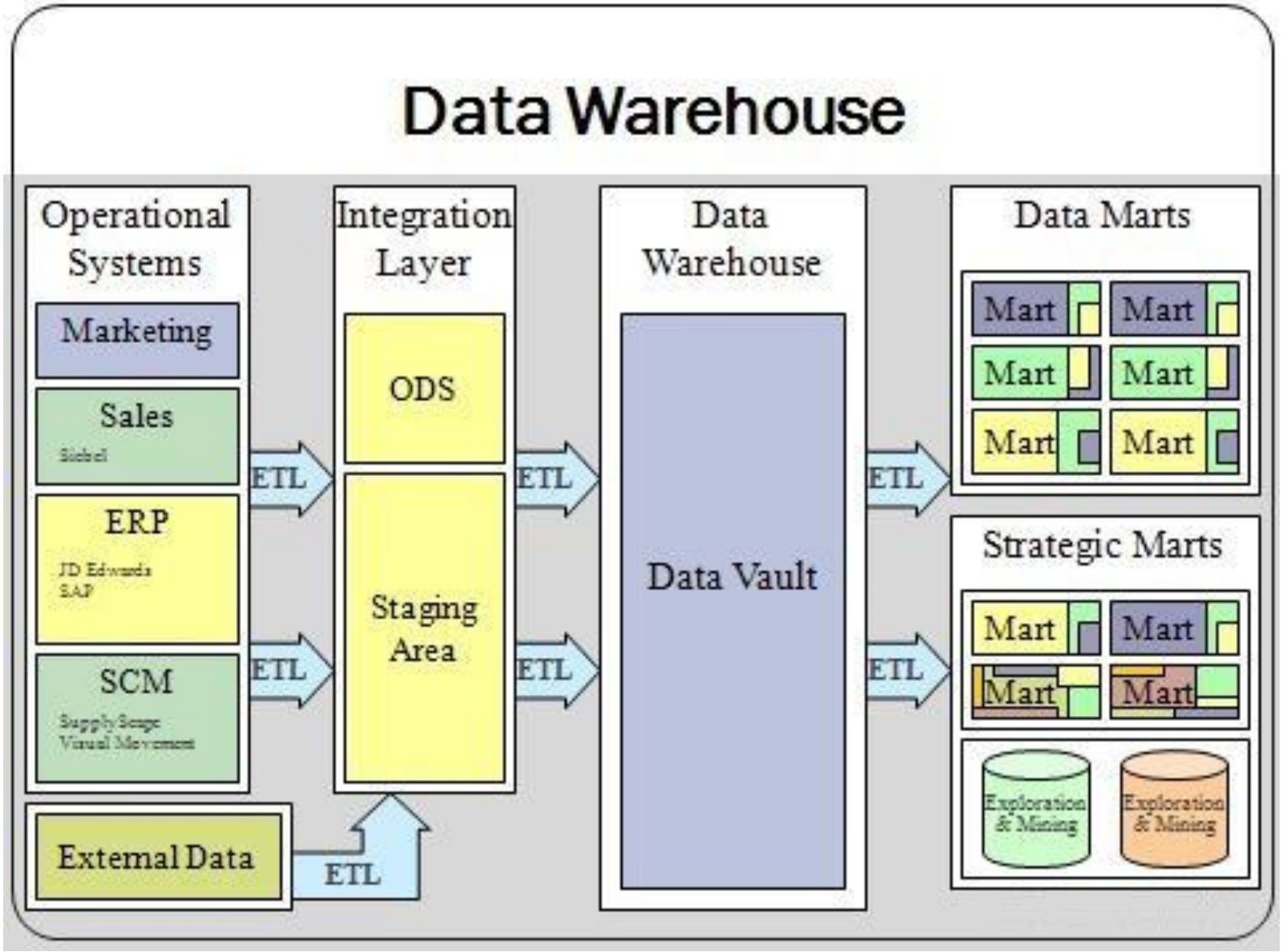
COURSE OUTCOMES

On completion of this course, the students shall be able to:-

CO3	To learn and understand Data Life Cycle, Data Preparation.

Introduction to data warehousing and data mart


Data Sources
Staging Area
Warehouse
Data Marts
Users
Operational System
Operational System
Flat Files
Metadata
Summary data
Raw data
Purchasing
Sales
Inventory
Analytics
Reporting
Mining


Data Warehouse
Operational Systems
Marketing
Sales
ERP
JD Edwards
SAP
SCM
External Data
ETL
ETL
ETL
ETL
ETL
ETL
Integration Layer
ODS
Staging Area
Data Warehouse
Data Vault
Data Marts
Mart
Mart
Mart
Mart
Mart
Mart
Strategic Marts
Mart
Mart
Mart
Mart
Exploration & Mining
Exploration & Mining

Definitions

- **Data Warehouse**
  - A subject-oriented, integrated, time-variant, non-updatable **collection of data** used in support of **management decision-making processes**.

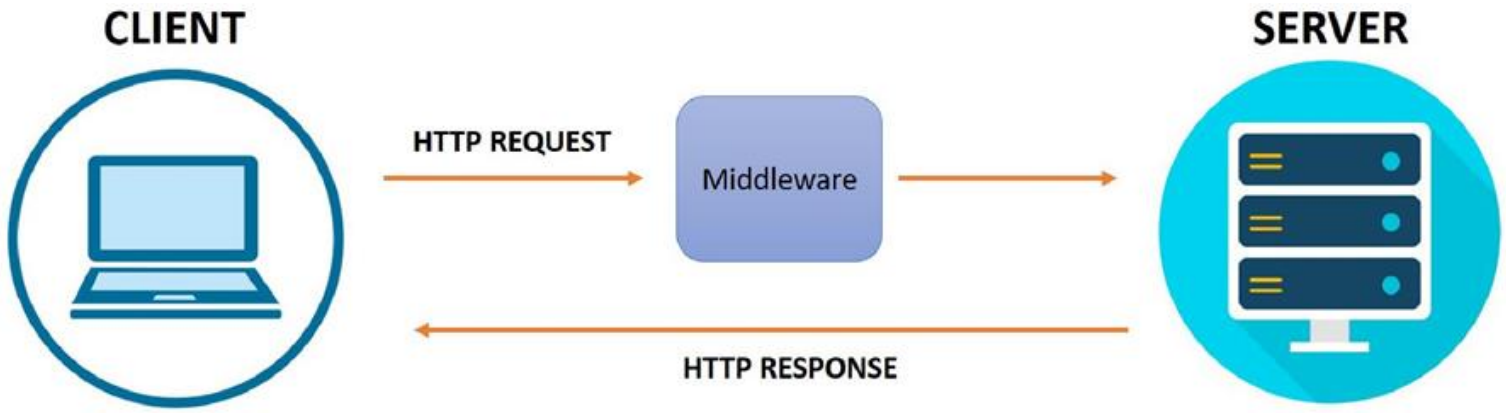
    - ***Subject-oriented:*** e.g. customers, patients, students, products
    - ***Integrated:*** consistent naming conventions, formats, and encoding structures; from multiple data sources
    - ***Time-variant:*** can study trends and changes
    - ***Non-updatable:*** read-only, periodically refreshed

Definitions

- **Data Mart**
  - A data mart is a **subset** of a data warehouse focused on a particular line of business, department, or subject area.
  - Data marts make **specific data** available to a **defined group of users**, which allows those users to **quickly access** critical insights **without wasting time** searching through an entire data warehouse.
  - A data warehouse that is limited in scope.

History Leading to Data Warehousing

- **Improvement in database technologies**, especially relational DBMSs.
- **Advances in computer hardware**, including mass storage and parallel architectures.
- **Emergence of end-user computing** with powerful interfaces and tools.
- **Advances in middleware**, enabling heterogeneous database **connectivity**.
- **Recognition** of the difference between **operational and informational systems**.



Need for Data Warehousing

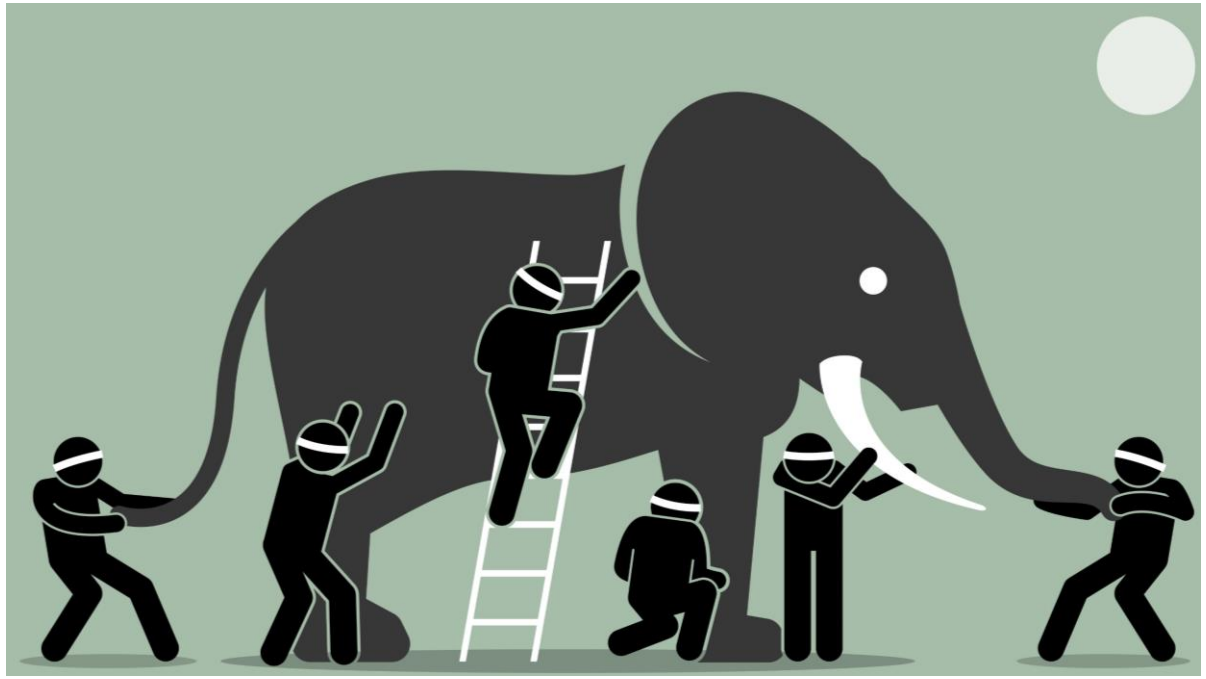
- **Integrated**, company-wide **view of high-quality information (from disparate databases)**

- **Separation of *operational* and *informational*** systems and data **(for improved performance)**

Issues with Company-Wide View

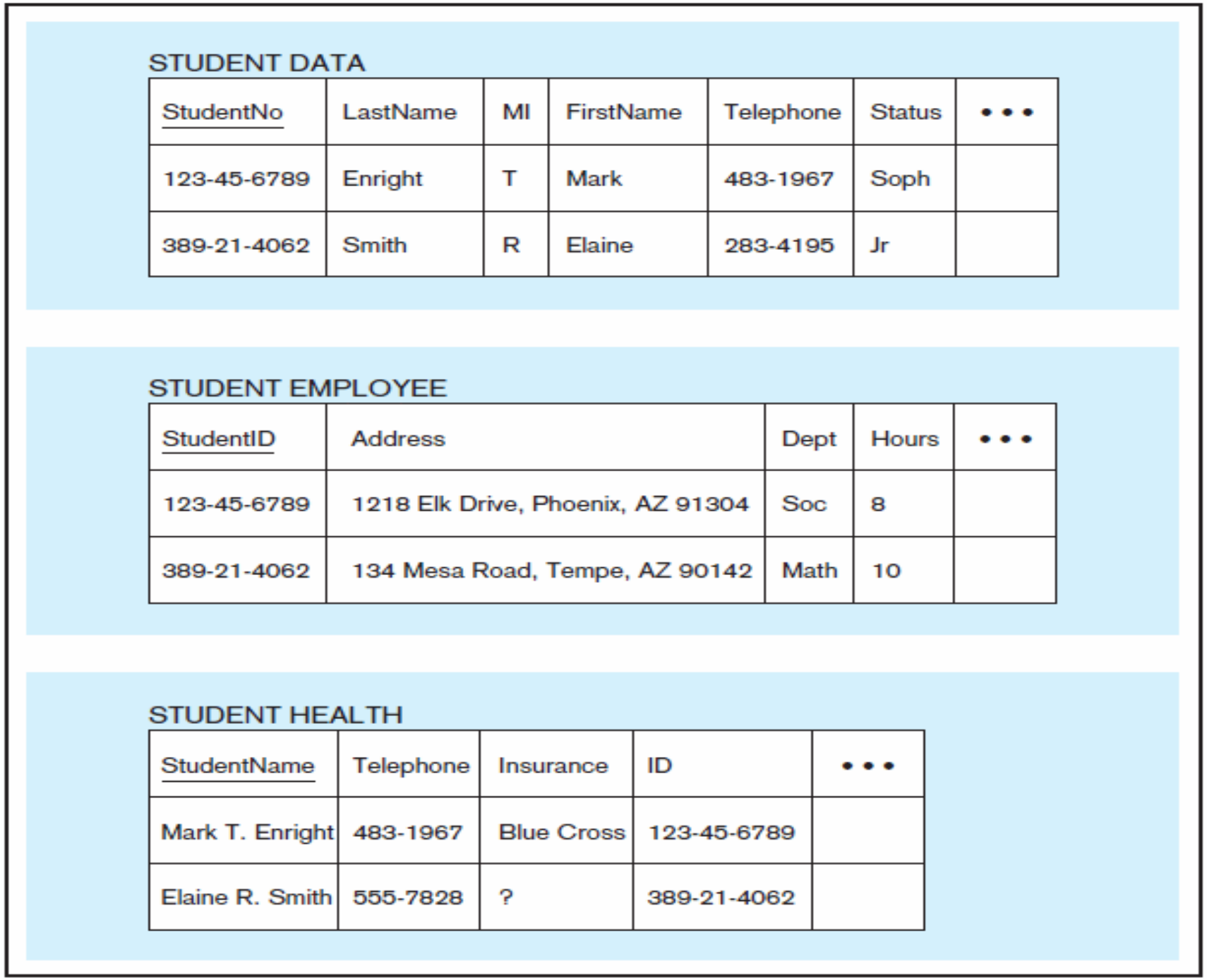
- ✖ Inconsistent key structures
- ✖ Synonyms
- ✖ Free-form vs. structured fields
- ✖ Inconsistent data values
- ✖ Missing data

The moral of the story, of course, is that people often form different conclusions based on a partial view of the same information.



Start Small To Create a Company-wide View of the Customer

Figure 1 Examples of heterogeneous data



STUDENT DATA

StudentNo	LastName	MI	FirstName	Telephone	Status	• • •
123-45-6789	Enright	T	Mark	483-1967	Soph	
389-21-4062	Smith	R	Elaine	283-4195	Jr	

STUDENT EMPLOYEE

StudentID	Address	Dept	Hours	• • •
123-45-6789	1218 Elk Drive, Phoenix, AZ 91304	Soc	8	
389-21-4062	134 Mesa Road, Tempe, AZ 90142	Math	10	

STUDENT HEALTH

StudentName	Telephone	Insurance	ID	• • •
Mark T. Enright	483-1967	Blue Cross	123-45-6789	
Elaine R. Smith	555-7828	?	389-21-4062	

Organizational Trends Motivating Data Warehouses

- No single system of records
- Multiple systems not synchronized
- Organizational need to analyze activities in a balanced way
- Customer relationship management
- Supplier relationship management

Separating Operational and Informational Systems

- **Operational system** – a system that is used to **run a business in real time**, based on current data; also called a system of record.

- **Informational system** – a system designed to support **decision-making** based on historical point-in-time and prediction data for complex queries or data-mining applications

TABLE 9-1 Comparison of Operational and Informational Systems

Characteristic	Operational Systems	Informational Systems
Primary purpose	Run the business on a current basis	Support managerial decision making
Type of data	Current representation of state of the business	Historical point-in-time (snapshots) and predictions
Primary users	Clerks, salespersons, administrators	Managers, business analysts, customers
Scope of usage	Narrow, planned, and simple updates and queries	Broad, ad hoc, complex queries and analysis
Design goal	Performance: throughput, availability	Ease of flexible access and use
Volume	Many constant updates and queries on one or a few table rows	Periodic batch updates and queries requiring many or all rows

Data Warehouse Architectures

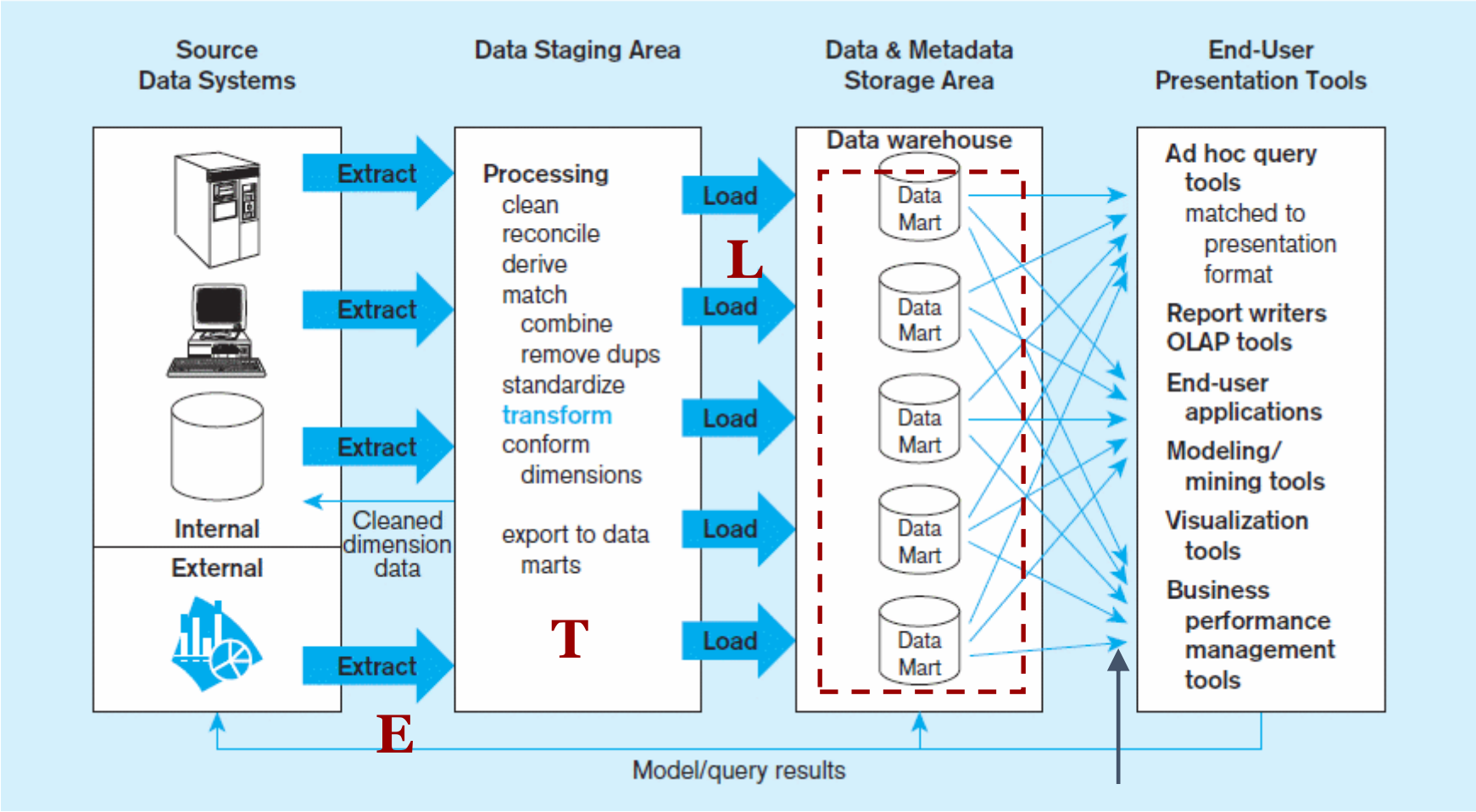
- Independent Data Mart
- Dependent Data Mart and Operational Data Store
- Logical Data Mart and Real-Time Data Warehouse
- Three-Layer architecture

All involve some form of ***extract***, ***transform*** and ***load*** (**ETL**)

Figure-2 Independent data mart data warehousing architecture

Data marts:

Mini-warehouses, limited in scope

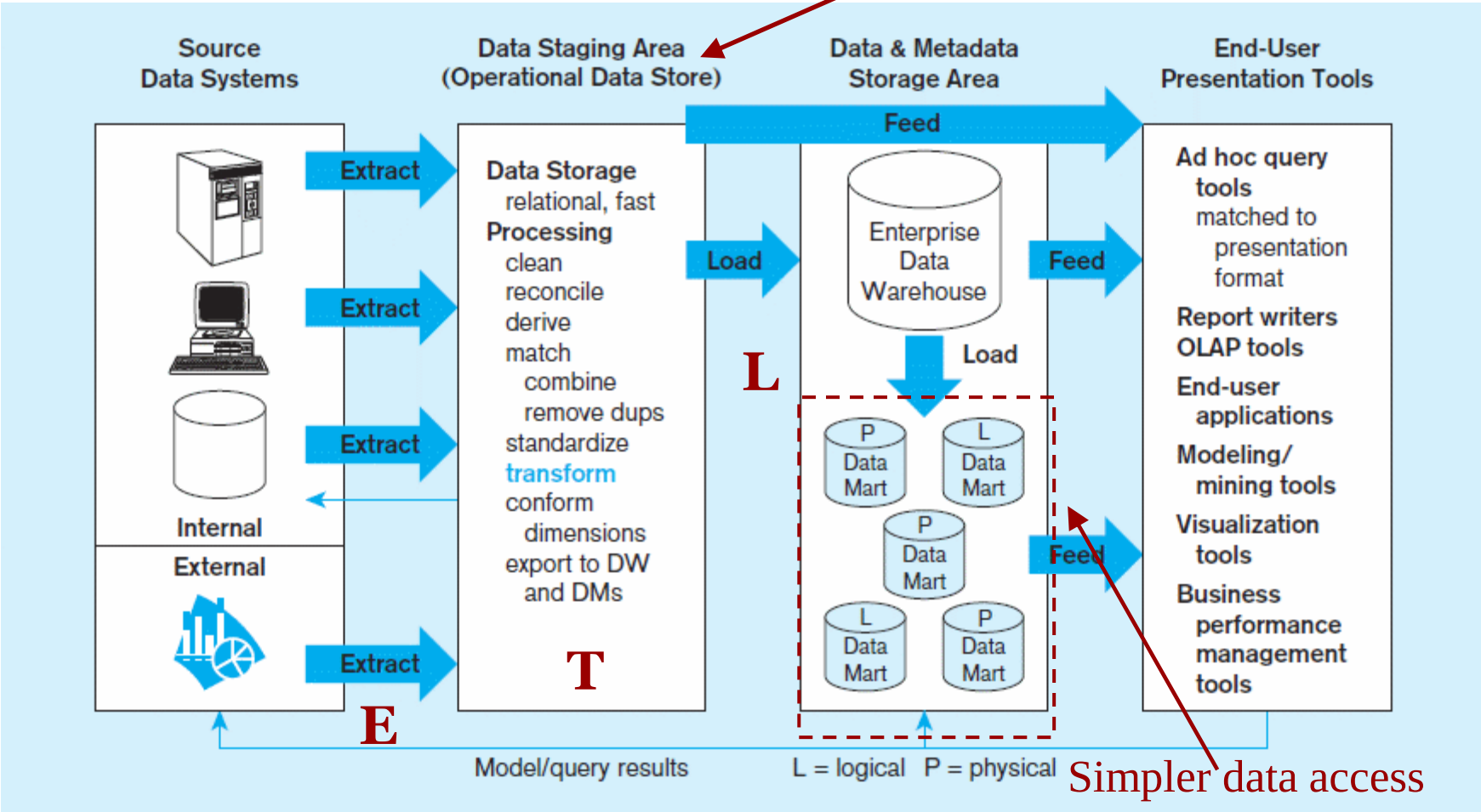


Separate ETL for each ***independent*** data mart

Data access complexity due to ***multiple*** data marts

Figure-3 Dependent data mart with operational data store: a three-level architecture

ODS provides option for obtaining ***current*** data

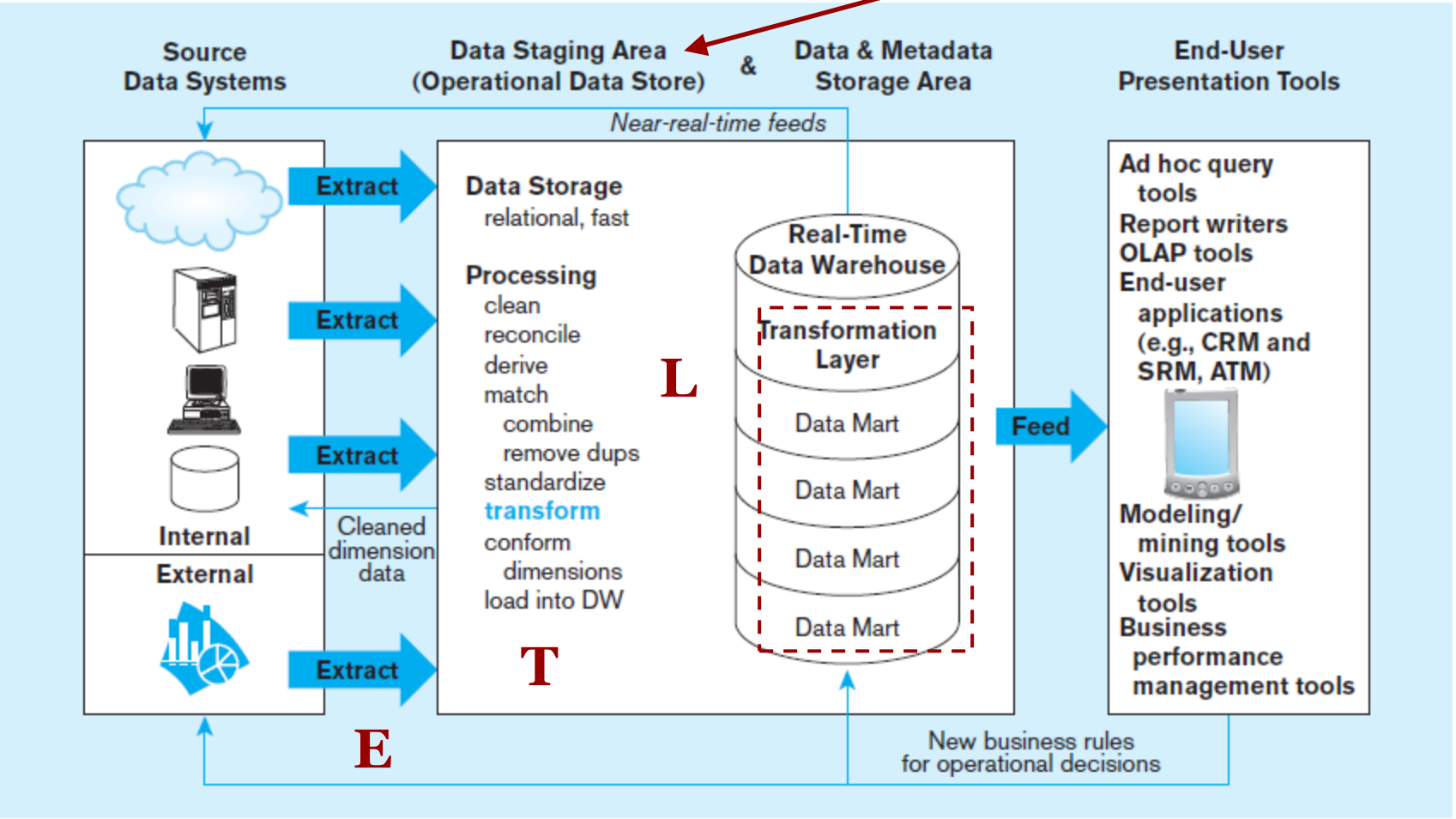


Single ETL for
enterprise data warehouse (EDW)

Dependent data marts loaded from EDW

Figure-4 Logical data mart and real-time warehouse architecture

ODS and **data warehouse** are one and the same



Near real-time ETL for ***Data Warehouse***

Data marts are NOT separate databases, but logical ***views*** of the data warehouse

➔ Easier to create new data marts

TABLE 9-2 Data Warehouse Versus Data Mart

Data Warehouse	Data Mart
Scope	**Scope**
• Application independent • Centralized, possibly enterprise-wide • Planned	• Specific DSS application • Decentralized by user area • Organic, possibly not planned
Data	**Data**
• Historical, detailed, and summarized • Lightly denormalized	• Some history, detailed, and summarized • Highly denormalized
Subjects	**Subjects**
• Multiple subjects	• One central subject of concern to users
Sources	**Sources**
• Many internal and external sources	• Few internal and external sources
Other Characteristics	**Other Characteristics**
• Flexible • Data oriented • Long life • Large • Single complex structure	• Restrictive • Project oriented • Short life • Start small, becomes large • Multi, semi-complex structures, together complex

THANK YOU