Data Mining

Asso. Profe. Dr. Raed Ibraheem Hamed

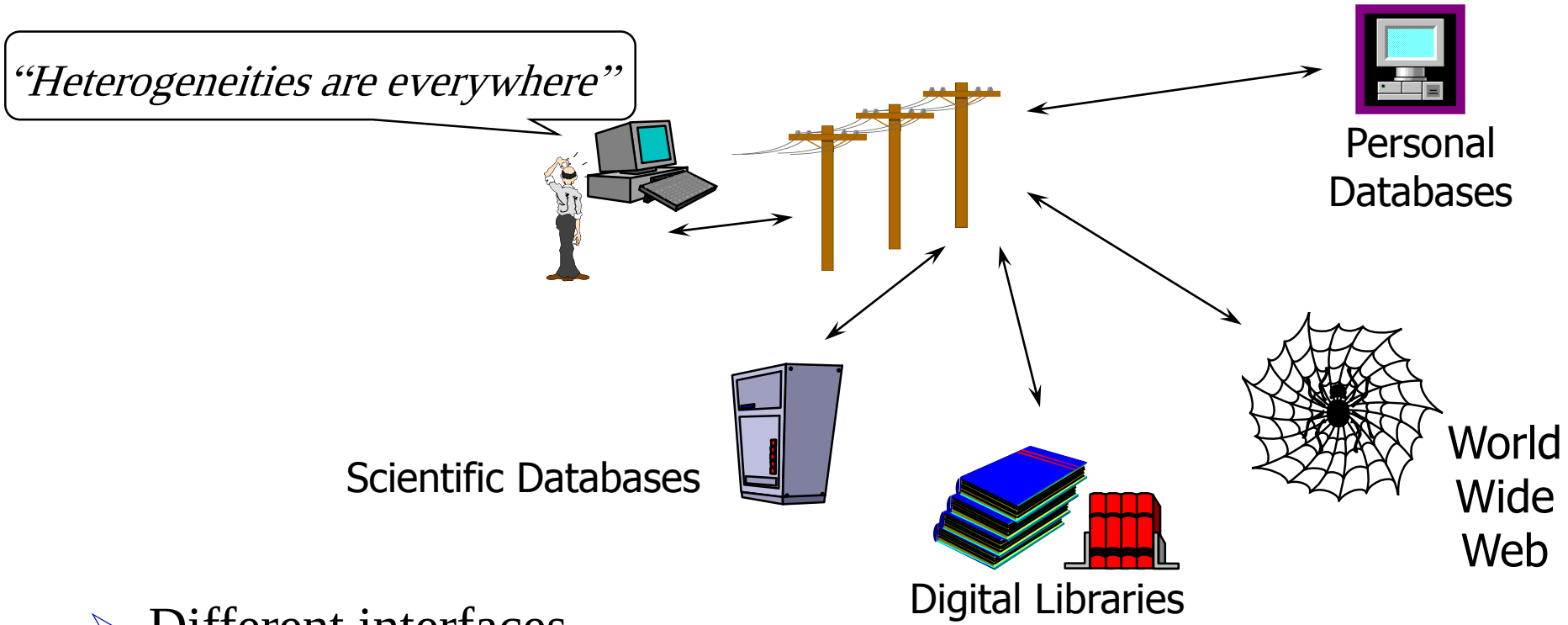
University of Human Development,
College of Science and Technology
Department of CS

2016 – 2017

Points to Cover

- Problem: Heterogeneous Information Sources
- Data integration
- Extract, Transform, Load (ETL)
- Data Warehouse With The ETL Process
- Problem: Data Management in Large Enterprises
- Goals of Data Integration
- Current Solutions
- Three-layer Architecture: Conceptual View
- Generic Warehouse Architecture

Problem: Heterogeneous Information Sources



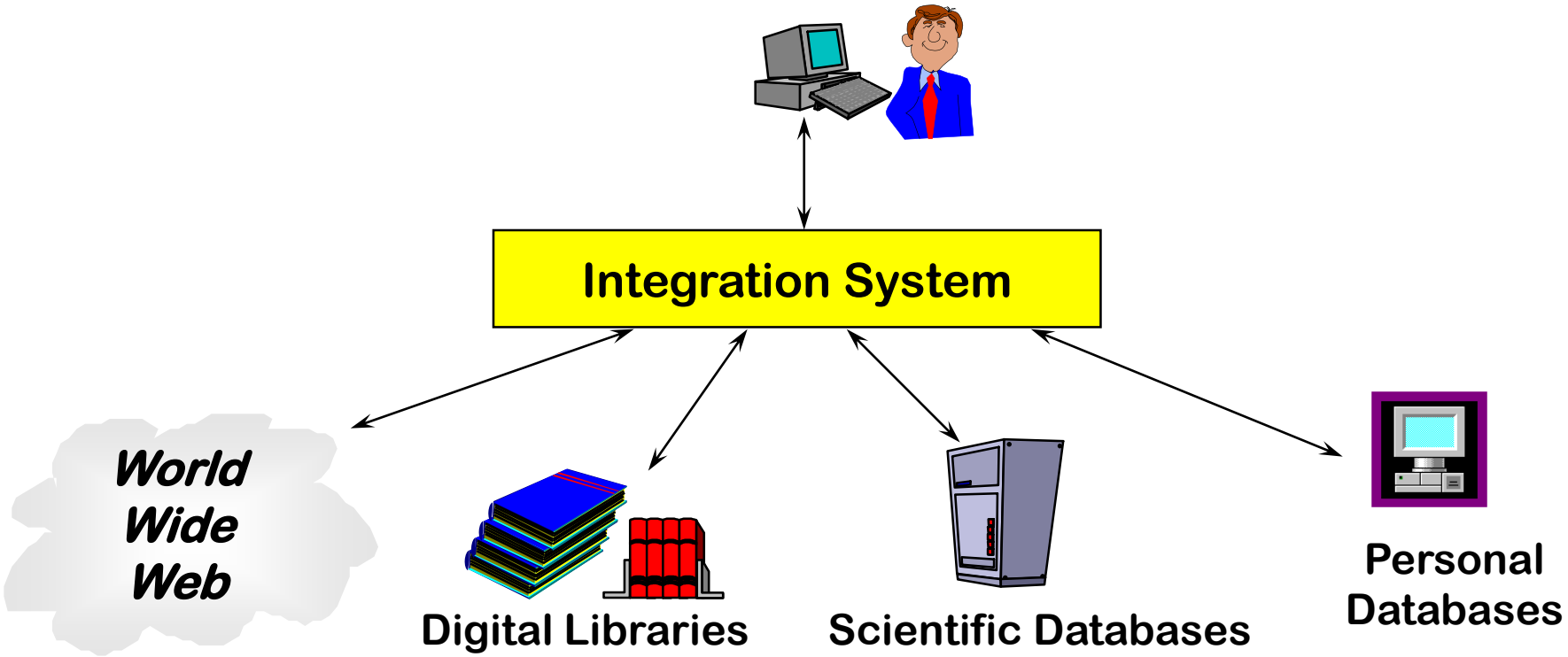
- Different interfaces
- Different data representations
- Duplicate and inconsistent information

Data integration

- **Data integration** involves combining data residing in different sources and providing users with a **unified** view of these data.

- This process becomes significant in a variety of situations, which include both **commercial** (when two similar companies need to merge their databases) and **scientific** research results.

Goal: Unified Access to Data



- Collects and combines information
- Provides integrated view, uniform user interface
- Supports sharing among different applications

Goals of Data Integration

- **Provide**
  1) Uniform (same query interface to all sources)
  2) Access to (updates the database)
  3) Multiple (we want many users at each time)
  4) Heterogeneous (data models are different)
  5) Distributed (over LAN, WAN, Internet)
  6) Data Sources (not only databases).

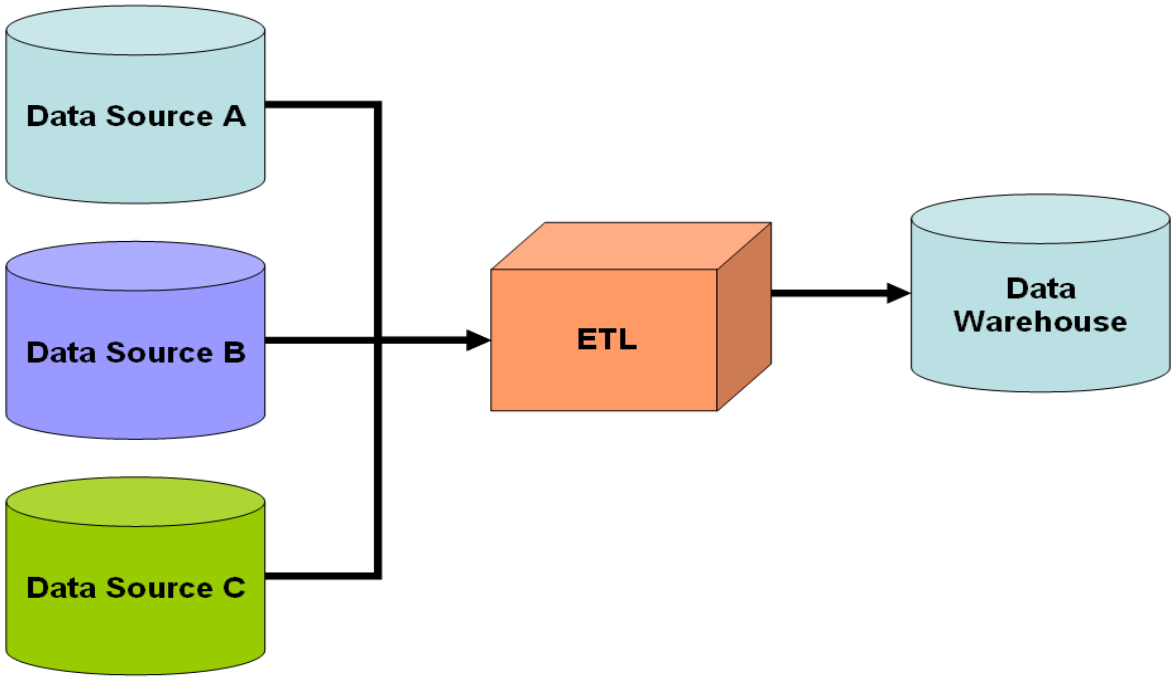
Extract, Transform, Load (ETL)

Extract, Transform, Load (ETL) refers to a process in database.

- **Data extraction** is where data is extracted from heterogeneous data sources;
- **Data transformation** where the data is transformed for storing in the proper format or structure.
- **Data loading** where the data is loaded into the final target more specifically, an **operational data store**, **data mart**, or **data warehouse**.

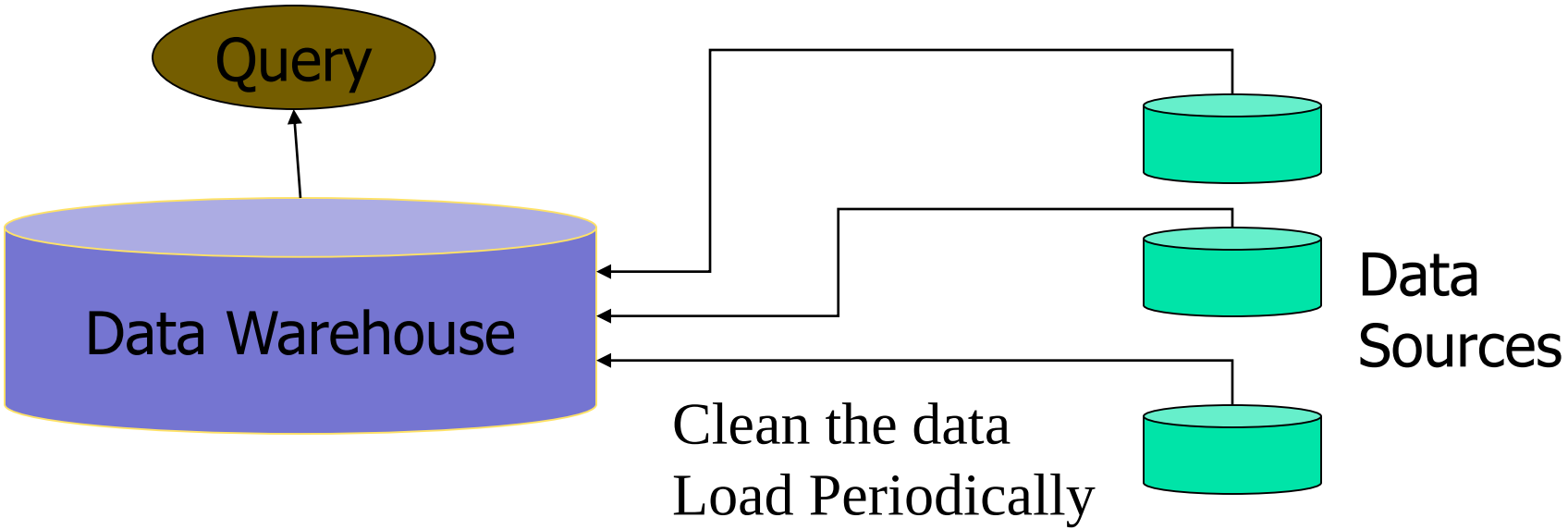
Data Warehouse With The ETL Process

A simple schematic for a data warehouse with the ETL process extracts information from the source databases, transforms it and then loads it into the data warehouse.



Extract all the data into a single data source

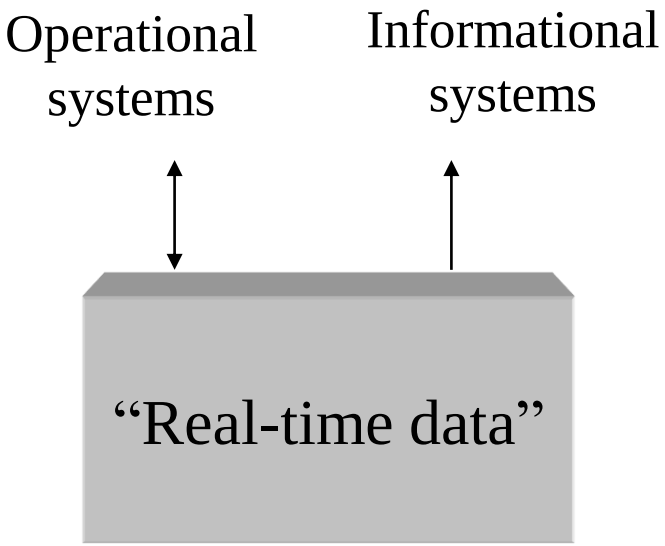
- Data Warehouse
  - Extract all the data into a single data source



Data Warehouse Architectures: Conceptual View

Single-layer

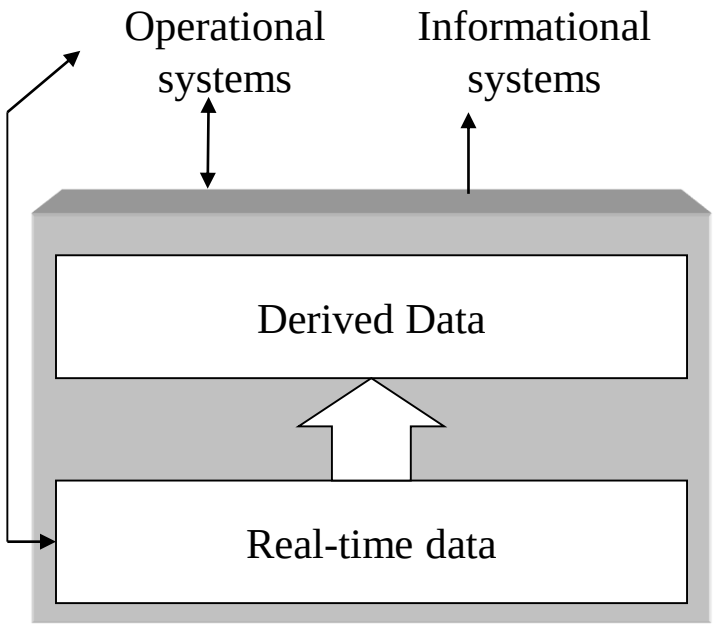
- Every data element is stored once only
- Virtual warehouse: is another term for a data warehouse.



Data Warehouse Architectures: Conceptual View

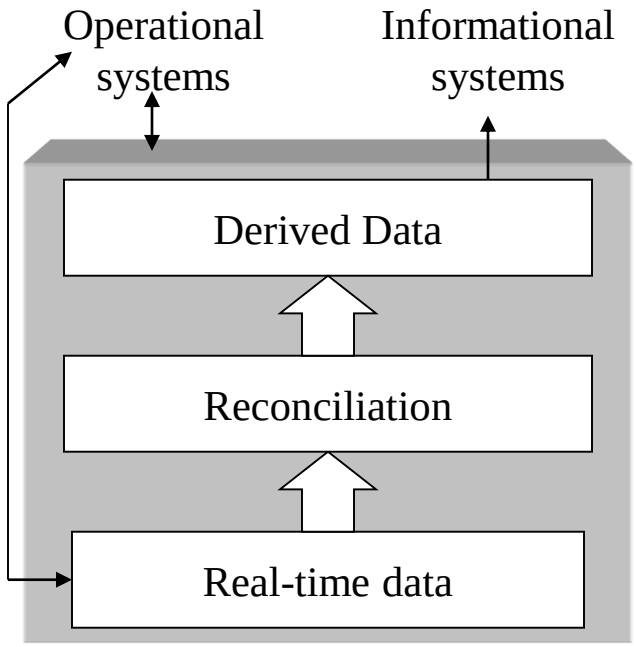
Two-layer

Real-time + derived data Most commonly used approach in industry today



Three-layer Architecture: Conceptual View

- Transformation of real-time data to derived data really requires reconciliation step



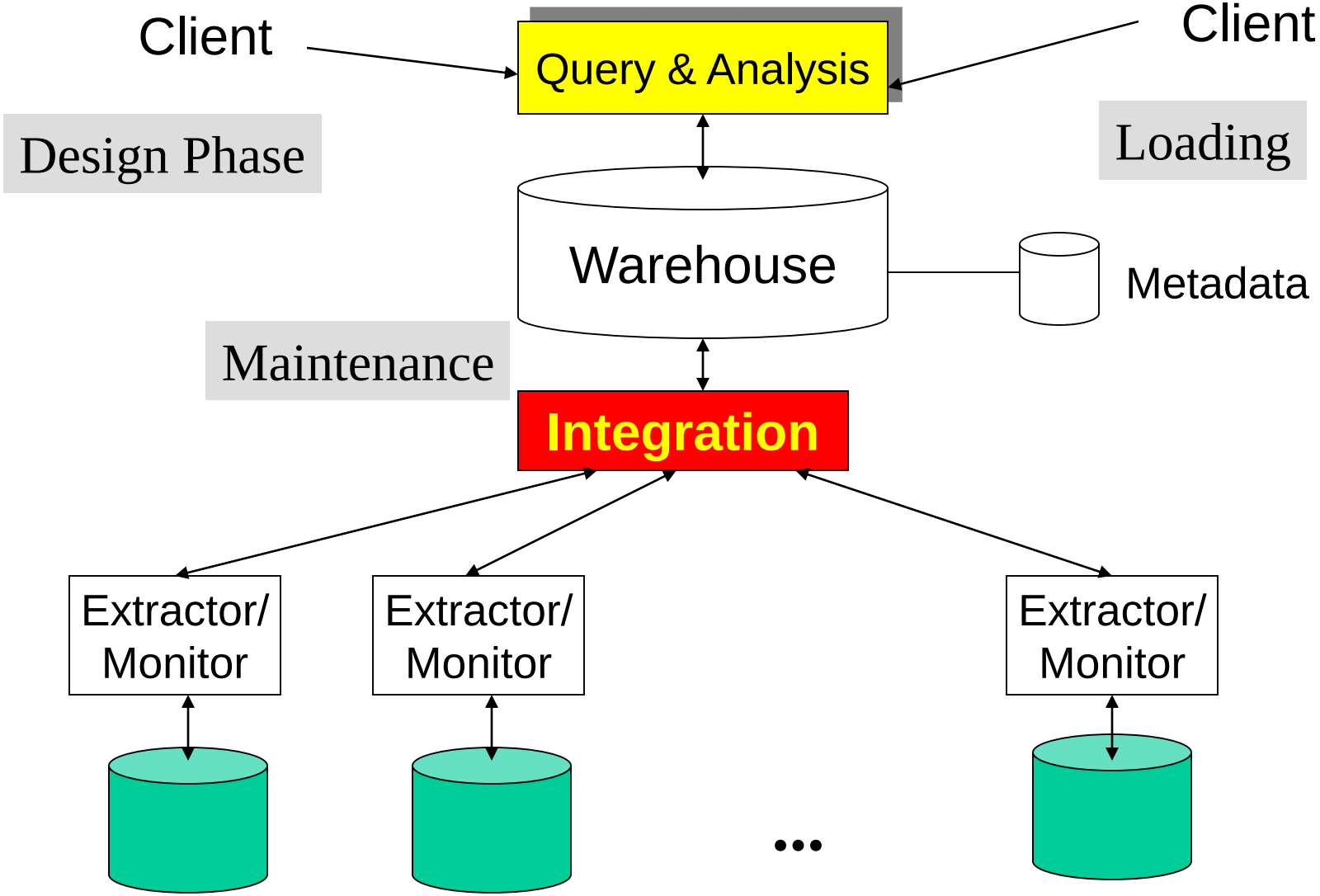
Reconciliation is the process of ensuring that two sets of records are in agreement.

Example:- Reconciliation is used to ensure that the money leaving an account matches the actual money spent.

Issues in Data Warehousing

- Warehouse Design
- Extraction
- Integration
  - Cleansing & merging
- Warehousing specification & Maintenance
- Optimizations
- Evolution

Generic Warehouse Architecture



Problems with DW Approach

- Data has to be cleaned – different formats
- Needs to store all the data from different data sources in single system
- Data needs to be updated periodically
  - Data sources are independent– content can change without notice
  - Expensive because of the large quantities of data and data cleaning costs


Thank
you