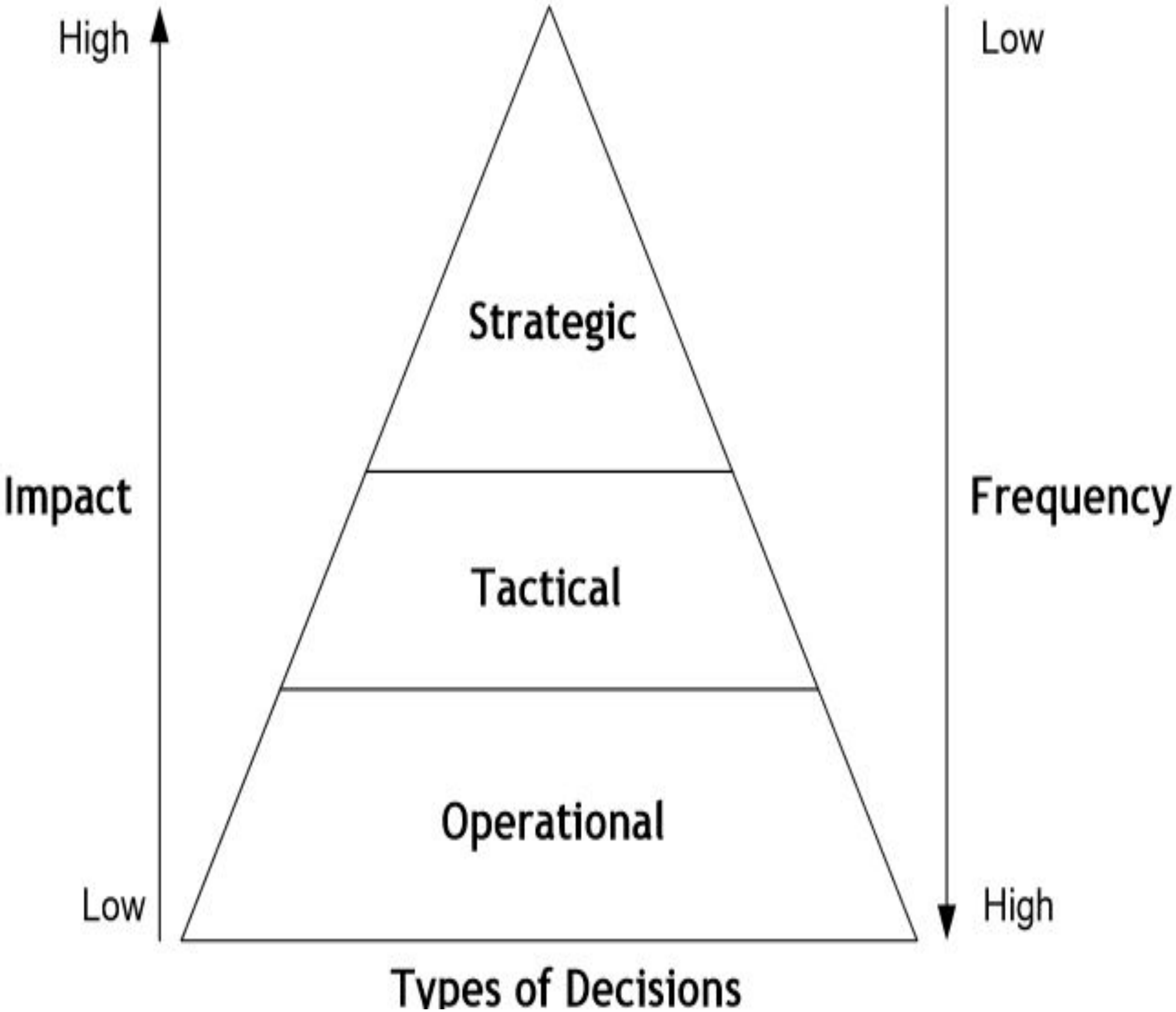
Data warehousing Concepts

- Why data warehouse?
- What’s data warehouse?
- What’s multi-dimensional data model?
- What’s difference between OLAP and OLTP?


High
Low
Impact
Strategic
Tactical
Operational
Low
High
Frequency
Types of Decisions

Operational – The information which is required to run day to day business operation activities. (Producing an invoice, make a shipment, settle a claim, post a withdrawal).

Strategic – This information is meant for executives and managers who are responsible for keeping the enterprise competitive.

- They need the information to make right decisions at the right time in the right format.
- Retain current customers of the business.
- Add to customer base by atleast 10% over next two years.
- Enhance the market share by 15%
- Launch new and better products in market
- Increase saes in north east region by 10%

- Characteristics of Strategic Information
- Integrated
- Data Integrity
- Accessible
- Timely

The Information Crisis

- Reasons –
- Organizations have huge amount of data.
- The Information systems that have are ineffective in turning this into useful strategic information.

- "Colossal amounts of data already exist which doubles every 18 months"

The Information Crisis

Reasons –

The data in corporation resides in various disparate systems and diverse structures.

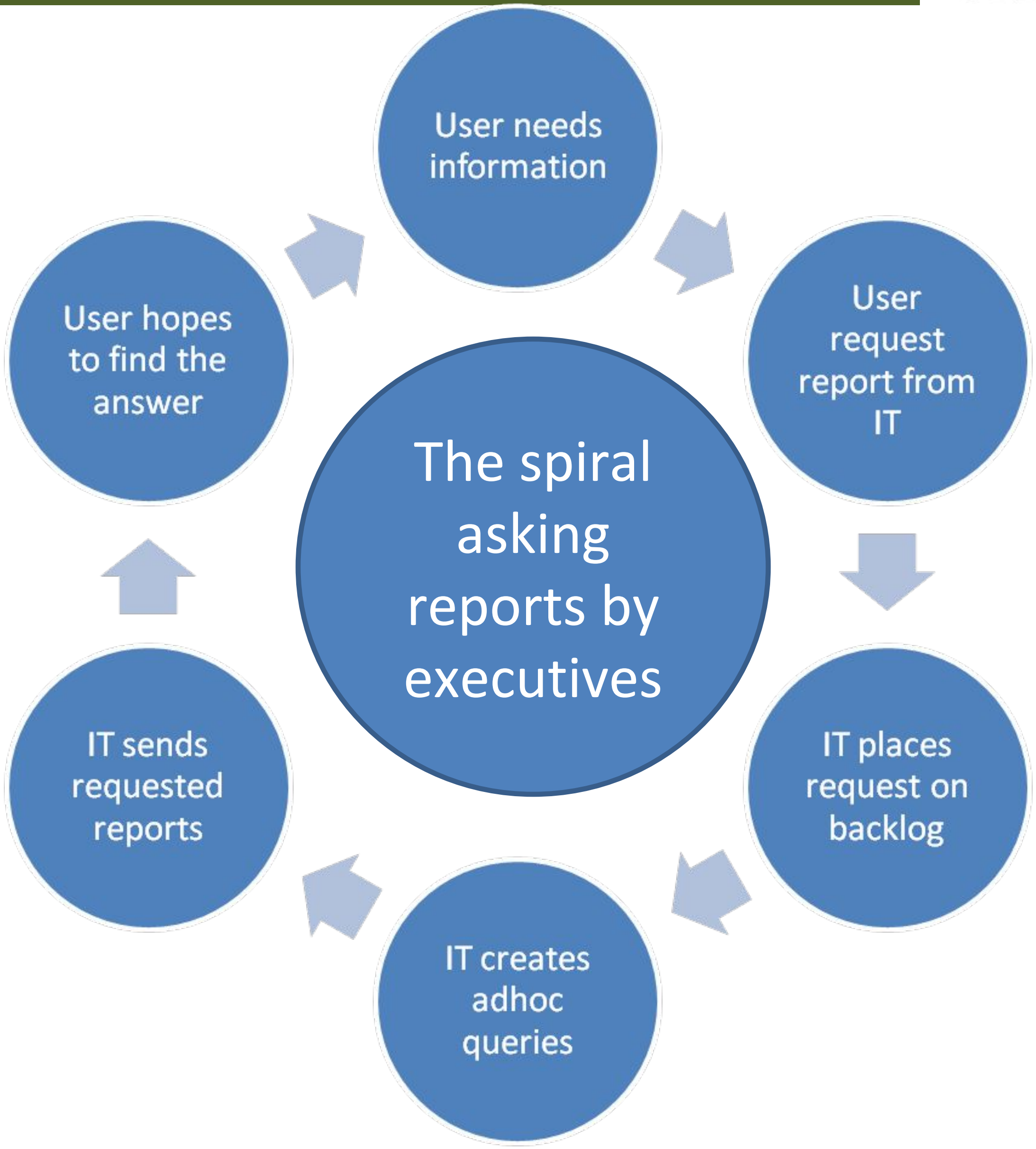
Data needed for making strategic decisions must be available in format that enables executives and managers to analyse trends in order to lead their companies in right direction.

The Information Crisis

Reasons –

Operation data is event driven (the data which you record in detail and for each and every transactions). This data is not useful for managers until and unless we transform.

Inadequate attempts by IT to provide strategic info


The spiral asking reports by executives
User needs information
User request report from IT
IT places request on backlog
IT creates adhoc queries
IT sends requested reports
User hopes to find the answer

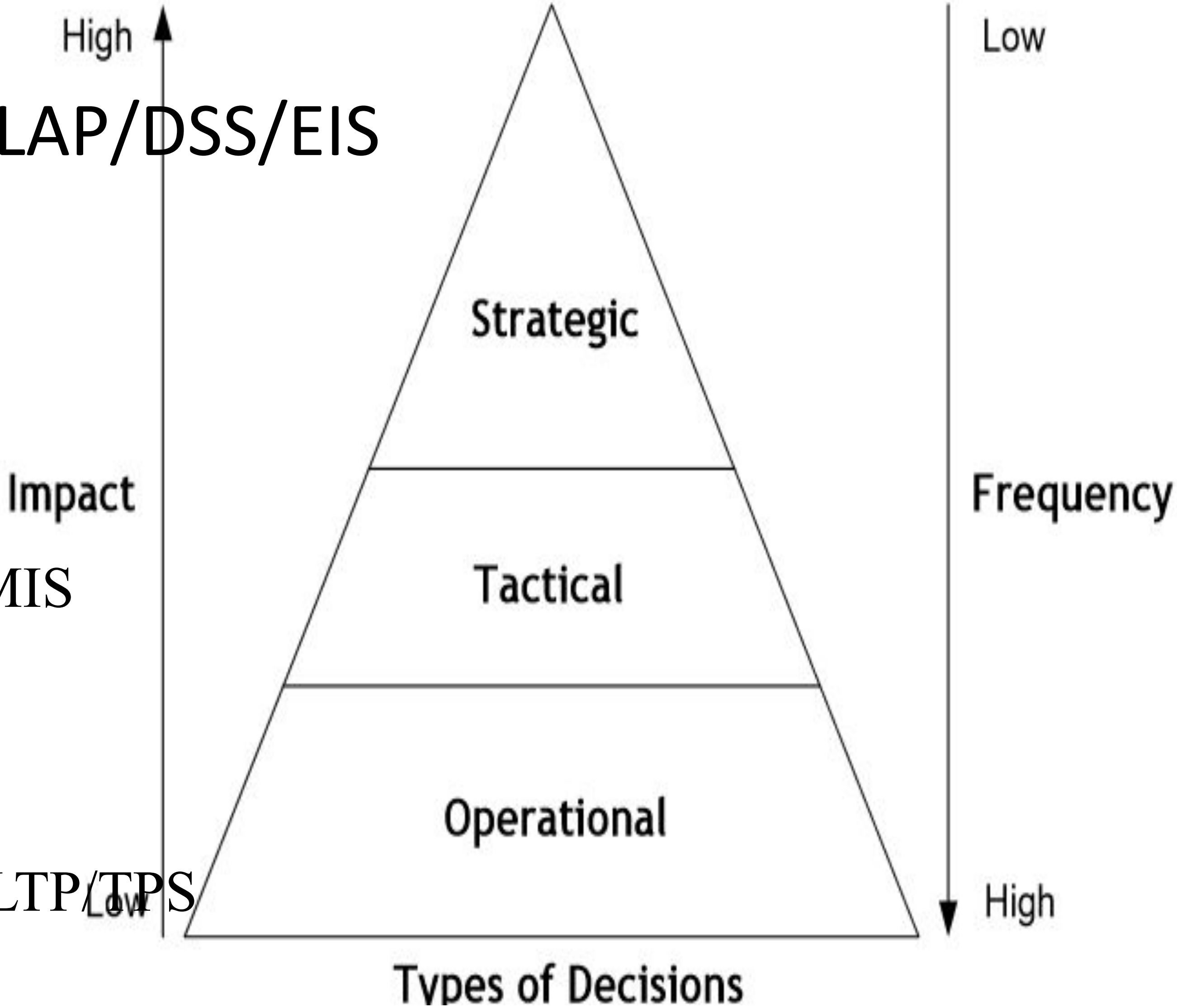
Decision Support Systems

- By the **1960s came a DBMS called *Decisions Support Systems* (DSS) which was a collection of software**.

- **In 1989 and early 1990s there were various such software in use like *Executive Information System* (EIS), *Online Analytical Processes* (OLAP).**

- **The term Business Intelligence (term used by Howard Dresner of Gartner Group)** started getting used as a general term encompassing all such methods and applications.

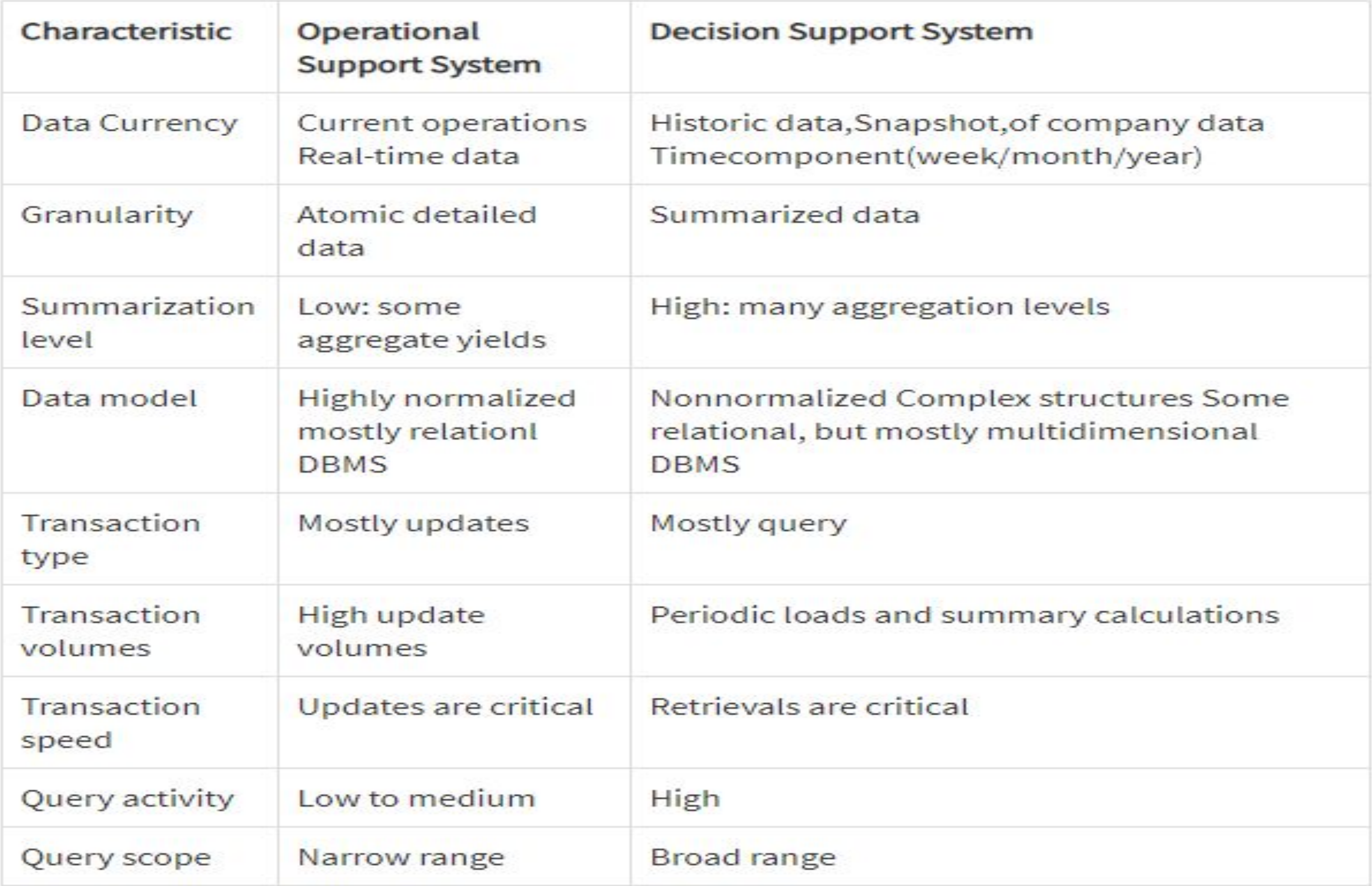
Information Systems



OLAP/DSS/EIS

MIS

OLTP/TPS



Characteristic	Operational Support System	Decision Support System
Data Currency	Current operations Real-time data	Historic data,Snapshot,of company data Timecomponent(week/month/year)
Granularity	Atomic detailed data	Summarized data
Summarization level	Low: some aggregate yields	High: many aggregation levels
Data model	Highly normalized mostly relationl DBMS	Nonnormalized Complex structures Some relational, but mostly multidimensional DBMS
Transaction type	Mostly updates	Mostly query
Transaction volumes	High update volumes	Periodic loads and summary calculations
Transaction speed	Updates are critical	Retrievals are critical
Query activity	Low to medium	High
Query scope	Narrow range	Broad range

University Tables

Student

matricNum	**fName**	**lName**	**gender**	**year reg**	***supervisor***
121212	Mary	Hill	F	2003	*1234*
232323	Steve	Gray	M	2005	*1234*
123456	Jimmy	Smith	M	2000	*1111*

Staff

staff Num	**first Name**	**last Name**	**gender**
1234	Jane	Smith	F
2323	Tom	Green	M
1111	Jim	Brown	M

Course

course code	**credit value**
c1	120
c3	60
c5	60

Enrolled

course code	***student Num***
c1	*121212*
c3	*121212*
c3	*123456*
c1	*232323*
Etc etc	*Etc etc*

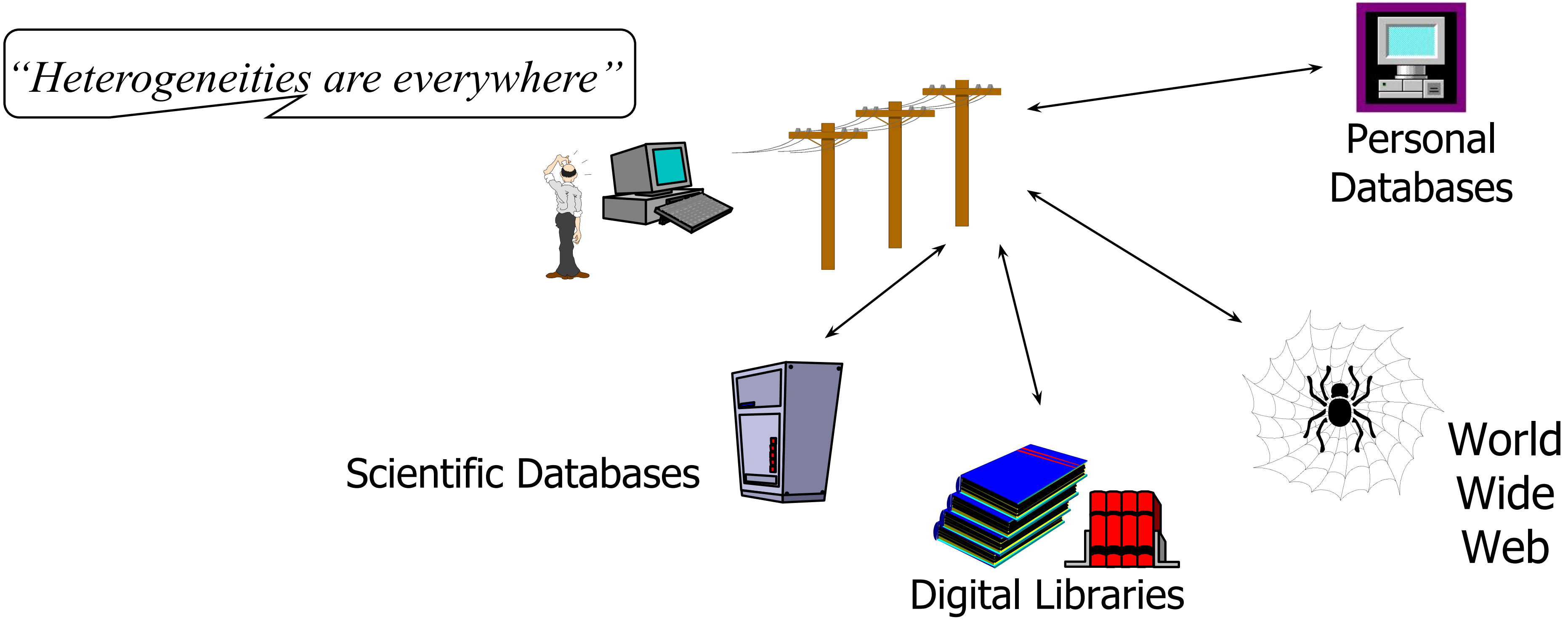
Relation Database Theory, cont'd

- The process of normalization generally breaks a table into many independent tables.
- A normalized database yields a flexible model, making it easy to maintain dynamic relationships between business entities.
- A relational database system is effective and efficient for operational databases – a lot of updates (aiming at optimizing update performance).

Problems

- A fully normalized data model can perform very inefficiently for queries.
- Historical data are usually large with static relationships:
  - Unnecessary joins may take unacceptably long time
- Historical data are diverse

Problem: Heterogeneous Information Sources



"Heterogeneities are everywhere"

Personal Databases

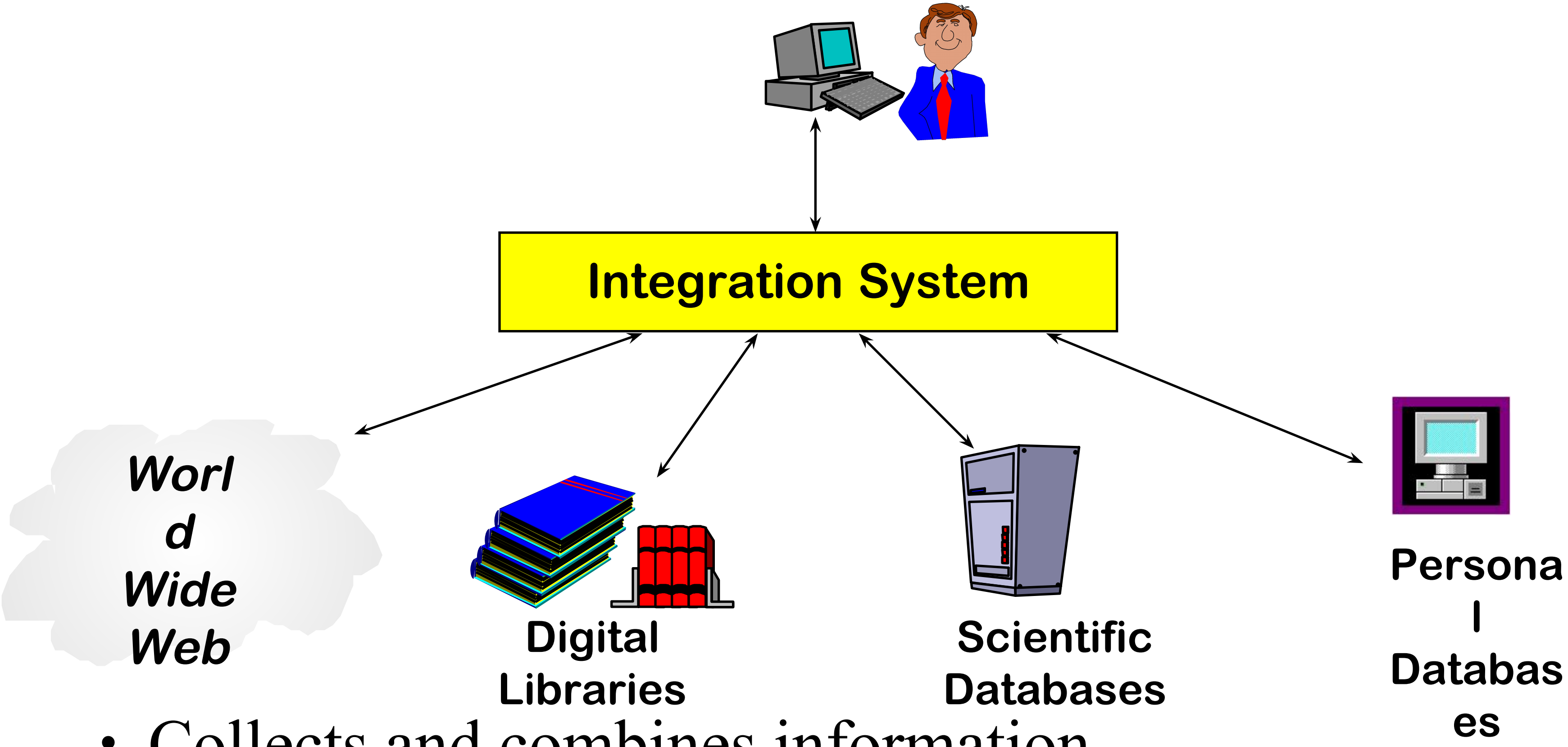
Scientific Databases

Digital Libraries

World Wide Web

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

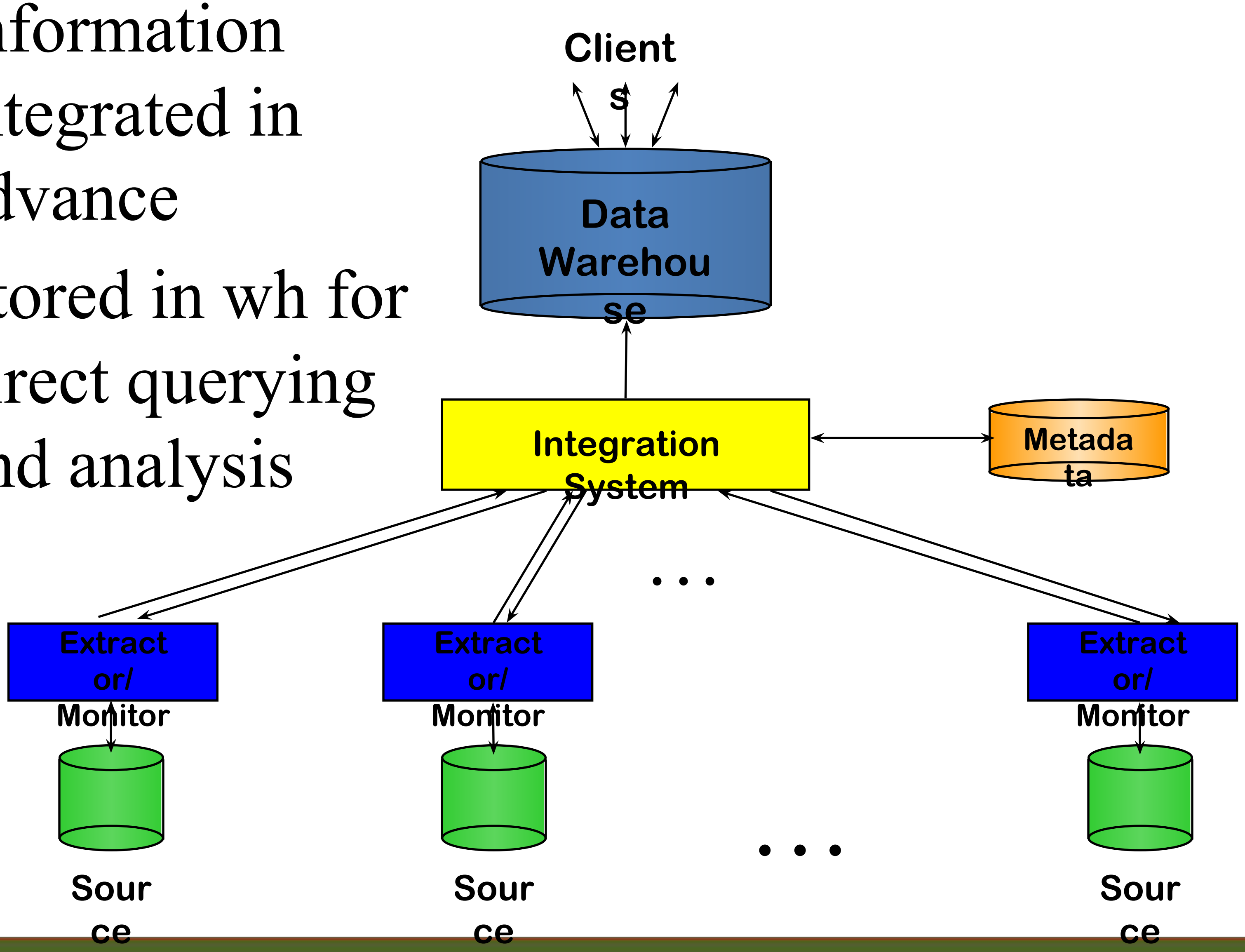
Goal: Unified Access to Data



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

The Warehousing Approach

- Information integrated in advance
- Stored in wh for direct querying and analysis



Advantages of Warehousing Approach

- High query performance
  - But not necessarily most current information
- Doesn’t interfere with local processing at sources
  - Complex queries at warehouse
  - OLTP at information sources
- Information copied at warehouse
  - Can modify, annotate, summarize, restructure, etc.
  - Can store historical information
  - Security, no auditing
- *Has* caught on in industry

What is a Data Warehouse?
A Practitioners Viewpoint

"A data warehouse is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use it in a business context."

-- Barry Devlin, *IBM Consultant*

What is a Data Warehouse?
An Alternative Viewpoint

"A DW is a

- subject-oriented,
- integrated,
- time-varying,
- non-volatile

collection of data that is used primarily in organizational decision making."

-- W.H. Inmon, Building the Data Warehouse, 1992

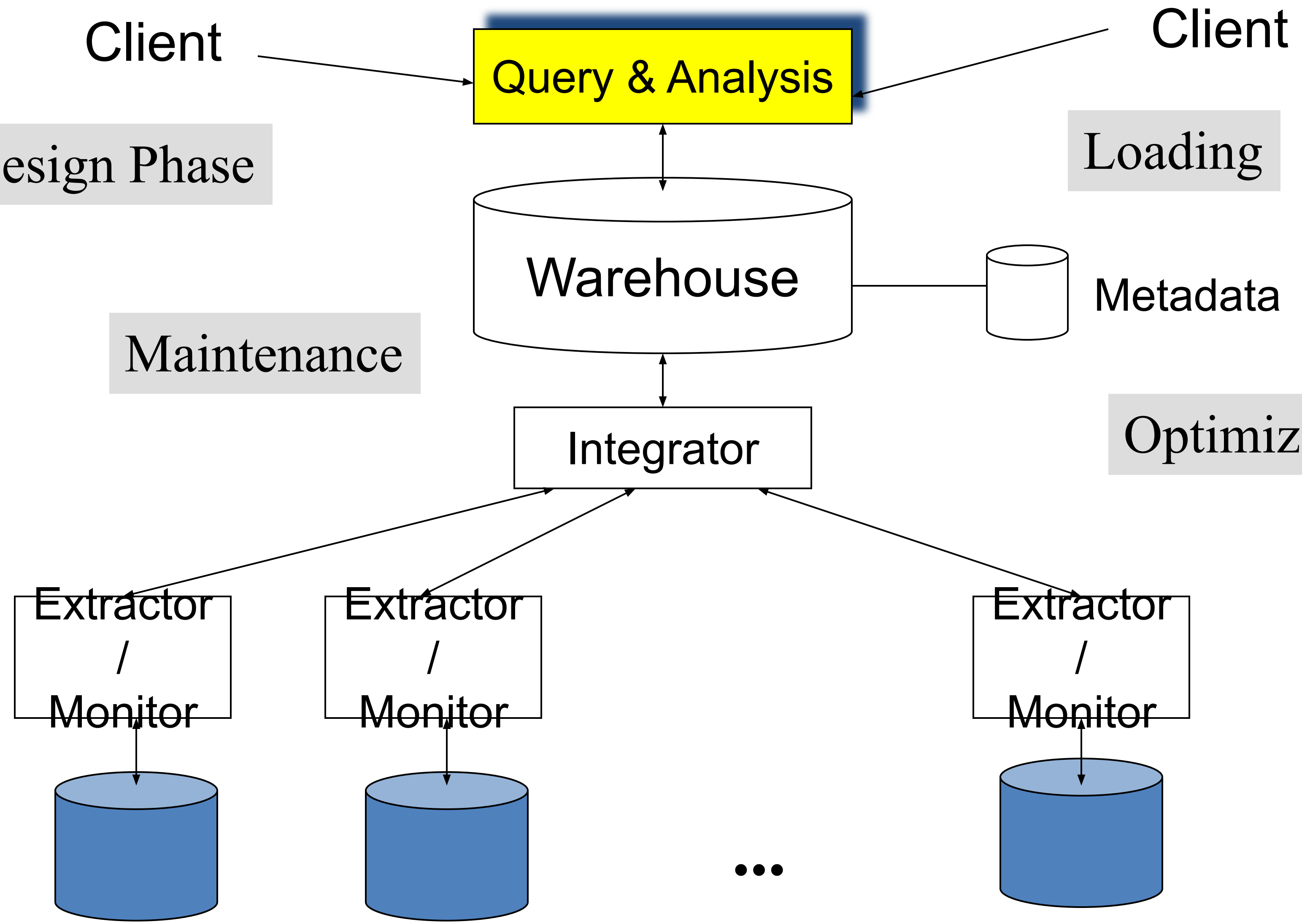
A Data Warehouse is...

- Stored collection of diverse data
  - A solution to data integration problem
  - Single repository of information
- Subject-oriented
  - Organized by subject, not by application
  - Used for analysis, data mining, etc.
- Optimized differently from transaction-oriented db
- User interface aimed at executive

… Cont’d

- Large volume of data (Gb, Tb)
- Non-volatile
  - Historical
  - Time attributes are important
- Updates infrequent
- May be append-only
- Examples
  - All transactions ever at Sainsbury’s
  - Complete client histories at insurance firm
  - LSE financial information and portfolios

Generic Warehouse Architecture



Design Phase

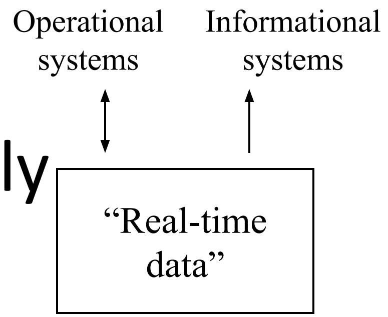
Loading

Maintenance

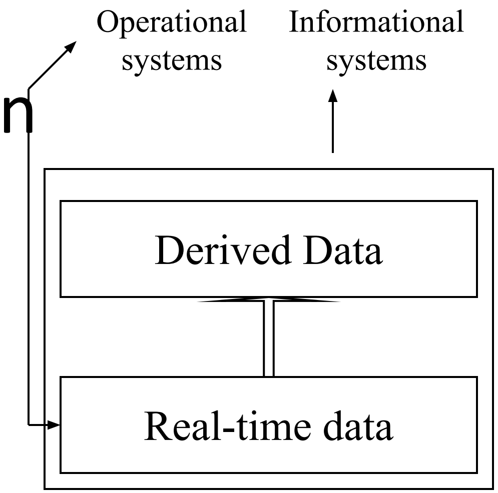
Optimization

Data Warehouse Architectures: Conceptual View

- Single-layer
  - Every data element is stored once only
  - Virtual warehouse

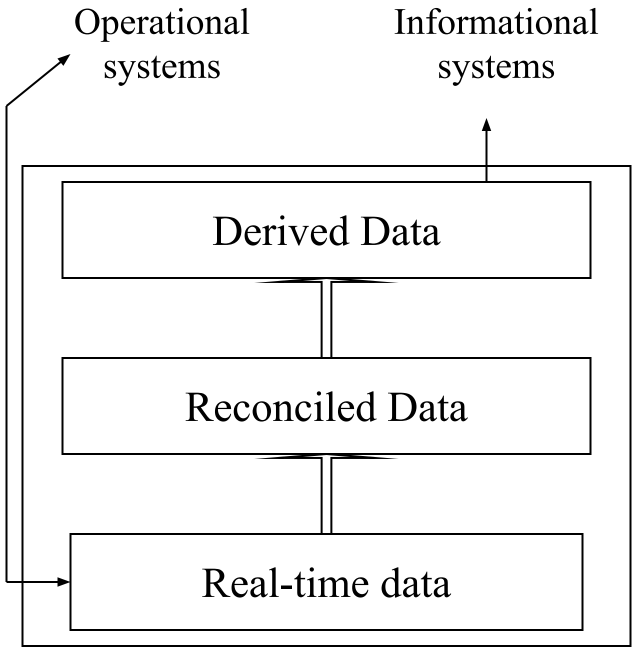


- Two-layer
  - Real-time + derived data
  - Most commonly used approach in industry today



Three-layer Architecture: Conceptual View

Transformation of real-time data to derived data really requires two steps



View level
"Particular informational needs"

Physical Implementation of the Data Warehouse

Data Warehousing: Two Distinct Issues

(1) How to get information into warehouse

"Data warehousing"

(2) What to do with data once it's in warehouse

"Warehouse DBMS"

- Both rich research areas
- Industry has focused on (2)

Issues in Data Warehousing

- Warehouse Design
- Extraction
  - Wrappers, monitors (change detectors)
- Integration
  - Cleansing & merging
- Warehousing specification & Maintenance
- Optimizations
- Miscellaneous (e.g., evolution)

OLTP vs. OLAP

- OLTP: On Line Transaction Processing
  - Describes processing at operational sites
- OLAP: On Line Analytical Processing
  - Describes processing at warehouse

Warehouse is a Specialized DB

Standard DB (OLTP)

- Mostly updates
- Many small transactions
- Mb - Gb of data
- Current snapshot
- Index/hash on p.k.
- Raw data
- Thousands of users (e.g., clerical users)

Warehouse (OLAP)

- Mostly reads
- Queries are long and complex
- Gb - Tb of data
- History
- Lots of scans
- Summarized, reconciled data
- Hundreds of users (e.g., decision-makers, analysts)

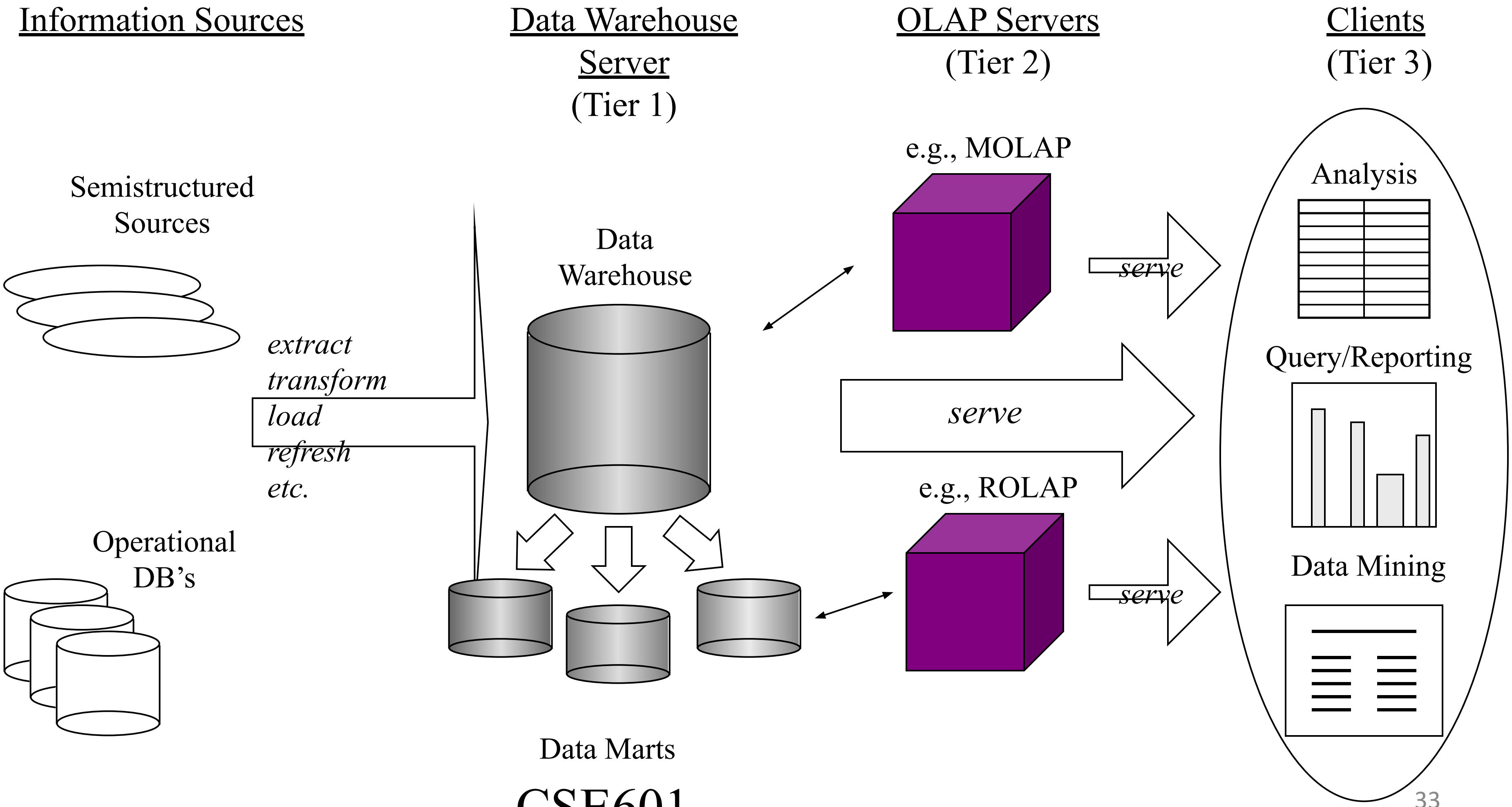
Decision Support

- Information technology to help the knowledge worker (executive, manager, analyst) make faster & better decisions
  - *“What were the sales volumes by region and product category for the last year?”*
  - *“How did the share price of comp. manufacturers correlate with quarterly profits over the past 10 years?”*
  - *“Which orders should we fill to maximize revenues?”*
- On-line analytical processing (OLAP) is an element of decision support systems (DSS)

Three-Tier Decision Support Systems

- Warehouse database server
  - Almost always a relational DBMS, rarely flat files
- OLAP servers
  - Relational OLAP (ROLAP): extended relational DBMS that maps operations on multidimensional data to standard relational operators
  - Multidimensional OLAP (MOLAP): special-purpose server that directly implements multidimensional data and operations
- Clients
  - Query and reporting tools
  - Analysis tools
  - Data mining tools

The Complete Decision Support System



CSE601

Data Warehouse vs. Data Marts

- *Enterprise warehouse*: collects all information about subjects (*customers,products,sales,assets, personnel*) that span the entire organization
  - Requires extensive business modeling (may take years to design and build)
- *Data Marts*: Departmental subsets that focus on selected subjects
  - Marketing data mart: customer, product, sales
  - Faster roll out, but complex integration in the long run
- *Virtual warehouse*: views over operational dbs
  - Materialize sel. summary views for efficient query processing
  - Easy to build but require excess capability on operat. db

OLAP for Decision Support

- OLAP = Online Analytical Processing
- Support (almost) ad-hoc querying for business analyst
- Think in terms of spreadsheets
    - View sales data by geography, time, or product
- Extend spreadsheet analysis model to work with warehouse data
    - Large data sets
    - Semantically enriched to understand business terms
    - Combine interactive queries with reporting functions
- <u>Multidimensional view of data is the foundation of OLAP</u>
    - Data model, operations, etc.

Approaches to OLAP Servers

- Relational DBMS as Warehouse Servers
- Two possibilities for OLAP servers
- (1) Relational OLAP (ROLAP)
  - Relational and specialized relational DBMS to store and manage warehouse data
  - OLAP middleware to support missing pieces
- (2) Multidimensional OLAP (MOLAP)
  - Array-based storage structures
  - Direct access to array data structures

OLAP Server: Query Engine Requirements

- Aggregates (maintenance and querying)
  - Decide what to precompute and when
- Query language to support multidimensional operations
  - Standard SQL falls short
- Scalable query processing
  - Data intensive and data selective queries