Data Mining and Warehousing

Week 1,2

Instructor : Dr. Natash Ali Mian

School of Computer and Information Technology

Beaconhouse National University

Switch off mobile phones during lectures, or put them into silent mode

Contents

- Class Introduction — **Not Required ☺**
- Instructor's Introduction — **Not Required ☺**
- Introduction to Course
- Student Guidelines — **Already Known ☺**

Course Objective

- To provide the Introduction of Datawarehouse and its purpose.
- To introduce the techniques, tools and applications of data mining,
- To apply DM techniques to a variety of research and application projects.

Key Topics

- Data Mining
- Knowledge Discovery Process
- Introduction to Warehouse
- Data Marts
- Description of Data Warehouse
- Operational Vs Information Systems
- Data Warehouse Architecture
- Decision Support System
- Dimensional Modeling
- Designing a Data warehouse

Key Topics

- Data Mining
- Data Reduction Techniques
- Statistical Methods in Data Mining
- Association rule mining
- Classification
- Cluster Analysis
- Advance topics in data mining
- Text mining, web mining, opinion mining, etc.

Reference Book

- ***Introduction to Data Mining with Case Studies* by G. K. Gupta**
- Mehmed Kantatardzic, *Data Mning: Concepts, Models, Methods, and Algorithms*, 2003, John Wiley and Sons.
- Margaret H. Dunham and S. Sridhar, *Data Mining*, *Introductory and Advanced Topics,* 2006, Pearson Education,
- David Hand, Heikki MAnnila and Padhraic Smyth, *Principles of Data Mining*, 2001, The MIT Press.
- Daniel T. Larose, *Data Mining Methods and Models*, 2006, John Wiley and Sons.
- Max Bramer, *Principles of Data Mining*, 2007, Springer-Verlag.
- Paulraj Ponniah, *Data Warehousing Fundamentals*, 2005, John Wiley and Sons.
- *Chuck Ballard Dirk Herreman Don Schau Rhonda Bell, Eunsaeng Kim Ann Valencic*, *Data Modeling Techniques for Data Warehousing*, 1999, IBM Corporation, International Technical Support Organization.

Grading Scheme

• Class Tests (min 3)	10
• Assignments (min 3)	10
• Term Presentation	20
• Mid Term Tests (2)	20
• Final Exam	40
Total	**100**

Guidelines for Students (1)

- No quiz will be dropped.
- Use of Mobile Phones is not allowed in the class, If mobile phone rings (due to: call, sms, alarm, reminder or any other), you will be requested to leave the class and you will be marked ABSENT.
- Course material will be uploaded on BNU-CMS

Guidelines for Students (2)

- Students are encouraged to discuss assignments but it is extremely important that everyone works on his/her own assignment
- The cases of plagiarism will be dealt ruthlessly & will be marked Zero
- Late comers should consult their class fellows for the missing topics, they will not be revised in the class
- No Extensions in deadlines will be given
- Be punctual, After 10minutes you will be marked late and after 20 minutes you will be marked Absent

Guidelines for Students (3)

- You should keep a track of your attendance yourself (From Moodle), no flexibility in attendance will be given.
- Remember minimum attendance required to appear in final exam is 75%
- Don't request any flexibility or you have to face embarrassment
- Further guidelines will be given time to time

Contents

- Why We need a warehouse?
- Why RDB of any conventional DB system is not enough?
- Is there a difference of size only?
- How Data Warehouse assists in analysis?
- How Data Warehouse helps to mine data?
- Is historical data important?
- Why data processing is required?

Introduction to Data Warehousing

Enrico Franconi

Data Warehouse and OLAP

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

Relational Database Theory

- Relational database modeling process – normalization, relations or tables are progressively decomposed into smaller relations to a point where all attributes in a relation are very tightly coupled with the primary key of the relation.
  - First normal form: data items are atomic,
  - Second normal form: attributes fully depend on primary key,
  - Third normal form: all non-key attributes are completely independent of each other.

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*

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*

Staff

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

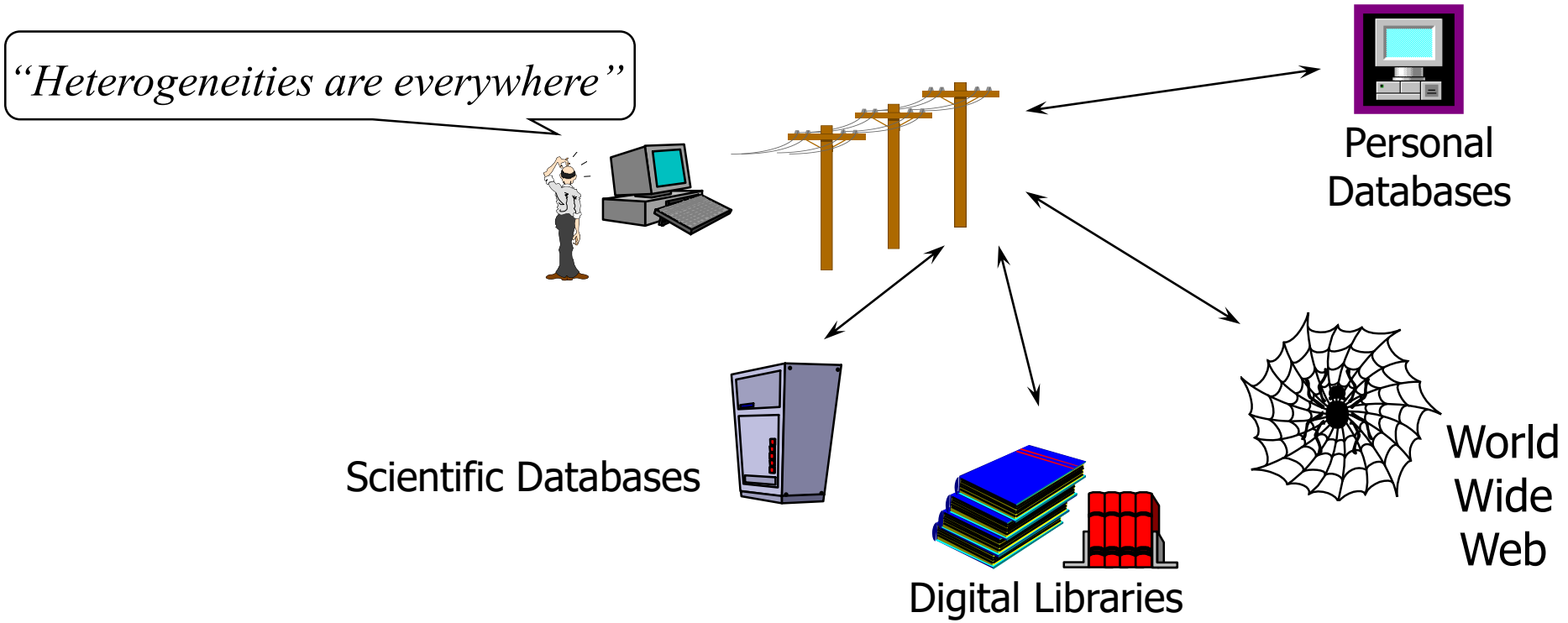
Relation Database Theory, cont'd

- 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

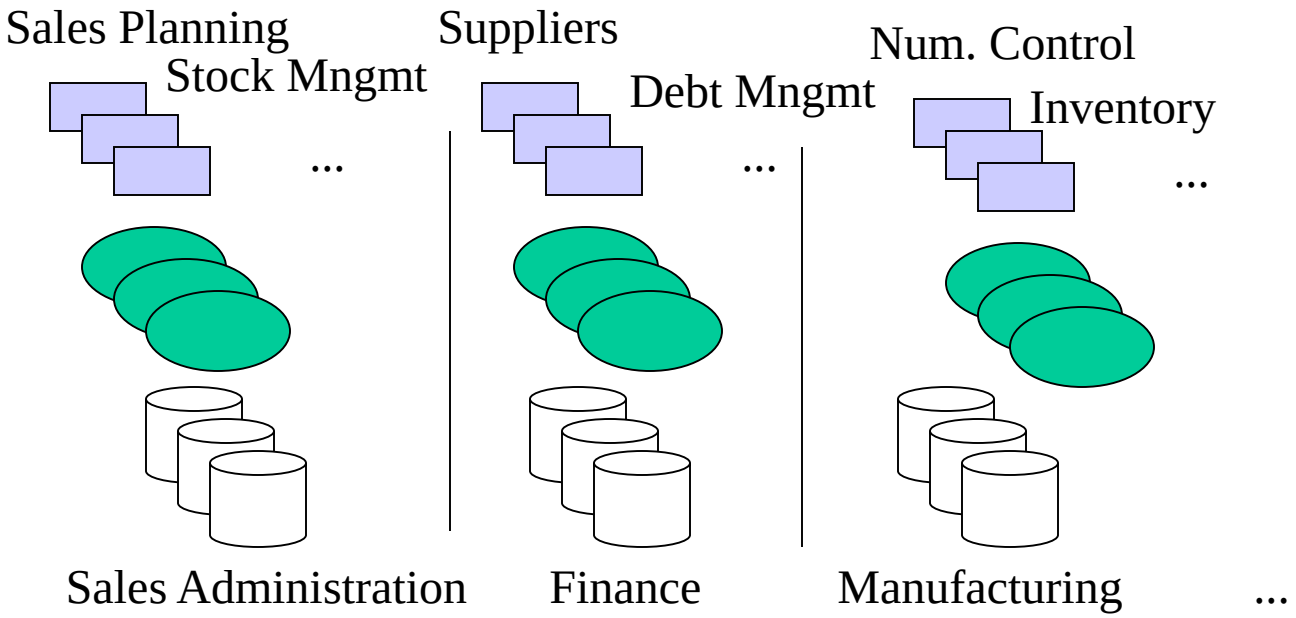
Problem: Heterogeneous Information Sources



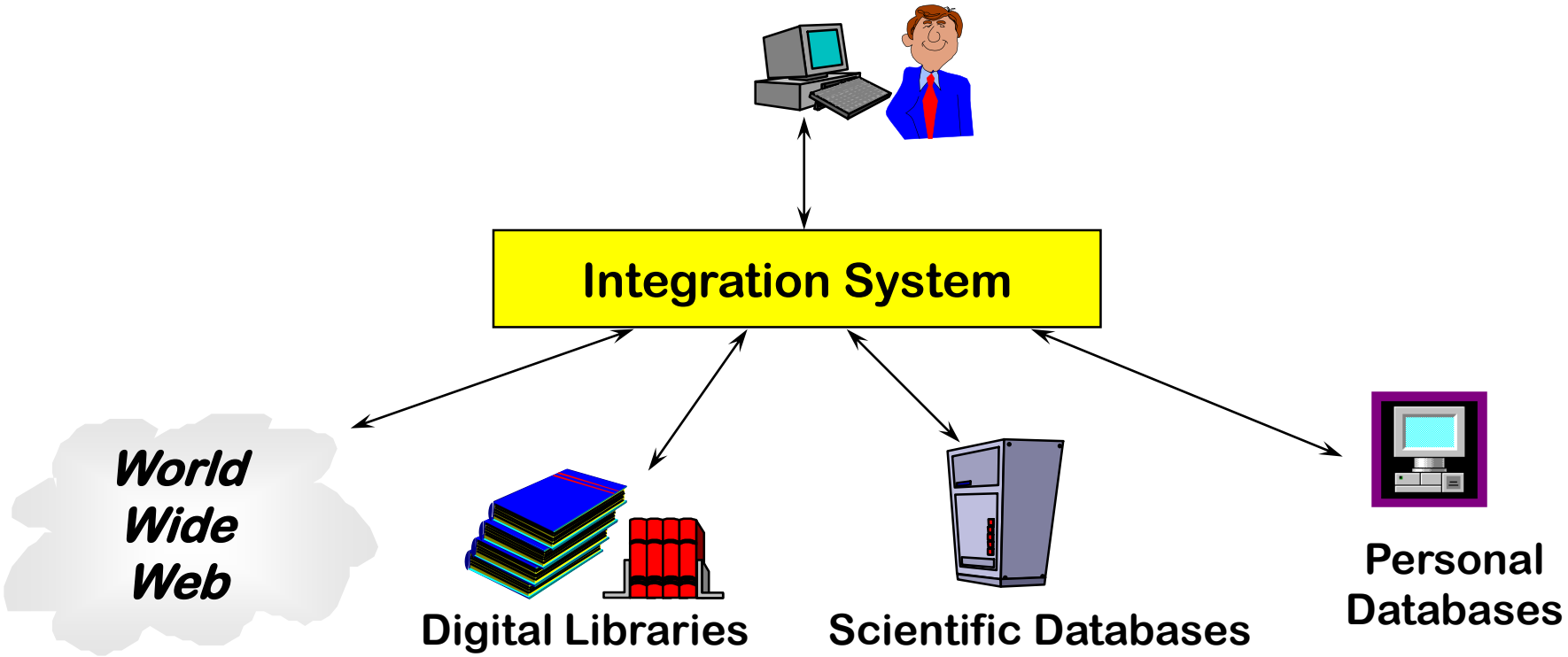
- Different interfaces
- Different data representations
- Duplicate and inconsistent information

Problem: Data Management in Large Enterprises

- Vertical fragmentation of informational systems (vertical stove pipes)
- Result of application (user)-driven development of operational systems



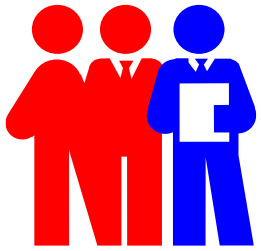
Goal: Unified Access to Data

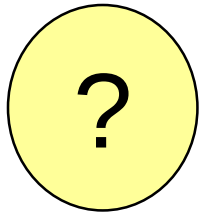


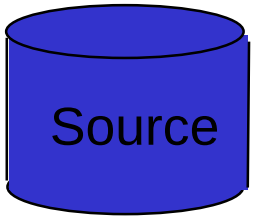
- Collects and combines information
- Provides integrated view, uniform user interface
- Supports sharing

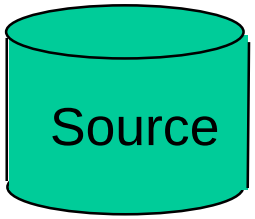
Why a Warehouse?

- Two Approaches:
  - Query-Driven (Lazy)
  - Warehouse (Eager)



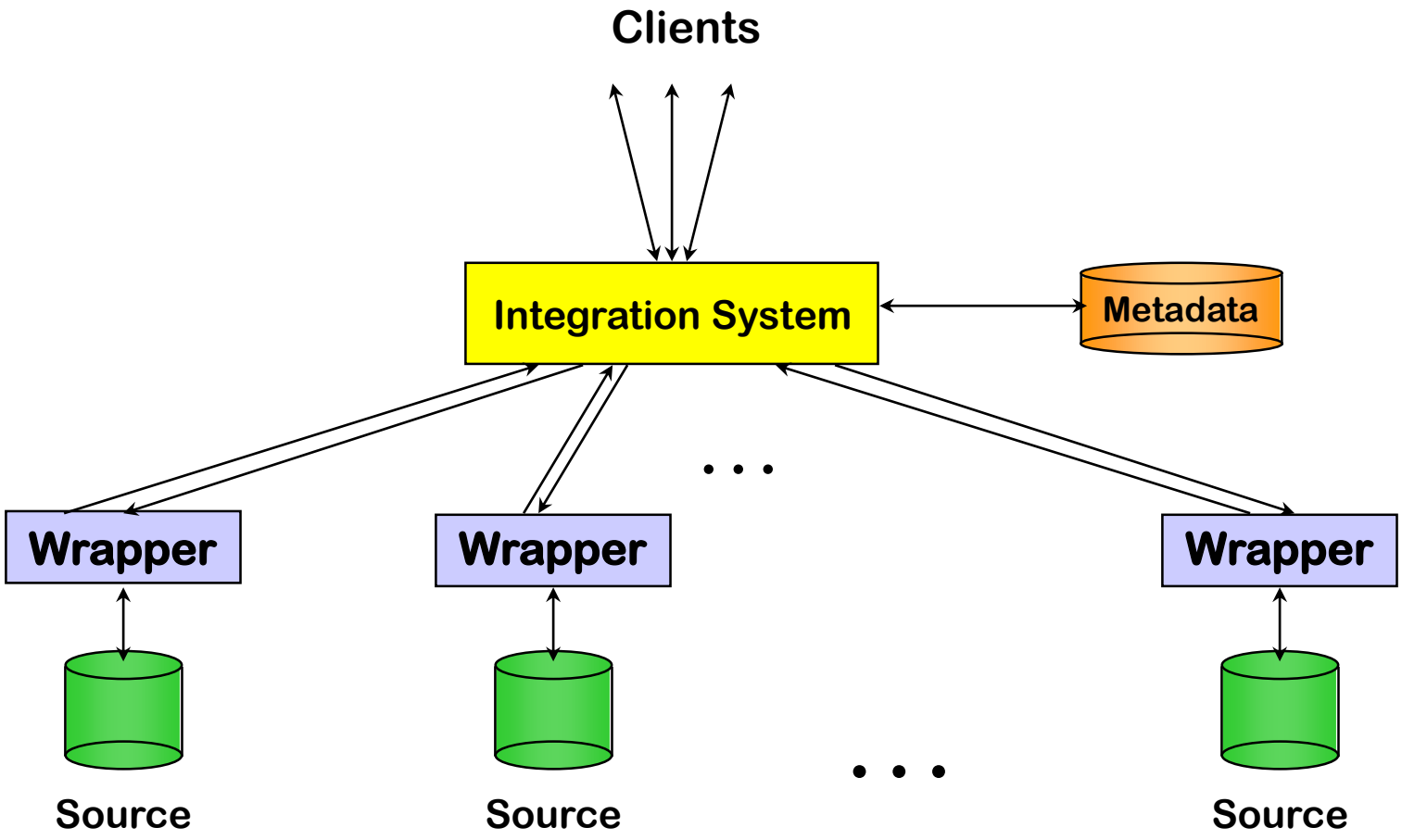






The Traditional Research Approach

- Query-driven (lazy, on-demand)

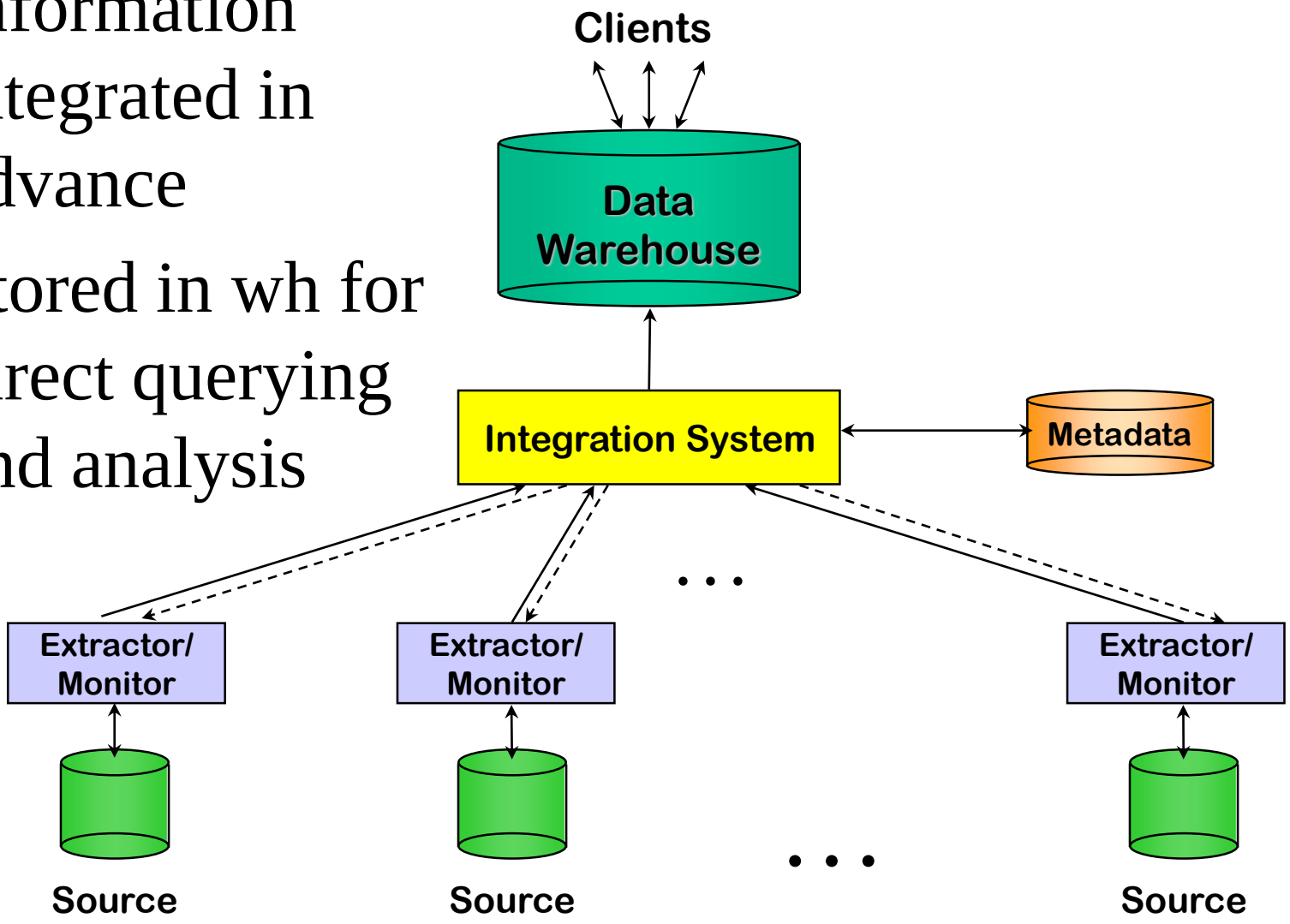


Disadvantages of Query-Driven Approach

- Delay in query processing
  - Slow or unavailable information sources
  - Complex filtering and integration
- Inefficient and potentially expensive for frequent queries
- Competes with local processing at sources
- *Hasn't* caught on in industry

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

Not Either-Or Decision

- Query-driven approach still better for
  - Rapidly changing information
  - Rapidly changing information sources
  - Truly vast amounts of data from large numbers of sources
  - Clients with unpredictable needs

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 servers

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

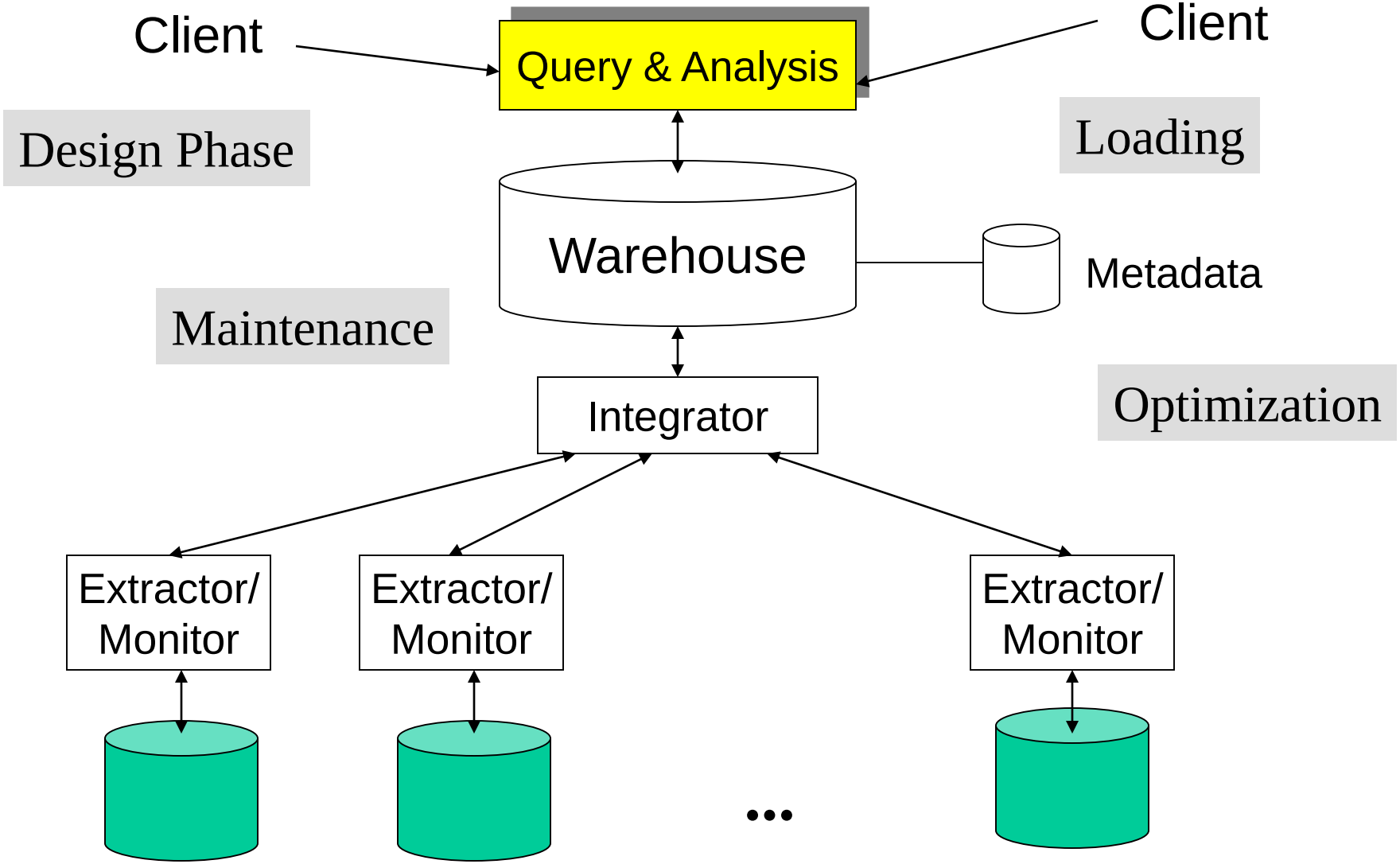
A Data Warehouse is...

- 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

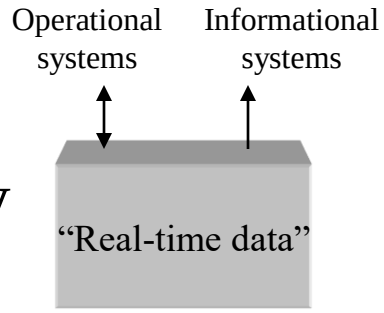
- Large volume of data (Gb, Tb)
- Non-volatile
  - Historical
  - Time attributes are important
- Updates infrequent
- May be append-only
- Examples
  - All transactions ever at Sainsbury's
  - Complete client histories at insurance firm
  - LSE financial information and portfolios

Generic Warehouse Architecture

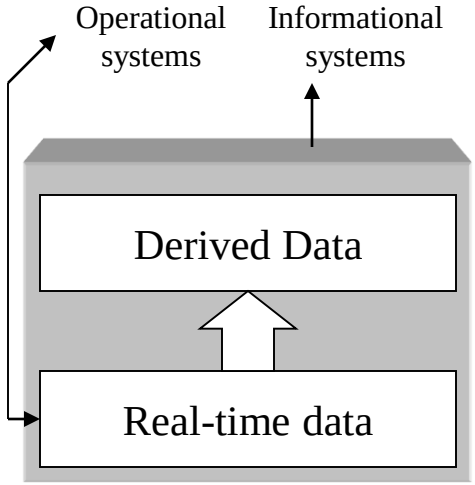


Data Warehouse Architectures: Conceptual View

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

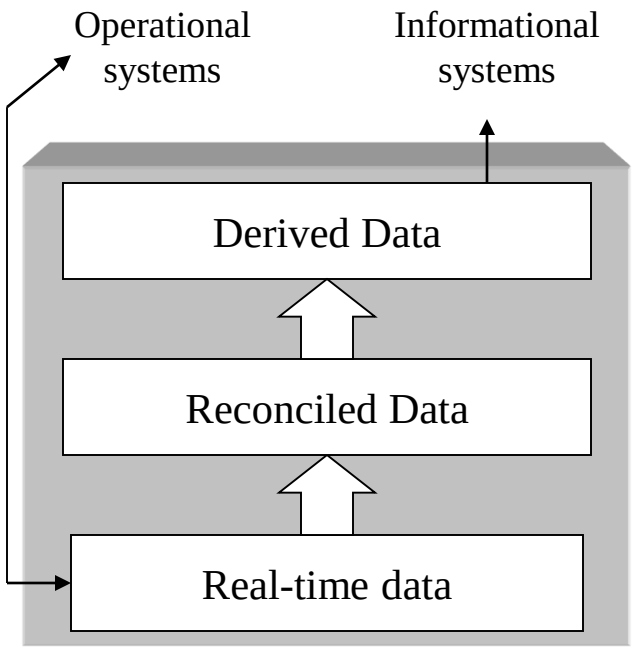


- Two-layer
  - Real-time + derived data
  - Most commonly used approach in industry today



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)

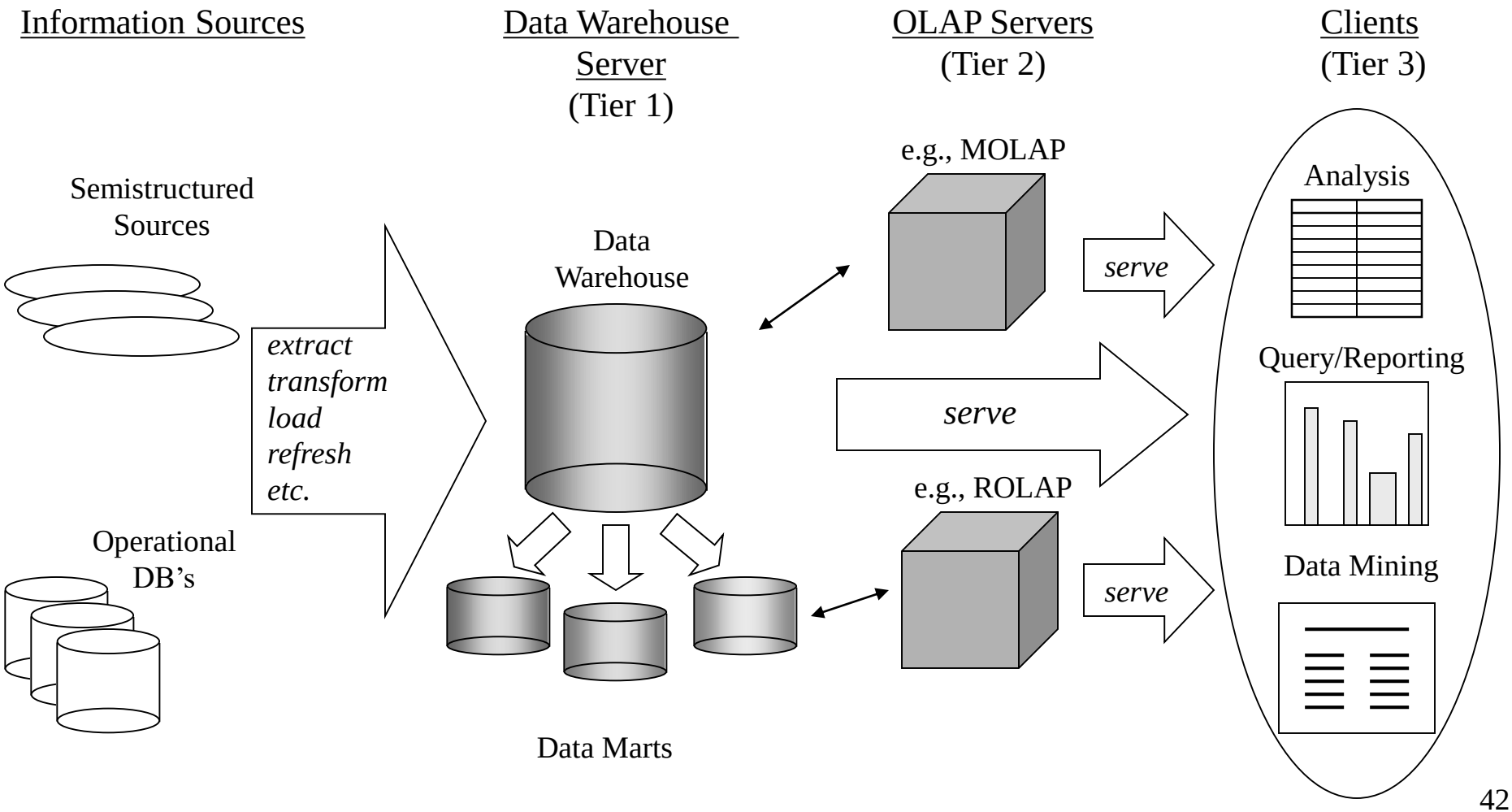
Decision Support

- 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



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
- Multidimensional view of data is the foundation of OLAP
  - 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