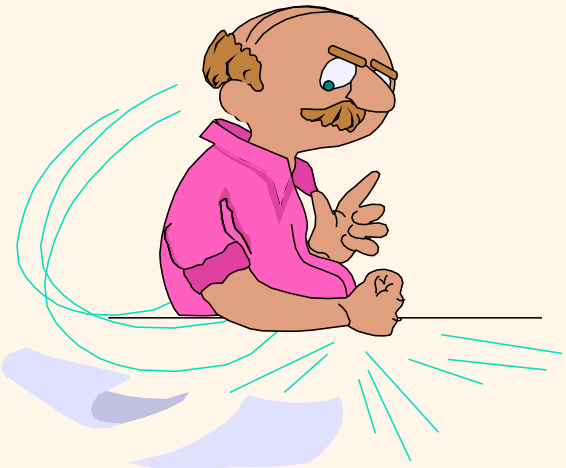
Data Warehousing/Mining

Data Warehousing Introduction

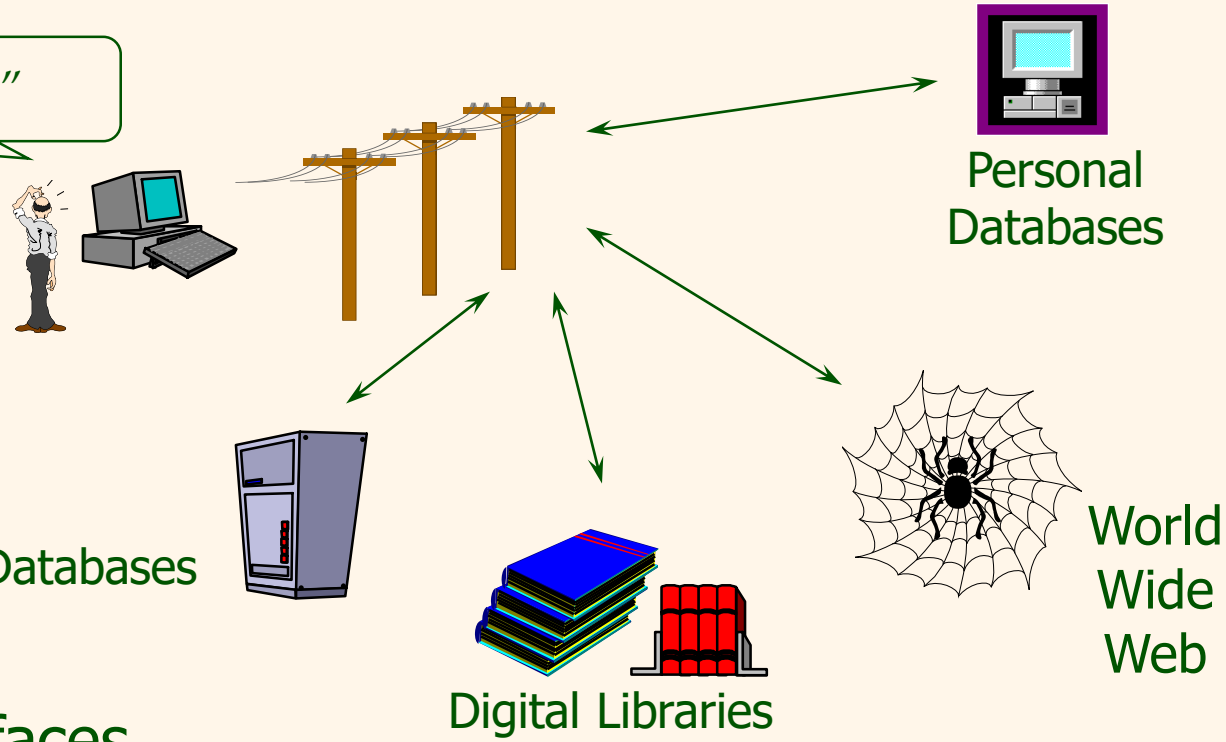


Outline of Lecture

- Data Warehousing and Information Integration
- Brief History of Data Warehousing
- What is a Data Warehouse?
- Types of Data and Their Uses
- Data Warehouse Architectures
- Issues in Data Warehousing

Problem: Heterogeneous Information Sources

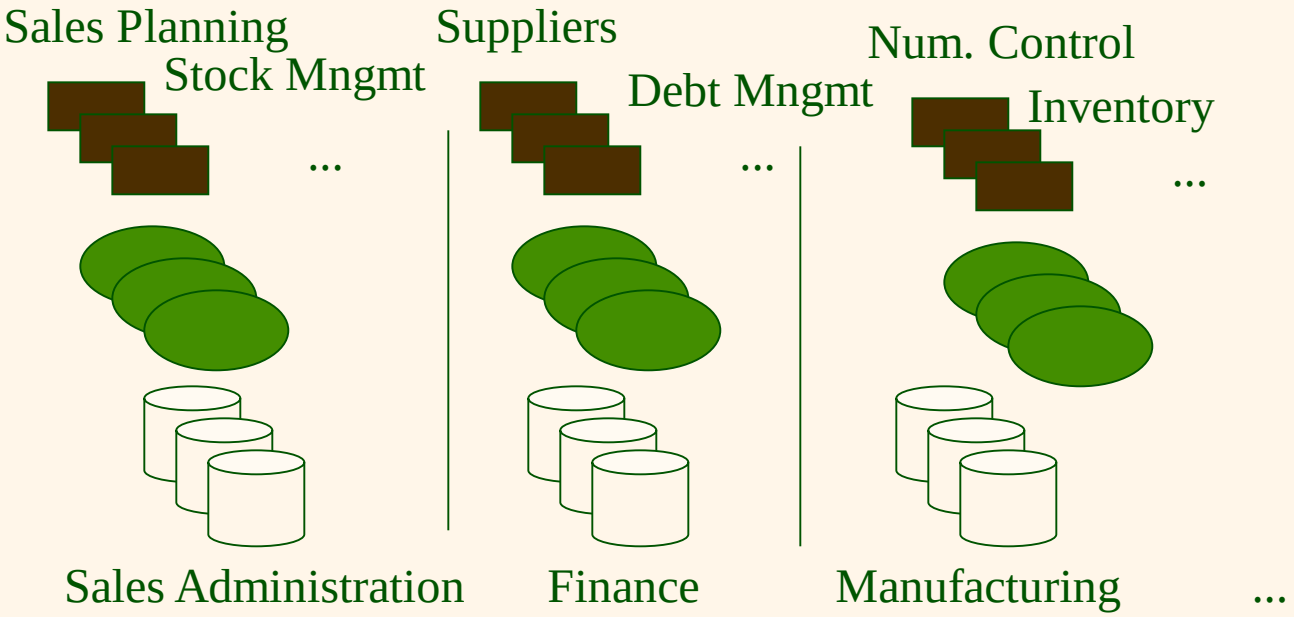
"Heterogeneities are everywhere"



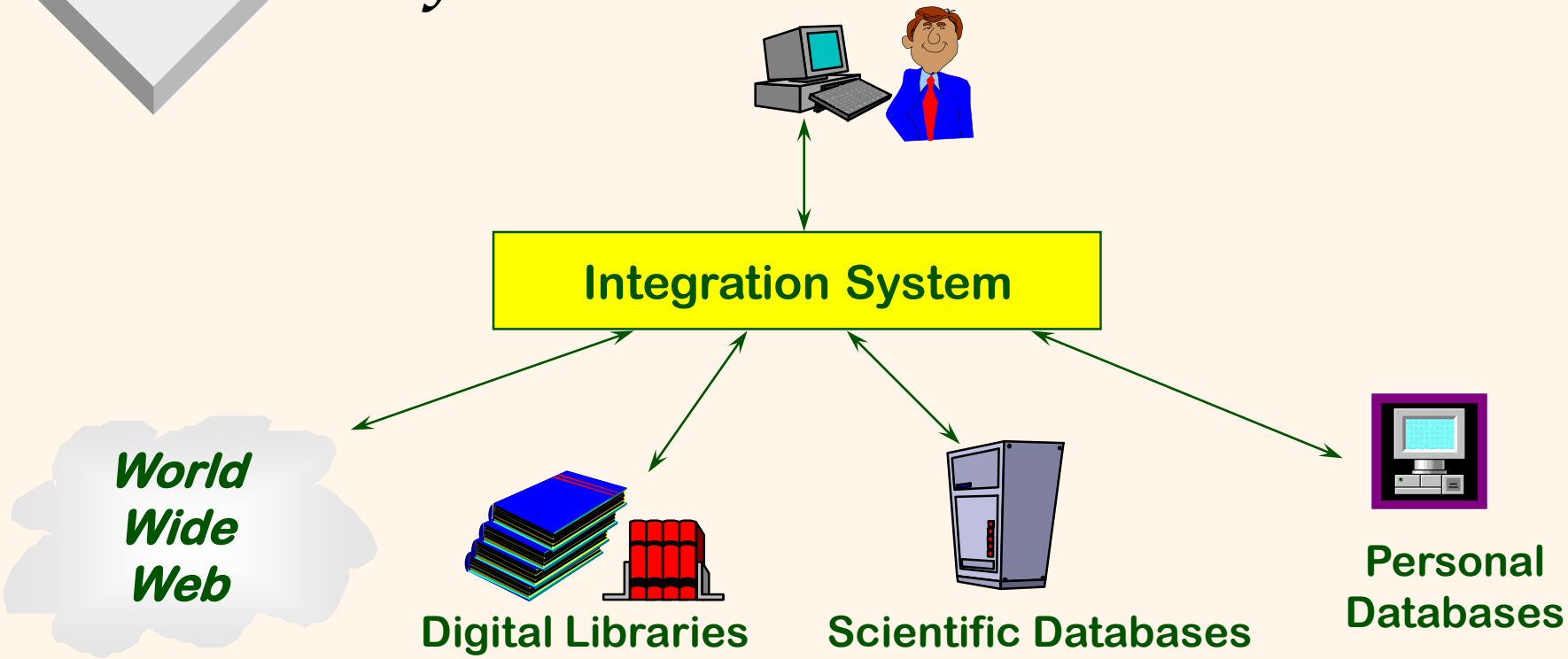
- Different interfaces
- Different data representations
- Duplicate and inconsistent information

Problem: Data Management in Large Enterprises

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



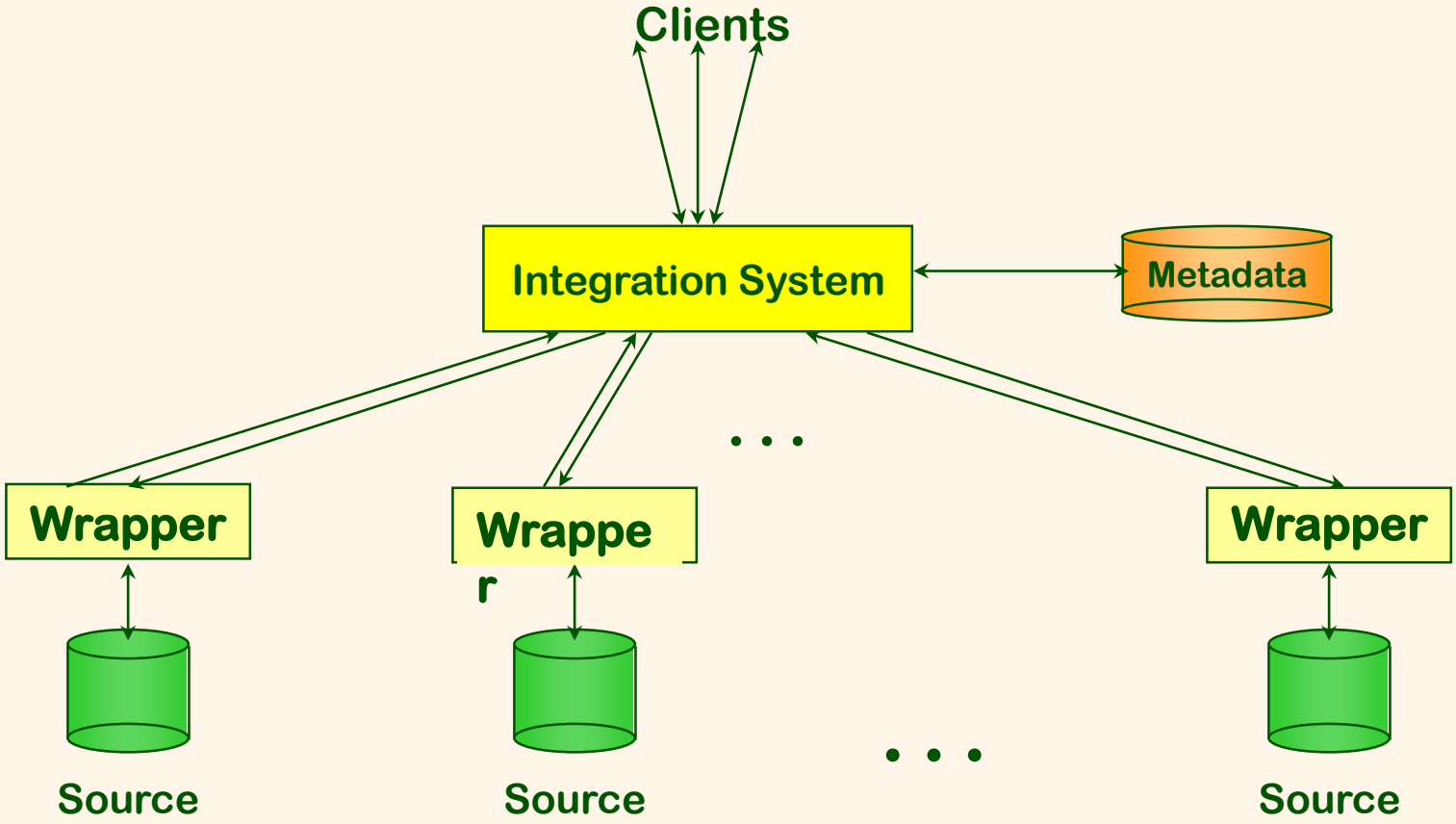
Goal: Unified Access to Data



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

The Traditional Research Approach

- Query-driven (lazy, on-demand)

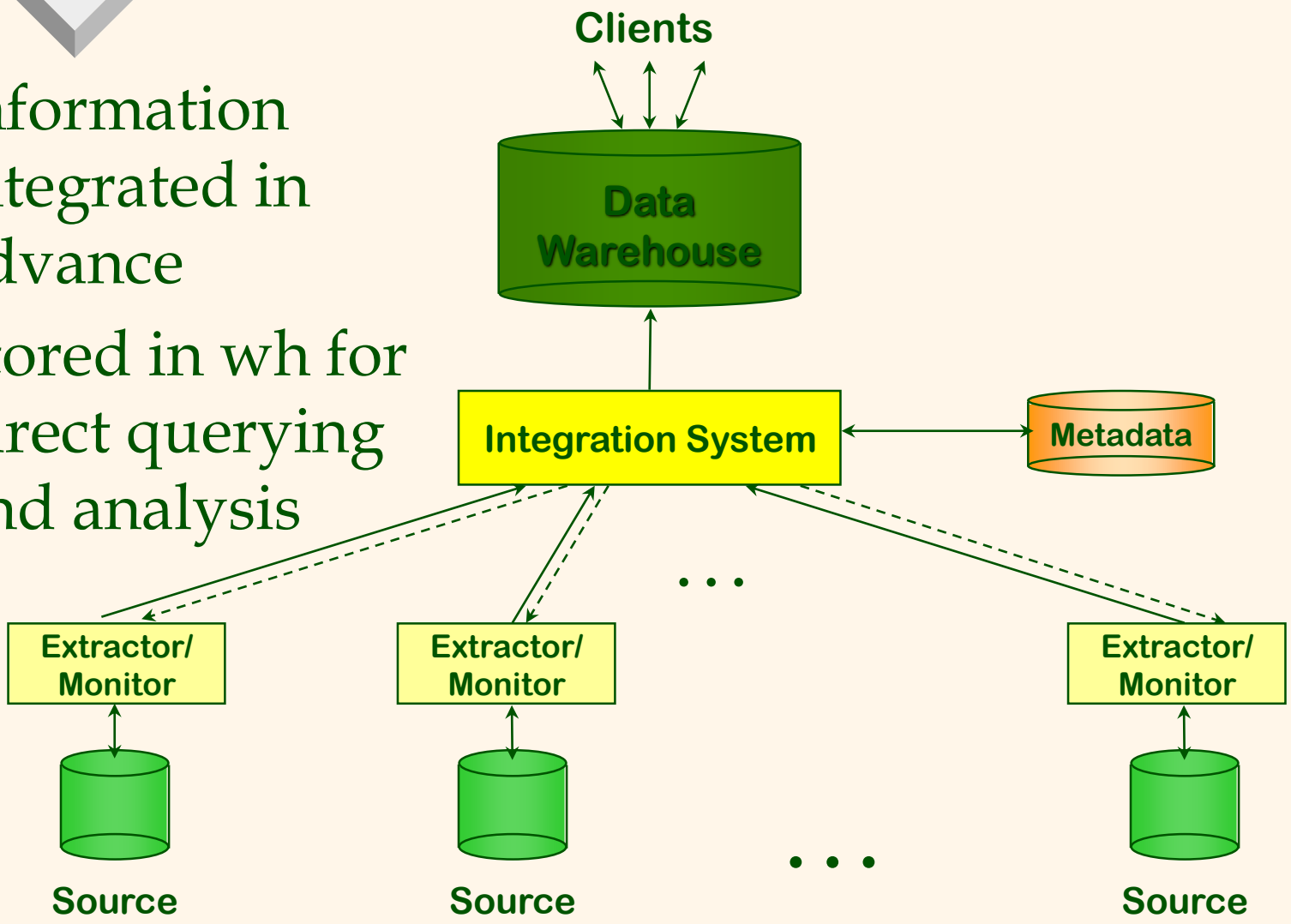


Disadvantages of Query-Driven Approach

- 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

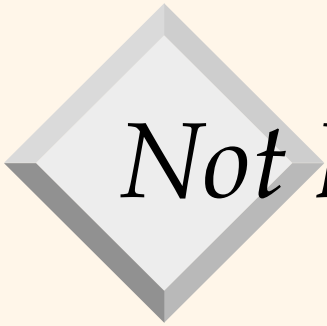
The Warehousing Approach

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



Advantages of Warehousing Approach

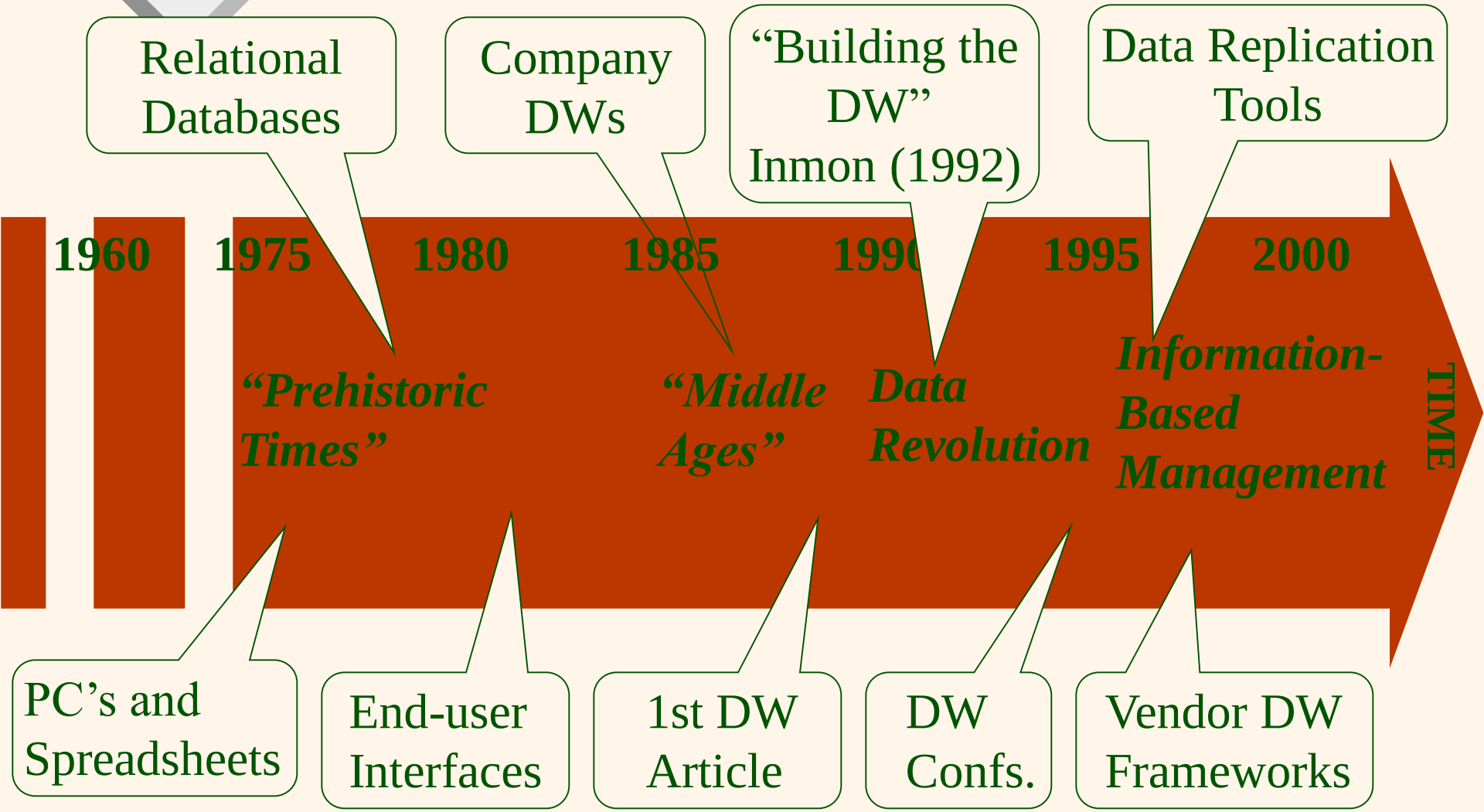
- High query performance
  - But not necessarily most current information
- Doesn't interfere with local processing at sources
  - Complex queries at warehouse
  - OLTP at information sources
- Information copied at warehouse
  - Can modify, annotate, summarize, restructure, etc.
  - Can store historical 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 Evolution



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*

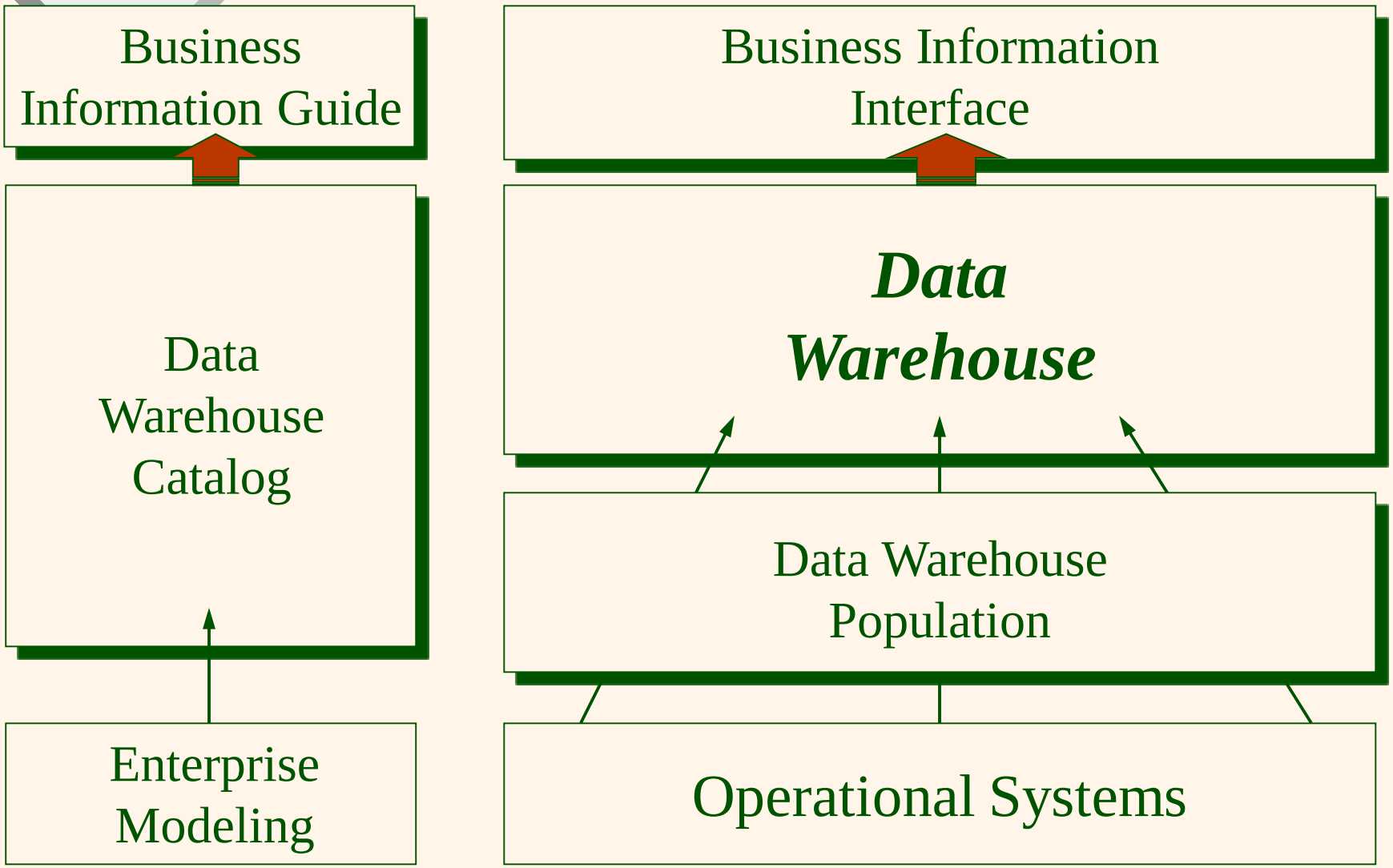
A Data Warehouse is...

- 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

A Data Warehouse is... (continued)

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

Summary


Business Information Guide
Business Information Interface
Data Warehouse Catalog
Data Warehouse
Data Warehouse Population
Enterprise Modeling
Operational Systems

Warehouse is a Specialized DB

Standard DB

- 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

- 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)

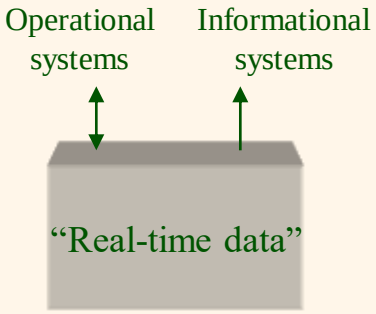
Warehousing and Industry

- Warehousing is big business
  - $2 billion in 1995
  - $3.5 billion in early 1997
  - Predicted: $8 billion in 1998 [Metagroup]
- WalMart has largest warehouse
  - 900-CPU, 2,700 disk, 23 TB Teradata system
  - ~7TB in warehouse
  - 40-50GB per day

Types of Data

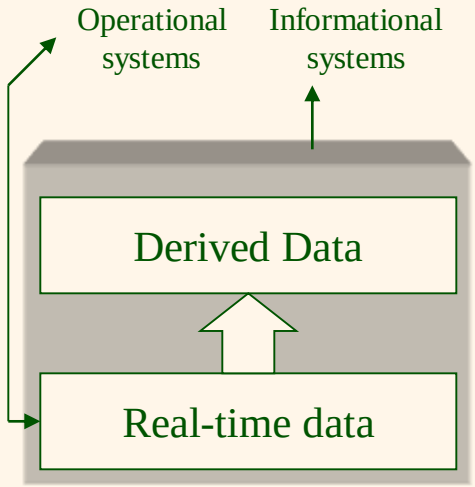
- Business Data - *represents meaning*
  - Real-time data (ultimate source of all business data)
  - Reconciled data
  - Derived data
- Metadata - *describes meaning*
  - Build-time metadata
  - Control metadata
  - Usage metadata
- Data as a product* - *intrinsic meaning*
  - Produced and stored for its own intrinsic value
  - e.g., the contents of a text-book

Data Warehouse Architectures: Conceptual View



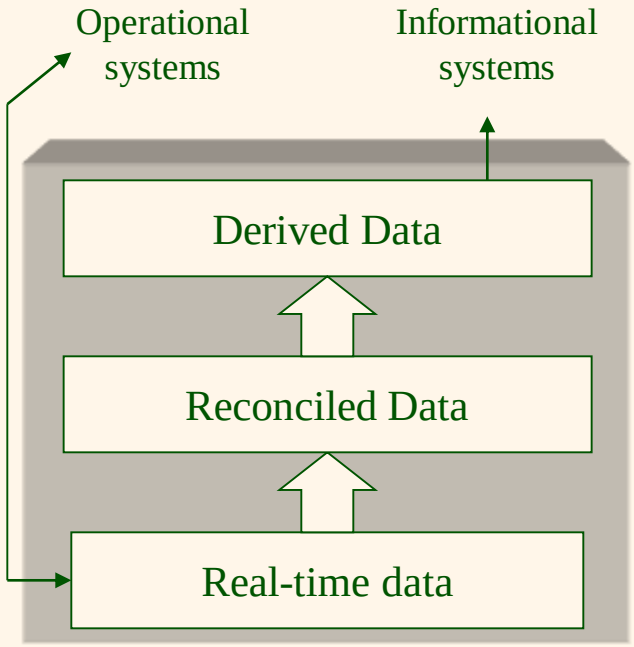
- Single-layer
  - Every data element is stored once only
  - Virtual warehouse

- 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"

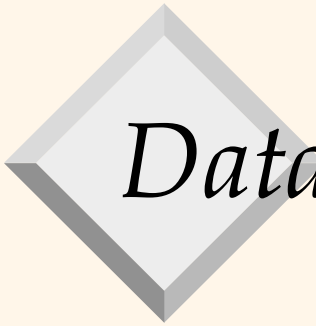
(2) What to do with data once it's in warehouse

"Warehouse DBMS"

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

Issues in Data Warehousing

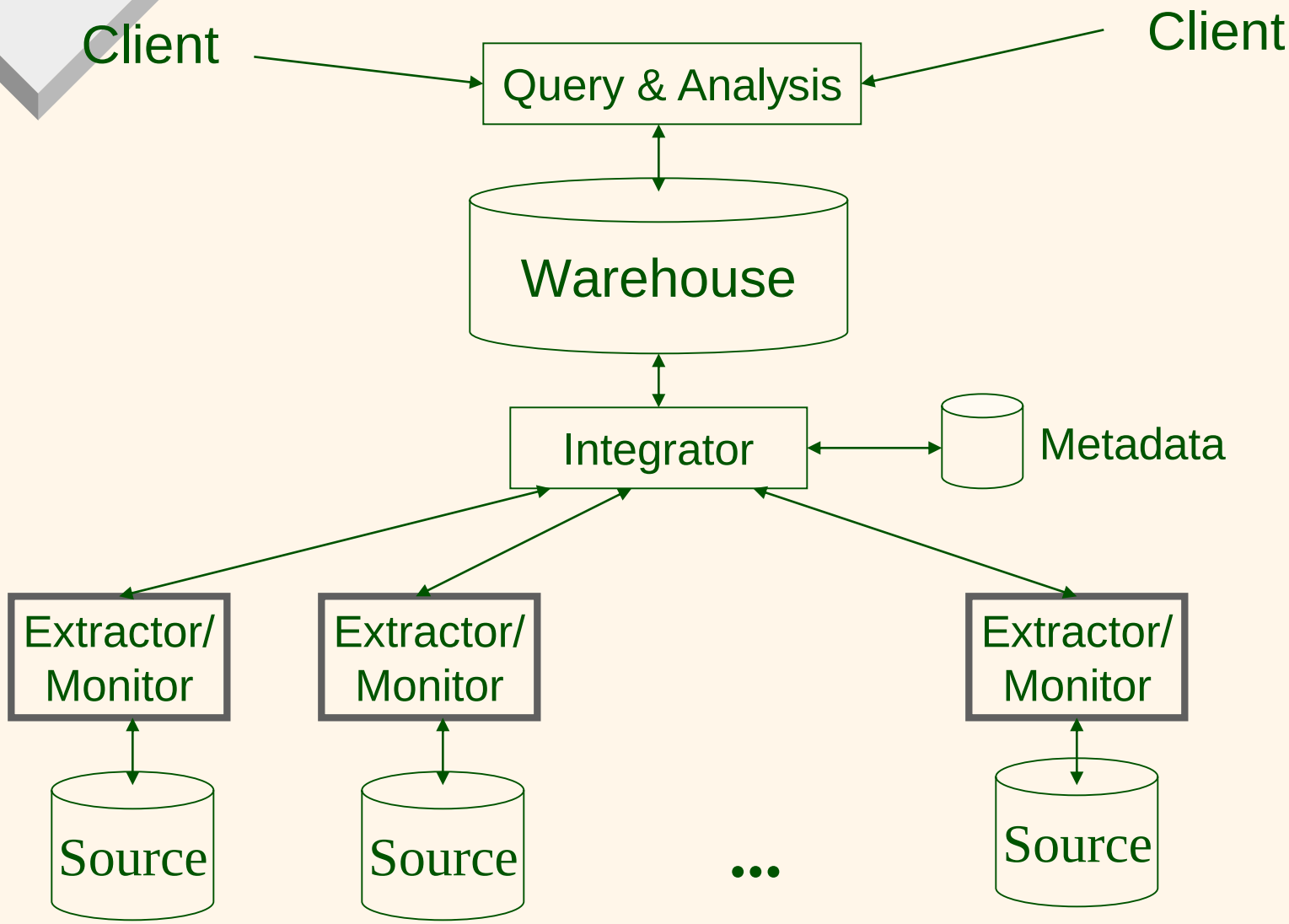
- Warehouse Design
- Extraction
  - Wrappers, monitors (change detectors)
- Integration
  - Cleansing & merging
- Warehousing specification & Maintenance
- Optimizations
- Miscellaneous (e.g., evolution)



Data Extraction

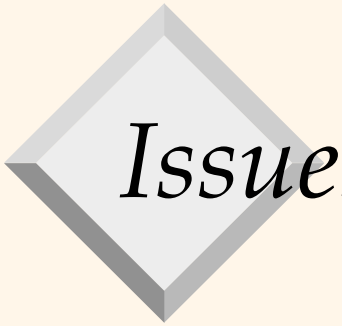
- Source types
  - Relational, flat file, WWW, etc.
- How to get data out?
  - Replication tool
  - Dump file
  - Create report
  - ODBC or third-party “wrappers”

Warehouse Architecture



Issues (1)

- Warehouse uses *relational data model* or *multi-dimensional data model* (e.g., data cube)
- On the other hand, source types
  - Relational, OO, hierarchical, legacy
  - Semistructured: flat file, WWW
- How do we get the data out?

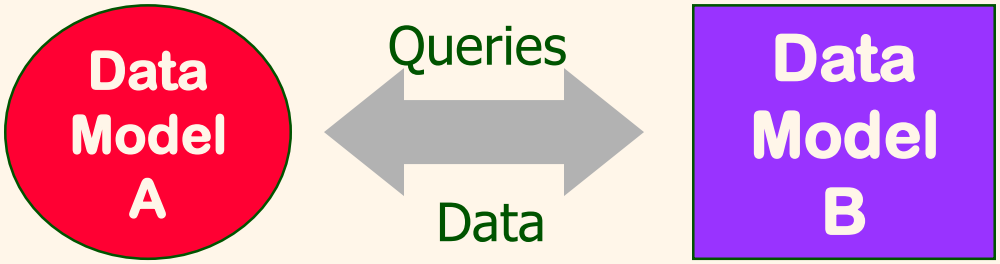


Issues (2)

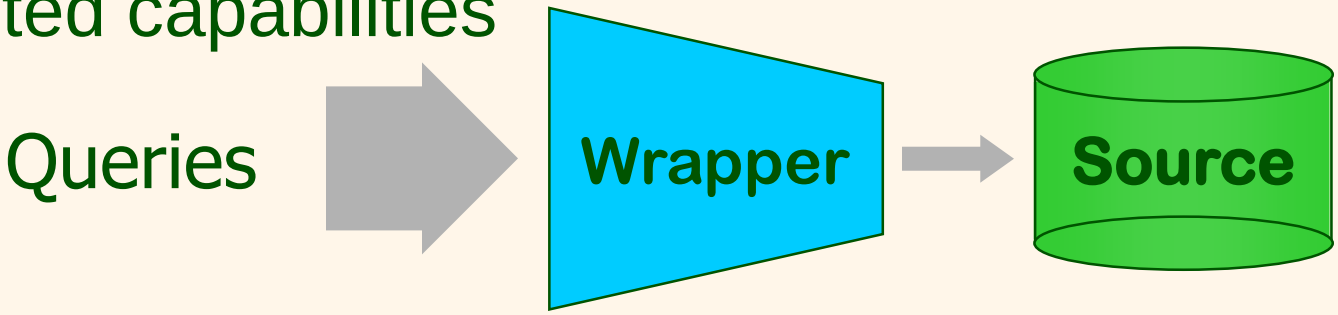
- Warehouse must be kept current in light of changes to underlying sources
- How do we detect updates in sources?

Wrapper

- Converts data and queries from one data model to another

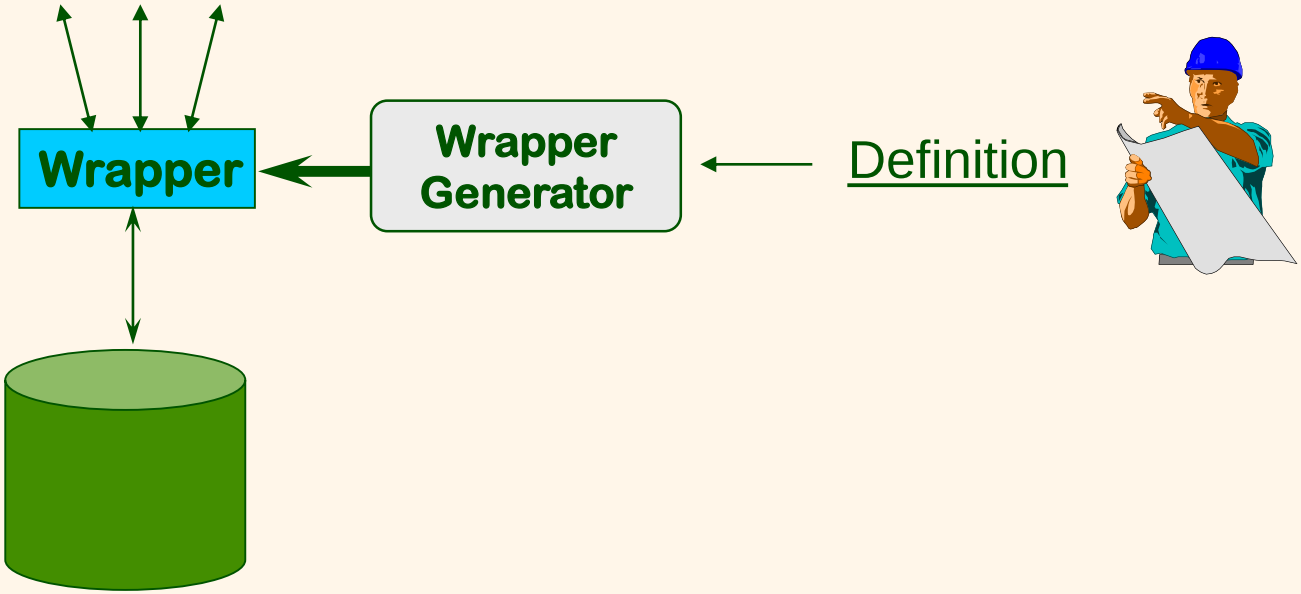


- Extends query capabilities for sources with limited capabilities



Wrapper Generation

- Solution 1: Hard code for each source
- Solution 2: Automatic wrapper generation



Wrapper Approach

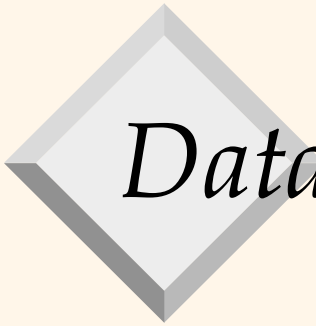
- *Source-specific adapter* (a.k.a. wrapper, translator)
- "Thickness" of adapter depends on source
  - Data model used (e.g. rel. schema vs. unstructured)
  - Interface (i.e., query language, API)
  - Active capabilities (i.e., triggers)
  - Degree of autonomy (e.g., same owner & modifiable vs. controlled by external entity & no changes possible)
  - Cooperation (e.g., friendly vs. uncooperative)

Routine When...

- Many tools for dealing with "standard situations"
  - Standard sources with full/many capabilities
    - e.g., most commercial DBMSs, all ODBC-compliant sources
  - Standard interactions
    - e.g., pass-through queries, extraction from rel. tables, replication
  - Cooperative sources or sources under our control
- Tools
  - Replication tools, ODBC, report writers, third-party "wrappers"

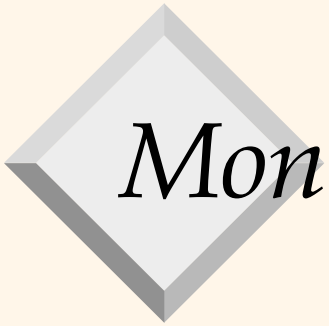
Not So Routine When...

- "Non-standard situations"
  - Unstructured or semistructured sources with little or no explicit schema
  - Uncooperative sources
  - Sources with limited capabilities (e.g., legacy sources, WWW)
- Few commercial tools
- Mostly research



Data Transformations

- Convert data to uniform format
  - Byte ordering, string termination
  - Internal layout
- Remove, add & reorder attributes
  - Add key
  - Add data to get history
- Sort tuples

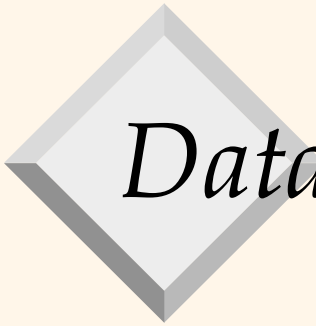


Monitors

- Goal: Detect changes of interest and propagate to integrator
- How?
  - Triggers
  - Replication server
  - Log sniffer
  - Compare query results
  - Compare snapshots/dumps

Data Integration

- Receive data (changes) from multiple wrappers/monitors and integrate into warehouse
- Rule-based
- Actions
  - Resolve inconsistencies
  - Eliminate duplicates
  - Integrate into warehouse (may not be empty)
  - Summarize data
  - Fetch more data from sources (wh updates)
  - etc.



Data Cleansing

- Find (& remove) duplicate tuples
  - e.g., Jane Doe vs. Jane Q. Doe
- Detect inconsistent, wrong data
  - Attribute values that don't match
- Patch missing, unreadable data
- Notify sources of errors found