Data Warehouse and OLAP

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

Operational Data Vs. DSS

- Operational data and decision support data serve different purposes.
- Most operational data are stored in a highly normalized relational database.
- Operational data storage is optimized to support transactions that represent daily operations.

For example, each time an item is sold, it must be accounted for, Customer data, inventory data, and so on, are in a frequent update mode.

To provide effective update performance, operational systems store data in many tables, each with a minimum number of fields.

Decision Support Data

- From the data analyst's point of view, decision support data differ from operational data in three main areas:

time span, granularity, and dimensionality.

- ***Time span***

Operational data cover a short time frame. In contrast, decision support data tend to cover a longer time frame.

Managers are seldom interested in a specific sales invoice to customer X; rather, they tend to focus on sales generated during the last month, the last year, or the last five years.

Granularity(level of aggregation)

Decision support data must be presented at **different levels of aggregation**, from highly summarized to near-atomic.

For example, if managers must analyse sales by region, they must be able to access data showing the sales by region, by city within the region, by store within the city within the region, and so on.

In that case, summarized data to compare the regions is required, and also data in a structure that enables a manager to drill down, or decompose, the data into more atomic components (that is, finer-grained data at lower levels of aggregation).

In contrast, when you roll up the data, you are aggregating the data to a higher level.

Dimensionality:

- Operational data focus on representing individual transactions rather than on the effects of the transactions over time.
- In contrast, data analysts tend to include many data dimensions and are interested in how the data relate over those dimensions.

For example, an analyst might want to know how product X fared relative to product Z during the past six months by region, state, city, store, and customer.

In that case, both place and time are part of the picture.

From the designer's point of view, the differences between operational and decision support data are as follows:

- Time: Operational data represent transactions as they happen in real time. Decision support data are a snapshot of the operational data at a given point in time. Therefore, decision support data are historic, representing a time slice of the operational data.
- Type and volume of transaction: Operational and decision support data are different in terms of transaction type and transaction volume. Whereas operational data are characterized by update transactions, decision support data are mainly characterized by query (read-only) transactions.

Decision support data also require periodic updates to load new data that are summarized from the operational data.

Finally, the concurrent transaction volume in operational data tends to be very high when compared with the low-to-medium levels found in decision support data.

- Operational data are commonly stored in many tables, and the stored data represent the information about a given transaction only.

Decision support data are generally stored in a few tables that store data derived from the operational data.

The decision support data do not include the details of each operational transaction.

Instead, **decision support data represent transaction summaries**; therefore, the decision support database stores data that are integrated, aggregated, and summarized for decision support purposes.

The degree to which decision support data are summarized is very high when contrasted with operational data. Therefore, you will see a great deal of derived data in decision support databases.

For example, rather than storing all 10,000 sales transactions for a given store on a given day, the **decision support database might simply store the total number of units sold and the total sales dollars generated during that day**. Decision support data might be collected to monitor such aggregates as total sales for each store or for each product. The purpose of the summaries is simple: they are to be used to establish and evaluate sales trends, product sales comparisons, and so on, that serve decision needs.

- The data models that govern operational data and decision support data are different.
- The operational database's frequent and rapid data updates make data anomalies a potentially devastating problem.
- Therefore, the data requirements in a typical relational transaction (operational) system generally require normalized structures that yield many tables, each of which contains the minimum number of attributes.

In contrast, the decision support database is not subject to such transaction updates, and the focus is on querying capability. Therefore, decision support databases tend to be non-normalized and include few tables, each of which contains a large number of attributes.

- Query activity (frequency and complexity) in the operational database tends to be low to allow additional processing cycles for the more crucial update transactions.
- Therefore, queries against operational data typically are narrow in scope, low in complexity, and speed-critical.
- In contrast, decision support data exist for the sole purpose of serving query requirements. Queries against decision support data typically are broad in scope, high in complexity, and less speed-critical.

- Finally, decision support data are characterized by very large amounts of data. The large data volume is the result of two factors.
- First, data are stored in non-normalized structures that are likely to display many data redundancies and duplications.
- Second, the same data can be categorized in many different ways to represent different snapshots.
- For example, sales data might be stored in relation to product, store, customer, region, and manager.

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*

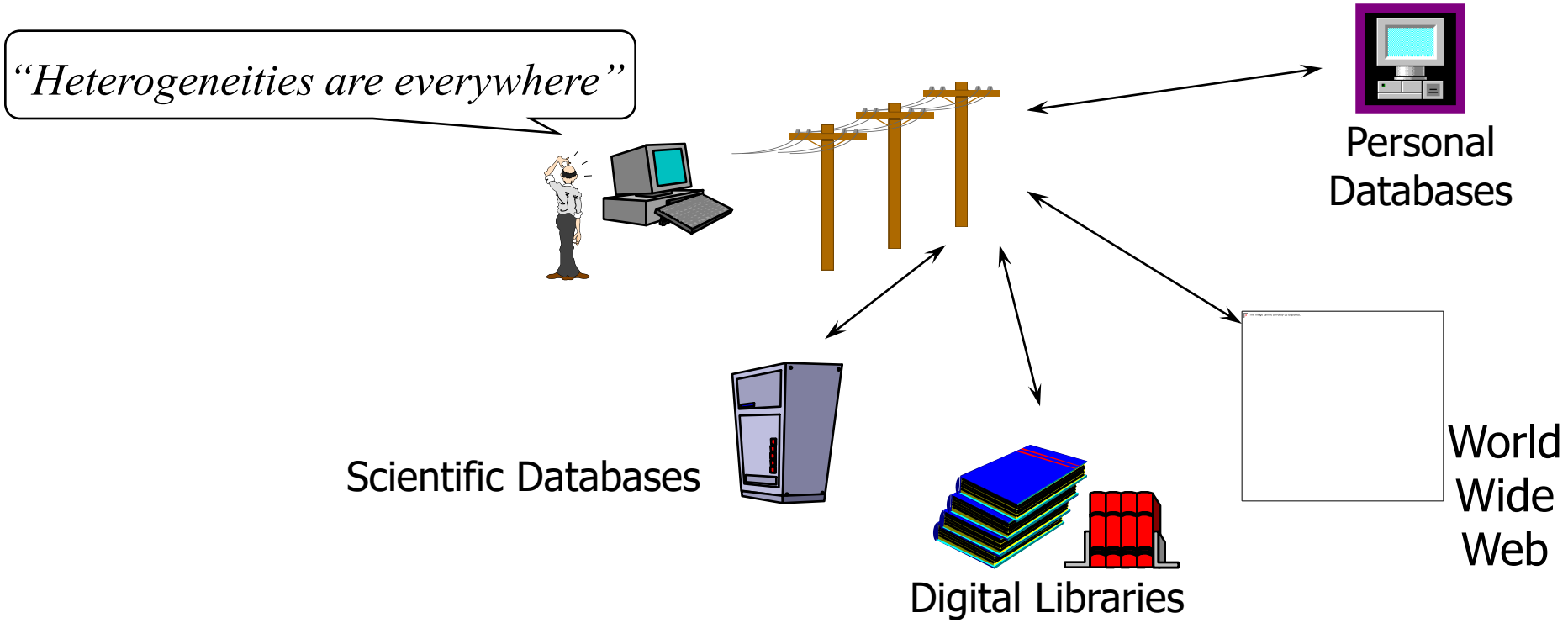
Staff

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

Problems

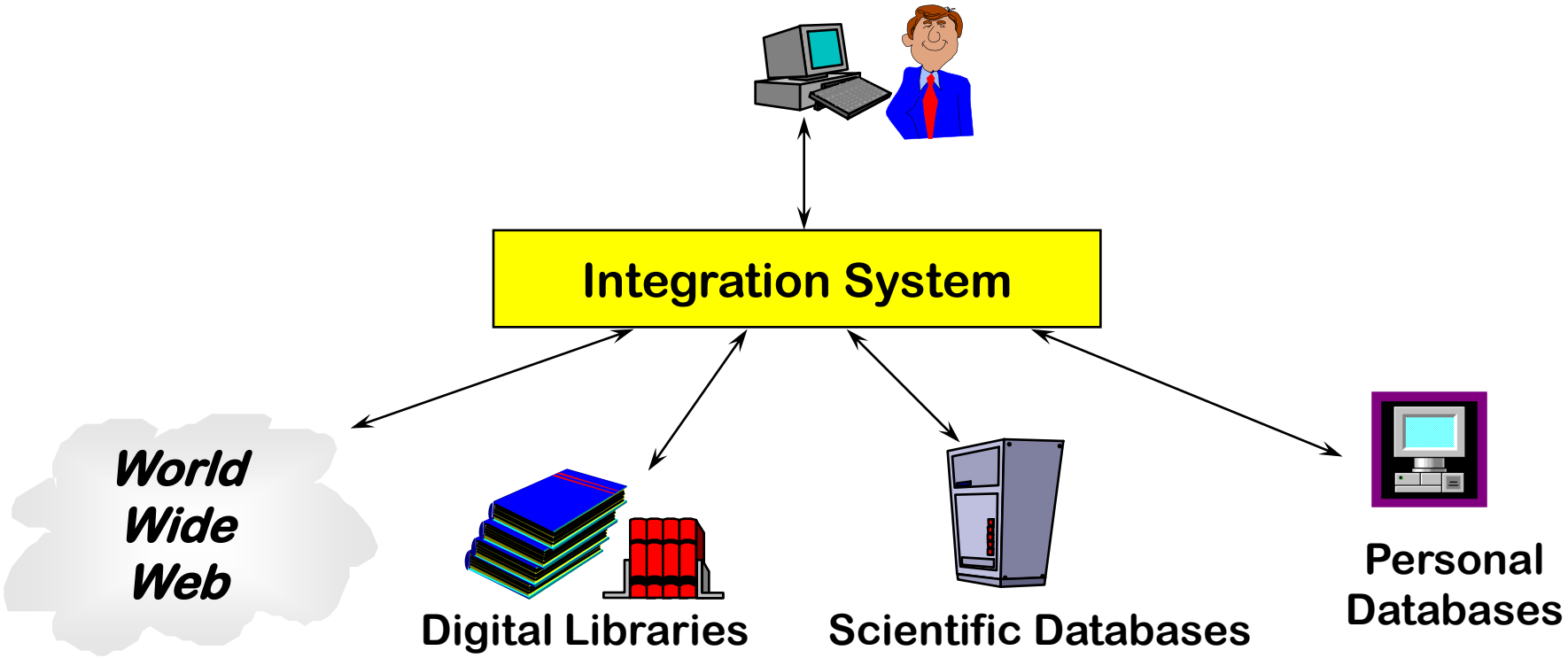
- A fully normalized data model can perform very inefficiently for queries.
- Historical data are usually large with static relationships:
  - Unnecessary joins may take unacceptably long time
- Historical data are diverse

Problem: Heterogeneous Information Sources



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

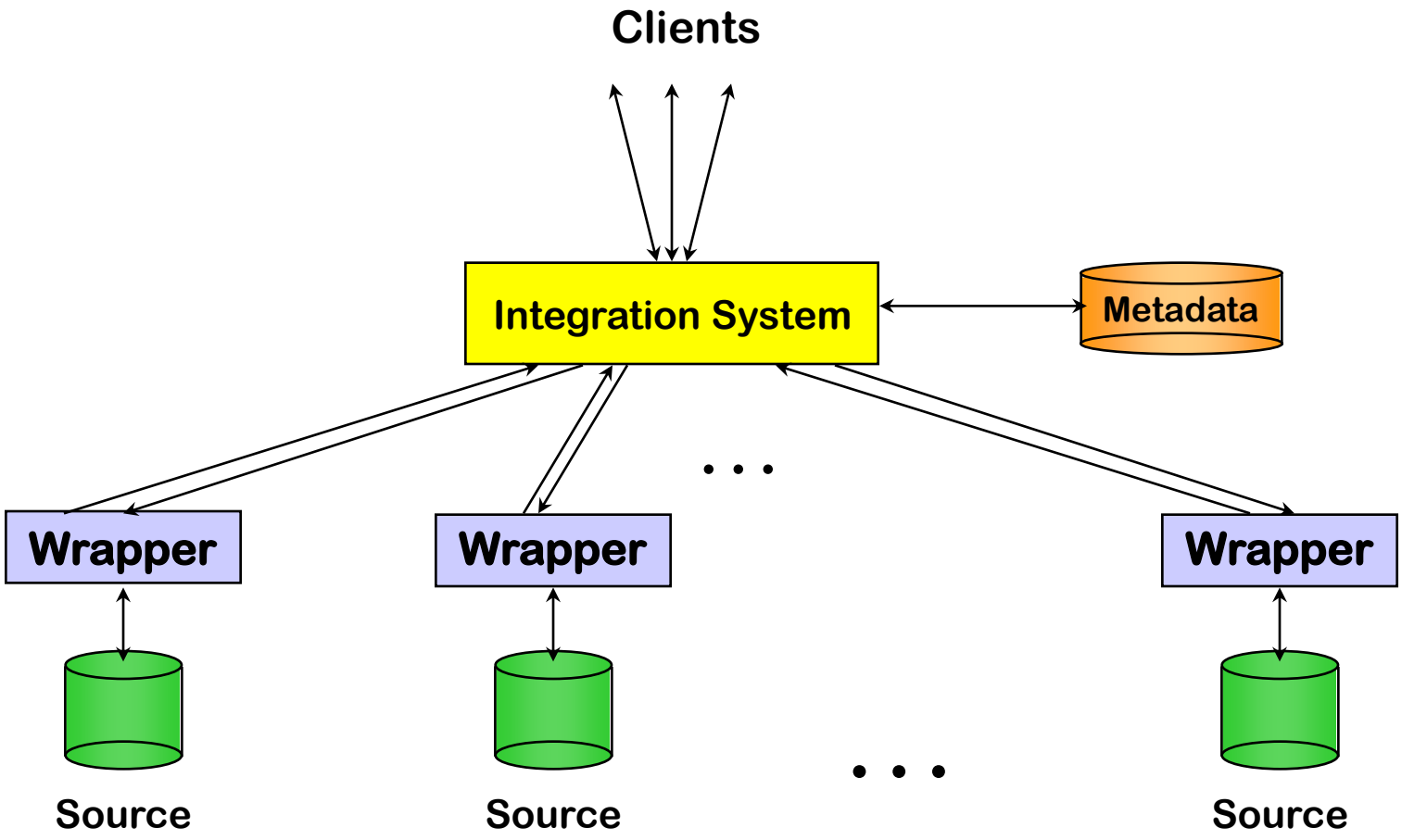
Goal: Unified Access to Data



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

The Traditional Research Approach

- Query-driven (lazy, on-demand)

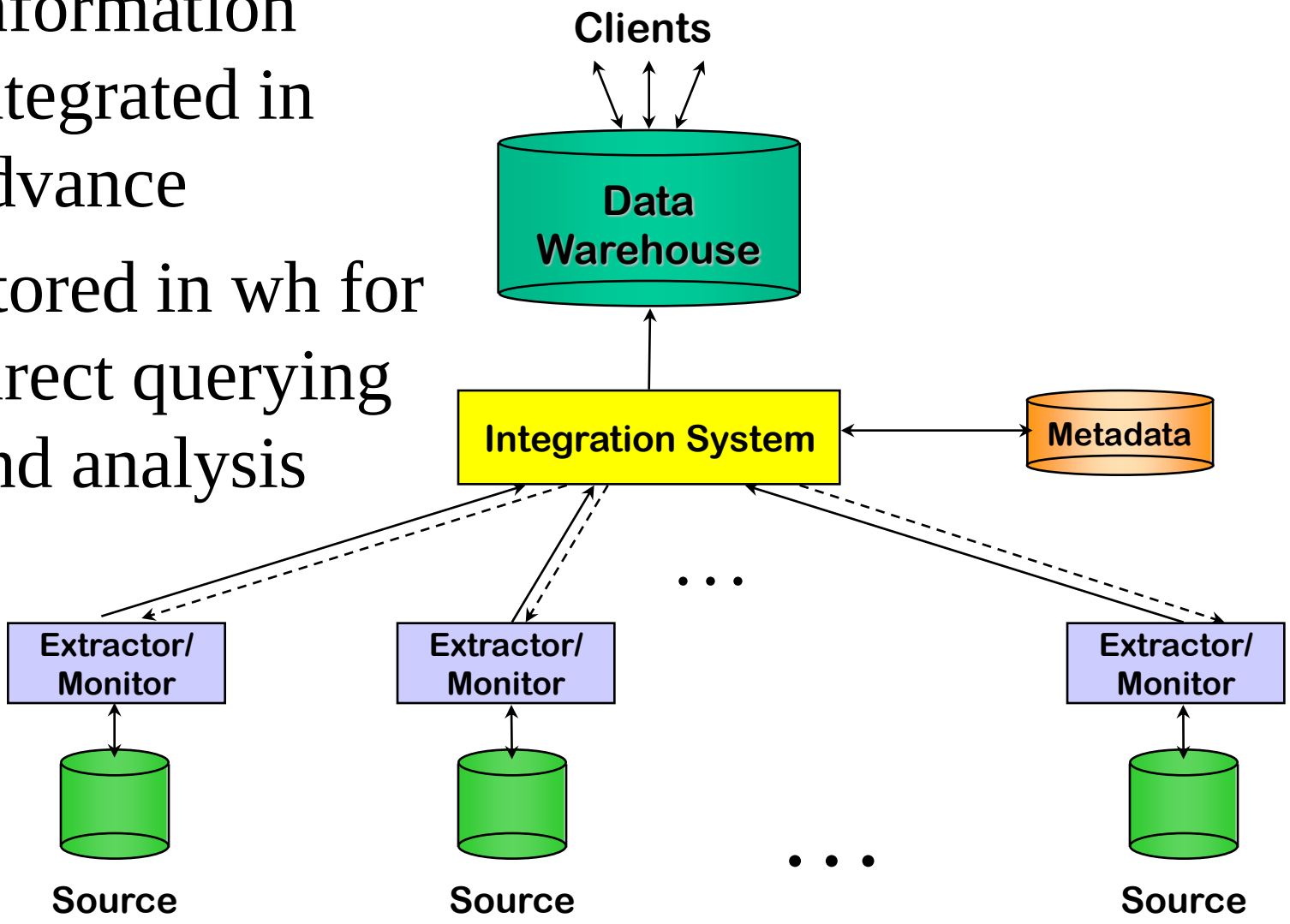


Disadvantages of Query-Driven Approach

- 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

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

- **Key Differences:**

1. **Data Types:**
    1. **Relational Databases:** Store transactional data with a focus on consistency.
    2. **Data Warehouses:** Store historical and aggregated data for analytical processing.
2. **Optimization:**
    1. **Relational Databases:** Optimize for transactional operations (OLTP).
    2. **Data Warehouses:** Optimize for analytical processing (OLAP).
3. **Schema Design:**
    1. **Relational Databases:** Use a normalized schema to minimize redundancy.
    2. **Data Warehouses:** Use a star or snowflake schema for efficient analytical queries.
4. **Performance Trade-Offs:**
    1. **Relational Databases:** Prioritize fast write operations and consistency.
    2. **Data Warehouses:** Prioritize fast read operations and analytical queries.

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

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

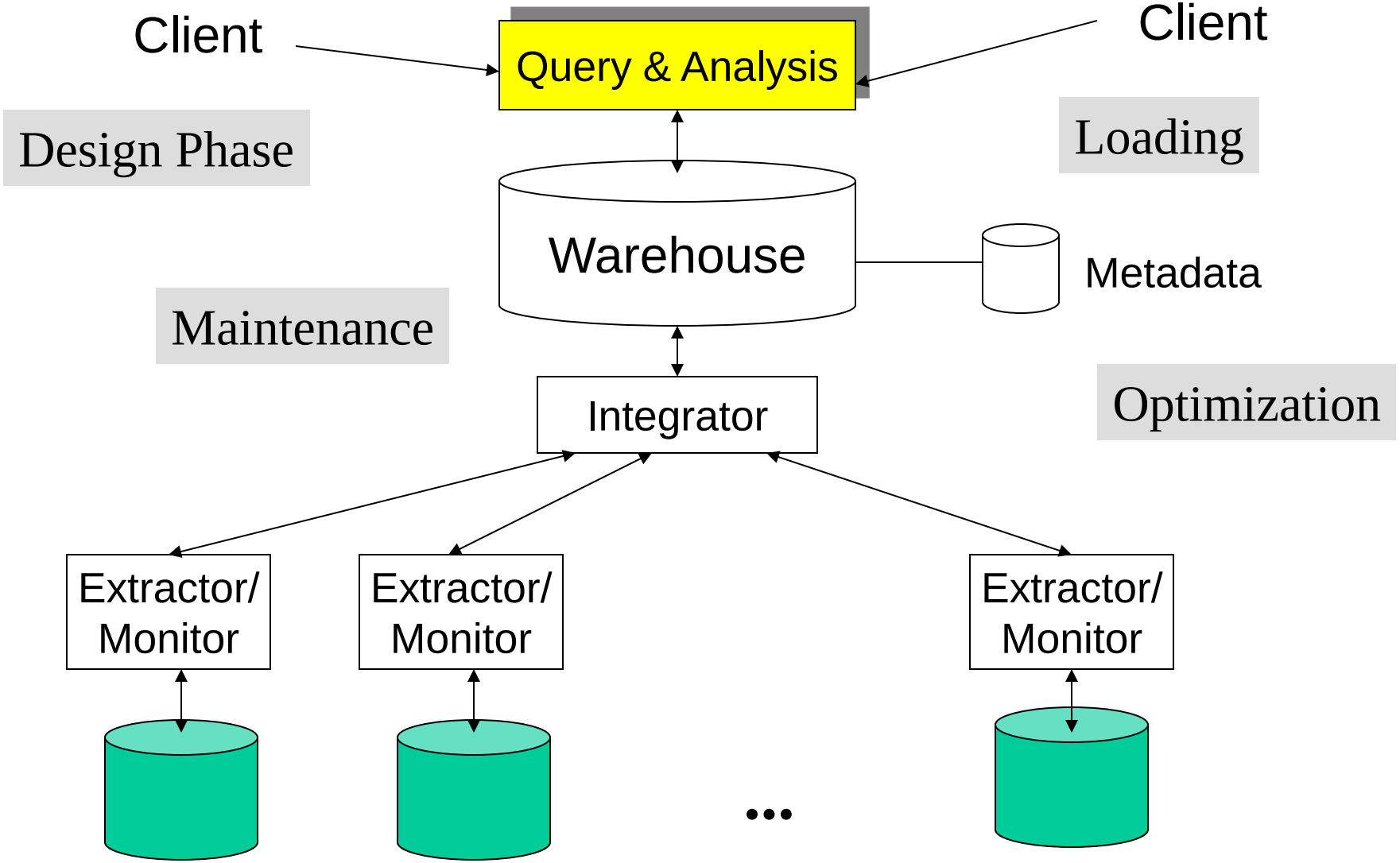
- transaction db is organized to collect the incoming data and data warehouse is not

- User interface aimed at executive
- A unified data model across all subjects

… Cont'd

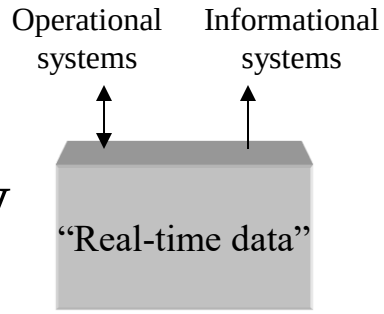
- Large volume of data (Gb, Tb)
- Flexible and scalable
- Non-volatile
  - Historical
  - Time attributes are important
- Updates infrequent
- May be append-only
- Examples
  - All transactions ever at Sainsbury's
  - Complete client histories at insurance firm
  - financial information and portfolios

Generic Warehouse Architecture

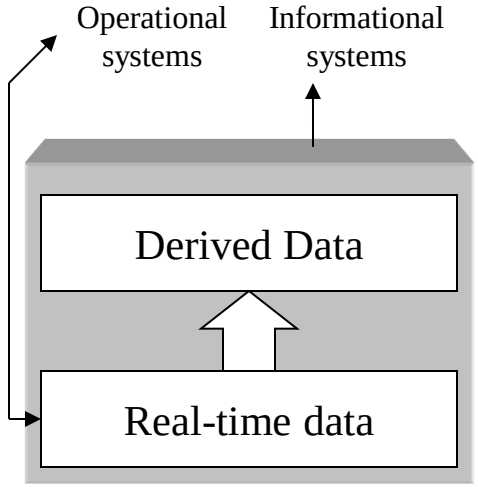


Data Warehouse Architectures: Conceptual View

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

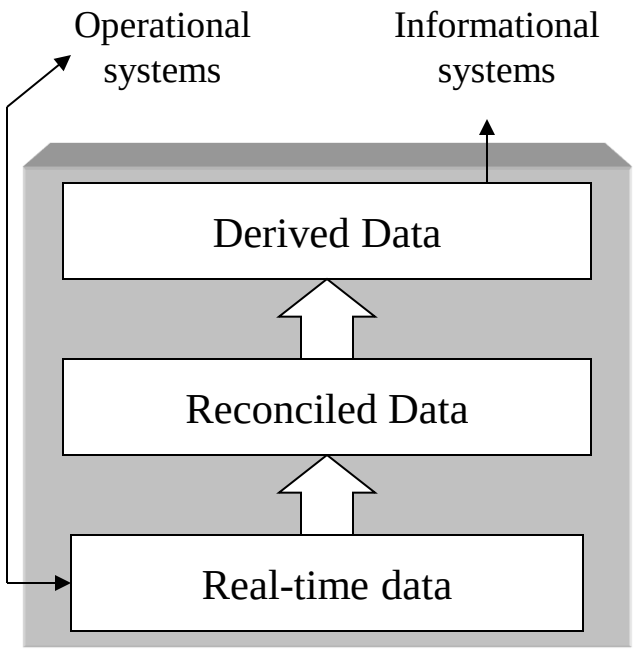


- Two-layer
  - Real-time + derived data
  - Most commonly used approach in industry today



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(for scraping), 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)	Warehouse (OLAP)
• Mostly updates	• Mostly reads
• Many small transactions	• Queries are long and complex
• Mb - Gb of data	• Gb - Tb of data
• Current snapshot	• History
• Index/hash on p.k.	• Lots of scans
• Raw data	• Summarized, reconciled data
• Thousands of users (e.g., clerical users)	• Hundreds of users (e.g., decision-makers, analysts)

.No.	Data Warehouse (OLAP)	Operational Database(OLTP)
1	It involves historical processing of information.	It involves day-to-day processing.
2	OLAP systems are used by knowledge workers such as executives, managers, and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
3	It is used to analyze the business.	It is used to run the business.
4	It focuses on Information out.	It focuses on Data in.
5	It is based on Star Schema, Snowflake Schema, and Fact Constellation Schema.	It is based on Entity Relationship Model.
6	It focuses on Information out.	It is application oriented.
7	It contains historical data.	It contains current data.
8	It provides summarized and consolidated data.	It provides primitive and highly detailed data.
9	It provides summarized and multidimensional view of data.	It provides detailed and flat relational view of data.
10	The number of users is in hundreds.	The number of users is in thousands.
11	The number of records accessed is in millions.	The number of records accessed is in tens.
12	The database size is from 100GB to 100 TB.	The database size is from 100 MB to 100 GB.
13	These are highly flexible.	It provides high performance.

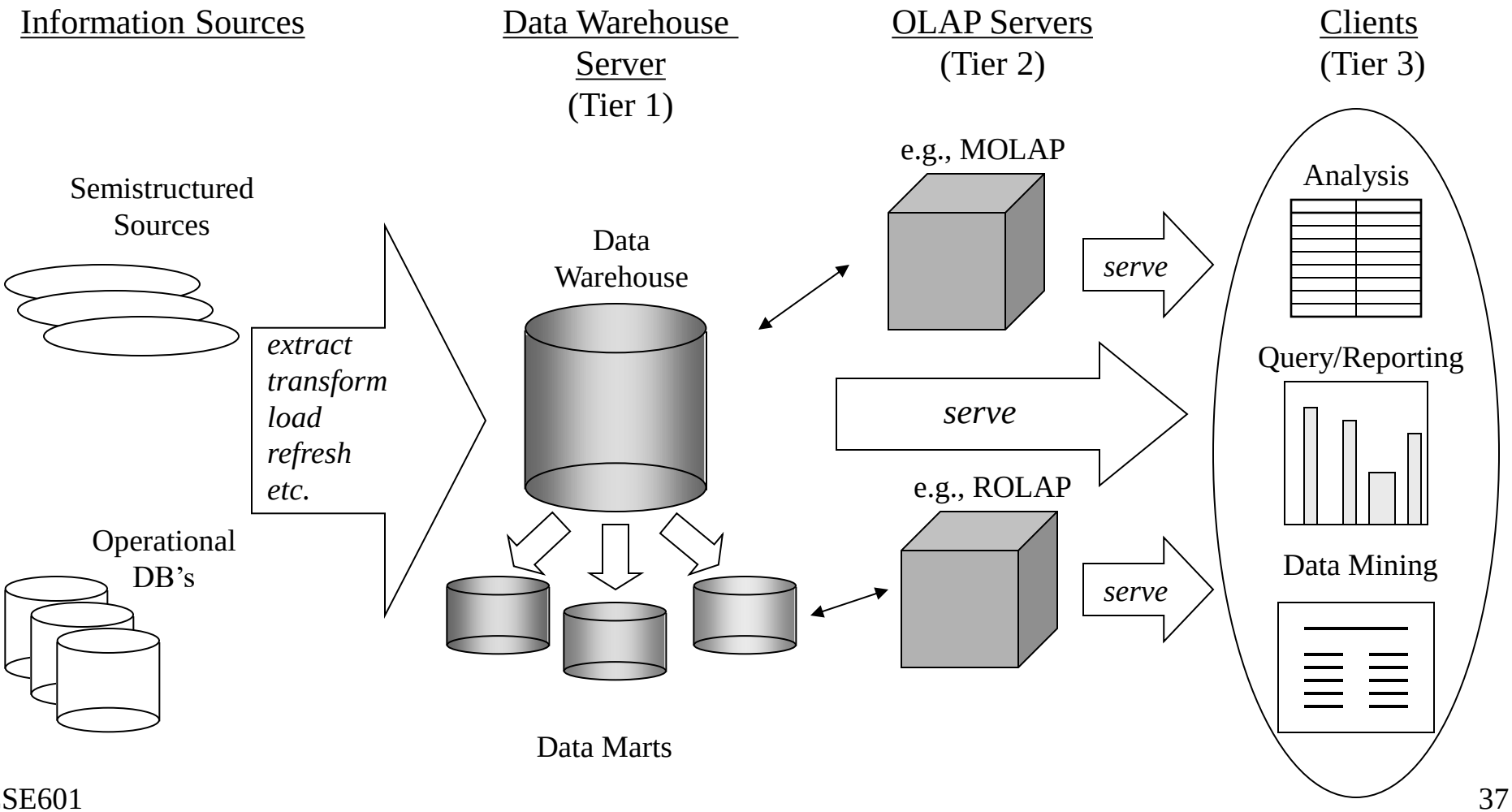
Decision Support

- 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(slice and dice)
  - Multidimensional OLAP (MOLAP): special-purpose server that directly implements multidimensional data and operations(cube,pivot)
- Clients
  - Query and reporting tools
  - Analysis tools
  - Data mining tools

The Complete Decision Support System



Data Warehouse Components

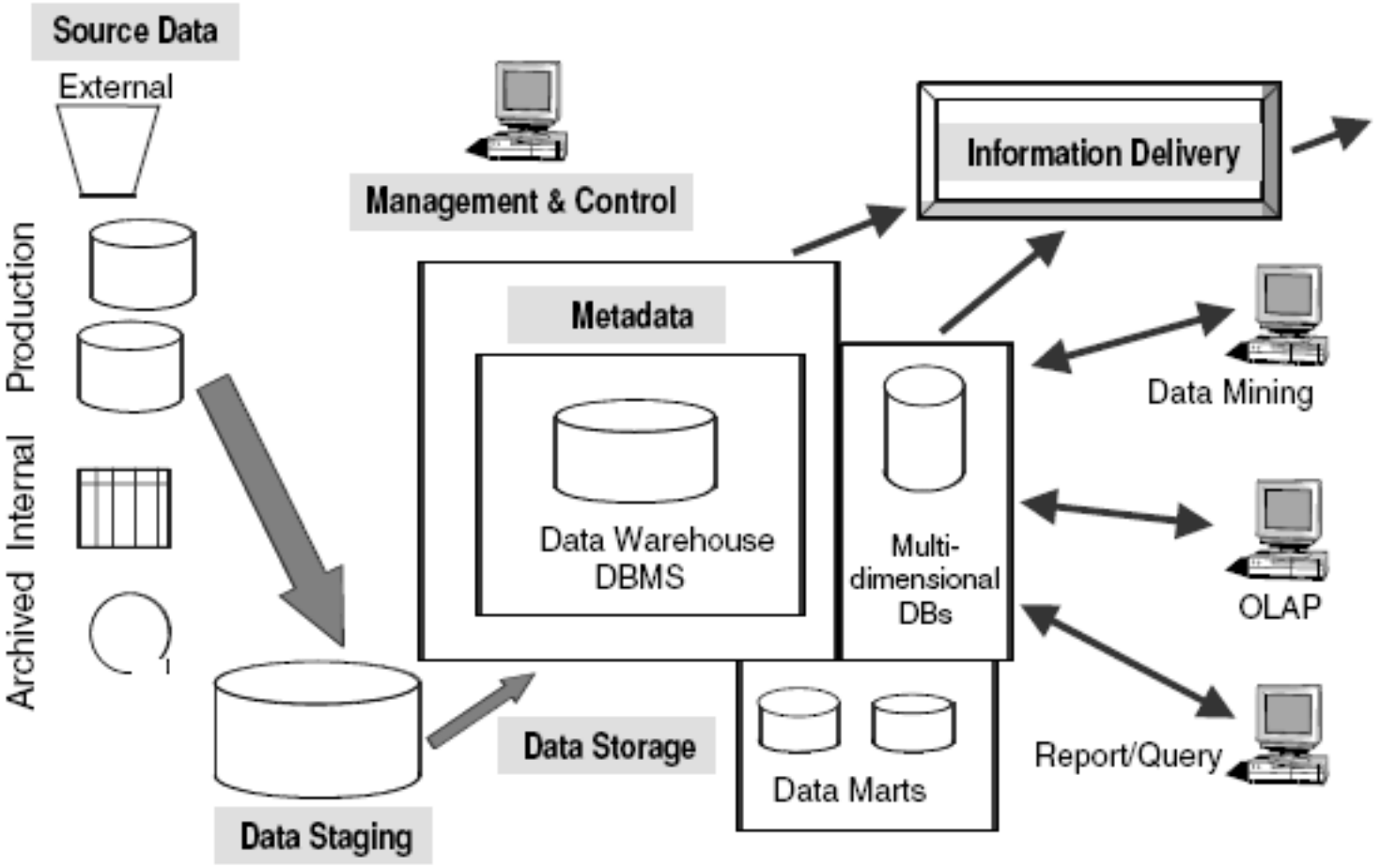


Figure 2-6 Data warehouse: building blocks or components.

Source data component

- Production systems
- Internal data (spreadsheets)
- Archived data (tapes)
- External data (stocks, interest rates, …)

Data Staging Component

- Data Extraction.
- Data Transformation.
- Data Loading.

Data Movement to the data Warehouse

- This function is time-consuming
- Initial load moves very large volumes of data
- The business conditions determine the refresh cycles

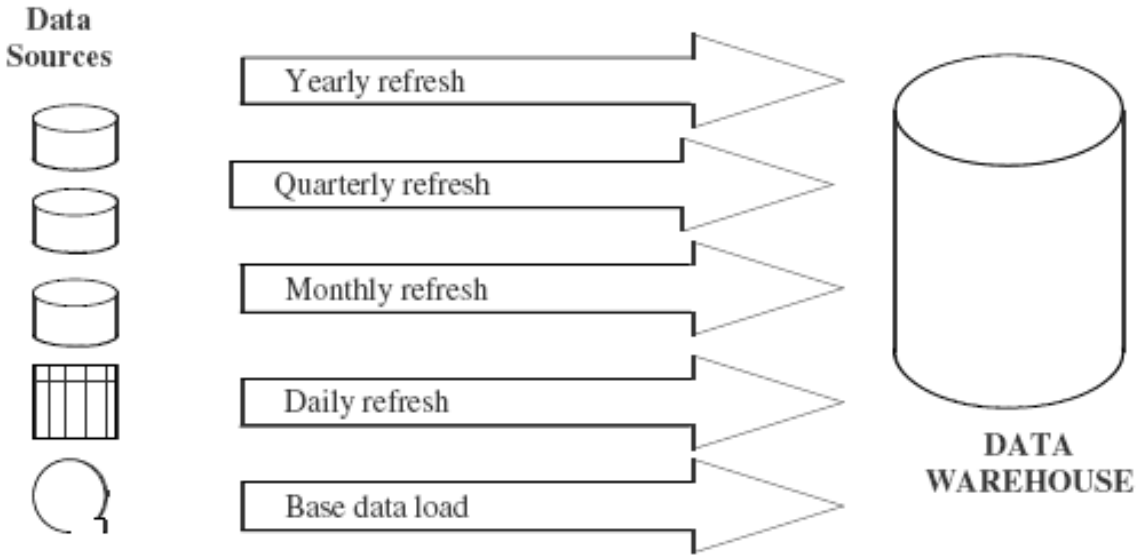


Figure 2-7 Data movements to the data warehouse.

Information Delivery Component

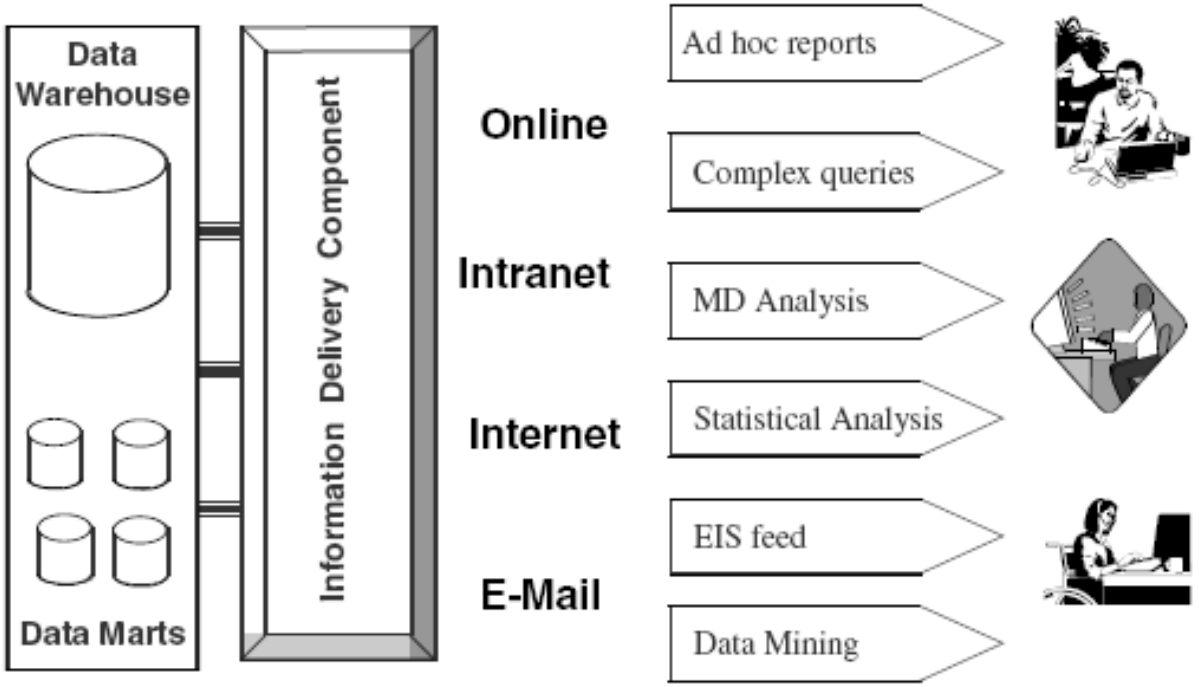


Figure 2-8 Information delivery component.

Data WH vs data mart

1. **Scope:**
    1. **Data Warehouse:** Enterprise-wide in scope, covering data from various business functions and departments.
    2. **Data Mart:** Departmental or business unit-specific, focusing on a subset of data relevant to a particular group.
2. **Size and Complexity:**
    1. **Data Warehouse:** Typically larger and more complex due to its enterprise-wide nature.
    2. **Data Mart:** Smaller and more focused, catering to the specific needs of a particular department.
3. **Integration:**
    1. **Data Warehouse:** Integrates data from multiple sources across the entire organization.
    2. **Data Mart:** Extracts and integrates data from the data warehouse or other local sources specific to the department.
4. **Autonomy:**
    1. **Data Warehouse:** Centralized control and management.
    2. **Data Mart:** May allow for more local autonomy in terms of data management and analysis.

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

Data marts are subsets of a data warehouse that focus on specific business functions, departments, or user groups.
They are designed to provide targeted and optimized data for analytical and reporting purposes.
There are several types of datamarts based on their scope, purpose, and the users they serve. Here are some common types:

1.Dependent Data Mart:

A dependent data mart relies on a centralized data warehouse for its data.
It extracts and processes a subset of the data warehouse's information to serve the specific needs of a particular business unit or department.

2.Independent Data Mart:

An independent data mart operates independently of a centralized data warehouse.
It is a standalone data repository designed to meet the specific needs of a particular business unit or functional area.
Independent data marts are often quicker to implement but may lack the consistency and integration benefits of centralized data warehouses.

Dependent Data Mart:

Sales Data Mart: A dependent data mart for the Sales department that relies on a centralized data warehouse.

It extracts and processes relevant data from the centralized warehouse, providing sales teams with information on customer purchases, product performance, and sales trends.

Finance Data Mart: A dependent data mart for the Finance department that pulls information from the centralized data warehouse.

It focuses on financial metrics, budgeting, and financial reporting, offering insights into revenue, expenses, and financial performance.

Human Resources Data Mart: This data mart serves the Human Resources department, relying on the centralized data warehouse for employee data.

It provides HR professionals with analytics related to workforce management, employee performance, and talent acquisition.

Independent Data Mart:

Marketing Data Mart: An independent data mart specifically designed for the Marketing department.

It operates autonomously, collecting and storing marketing-related data without direct reliance on a centralized data warehouse. T

his allows the Marketing team to have a more agile and specialized data repository for campaign analysis, customer segmentation, and marketing performance.

Research and Development Data Mart: An independent data mart for the Research and Development department.

This mart collects and stores data related to research projects, product development, and innovation without relying heavily on a centralized warehouse.

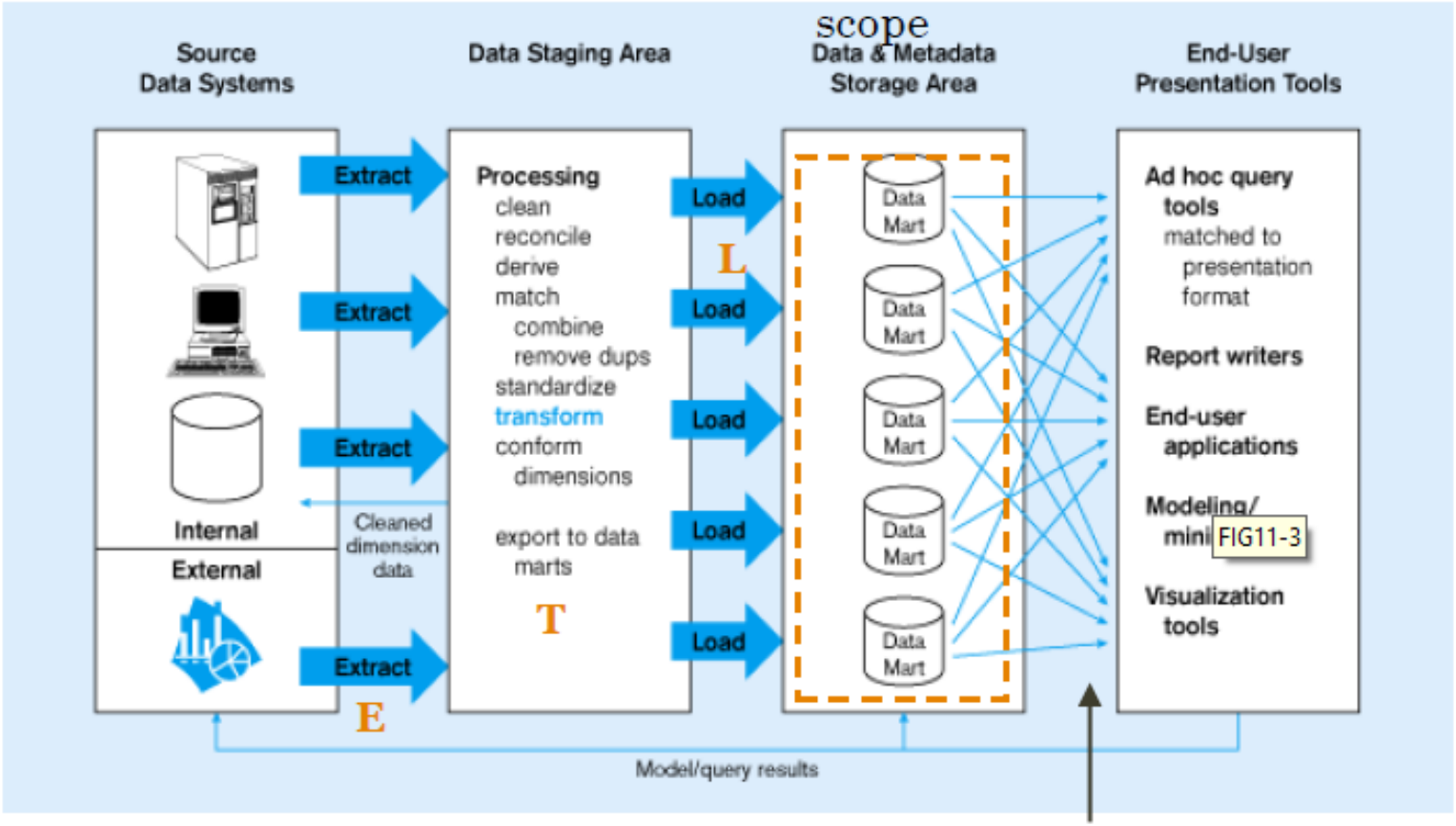
It provides R&D teams with quick access to project-specific data for analysis.

Customer Support Data Mart: An independent data mart catering to the Customer Support department.

It focuses on customer service metrics, ticket resolution times, and customer feedback, operating independently to meet the specific needs of the support team.

Figure 11-3: Independent Data Mart

Data marts:
Mini-warehouses, limited in scope



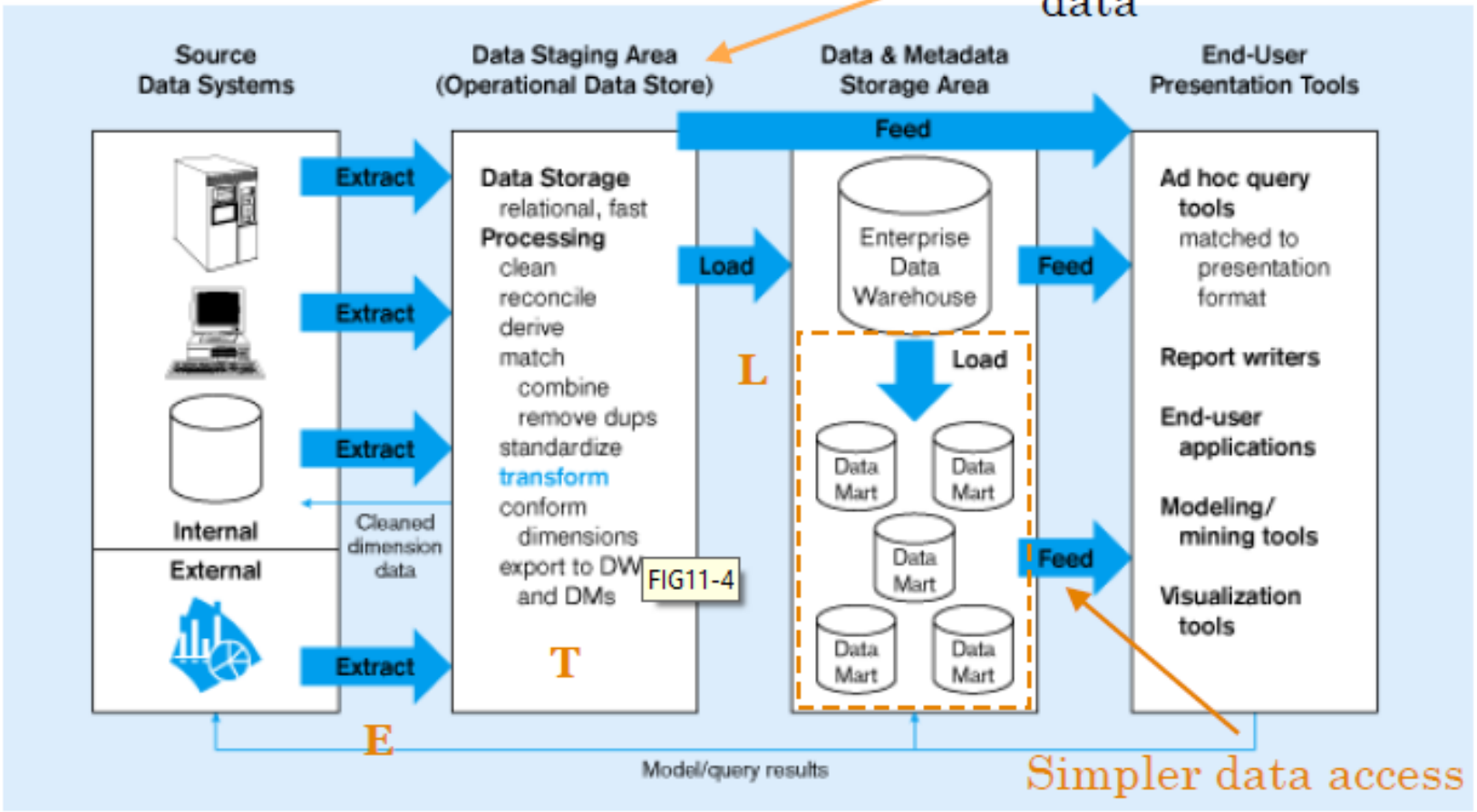
Separate ETL for each ***independent*** data mart

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

20

Figure 11-4:
Dependent data mart with *operational data store*

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



KJSCE

Single ETL for ***enterprise data warehouse (EDW)***

Dependent data marts loaded from EDW

Data Warehousing Design Strategies.

Bottom up approach

Data Marts → Data warehouse

Top down approach

Data Warehouse → Data Marts

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

What is Metadata?

- Metadata is simply defined as data about data.
- The data that is used to represent other data is known as metadata.
- For example, the index of a book serves as a metadata for the contents in the book. In other words, we can say that metadata is the summarized data that leads us to detailed data. define metadata as follows.

METADATA IN THE DATA WAREHOUSE

WHY METADATA IS IMPORTANT

Users to compose and run the query can have several important questions:

- Are there any predefined queries I can look at?
- What are the various elements of data in the warehouse?
- Is there information about unit sales and unit costs by product?
- How can I browse and see what is available?
- From where did they get the data for the warehouse? From which source systems?
- How did they merge the data from the telephone orders system and the mail orders system?
- How old is the data in the warehouse?
- When was the last time fresh data was brought in?
- Are there any summaries by month and product?

- Metadata in a data warehouse contains the answers to questions about the data in the data warehouse.

Different definitions for metadata

- Data about the data
- Table of contents for the data
- Catalog for the data
- Data warehouse atlas
- Data warehouse roadmap
- Data warehouse directory
- Glue that holds the data warehouse contents together
- The nerve center

Metadata in OLTP

- In operational systems we do not really have any easy and flexible methods for knowing the nature of the contents of the database.
- There is no great need for user-friendly interfaces to the database contents.
- **The data dictionary or catalog is meant for IT uses only**.

Metadata in DWH

- Users need sophisticated methods for browsing and examining the contents of the data warehouse.
- Users need to know the meanings of the data items.
- Users have to prevent them from drawing wrong conclusions from their analysis through their ignorance about the exact meanings.
- Without adequate metadata support, users of the larger data warehouses are totally handicapped.

Types of Metadata

- Metadata in a data warehouse fall into three major categories:
- Operational Metadata
- Extraction and Transformation Metadata
- End-User Metadata

Operational Metadata(technical)

- Data for the data warehouse comes from several operational systems of the enterprise.
- These source systems contain **different data structures**.
- The data elements selected for the data warehouse have **various field lengths and data types**.
- In selecting data from the source systems for **the data warehouse, you split records, combine parts of records from different source files, and deal with multiple coding schemes and field lengths**.
- When you deliver information to the end-users, you must be able to tie that back to the original source data sets.
- Operational metadata contain all of this information about the operational data sources.

Extraction and Transformation Metadata

- Extraction and transformation metadata contain data about the extraction of data from the source systems, namely, the **extraction frequencies, extraction methods, and business rules for the** data extraction. Also, this category of metadata contains **information about all the data transformations that take place in the data staging area**.

End-User Metadata

- The end-user metadata is **the navigational map of the data warehou**se. It enables the end-users to find information from the data warehouse. The end-user metadata allows the end-users to use their own business terminology and look for information in those ways in which they normally think of the business.

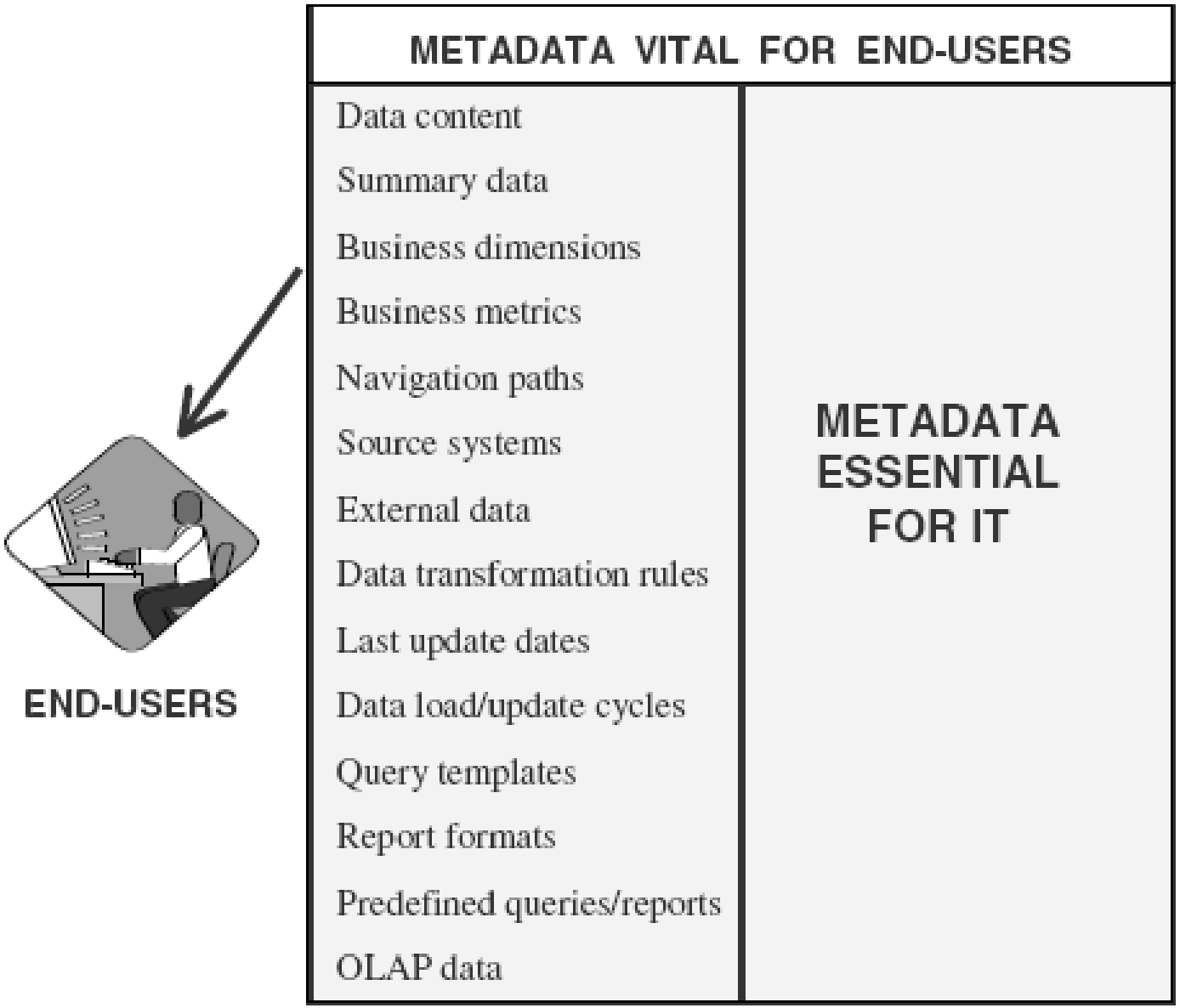


Figure 9-5 Metadata vital for end-users.

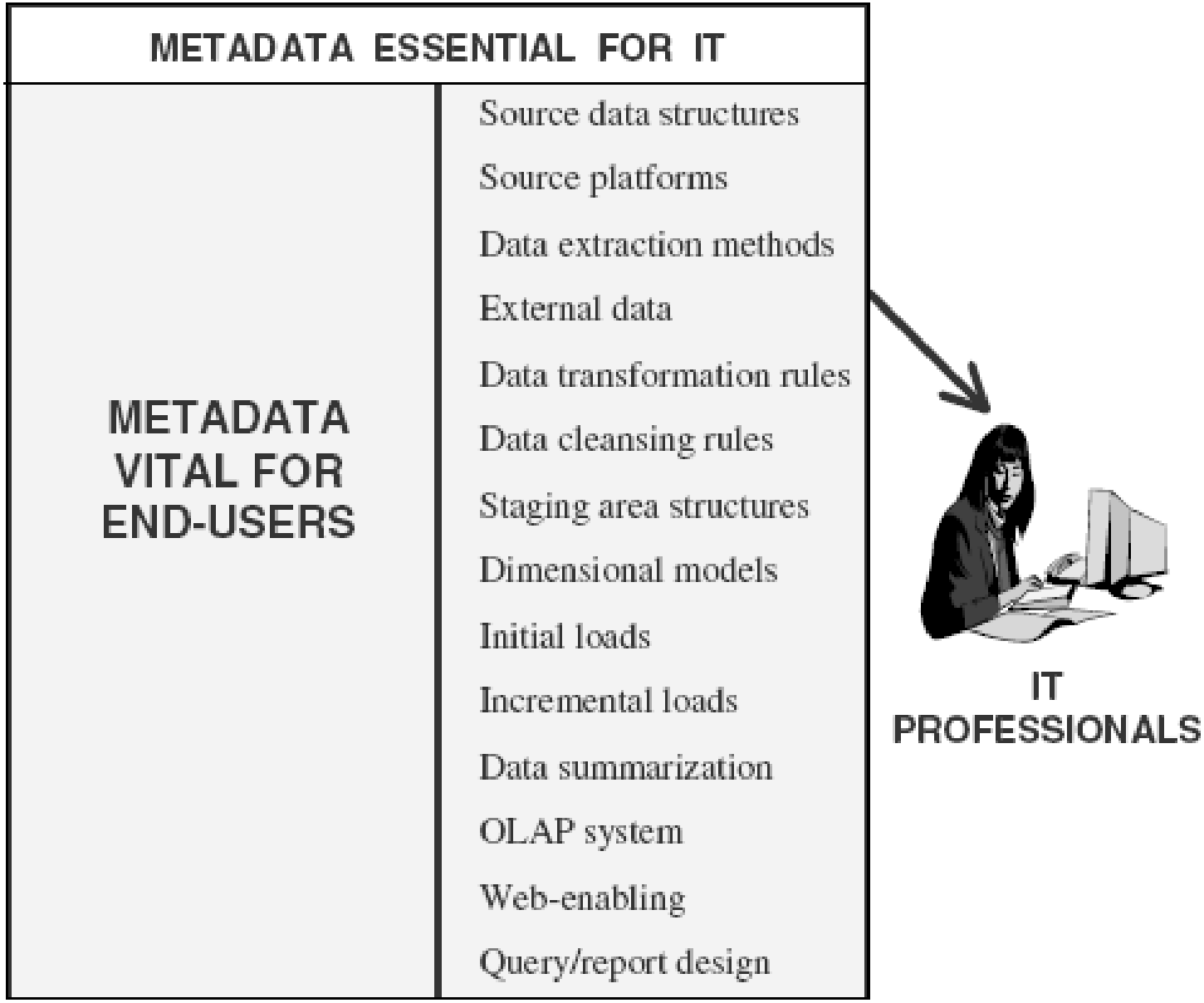


Figure 9-6 Metadata essential for IT.

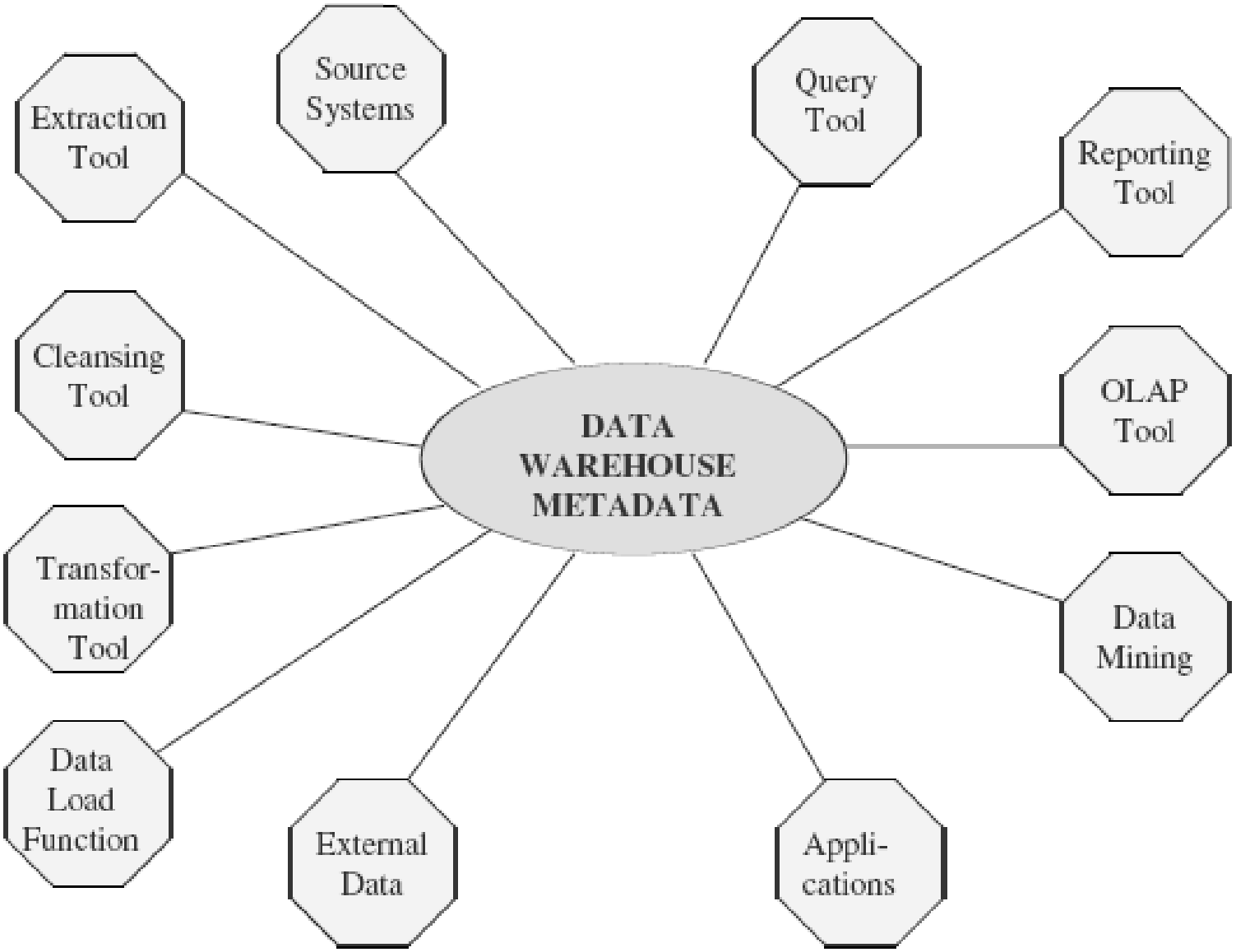


Figure 9-4 Metadata acts as a nerve center.

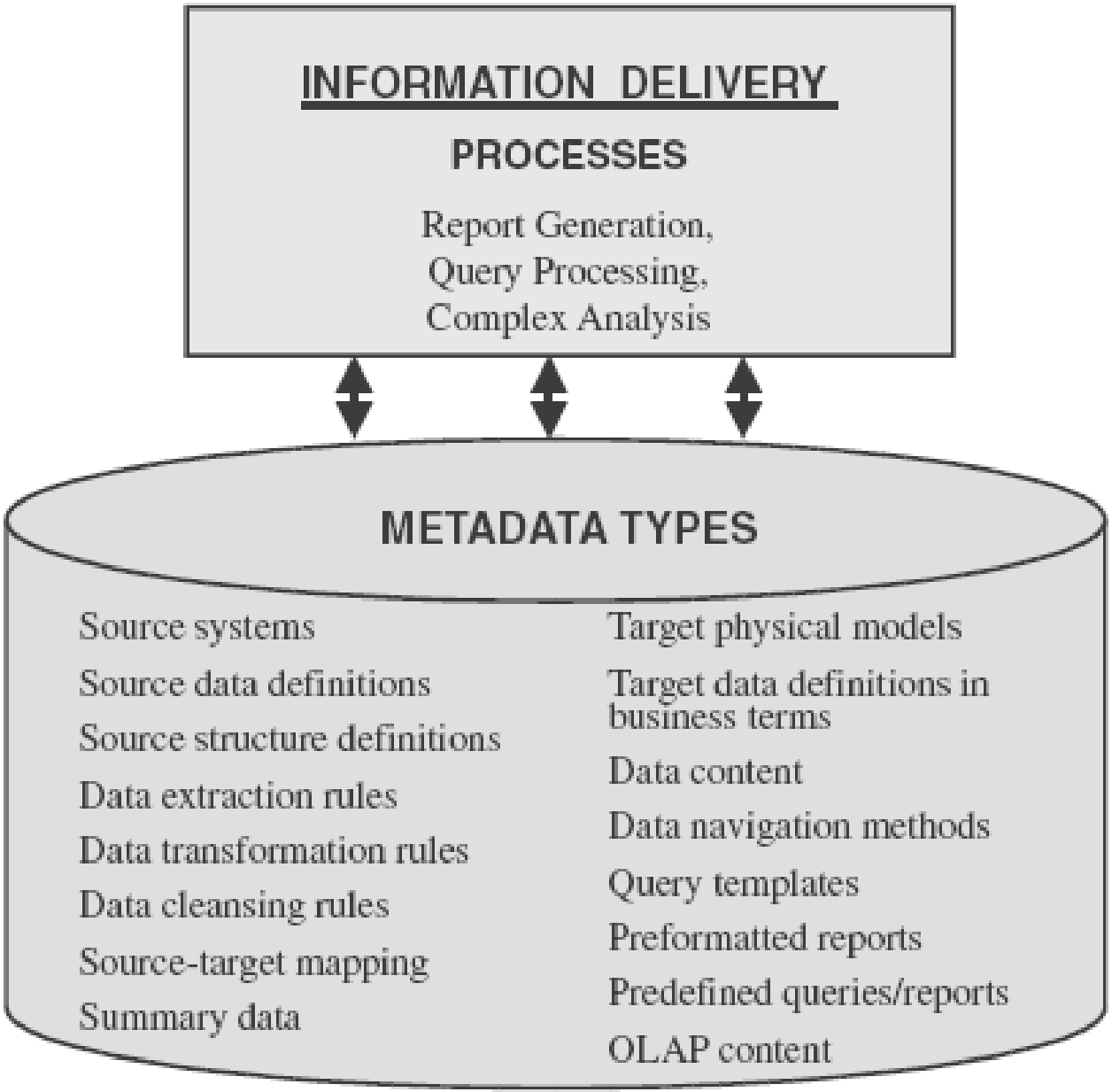


Figure 9-10 Information delivery: metadata types.

Metadata Management

- Metadata management answers these questions:
- What is Metadata?
- How can Metadata be Managed?
- Extracting Metadata from Legacy Systems

Metadata is your control panel to the data warehouse. It is data that describes the data warehousing and business intelligence system:

- Reports
- Cubes
- Tables (Records, Segments, Entities, etc.)
- Columns (Fields, Attributes, Data Elements, etc.)
- Keys
- Indexes

Metadata is often used to control the handling of data and describes:

- Rules
- Transformations
- Aggregations
- Mappings

The power of metadata is that enables data warehousing personnel to develop and control the system without writing code in languages such as: Java, C# or Visual Basic. This saves time and money both in the initial set up and on going management.

- Data warehousing has specific metadata requirements. Metadata that describes tables typically includes:
- Physical Name
- Logical Name
- Type: Fact, Dimension, Bridge
- Role: Legacy, OLTP, Stage,
- DBMS: DB2, Informix, MS SQL Server, Oracle, Sybase
- Location
- Definition
- Notes

Metadata describes columns within tables:

- Physical Name
- Logical Name
- Order in Table
- Datatype
- Length
- Decimal Positions
- Nullable/Required
- Default Value
- Edit Rules
- Definition
- Notes

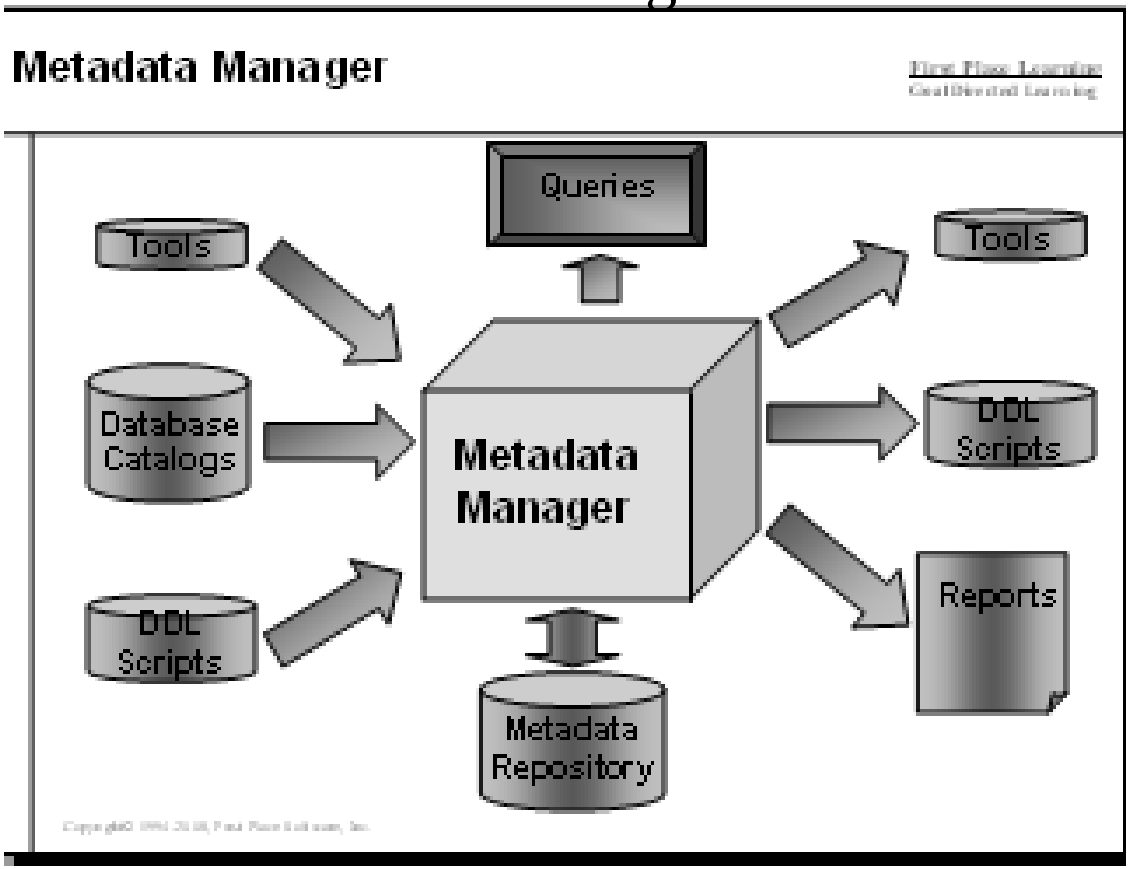
How can Data Warehousing Metadata be Managed?

- Data warehousing and business intelligence metadata is best managed through a combination of **people, process and tools.**
- The people side requires that people be trained in the importance and use of metadata. They need to understand how and when to use tools as well as the benefits to be gained through metadata.
- The process side incorporates metadata management into the data warehousing and business intelligence life cycle. As the life cycle progresses metadata is entered into the appropriate tool and stored in a metadata repository for further use.
- Metadata can be managed through individual tools:
    1. Metadata manager / repository
    2. Metadata extract tools
    3. Data modeling
    4. ETL

Metadata Manager / Repository

Metadata can be **managed through a shared repository** that combines information from multiple sources.

The metadata manager can be **purchased as a software package or built as "home grown" system.** Many organizations start with a spreadsheet containing data definitions and then grow to a more sophisticated approach.



Extracting Metadata from Input Sources

Metadata can be obtained through **a manual process** of keying in metadata or **through automated processes**.
Scanners can extract metadata from text such as SQL DDL or COBOL programs.
Other tools can **directly access metadata through SQL catalogs and other metadata** sources.
Picking the appropriate metadata extract tools is a key part of metadata management.
Many data modeling tools include a metadata extract capability - otherwise known as "reverse engineering".
Through this tool, database information about tables and columns can be extracted.
The information can then be exported from the data modeling tool to the metadata manager.

Challenges for Metadata Management

- Metadata in a big organization is s**cattered** across the organization. This metadata is spread in spreadsheets, databases, and applications.
- Metadata could be present in **text files or multimedia files**. To use this data for information management solutions, it has to be correctly defined.
- There are **no industry-wide accepted standards**. Data management solution vendors have narrow focus.
- There are no easy and accepted methods of passing metadata.