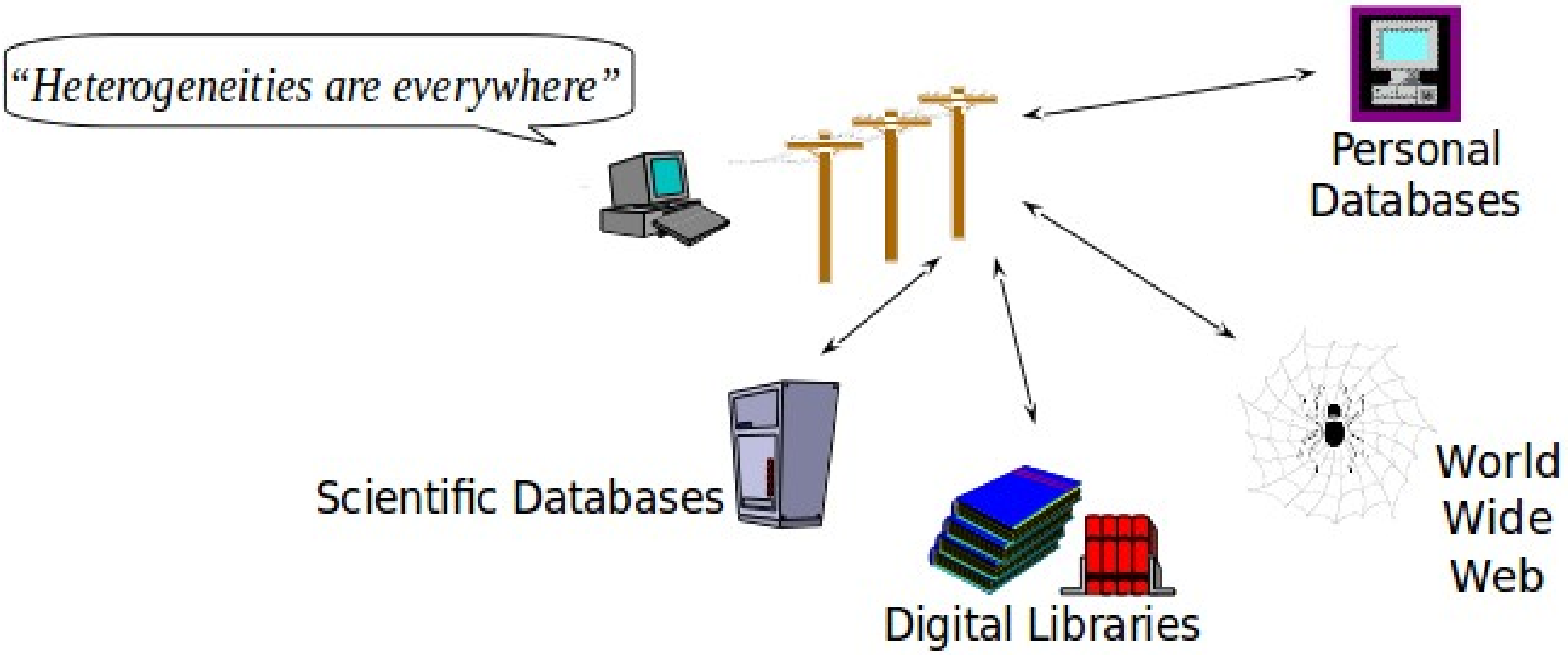
Data Mining

Ch2. Introduction to Data Warehousing

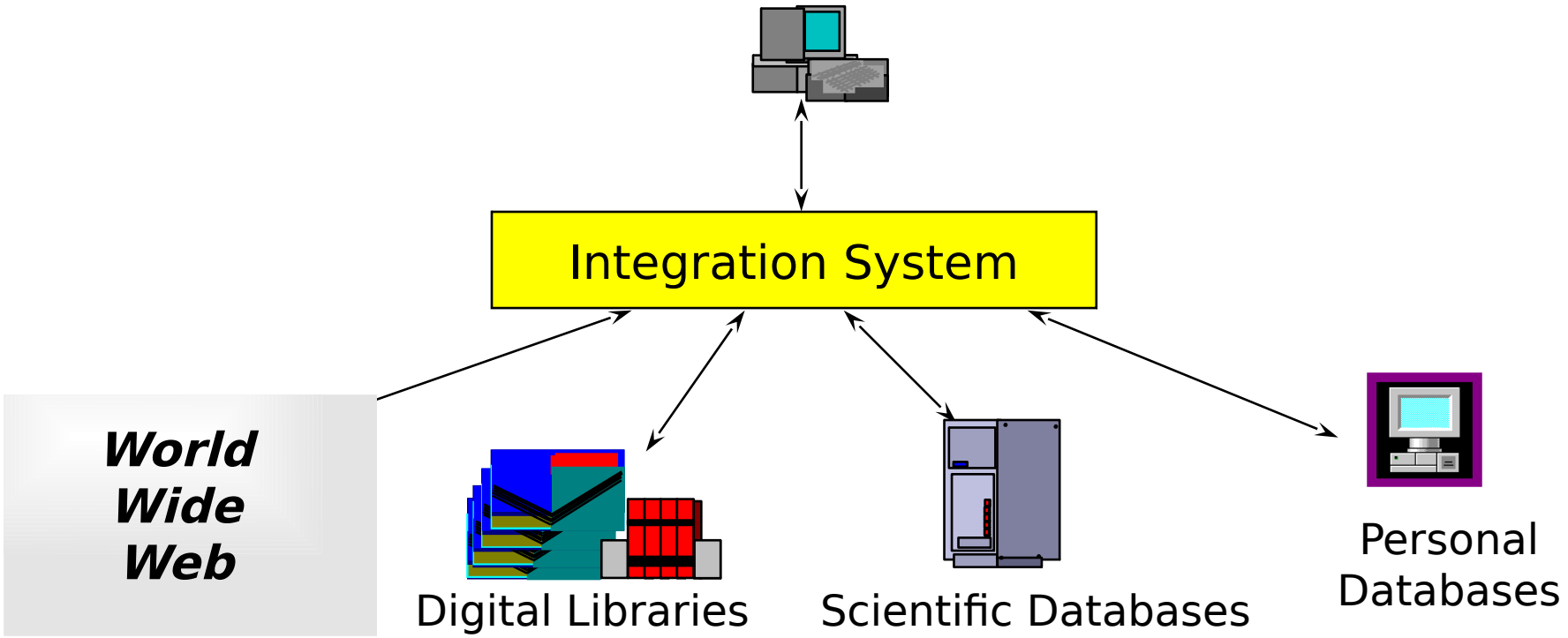
Er. Bidur Devkota
Gandaki College of Engineering and Science
Pokhara-16

Heterogeneity as a problem



- 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

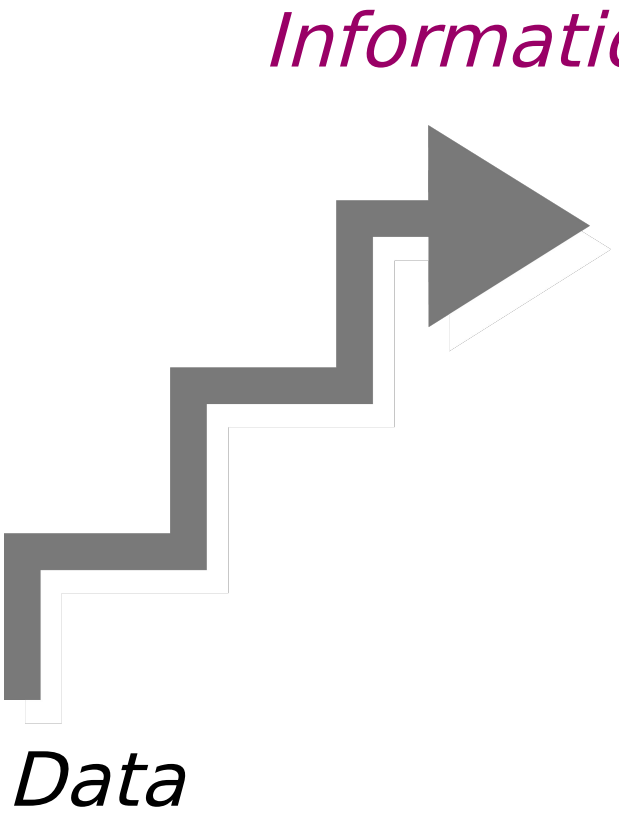
Evolution

- 60’s: Batch reports
  - hard to find and analyze information
  - inflexible and expensive, reprogram every new request
- 70’s: Terminal-based DSS (Decision Support Systems) and EIS (Executive Information Systems)
  - still inflexible, not integrated with desktop tools
- 80’s: Desktop data access and analysis tools
  - query tools, spreadsheets, GUIs
  - easier to use, but only access operational databases
- 90’s: Data warehousing with integrated OLAP engines and tools

Data Warehouse: user expectation

- Data should be integrated across the enterprise
- Summary data has a real value to the organization
- Historical data holds the key to understanding data over time
- What-if capabilities are required

Data Warehousing as a process



A **process** of transforming **data** into **information** and making it **available** to users in a **timely** enough manner to make a **difference**

[Forrester Research, April 1996]

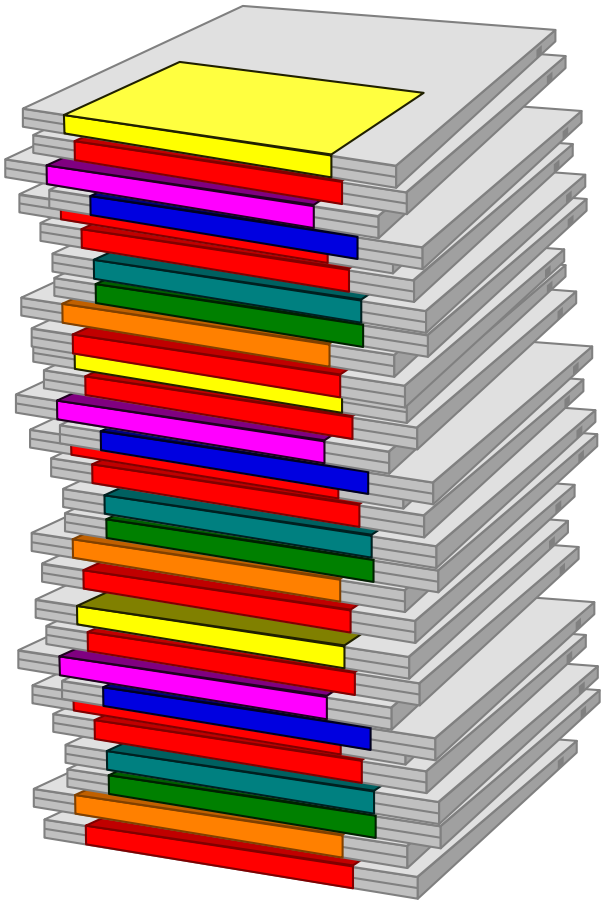
Data Warehousing as a process

- Technique for assembling and managing data from various sources for the purpose of answering business questions.
- Thus making decisions that were not possible previously
- A decision support database maintained separately from the organization’s operational database

Data Warehouse

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

[Barry Devlin]



Data Warehouse

- A data warehouse refers to a database that is maintained separately from an organization's operational databases.
- Data warehouse systems allow for the integration of a variety of application systems.
- They support information processing by providing a solid platform of consolidated historical data for analysis.

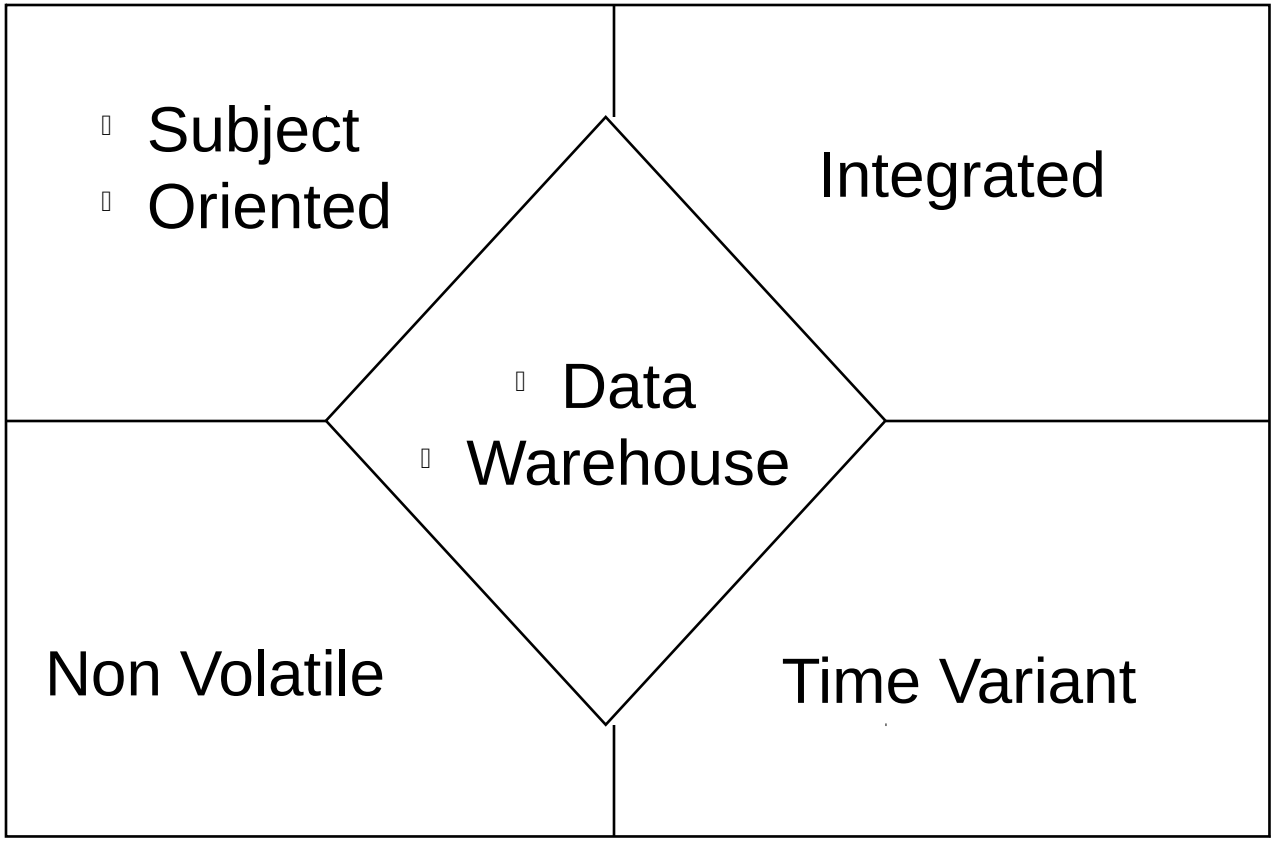
Data Warehouse

- A data warehouse is a
  - subject-oriented
  - integrated
  - time-varying
  - non-volatile

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

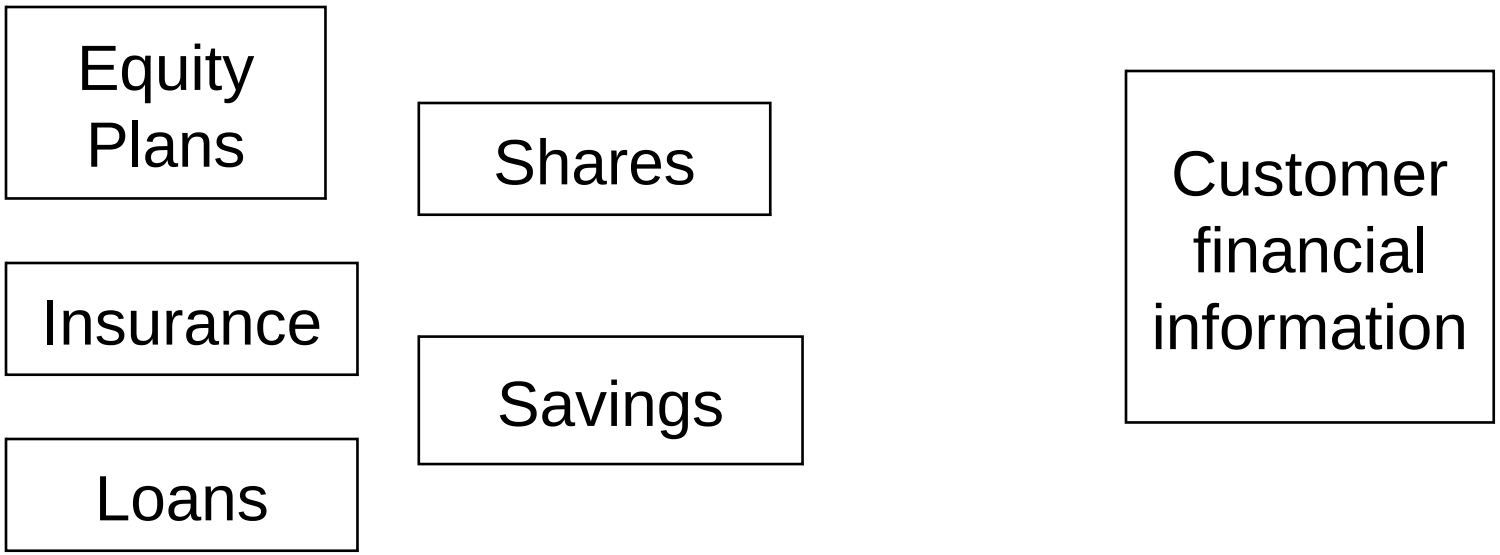
-- Bill Inmon, Building the Data Warehouse 1996

Data Warehouse Properties



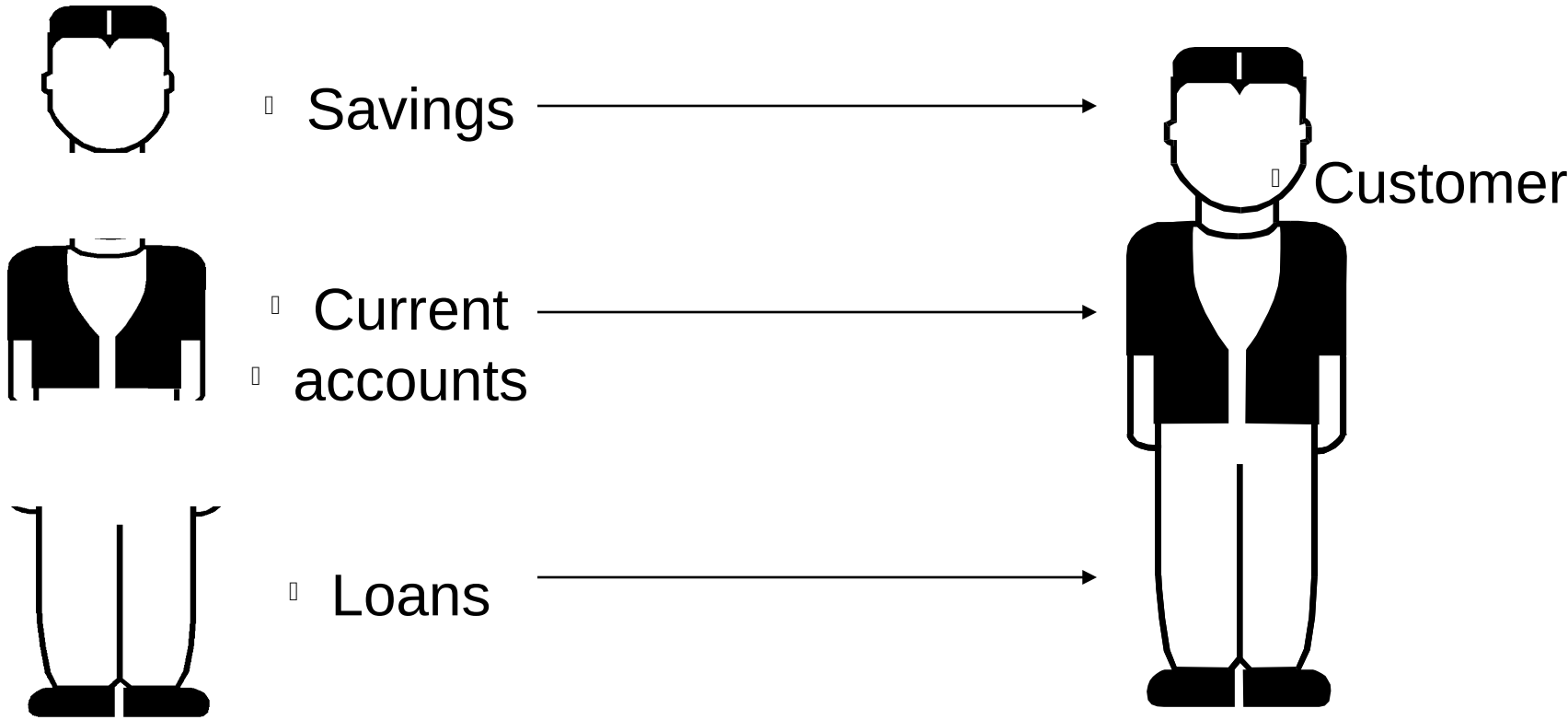
Subject-Oriented

- Data warehouses provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.
- Data is categorized, stored and Organized by subject not by application
- Used for analysis, data mining, etc.
  - **OLTP Applications**
  - **Data Warehouse Subject**



Integrated

- Data on a given subject is defined and stored once.



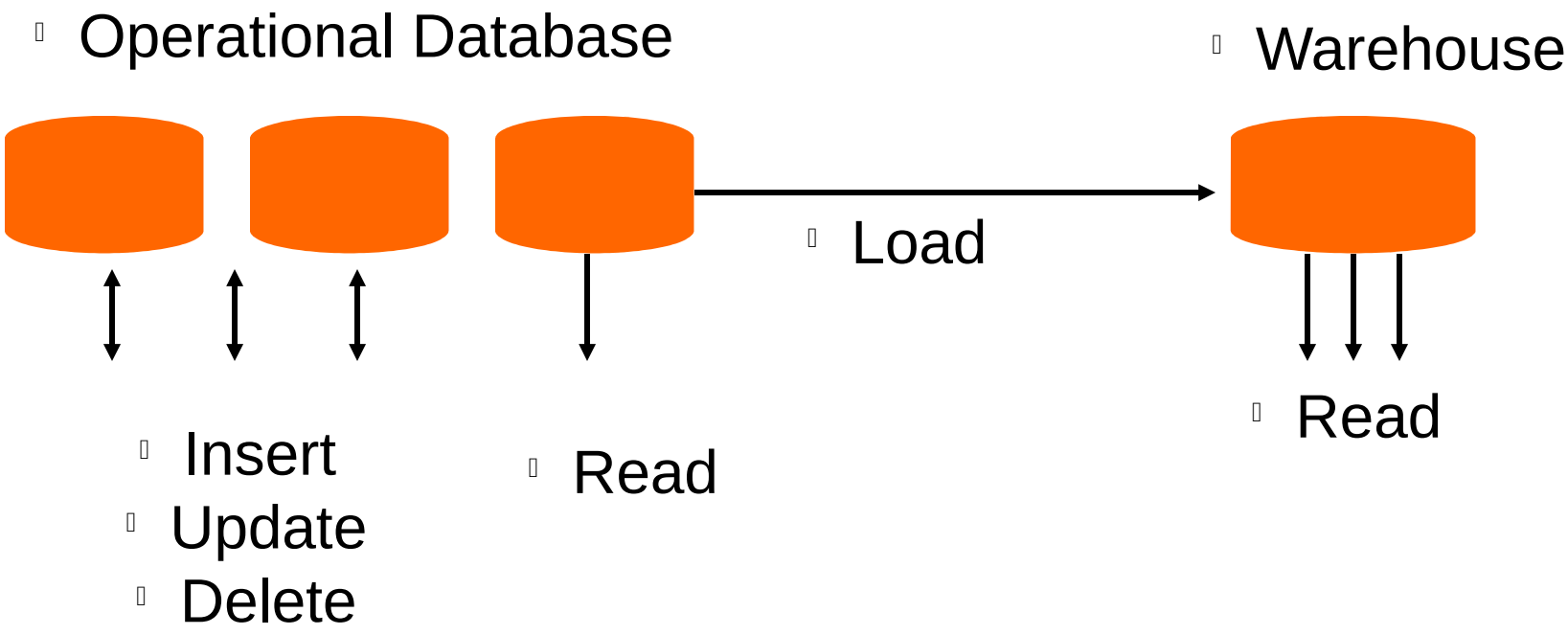
- **OLTP Applications**
- **Data Warehouse**

Time-Variant

- Data are stored to provide information from a historical perspective (e.g., the past 5–10 years).
- Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.
- Data is stored as a series of snapshots, each representing a period of time.

Nonvolatile

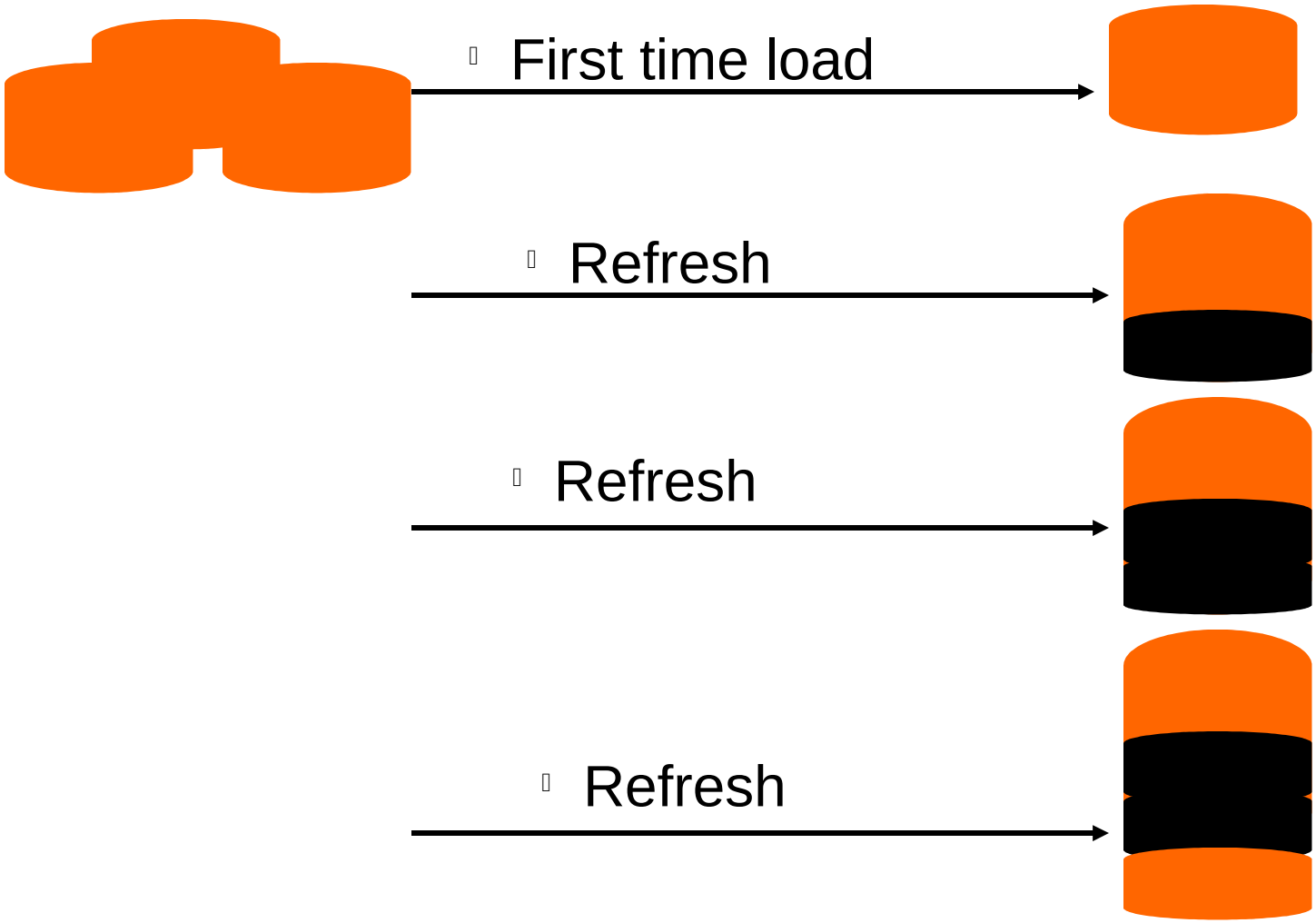
- Typically data in the data warehouse is not updated or deleted



Changing Data

Operational Database

- **Warehouse Database**

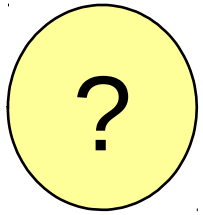


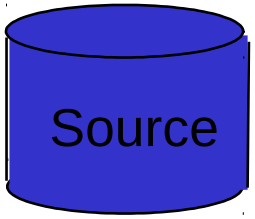
Why Separate Data Warehouse?

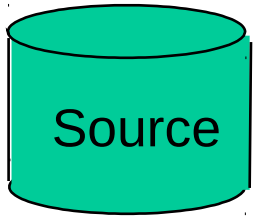
- **High performance** for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation
- **Different functions and different data**:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: Decision Support requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled
- Note: There are more and more systems which perform OLAP analysis directly on relational databases

Warehousing Approaches

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

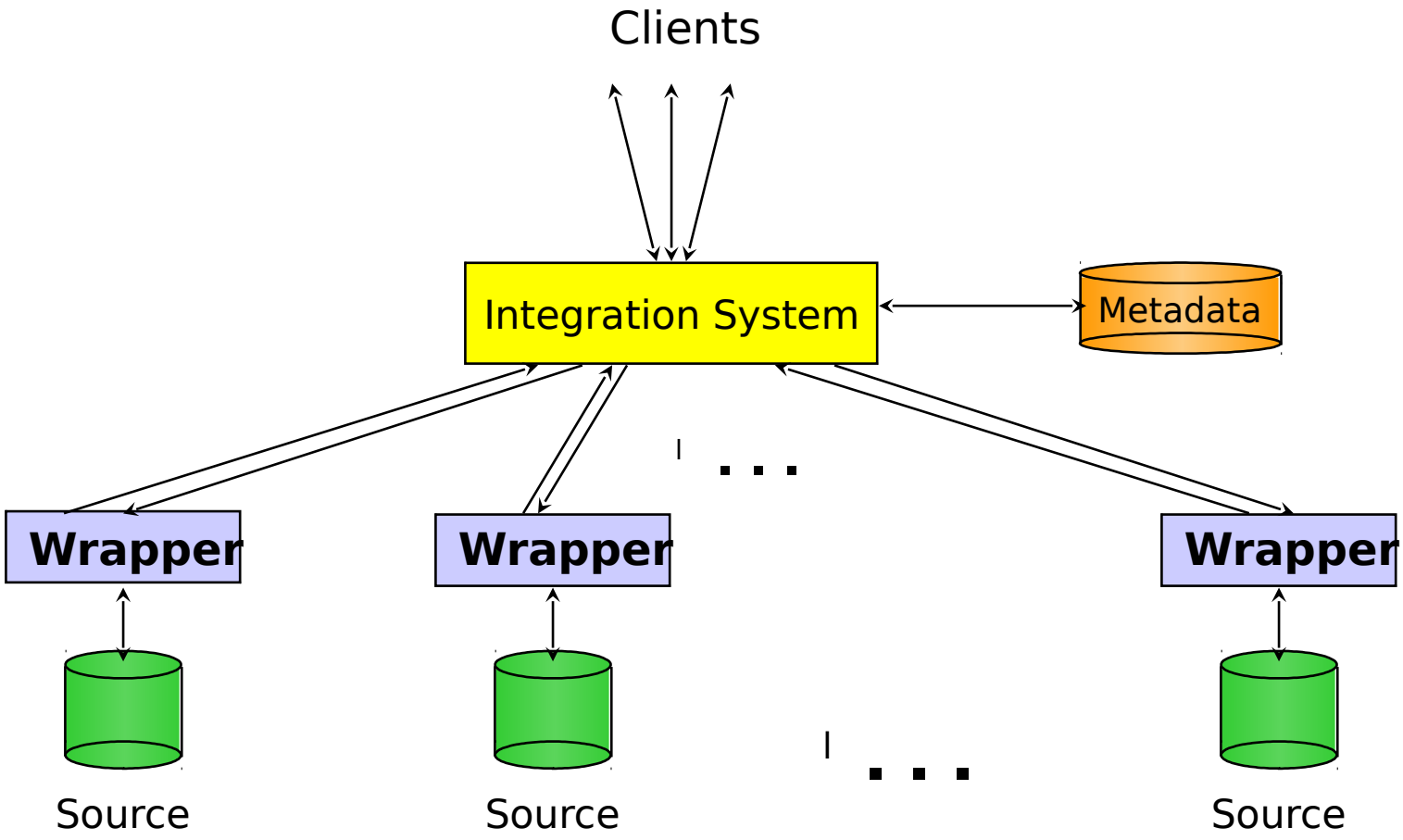






The Traditional Research Approach

Query-driven (lazy, on-demand)



Wrappers are adapters which transform the local query results (those returned by the respective databases) into an easily processed form for the data integration solution

Merits of Query driven approach

- 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

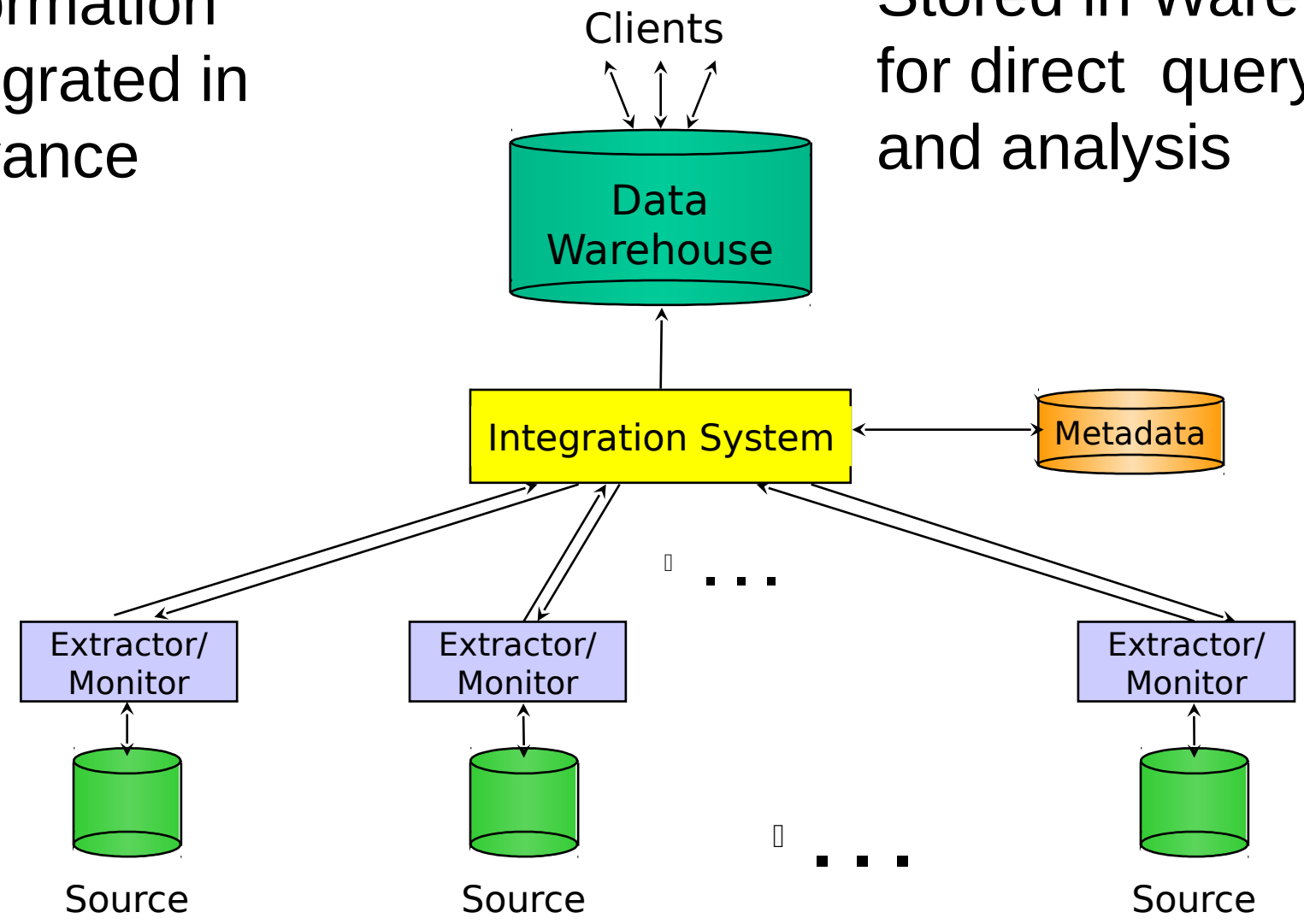
Demerits of Query-Driven Approach

- 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

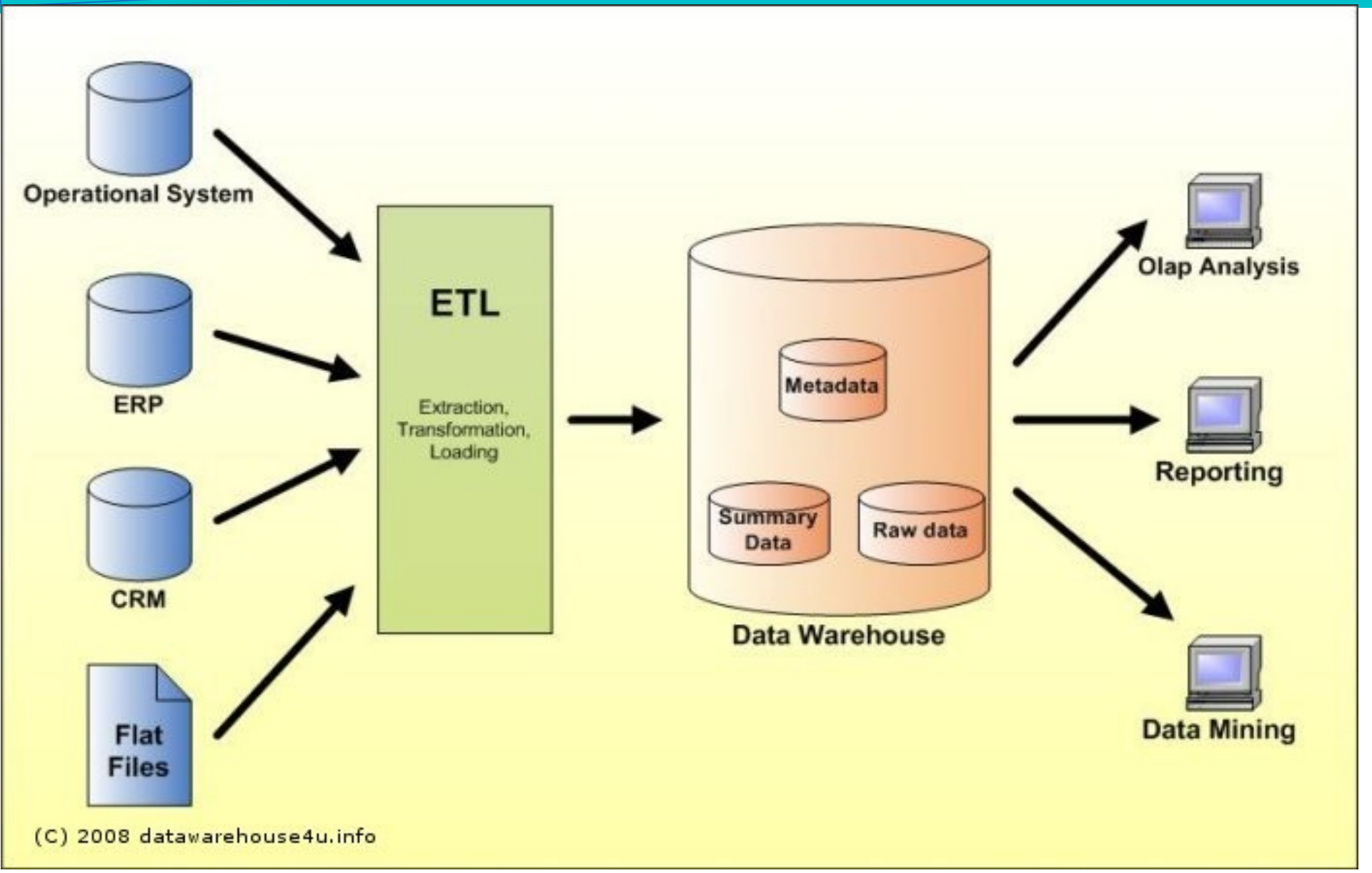
Stored in Warehouse for direct querying and analysis



Merits 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

Data Warehouse Architecture



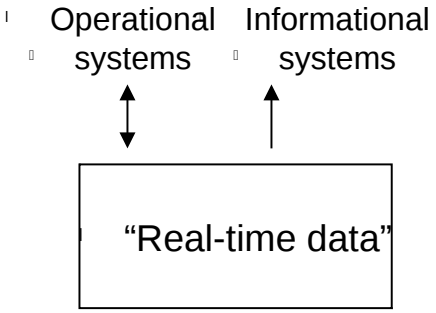
Data Warehouse Architecture

The essential architecture properties:

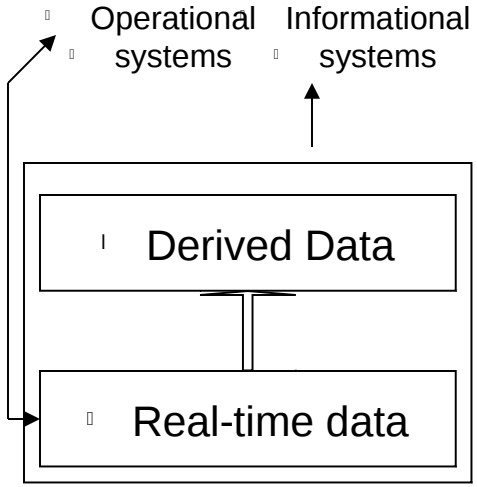
- **Separation**: *Analytical and transaction processing* should be kept apart as much as possible.
- **Scalability:** Hardware and software architectures should be easy to *upgrade* as the data volume, which has to be managed and processed, and the number of users' requirements, which have to be met, progressively increase.
- **Extensibility**: The architecture should be able to host new applications and technologies *without redesigning* the whole system.
- **Security**: Monitoring *accesses* is essential because of the strategic data stored in the warehouses.
- **Administerability**: Data warehouse management should not be difficult.

Data Warehouse Architectures: Conceptual View

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



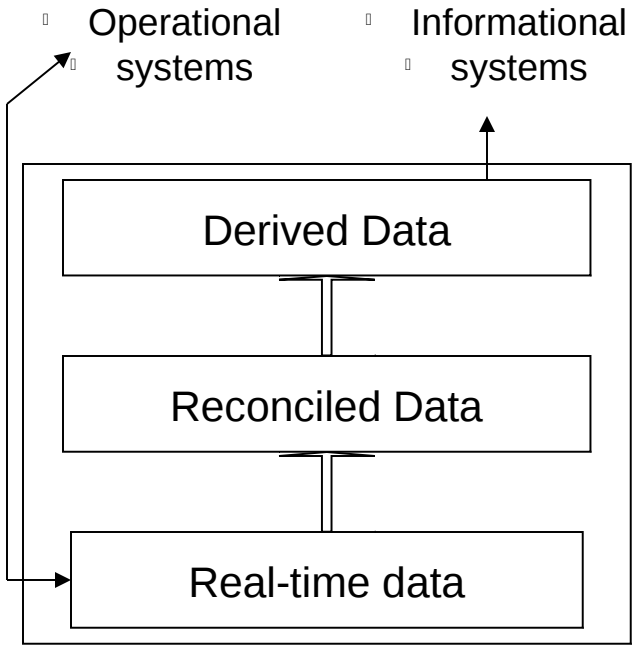
- **Two-layer**
  - Real-time + derived data
  - Most commonly used approach
  - in industry today



Data Warehouse Architectures: Conceptual View

Three-layer

- 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)

The Meta data

- The name suggests some high-level technological concept, but it really is fairly simple.
- Metadata is ***"data about data***".
- With the emergence of the data warehouse as a decision support structure, the metadata are considered as an important resource (like the business data they describe)
- Metadata are **abstractions** -- they are high level data that provide **concise descriptions** of lower-level data.

The Metadata

- For eg, a line in a sales database may contain:
  - 1024 P02 1000
- The meaning of this line is provided by referring the metadata:
  - store number 1024, product P02 and sales of $1000
- The metadata are essential ingredients in the transformation of *raw data into knowledge*.
- They are the “keys” that allow us to handle the raw data.

General Metadata Issues

- What tables, attributes and keys does the DW contain?
- Where did each set of data come from?
- What transformations were applied with cleansing?
- How have the metadata changed over time?
- How often do the data get reloaded?
- Are there so many data elements that we need to be careful what we ask for?

OLAP

- Online Analytical Processing - coined by EF Codd in 1994 paper
- Generally synonymous with earlier terms such as :
  - Decisions Support,
  - Business Intelligence,
  - Executive Information System
- OLAP = Multidimensional Database

* Reference: http://www.arborsoft.com/essbase/wht_ppr/coddTOC.html

Typical OLAP Queries

- Write a multi-table join to compare sales for each product line this year vs. last year.
- Repeat the above process to find the top 5 product contributors to margin.
- Repeat the above process to find the sales of a product line to new vs. existing customers.
- Repeat the above process to find the customers that have had negative sales growth.

Typical OLAP Examples

1. Amazon analyzes purchases by its customers to come up with an individual screen with products of likely interest to the customer.

2. Analysts at Wal-Mart look for items with increasing sales in some region.

Strengths of OLAP

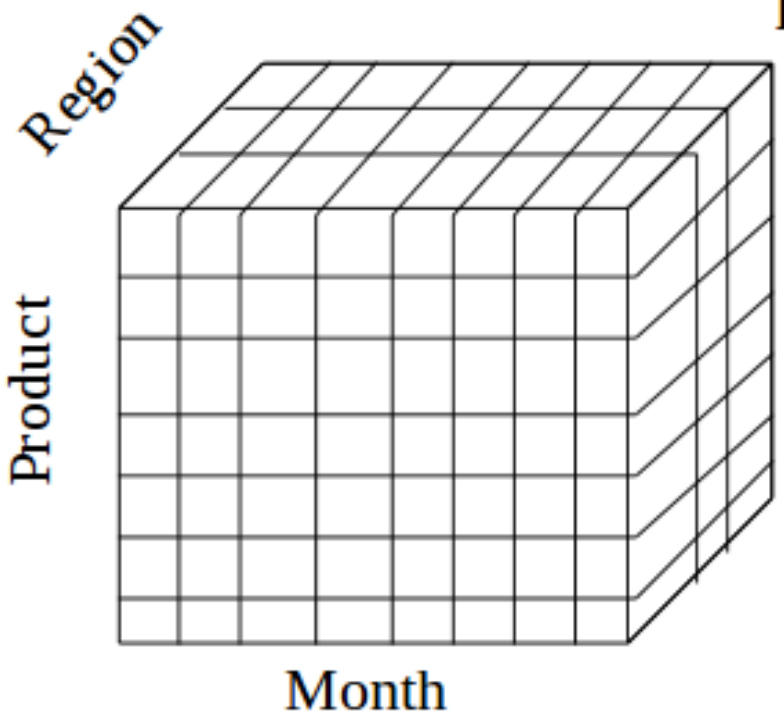
- It is a powerful visualization paradigm
- It provides fast, interactive response times
- It is good for analyzing time series
- It can be useful to find some clusters and outliers
- Many vendors offer OLAP tools

Nature of OLAP Analysis

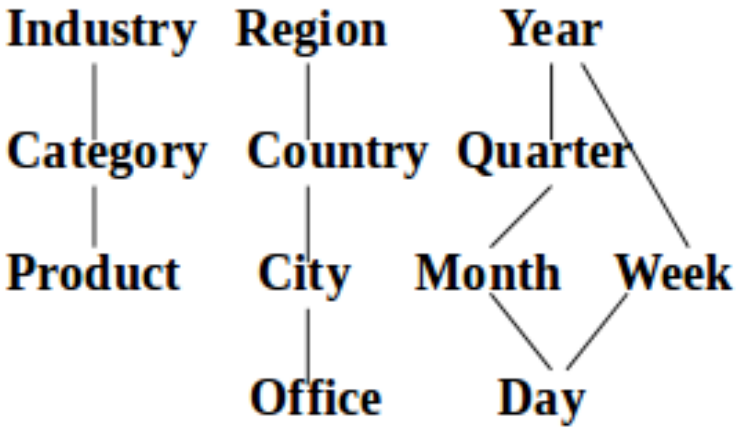
- Aggregation -- (total sales, percent-to-total)
- Comparison -- Budget vs. Expenses
- Ranking -- Top 10, quartile analysis
- Access to detailed and aggregate data
- Complex criteria specification
- Visualization

Multidimensional OLAP

Sales volume as a function of product, month, and region



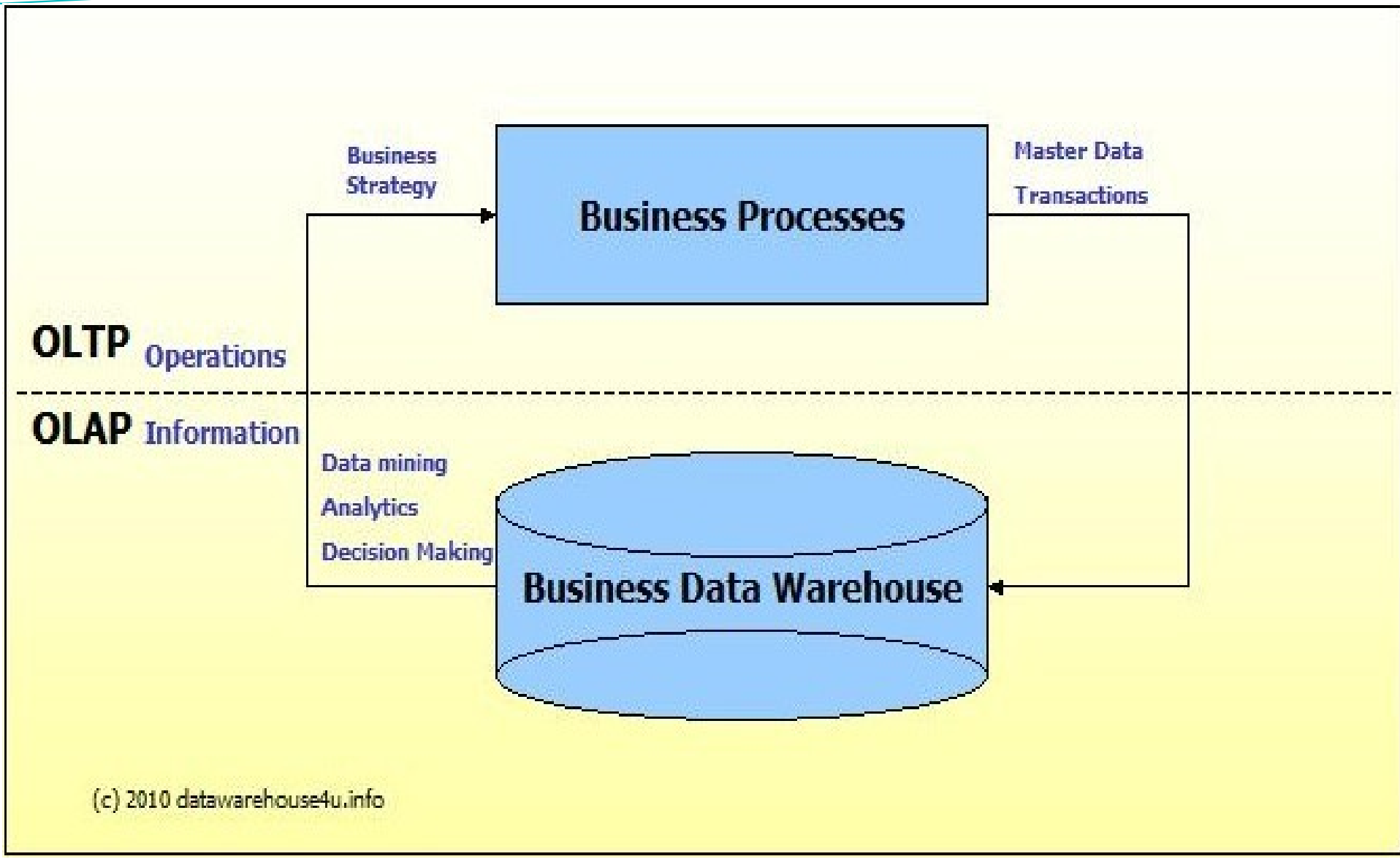
Dimensions: Product, Location, Time
Hierarchical summarization paths



OLTP: On Line Transaction Processing

- Supporting transaction-oriented applications on the Internet.
- Operational Databases are used
- OLTP systems are used for order entry, financial transactions, CRM and retail sales.
- Such systems have a large number of users who conduct short transactions.
- Database queries are usually simple, require sub-second response times and return relatively few records.
- An important attribute of an OLTP system is its ability to maintain concurrency.
- To avoid single point of failure, OLTP systems are often decentralized.

OLTP vs. OLAP


Business Strategy
Business Processes
Master Data
Transactions
OLTP Operations
OLAP Information
Data mining
Analytics
Decision Making
Business Data Warehouse
(c) 2010 datawarehouse4u.info

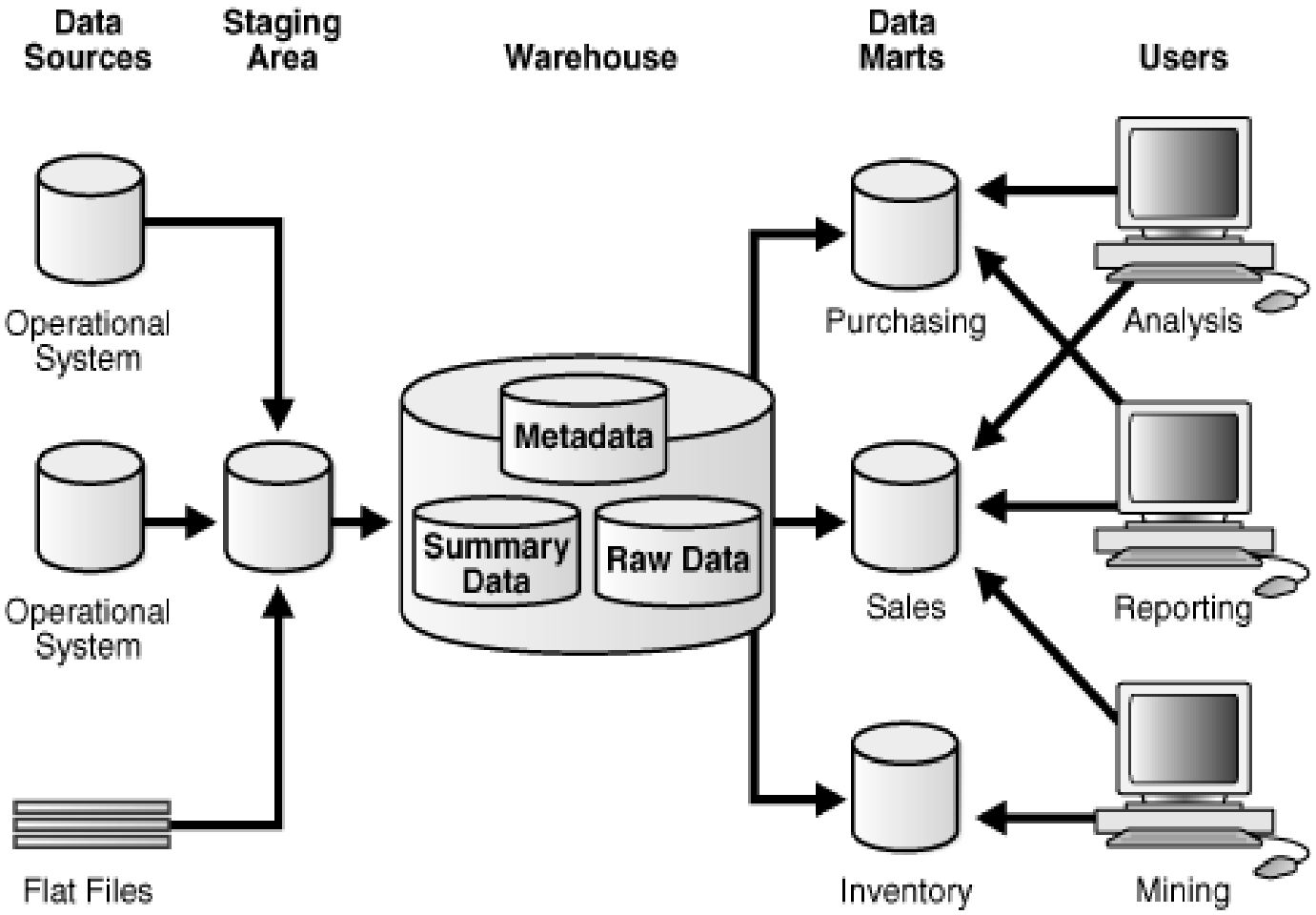
Warehouse vs Operational DB

The usage in Operational DB system is predictable whereas it is unpredictable and random in data warehouse.

Property	Operational DB	Data Warehouse
Response Time	Sub seconds to seconds	Seconds to hours
Operations	DML	Primarily read only
Nature of Data	30-60 days	Snapshots over time
Data Organization	Applications	Subject, time
Size	Small to large	Large to very large
Data Source	Operational, Internal	Operational, Internal, External
Activities	Processes	Analysis

Feature	*OLTP*	*OLAP*
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements, decision support
DB design	ER based, application-oriented	star/snowflake, subject-oriented
Data	current; guaranteed up-to-date	historical; accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	100 MB to GB	100 GB to TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

The Data Mart



The Data Mart

- It is lower-cost, scaled down version of the DW.
- Data Mart offer a targeted and less costly method to exploit the advantages of data warehousing and can be scaled up to a full DW environment over time.
- In addition to the main DW, there can be multiple Data Marts (DM).

The Data Mart

- A data mart is a subset of the data warehouse, which concentrates on a specific business unit.
- Often holds only one subject area- for example, Finance, or Sales
- May hold more summarized data (although many hold full detail)
- A Data Mart is the implementation of a Data Warehouse with a range restricted to a functional area, particular problem, department, subject or group of needs.

The Data Mart

Benefits of data mart:

- Frequently needed data can be accessed very easily.
- Performance improvement.
- Data marts can be created easily.
- Lower cost in implementing data mart than a data warehouse.

Data Warehouses Versus Data Marts

Data Mart	Data Ware House
Here the data collected for analysis is based on single subject	Here the data collected for analysis is based on multiple subjects.
Here each subject area works independently.	Here integration is possible between various subjects.
It occupies less storage area	It occupies more storage area.
It can be normalized	It is highly denormalized
They are developed 6-8 months of duration.	They are developed with in duration of 1-2 years.
Maintenance over head is less	Maintenance over head is more.

Data Warehouse Schema

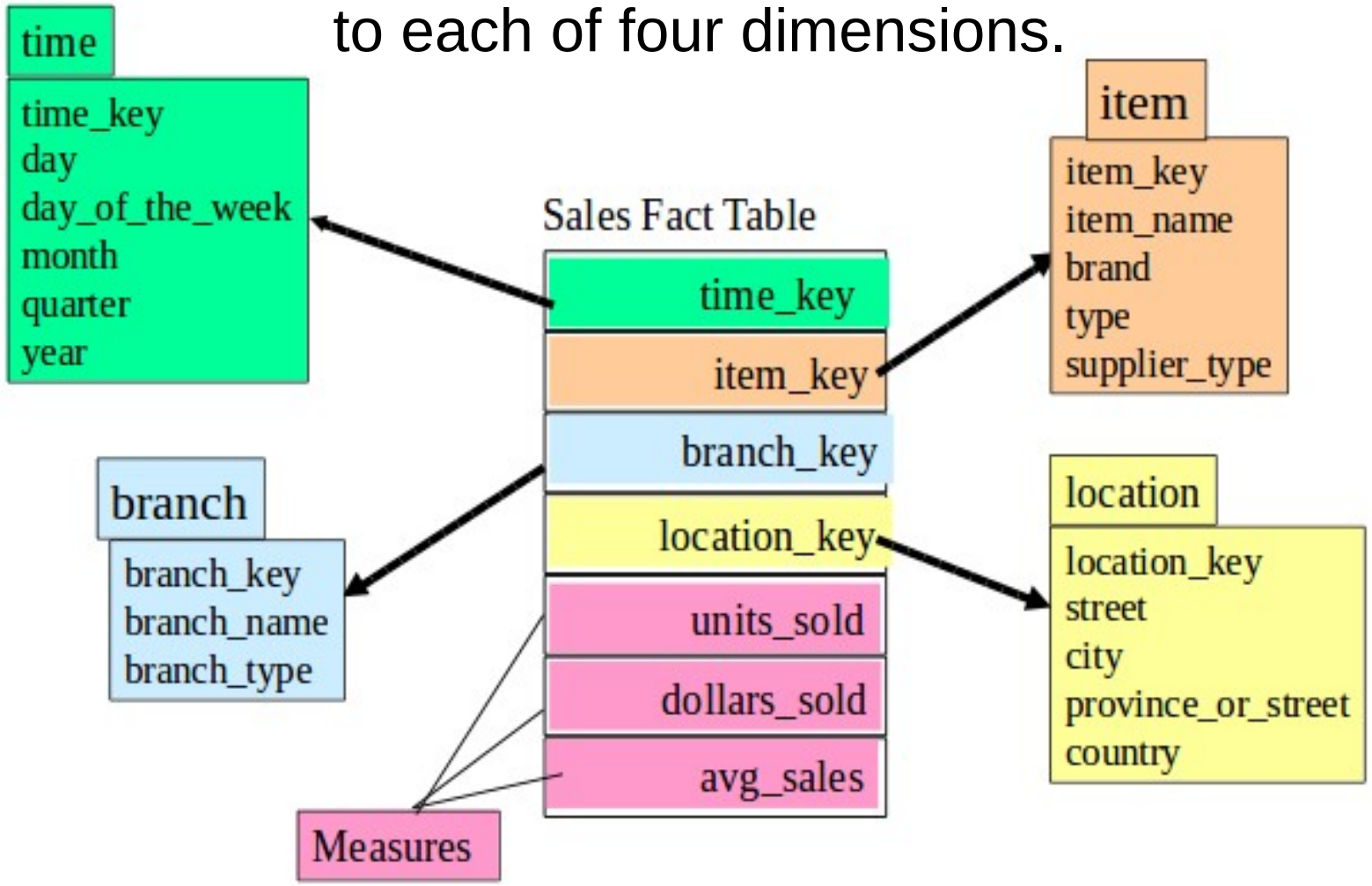
- Schema is a logical description of the entire database.
- It includes the name and description of records, their types and all associated data-items and aggregates.
- Similar to a database, a data warehouse also needs to maintain a schema.
- It provides the Conceptual Modeling of Data Warehouses
- A database uses relational model, while a data warehouse uses Star, Snowflake, and Fact Constellation schema.

Star Schema

- A fact table in the middle connected to a set of dimension tables
- Each dimension in a star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.

Example of Star Schema

There is a fact table at the center. It contains the keys to each of four dimensions.



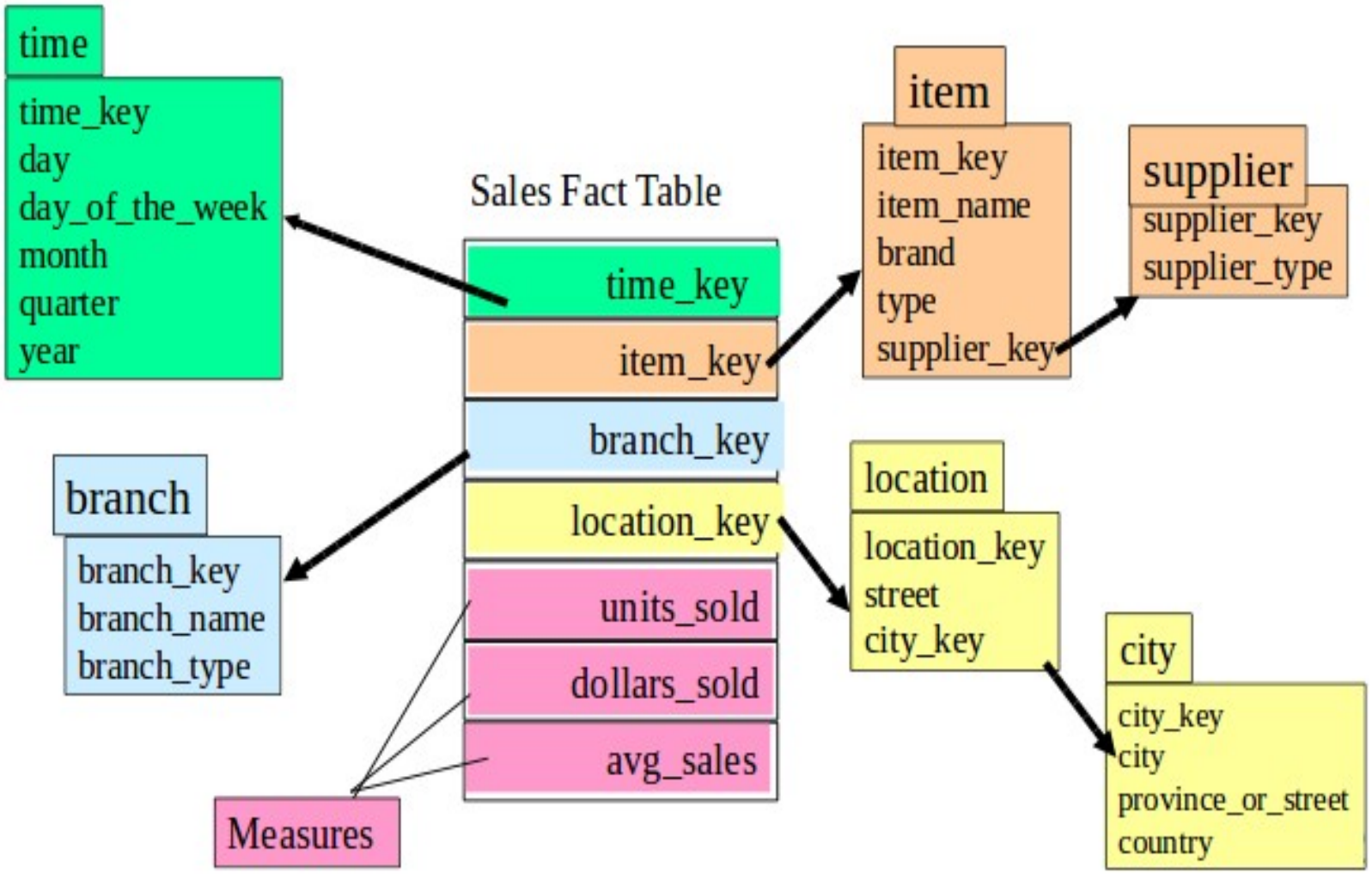
Star Schema

- Each dimension has **only one** dimension table and each table holds a set of attributes.
- For instance, the attributes in the location dimension table: {location_key, street, city, province_or_state,country}.
- This constraint may cause data redundancy.
- For e.g., "Pokhara" and "Baglung" both the cities are in the the province of Western Region in country Nepal.
- The entries for such cities may cause data redundancy along the attributes province_or_state and country.

Snowflake Schema

- A refinement of star schema
- Forming a shape similar to snowflake
- Allows the **normalization** of some dimensions
- The normalization splits up the data into additional tables.
- For example, the **item** dimension table in star schema is normalized and split into two dimension tables, namely **item** and **supplier** table.
- Here, the item dimension table contains the attributes item_key, item_name, type, brand, and supplier-key.
- The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier_key and supplier_type.
- Hence, as a result of normalization in the Snowflake schema, the redundancy is reduced and therefore, it becomes easy to maintain and the save storage space.

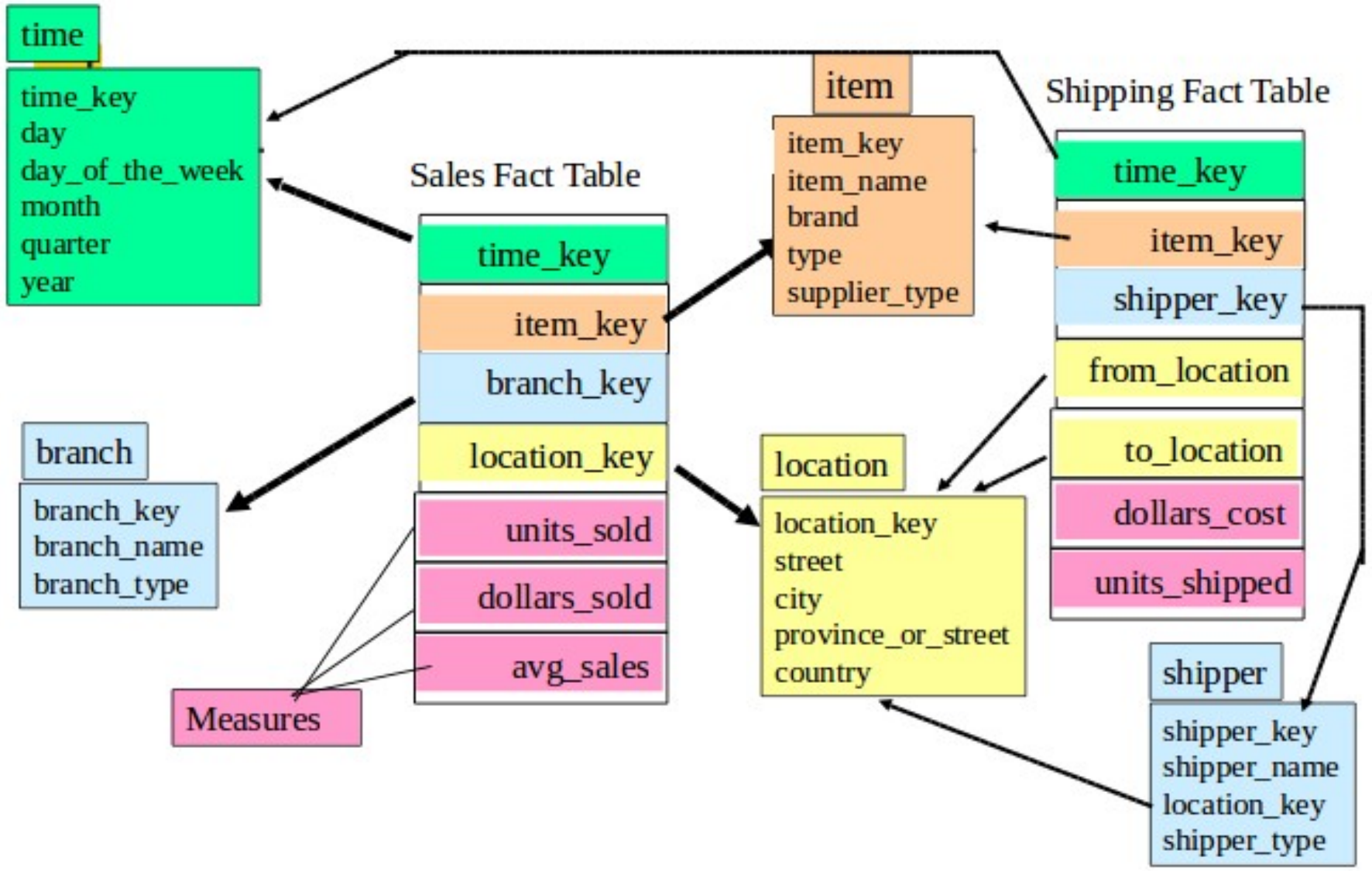
Example of Snowflake Schema



Fact Constellations Schema

- Multiple fact tables share dimension tables
- Viewed as a collection of stars and called galaxy schema
- For example, the previous schema is modified to shows two fact tables, namely **sales** and **shipping**.
- The sales fact table is same as that in the star schema.
- The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location.
- The shipping fact table also contains two measures, namely dollars sold and units sold.
- It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table.

Example of Fact Constellation



Data Warehouse Backend Tools

- use back-end tools to populate and refresh the data:
  - **Data extraction:** gathers data from multiple, heterogeneous, and external sources
  - **Data cleaning: d**etects errors in the data and rectifies them when possible
  - **Data transformation:** converts data from legacy or host format to warehouse format
  - **Load:** sorts, summarizes, consolidates, computes views, checks integrity, and builds indices and partitions
  - **Refresh:** propagates the updates from the data sources to the warehouse

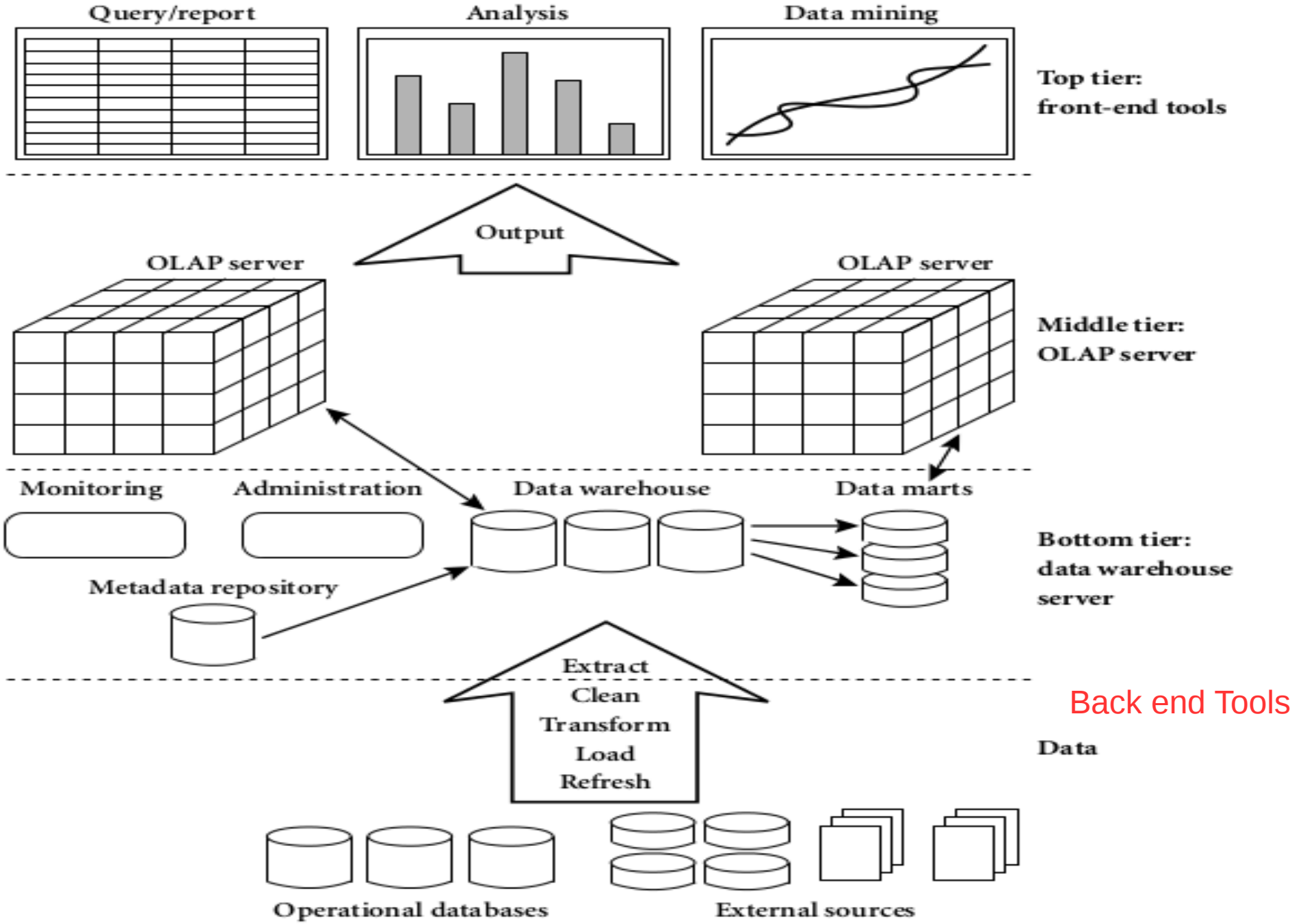


Fig: A three-tier data warehousing architecture with front end and back end tools.

Data Warehouse Design Process

- **Guideline**:
  - take all the enterprise data and build a data warehouse, so that the management can get answers to their questions
  - start somewhere and get going
  - common technique is to develop a datamart and gradually blow it to a full fledged data warehouse.

Data Warehouse Design Process

Ralph Kimball strategy to build a data mart:

1. Choose the subject matter (one subject at a time)
2. Decide what the fact table represents
3. Identify and conform the dimensions
4. Choose the facts
5. Store pre-calculations in the fact table
6. Define the dimensions and tables.
7. Decide the duration of the database and the periodicity of updation
8. Track slowly the changing dimensions
9. Decide the query priorities and query models
10. Build a few simple data marts and
11. Integrate them in stages

References:

- Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*, 2nd Edition, 2006, Morgan Kaufmann.
- Data Mining By Pieter Adriaans, Dolf Zantinge
- Introduction to Data Mining by Tan, Steinbach, Kumar
- http://web.calstatela.edu/faculty/ppartow/warehouse_files/warehouse_chapter3.ppt
- Data Warehousing and Data Mining, S. Sudarshan and Krithi Ramamritham, IIT Bombay
- http://www.cse.buffalo.edu/DBGROUP/nachi/ecopres/kalyani.ppt
- http://www.tutorialspoint.com/dwh/dwh_schemas.htm
- https://docs.oracle.com/database/121/DWHSG/img/dwhsg064.gif