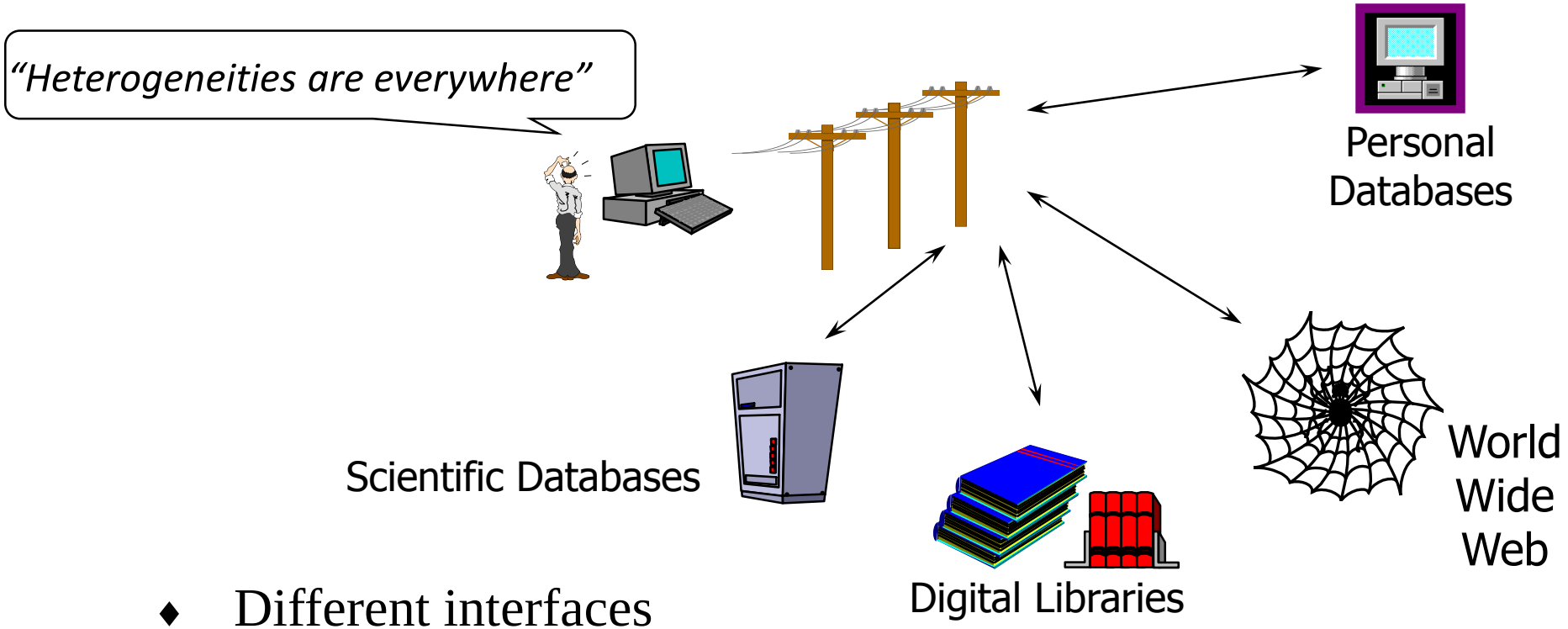
Data Mining and Data Warehousing

Chapter 2

Data warehousing

Instructor: Suresh Pokharel
ME in ICT (Asian Institute of Technology, Thailand)
BE in Computer (NCIT, Pokhara University)

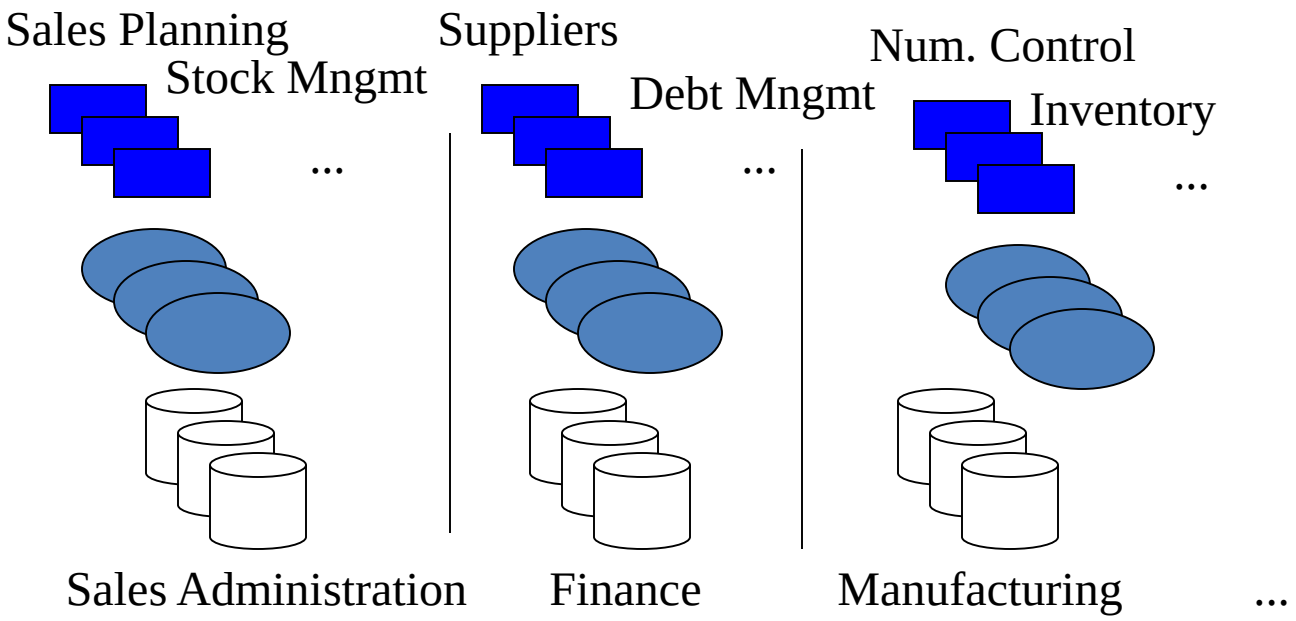
Problem: Heterogeneous Information Sources



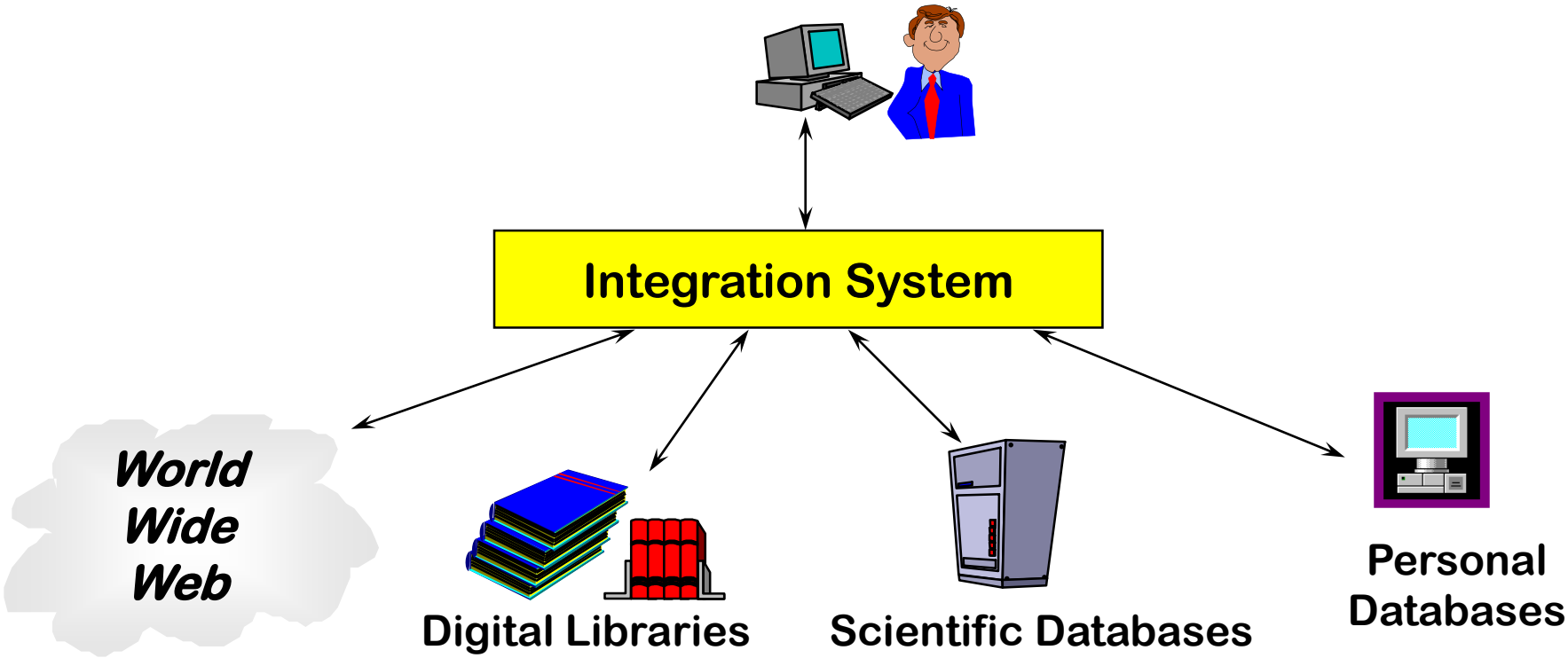
- Different interfaces
- Different data representations
- Duplicate and inconsistent information

Problem: Data Management in Large Enterprises

- Vertical fragmentation of informational systems (vertical stove pipes)

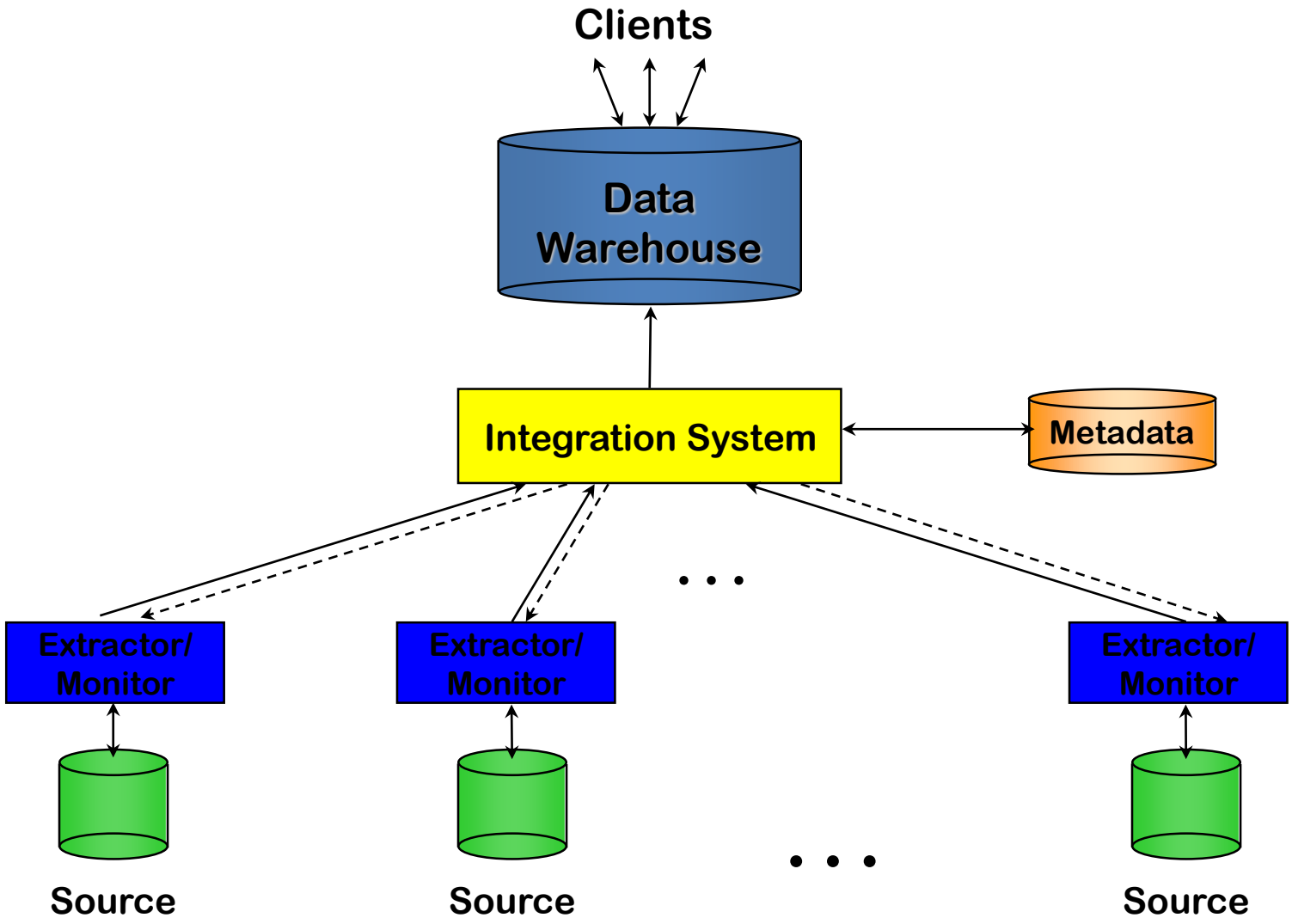


Goal: Unified Access to Data

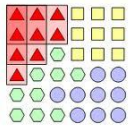


- Collects and combines information
- Provides integrated view, uniform user interface
- Supports sharing

The Warehouse

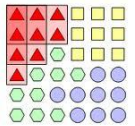


What is Data Warehouse?



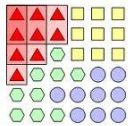
- Defined in many different ways:
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon
- Data warehousing:
  - The process of constructing and using data warehouses

Data Warehouse—Subject-Oriented



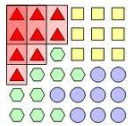
- Organized around major subjects, such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Data Warehouse—Integrated



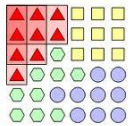
- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

Data Warehouse—Time Variant



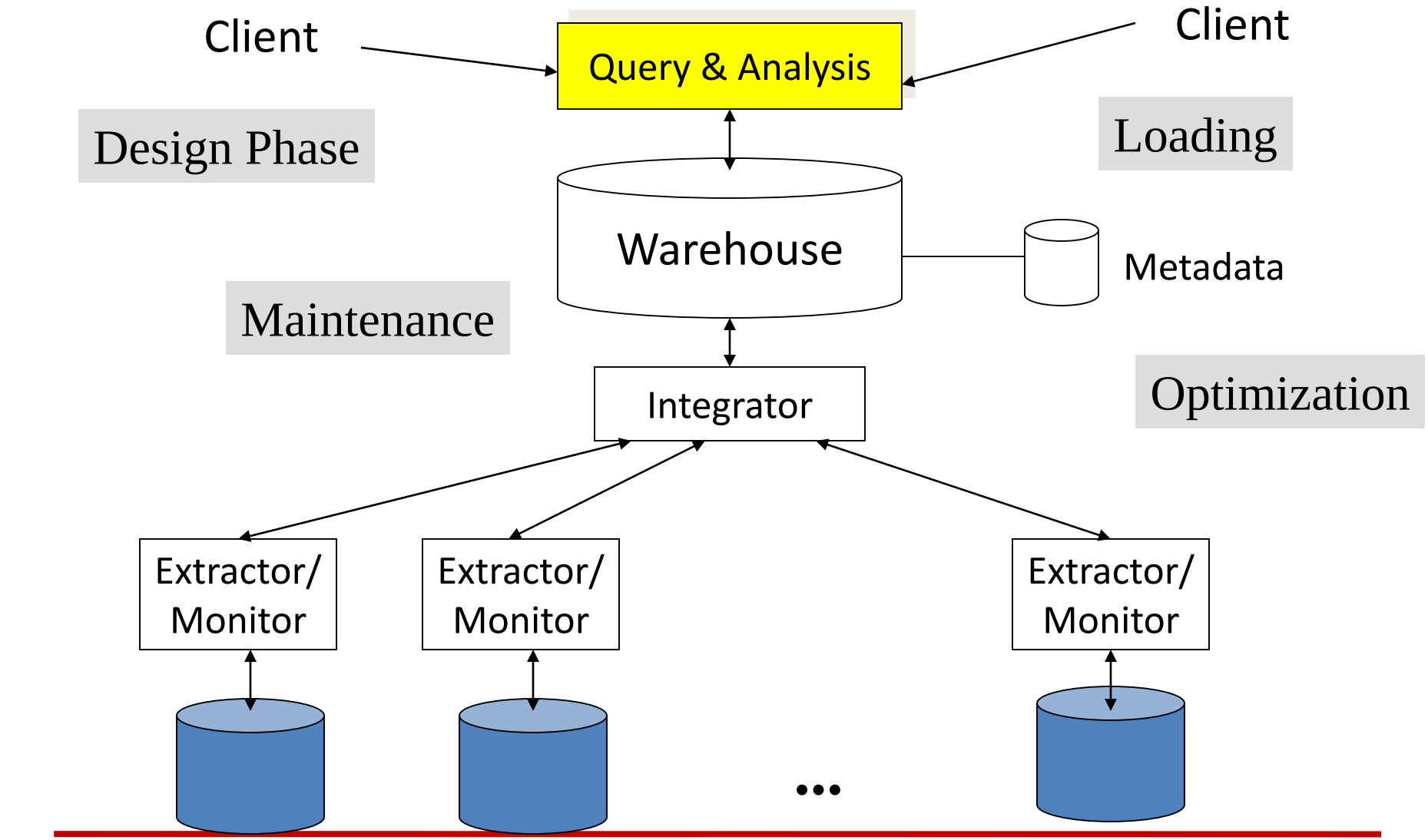
- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)

Data Warehouse—Non-Volatile

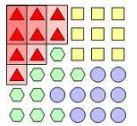


- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*.

Generic Warehouse Architecture



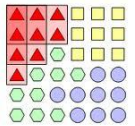
Data Warehousing: Two Distinct Issues



1. How to get information into warehouse *"Data warehousing"*
2. What to do with data once it's in warehouse *"Warehouse DBMS"*

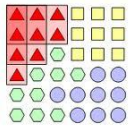
- Both rich research areas
- Industry has focused on 2

Data Warehouse & Database



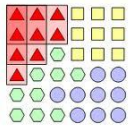
	Data Warehouse	Database
Purpose	Analysis, Decision making	Day to day use
Support For	OLAP(on-line analytical processing)	OLTP(on-line transaction processing)
Data model	Multi-dimentional	Rational
Age of data	Current & time series	Current & real time
Data modification	Read/access only	Insert, update, delete
Type of data	Static	Dynamic
Amount of data per transaction	Larger	Smaller
Schema design	Denormalization	normalization

Data Warehousing: Benefit



- Provides organizing framework
- Allows simplified maintenance
- Speeds up future development by aiding understanding of DM
- Communication tool for roles and requirements
- Coordinate data marts

Data Warehouse and data Mart



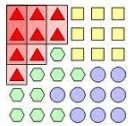
Data warehouse: **enterprise based,** collects all information about subjects (*customers, products, sales, assets, personnel*) that span the entire organization

- Concerns with decision subjects of the whole enterprise or organization
- Requires extensive business modeling (may take years to design and build)

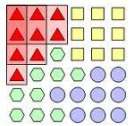
Data mart: **department based,** Departmental subsets that focus on selected subjects

- Specialized single line of business warehouses e.g. within departments or groups of people
- Marketing data mart: customer, product, sales

Decision Support System



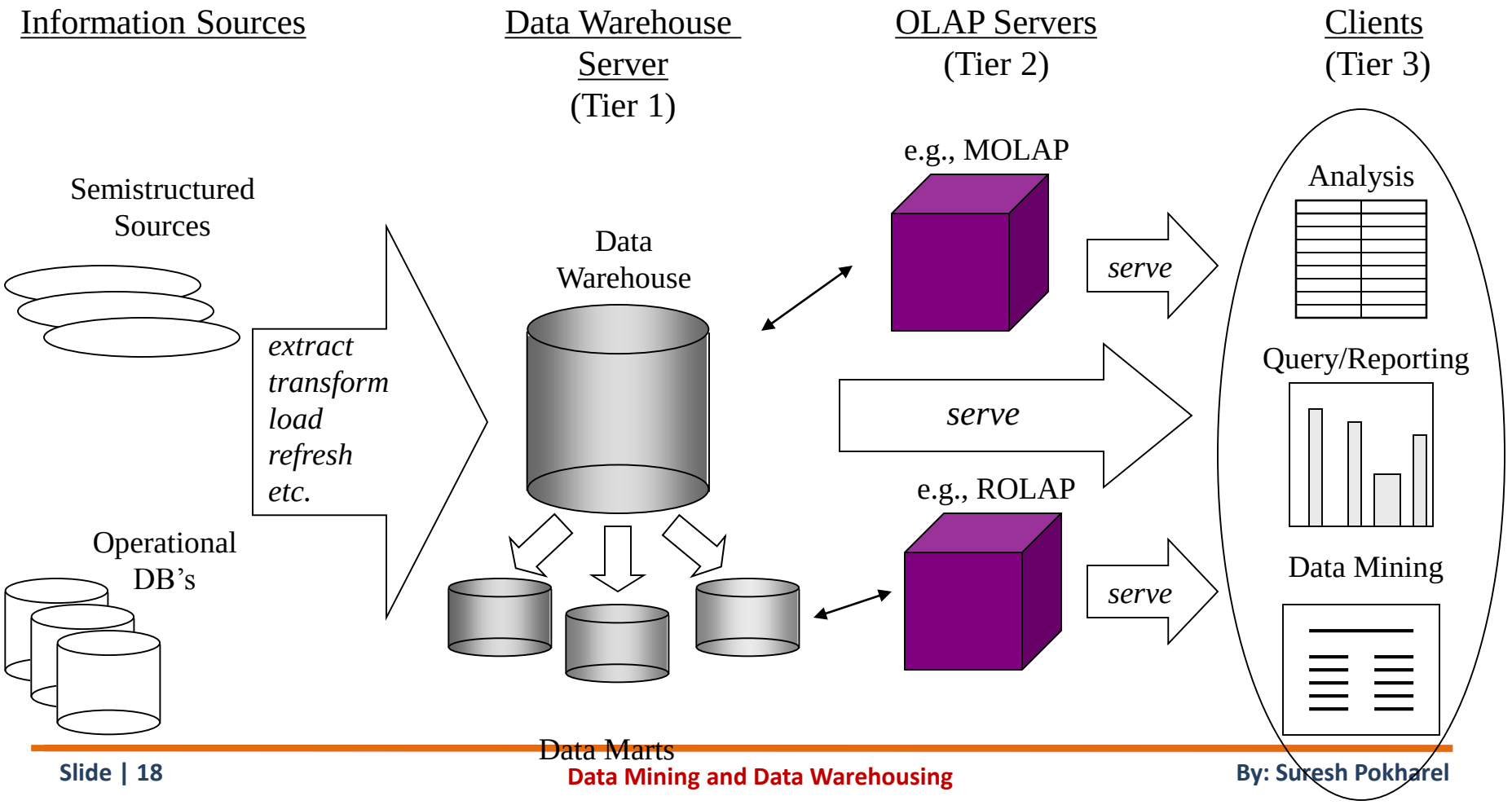
- Information technology to help the knowledge worker (executive, manager, analyst) make faster & better decisions
  - *"What were the sales volumes by region and product category for the last year?"*
  - *"How did the share price of comp. manufacturers correlate with quarterly profits over the past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



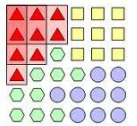
Three-Tier Decision Support Systems

- Warehouse database server
  - Almost always a relational DBMS, rarely flat files
- OLAP servers(p.p.135)
  - 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



Approaches to OLAP 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
- have greater scalability

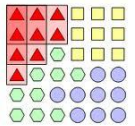
(2) Multidimensional OLAP (MOLAP)

- Array-based storage structures
- Direct access to array data structures
- Fast indexing to pre-computed summarized data

(3) Hybrid OLAP (HOLAP) server

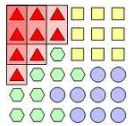
- Combine both ROLAP and MOLAP
- E.g. Microsoft SQL Server 2000

Data Preprocessing

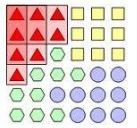


- Real world data : Noisy, missing and inconsistent (why??)
- Low quality data => Low quality mining result
- Data Cleaning
- Data integration
- Data transformations
- Data reduction

Data Cleaning



- Missing values
  - No record value for several attributes such as income
  - How can fill missing data?
  - E.g. manually, fill with mean, fill with probable
- Noisy Data
  - containing errors, or outlier values
  - How can smooth data ?
  - E.g. Binning, regression, clustering



Binning

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

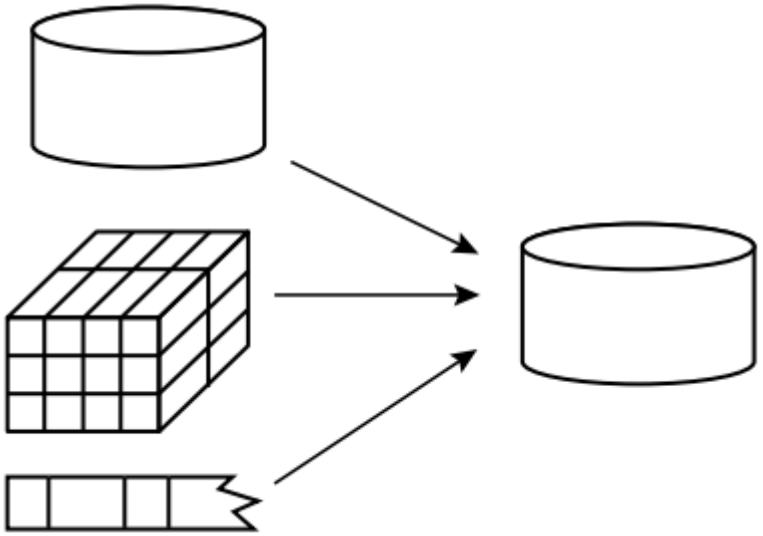
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

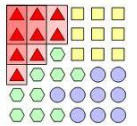
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Data Integration

- Combines data from multiple sources(e.g. databases, data cubes or flat files) into data warehouse



Data Transformation



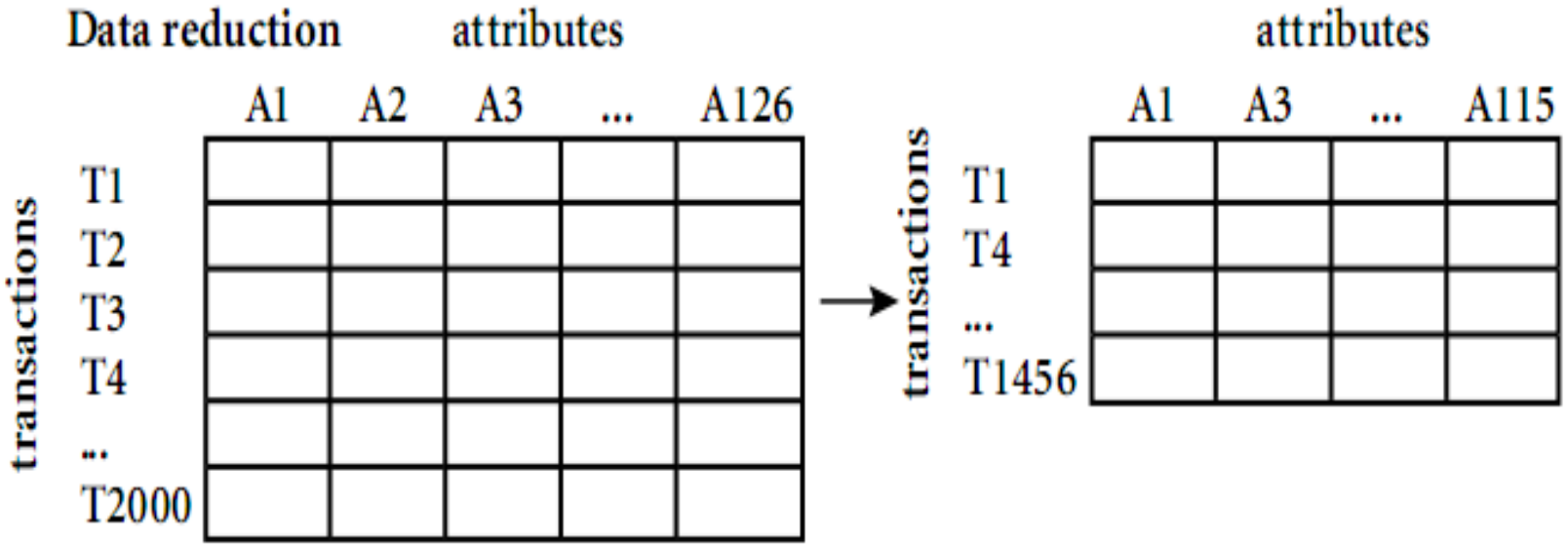
- Data transforms into appropriate form for mining
- Some of the methods:
- Smoothing: remove noise
- Aggregation : summary or aggregation operations are applied to the data.
- Generation : low-level =>high level concepts e.g. age => youth, middle-aged, senior
- Normalization : attribute data are scaled into specified range such as -1.0 to 1.0 or 0.0 to 1.0 (e. g. how??)

  e.g. $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

- Attribute construction : New features are constructed and added from the given set of attributes to help the mining process

Data Reduction

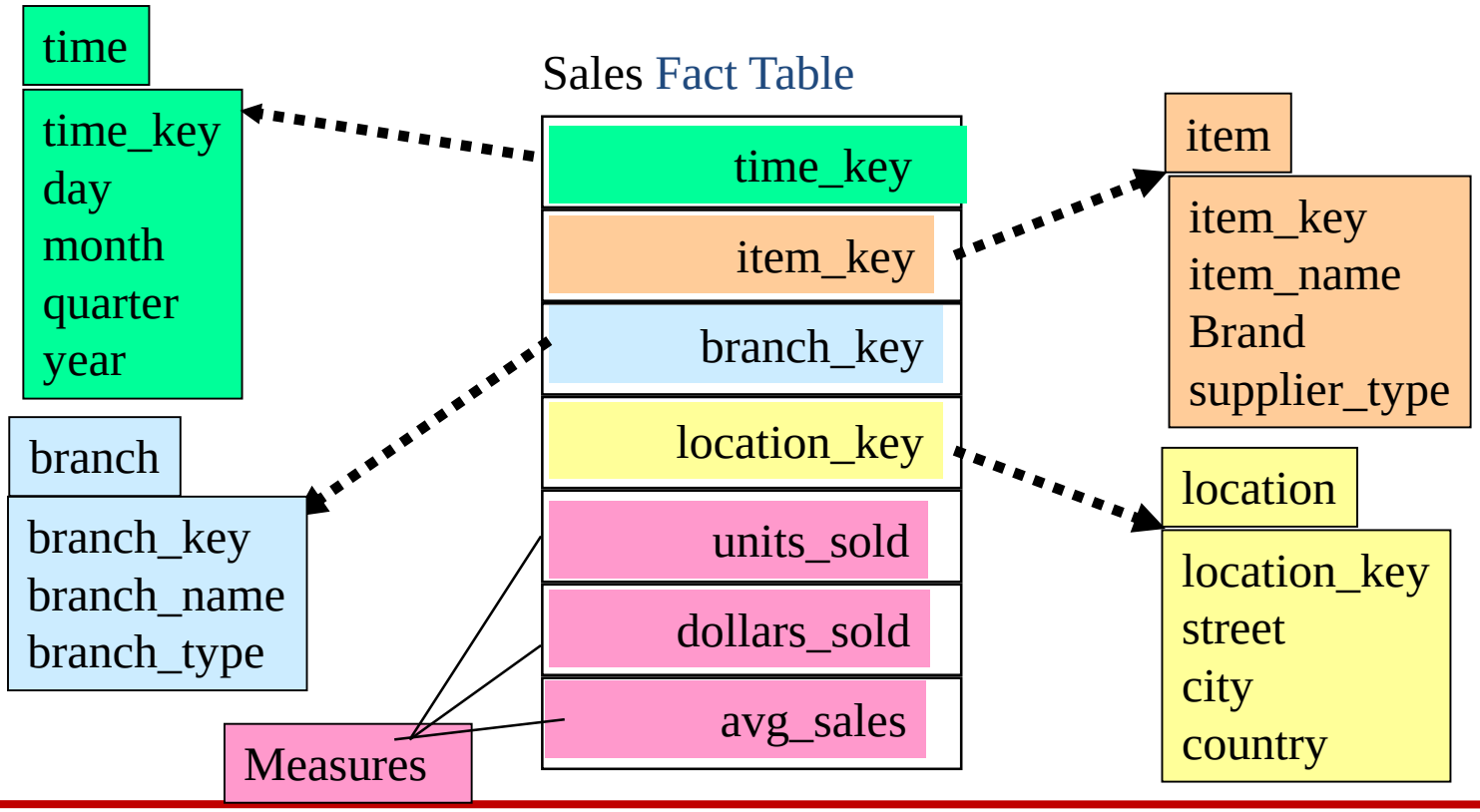
- Goal : Making mining process more efficient with out losing quality
- E.g.



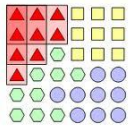
Conceptual Modeling of DW

- Dimensions & Measures

Star schema: A fact table in the middle connected to a set of dimension tables

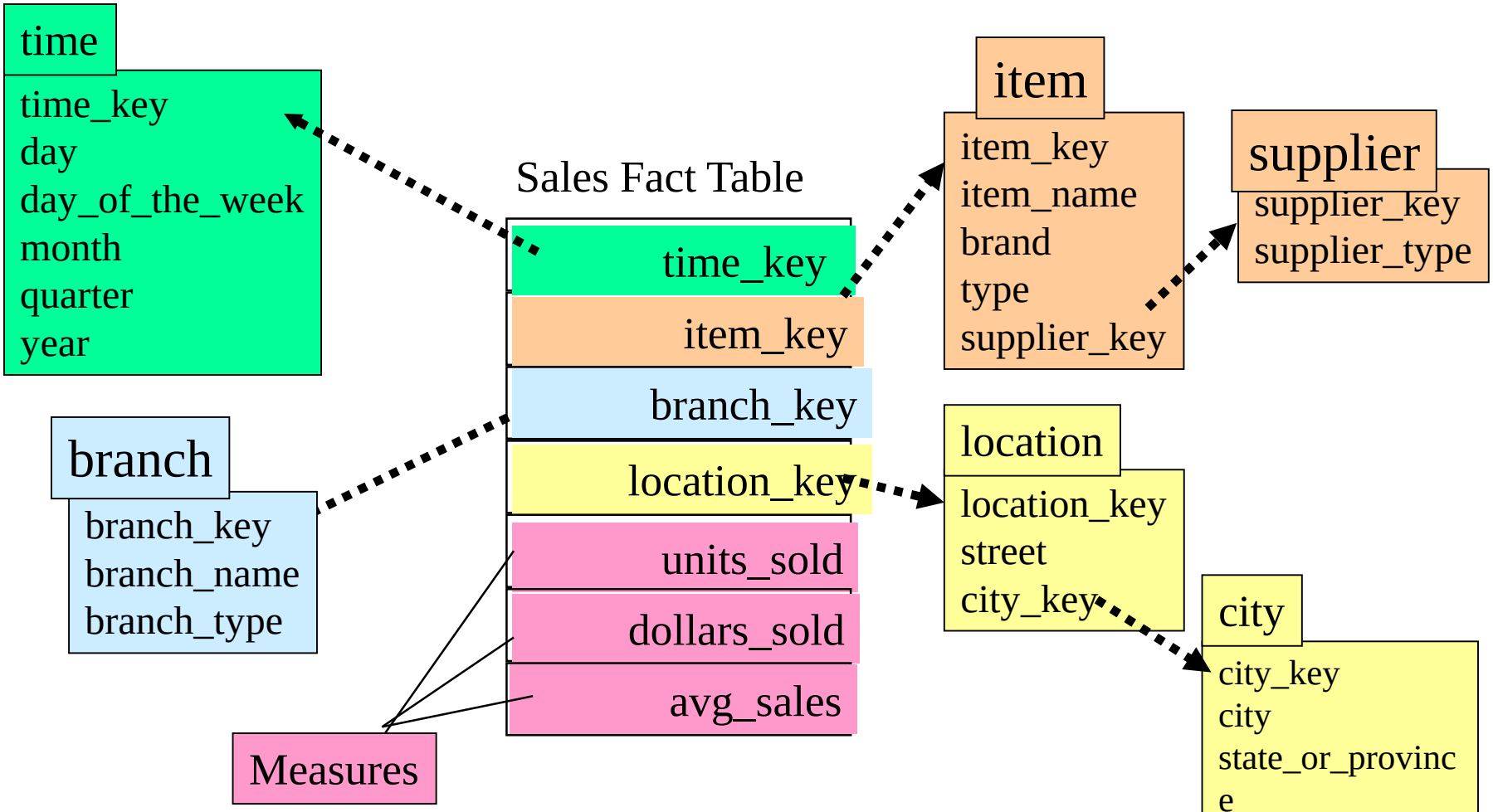


Conceptual Modeling of DW

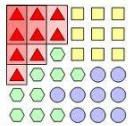


Snowflake schema

A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake.

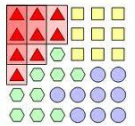

time
time_key
day
day_of_the_week
month
quarter
year
Sales Fact Table
time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales
item
item_key
item_name
brand
type
supplier_key
supplier
supplier_key
supplier_type
branch
branch_key
branch_name
branch_type
location
location_key
street
city_key
city
city_key
city
state_or_provinc
e
country
Measures

Conceptual Modeling of DW



Fact constellations:

Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation



time

time_key
day
day_of_the_week
month
quarter
year

item

item_key
item_name
brand
type
supplier_type

Shipping Fact Table

time_key
item_key
shipper_key
from_location
to_location
dollars_cost
units_shipped

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

branch

branch_key
branch_name
branch_type

location

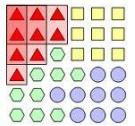
location_key
street
city
province_or_state
country

shipper

shipper_key
shipper_name
location_key
shipper_type

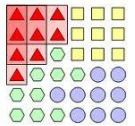
Measures

Data Discretization



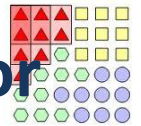
- **Three types of attributes:**
  - Nominal — values from an unordered set, e.g., color, profession
  - Ordinal — values from an ordered set, e.g., military or academic rank
  - Continuous — real numbers, e.g., integers or real numbers
- **Data discretization:**
  - Divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization
  - Prepare for further analysis

Discretization and Concept Hierarchy



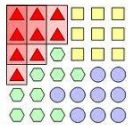
- **Discretization**
  - Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Supervised (use class information) vs. Unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
- **Concept hierarchy formation**
  - Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for age) by higher level concepts (such as young, middle-aged, or senior)

Discretization and Concept Hierarchy Generation for Numeric Data



- **Typical methods:**
  - **Binning**
  - Entropy-based discretization: supervised, top-down split

Entropy-Based Discretization



- Given a set of samples S, if S is partitioned into two intervals S_1 and S_2 using boundary T, the information gain after partitioning is

$$I(S,T) = \frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2)$$

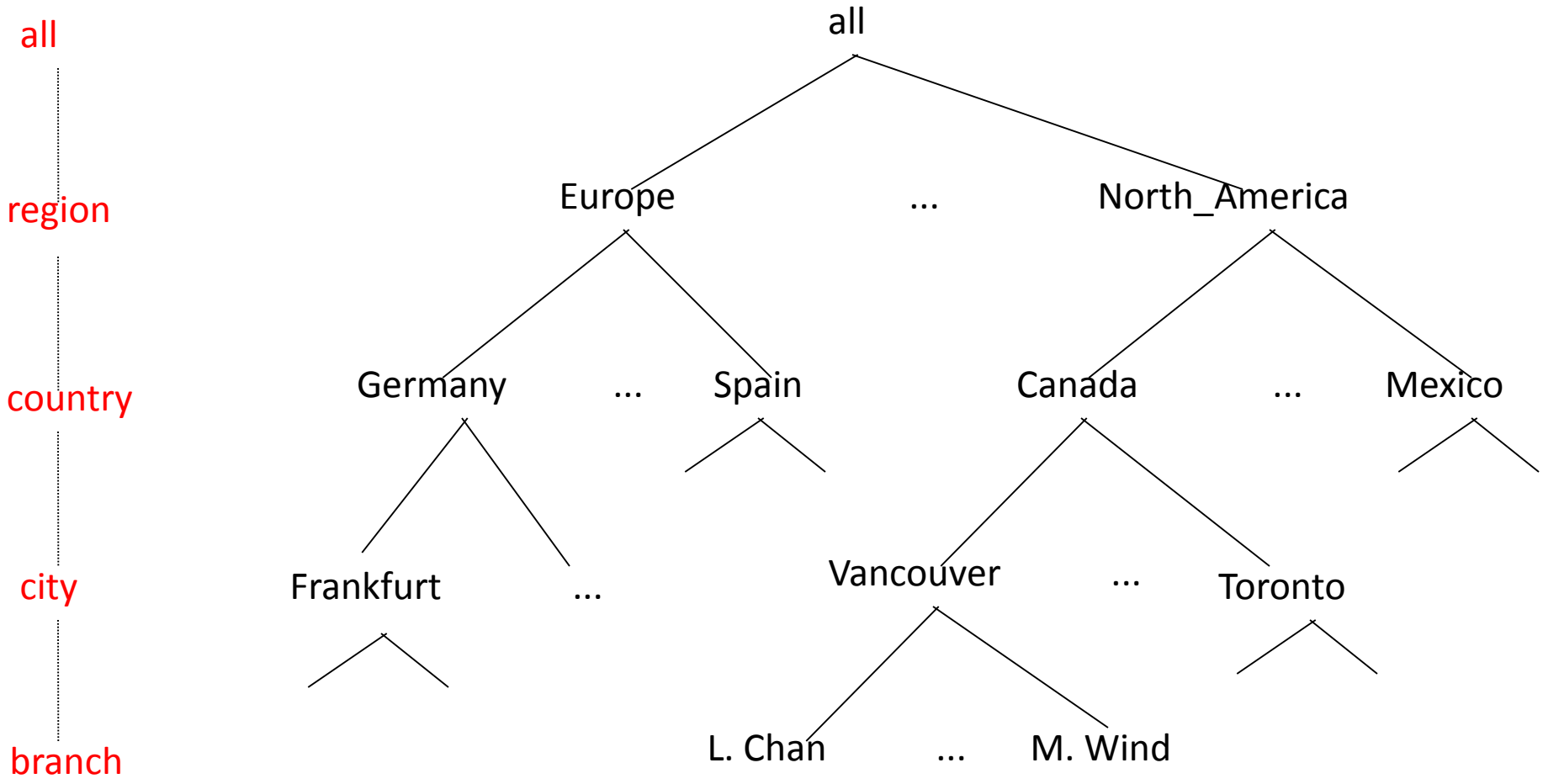
- Entropy is calculated based on class distribution of the samples in the set. Given *m* classes, the entropy of S_1 is

$$Entropy(S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

where p_i is the probability of class *i* in S_1

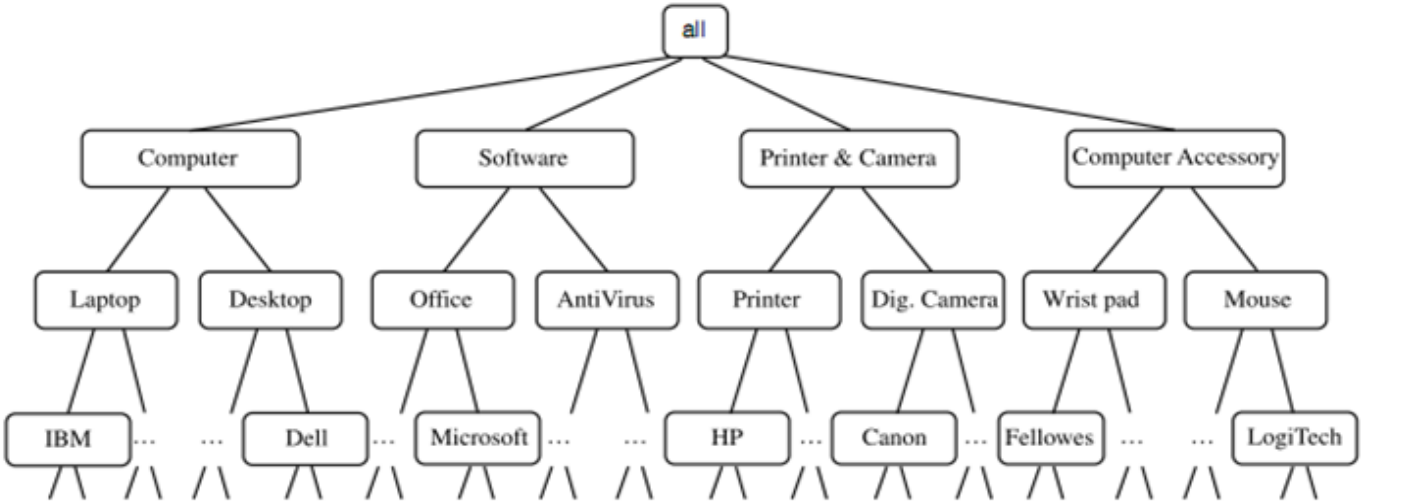
- The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization
- The process is recursively applied to partitions obtained until some stopping criterion is met
- Such a boundary may reduce data size and improve classification accuracy

A Concept Hierarchy: Dimension (location)



A Concept Hierarchy

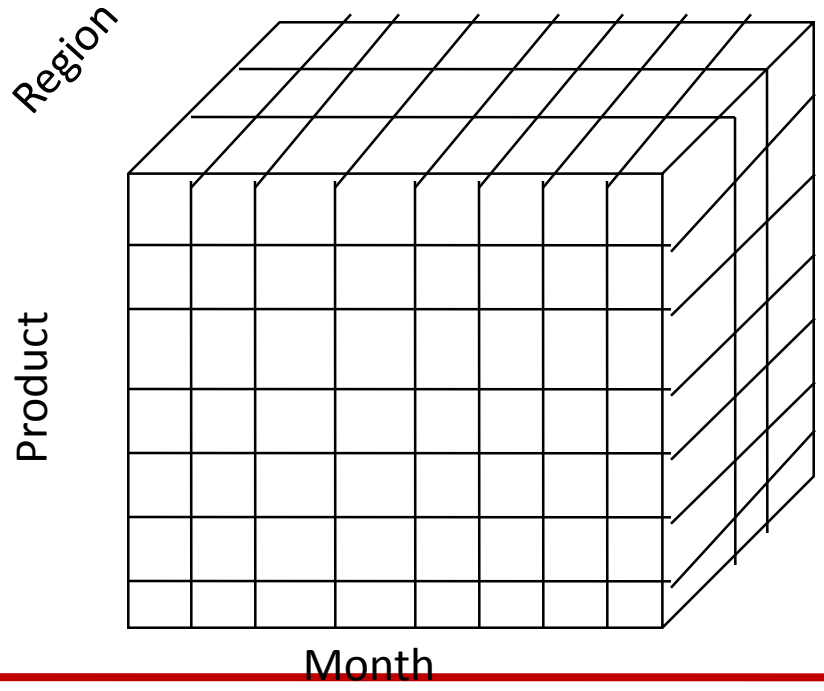
TID	Items Purchased
T100	IBM-ThinkPad-T40/2373, HP-Photosmart-7660
T200	Microsoft-Office-Professional-2003, Microsoft-Plus!-Digital-Media
T300	Logitech-MX700-Cordless-Mouse, Fellowes-Wrist-Rest
T400	Dell-Dimension-XPS, Canon-PowerShot-S400
T500	IBM-ThinkPad-R40/P4M, Symantec-Norton-Antivirus-2003
...	...

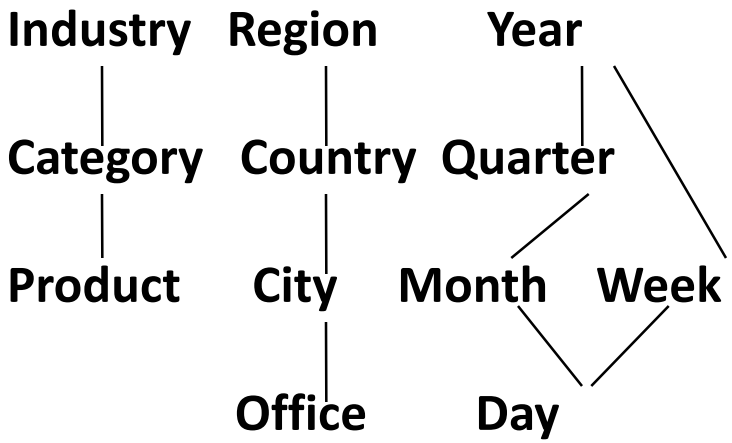


Multidimensional Data

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

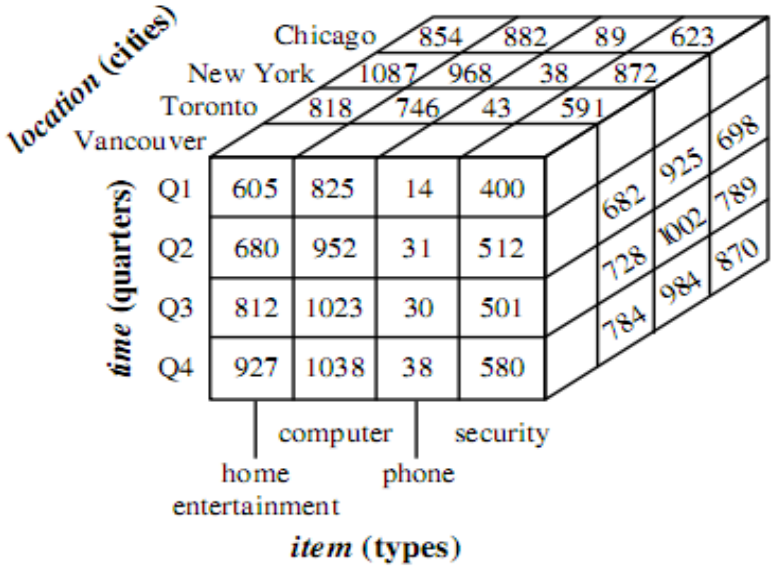
Dimensions: Product, Location, Time
Hierarchical summarization paths



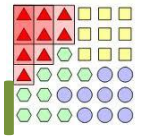


3-D data cube representation from table

	location = "Chicago"				location = "New York"				location = "Toronto"				location = "Vancouver"			
	item				item				item				item			
time	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

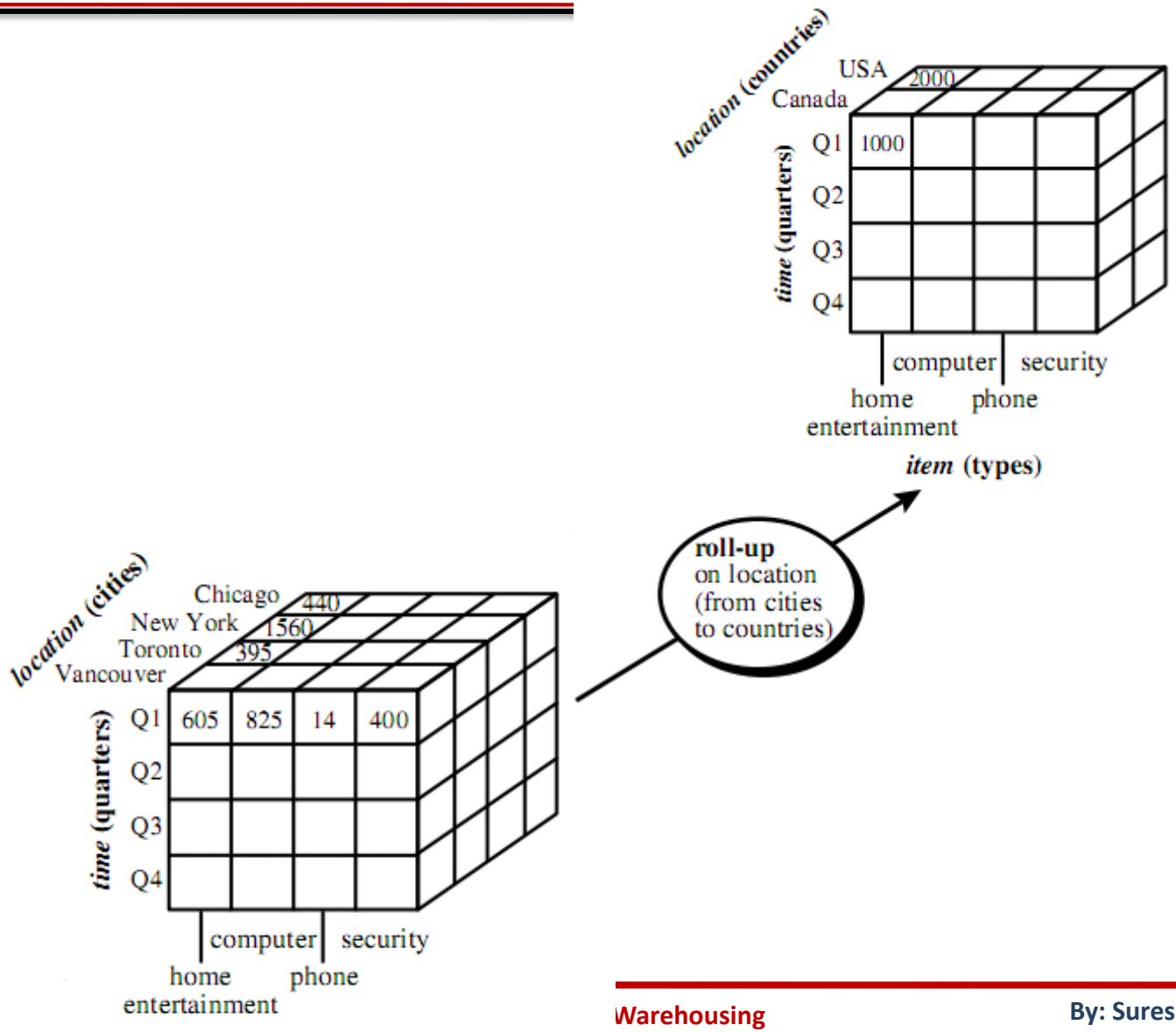


OLAP Operations in Multidimensional Data Model

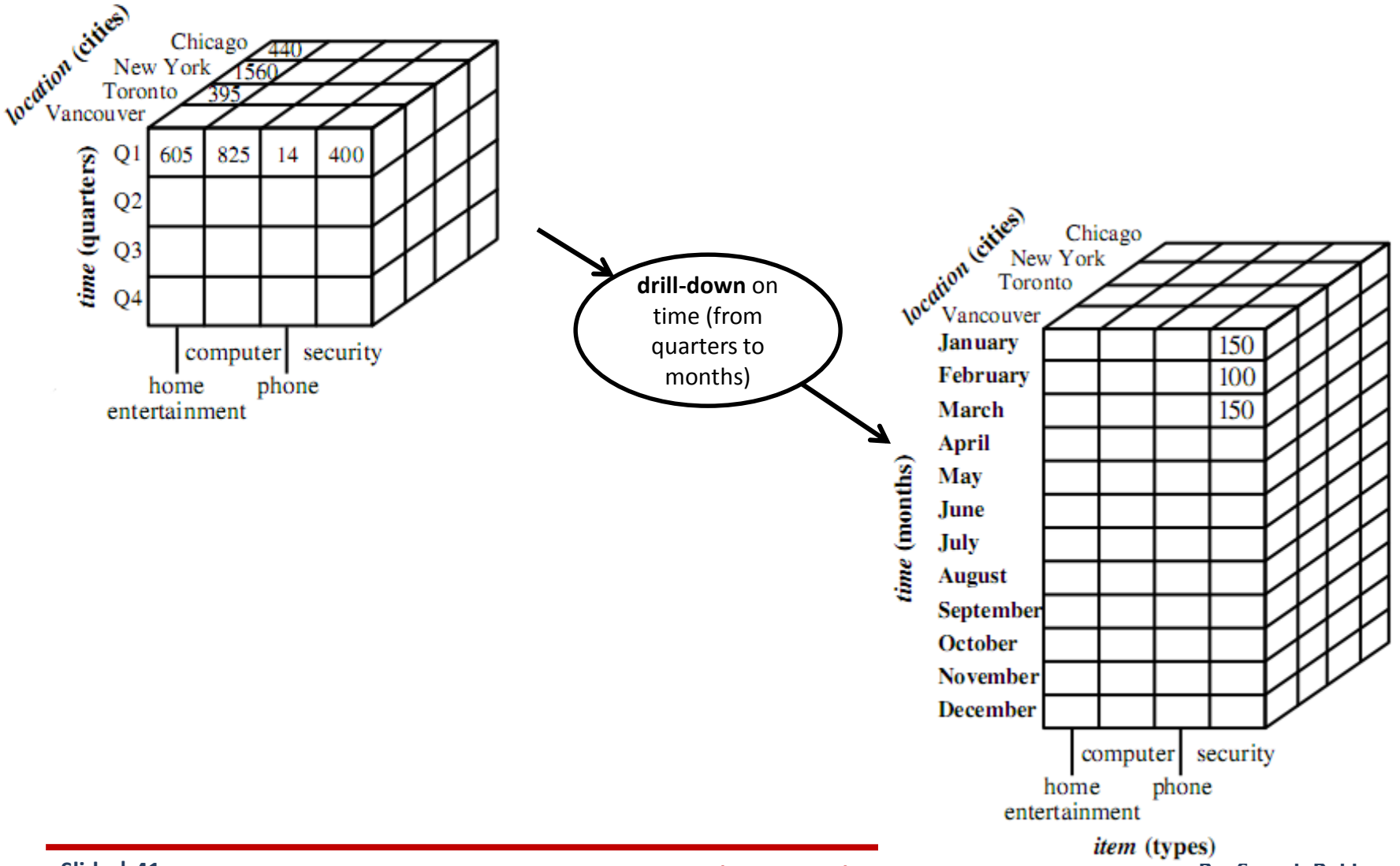


- Roll-up :
- Drill- down :
- Slice and dice :
- Pivot (rotate) :

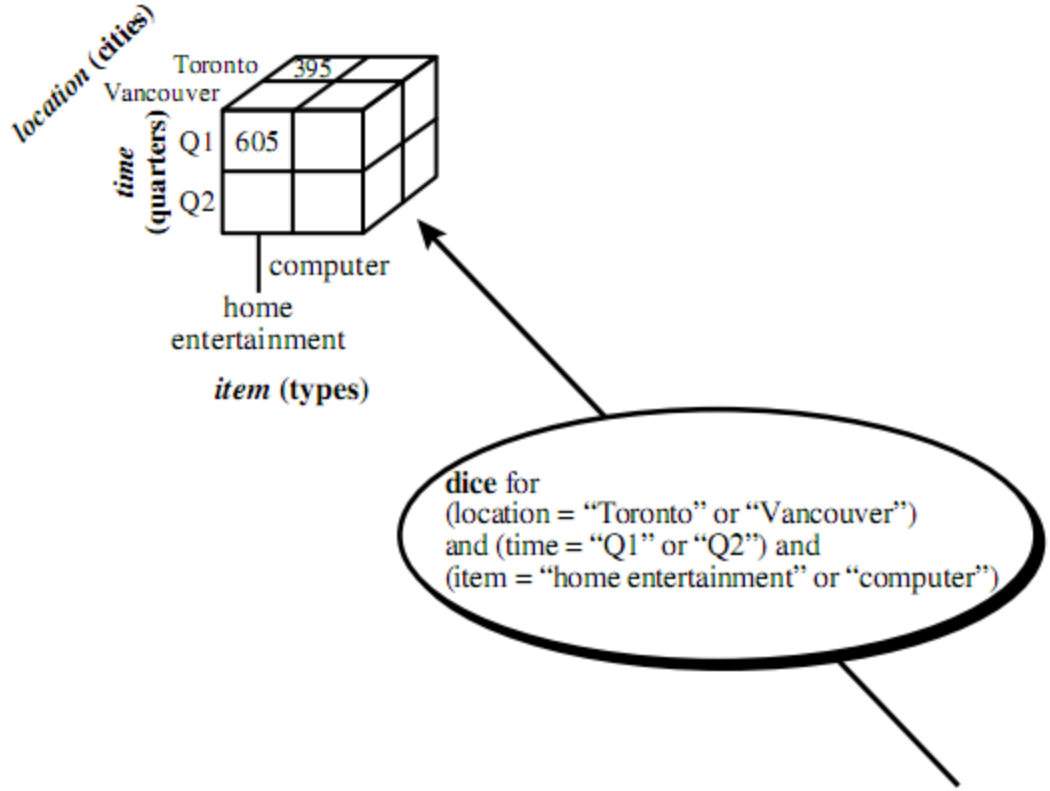
OLAP Operations : roll-up

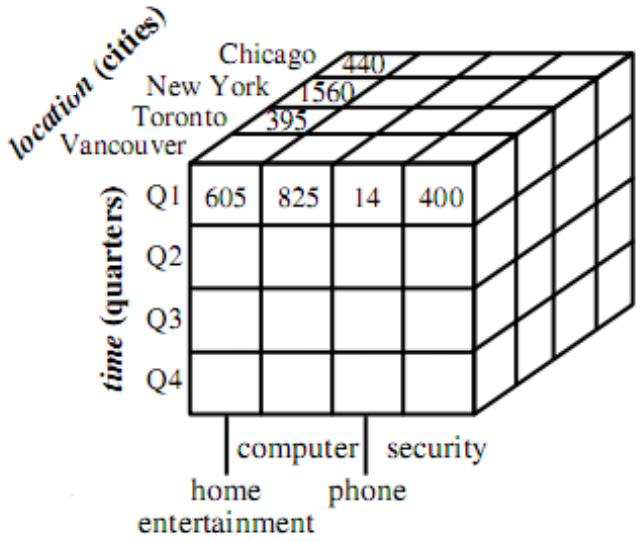


OLAP Operations : drill-down

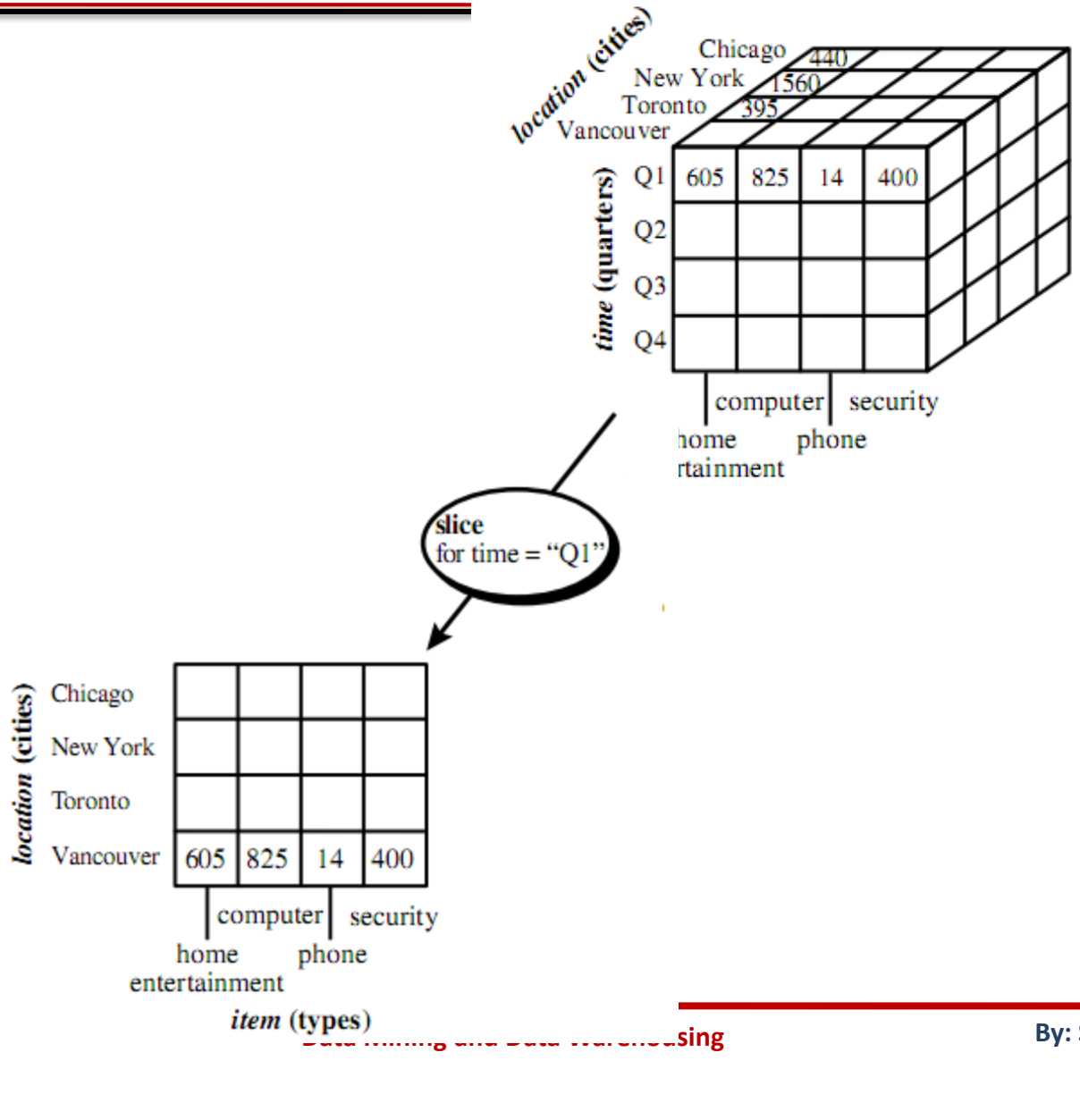


OLAP Operations : dice

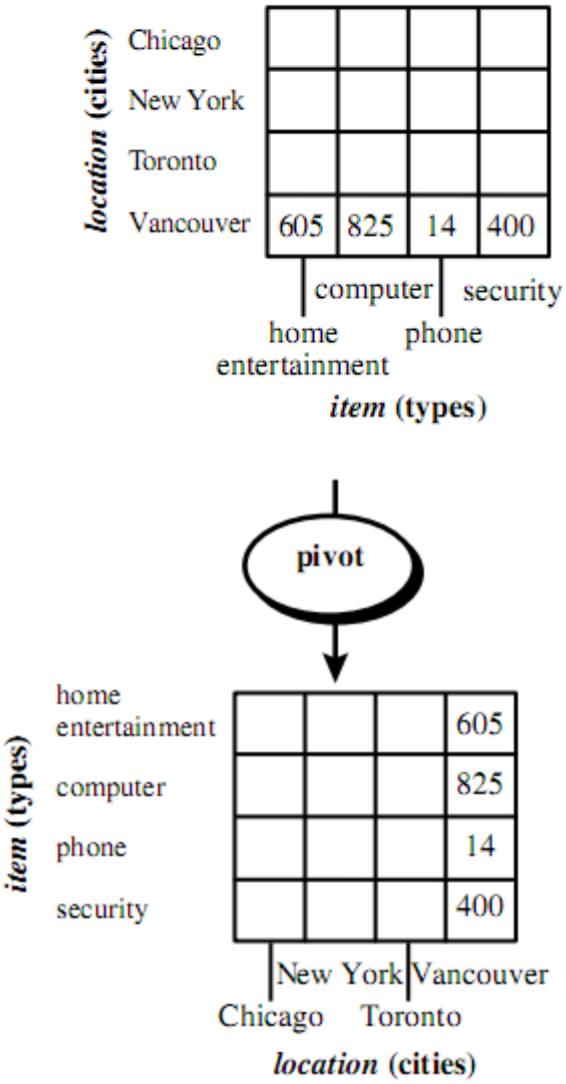


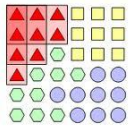


OLAP Operations : slice



OLAP Operations : pivot





- Review
  - Data mining definitions, applications, issues, classifications
  - Data warehouse, architecture, benefits, DSS, preprocessing
  - data cube, OLAP operations
- Read Chapter 1, 2, 3
- Questions