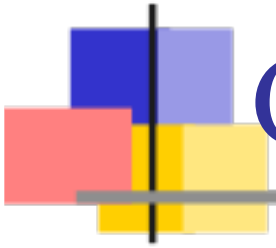
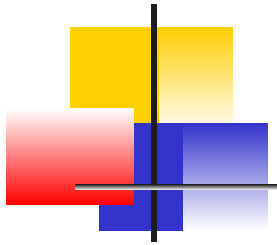


Ch3 Data Warehouse

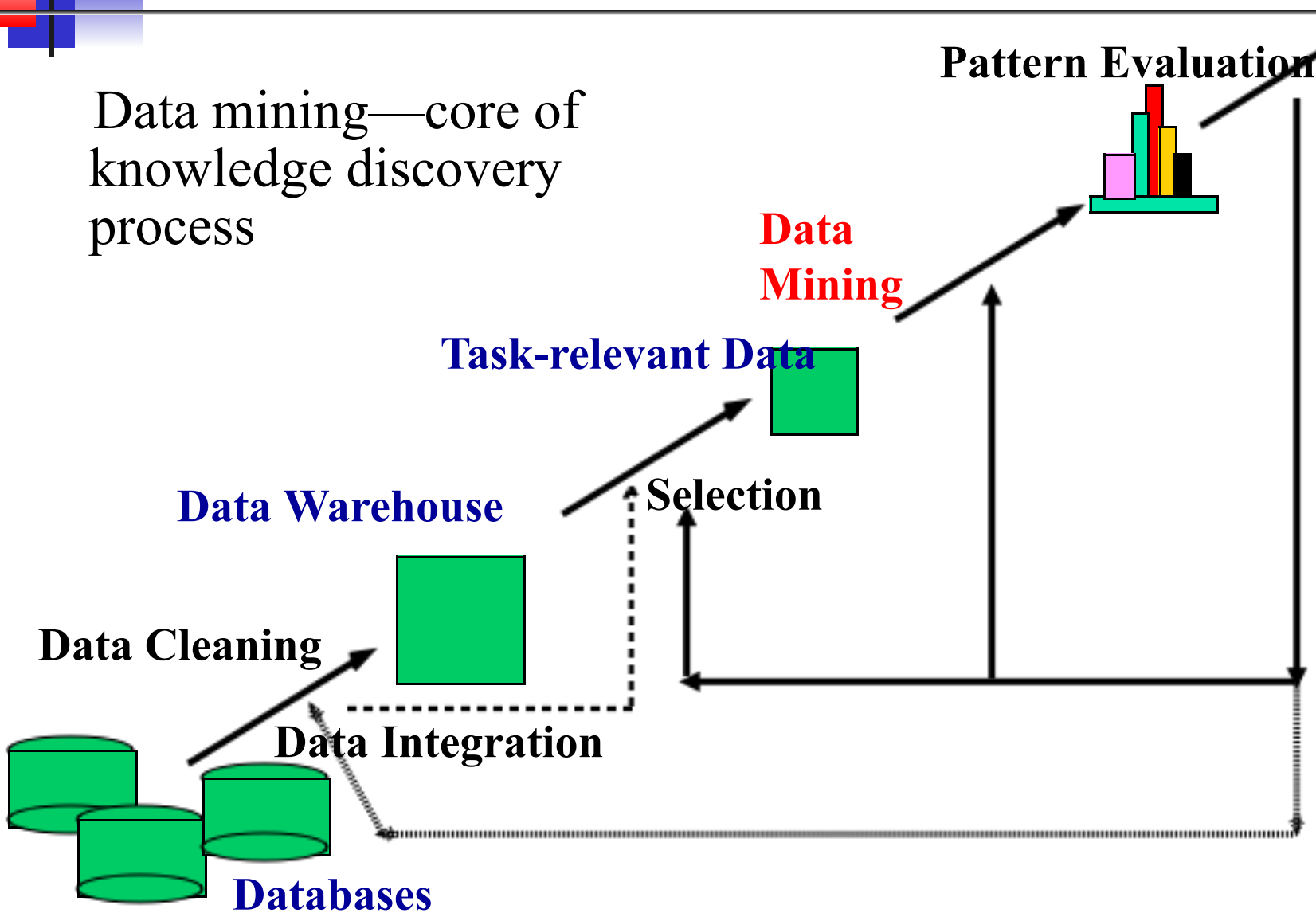
Dr. Bernard Chen Ph.D.
University of Central Arkansas
Fall 2010

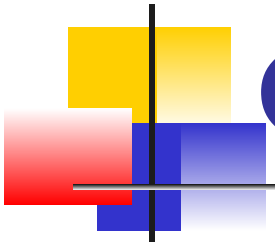
Knowledge Discovery (KDD) Process



Knowledge

- Data mining—core of knowledge discovery process



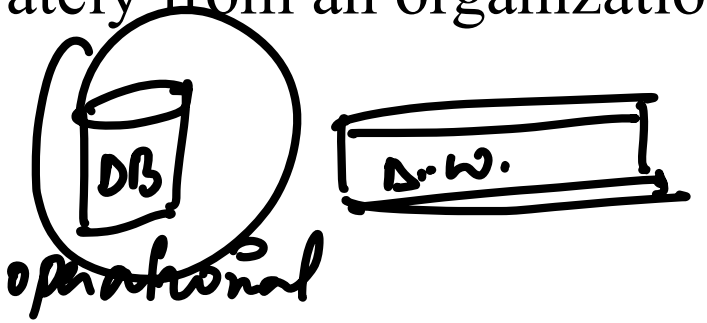


Outline

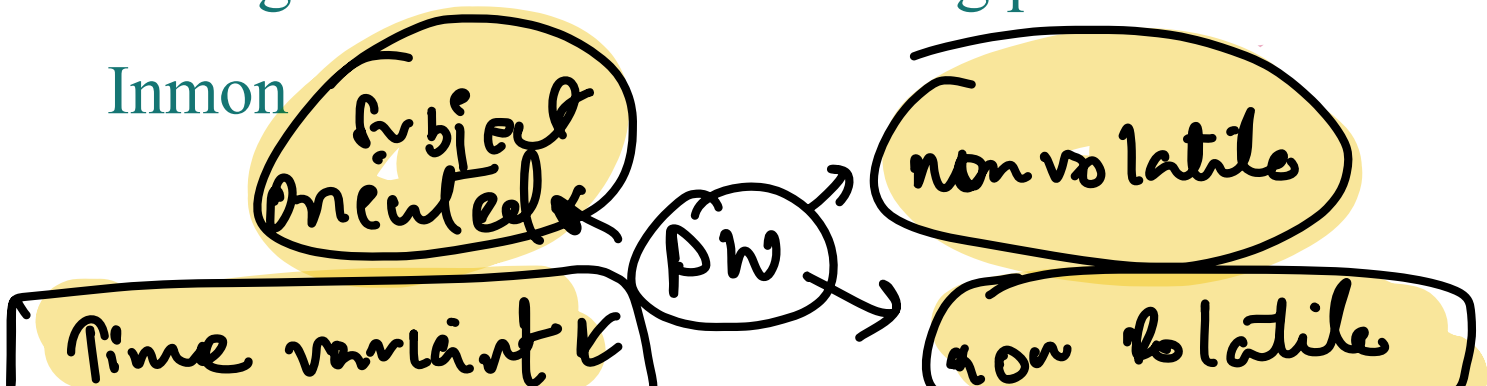
- What is Data Warehouse?
- Data Warehouse: A multidimensional data model
- Data Warehouse Architecture

What is Data Warehouse?

- Loosely speaking, a data warehouse refers to a database that is maintained separately from an organization's operational database



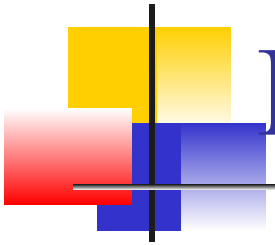
- Officially speaking:
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—William. H. Inmon



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

Multiple datasources.



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
  - When data is moved to the warehouse, it is converted.

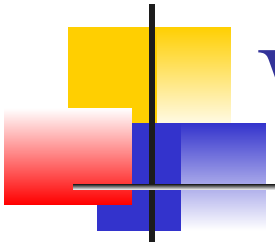
Data Warehouse—Time Variant

- 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—Nonvolatile

- 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*

↳ Subject oriented, Time Variant, Nonvolatile & Integrated

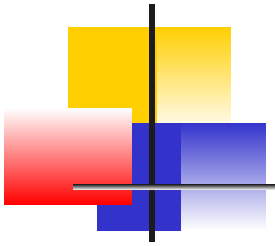


What is Data Warehouse?

- In sum, data warehouse is a semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisioins

data on which enterprise needs to take strategic decisions

Data Warehouse vs. Heterogeneous DBMS



- Traditional heterogeneous DB integration: A query driven approach
  - Build wrappers/mediators on top of heterogeneous databases
- Data warehouse: update-driven, high performance
  - Information from heterogeneous sources is integrated in advance and stored in warehouses for direct query and analysis

- heterogeneous: Query driven
- data warehouse: update driven

Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction (query) processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.

- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making

OLTP (Online Txn processing) | OLAP (Online Analytical Processing)

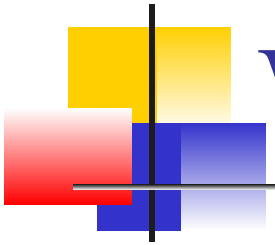
↳ Task of traditional Rel. DBMS | ↳ task of data warehouse system

D2D operations | ↳ data analysis & decision making

Order Mgmt – Payroll, purchase accounts etc

Data Warehouse vs. Operational DBMS

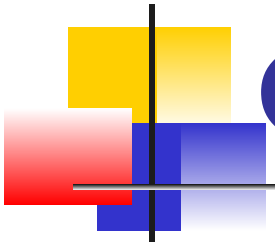
	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response



Why Separate Data Warehouse?

- Question: why not perform on-line analytical processing directly on such database instead of spending additional time and recourses to construct a 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



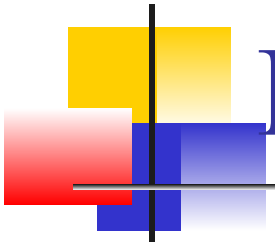
Outline

- What is Data Warehouse?
- Data Warehouse: A multidimensional data model
- Data Warehouse Architecture

A multi-dimensional data model

- A data warehouse is based on a multidimensional data model which views data in the form of a data cube

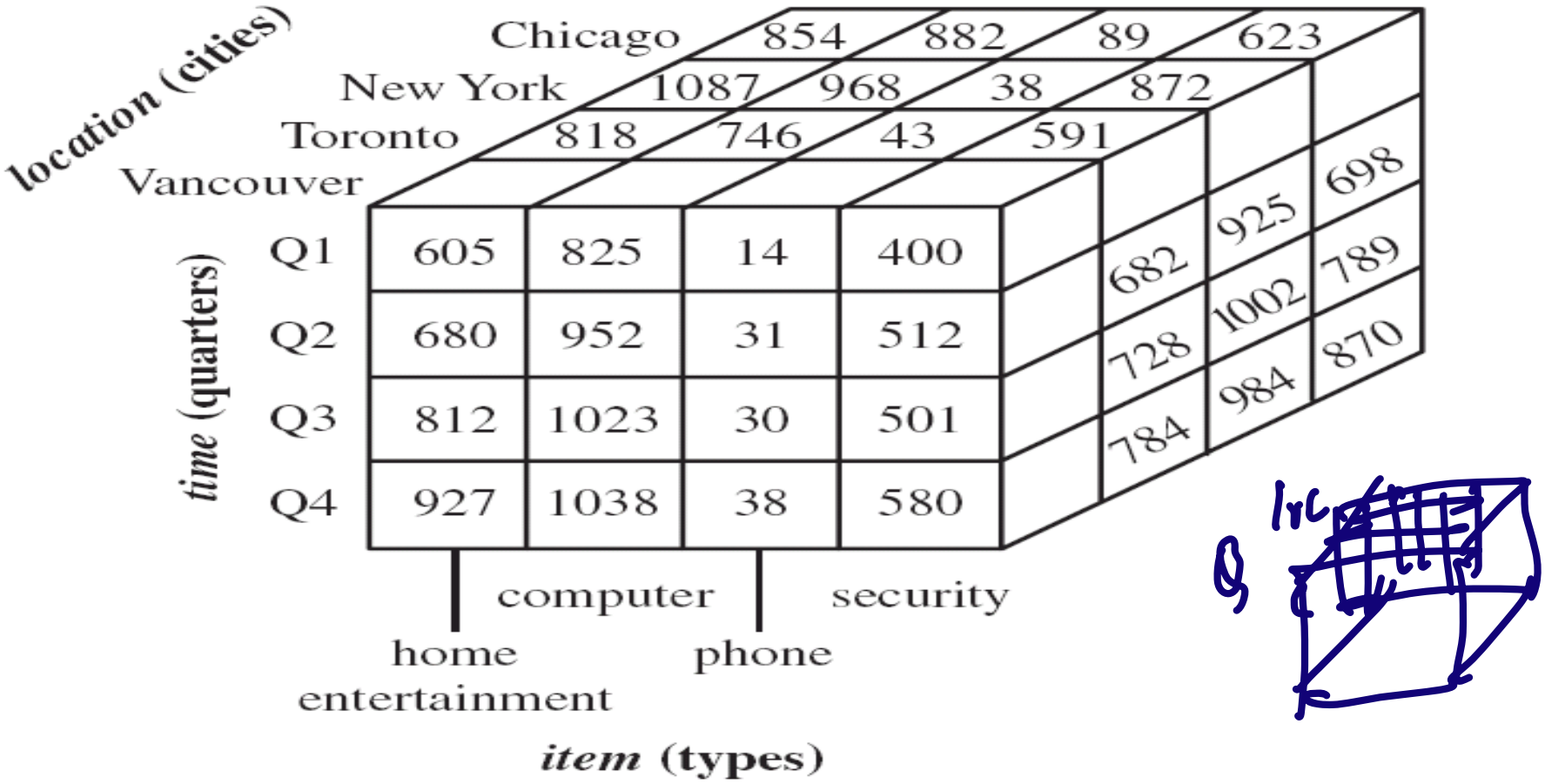
Multidimensional data model → data ↓ data cube

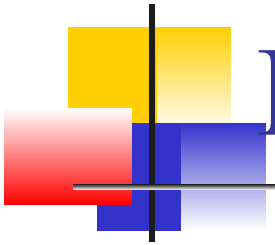


Data cube

- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
- Suppose ALLELETRONICS create a *sales* data warehouse with respect to dimensions
  - Time
  - Item
  - Location

3D Data cube Example

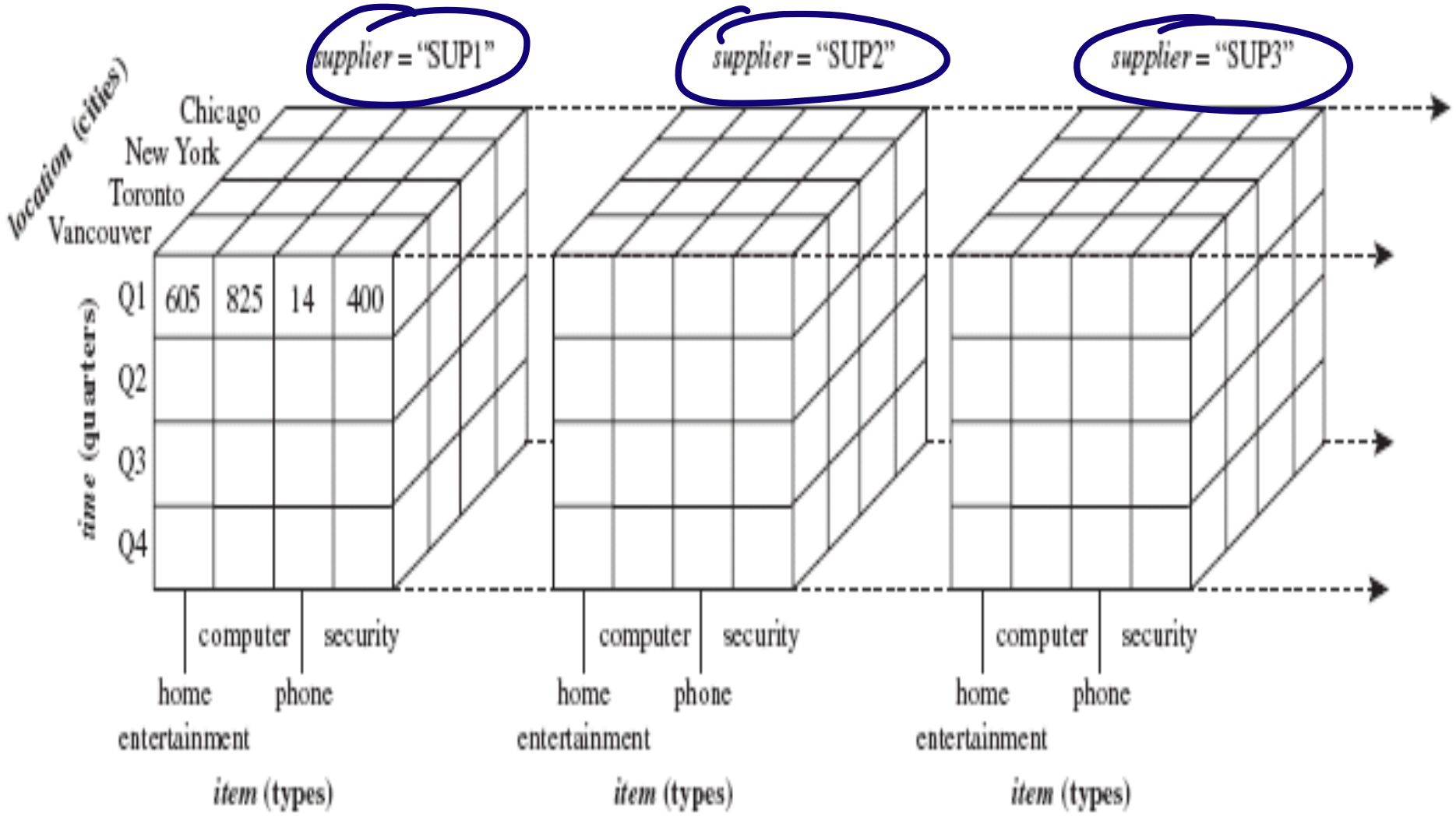


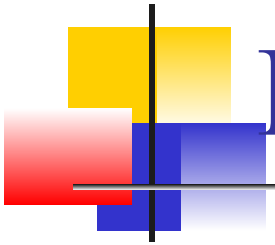


Data cube

- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
- Suppose ALLELETRONICS create a *sales* data warehouse with respect to dimensions
  - Time
  - Item
  - Location
  - Supplier

4D Data cube Example

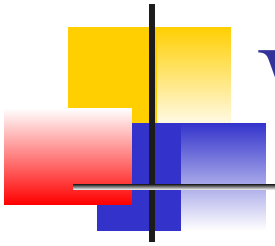




Practice Question

- What is a 5D cube looks like?

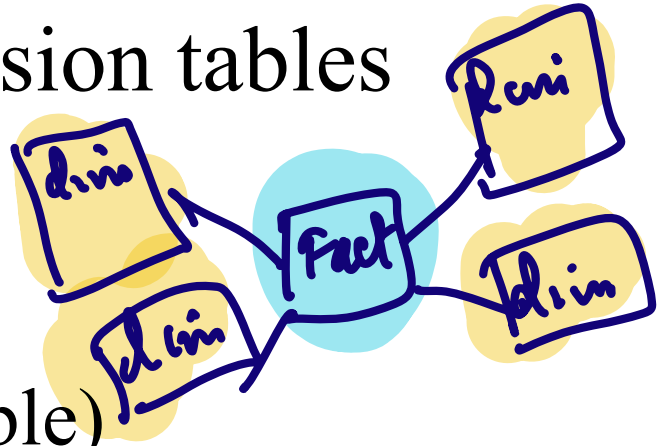
Conceptual Modeling of Data Warehouses



- The most popular data model for a data warehouse is a multi-dimensional model
  - types → Star schema, Snowflake, Fact constellation
- Such a model can exist in the form of:
  - Star schema
  - Snowflake schema
  - Fact constellations

Conceptual Modeling of Data Warehouses

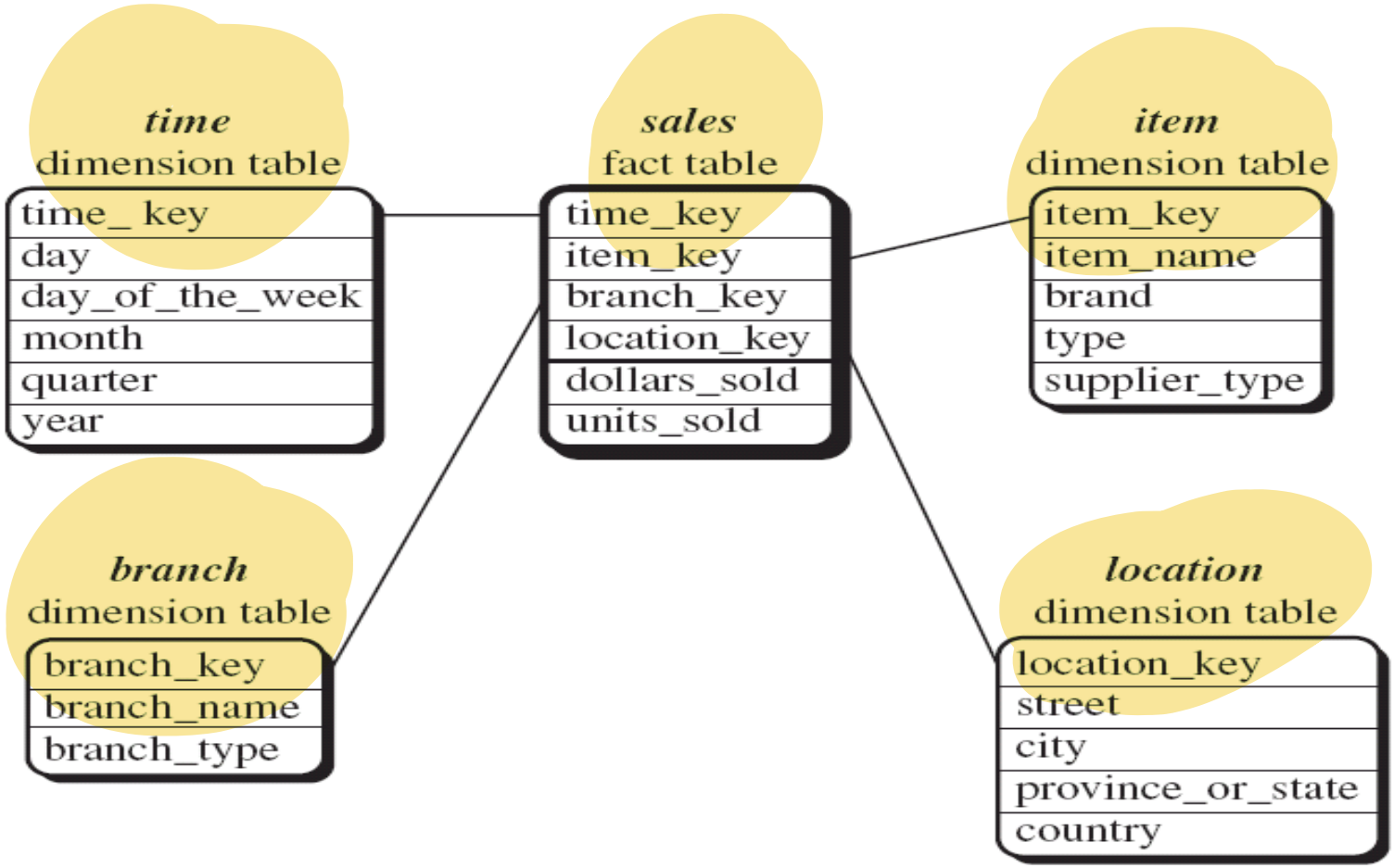
- Star schema: A fact table in the middle connected to a set of dimension tables



- It contains:
  - A large central table (fact table)
  - A set of smaller attendant tables (dimension table), one for each dimension

fact: large central table
dim: small tables, 1 for each dimensions

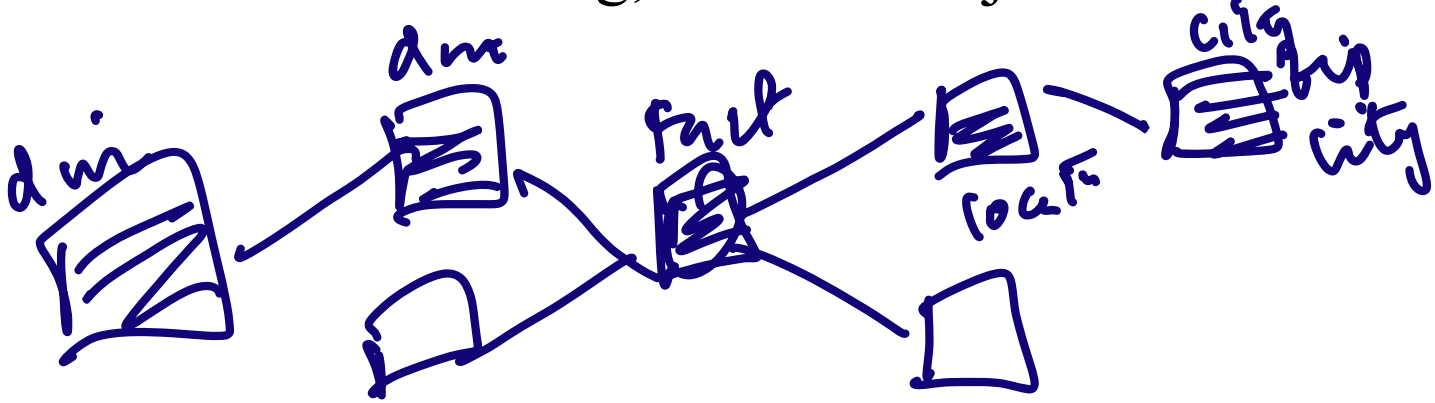
Star schema


time
dimension table
time_ key
day
day_of_the_week
month
quarter
year
sales
fact table
time_key
item_key
branch_key
location_key
dollars_sold
units_sold
item
dimension table
item_key
item_name
brand
type
supplier_type
branch
dimension table
branch_key
branch_name
branch_type
location
dimension table
location_key
street
city
province_or_state
country

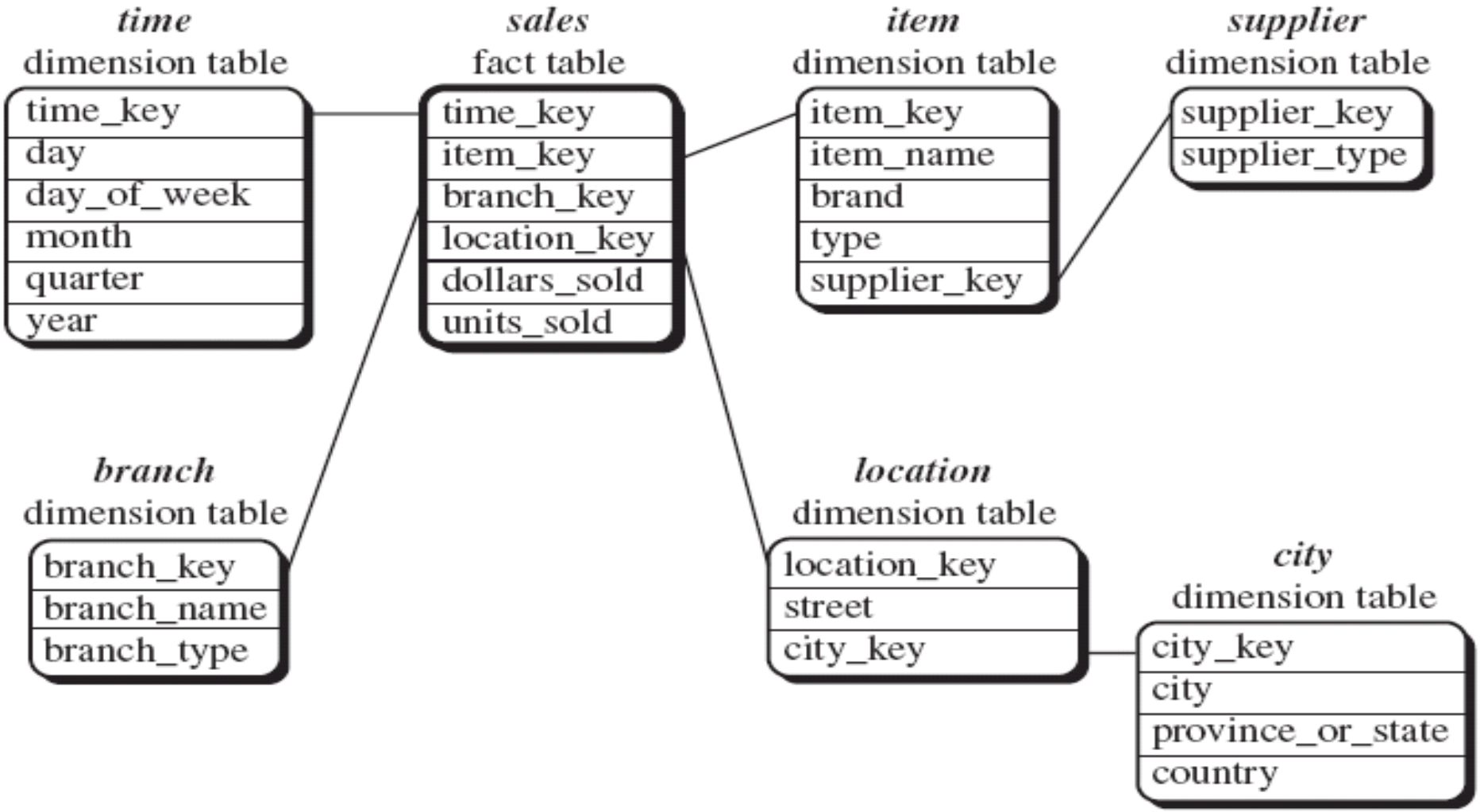
Conceptual Modeling of Data Warehouses

- Snowflake schema: A refinement of star schema where some dimensional hierarchy is further splitting (normalized) into a set of smaller dimension tables, forming a shape similar to snowflake

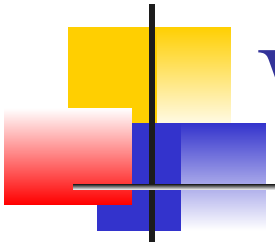
- However, the snowflake structure can reduce the effectiveness of browsing, since more joins will be needed



Snowflake schema



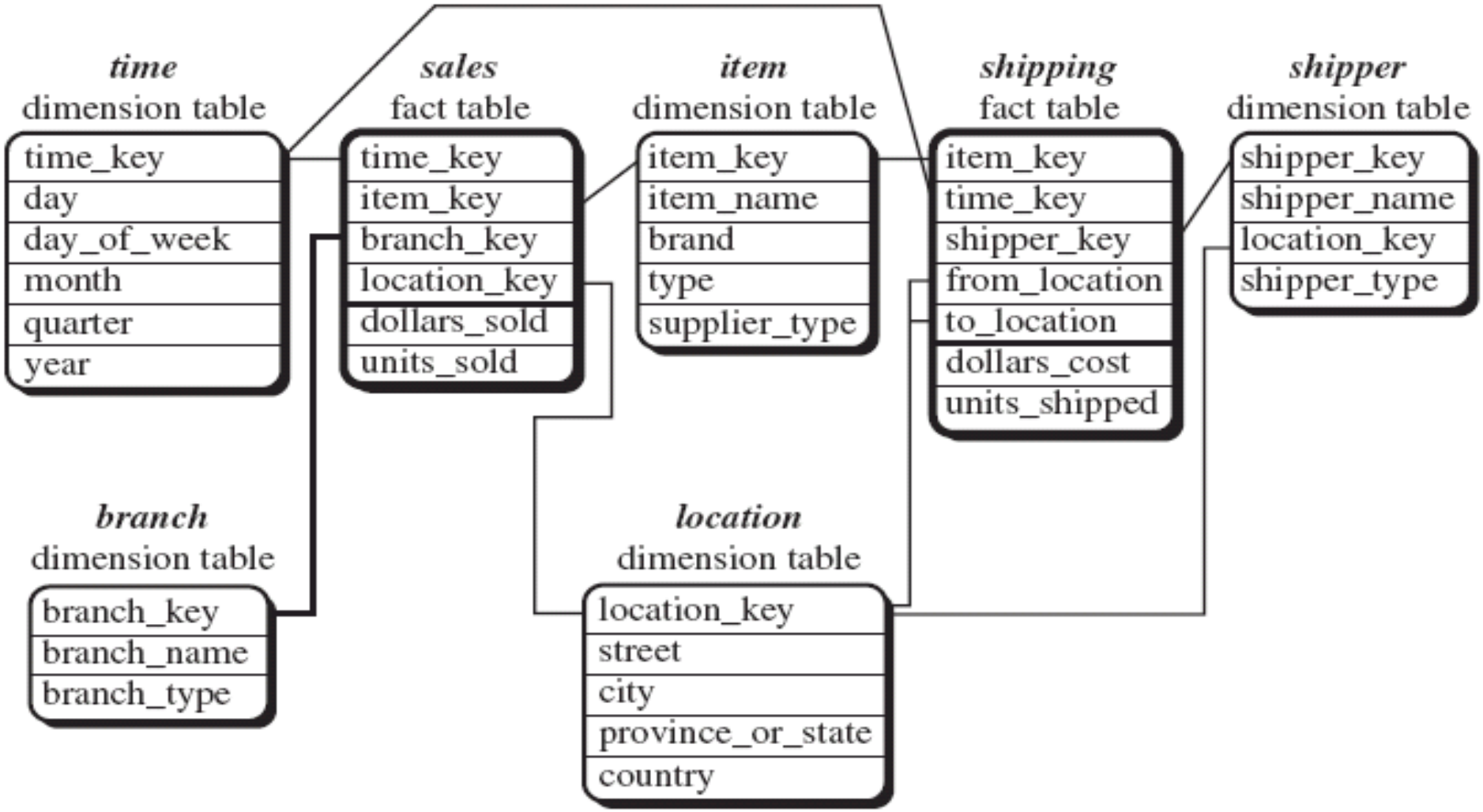
Conceptual Modeling of Data Warehouses



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

Galaxy Schema

Fact constellations

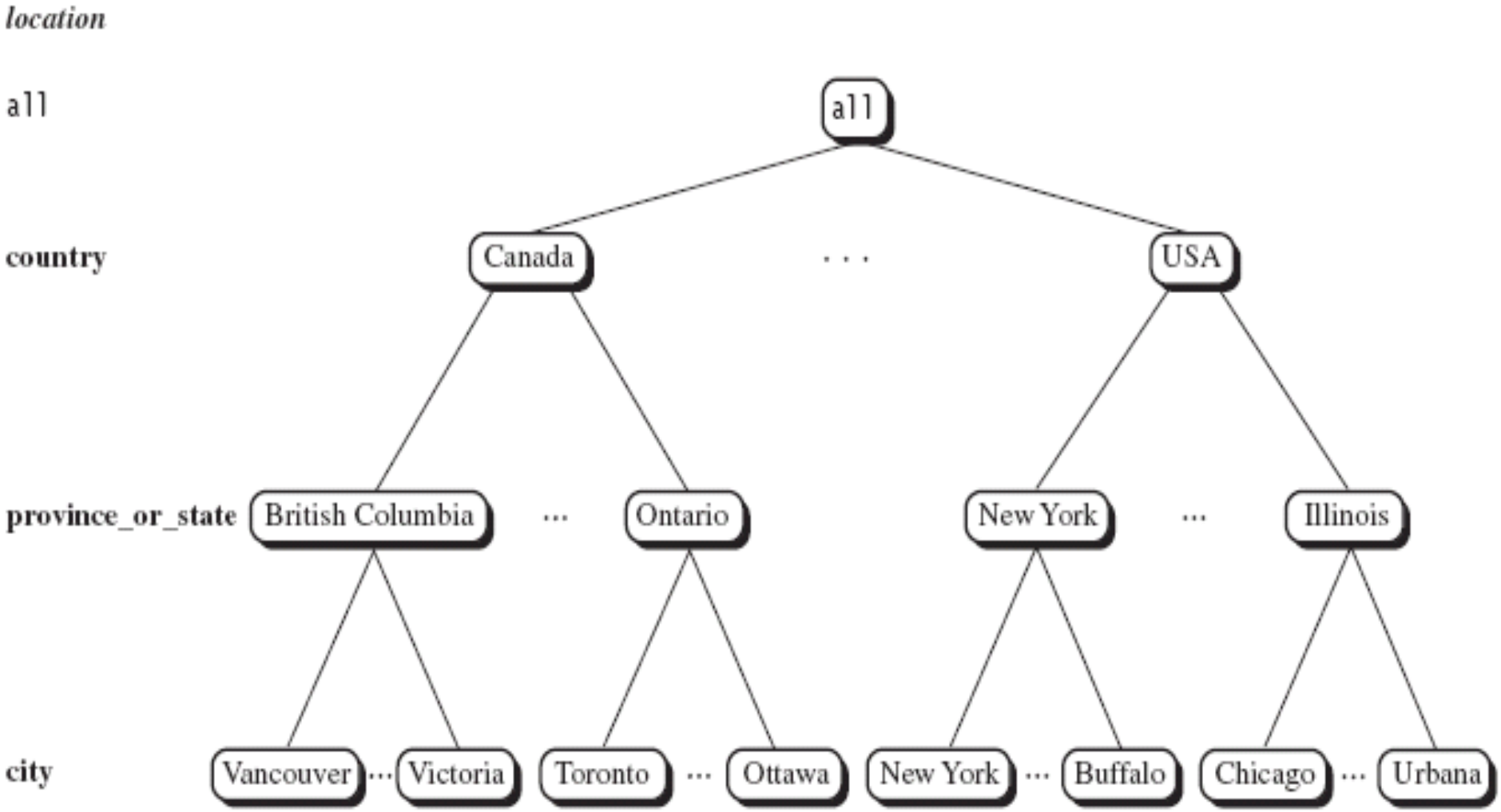


Concept Hierarchies

- A **Concept Hierarchy** defines a sequence of mappings from a set of low-level concepts to high-level

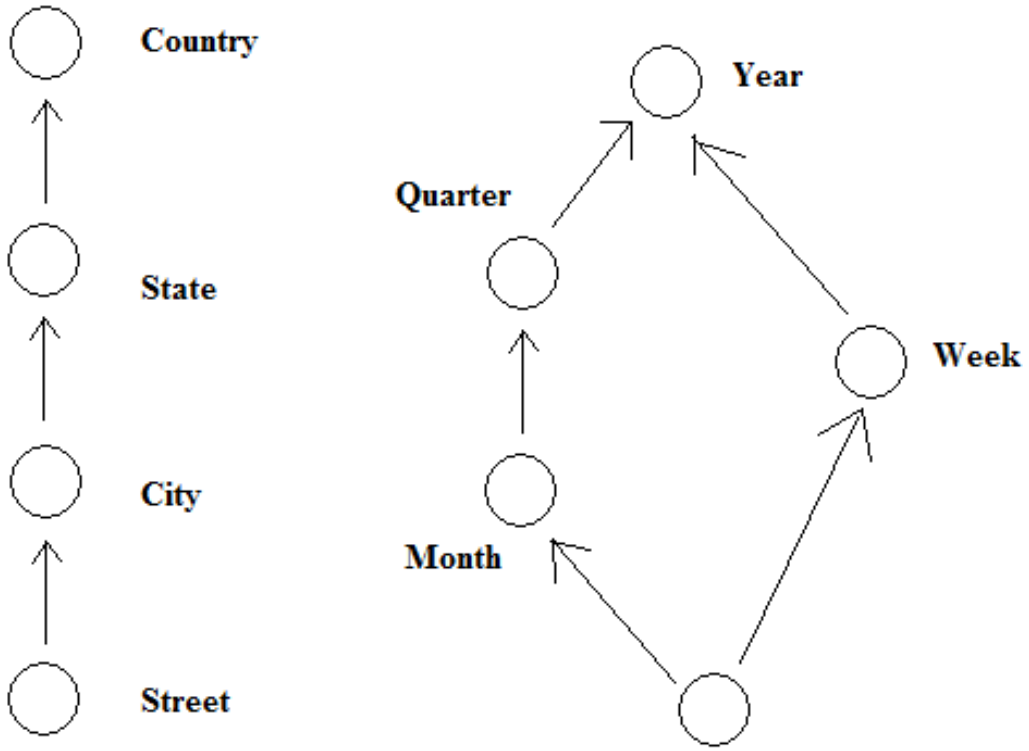
- Consider a concept hierarchy for the dimension **"Location"**

Concept Hierarchies



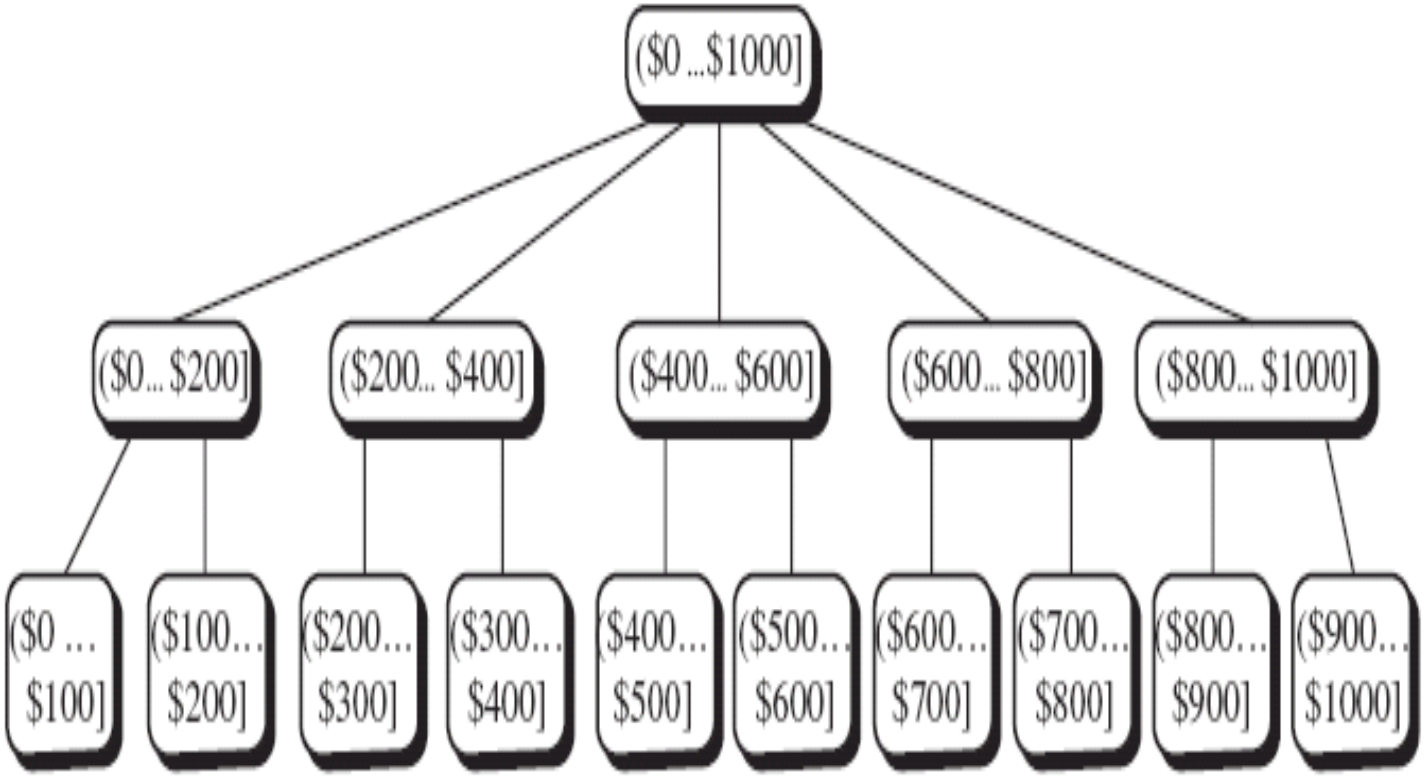
Concept Hierarchies

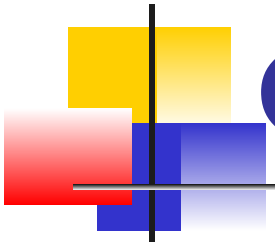
- Many concept hierarchies are implicit within the database system



Concept Hierarchies

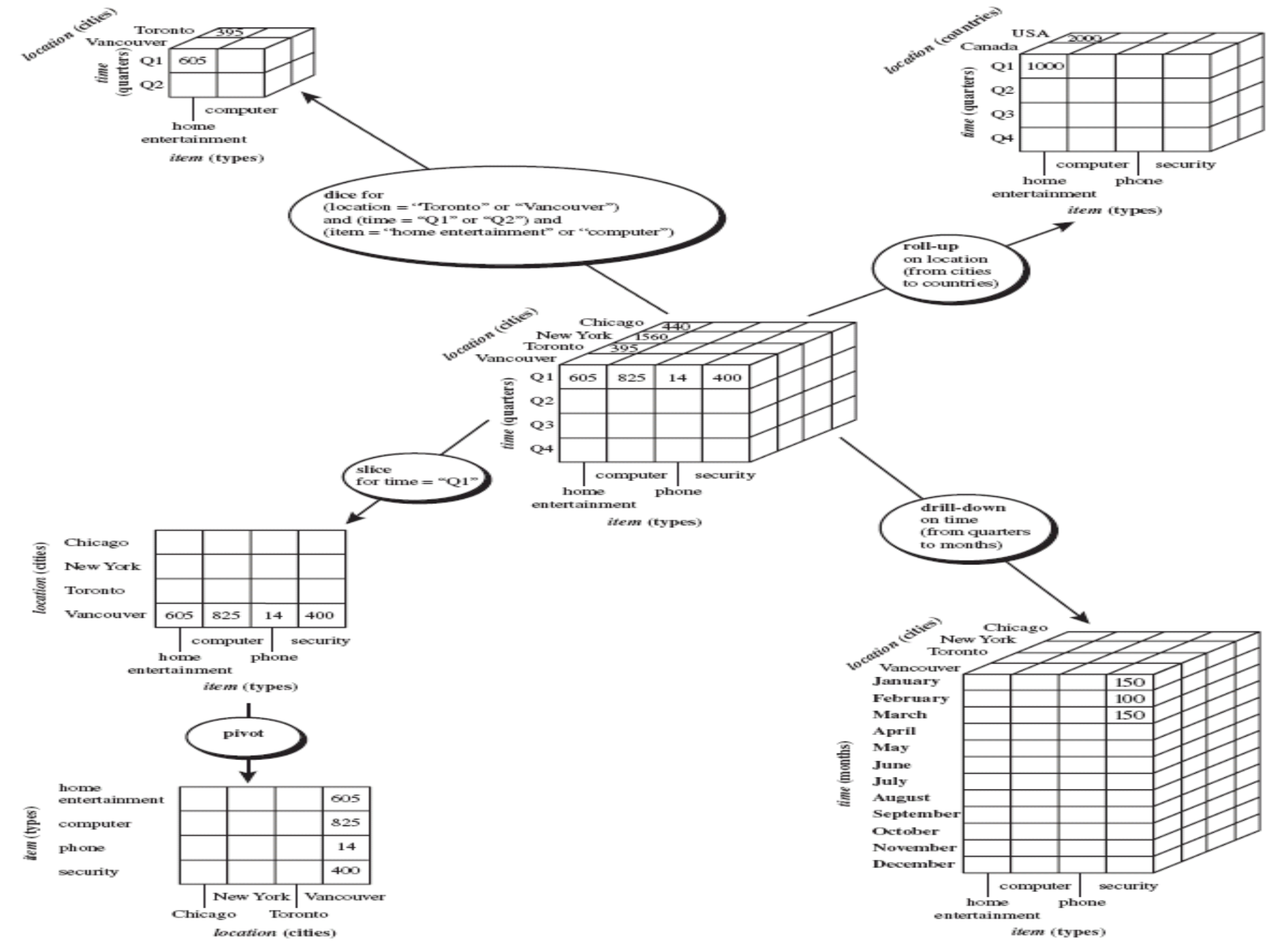
- Concept hierarchies may also be defined by grouping values for a given dimension or attribute, resulting in a **set-grouping hierarchy**





OLAP Operation

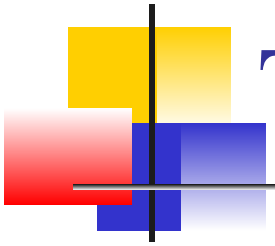
- So, how are *concept hierarchies* useful in OLAP?
- In the multidimensional model, data are organized into multiple dimensions,
- And each dimension contains multiple levels of abstraction defined by concept hierarchies


location (cities)
Toronto
Vancouver
395
time (quarters)
Q1
Q2
605
computer
home entertainment
item (types)
dice for
(location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "home entertainment" or "computer")
location (countries)
USA
Canada
2000
Q1
Q2
Q3
Q4
1000
computer
security
home entertainment
phone
item (types)
roll-up
on location
(from cities
to countries)
location (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Q1
Q2
Q3
Q4
605
825
14
400
computer
security
home entertainment
phone
item (types)
slice
for time = "Q1"
drill-down
on time
(from quarters
to months)
location (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
computer
security
home entertainment
phone
item (types)
pivot
item (types)
home entertainment
computer
phone
security
605
825
14
400
New York
Vancouver
Chicago
Toronto
location (cities)
location (cities)
Chicago
New York
Toronto
Vancouver
January
February
March
April
May
June
July
August
September
October
November
December
150
100
150
time (months)
computer
security
home entertainment
phone
item (types)

Typical OLAP Operations

- Roll up (drill-up): summarize data
  - *by climbing up hierarchy or by dimension reduction*

- Drill down (roll down): reverse of roll-up
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*

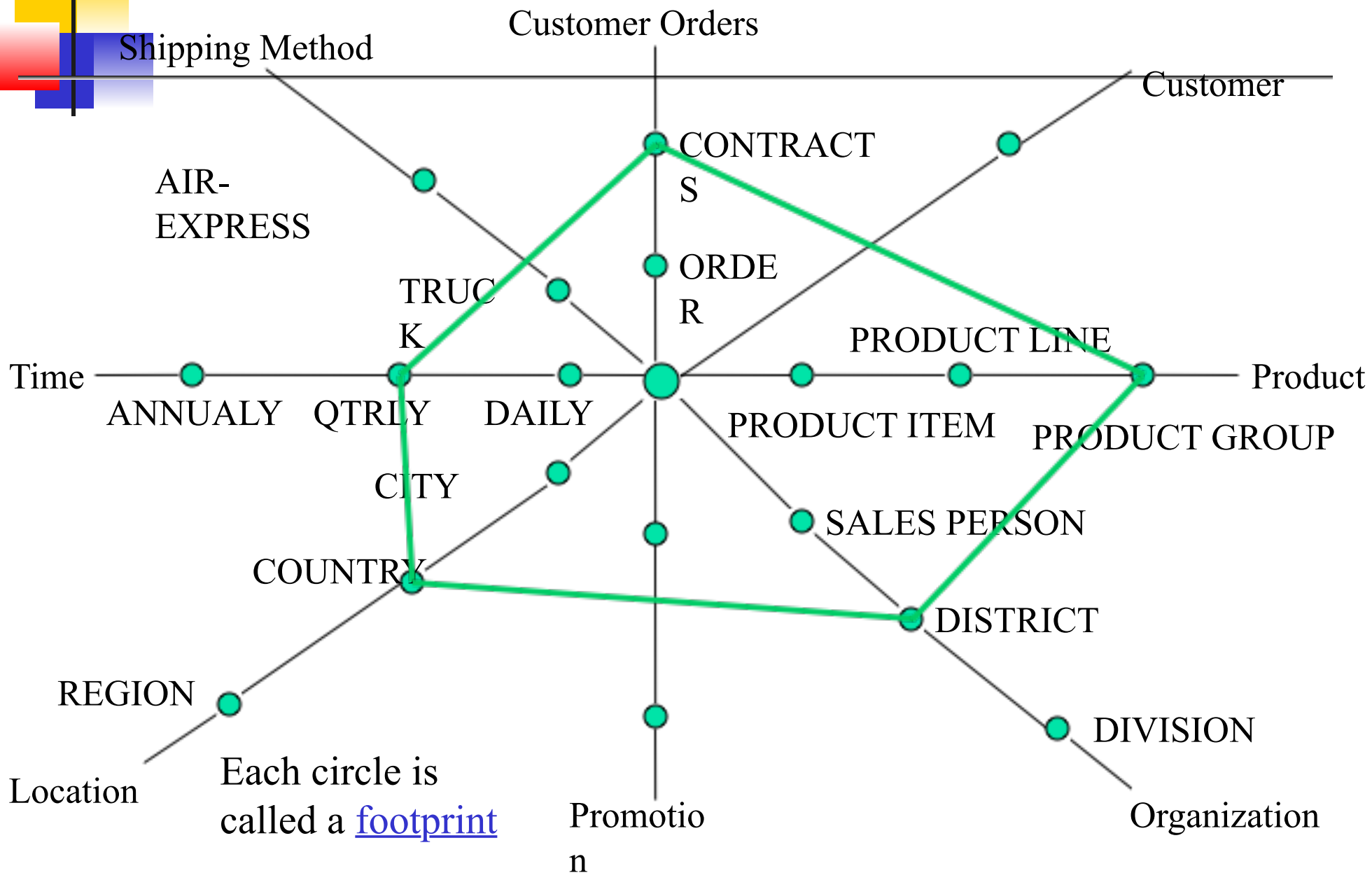


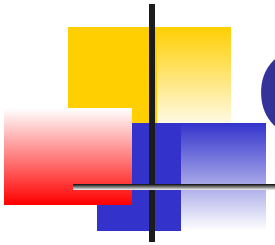
Typical OLAP Operations

- Slice and dice: *project and select*

- Pivot (rotate):
  - *reorient the cube, visualization, 3D to series of 2D planes*

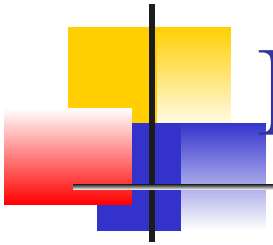
A Star-Net Query Model





Outline

- What is Data Warehouse?
- Data Warehouse: A multidimensional data model
- Data Warehouse Architecture

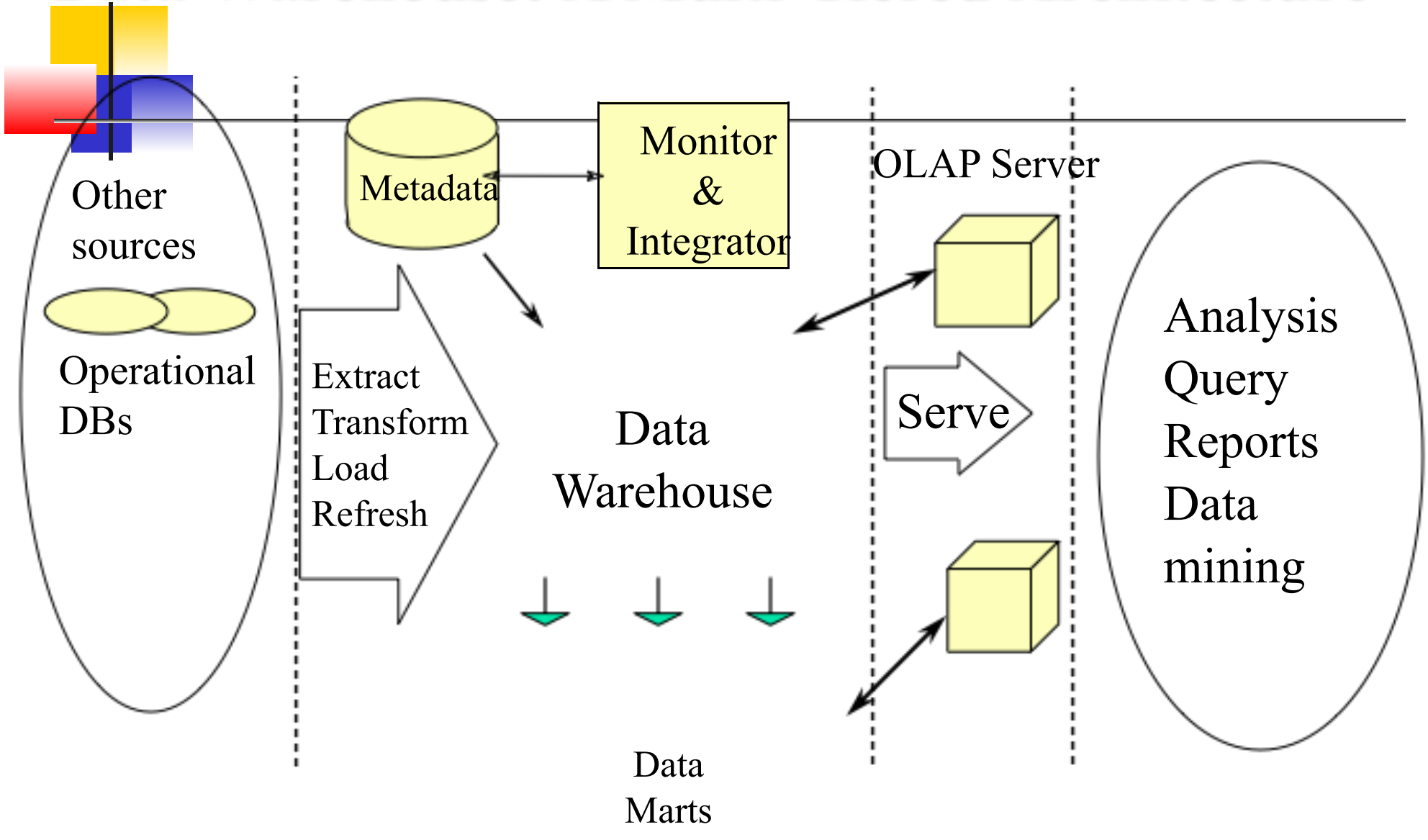


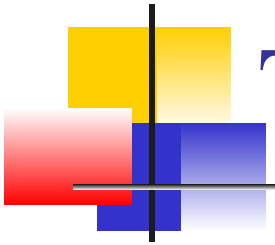
Design of Data Warehouse

- Four views regarding the design of a data warehouse MUST be conserned

  - Top-down view
    - allows selection of the relevant information necessary for the data warehouse
  - Data source view
    - exposes the information being captured, stored, and managed by operational systems
  - Data warehouse view
    - consists of fact tables and dimension tables
  - Business query view
    - sees the perspectives of data in the warehouse from the view of end-user

Data Warehouse: A Multi-Tiered Architecture

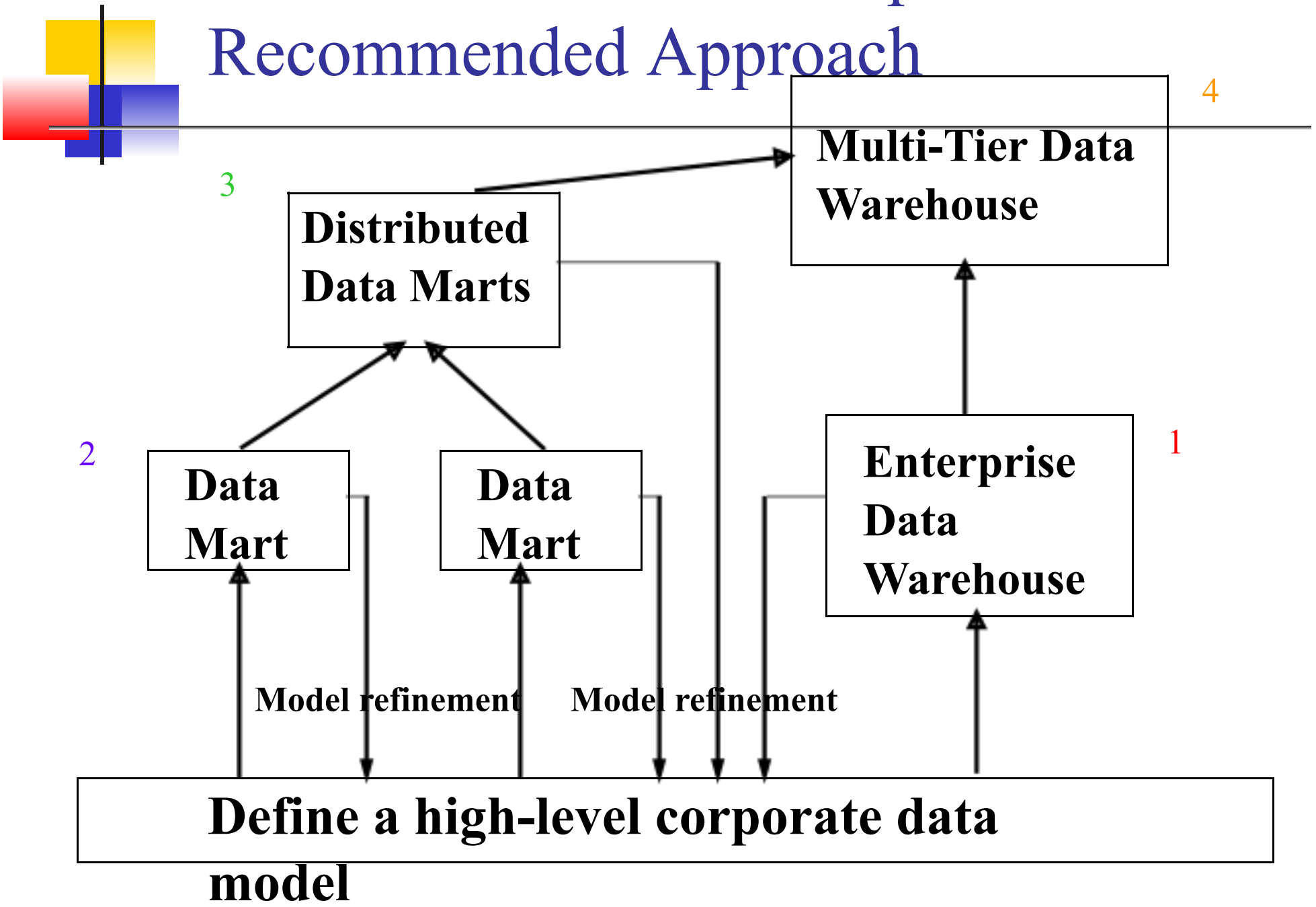


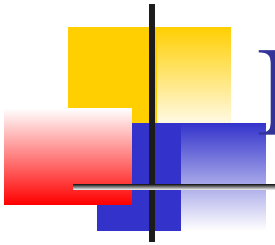


Three Data Warehouse Models

- Enterprise warehouse
  - collects all of the information about subjects spanning the entire organization
- Data Mart
  - a subset of corporate-wide data that is of value to a specific groups of users. Its scope is confined to specific, selected groups, such as marketing data mart
    - Independent vs. dependent (directly from warehouse) data mart
- Virtual warehouse
  - A set of views over operational databases. Only some of the possible summary views may be materialized.

Data Warehouse Development: A Recommended Approach

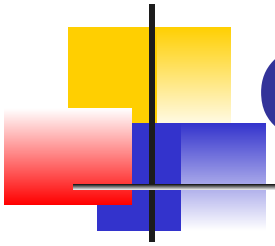




Metadata Repository

- Meta data is the data defining warehouse objects. It stores:
- Description of the structure of the data warehouse
  - schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents
- Operational meta-data
  - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
- The algorithms used for summarization
- The mapping from operational environment to the data warehouse
- Data related to system performance
  - warehouse schema, view and derived data definitions
- Business data
  - business terms and definitions, ownership of data, charging policies

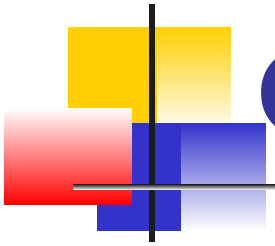
Efficient Processing OLAP Queries



- Determine which operations should be performed on the available cuboids

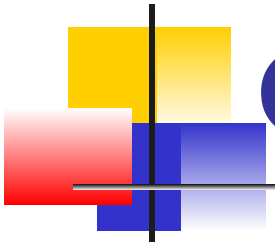
- Determine which materialized cuboid(s) should be selected for OLAP op.

Efficient Processing OLAP Queries



- Suppose we define a data cube of the form of
  - Day<month<quarter<year
  - Item_name<brand<type
  - Street<city<province or state<country

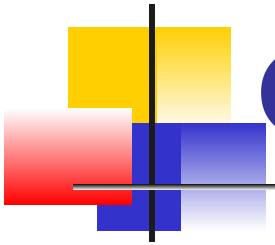
Efficient Processing OLAP Queries



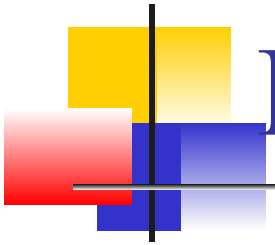
- Let the query to be processed be on {brand, province_or_state} with the condition “year = 2004”, and there are 4 materialized cuboids available:
  1) {year, item_name, city}
  2) {year, brand, country}
  3) {year, brand, province_or_state}
  4) {item_name, province_or_state} where year = 2004

  Which should be selected to process the query?

Efficient Processing OLAP Queries

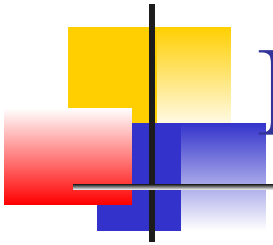


- Cuboid 1,3,4 can be applied
- Finer granularity data cannot be generated from coarser-granularity data



Data Warehouse Usage

- Three kinds of data warehouse applications
  - Information processing
    - supports querying, basic statistical analysis, and reporting using crosstabs, tables, charts and graphs
  - Analytical processing
    - multidimensional analysis of data warehouse data
    - supports basic OLAP operations, slice-dice, drilling, pivoting



Data Warehouse Usage

- Data mining
  - knowledge discovery from hidden patterns
  - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools