

---

Università degli Studi di Milano  
Master Degree in Computer Science

# Information Management course

Teacher: Alberto Ceselli

Lecture 07 : 06/11/2012

# Data Mining:

---

## Concepts and Techniques

(3<sup>rd</sup> ed.)

### — Chapter 4 —

Jiawei Han, Micheline Kamber, and Jian Pei  
University of Illinois at Urbana-Champaign &  
Simon Fraser University

©2011 Han, Kamber & Pei. All rights reserved.

# Chapter 4: Data Warehousing and On-line Analytical Processing

---

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage
- *Data Warehouse Implementation*
- *Data Generalization by Attribute-Oriented Induction*
- Summary

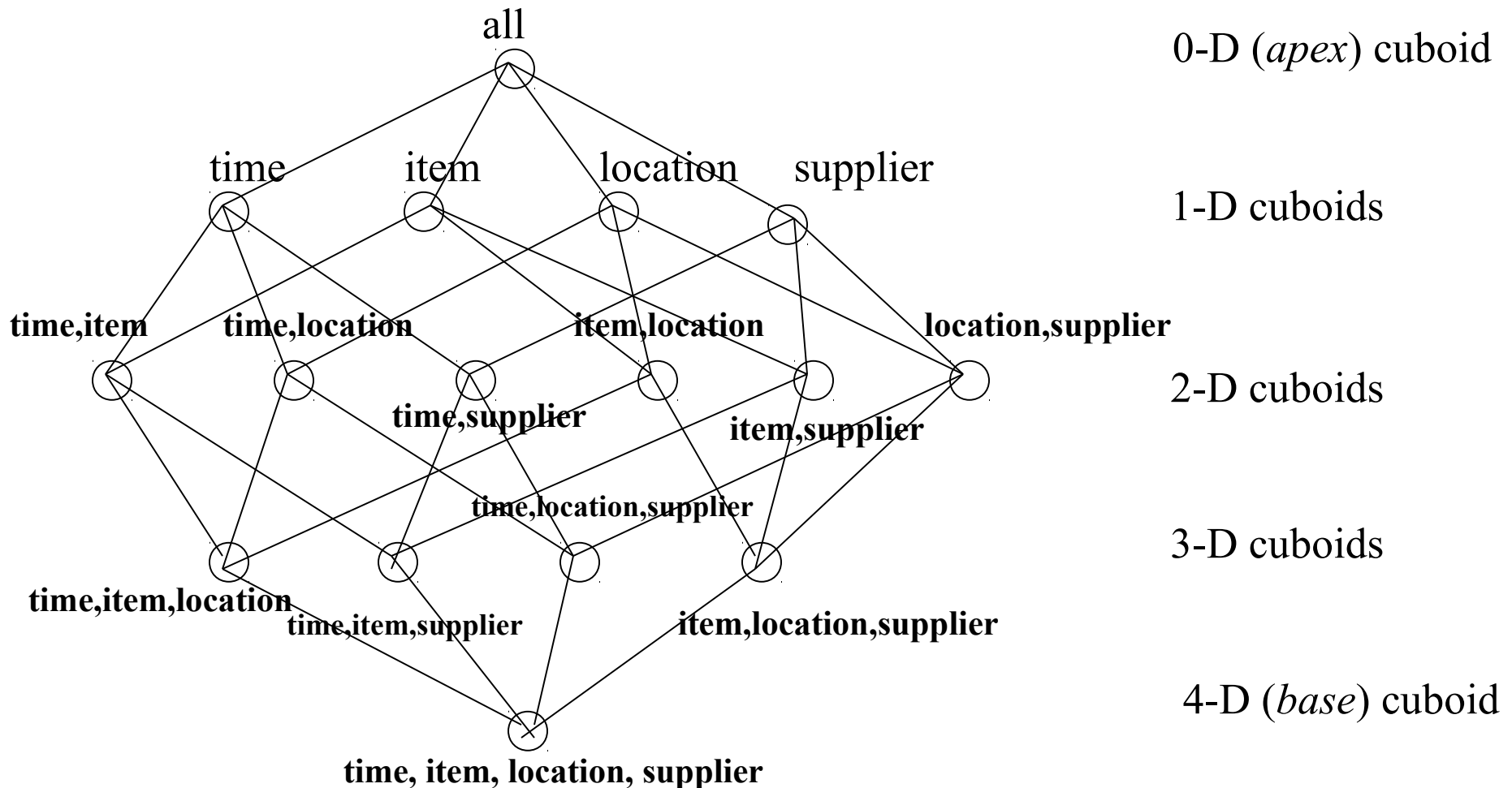


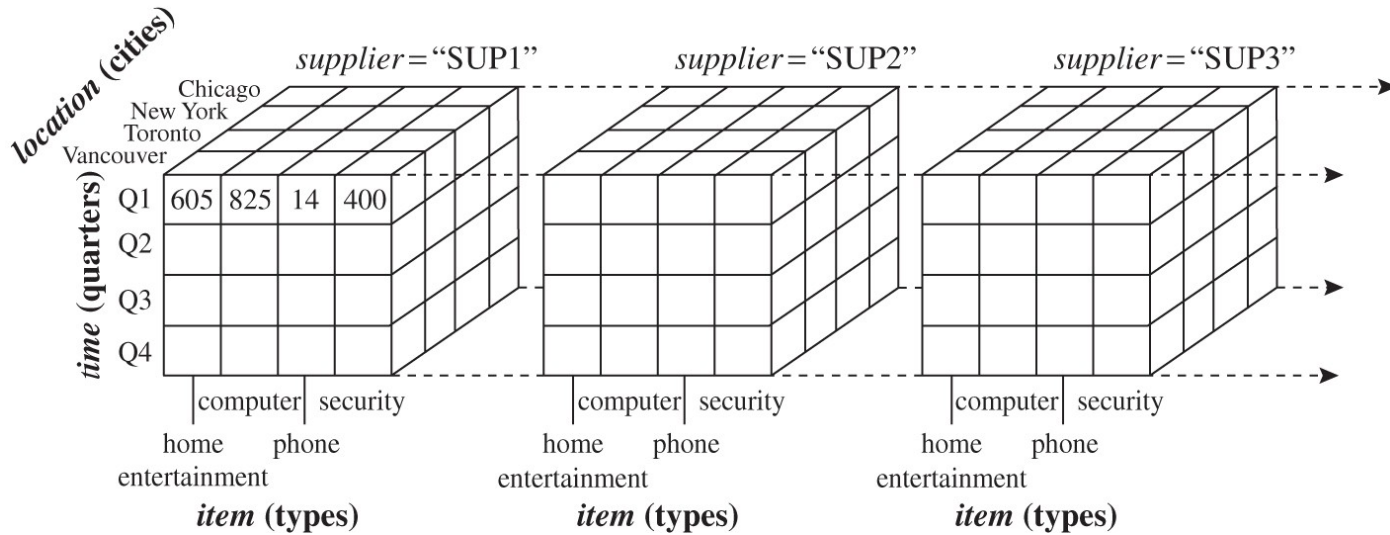
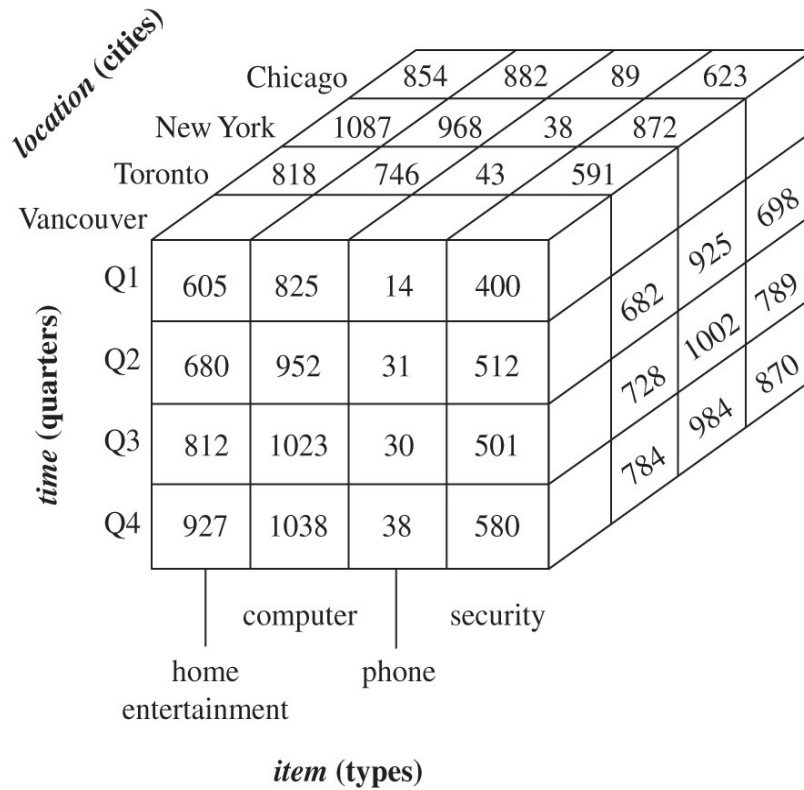
# From Tables and Spreadsheets to Data Cubes

---

- A **data warehouse** is based on a **multidimensional data model** which views data in the form of a data cube
- A data cube, such as **sales**, allows data to be modeled and viewed in multiple dimensions
  - **Dimension tables**, such as **item** (item\_name, brand, type), or **time**(day, week, month, quarter, year)
  - **Fact table** contains **measures** (such as **dollars\_sold**) and keys to each of the related dimension tables
- In data warehousing literature, an n-D base cube is called a **base cuboid**. The top most 0-D cuboid, which holds the highest-level of summarization, is called the **apex cuboid**. The lattice of cuboids forms a **data cube**.

# Data Cube: A Lattice of Cuboids



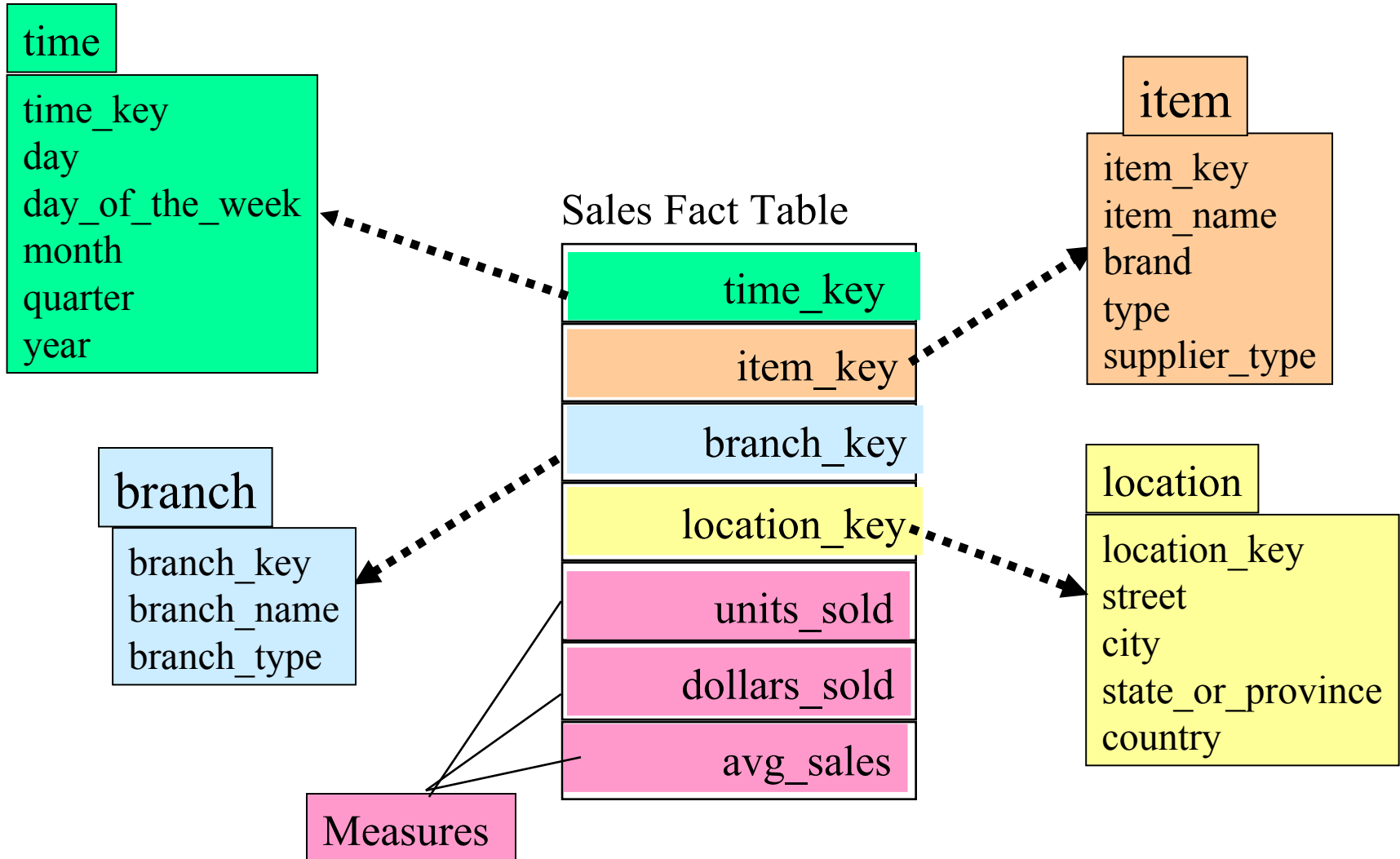


# Conceptual Modeling of Data Warehouses

---

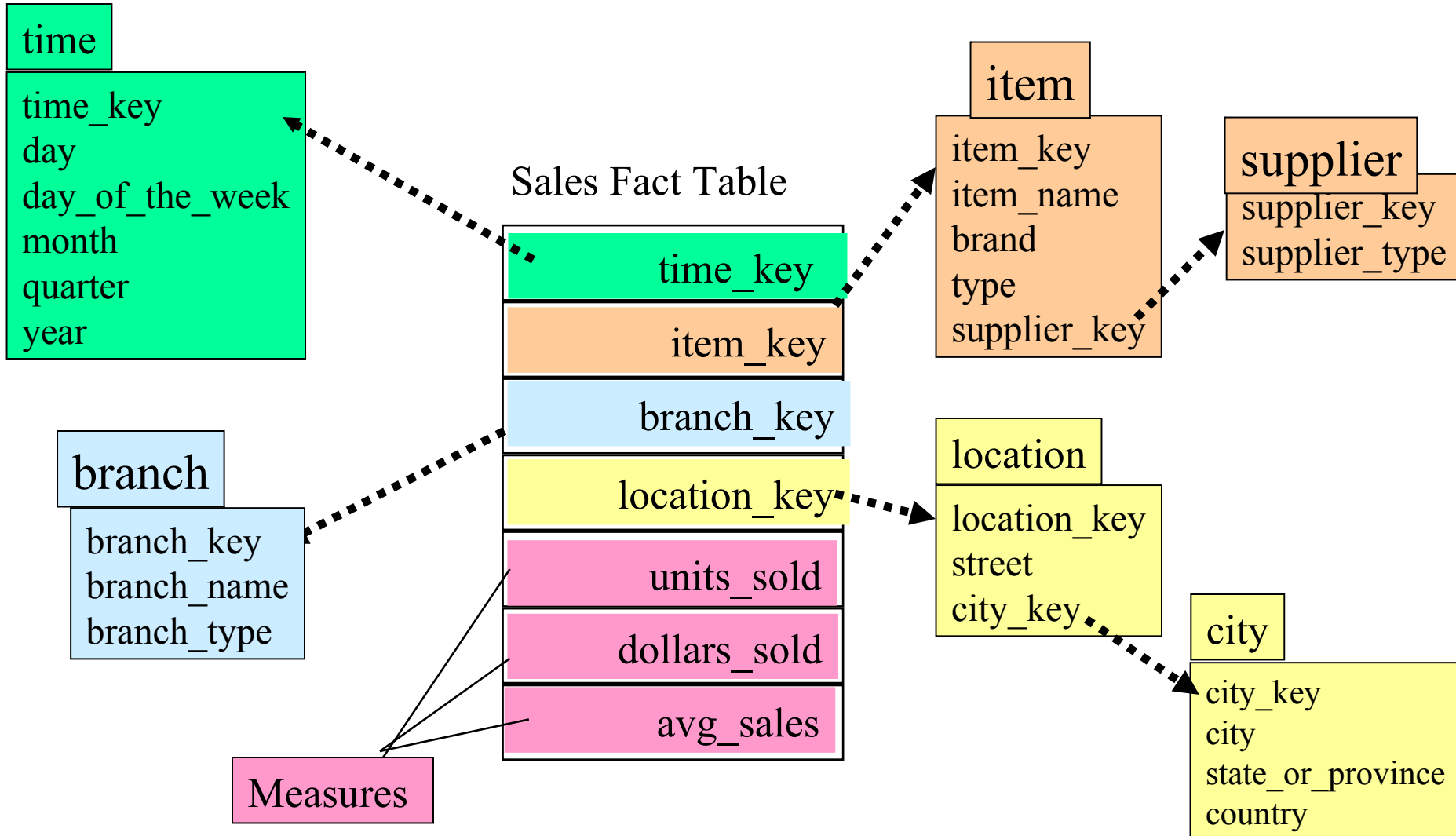
- Modeling data warehouses: dimensions & measures
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is **normalized** into a set of smaller dimension tables, forming a shape similar to snowflake
  - Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called **galaxy schema** or fact constellation

# Example of Star Schema

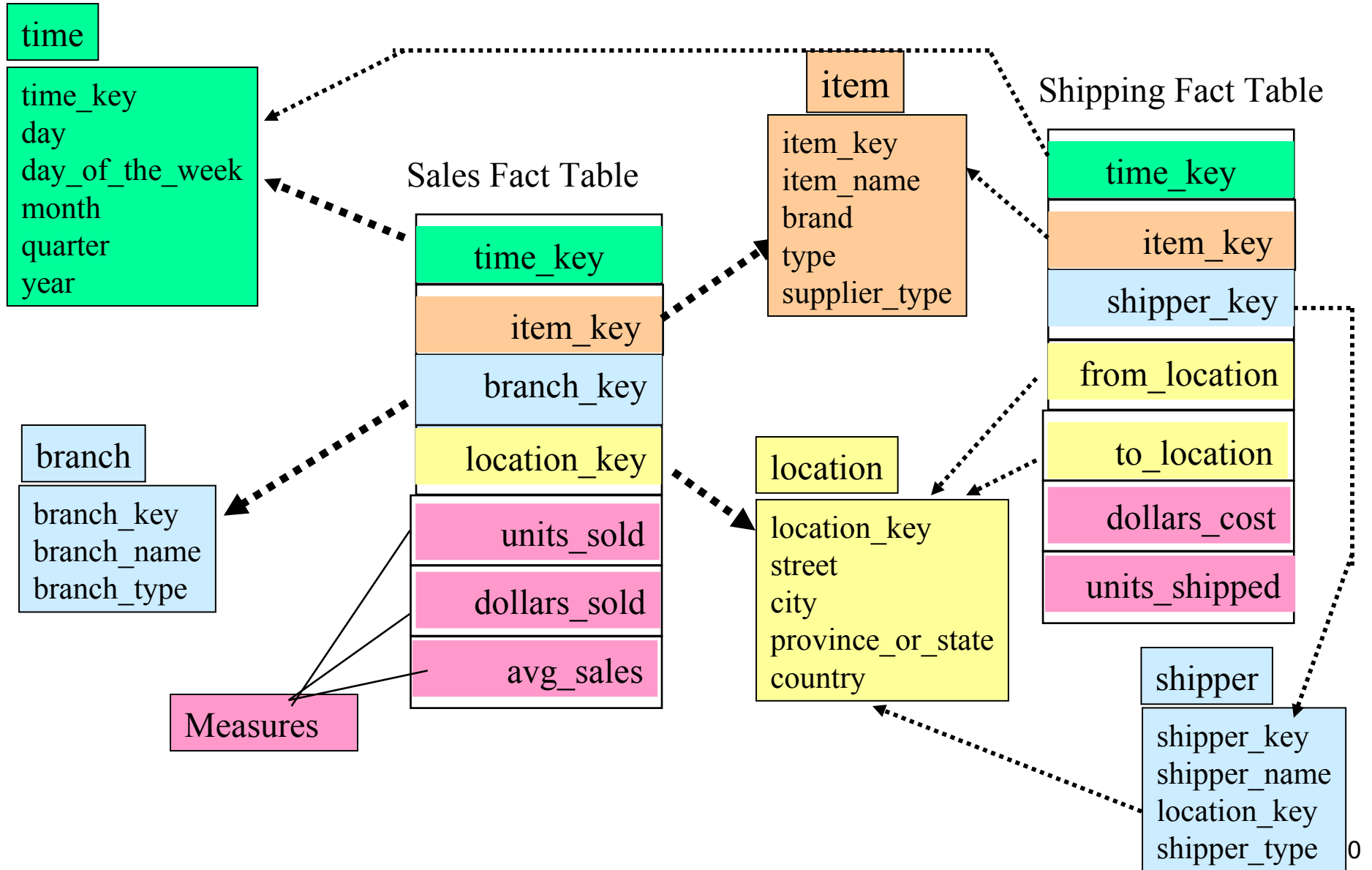




# Example of Snowflake Schema

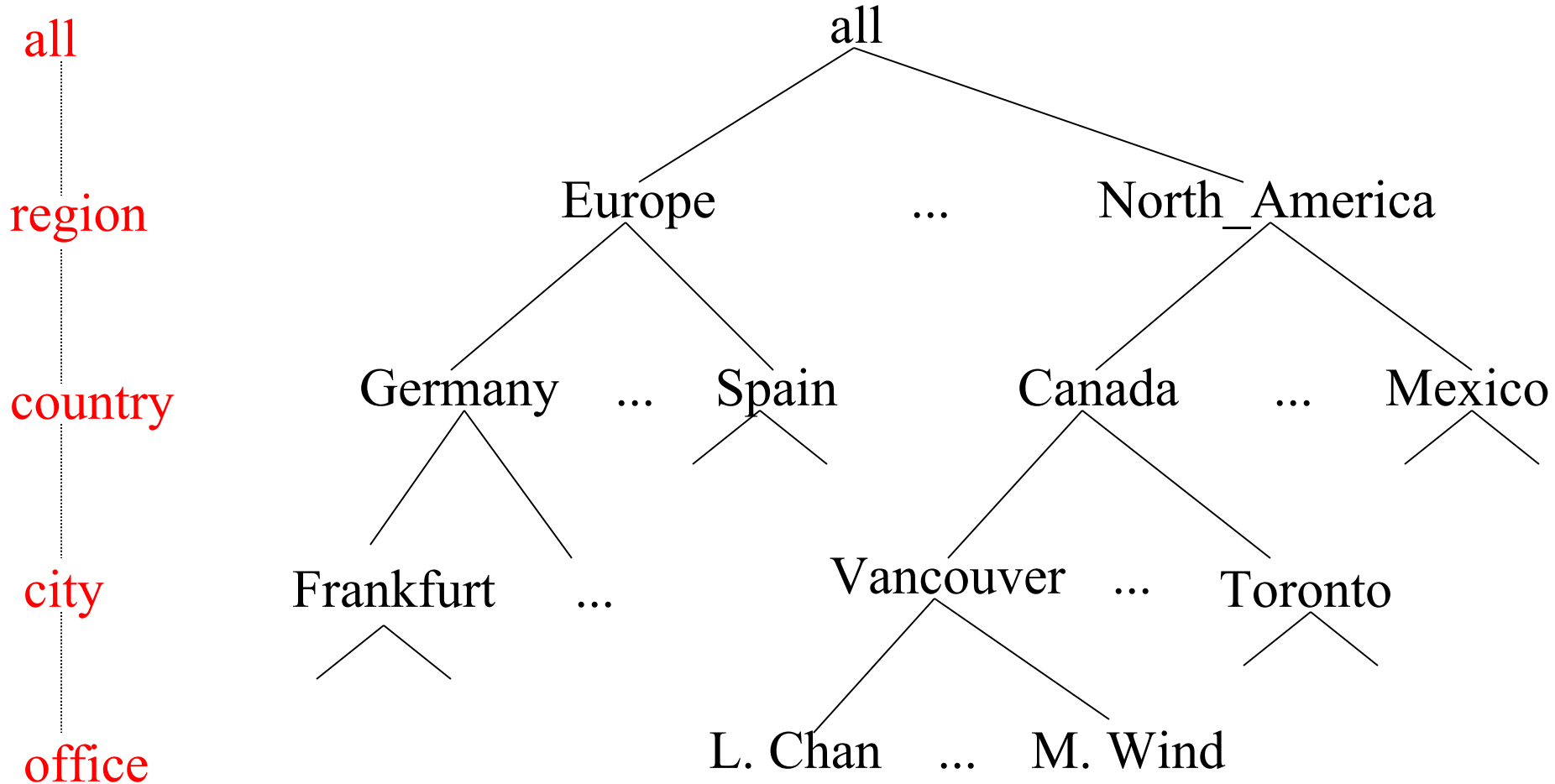


# Example of Fact Constellation



# A Concept Hierarchy: **Dimension** (location)

---



# Data cube measures

---

- Measure: a numeric function that can be evaluated at each point in the data cube space:
  - Fact
  - Aggregation of facts

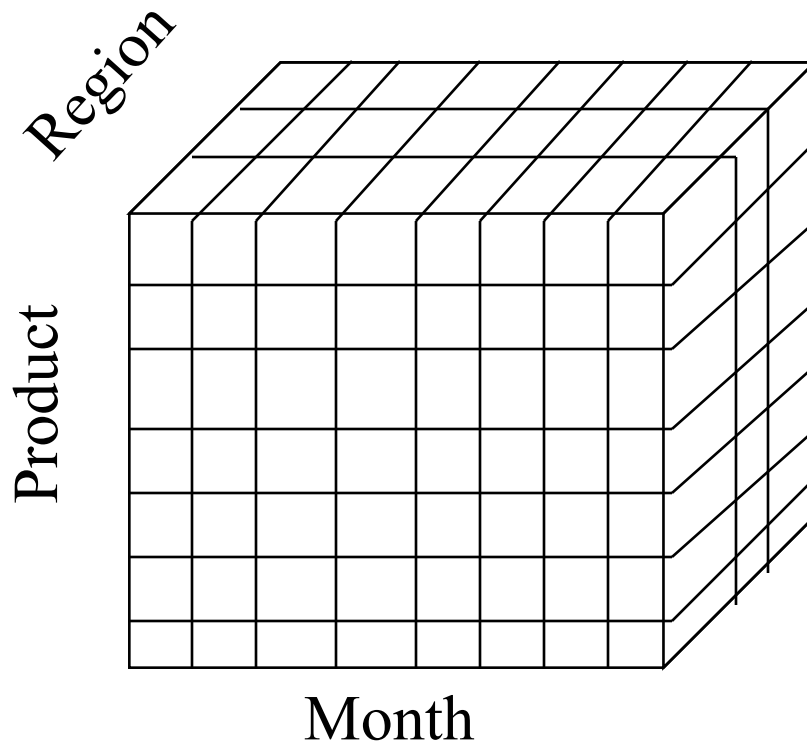
# Data Cube Measures: Three Categories

---

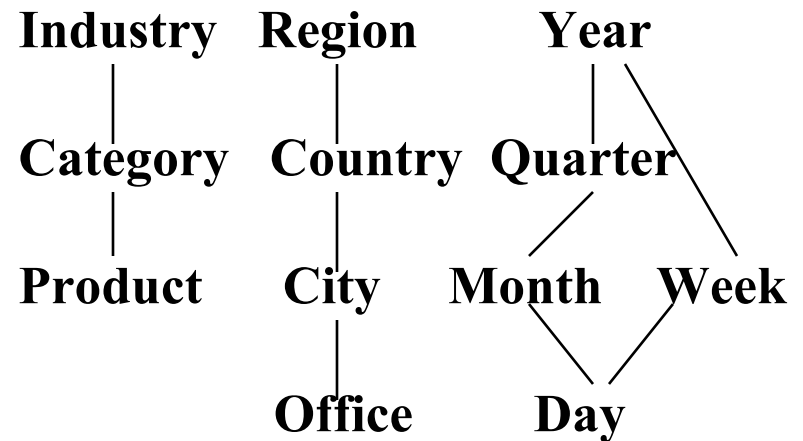
- **Distributive**: if the result derived by applying the function to  $n$  aggregate values is the same as that derived by applying the function on all the data without partitioning
  - E.g., `count()`, `sum()`, `min()`, `max()`
- **Algebraic**: if it can be computed by an algebraic function with  $M$  arguments (where  $M$  is a bounded integer), each of which is obtained by applying a distributive aggregate function
  - E.g., `avg() = sum() / count()`, `min_N()` ...
- **Holistic**: if there is no constant bound on the storage size needed to describe a subaggregate.
  - E.g., `median()`, `mode()`, `rank()`

# Multidimensional Data

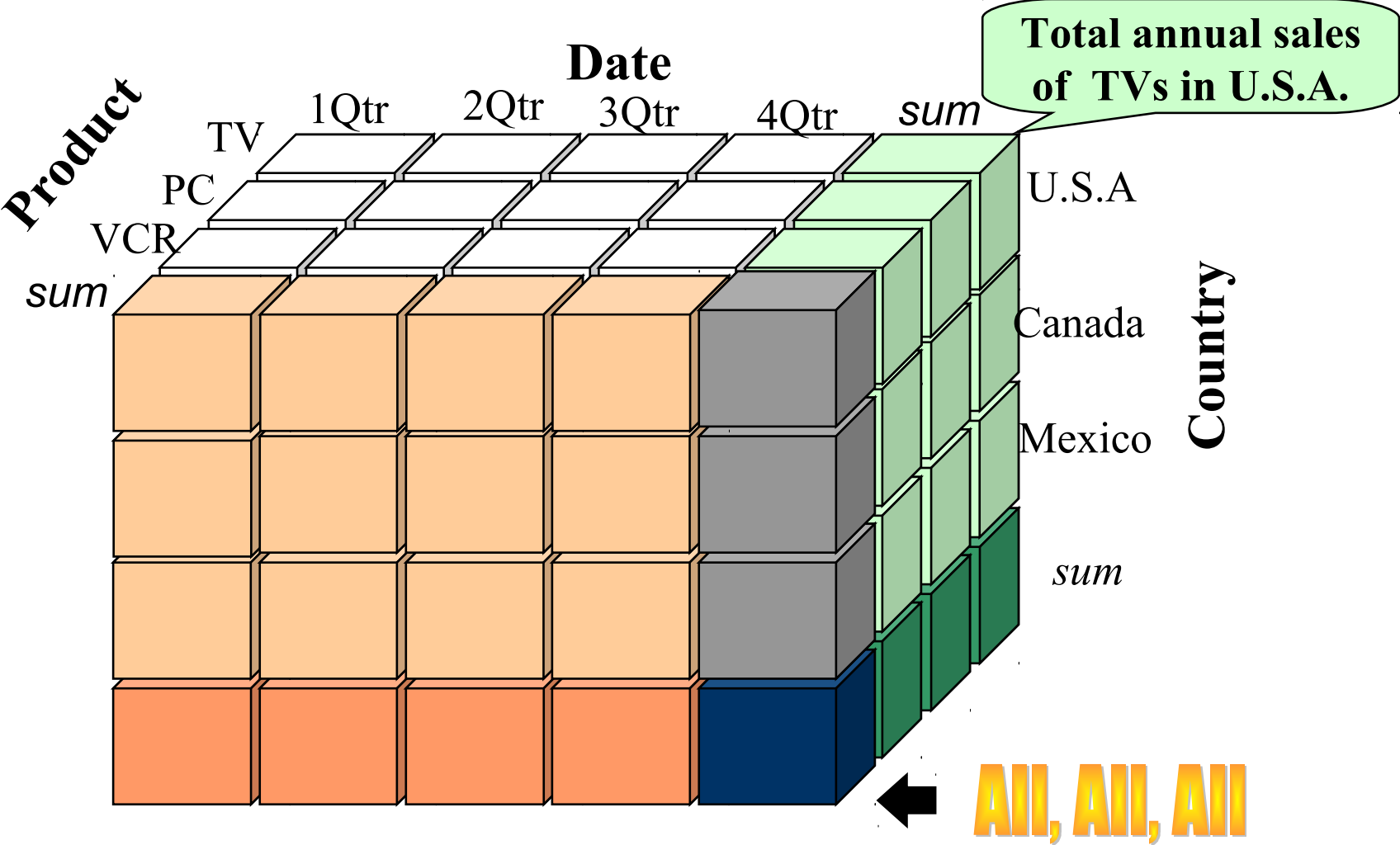
- Sales volume as a function of product, month, and region



**Dimensions:** *Product, Location, Time*  
**Hierarchical summarization paths**

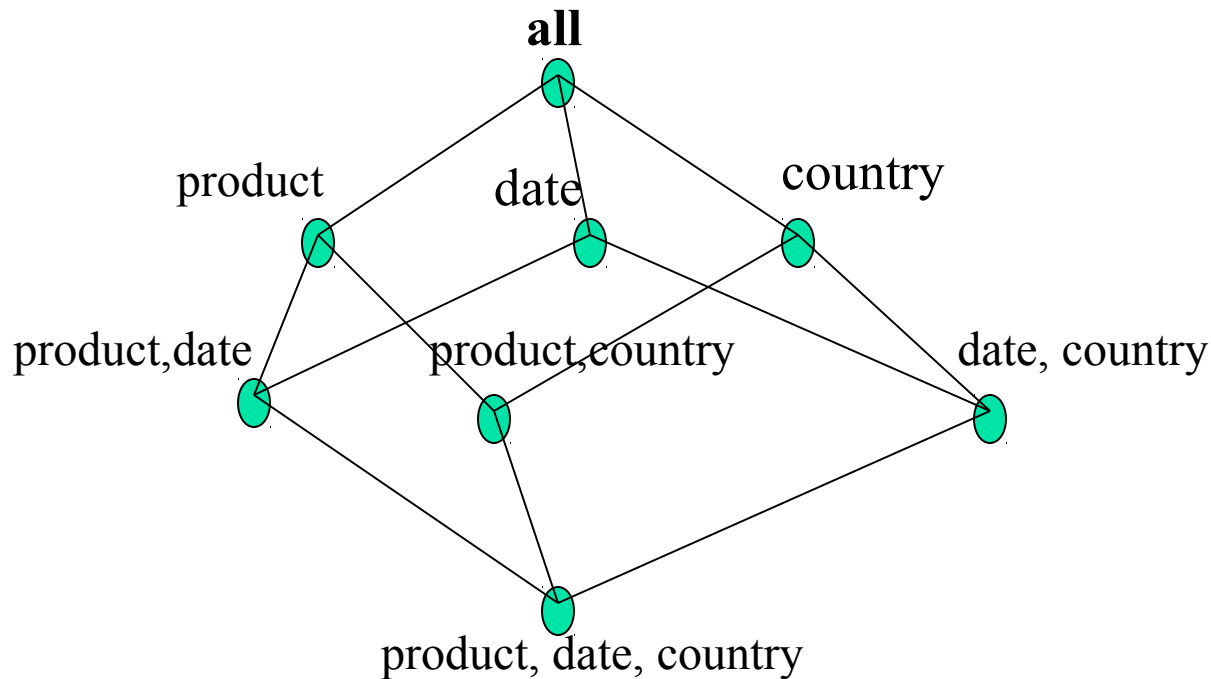


# A Sample Data Cube



# Cuboids Corresponding to the Cube

---



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

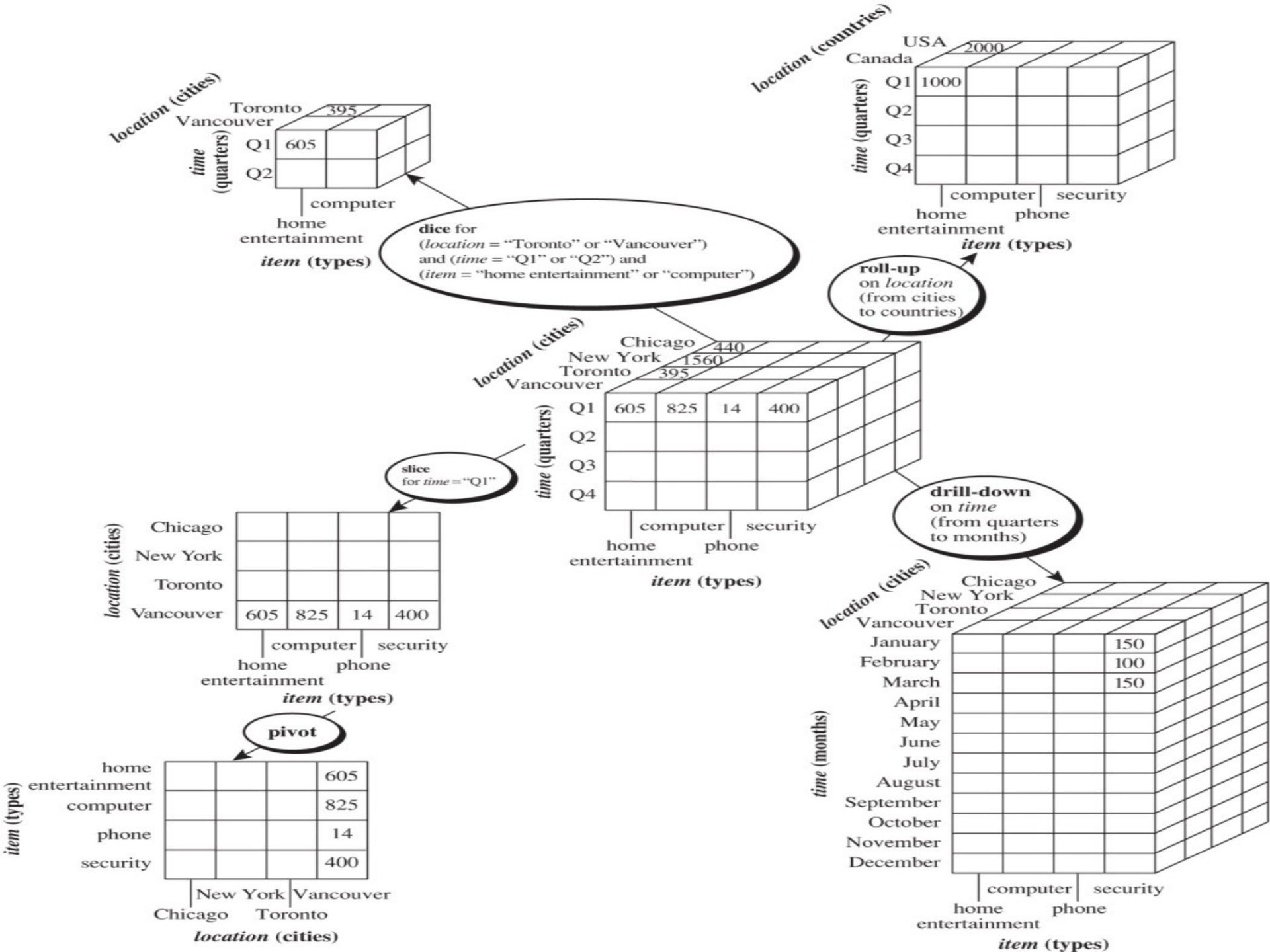
3-D (*base*) cuboid

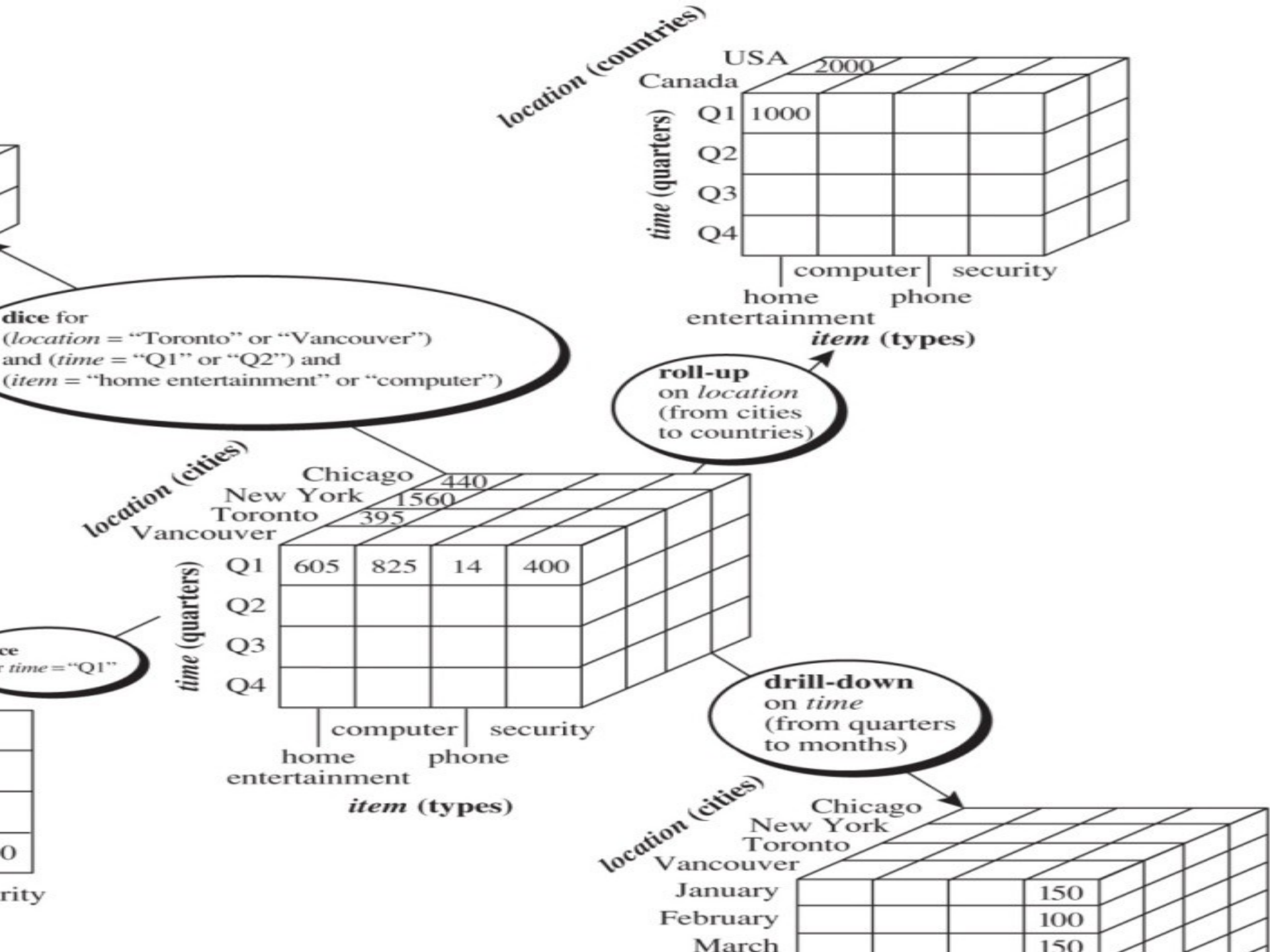


# Typical OLAP Operations

---

- **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*
- **Slice and dice:** *project and select*
- **Pivot (rotate):**
  - *reorient the cube, visualization, 3D to series of 2D planes*
- Other operations
  - ***drill across:*** *involving (across) more than one fact table*



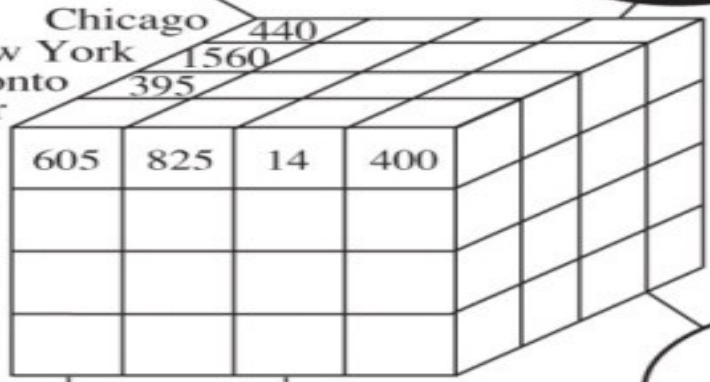


(item = "home entertainment" or "computer")

drill-up  
on location  
(from cities  
to countries)

location (cities)  
Chicago  
New York  
Toronto  
Vancouver

time (quarters)  
Q1  
Q2  
Q3  
Q4



computer security  
home entertainment phone  
item (types)

slice  
for time = "Q1"

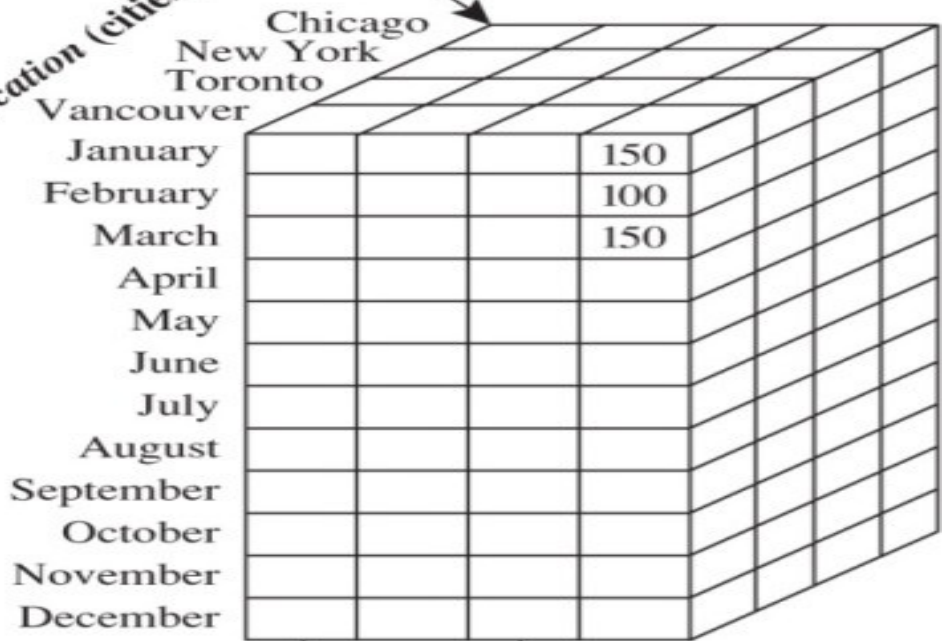


security

drill-down  
on time  
(from quarters  
to months)

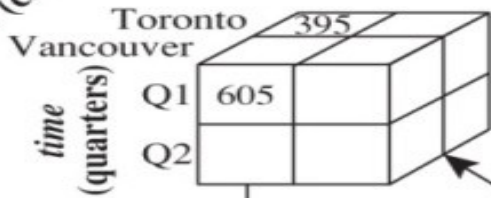
location (cities)  
Chicago  
New York  
Toronto  
Vancouver

time (months)  
January  
February  
March  
April  
May  
June  
July  
August  
September  
October  
November  
December



computer security  
home entertainment phone  
item (types)

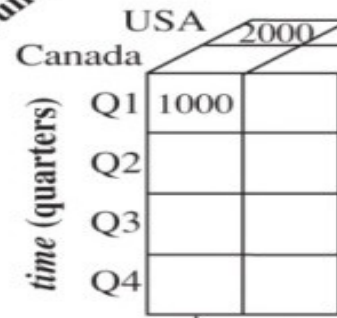
location (cities)



computer  
home entertainment  
item (types)

**dice** for  
 (location = "Toronto" or "Vancouver")  
 and (time = "Q1" or "Q2") and  
 (item = "home entertainment" or "computer")

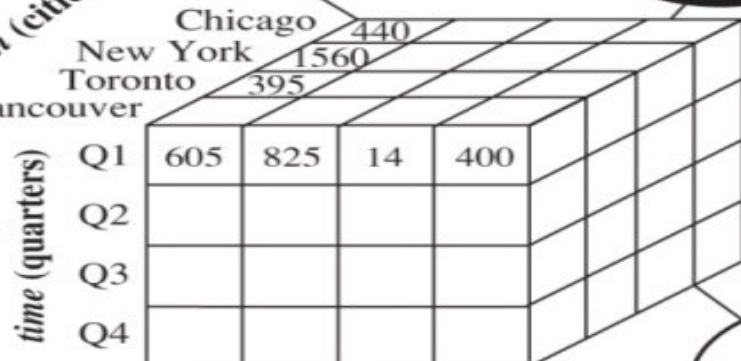
location (countries)



comput  
home entertainment  
item (types)

**roll-up**  
 on location  
 (from cities to countries)

location (cities)

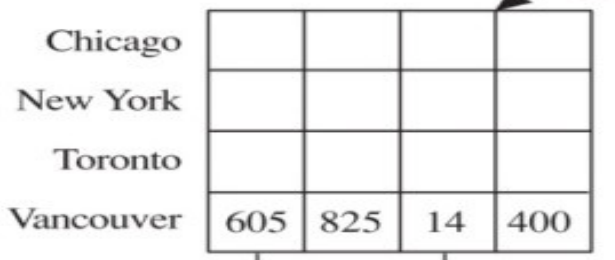


computer security  
home entertainment phone  
item (types)

**drill-down**  
 on time  
 (from quarters to months)

**slice**  
 for time = "Q1"

location (cities)



computer security  
home entertainment phone  
item (types)

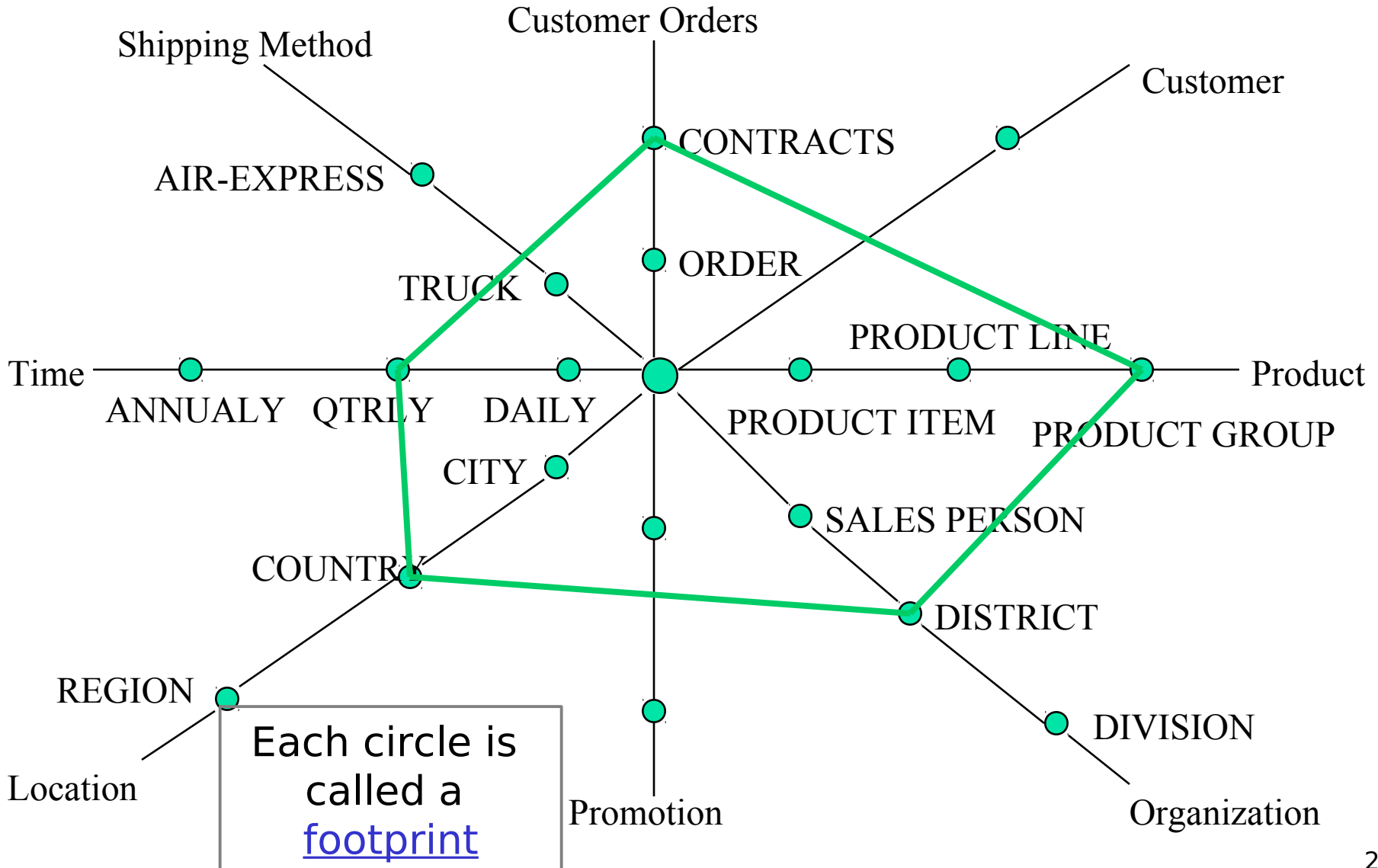
location (cities)



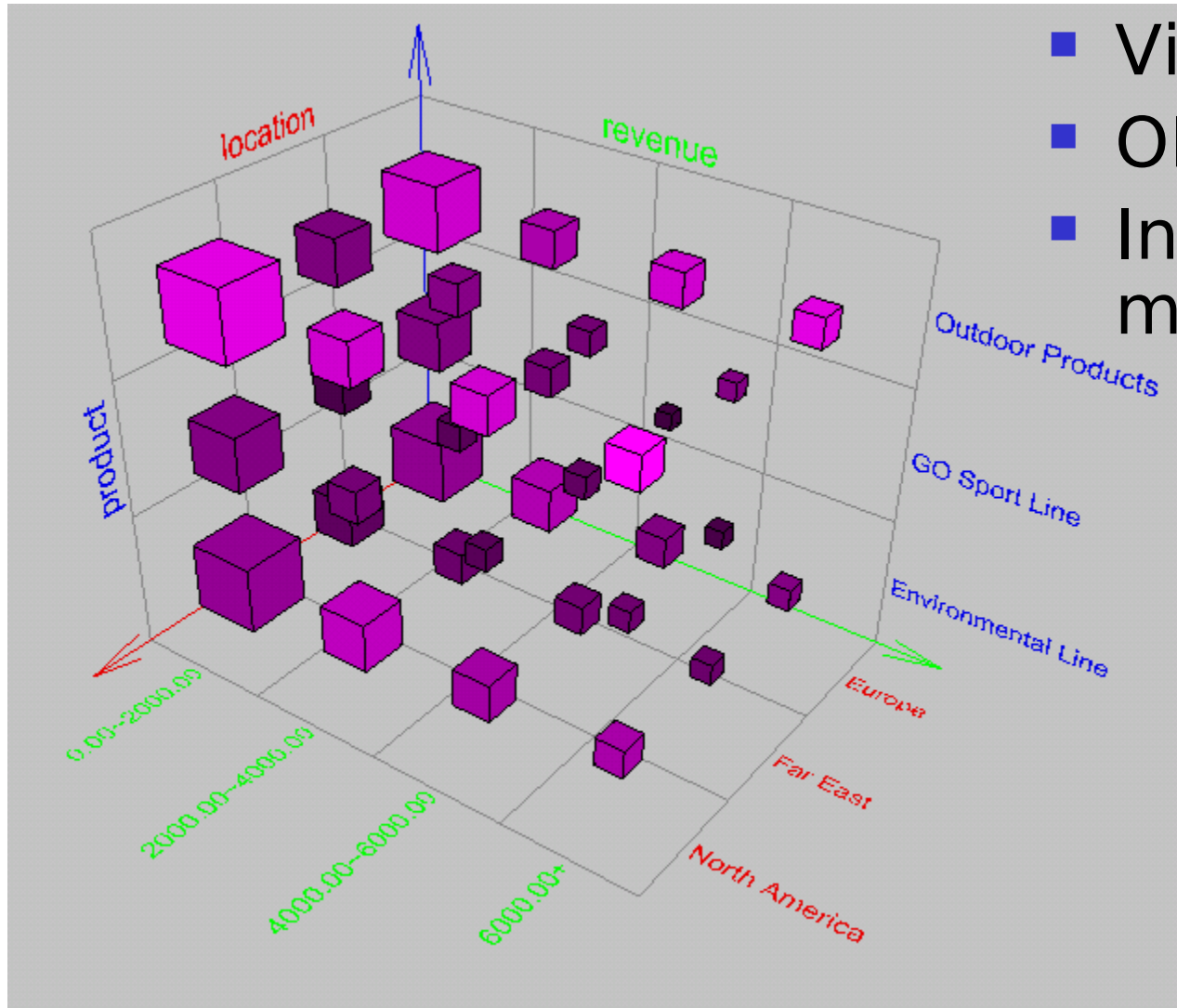




# A Star-Net Query Model



# Browsing a Data Cube




- Visualization
- OLAP capabilities
- Interactive manipulation



# Chapter 4: Data Warehousing and On-line Analytical Processing

---

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage 
- *Data Warehouse Implementation*
- *Data Generalization by Attribute-Oriented Induction*
- Summary

# Design of Data Warehouse: A Business Analysis Framework

---

- Four views regarding the design of a data warehouse
  - **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 Design Process

---

- **Top-down, bottom-up approaches or a combination** of both
  - Top-down: Starts with overall design and planning (mature)
  - Bottom-up: Starts with experiments and prototypes (rapid)
- **From software engineering point of view**
  - (1) planning (2) requirements study (3) problem analysis (4) warehouse design (5) data integration and testing (6) deployment
  - Waterfall: structured and systematic analysis at each step before proceeding to the next (better for data warehouse)
  - Spiral: rapid generation of increasingly functional systems, short turn around time, quick turn around (better for data marts)

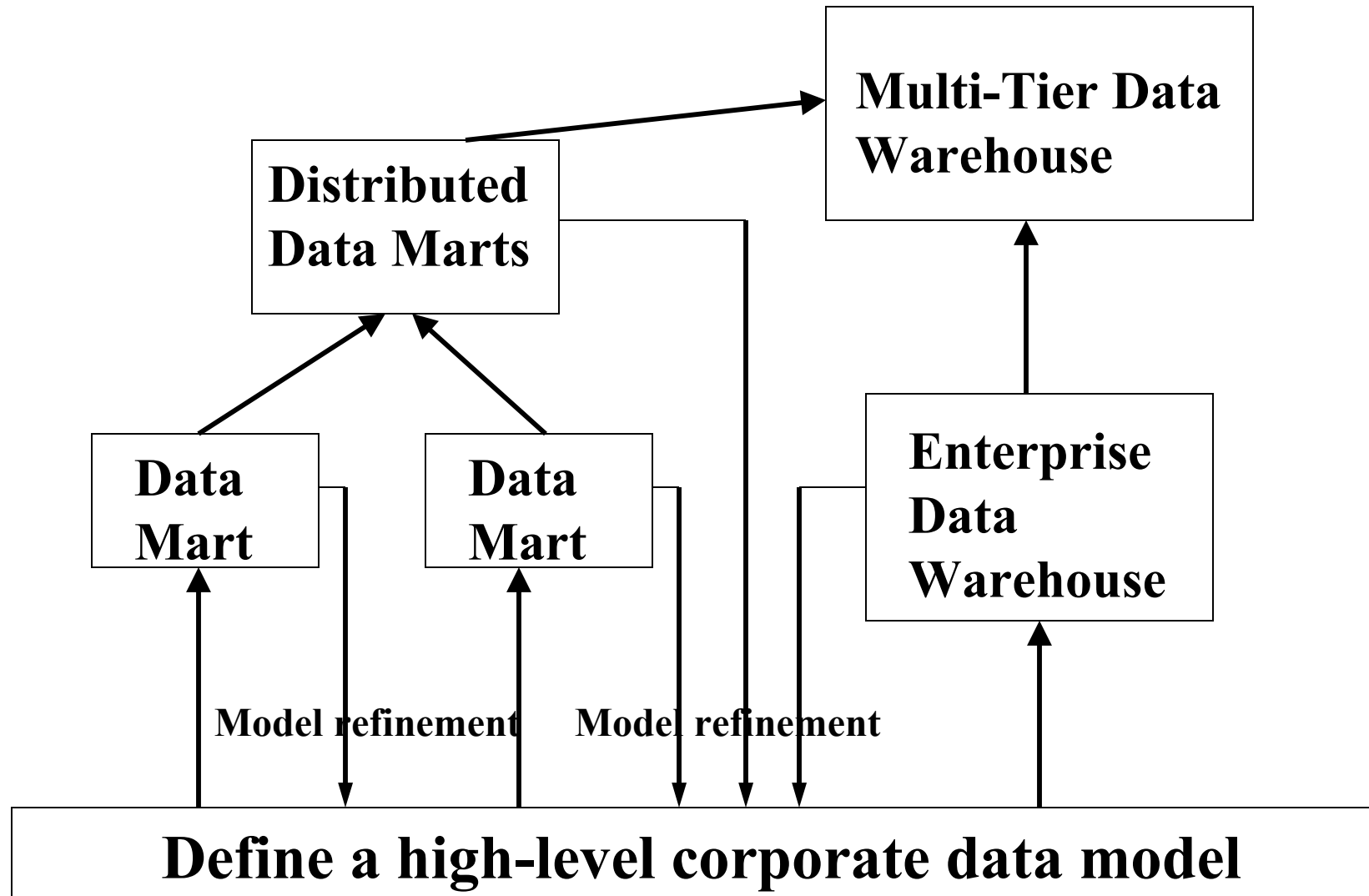
# Data Warehouse Design Process

---

- **Typical data warehouse design process**
  - Choose a **business process** to model, e.g., orders, invoices, etc.
  - Choose the ***grain (atomic level of data)*** of the business process
  - Choose the **dimensions** that will apply to each fact table record
  - Choose the **measure** that will populate each fact table record

# Data Warehouse Development: A Recommended Approach

---



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

# From On-Line Analytical Processing (OLAP) to On Line Analytical Mining (OLAM)

---

- Why **online analytical mining**?
  - High quality of data in data warehouses
    - DW contains integrated, consistent, cleaned data
  - Available information processing structure surrounding data warehouses
    - ODBC, OLEDB, Web accessing, service facilities, reporting and OLAP tools
  - OLAP-based exploratory data analysis
    - Mining with drilling, dicing, pivoting, etc.
  - On-line selection of data mining functions
    - Integration and swapping of multiple mining functions, algorithms, and tasks