CS 412 Intro. to Data Mining

Chapter 4. Data Warehousing and On-line Analytical Processing

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017

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
- Summary

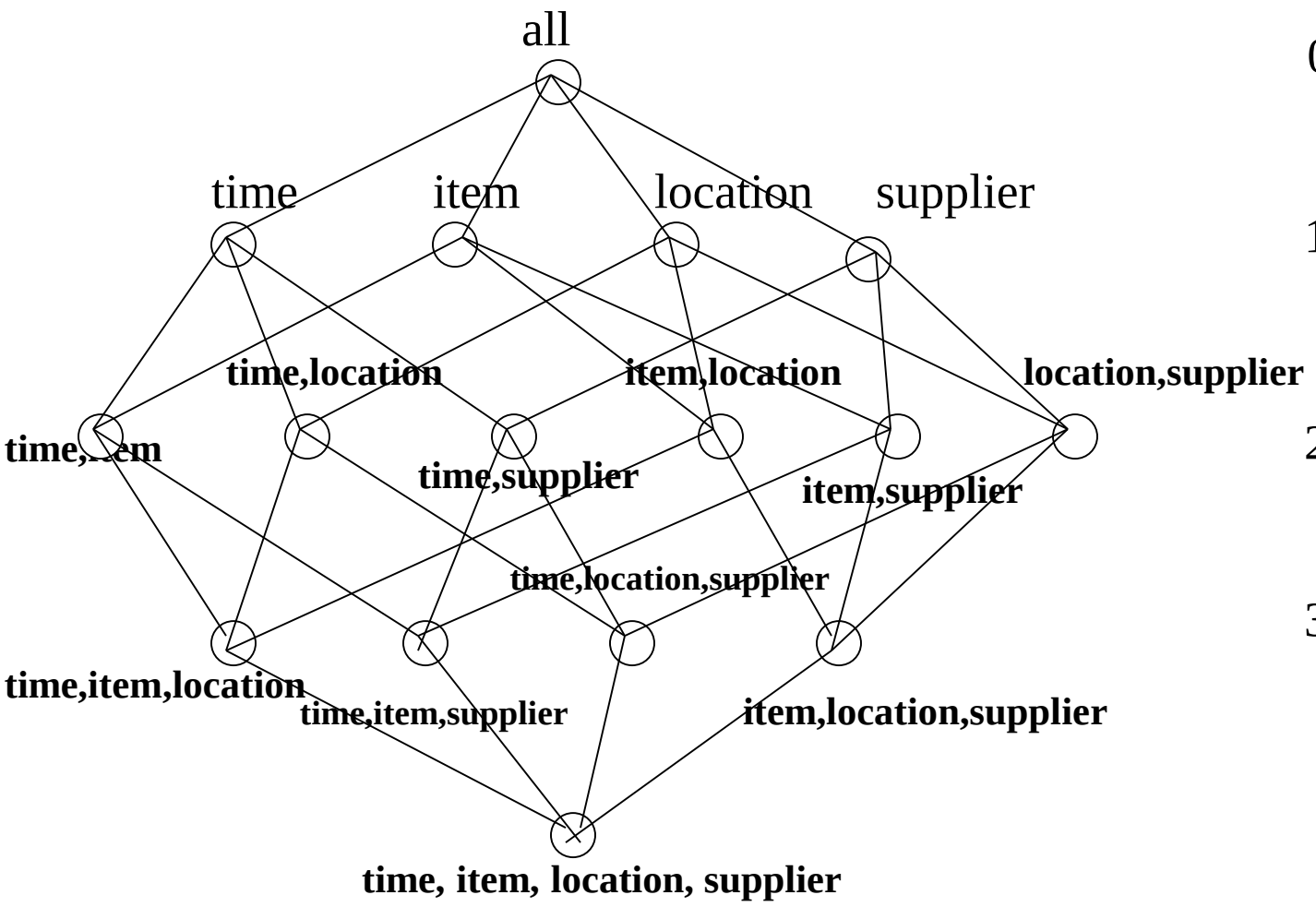
Chapter 4: Data Warehousing and On-line Analytical Processing

- Data Warehouse: Basic Concepts
- Data Warehouse Modeling: Data Cube and OLAP
- Data Warehouse Design and Usage
- Data Warehouse Implementation
- 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) → ใช้อธิบายซ้ำ สามารถจำแนกได้มีเป็นอย่างไร
  - **Fact table** contains **measures** (such as dollars_sold) and keys to each of the related dimension tables
    ↳ การเก็บตัวเลข
- **Data cube**: A lattice of cuboids
  - 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



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

3-D cuboids

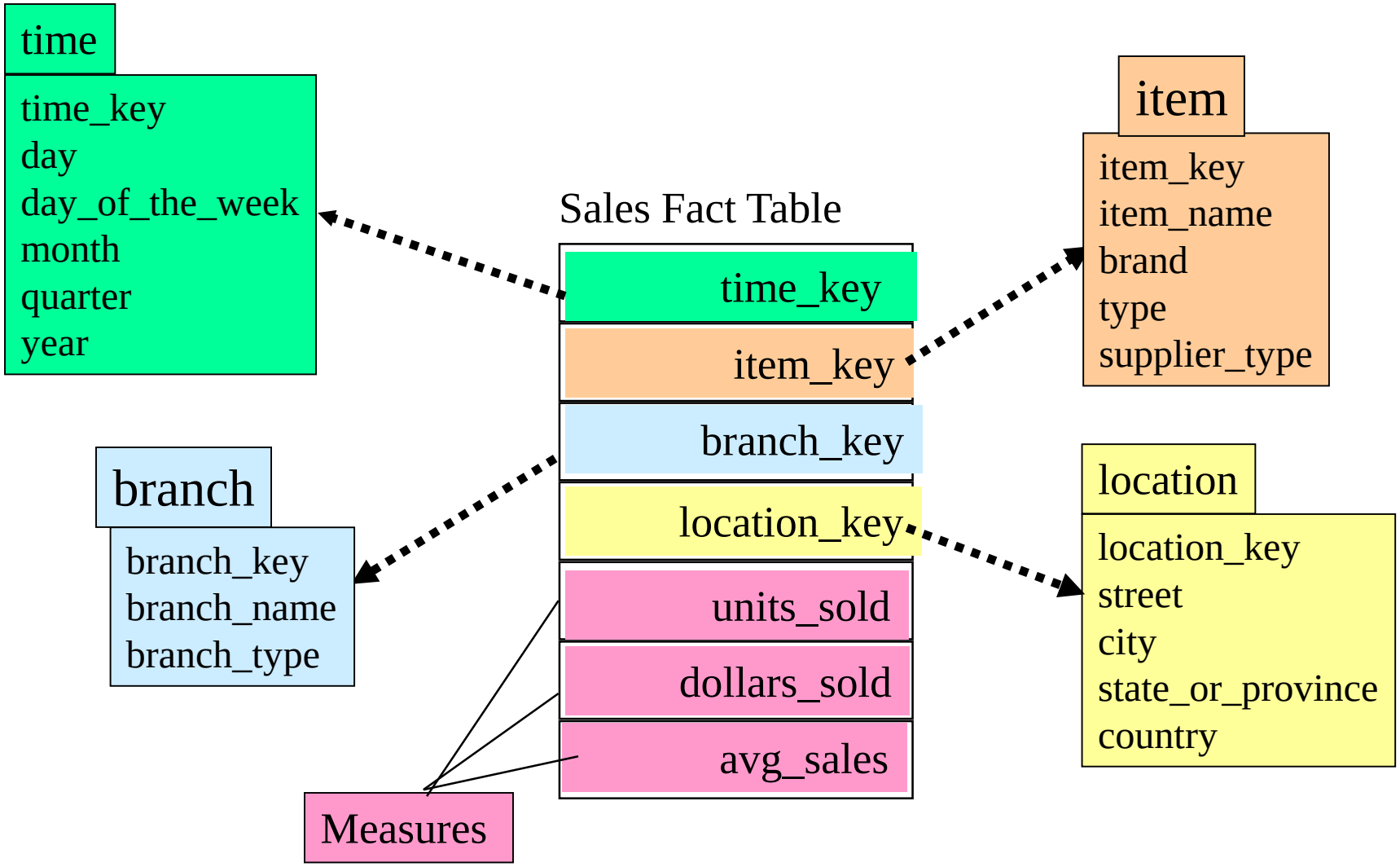
4-D (*base*) cuboid

17

Conceptual Modeling of Data Warehouses

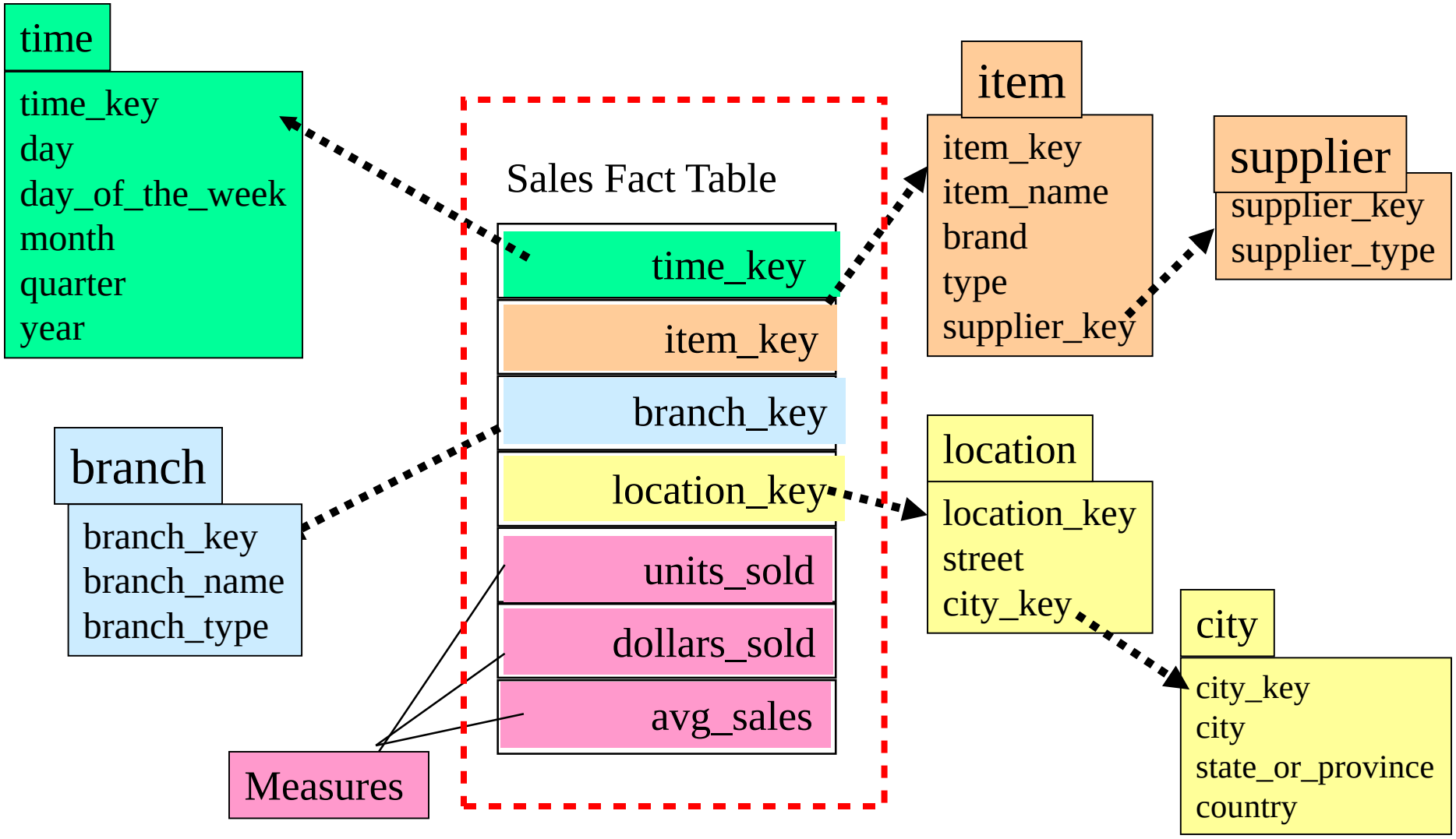
- Modeling data warehouses: dimensions & measures
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake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

Star Schema: An Example



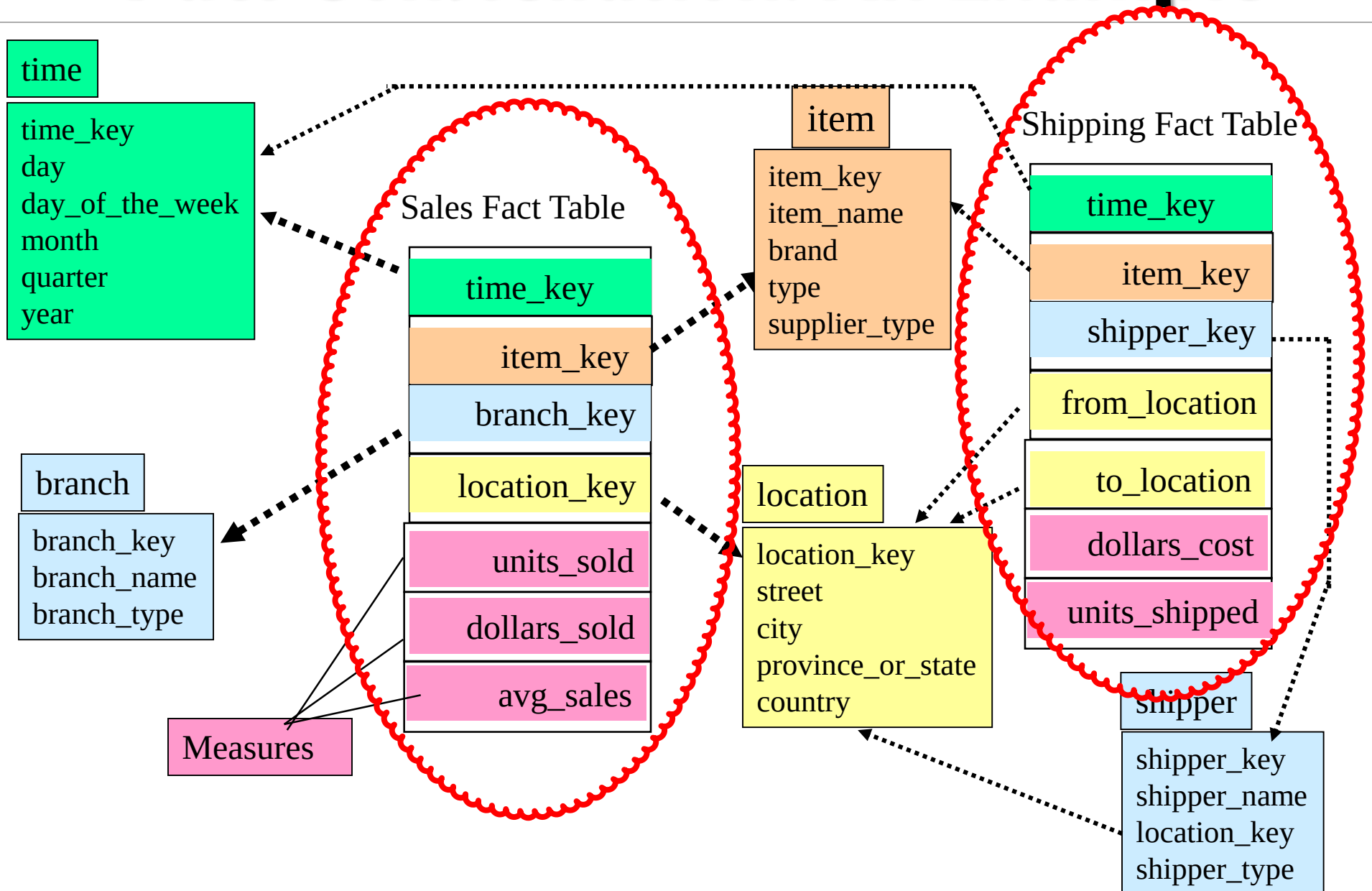
19

Snowflake Schema: An Example



กลุ่มดาว

Fact Constellation: An Example



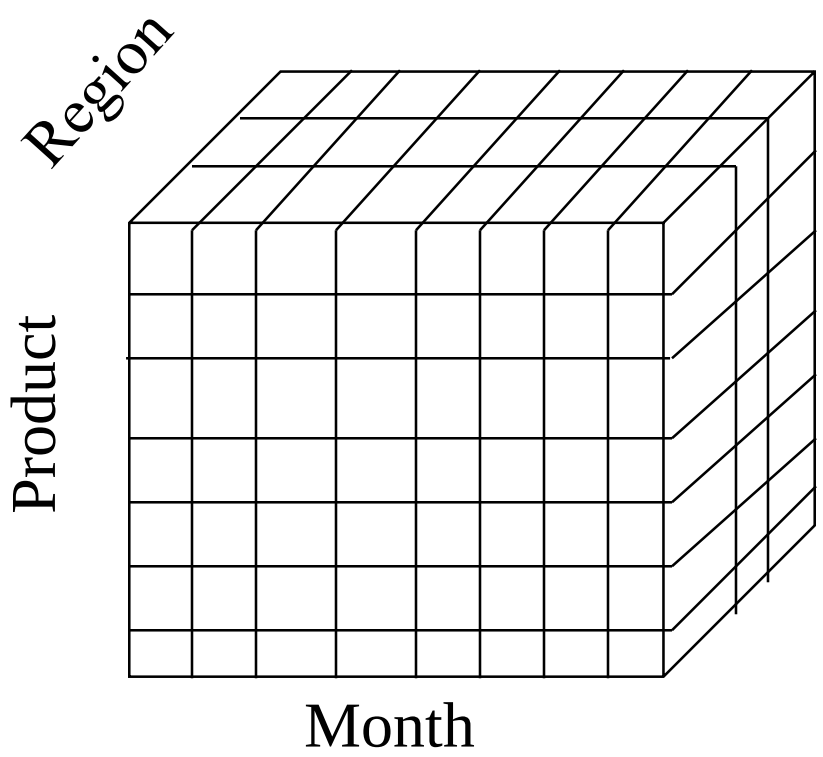
Data Cube Measures: Three Categories

- 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
  - avg(x) = sum(x) / count(x)
  - Is min_N() an algebraic measure? How about standard_deviation()?
- Holistic: if there is no constant bound on the storage size needed to describe a subaggregate.
  - E.g., median(), mode(), rank()

Multidimensional Data

สมมุติเป็นฟังก์ชันของ สินค้า เดือน ชาติ

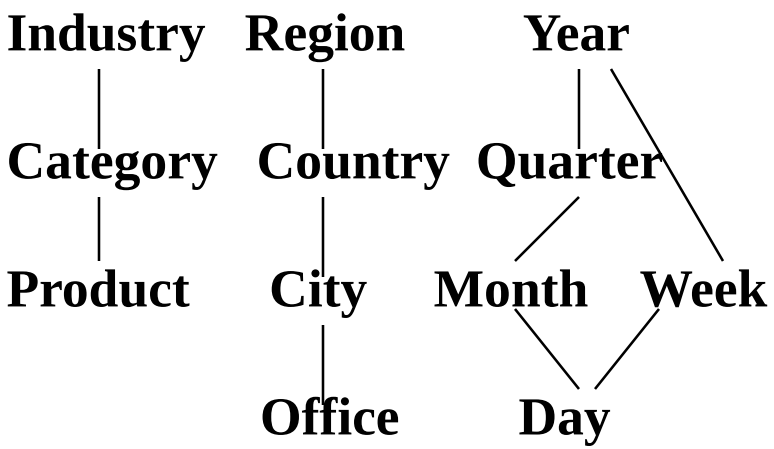
- Sales volume as a function of product, month, and region



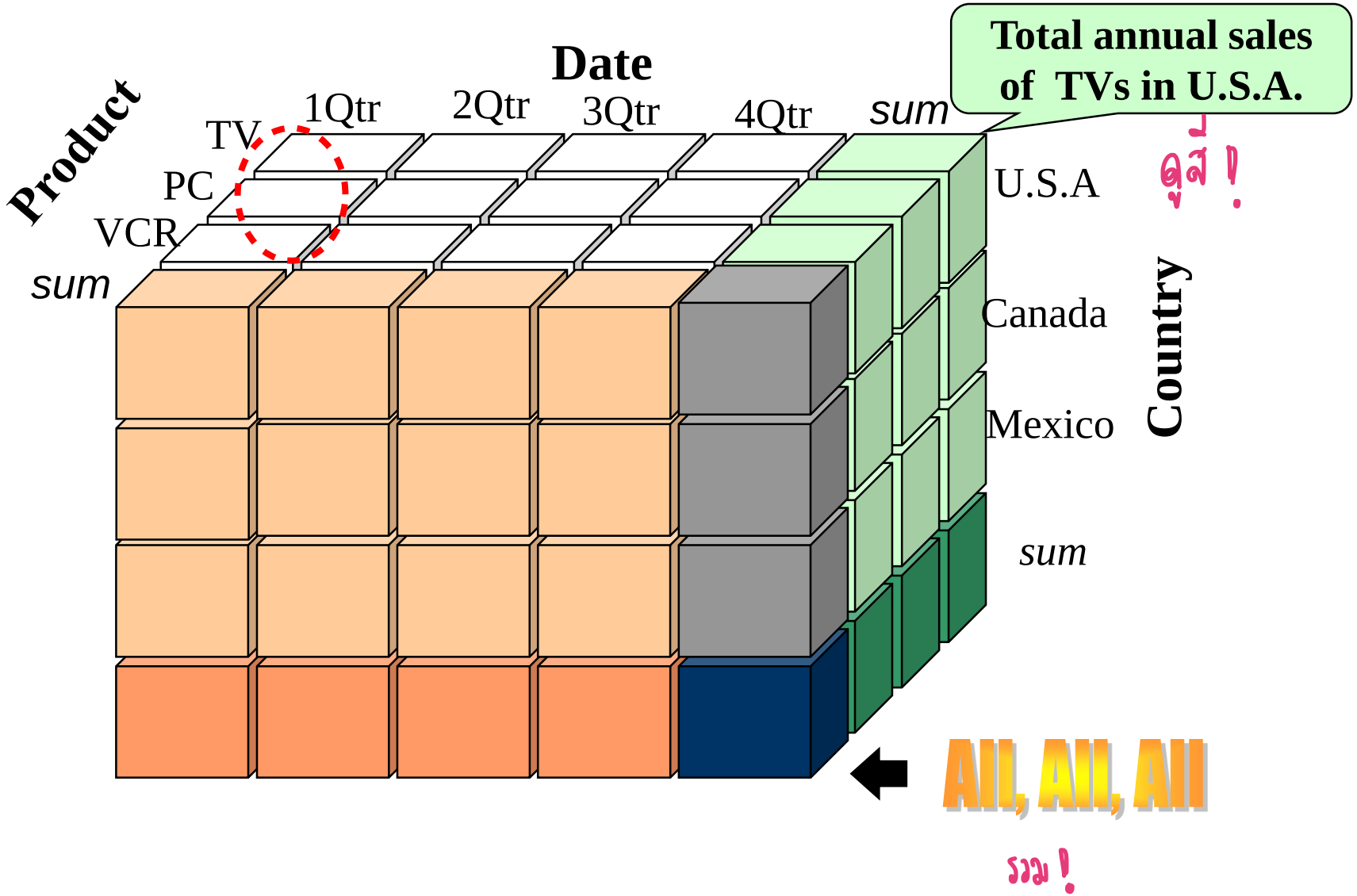
Dimensions: ***Product, Location, Time***

Hierarchical summarization paths

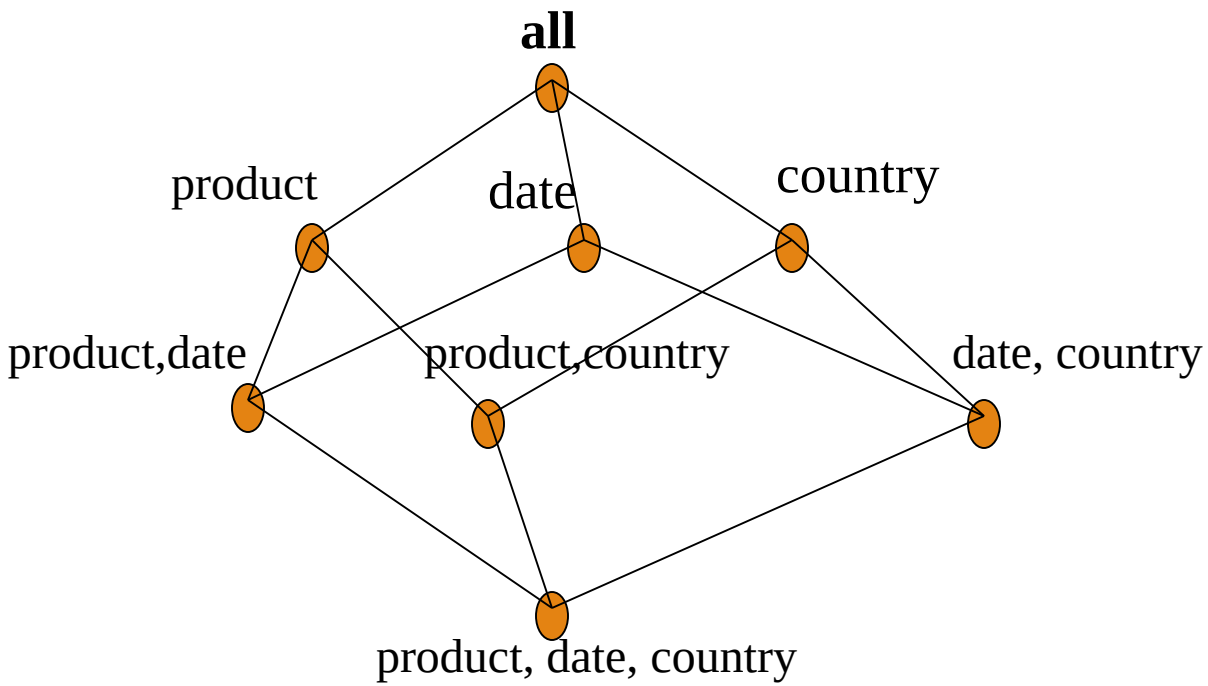
↳ มองให้ละเอียดมากขึ้น



A Sample Data Cube



Cuboids Corresponding to the Cube



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

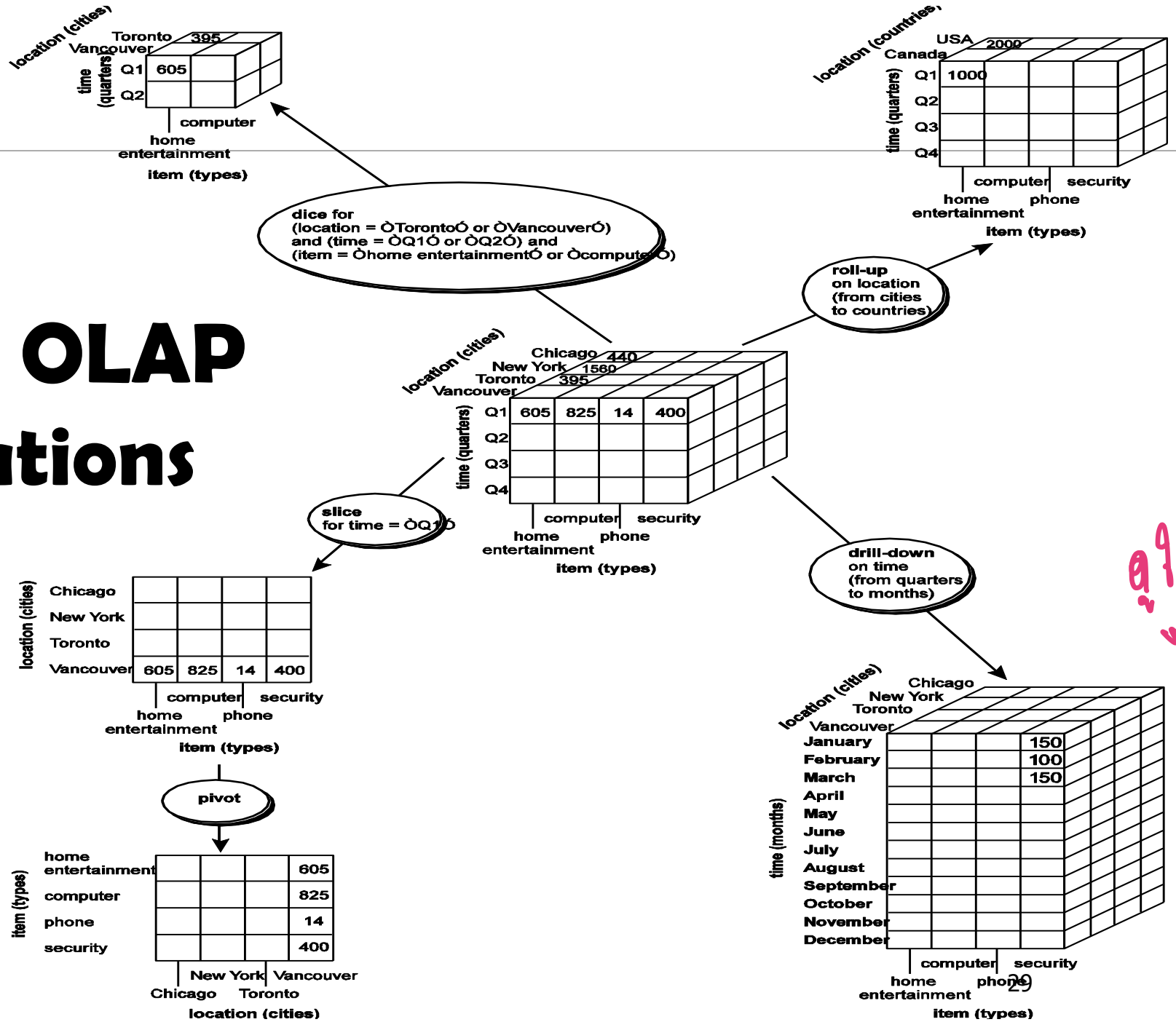
3-D (*base*) cuboid

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*
  - *Drill through: through the bottom level of the cube to its back-end relational tables (using SQL)*

- ข้อมูลใหญ่ๆ ที่ถูกตัดออกให้เป็นข้อมูลเล็กๆ "dice"

Typical OLAP Operations



จากที่แยกอยู่ จะถูกยุบรวม "Roll-up"

สนใจเฉพาะ "slice"

ดูให้ละเอียดขึ้น "drill-down"

A Star-Net Query Model