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

What is a Data Warehouse?

- ❑ Defined in many different ways, but not rigorously
  - ❑ 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

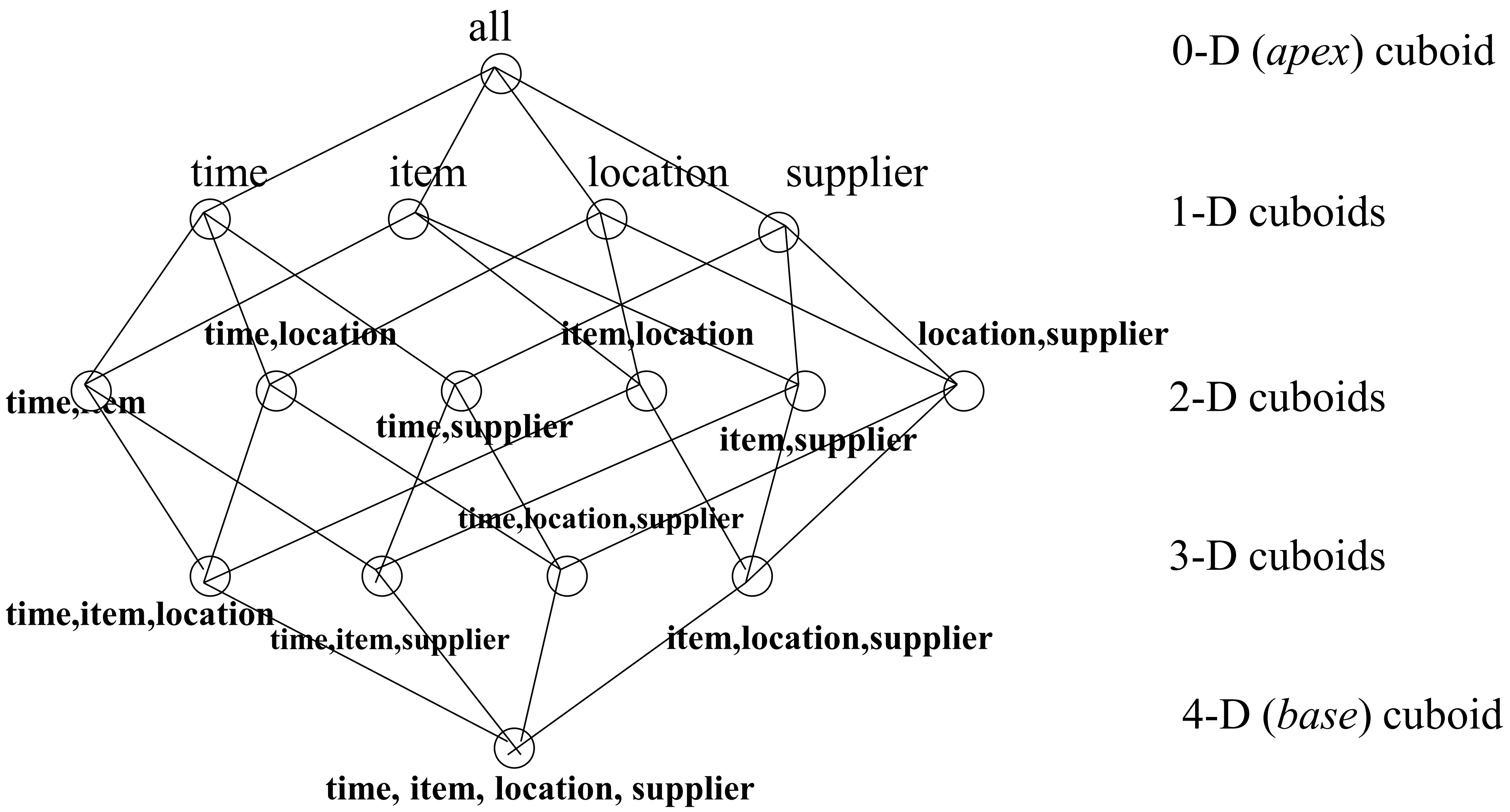
สร้างเพื่ออะไร มาจากหลายแหล่ง

- ❑ “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

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
- **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

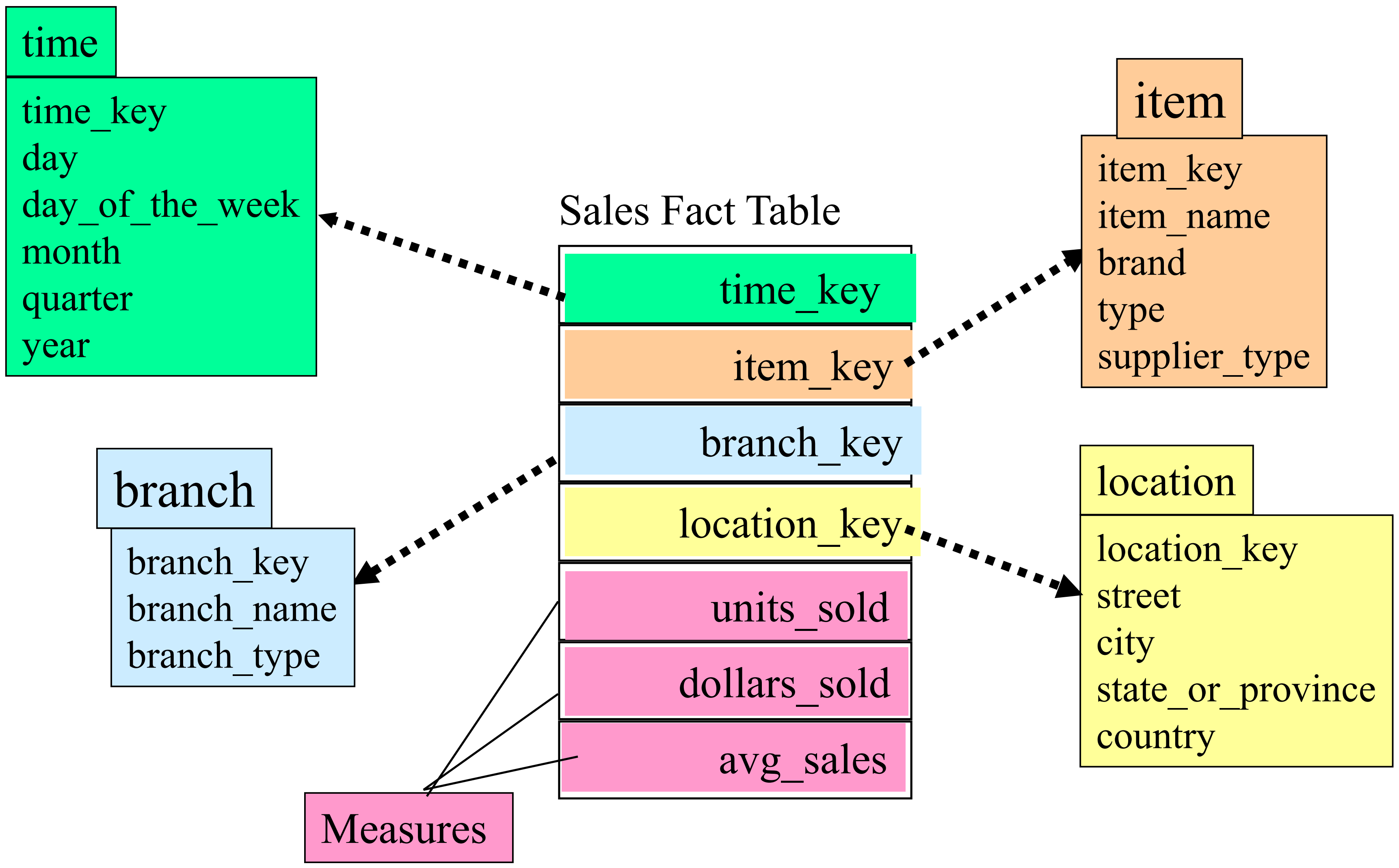


Conceptual Modeling of Data Warehouses

- 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

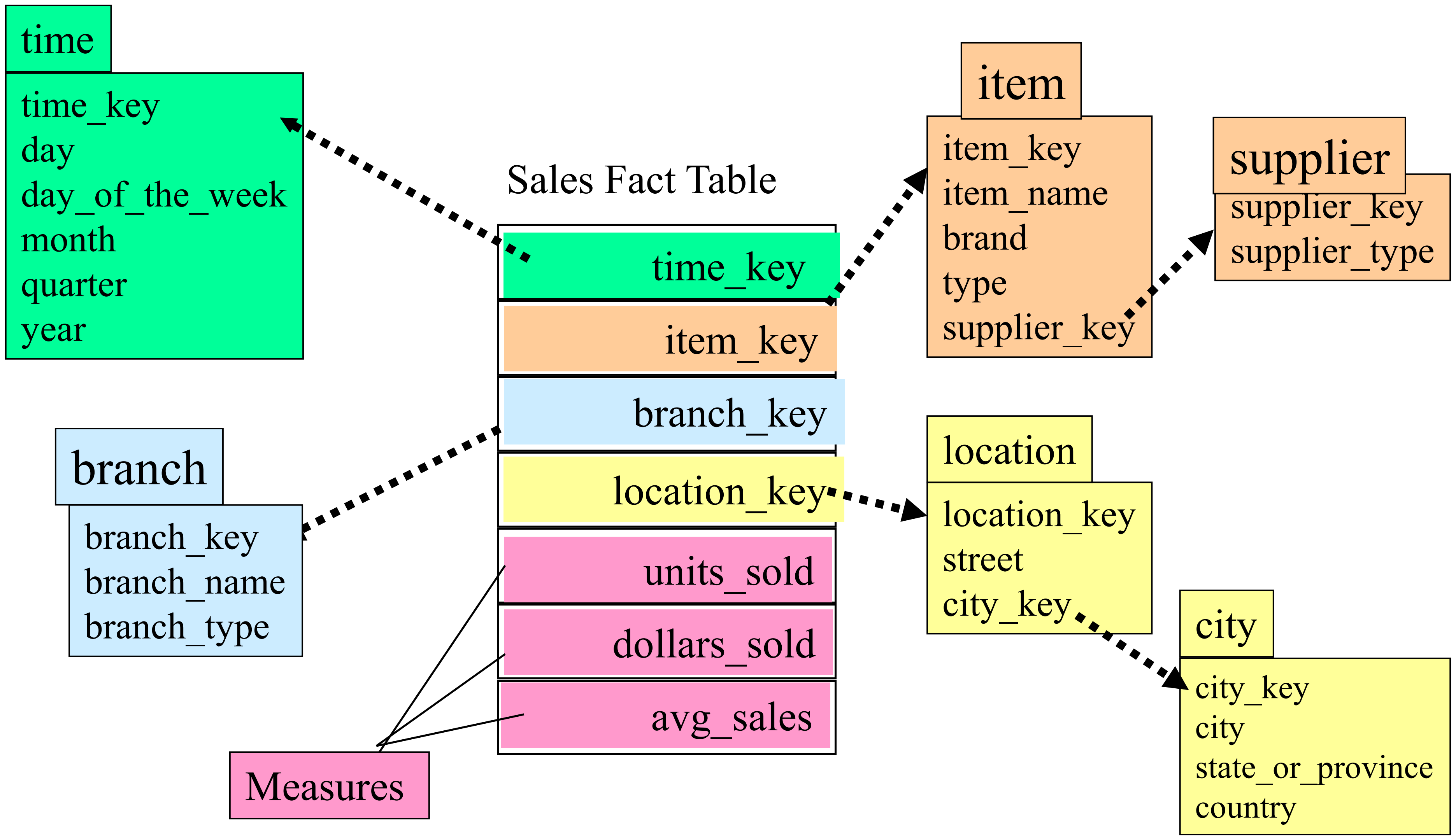
Star Schema: An Example

ศูลดตุนร์ กลางทุกอย่าง แตกออกมาหมด ศูนย์กลาง



Snowflake Schema: An Example

คล้ายเกล็ดหิมะ



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

Measures

branch
branch_key
branch_name
branch_type

item
item_key
item_name
brand
type
supplier_key

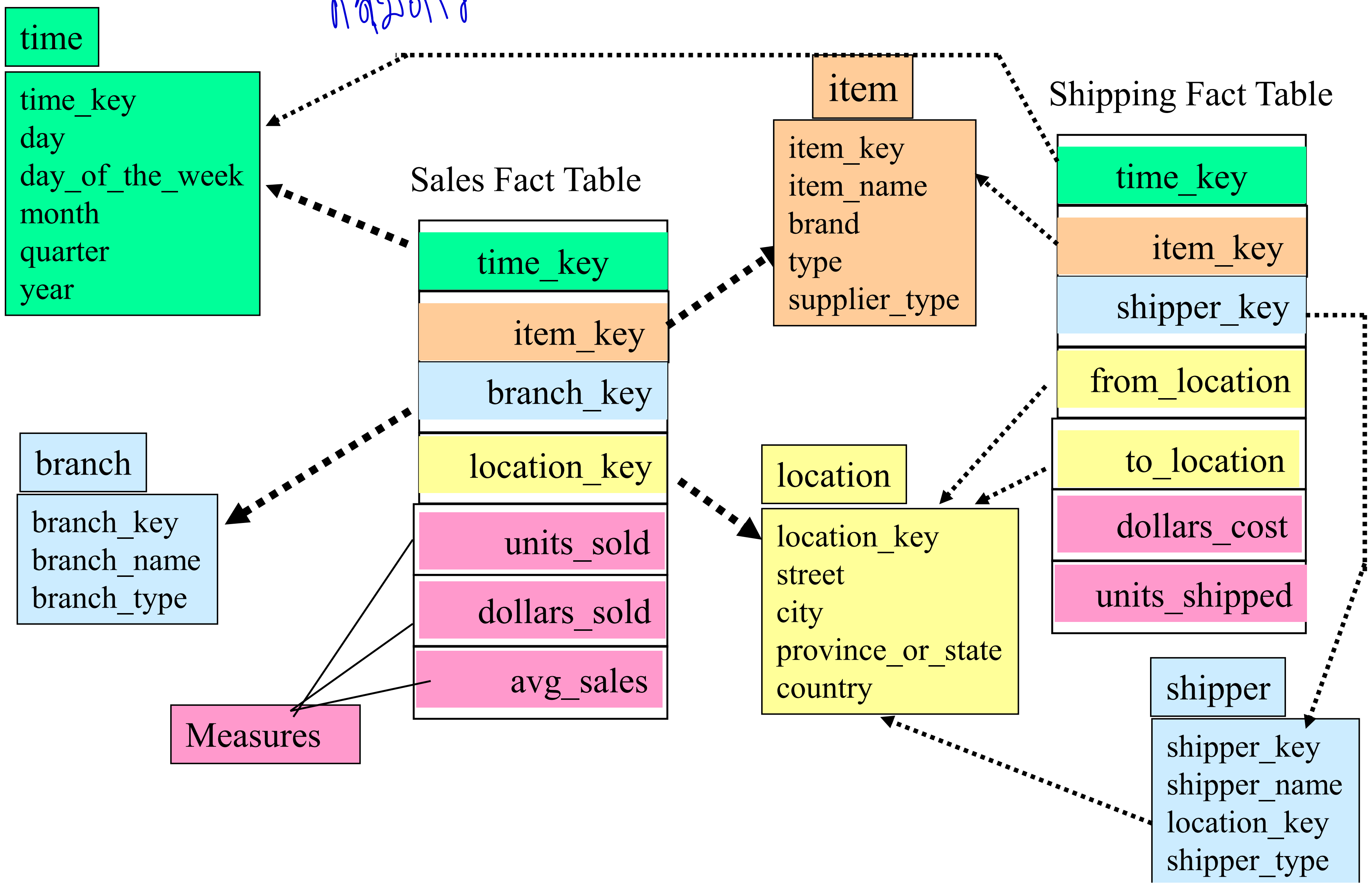
supplier
supplier_key
supplier_type

location
location_key
street
city_key

city
city_key
city
state_or_province
country

Fact Constellation: An Example

กลุ่มดาว



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

Measures

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

branch

branch_key
branch_name
branch_type

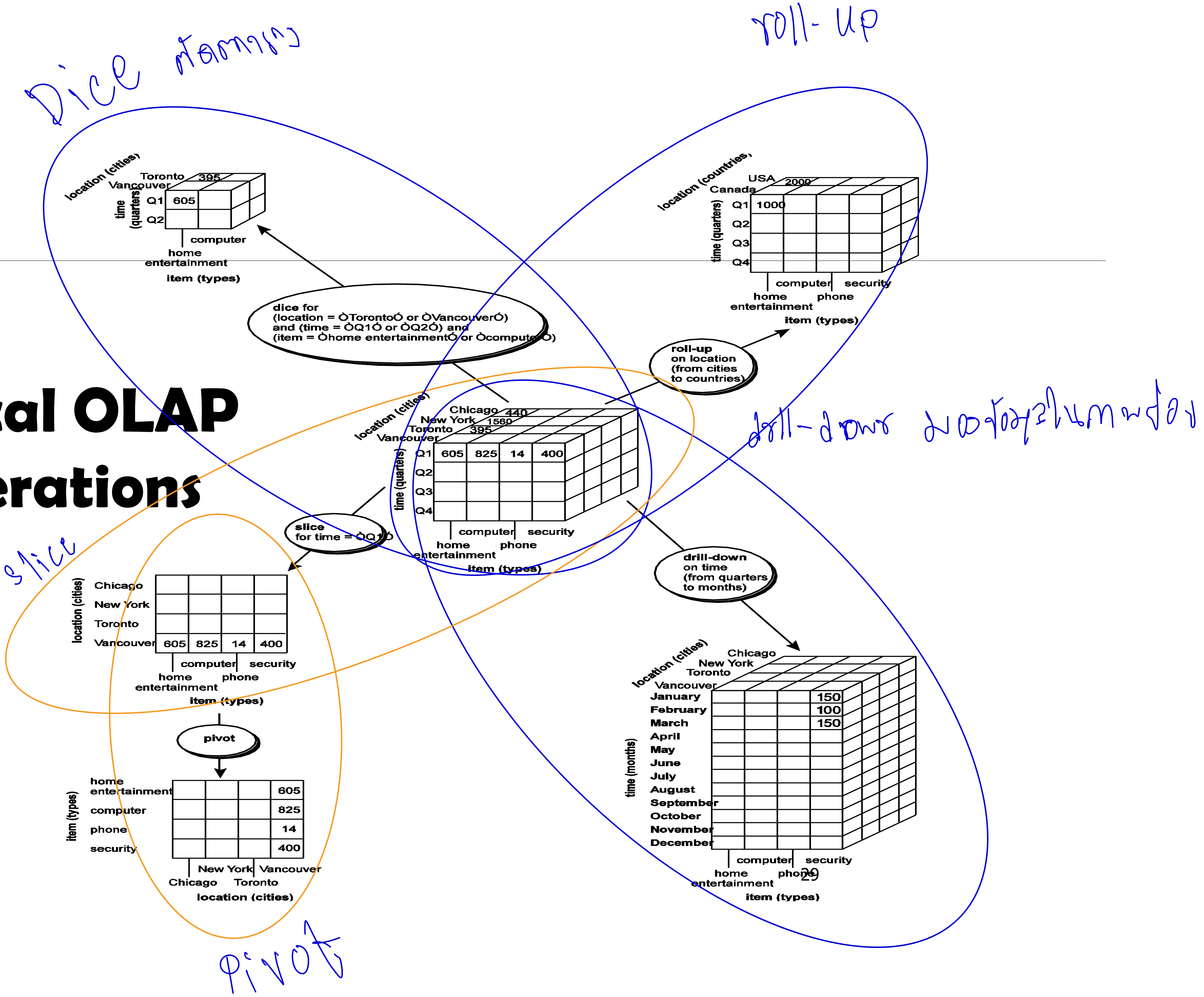
location

location_key
street
city
province_or_state
country

shipper

shipper_key
shipper_name
location_key
shipper_type

Typical OLAP Operations



location (cities)	home entertainment	computer	phone	security
Chicago				
New York				
Toronto				
Vancouver	605	825	14	400

item (types)

item (types)	Chicago	New York	Toronto	Vancouver
home entertainment				605
computer				825
phone				14
security				400

location (cities)