~ Chapter 4 ~

why is Business Intelligence useful for a Data Scientist ?

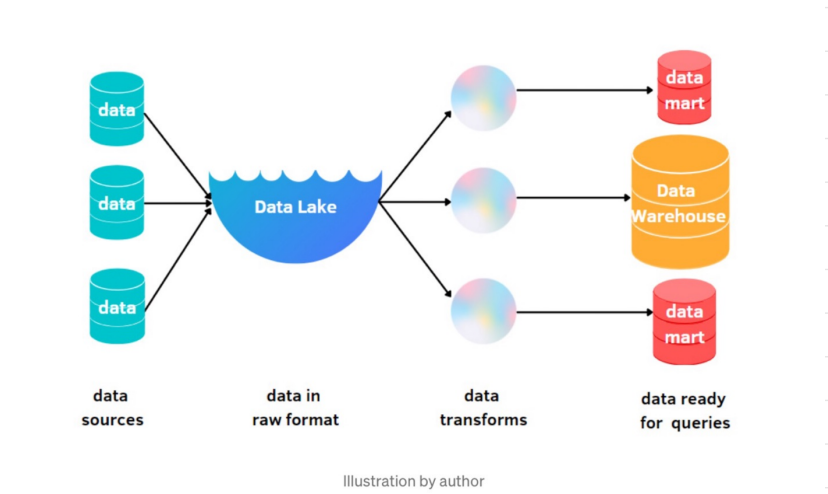


Illustration by author

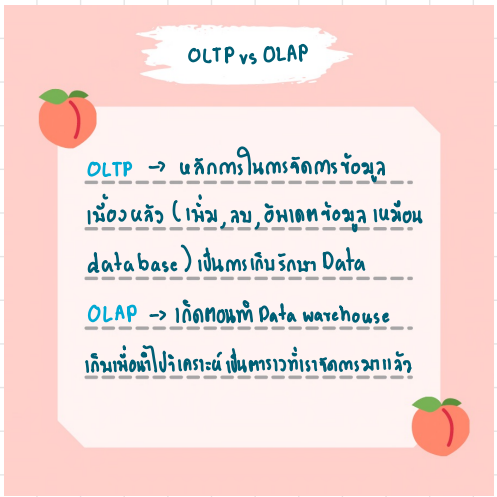
structured → Data ที่ถูกเก็บไว้ใน Warehouse

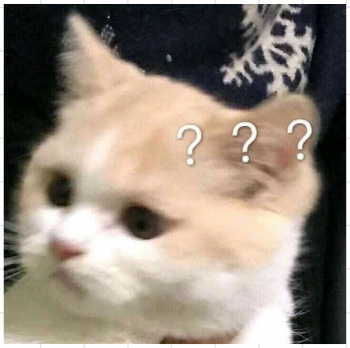
Unstructured → Data ที่คอมพิวเตอร์ไม่สามารถเข้าใจได้ทันที เช่น รูป เสียง ข้อความ (คอมเมนต์)

ศัพท์ที่ควรรู้

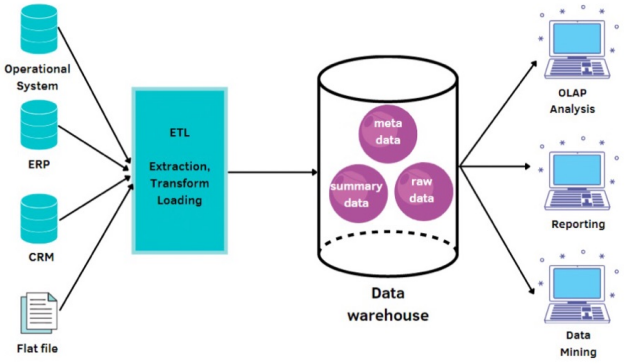
Table of Content:

1. Data Warehouse
2. Data Mart
3. OLTP vs OLAP
4. ETL
5. Star vs Snowflake schemas
6. Data Lake
7. From ETL to ELT
8. Batch vs stream Processing





1. Data Warehouse

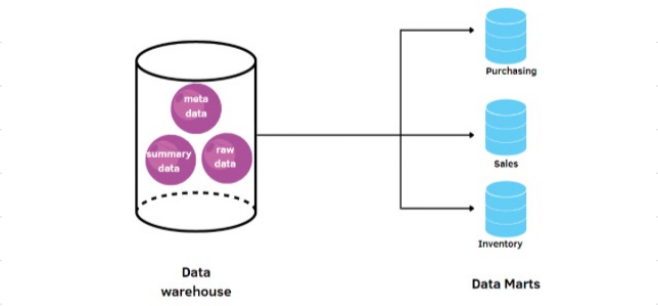


Data Warehouse architecture. Illustration by author.

The data warehouses store three types of data :

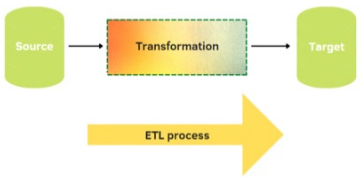
- **Metadata** → ข้อมูลที่อธิบายข้อมูลเกี่ยวกับคลังข้อมูลและรายละเอียดเพื่อให้เข้าใจคลังข้อมูลที่จัดเก็บ
- **Summary data** → ข้อมูลที่รวบรวม / สรุปที่สร้างโดยผู้จัดการของคลังข้อมูล ช่วยเร่งประสิทธิภาพการสืบค้นข้อมูล
- **Raw data** → ข้อมูลดิบเป็นข้อมูลที่ไม่ได้ผ่านการประมวลผล

2. Data Marts



Data warehouse vs Data Marts. Illustration by author.

3. ETL

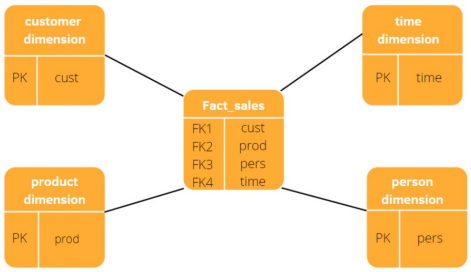


ETL process. Illustration by author

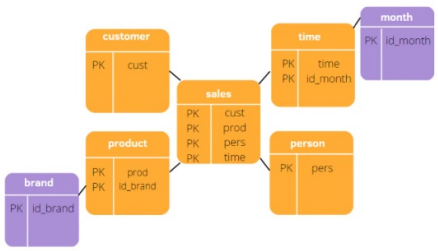
ETL : เป็นกระบวนการรวบรวมและประมวลผลที่ช่วยให้สามารถสร้างคลังข้อมูลได้

- **Extract** : ขั้นตอนการดึงข้อมูลจากแหล่งต่างๆ เช่น ไฟล์ csv, json, xml
- **Transform** : การแปลงข้อมูลให้อยู่ในรูปแบบที่เป็นประโยชน์สำหรับการวิเคราะห์ รวมถึง การล้าง ตรวจสอบ การรวม และการลบรายการที่ซ้ำกัน
- **Load** : ข้อมูลจะถูกบันทึกลงในปลายทางสุดท้ายนั่นคือคลังข้อมูล เกิดขึ้นได้ทุกนาที ชั่วโมง วัน สัปดาห์ ขึ้นอยู่กับว่าข้อมูลมีการเปลี่ยนแปลงบ่อยแค่ไหน การวิเคราะห์ก็จะแม่นยำมากขึ้นเท่านั้น

4. Star vs Snowflake Schemas



Star Schema. Illustration by author



Snowflake Schema. Illustration by author

Data warehouse มันเก็บข้อมูลโดยใช้ Schemas นายมีที่ Schemas มีประโยชน์ในการจัดการข้อมูลจำนวนมากเพื่อวัตถุประสงค์ในการวิเคราะห์ มี schemas ที่ใช้แพร่หลาย สองแบบ

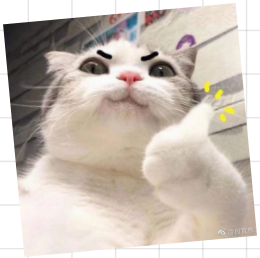
- **Star Schema** → มีโครงสร้างคล้ายดาวตรงกลางมี fact table โดยมี Dimension tables อยู่รอบๆ
- **Snowflake Schema** → มีโครงสร้างเหมือนเกล็ดหิมะ ข้อดีคือช่วยเพิ่มขนาดและใช้พื้นที่จัดเก็บน้อยลง

5. Data Lake

DATA WAREHOUSE	DATA LAKE
structured, processed	structured / semi-structured/ unstructured, raw
schema-on-write	schema-on-read
expensive for large data volumes	designed for low-cost storage
less agile, fixed configuration	highly agile, possible updates
mature	maturing
business professionals	data scientists

Comparison of Data Warehouse and Data Lake. Illustration by author.

- **Data Lake** มีความซับซ้อนน้อยกว่า Data warehouse เนื่องจากเก็บข้อมูลทั้งข้อมูลดิบ ที่มีโครงสร้าง และไม่มีโครงสร้าง อนุญาตให้เก็บข้อมูลโดยไม่มี schemas อนุญาตให้ทำการวิเคราะห์ข้อมูลในอดีต ปัจจุบัน และอนาคต เนื่องจากข้อมูลจะไม่ถูกลบ



6. From ETL to ELT

ETL	ELT
data is transformed and then transferred to Data Warehouse DB	Data remains in the DB of Data Warehouse
At early stages, easier to implement	To implement ELT process deep knowledge of tools and expert skills are needed.
Supports relational and structured data.	Supports structured, unstructured data sources.
Does not support Data Lake	Allows use of Data Lake
High costs for small and medium businesses.	Low entry costs using online Software as a Service Platforms.
Complexity increase with the additional amount of data in the dataset.	Power of the target platform can process significant amount of data quickly.
The process is used for over two decades.	Relatively new concept and complex to implement.

Comparison of ETL and ELT. Illustration by author

- **ETL** หรือ Extract-transform-load เป็นการเอาข้อมูลใหม่ๆไปลงในคลัง เหมาะกับการย้ายข้อมูลที่มีไม่มาก เพราะจะมีการ transform ข้อมูลก่อน ทำให้ผลลัพธ์ที่ออกมานั้นข้อมูลมีความซับซ้อนมากขึ้น

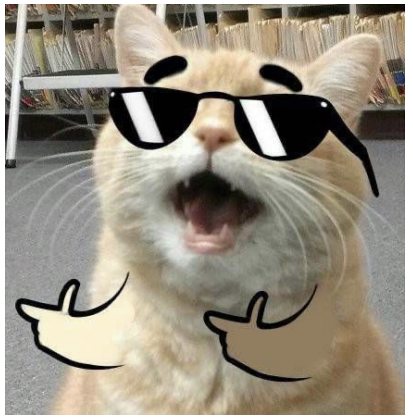
- **ELT** หรือ Extract (สกัดข้อมูล) load (ตั้งข้อมูล) transform (เปลี่ยนรูปแบบข้อมูล) คือการเก็บข้อมูลจากแหล่งแหล่ง หรือ หลายๆ แหล่งมาเก็บไว้ในคลังโดยในกระบวนการจะมีการ transform ข้อมูลตามที่กำหนดไว้

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

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



- "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

รู้ว่าสร้างเพื่ออะไร ต้องมาจากหลายแหล่ง แต่ละตารางต้องมีช่วงเวลาเดียวกัน

เป็น Data ที่ไม่เคลื่อนไหว เป็น Bath

→ เราต้องมีเป้าหมายว่าจะสร้าง DW มาเพื่ออะไร สร้างมาเพื่อตอบคำถาม สร้างเพื่อ support การตัดสินใจของผู้บริหาร

- Data warehousing:
  - The process of constructing and using data warehouses

From Tables and Spreadsheets to Data Cubes

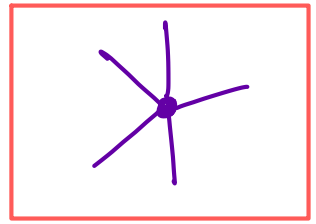
- A **data warehouse** is based on a multidimensional data model which views data in the form of a data cube
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - **Dimension tables**, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - **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**.

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

มีจุดศูนย์กลาง ทุกอย่างแตกออกจากศูนย์กลาง

Star Schema: An Example



time เวลาเก็บได้เป็น

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

branch

branch_key
branch_name
branch_type

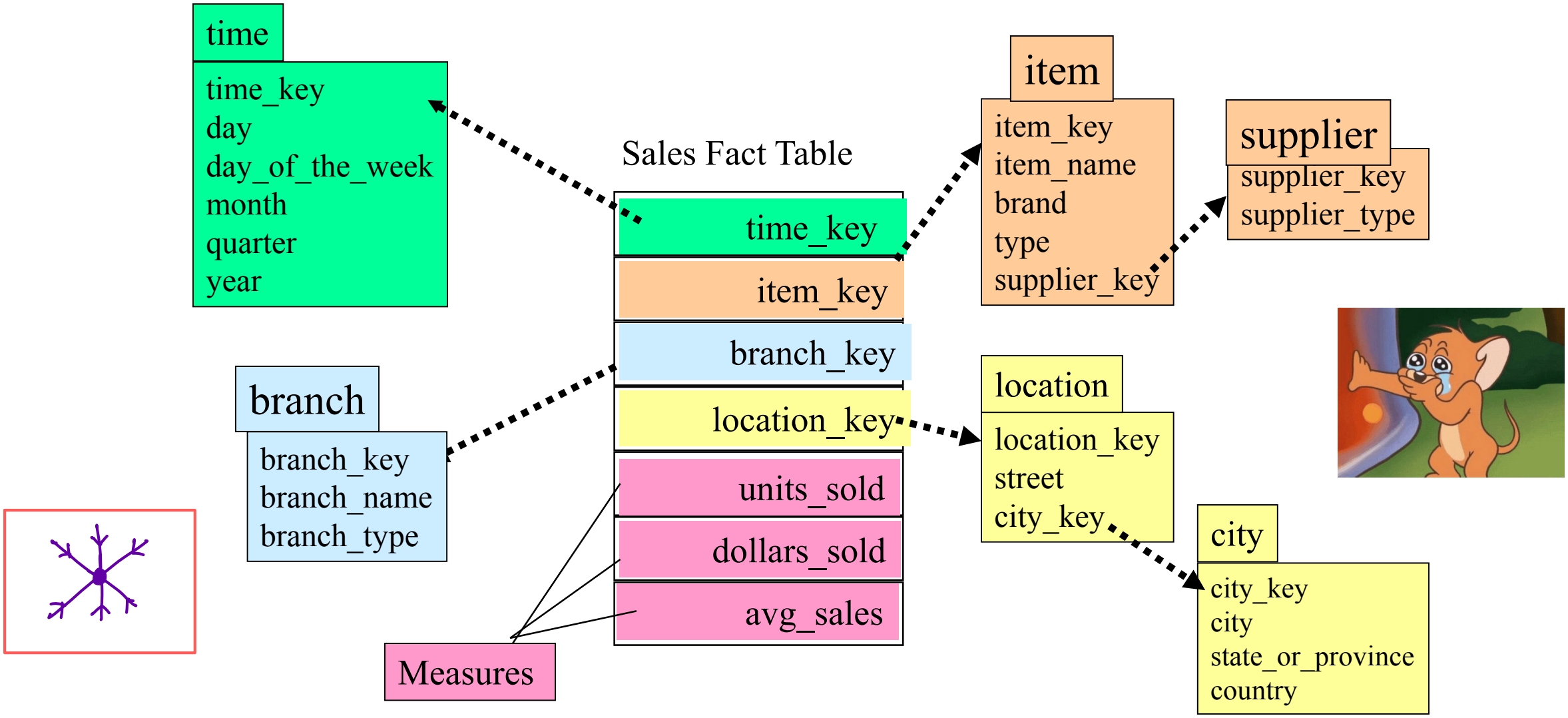
location

location_key
street
city
state_or_province
country

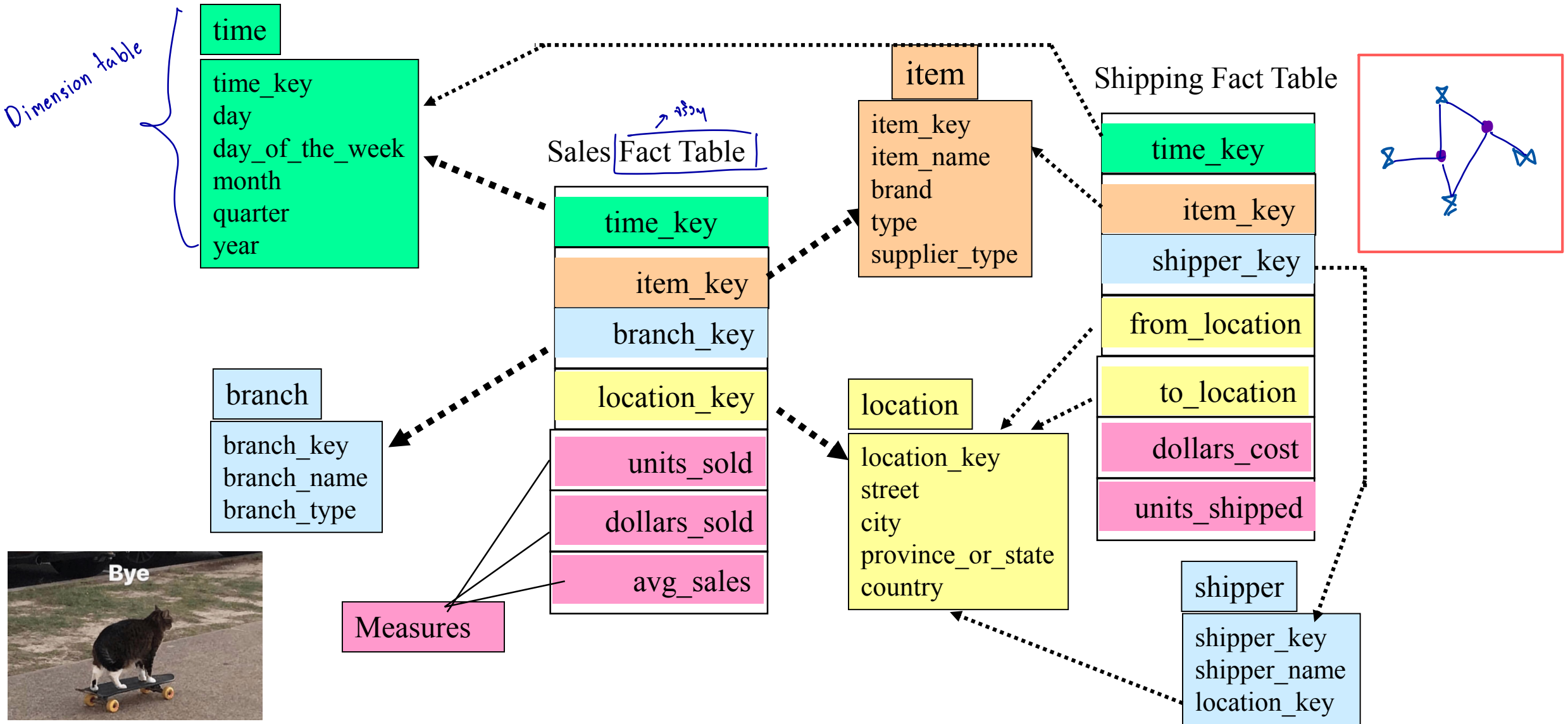
19

มีลักษณะคล้ายเกล็ดหิมะ !!

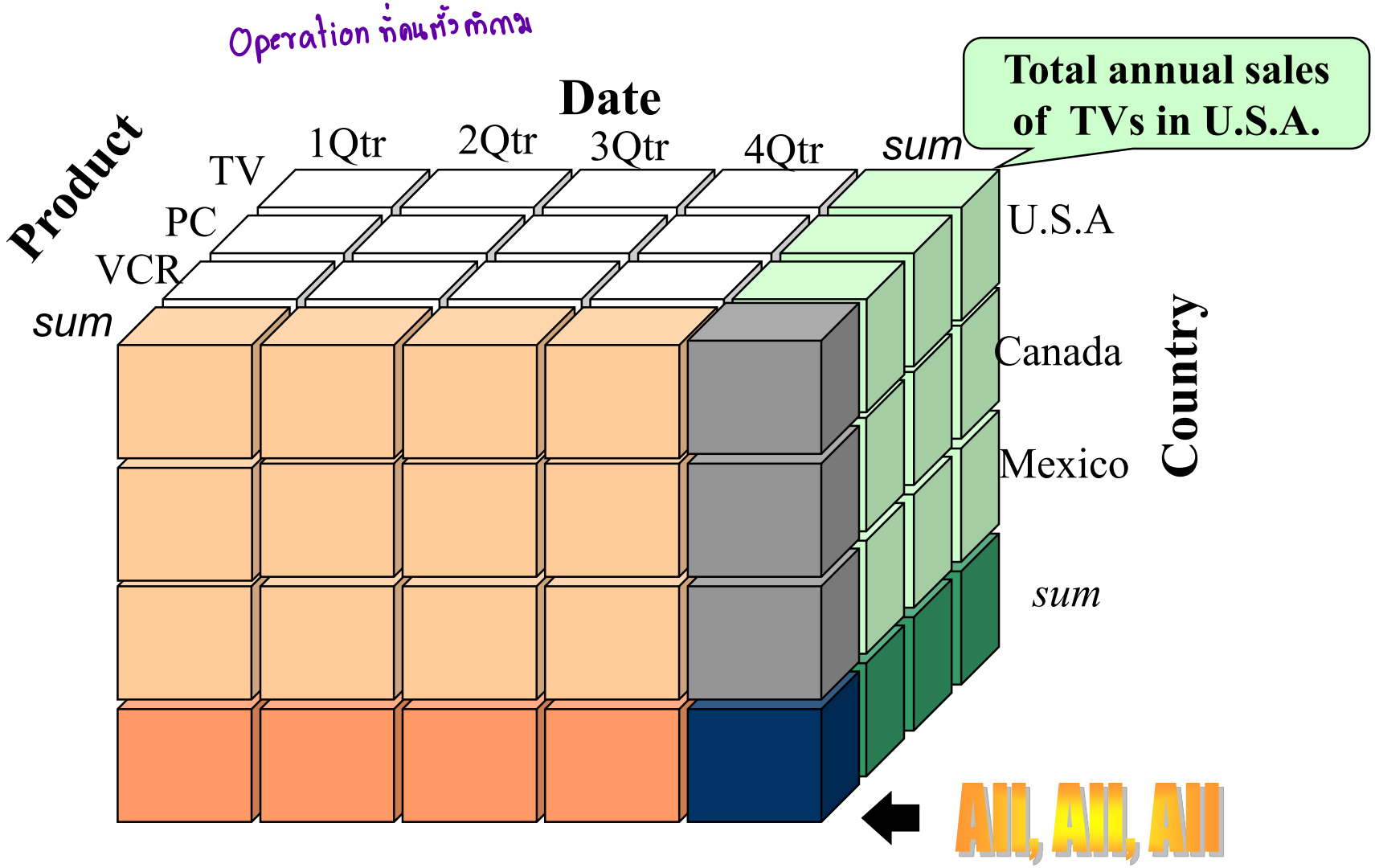
Snowflake Schema: An Example



Fact Constellation: An Example



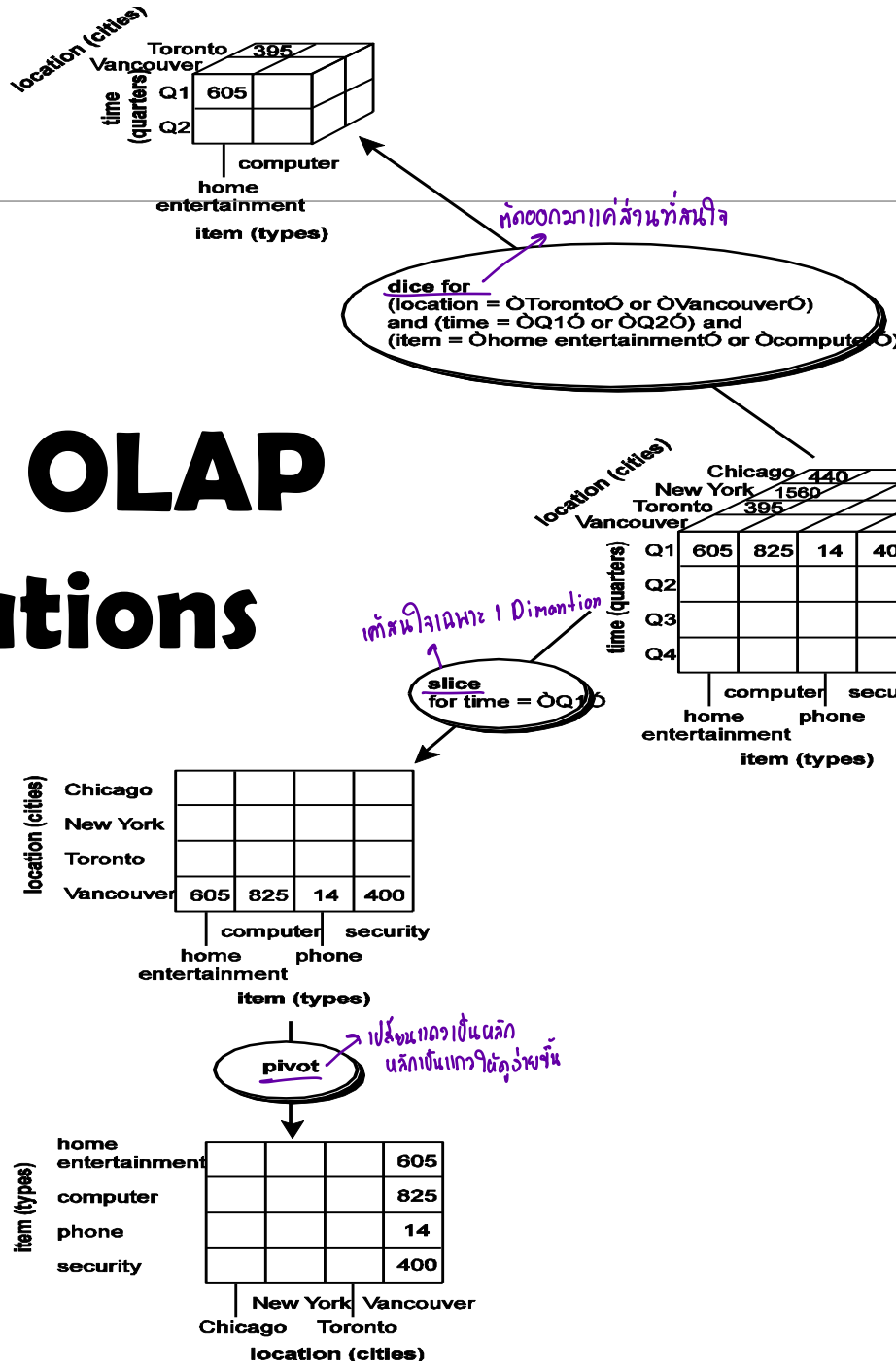
A Sample Data Cube



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)*

Typical OLAP Operations



the end...

29