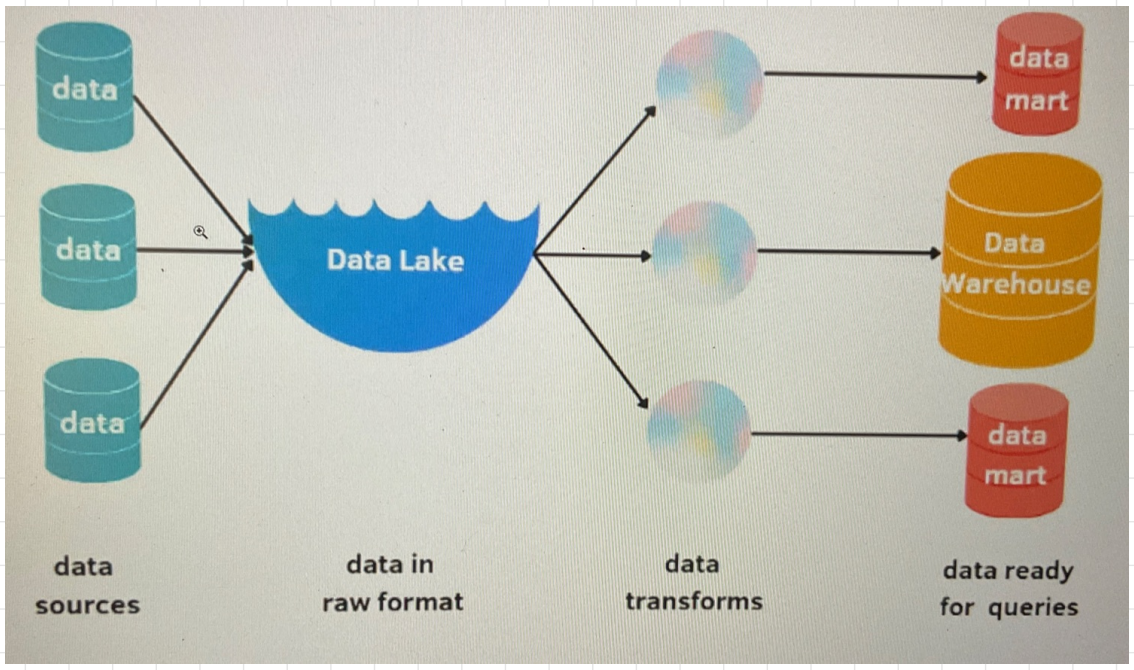
CS 412 Intro. to Data Mining

Chapter 4. Data Warehousing and On-line Analytical Processing

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

Why is Business Intelligence useful for a Data Scientist?



ที่ยังไม่สามารถปกามวลผลได้ตรงๆ เช่น เสียง จะเก็บไว้ใน Data Lake ก่อน

คำศัพท์ที่ควรรู้และควรทำความเข้าใจ

Table of Content:

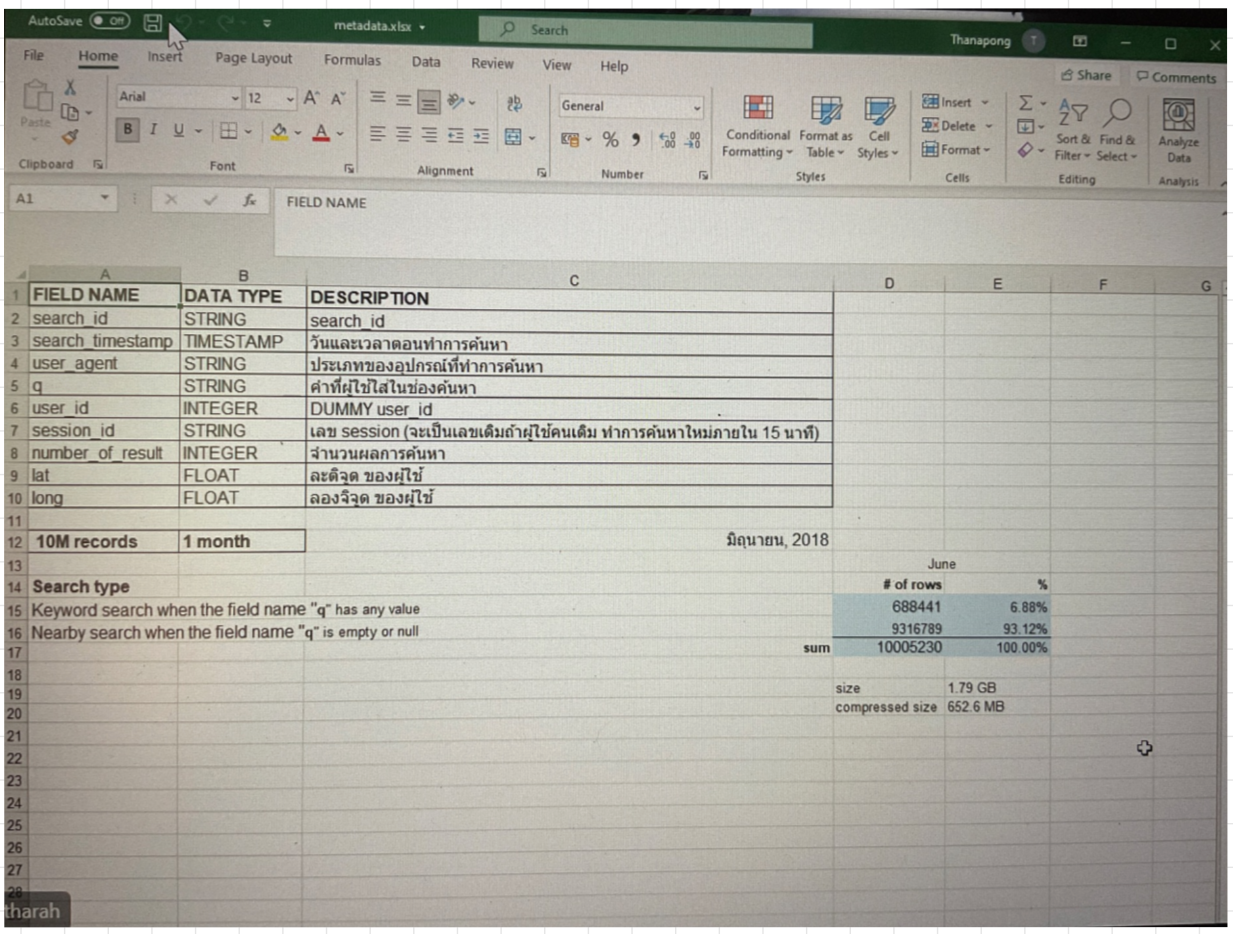
1. Data Warehouse
2. Data Mart
3. OLTP vs OLAP
4. ETL
5. Star vs Snowflake schemas
6. Data Lake
7. From ETL to ELT
8. Batch vs Stream Processing

Structured => Data ที่ถูกเก็บไว้ใน Warehouse

Unstructured => Data ที่คอมพิวเตอร์ไม่สามารถทำความเข้าใจกับข้อมูลได้ทันที เช่น รูป เสียง ข้อความเคมเมนต์

3. OLTP vs OLAP → เกิดตอนที่ทำ Data warehouse เก็บเพื่อนำไปใช้วิเคราะห์เป็นตารางที่เราจัดการมาแล้ว

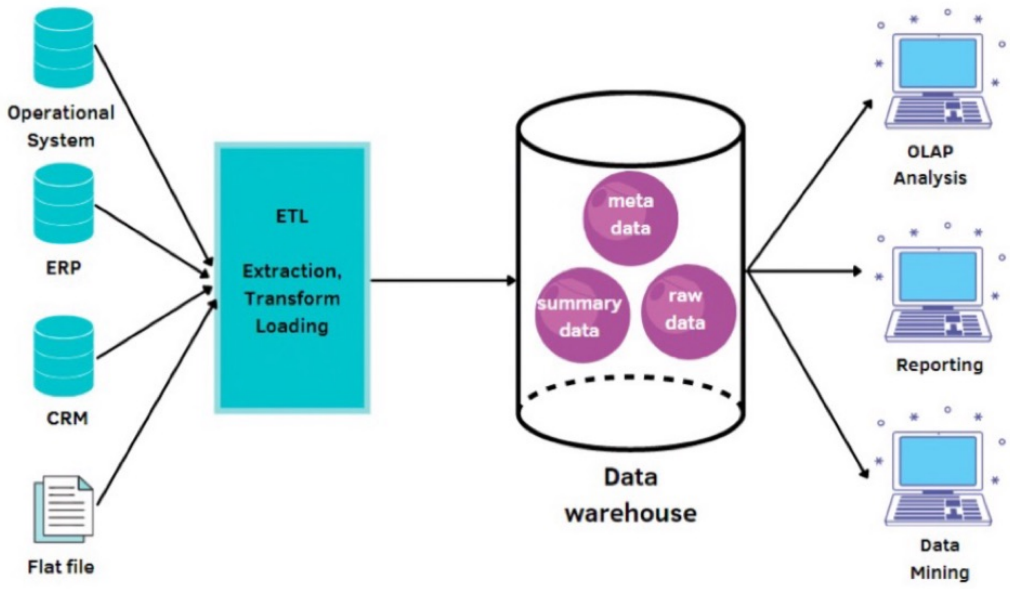
↳ หลักการในการจัดการข้อมูลเบื้องหลัง (เพิ่ม, ลบ, อัพเดท เหมือน Database) เป็นการเก็บรักษา Data



	FIELD NAME	DATA TYPE	DESCRIPTION
2	search_id	STRING	search_id
3	search_timestamp	TIMESTAMP	วันและเวลาตอนทำการค้นหา
4	user_agent	STRING	ประเภทของอุปกรณ์ที่ทำการค้นหา
5	q	STRING	ค่าที่ผู้ใช้ใส่ในช่องค้นหา
6	user_id	INTEGER	DUMMY user_id
7	session_id	STRING	เลข session (จะเป็นเลขเดิมถ้าผู้ใช้คนเดิม ทำการค้นหาใหม่ภายใน 15 นาที)
8	number_of_result	INTEGER	จำนวนผลการค้นหา
9	lat	FLOAT	ละติจุด ของผู้ใช้
10	long	FLOAT	ลองจิจุด ของผู้ใช้

12	**10M records**	**1 month**	มิถุนายน, 2018	
14	**Search type**		June **# of rows**	**%**
15	Keyword search when the field name "q" has any value		688441	6.88%
16	Nearby search when the field name "q" is empty or null		9316789	93.12%
17		**sum**	10005230	100.00%
19			size	1.79 GB
20			compressed size	652.6 MB

1. Data Warehouse



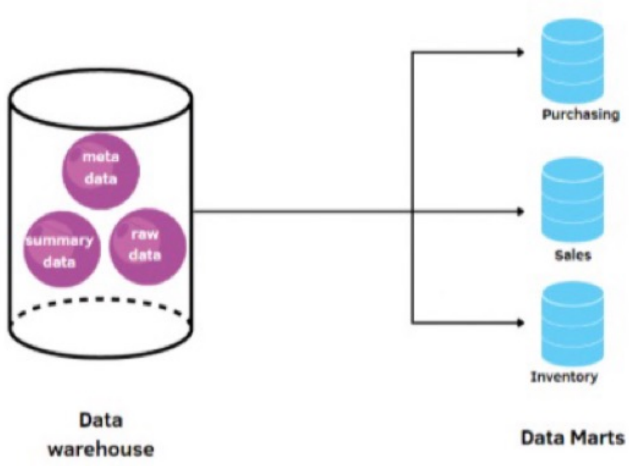
Data Warehouse architecture. Illustration by author.

The data werehouses store three types of data :

- ⭐ Meta data ⇒ ข้อมูลที่อธิบายเกี่ยวกับคลังข้อมูล และรายละเอียดเพื่อให้เข้าใจถึงข้อมูลที่จัดเก็บ
- ⭐ Summary data ⇒ ข้อมูลที่รวบรวม หรือสรุปที่สร้างโดยผู้ใช้คลังข้อมูล ช่วยให้ประสิทธิภาพการสืบค้นข้อมูล
- ⭐ Raw data ⇒ ข้อมูลดิบเป็นข้อมูลที่ไม่ได้ผ่านการวิเคราะห์หรือปรับเปลี่ยน

2. Data Marts

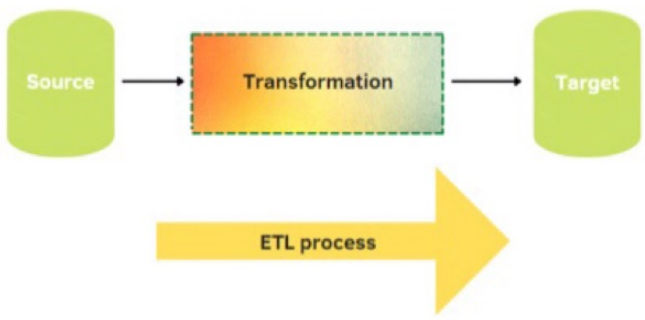
→ เป็นที่เก็บข้อมูลขนาดเล็กย่อยๆ ลงมา ส่วนใหญ่ใช้เก็บข้อมูลแยกเป็นแผนก เช่น Sales, Inventory



Data warehouse vs Data Marts. Illustration by author.

3. ETL

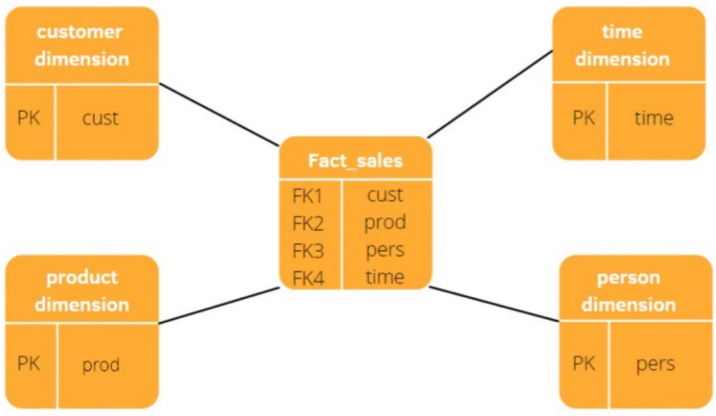
⇒ เป็นกระบวนการรวบรวม และประมวลผลที่ช่วยให้สามารถสร้างคลังข้อมูลได้



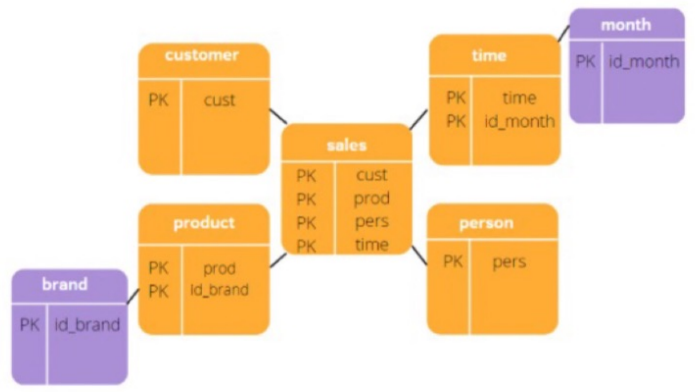
ETL process. Illustration by author

- ⭐ Extract : ขั้นตอนการเข้าถึงข้อมูลจากแหล่งต่างๆ เช่น ไฟล์ csv, JSON, XML
- ⭐ Transform : การแปลงข้อมูลให้อยู่ในรูปแบบที่เป็นประโยชน์สำหรับการวิเคราะห์ รวมถึงการล้าง ตรวจสอบ การรวบรวม และการลบรายการที่ซ้ำกันออก
- ⭐ Load : ข้อมูลจะถูกโอนย้ายไปยังปลายทางสุดท้าย นั่นคือ คลังข้อมูล เกิดขึ้นได้ทั้งแบบ ชั่วโมง วัน สัปดาห์ ขึ้นอยู่กับข้อมูลมากน้อยแค่ไหน การวิเคราะห์จะยิ่งแม่นยำมากขึ้นเท่านั้น

4. Star vs Snowflake Schemas



Star Schema. Illustration by author



Snowflake Schema. Illustration by author

Data warehouse ส่วนที่ทำข้อมูลโดยใช้ Schemas หมายถึง Schemas มีประโยชน์ในการจัดการข้อมูลจำนวนมาก เพื่อความปลอดภัยในการวิเคราะห์ มี Schemas ที่ใช้อย่างแพร่หลาย 2 แบบ คือ

⭐ Star Schema ⇒ มีโครงสร้างคล้ายดาวตรงกลางมี fact table โดยมี Dimension tables อยู่รอบๆ

⭐ Snowflake Schema ⇒ มีโครงสร้างเหมือนเกล็ดหิมะ ข้อดีคือ ช่วยเพิ่มขนาดและใช้พื้นที่จัดเก็บน้อยลง

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.

⭐ มีความซับซ้อนน้อยกว่า Data warehouse เนื่องจากเก็บข้อมูลทั้งข้อมูลดิบที่มีโครงสร้าง 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-tranform-load เป็นการเอาข้อมูลใหม่ๆไปลงในคลัง เหมาะกับการย้ายข้อมูลที่มีไม่มาก เพราะจะมีการ transform ข้อมูลก่อน ทำให้ผลลัพธ์ที่ออกมานั้นข้อมูลมีความซับซ้อนมากยิ่งขึ้น

⭐ ELT หรือ Extract เรกัดข้อมูล load (ดึงข้อมูล) transform (เปลี่ยนรูปแบบข้อมูล)

คือ การนำข้อมูลจากแหล่งหนึ่ง หรือ หลายๆ แหล่งมาเก็บไว้ในคลัง โดยไม่ต้องการจะมีการ 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

Model ของ Data Warehouse คือ

1. Data Cube คือ การมองแบบ Dimention มี 3 แบบ คือ Star Schema, Snowflake Schema, Fact Constellation

2. OLAP (on-line Analytical Processing)

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
- 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

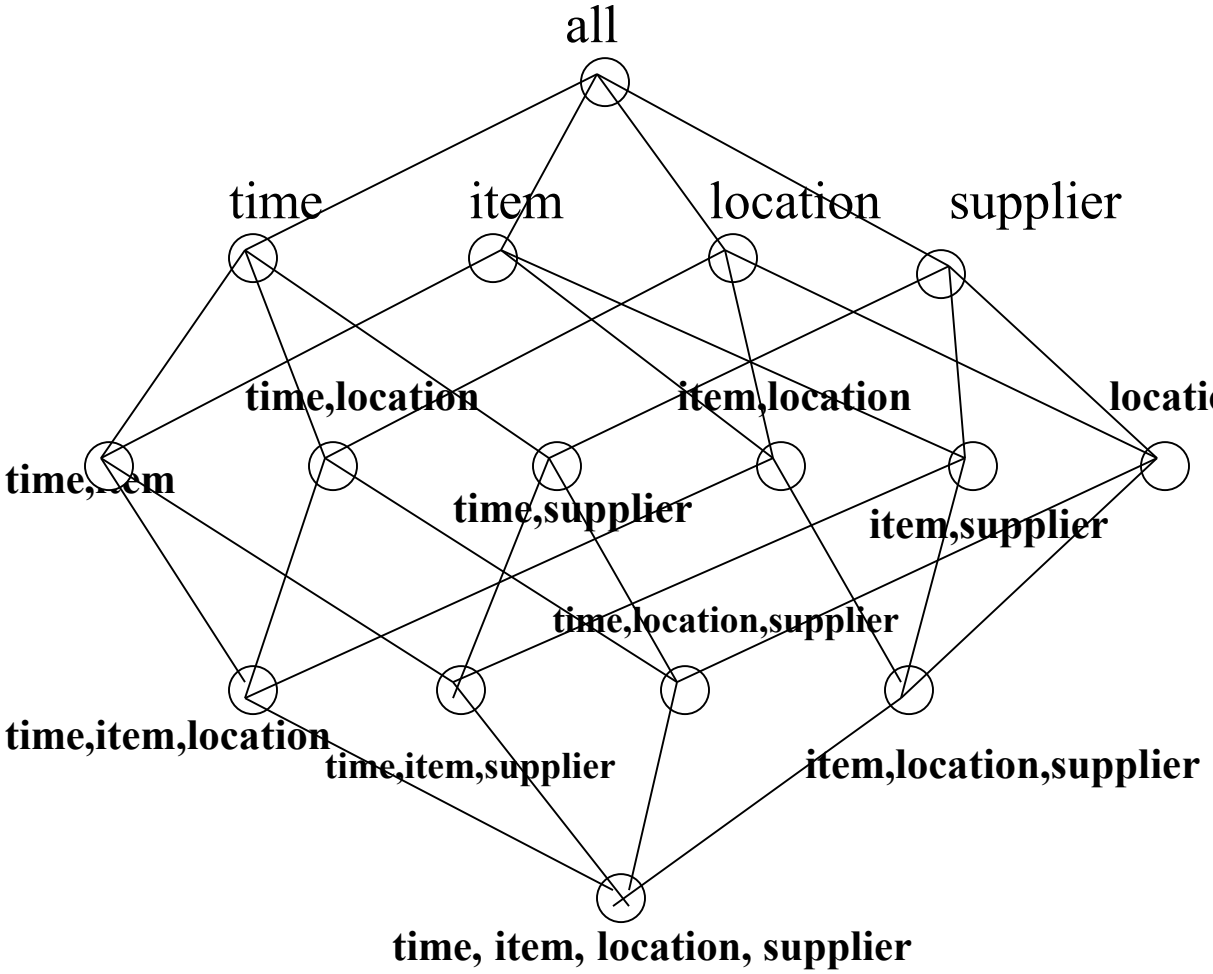
ตัวที่อธิบายว่าแต่ละ Dimension ของ ข้อมูล ที่แยกเป็นอะไรบ้าง

  - **Dimension tables**, such as item (item_name, brand, type), or time(day, week, month, quarter, year)

ตารางที่เก็บตัวเลขต่างๆ

  - **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



0-D (*apex*) cuboid

1-D cuboids

2-D cuboids

3-D cuboids

4-D (*base*) cuboid

17

2D

Table 4.2: A 2-D view of sales data for *AllElectronics* according to the dimensions *time* and *item*, where the sales are from branches located in the city of Vancouver. The measure displayed is *dollars_sold* (in thousands).

location = "*Vancouver*"

	item *(type)*			
time *(quarter)*	*home entertainment*	*computer*	*phone*	*security*
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q4	927	1038	38	580

⇒ 2 มิติ
↳ แสดงในแนวนอน
- มุมมองด้านเวลา time (quarter)
- มุมมองด้านชนิดของสินค้า item (type)

3D

Table 4.3: A 3-D view of sales data for *AllElectronics*, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).

	location = "Chicago"				***location*** = "New York"				***location*** = "Toronto"				***location*** = "Vancouver"			
	item				***item***				***item***				***item***			
time	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*	*home ent.*	*comp.*	*phone*	*sec.*
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	400
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	512
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	501
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580

⇒ 3 มิติ
↳ ตัวนี้มีมุมมองอีกมุมหนึ่งด้วย
แต่ละตารางตัวเลขเป็นอย่างไร

3D

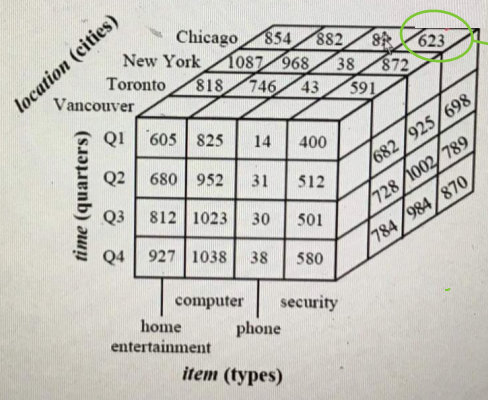


Figure 4.3: A 3-D data cube representation of the data in Table 4.3, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).

⇒ 3 มิติ

Q_1 ของ location
Chicago = 123

4D

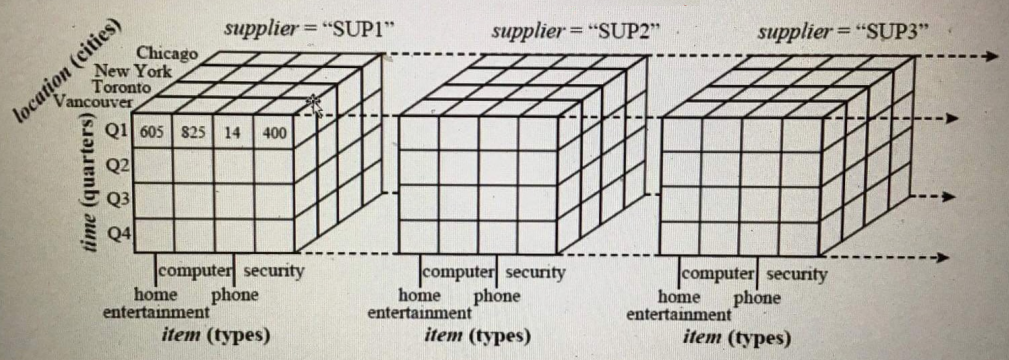


Figure 4.4: A 4-D data cube representation of sales data, according to the dimensions *time*, *item*, *location*, and *supplier*. The measure displayed is *dollars_sold* (in thousands). For improved readability, only some of the cube values are shown.

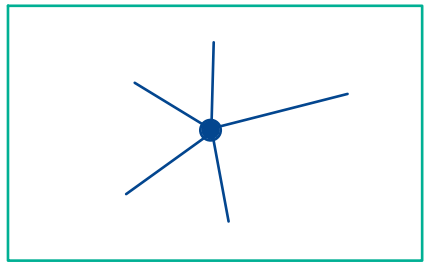
⇒ 4 มิติ
↳ จะมีมุมมองของ Supplier เพิ่มมาอีกมุมมอง

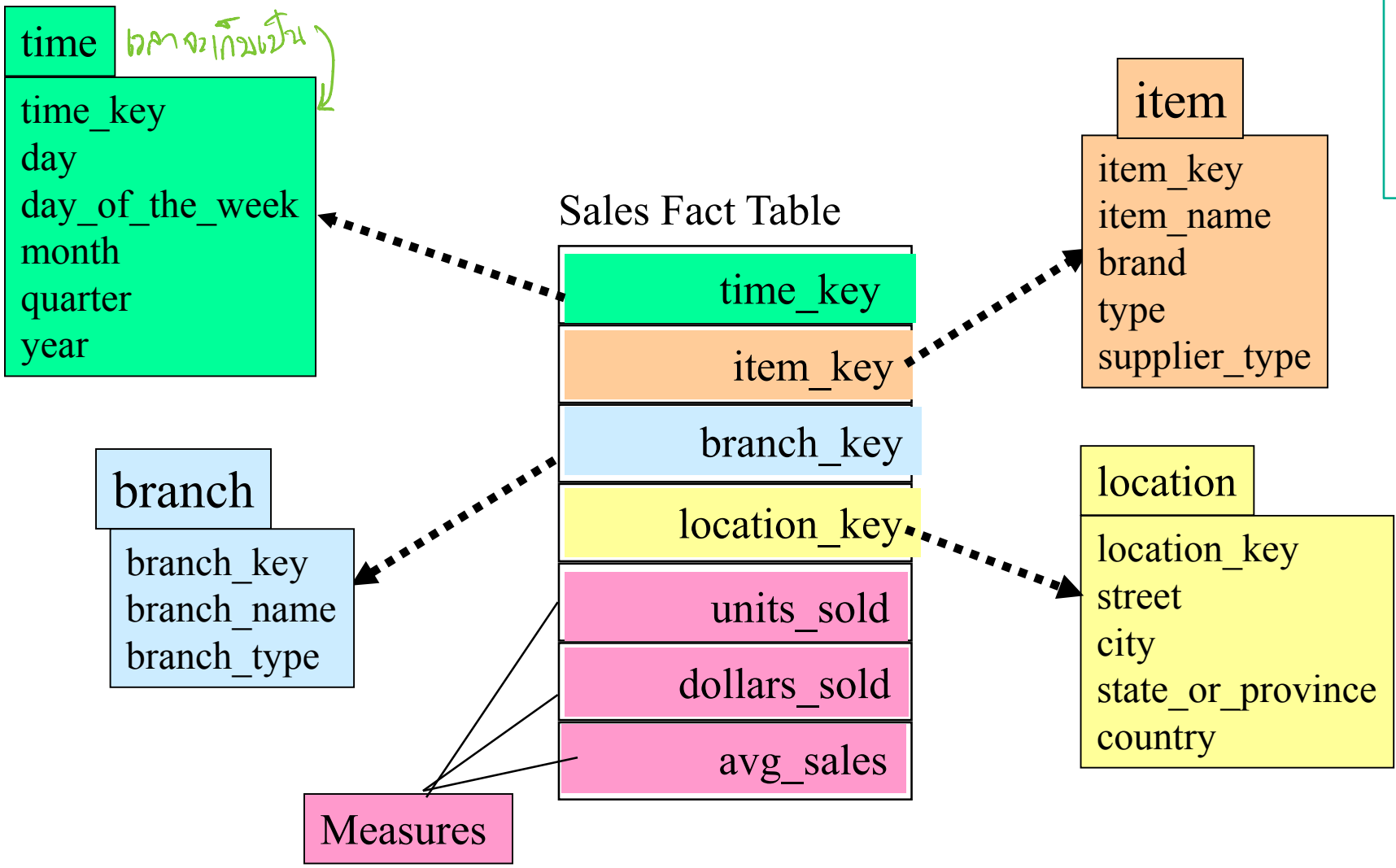
Conceptual Modeling of Data Warehouses

- 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

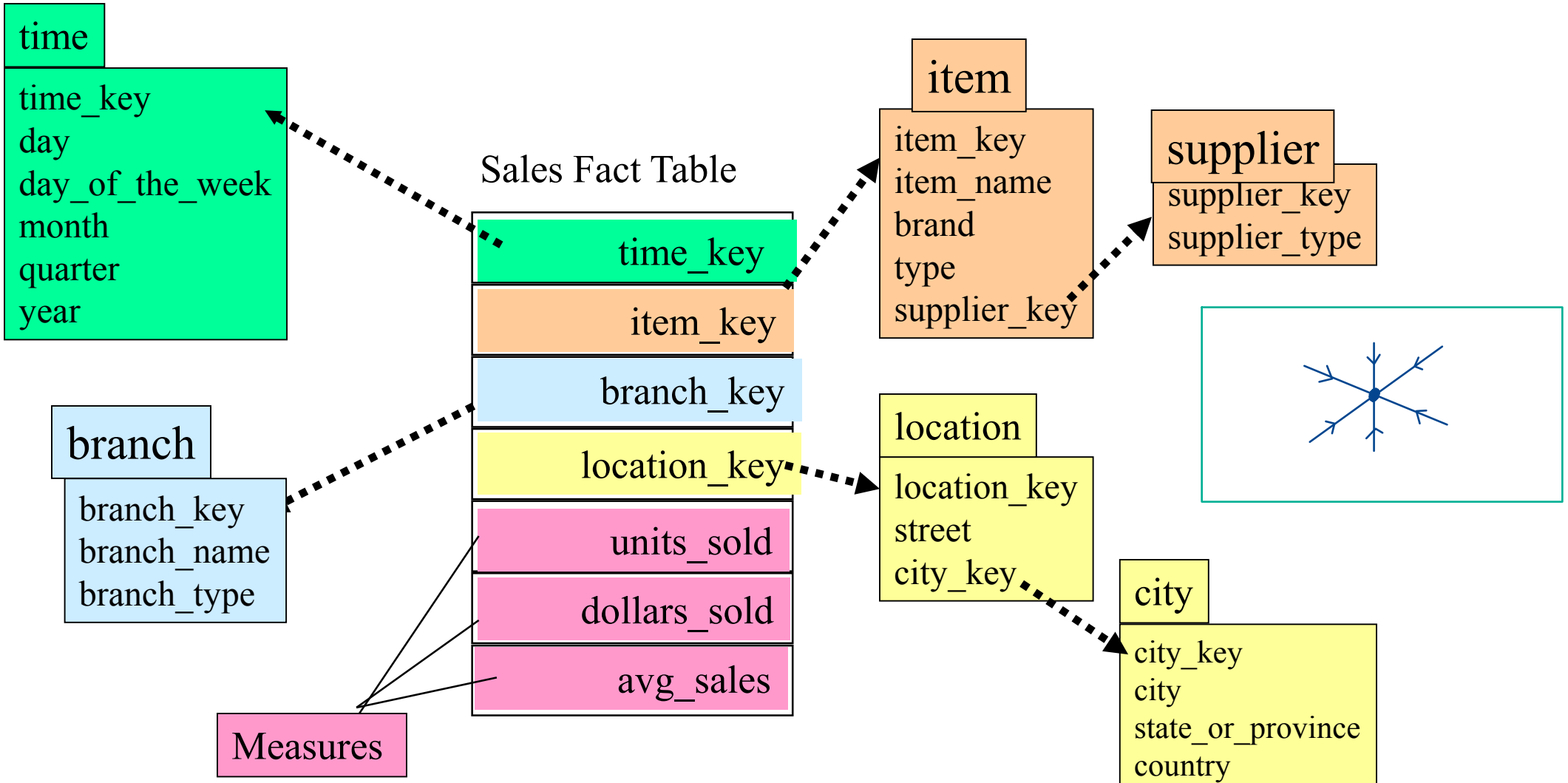




19

มีลักษณะคล้ายเกล็ดหิมะ

Snowflake Schema: An Example



Fact Constellation: An Example

Dimension Table

time
- time_key
- day
- day_of_the_week
- month
- quarter
- year

item
- item_key
- item_name
- brand
- type
- supplier_type

Sales Fact Table
- time_key
- item_key
- branch_key
- location_key
- units_sold
- dollars_sold
- avg_sales

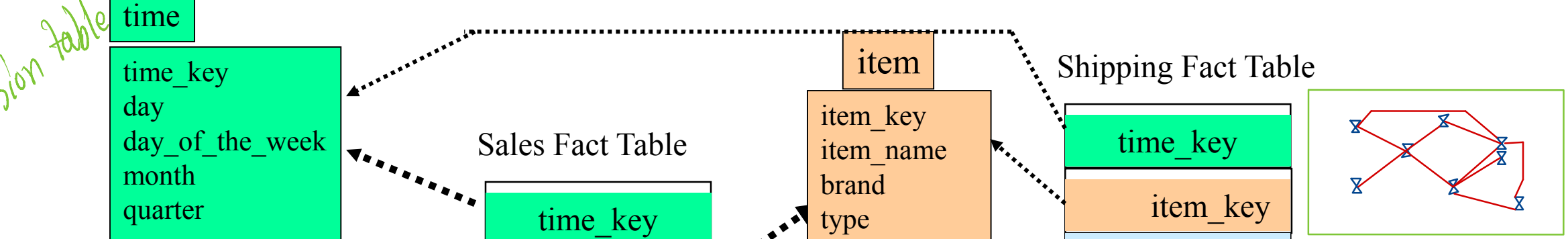
Measures

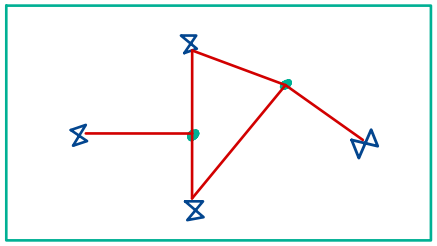
branch
- branch_key
- branch_name
- branch_type

location
- location_key
- street
- city
- province_or_state
- country

Shipping Fact Table
- time_key
- item_key
- shipper_key
- from_location
- to_location
- dollars_cost
- units_shipped

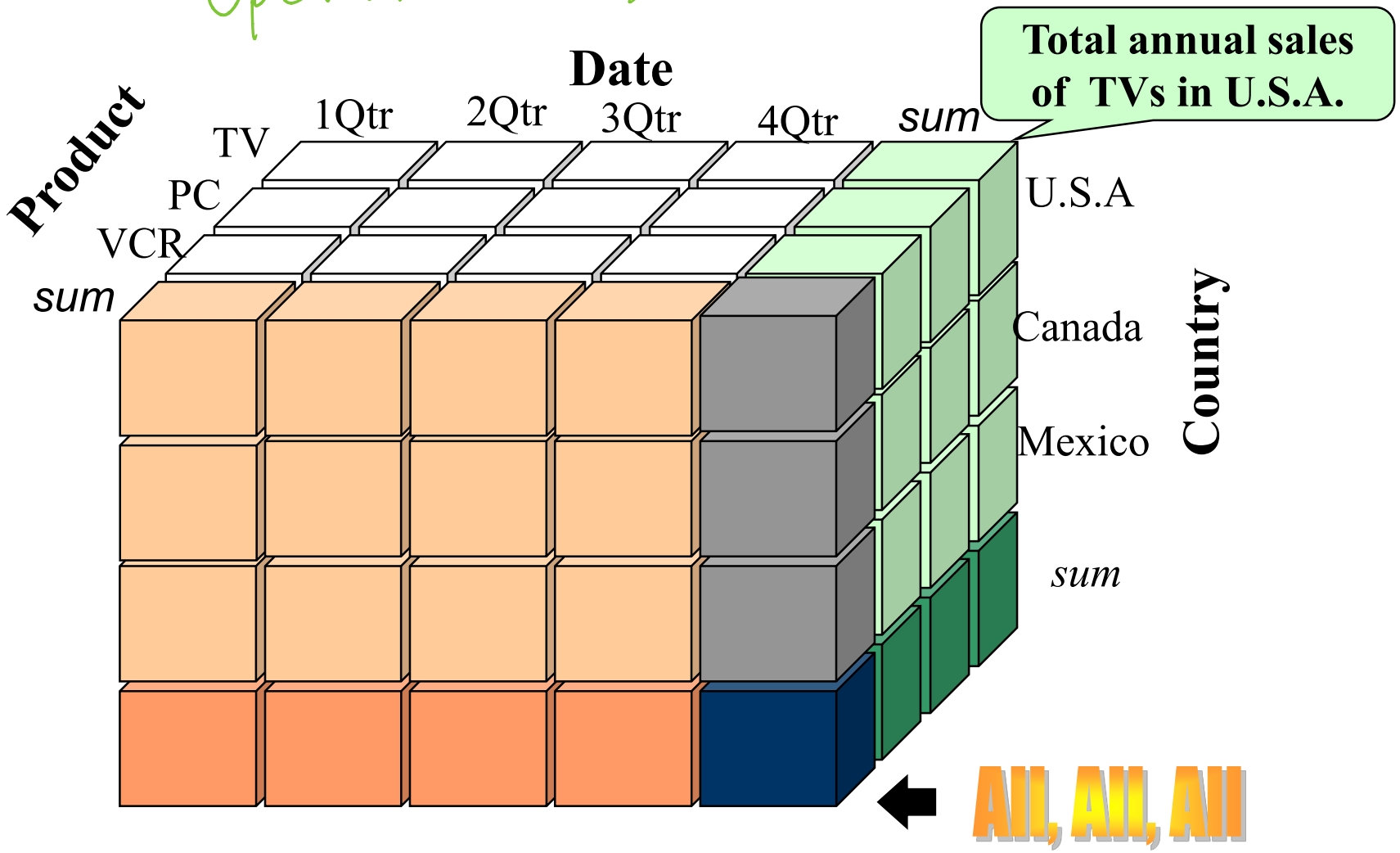
shipper
- shipper_key
- shipper_name
- location_key
- shipper_type





A Sample Data Cube

operation ที่ใช้ในการคำนวณค่าต่างๆ



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