

ทำขึ้นมาเพื่ออะไร?

## 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

สิ่งก่อนที่เราจะเข้ามาเกี่ยวข้องกับ

## Data Warehouse—Subject-Oriented

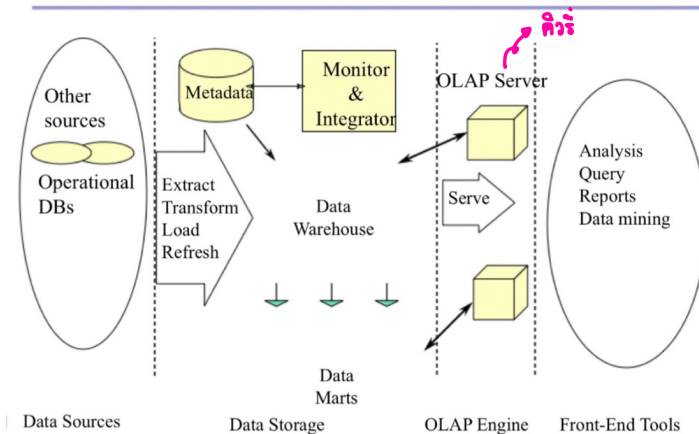
---

- Organized around major subjects, such as **customer**, **product**, **sales**
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
- Provide **a simple and concise** view around particular subject issues by **excluding data that are not useful in the decision support process**

## Data Warehouse—Nonvolatile

- A **physically separate store** of data transformed from the operational environment
- Operational **update of data does not occur** in the data warehouse environment
- Does not require transaction processing, recovery, and concurrency control mechanisms
- Requires only two operations in data accessing:
  - *initial loading of data* and *access of data*

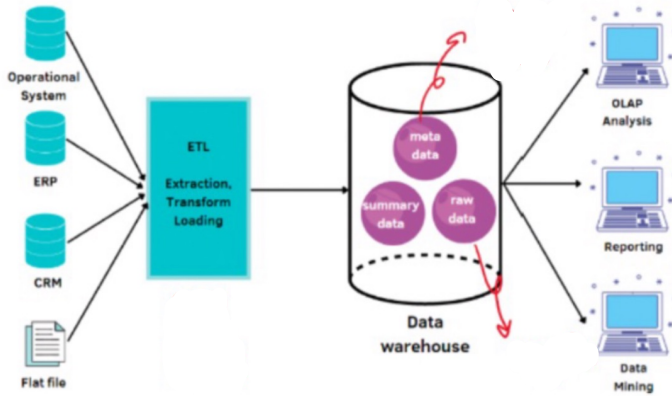
### Data Warehouse: A Multi-Tiered Architecture



## 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) สามารถอธิบายได้ว่า จำแนกเป็นช่วงเวลา
- **Fact table** contains **measures** (such as **dollars\_sold**) and keys to each of the related dimension tables → เก็บตัวเลข
- 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 warehouse



Data Lake → เก็บทุกอย่าง

Data warehouse → จัดระเบียบ

### ① ประเภทของ Data warehouse

structured → จัดระเบียบ

unstructured → ไม่สามารถทำได้ ex. รูป, วิดีโอ, ไฟล์เสียง

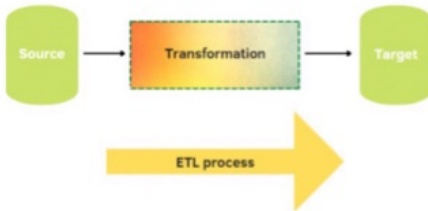
Data marts → จัดระเบียบ (ข้อมูลเฉพาะ)

### ② OLTP → ใช้ในการจัดการข้อมูลเบื้องต้น ex. update ข้อมูล

แก้ไขข้อมูล, จัดข้อมูล

เป็นการจัดการข้อมูลถึงข้อมูลเพื่อไปวิเคราะห์ ex. query, filter

### ③ ETL → รวม Data



Data Lake → ELT → AI von

E - Extration

T- Transform

L- Loading → update data → run

④ star vs snowflake schema

⑤ Data Lake → เก็บทั้ง structured อื่นๆ

⑥ การประมวลผล มี 2 แบบ

1. Batch Processing → รวบรวมเป็น 1 ชุดส่งเข้าประมวลผล
2. streaming Processing → ประมวลผลทีละชิ้นๆ เข้ามาในเฟรมเวลา