Data Warehousing

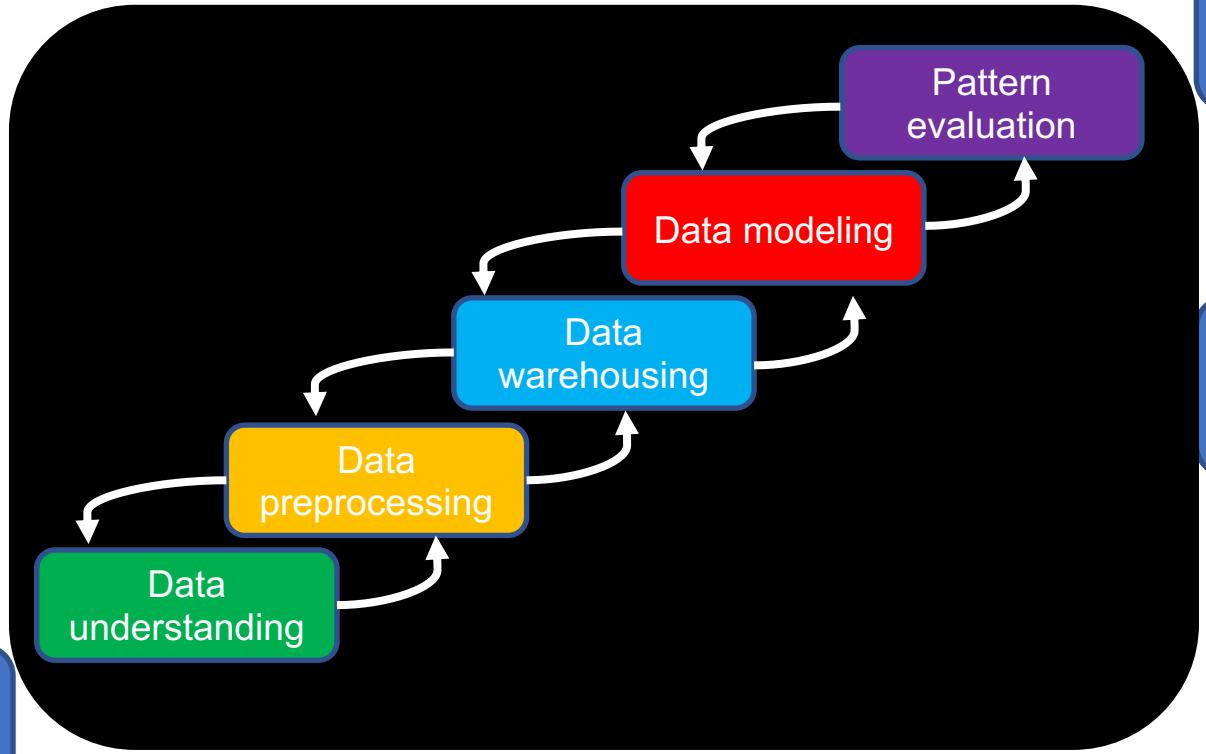
Data Mining:
Data Mining Pipeline

with **Dr. Qin Lv**

Master of Science in Data Science
UNIVERSITY OF COLORADO **BOULDER**

Learning objective: Identify key characteristics of data warehousing. Apply data warehousing techniques for data mining tasks.

Data Mining Pipeline



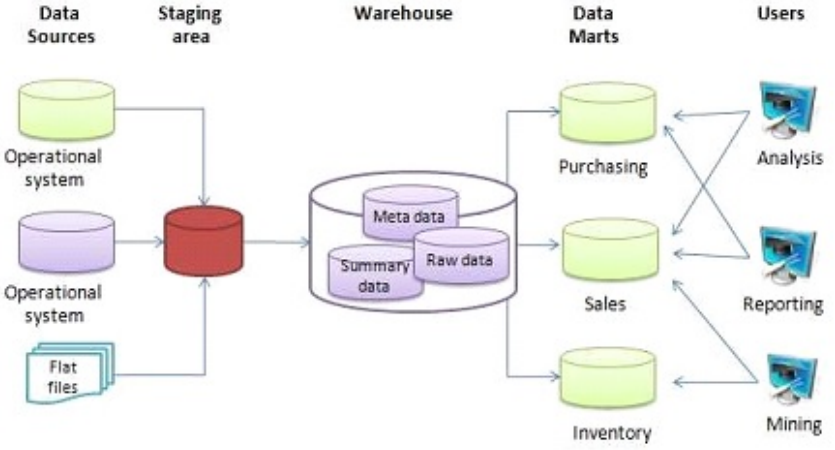
Application

Knowledge

Technique

Data

Data Warehousing



- ➢ Data warehouse
  - Vs. operational data
- ➢ Data cube and OLAP
  - Multi-dimensional data management
- ➢ Data warehouse architecture

Why Data Warehousing?

- ➢ Data warehouse
  - William H. Inmon -- “a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”
- ➢ Separated from operational data
- ➢ Focused on data-driven decision support

Key Characteristics (1)

➢ Subject-oriented

- E.g., stores, customers, products, students, courses
- Focuses on specific subjects, ignore unrelated ones

➢ Integrated

- E.g., new store location: customers, businesses, traffic
- Heterogeneous sources, data cleaning and integration

Key Characteristics (2)

➢ Time-variant

- E.g., traffic information in the past 5-10 years
- Historical data, longer time span, timestamped data

➢ Nonvolatile

- Typical data operations: initial loading, append, read
- Separate from operational data, not updated in place

OLTP vs. OLAP (1)

➢ Online Transactional Processing (OLTP)

- Transaction-oriented tasks: bank transfer, purchase, …
- Daily operations: insert, update, delete

➢ Online Analytical Processing (OLAP)

- Complex queries on historical data
- Data analysis for insights and decision making

OLTP vs. OLAP (2)

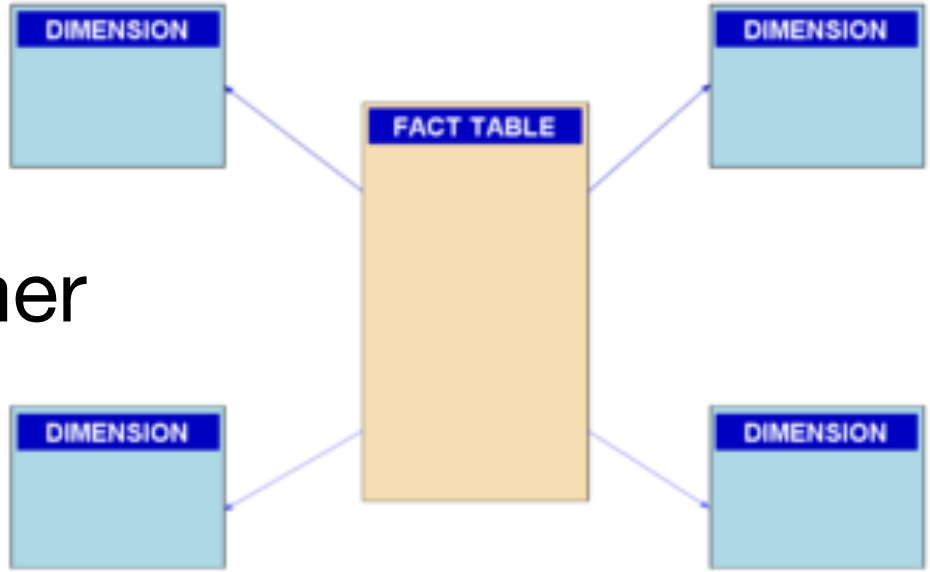
	OLTP System Online Transaction Processing (Operational System)	OLAP System Online Analytical Processing (Data Warehouse)
Source of data	Operational data; OLTP's are the original source of the data	Consolidation data; OLAP data comes from the various OLAP Databases
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support
What the data Reveals	A snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and Updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Relatively standardized and simple queries Returning relatively few records	Often complex queries involving aggregations
Processing Speed	Typically very fast	Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes
Space Requirements	Can be relatively small if historical data is achieved	Larger due to the existence of aggregation structures and history data; requires more indexes than OLTP
Database Design	Highly normalized with many tables	Typically de-normalized with fewer tables; use of star and/or snowflake schemas
Backup and Recovery	Backup religiously: operational data is critical to run the business, data loss is likely to entail significant monetary loss and legal liability	Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method

Data Warehouse Model

- Fact (e.g., sales) vs. dimension (e.g., item)
- Star schema: one fact table, multiple dimension tables
- Snowflake schema
  - one fact table, multiple levels of dimension tables
- Fact constellation schema
  - multiple fact tables, shared dimension tables

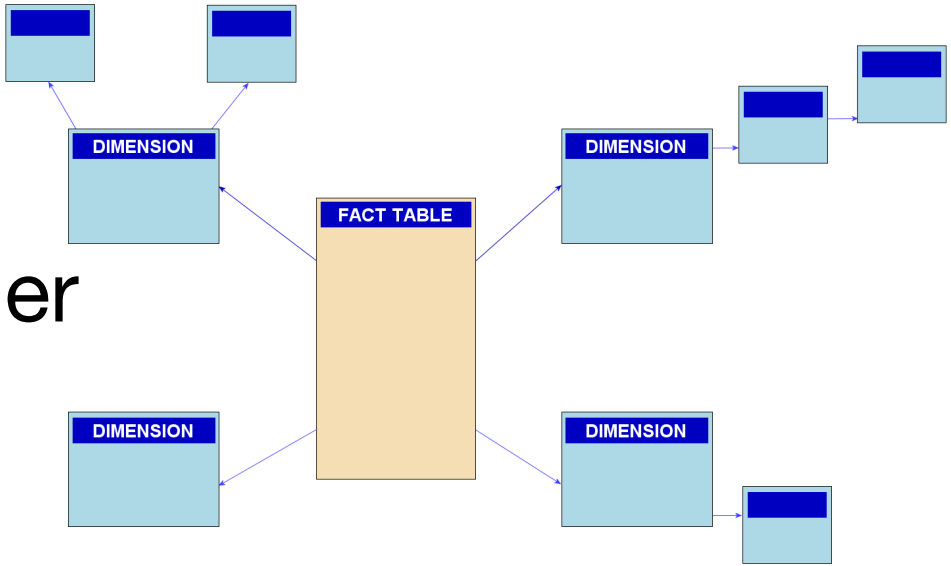
Star Schema

- ➢ Fact: Sales
  - Customer, item, time
- ➢ Dimension: Customer
  - Name, address, DOB
- ➢ Dimension: Time
  - Year, month, date



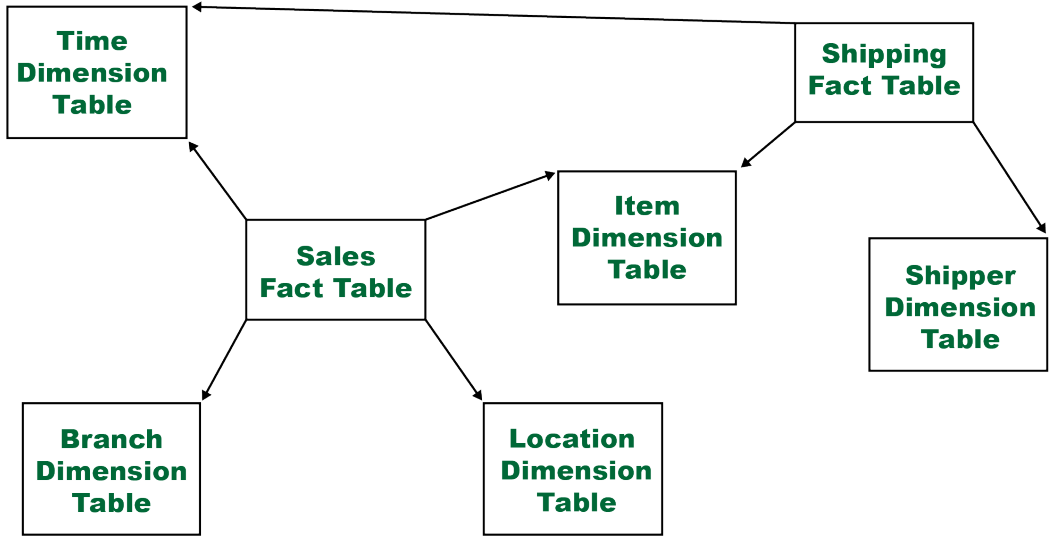
Snowflake Schema

- ➢ Fact: Sales
  - Customer, item, time
- ➢ Dimension: Customer
  - Name, address, card
- ➢ Dimension: Card
  - Number, exp date, CVC



Fact Constellation Schema

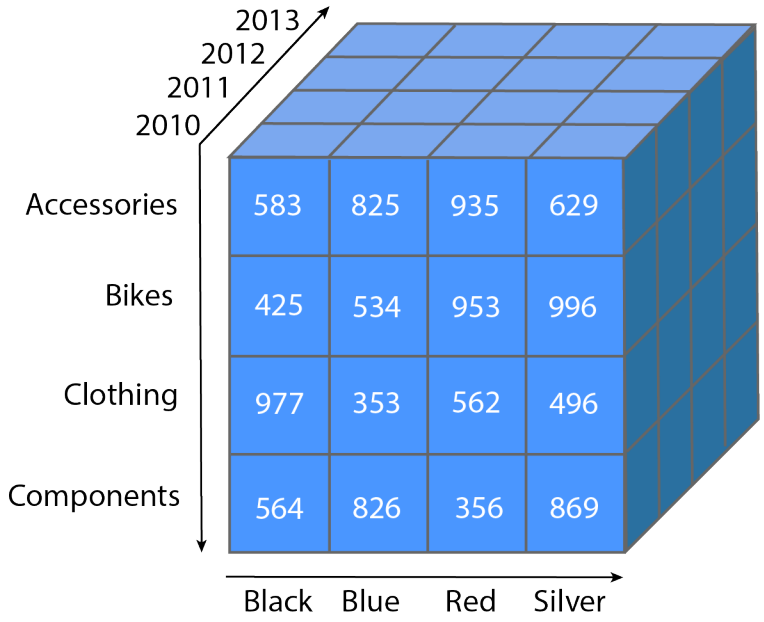
- Fact: Sales
- Dimension: Item
- Fact: Shipping



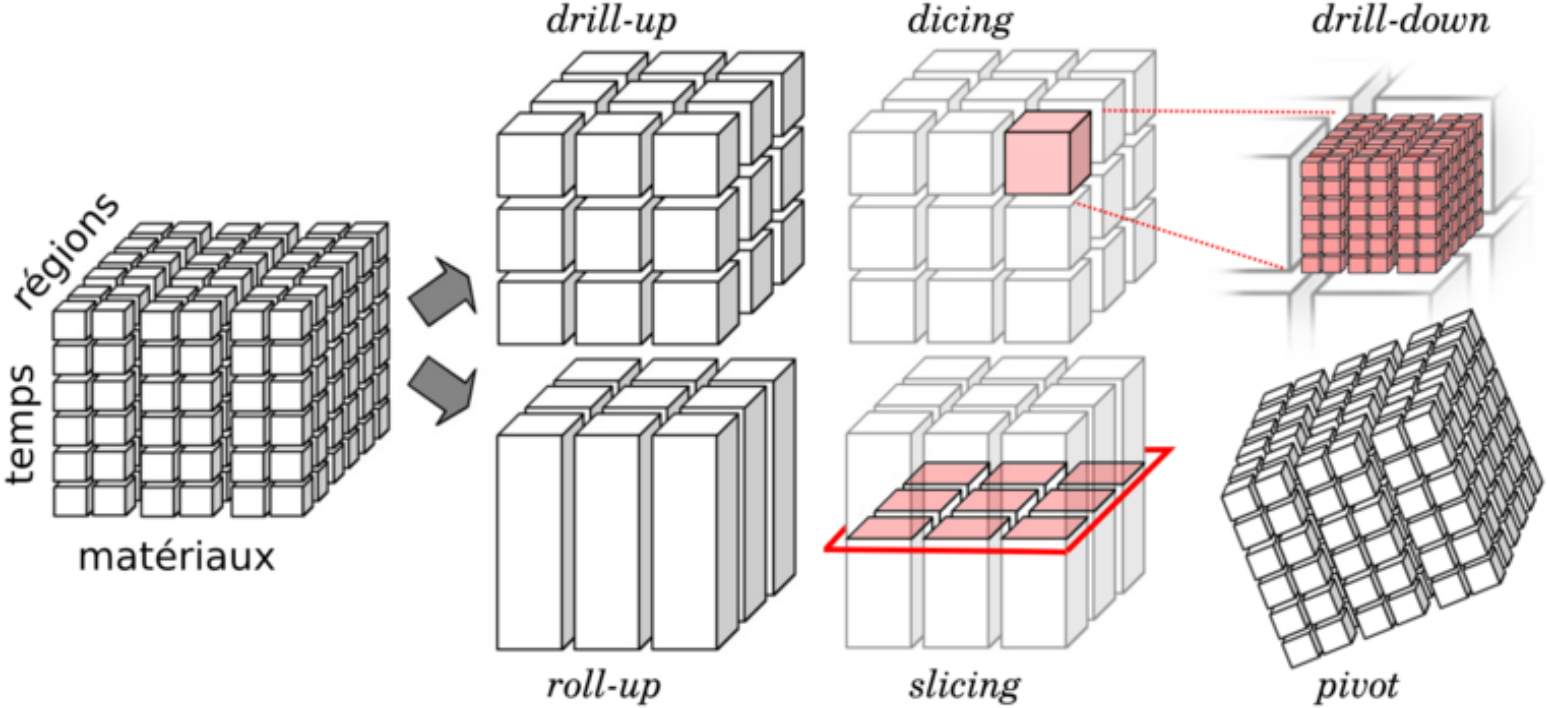
Data Cube

➢ Multi-dimensional data model

- Dimensions: cube attribute
- E.g., year, product, color
- Facts: numeric measure
- E.g., sales volume/value



Data Cube Operations (1)



Data Cube Operations (2)

- ➢ Roll up: aggregation
  - E.g., daily => monthly
- ➢ Drill down: reverse of roll up
  - E.g., North America => USA, Mexico, Canada, …
- ➢ Pivot: rotate (visualization)
  - E.g., <country, item> => <item, country>

Data Cube Operations (3)

- ➢ Slicing: select along a single dimension
  - E.g., country = “USA”

- ➢ Dicing: select along multiple dimensions
  - E.g., county = “USA”, year = “2011 – 2020”