Data Warehousing

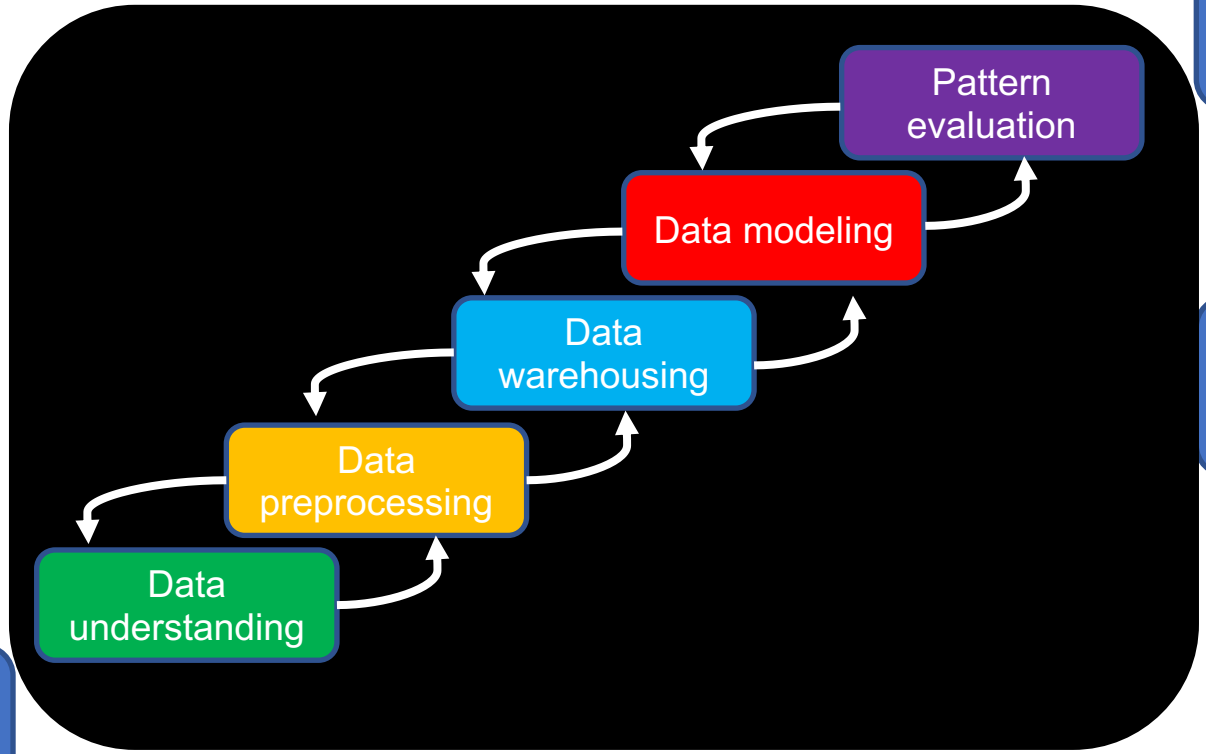
Data Mining:
Data Mining Pipeline

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



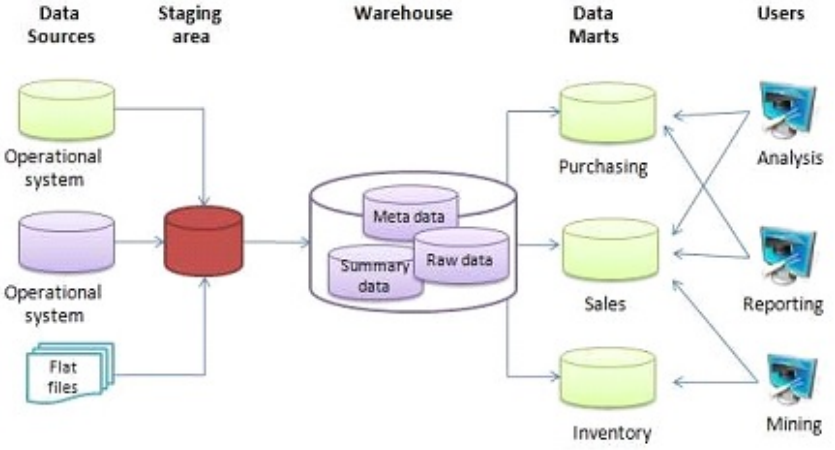
Application

Knowledge

Technique

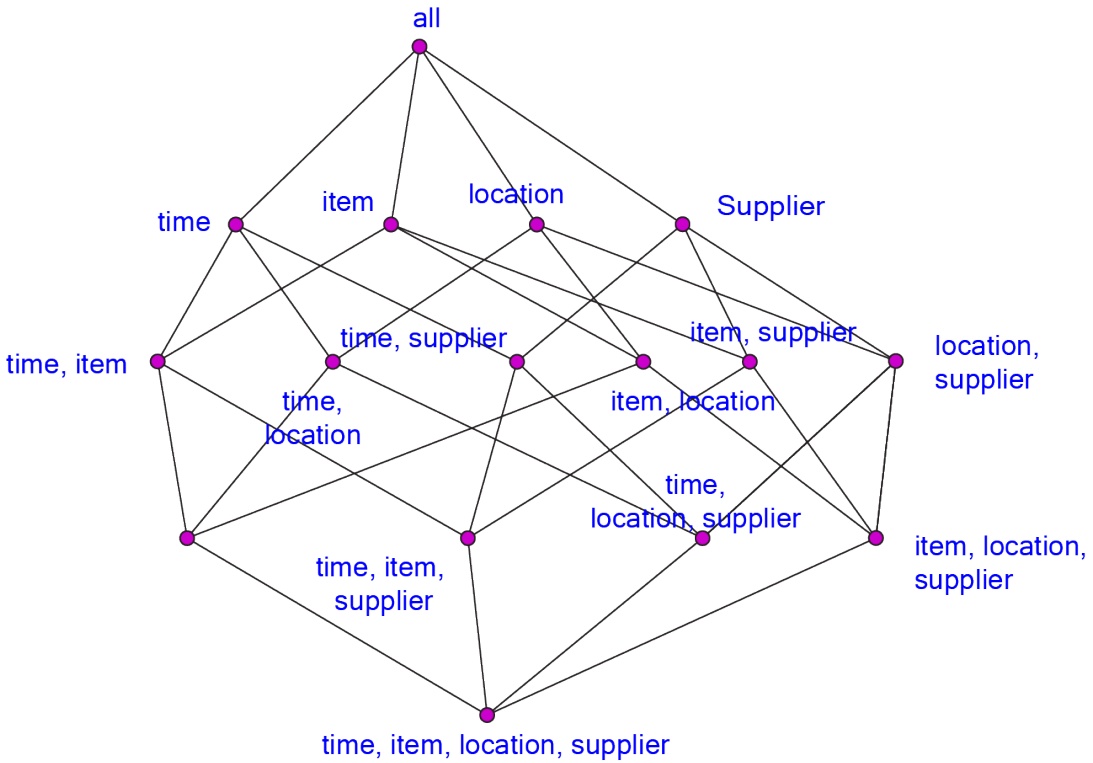
Data

Data Warehousing



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

Data Cube: Lattice of Cuboids



0-D (Apex) Cuboid

1-D Cuboids

2-D Cuboids

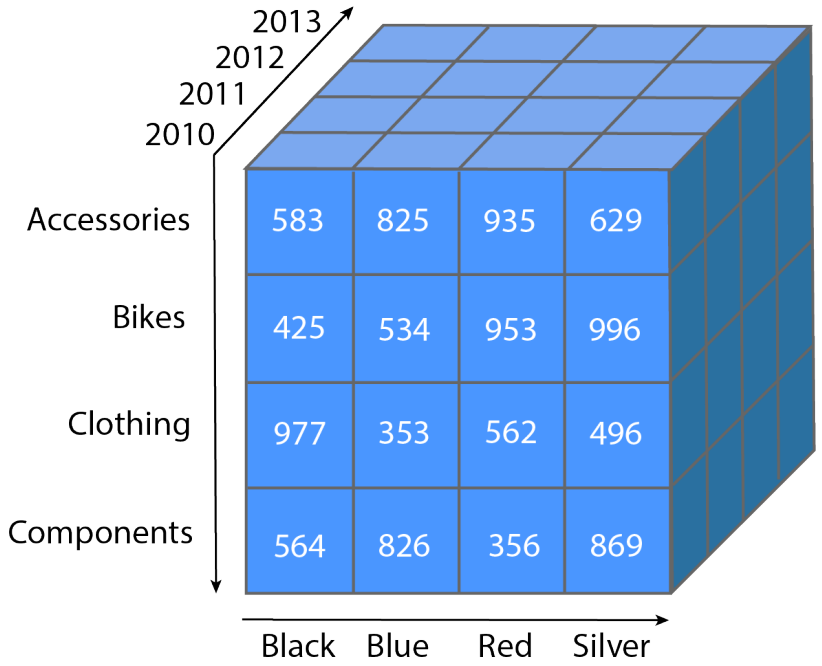
3-D Cuboids

4-D (Base) Cuboid

Cuboid Cells

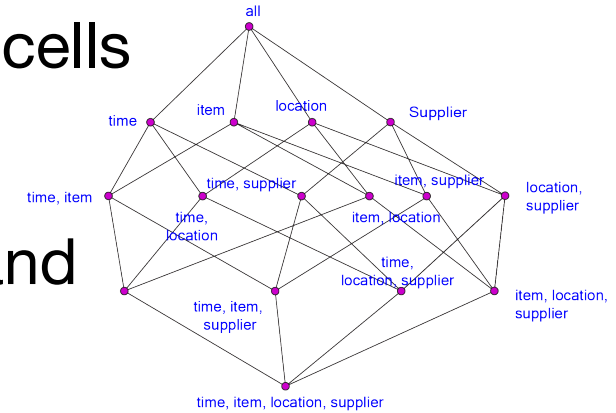
- Base cuboid
- Aggregation (roll-up)

- E.g., (year, item, color)
- a (2010, bikes, red) = 953
- b (2010, bikes, *)
- = 425 + 534 + 953 + 996 = 2908



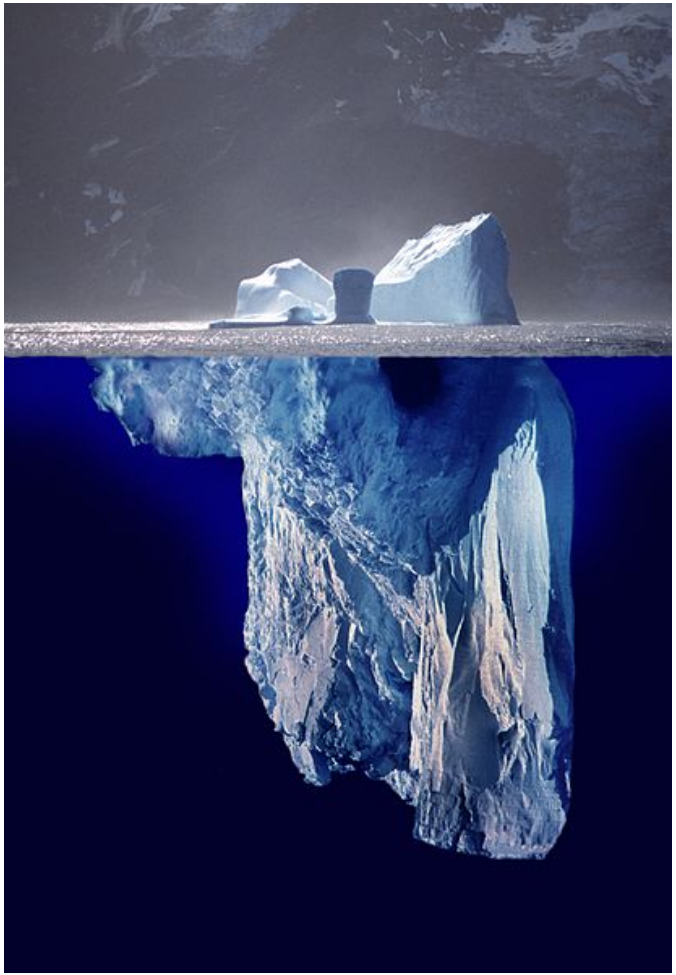
Data Cube Materialization

- ➢ Full materialization
  - Pre-compute all cuboids and cells
- ➢ No materialization
  - No precomputation, on-demand
- ➢ Partial materialization
  - (heuristically) pre-compute some cuboids and cells



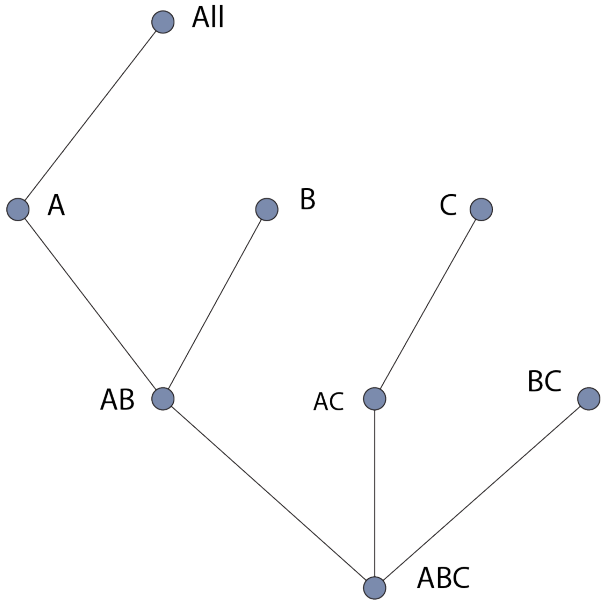
“Iceberg” Cube

- ➢ Only a small tip may be “above water”
- ➢ Minimum support
  - Only compute cuboid cells that are above a certain threshold
  - E.g., > $10K in sales



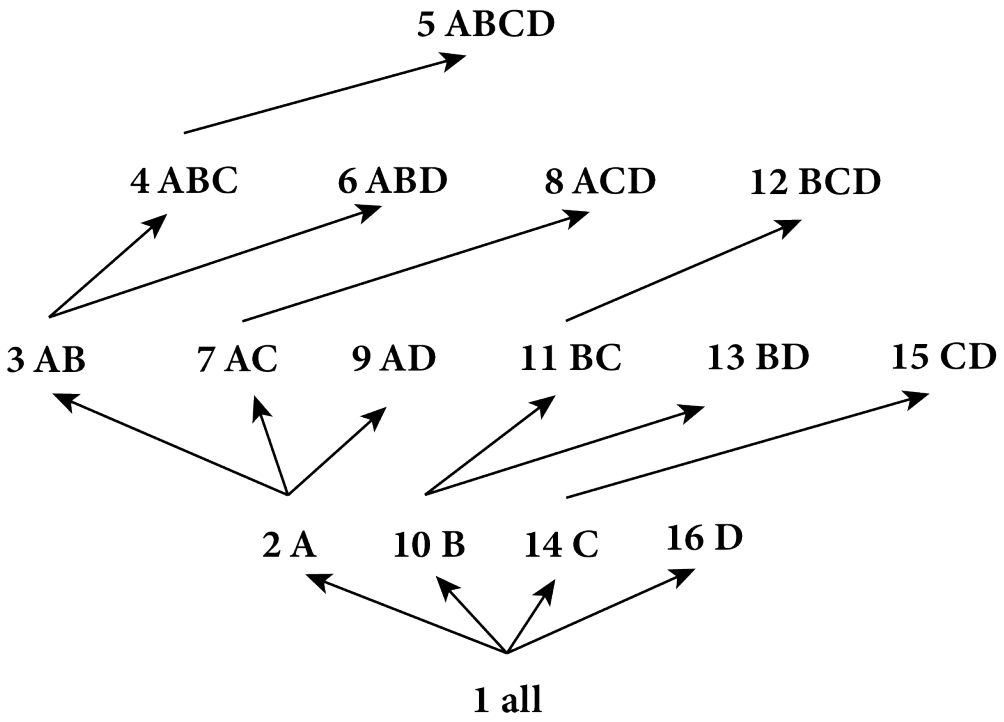
Multiway Array Aggregation

- ➢ Bottom-up computation
  - Start from base cuboid (e.g., ABC)
- ➢ Simultaneously aggregate along multiple dimensions
  - E.g., ABC => AB, AC, BC
- ➢ Full cube materialization
  - Not scalable for high dimensions



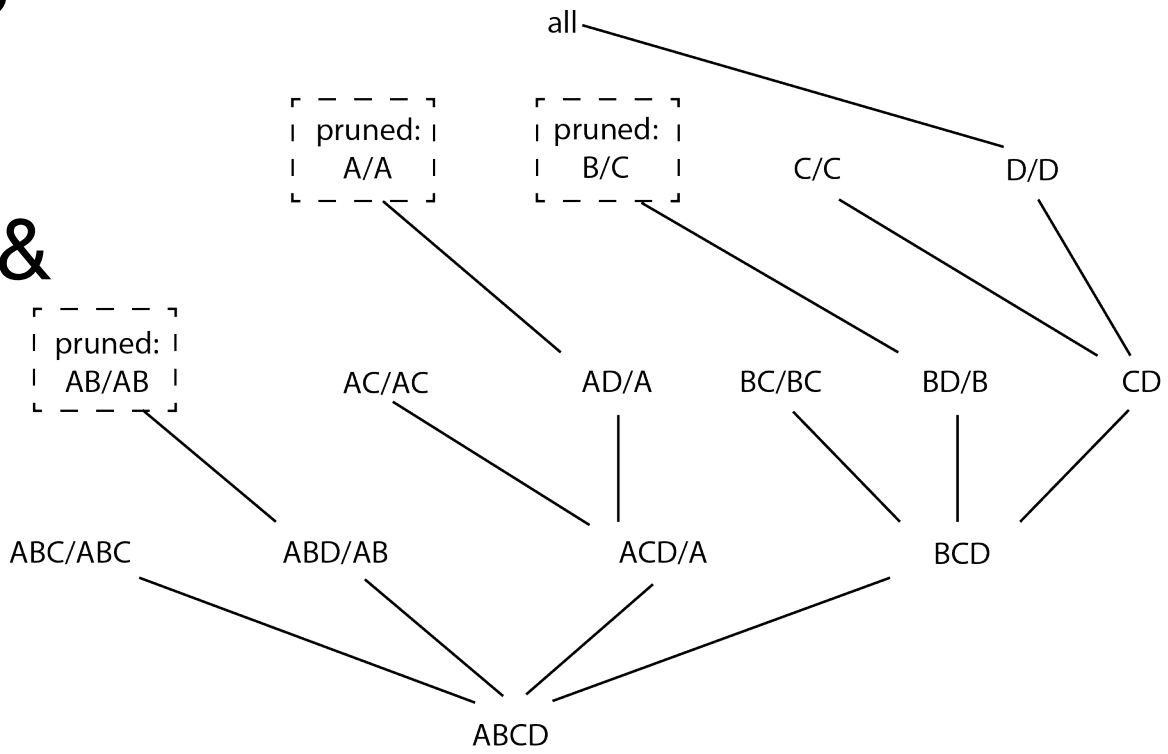
Bottom-Up Computation (BUC)

- Top-down computation
  - Starting from “1 all”
- Iceberg pruning
  - Divide dimensions into partitions while > threshold

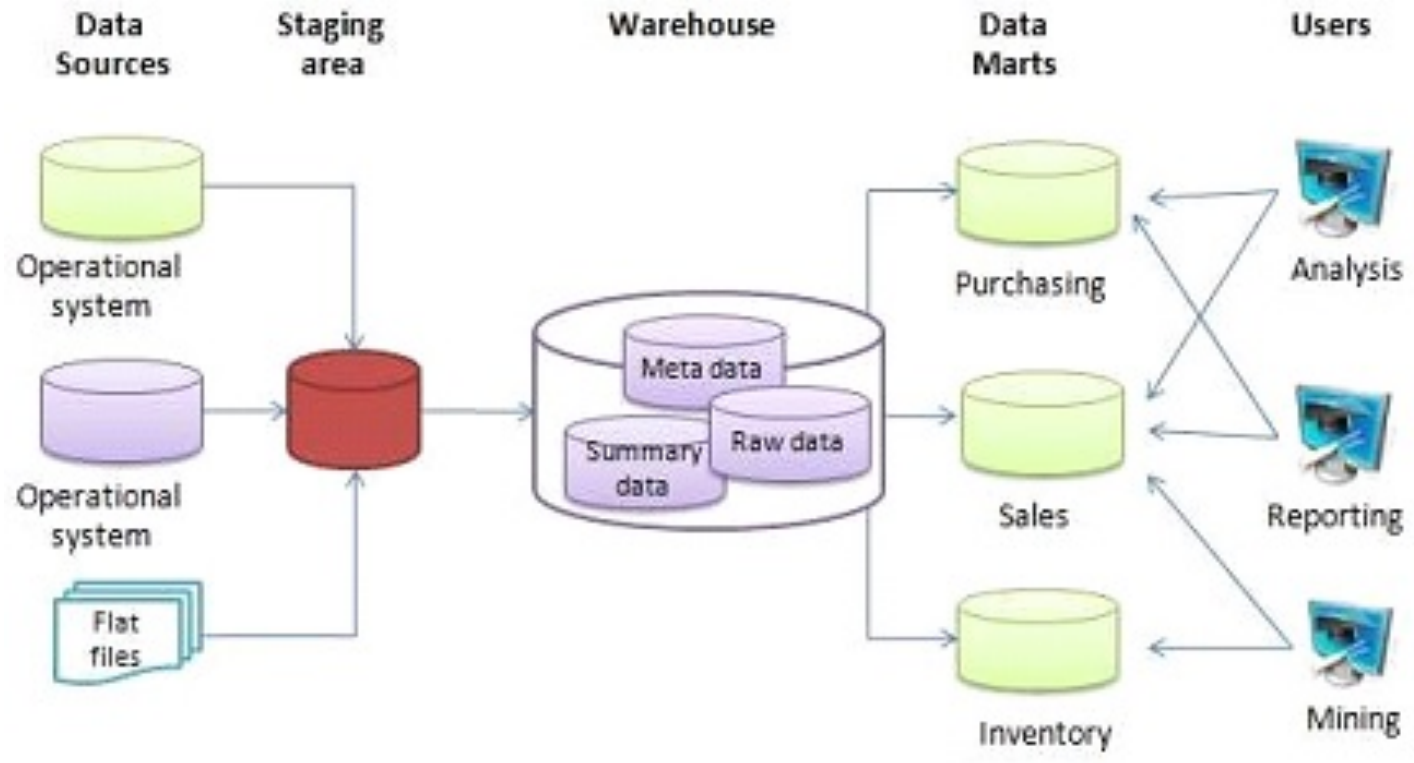


Star-Cubing

- Bottom-up computation & top-down expansion of shared dimensions



Data Warehouse Architecture



Data Sources

- Flat files: e.g., CSV
- Operational systems
- CRM (Customer Relationship Management
- ERP (Enterprise Resource Planning)
- Structured, semi-structured, unstructured

Staging (ETL)

- Extract data from various data sources
- Transform data
- Load data into the data warehouse

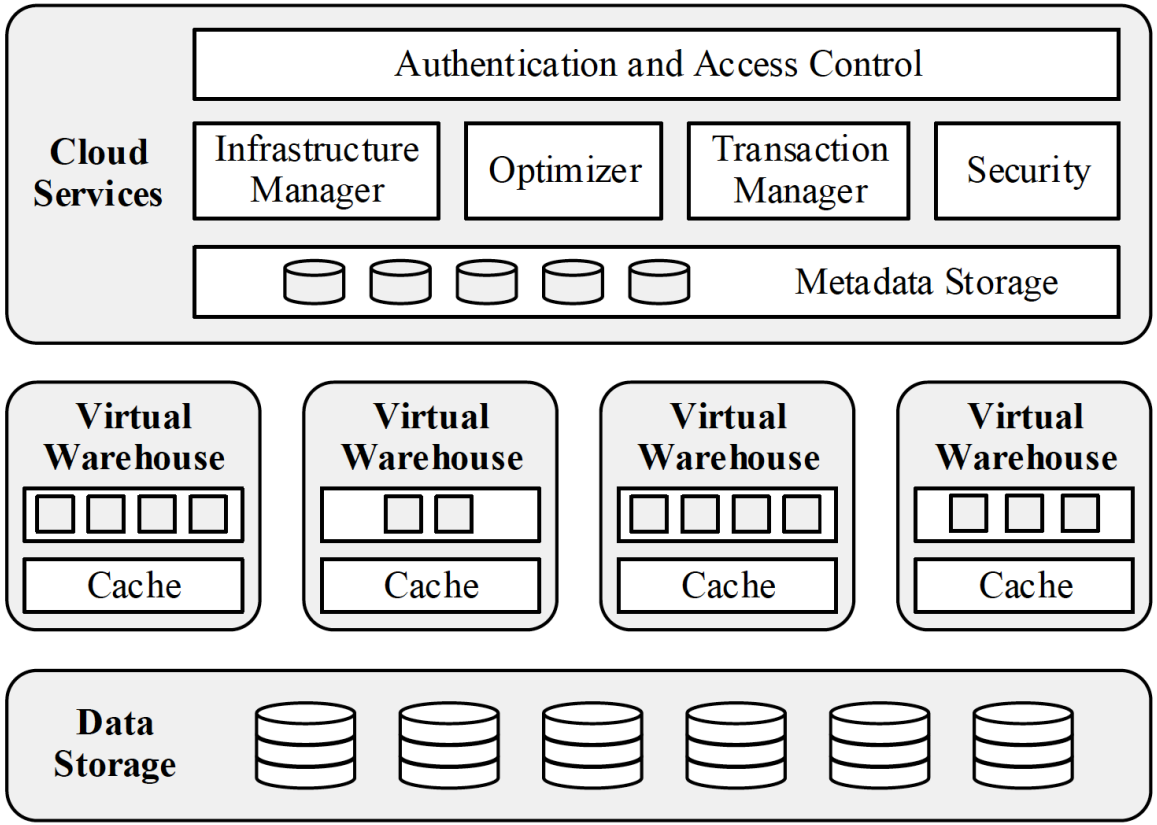
Data Warehouse

- Raw data, meta data, summary data
- Data marts: subsets with specific focuses
- Supports analysis, reports, data mining

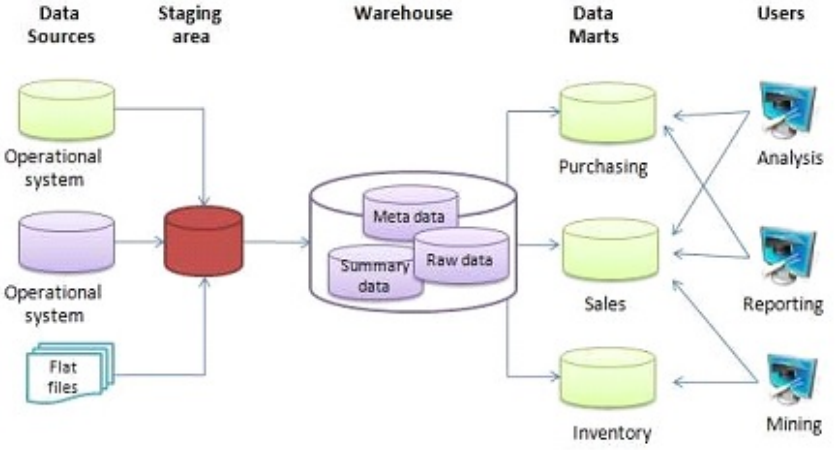
Cloud-based Data Warehousing

- Scalability
- Elasticity

- E.g.,
- Snowflake

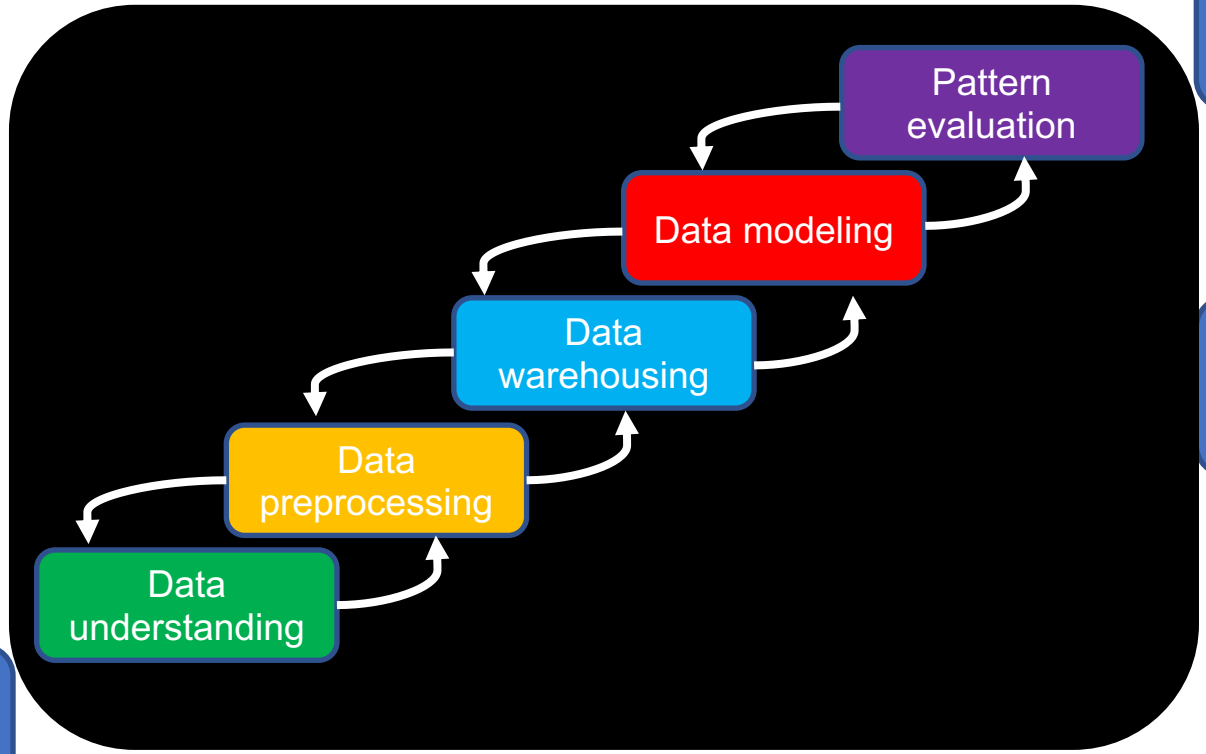


Data Warehousing



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

Data Mining Pipeline



Application

Knowledge

Technique

Data