Data for Data Mining

Project ENABLE

May 23, 2019

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Suggested Reading

- DMCT 1.3.
- DMCT 2.1.

What is Data?

- Definition
  - “Factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation.” (Merriam-Webster Dictionary)
  - “Facts and statistics collected together for reference or analysis.” (Oxford Dictionaries)
  - “A set of values of subjects with respect to qualitative or quantitative variables.” (Wikipedia)

An Example of Dataset

attributes (features)

concept

Instances (tuples)

DUPERSID	REGION	AGE	SEX	WLKDIF	ASPRIN	BMINDX	TOTEXP
40001101	2	28	MALE	-1	2	26.6	1173
40001103	2	2	MALE	-1	-1	-1	3215
40044101	2	79	MALE	3	1	34.3	14951
40050101	2	41	MALE	-1	2	31.7	3791
40052101	3	69	FEMALE	4	2	25.3	58604
40084101	2	62	FEMALE	4	2	45.9	3626
40088102	3	50	MALE	-1	2	28.3	6308
40105105	1	5	FEMALE	-1	-1	-1	1958
40148102	3	38	FEMALE	2	2	28.8	701

Types of Variable

- There are two types of variables in high-level: categorical vs. numeric.
- Categorical variables
  - Nominal: values that are distinct symbols (e.g., gender). It might be numerical in form (e.g., REGION), but these values should not be interpreted mathematically. No order or distance.
  - Binary: a special case of a nominal variable which takes only two values (e.g. True or False).
  - Ordinal: ranked order of the categories (e.g., hot, mild, and cool). No distance.

Types of Variable (cont ...)

- Numeric variables
  - Interval: ordered and measured in fixed and equal units (e.g., Fahrenheit and Celsius temperature scales). 0 is arbitrary.
  - Ratio: similar to interval variables. A zero point inherently reflect the absence of measured characteristic (e.g., Kelvin temperature and molecular weight).

Data Sources

- Types of data sources: databases, data warehouses, the Web, other information repositories, or transactional data.
  - Data warehouse: a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site (e.g., Carolina Data Warehouse for Health).
- Mining multiple data sources can lead to interesting findings.
  - For instance, genomic sequence, biological networks, and 3-D spatial structures of genomes can be combined for certain biological objects.
  - At the same time, we need to be careful about data cleaning and data integration.

Database Management System (DBMS) Architecture

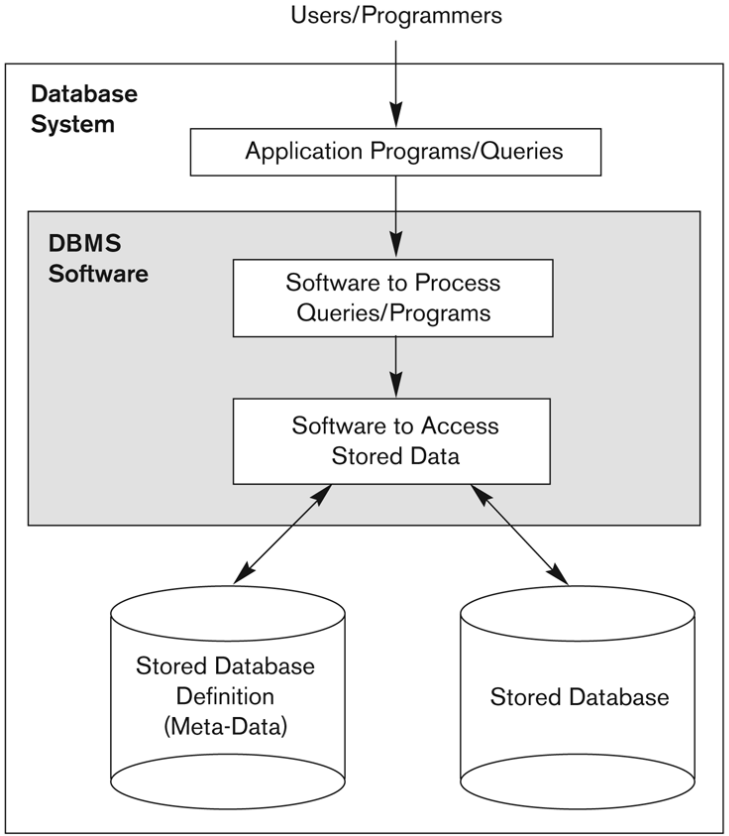
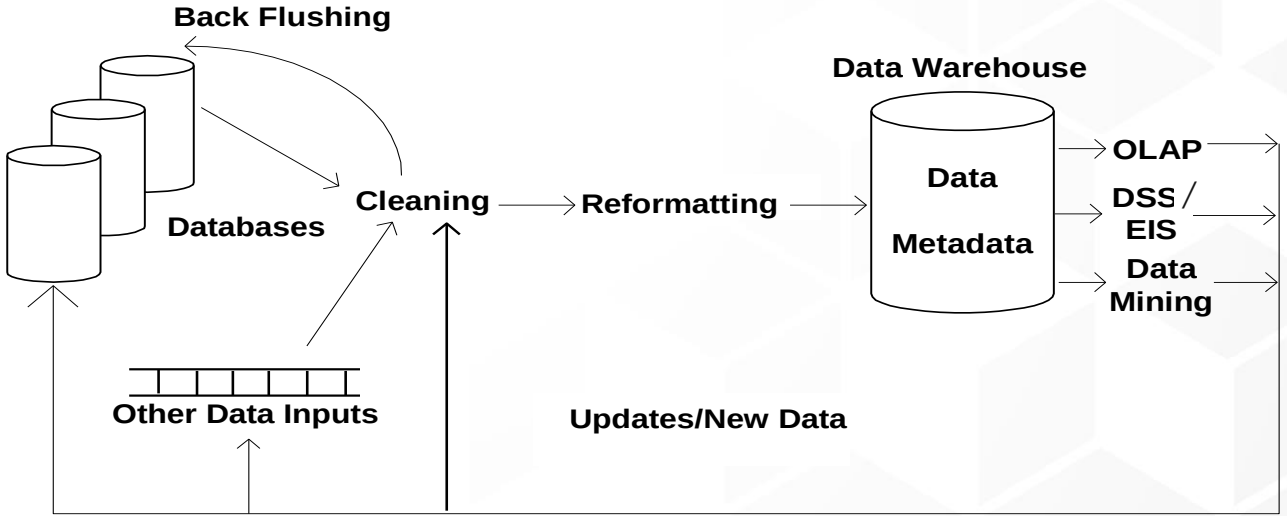


Figure 1.1
A simplified database system environment.

Source: *Electronic Health Records: Understanding and Using Computerized Medical Records.* Gartee, R. Pearson. 2017.

Conceptual Structure of Data Warehouse

- Data Warehouse processing involves
  - Cleaning and reformatting of data
  - Online Analytical Processing (OLAP)
  - Data Mining



Building A Data Warehouse

- The Design of a Data Warehouse involves following steps.
  - Acquisition of data for the warehouse.
  - Ensuring that data storage meets the query requirements efficiently.
  - Giving full consideration to the environment in which the data warehouse resides.

Building A Data Warehouse

- Acquisition of data for the warehouse
  - The data must be extracted from multiple, heterogeneous sources.
  - Data must be formatted for consistency within the warehouse.
  - The data must be cleaned to ensure validity.
    - Difficult to automate cleaning process.
    - Back flushing, upgrading the data with cleaned data.
  - The data must be fitted into the data model of the warehouse.
  - The data must be loaded into the warehouse.
    - Proper design for refresh policy should be considered.

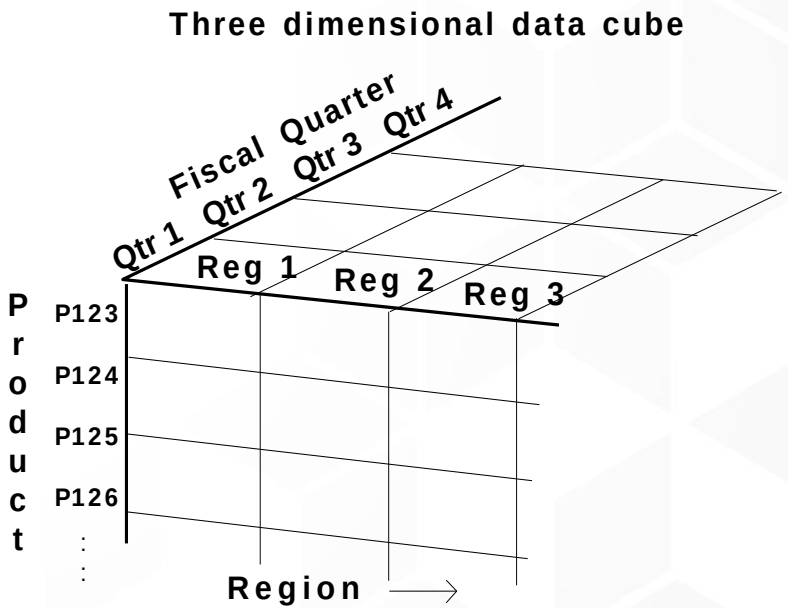
Multi-dimensional Relational Structure

- Example of Two- Dimensional vs. Multi- Dimensional

Two Dimensional Model

REGION

PRODUCT	REG1	REG2	REG3
P123			
P124			
P125			
P126			
⋮			



Multi-dimensional Schemas

- Multi-dimensional schemas are specified using:
  - Fact table
    - Each tuple is a recorded fact. This fact contains some measured or observed variable (s) and identifies it with pointers to dimension tables. The fact table contains the data and the dimensions associated with the data.
  - Dimension table
    - It consists of tuples of attributes of the dimension.

Multi-dimensional Schemas

- Star schema:
  - Consists of a fact table with a single table for each dimension.

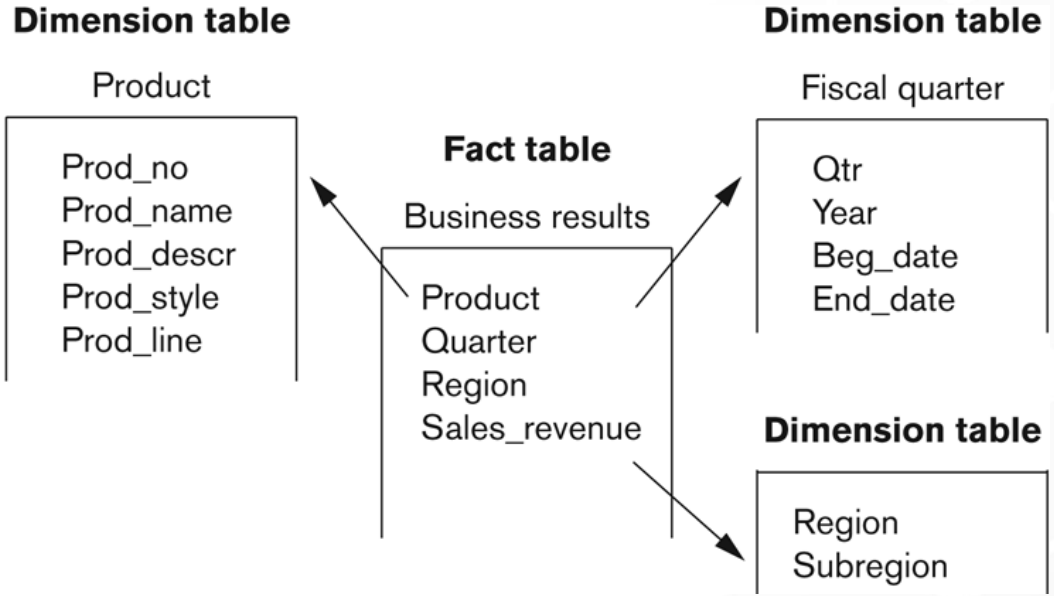


Figure 29.7
A star schema with fact and dimensional tables.

Data Source in Our Class

- Medical Expenditure Panel Survey (MEPS)
  - A set of large-scale surveys of families and individuals, their medical providers, and employers about the cost and use of health care and health insurance coverage across the United States.
  - MEPS collects data on the specific health services that Americans use, how frequently they use them, the cost of these services, and how they are paid for, as well as data on the cost, scope, and breadth of health insurance held by and available to U.S. workers.
  - MEPS is one of the core data resources used to health and healthcare policy decisions; including the evaluation of HealthyPeople CDC goals, and several key provisions of the Affordable care Act.
  - Link: https://meps.ahrq.gov/mepsweb/about_meps/survey_back.jsp

Other Data Resources

- Healthcare Cost and Utilization Project (HCUP)
- Statistics Online Computational Resource (SOCR)
- Database of Genotypes and Phenotypes (dbGaP)
- Data.gov
- Kaggle
- UCI Machine Learning Repository

Data Processing

1. Search for data.
2. Combine different datasets as necessary and convert into a standard format.
3. Clean and normalize data as necessary.
4. Apply an data mining algorithm (e.g., linear regression, SVM, and deep learning) to the prepared data.
5. Find patterns and/or knowledge.
6. Interpret the patterns and/or knowledge and do error analysis.
7. If necessary, iterate the process.

Any Questions?

Descriptive Analysis

Next Class

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL