Data and Databases

1. Data and Information

1.1 Data versus Information

- ***Data*** refers to raw, unprocessed and unorganized fact, figures or symbols that do not carry any specific meaning or context. It may be in the form of numbers, text, images, audio, video, or other.
- ***Information*** refers to processed, organized and structured data that has been given context, meaning and relevance.

Data may be transformed into data through analyzing, interpreting, or summarizing.

1.2 Structure

Unstructured data sources include:

- email, forums, newsgroups,
- letters, newspapers, books
- computer files: pdf, web pages, word processor files, …
- multimedia files: photos, audio, video

Although some of these data sources may have some inherent internal structure, they're not considered structured unless all their content can be processed by data mining tools.

A ***database*** may be used to structure data.

2. Databases

2.1 Features

Database Management Systems (**DBMS**) provide a systematic, organized and efficient approach to storing, managing, and retrieving information allowing for easy data analysis and reporting. They provide:

- ***Concurrency*** – control at the level of the data item, so multiple users can access the database; access may be temporarily prevented, but only while another user is accessing the same specific data item.
- ***Efficiency*** – the complexity of data storage structures and search algorithms are encapsulated within the database management software, allowing the casual user to benefit from years of research and optimization.

- ***Scalability*** – the interface of the database may remain essentially the same whether a database is stored in a small embedded system in a single device, or distributed across a large array of servers in diverse geographical locations. Database systems are able to handle increasing amounts of data without significant loss of performance. (Think about databases for WeChat, Weibo, Facebook, Twitter).
- ***Persistence*** – With a typical computer program, the program waits for data input, and when the program terminates, the data "goes away". For a database, the program, it is the data in the database that waits for the computer program; the program starts up, operates on the data, then the program "goes away".
- ***Data Security*** – prevention unauthorized access, disclosure, alteration or destruction. This includes ***authentication*** and ***authorization***.
- ***Data Integrity*** – accuracy and reliability of data, ensuring it remains uncorrupted during storage and processing. (What happens if there is some fault during a write, such as the device breaking or the power going out?)
-

2.2 Structure

tables, primary keys, foreign keys, records, fields

one-to-one, one-to-many, many-to-many; entity relationship diagrams

2.3 Relational Databases

A relational database is based on a relational model – a straightforward, intuitive way of representing data in tables. It uses a structured ***schema*** to define the data model and relationships between pieces of data.

2.4 Software Query Language (SQL)