Data and database

Aniel Nieves-González

Fall 2015

Data I

In the context of information systems, the following definitions are important:

1. *Data* refers simply to raw facts, i.e., facts obtained by measuring a property of an object.
2. *Information* is data presented in a context so that it can answer a question or support decision making.

Database I

1. A *a database* is an organized collection of data. In some instances can be seen as a collection of tables that are related to one another.
2. A database management system (DBMS) is the software that creates, maintains, and manipulates a database.
3. Some examples of DBMS are: MySQL, Microsoft SQL server, and Oracle.
4. A database model determines the logical structure of a database and the way data is going to be stored, organized, and manipulated.
5. Some examples of database model are:

Database II

- Relational: It was first proposed by E.F. Codd (1969). All data is represented by tuples and grouped into relations. It is based on type of *mathematical logic* that yields a method to specify data and queries. (Tuples, mathematically, are ordered lists).
- Hierarchical: In this model the data is organized in a tree-like structure that contains records. The records correspond to the tuples in relational model.
- Object: This model represents data as objects, in the sense of object-oriented programming.

6. Most DBMS are built around a specific model, but some support more than one model.
7. *Structured query language* (SQL) is the most common programming language for creating and manipulating relational databases.

Database III

8. Other important terminology is:
   1. A table or file refers to a list of data.
   2. A database can be a single table or a collection of related tables.
   3. A column or field defines the data that a table can hold.
   4. A row or record represents a single instance of whatever the table keeps track of.
   5. A key is the field used to relate tables in a database.
   6. In a relational database multiple tables are related based on common keys.
9. Let's see a very simple example of a relational database that makes use of a web page (xxx.xxx.xxx/phpMyAdmin) to collect data and to administer the database (of course, using SQL).

Collecting data I

1. There are several ways to collect data in the context of a business.
   - Transactions processing systems (TPS).
   - Loyalty card systems.
   - Other?
2. There are firms that collect data for resale. This are called data aggregators. Any concern about that? Any privacy concern?
3. Sometimes the data, abeit collected, cannot be properly used.
   - Incompatible systems. This includes the legacy systems, which are older information systems not compatible with newer technologies.
   - Transactional data that cannot be accessed for analysis.

Big Data I

The term big data is widely used in many areas and it encompasses several things.

- *Business Intelligence* combines aspects of reporting, data exploration and *ad hoc* queries, and sophisticated data modeling and analysis.
- The term *analytics* describes the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions.
- On occassions data can be a true strategic asset. This happens when it is rare, valuable, imperfectly imitable, and lacking in substitutes.

Big Data II

- Advantages based on analytics and modeling are sustainable if they result in differentiation as opposed to just operational efficiency. Advantages based on capabilities that others can acquire will be short-lived.
- A *data warehouse* is a set of databases designed to support decision making in an organization. It is structured for fast online queries and exploration and may aggregate enormous amounts of data from many different operations.
- A *data mart* is a database focused on addressing the concerns of a specific problem or business unit.
- Hadoop: it is an open-source project created to analyze massive amounts of raw information better than traditional, highly-structured databases. Scalability, flexibility, cost, and fault tolerance are its main advantages.

Big Data III

- In essence Hadoop is a software framework for parallel computing and parallel storage in computer clusters.
- Data mining is the process of using computers to identify hidden patterns and to build models from large data sets. Data used for data mining must:
  - be clean and consistent
  - reflect current and future trends.

Discussion

- Walmart.

Discussion

- Walmart.
- Ceasar's Casinos.