Data Management

**Big Data**

* The superabundance of data available today
* Allows us to spot trends more rapidly
* Prevent disease
* Track crime

**Big Data Problems**

* More data than space to store it
* Quantity grows faster than networks can handle
* Hard to protect data

**Data Quality**

* IT and systems support organizations by interpreting, analyzing, storing, and organizing the data
* The data needs to be accurate, complete, consistent, accessible, relevant, concise
* It’s difficult to have high quality data.

**Difficulties in managing data**

* Data increases exponentially
* Data is scattered and collected by many individuals
* Data comes from many sources
* Security

**Source Examples**

* Credit card swipes
* Video surveillance
* Blogs
* Radiology scans
* Emails

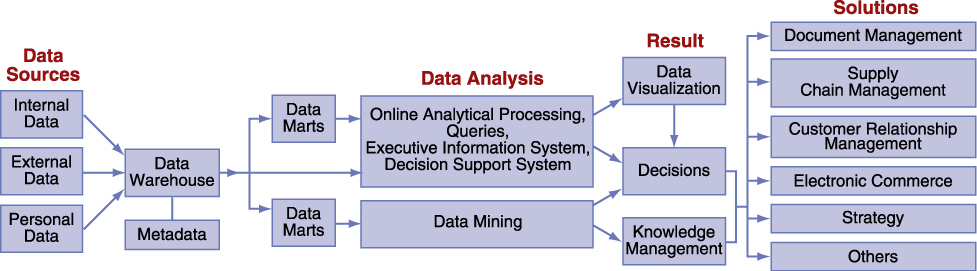
**Data Degrades Over Time**

* Customers have new addresses
* Employees are hired or fired
* New products
* Companies go in and out of business
* **Data Rot –** problems with the media on which data is stored
  + Temperature, humidity, etc. can cause physical problems
  + Technology changes can cause problems on accessing that data
* **Data Errors** – data gets out of date, inaccurate, technically corrupt

**Data Life Cycle**

* Data -> Information -> Knowledge -> Wisdom
* Businesses run on processed data such as information and knowledge (look back at *Information Technology* for definitions.
* Managers apply this knowledge to business problems and opportunities
* Businesses transforms data into knowledge and solutions in several ways

**Data Processing**



**The Database Approach**

* A database management system (DBMS) is a set of programs that provide tools to add, delete, access and analyze data stored in one location
* DBMS interfaces with database which provides all users integrated access to the data
* Different models but the relational database is the most popular and easy to use
* Other models: hierarchical, network, object-oriented

**Popular Products**

* DB2 – from IBM
* Access, SQL server – Microsoft
* Oracle – Oracle Corp.

**DBMS Benefits**

* DBMS will minimize
  + **Data Redundancy:** same data stored in many places
  + **Data Isolation:** applications cannot access data associated with other applications
  + **Data Inconsistency:** Various copies of data don’t agree.

**DBMS Issues**

* Maximize the following issues
  + **Data Security:** Databases have high security to deter mistakes or attacks
  + **Data Integrity:** Data meets constraints such as no Alphabetic characters in SIN numbers
  + **Data Independence:** Applications and data are independent of one another (apps can be designed to access the same data)

**Relational DBMS (RDBMS)**

* Query languages, data dictionary, normalization
* Based on concept of 2D tables
* **NOT** one huge tables, but rather multiple tables that relate
* If there were only one big table that would create data redundancy
* **Query Languages**
  + Requests information from a database
  + Most popular language used to request information
  + SQL allows people to perform complicated searches by using simple statements or key words.
* **Data Dictionary**
  + Defines the format necessary to enter the data into the database
  + Provides information on each attribute such as its name, whether it’s a key value or not, etc.
* **Normalization**
  + Method for analyzing and reducing the database to its most streamlined form for:
    - Minimum Redundancy
    - Maximum Data Integrity
    - Best Processing Performance
  + Normalized data is when attributes in the table depend only on the primary key.

**RDBMS Benefits**

* Increased flexibility
  + Handle changes quickly
  + Provides users with different views
  + **1 physical view** (physical storage of information of storage device)
  + **Multiple logical views** (how user logically accesses information)
* Increased performance & Scalability
  + A database must increase or decrease in size to meet demands
  + **Scalability:** how a systems adapts to its scalability
  + **Performance:** How quickly a systems performs certain processes
* Reduced redundancy
  + **Minimizes the following problems**
    - Inconsistency of data describing the same thing
    - Waste of space
    - Difficulty securing data in many places
* Increased Integrity
  + Measures the quality of information
  + **Integrity Constraint:** rules that help ensure the quality of information
  + **Relational Integrity constraint:** Rules that enforce basic and fundamental
  + **Business-critical integrity constraint:** Enforces business rules vital to an organization’s success and often requires more insight and knowledge than relational integrity.
* Increased security
  + **Security features**
    - Passwords (authentication)
    - Access level (Who has access)
    - Access Control (types of users)

**Data Driven Websites**

* Interactive website which uses a database to keep it updated and relevant (Google)
* **Advantages**
  + **Development:** changes can be made anytime without reliance on developers
  + **Content Management:** No programmer required to update, which reduces costs
  + **Future Expandability:** Able to grow faster since it’s adaptable
  + **Human Error is Minimized:** Contains error trapping mechanisms to ensure data entered is correct
  + **Reduced Production and Update Costs**
  + **Increased Efficiency:** Because sites are computer driven
  + **Improved Stability:** Data resides in a database and not organized by someone who may leave at any time