Data Storage & Databases

Were does all this stuff go?

- Each source systems usually has its own storage, but ...
  - Optimized for functional performance, not data extraction & analysis
    - Online Transactional Processing (OLTP) vs.
    - Online Analytical Processing (OLAP)
  - Typically has a lot more stuff than we are interested in
  - Risky to access directly; 'back end' load can impact 'front end' stability
  - Retention times vary; data may not be stored locally for very long
- Sometimes we actually do connect directly to source systems, or even intercept data as it 'streams' through a connection
- However, the solution is usually to gather data into a separate storage location
  - May be centralized, semi-centralized, or 'virtualized'

Data Storage & Databases

Data Files

Databases

Data Storage & Databases – File Systems

File Systems

- Think of your own computer; can essentially put anything we want in there and just note it's name and location
- Handles all sorts of information, including 'unstructured' data really well
- Primary limitation is in 'readiness' for use and the ability to interconnect different elements in a meaningful way
- The Hadoop Distributed File System (HDFS) is a 'Big Data' manifestation of the idea, using massively parallel processing on relatively inexpensive infrastructure to efficiently store large amounts of varied information

Data Storage & Databases – Data Files

- **Delimited Text Files**
  - Data stored as text, with breaks between fields & rows defined by ‘delimiters’ - specific characters or formatting codes
  - Comma-separated value (CSV), Tab-delimited and Pipe-delimited (|) most common
- **Extensible Markup Language (XML) Files**
  - Flexible structure for encoding documents & data, especially for Web applications
- **Log Files**
  - Largely nonstandard output from machine data sources, including the Web
  - Generally require some sort of parser to interpret
- **Application-Specific Files**
  - Excel Files
  - Specialized files like SAS,SPSS or Tableau files

Data Storage & Databases – Database Systems

- **Databases**
  - Simply an organized collection of data
  - Usually refers to the structure/design itself as well as the actual data that resides in the structure
- **Database Management System (DBMS)**
  - Software used for creating, maintaining and accessing databases
- **Relational Database**
  - Invented by E. F. Codd at IBM in 1969-70
  - Far and away the most common type of database system
  - Stores information in two dimensional tables with defined set of relationships among them
  - Highly efficient and intuitive way of storing information

Data Storage & Databases – Other Types

There are a variety of emerging database types, most designed to handle ‘big-data’ applications and/or ‘unstructured’ data

- Graph Databases
  - Based on graph theory; tends to work well with highly interconnected data (geographic, network, etc.)
- Document Store
  - As name suggests, generally designed to store documents and key pieces of metadata
- Columnar Databases
  - Improves performance by storing data in ‘columns’ of similar types vs. the ‘rows’ of relational databases
- Key-Value Store
  - Simple database system which stores information in pairs (key & value)
  - Can be used to achieve very high speed in certain types of operations

Data Storage & Databases

Data Files

- Delimited Text Files
- XML Files
- Log Files
- Application-specific Files

Databases

- Relational Databases
- Graph Databases
- Document Stores
- Columnar Databases
- Key-Value Stores