Introduction to Data Management
Course Description
Draft of May 19, 2009

Structural place in the curriculum

- 4 credits (3 weekly lectures, 1 weekly section, no lab)
- Pre-requisites: 143
- Subsequent courses: The following courses would have this course as a pre-requisite: the new 444 (Database management implementation/internals) and the 400-level parallel data processing class (currently the Hadoop class).
- Taken by: Optional for CS and CE students.
- Catalog description: To be determined

Course Overview / Goals

Advances in our capability to generate and collect information coupled with decreasing disk-space prices are pushing us toward a world centered around data management. Databases are at the heart of modern commercial application development. Their use extends beyond this to many other environments and domains where large amounts of data must be stored for efficient update and retrieval. The purpose of this course is to provide a comprehensive introduction to the use of database management systems for applications.

Description of Possible Homework, Etc.

The course would have a mix of homeworks, mini-projects, and exams.

We envision the following mini-projects:

1. Using a relational database management system (DBMS): SQL queries (e.g., 444 project 1), creating a database, creating tables, inserting/deleting data.
2. Building an application that uses a DBMS (e.g., 444 project 2). This lab includes transactions.
3. Perhaps we could have the students add a web interface to the previous app.
4. Becoming a power-user: indexes, views, security, (e.g., 444 project 3)
5. Other data models: XML, text, and multimedia data management.

We envision the following homeworks:

1. How to go from real-world entities to a relational schema (e.g., 444 hw1)

2. Other homeworks will either cover the theoretical side of what the mini-projects cover or they will otherwise follow the lectures.

There will be a midterm and a final.

Possible Textbook

Database Systems: the Complete Handbook, by Hector Garcia-Molina, Jennifer Widom, and Jeffrey Ullman.

Approximate Topic List

1. Week 1 - Motivation and relational DBMS overview
    a. Motivation: the importance of data, data management, and data management systems.
    b. Overview: different types of data, data models, and data management systems (relational, XML, text, small/large), etc.
    c. Relational DBMSs overview: the lifetime of a query, DBMS architecture overview
2. Week 2 – Using relational DBMSs
    a. SQL: how to create a database, load data, insert/delete, and ask queries (including aggregation, subqueries, etc).
3. Week 3 – Modeling a real problem using a database
    a. Conceptual design: ER diagrams, functional dependencies
    b. Views and integrity constraints.
4. Week 4 - Transactions
    a. Transactions: what they are and how to use them
    b. Different levels of isolation and possible anomalies
5. Week 5 – Advanced features
    a. Indexes and how to use them
    b. Security and access control
6. Week 6 – Other data models: XML, Text, and multimedia
7. Week 7 – Data on the web
    a. Data integration
    b. Information retrieval
    c. Asking structured queries over the web (?)

8. Week 8 – Data warehousing (OLAP) and data mining
    a. Data cleaning and extract-transform-load pipelines
    b. Data cubes
9. Week 9 – Big data processing and cloud computing
    a. Parallel DBMSs and MapReduce
    b. Databases as a service (SimpleDB, etc.)
10. Week 10 – Buffer.

Stuff that we won't have time to cover and that will go into the next 400-level database class:

1. How a DBMS is implemented; how a DBMS works inside.

Stuff that we do not cover in either course (should we?) but cover in 544.

    a. Scientific data
    b. Sensor data
    c. Data streams
    d. Probabilistic data

Background / Correspondence to Old Curriculum

This course would correspond to the first part of 444, but would go deeper into using different types of data management systems. It would NOT discuss DBMS internals.

Similar Courses Elsewhere

Stanford's CS145 - Introduction to Databases (also on quarter system)
http://infolab.stanford.edu/~widom/cs145/

CS145 provides the student with a comprehensive introduction to the design of databases and the use of database management systems for applications. We will cover the relational model, relational algebra, and SQL, the standard language for creating, querying, and modifying relational and object-relational databases. We will also learn about XML data, including the XML languages XPath, XQuery, and XSLT. The UML approach to database design will be covered, as well as relational design principles based on functional dependencies and normal forms. A variety of other issues important to database designers and users will be covered, including indexes, views, transactions, authorization, integrity constraints, and triggers. Finally, we will cover several advanced topics such as data warehousing, data mining, data stream processing, and uncertain data.