EE382V Data Engineering

Fall '18
Date: 8/30/18
Version: 1.1

Syllabus

Instructor Information: Professor Daniel Miranker

GDC 5.436

mailto:miranker@cs.utexas.edu

Office Hours: By appointment made by email

T.A.: Zeyuan Hu

T.A. Offices hours: By appointment made by email, iamzeyuanhu@utexas.edu

Course Home Page: Canvas will be used for grading and distribution of materials. Piazza will be used for class discussions.

Required Text: "Database Systems: The Complete Book", Garcia-Molina, Ullman, Widom, Prentice Hall.; 2nd edition. See http://infolab.stanford.edu/~ullman/pub/dscbtoc.txt

Additional Readings: Technical papers linked into the course web site.

Quick Syllabus: The course spans classic, core graduate database material by covering nearly all of text chapters, 7, 8, 16-18, and parts of text chapters 11,12 and 21. Contemporary (big/cloud) database practice and advanced cutting edge issues, such as, contemporary data models, e.g. graphs, Corresonding indexing and query processsing techniques will also be covered, as well as distributed data integration. Precise selection of contemporary papers, TBD.

Preequesites: A previous database class is not assumed. However, it is assumed that you are familiar with the rudiments of SQL programming, (any job exposure, whatsoever, will be ample), but it is also assumed that you have a basic comfort with relational algebra and associated formalism. This is precisely just Chapter 2 of the text. It comprises learning a chunk of vocabulary and a simple algebra (yes, just like high school). So, it is easy, but very important to catch-up. It will be common for expressions in relational algebra to appear across the course.

Grading:

There will be a midterm and a final. Among the homework will be a short, integrated series of implementaiton assignments that build on each other and, effectivelly, form a project on database query optimization. Thus, total homework is heavily weighted and individual homeworks may have different weight.

The three grading components will be weighted as follows, midterm 30%, final 35%, homework 35%.

Late Homework policy: Each student is permitted one late homework no questions asked, provided it is turned in before the solution set is posted. (Typically 3 or 4 days). More than one homework late, don't ask.

Topics – Not in Temporal Order

EE382V Data Engineering Spring 16
Created on 8/30/18
Version: 1.0

I. Introduction:
   A. Architecture of Database Management Systems
      I. Classic relational database managementsystems
      II. Contemporary, cloud database infrastructure
      III. Definition and role of transactions
   B. The rapidbly changing landscape of memory technology, and its implications
      I. Solid State Disks (are finally here and taking over).
      II. Non-volatile main-memory is very close

II. Data Models
   A. Introduction: Using UML class models as technology agnostic data model
   B. Comparng and compiling UML to classic relational (entity-relationship) E-R models
   C. Generalizing to contemporary data models. (most likely MongoDB as a case study).

III. Views, Constraints, Rule Systems Management (Ch. 7, 8)
   A. Views
   B. Integrity constraints
   C. Active-Database Systems
   D. Deductive Database Systems (Datalog) (time permitting)

III. Index Structures (Ch. 14 + supplementary material)
   A. Tree-based (classcal indexing methods)
      a. B-trees (as they underpin relational database)
      b. R-trees (as they underpin geospatial databaes)
   B. Contemporary methods of indexing (those exploited by cloud databases)
      a. Bit-vector indexes, and the role of compression
      b. Bloom-filters

IV. Query Execution (Ch. 15, 16)
   A. Basic Join Algorithms
   B. Expression tree representation
   C. Cost functions
   D. Transformation rules
   E. System R Optimization Algorithm
   F. Parallel query processing

V. Transaction Management (Ch. 17, 18.1-18.5, 19.1-19.2)
   A. ACID Properties
   B. Logging and error recovery
   C. Serializability
   D. Two phase locking
   E. NoSQL/Cloud Transaction Models

VI. Data Integration (Ch. 21.1-21.3, but mostely supplementary material)
   A. A large-scale architectures
      (Data warehouse, vs, Big Data vs. Data Lake vs. Mediator (virtual) Architectures
   B. Challenges and approaches to homogenizing data
   C. Graph-based methods (a.k.a The Semantic Web and Knowledge Graphs)