COSC6340: Database Theory
Models, Queries, Transactions, Algorithms and Data Structures

Catalog Description

First order logic, set theory. Relational algebra, advanced database normalization (BCNF, 4NF and 5NF). Query optimization: query plans, models, NP-hard and NP-complete problems. Transactions: serializability, recovery. Algorithms and data structures for secondary storage. Parallel and distributed database computations. Applied math for large data sets.

Prerrequisites

The course is self-contained, but assuming the student has a strong CS background and reasonable math background.

Students should have basic knowledge on: relational database systems, SQL, normalization up to 3NF, discrete math, algorithms and evaluating programming languages (i.e. compilers theory). On the programming side, students should be familiar with C++, SQL, Unix (Linux) development environment including ssh, vi, g++, python and node.

Course contents

This is a graduate course covering theory of data systems (relational databases, logic and non-relational) and applied mathematics to store, update, search, query and analyze large data sets. On the other hand, the course will cover important theory from statistics, machine learning and numerical methods, which is helpful to compute machine learning models at large scale.

There are two main textbooks [1, 2]. The course will require reading specific textbook chapters, but also some research papers.

This is a detailed list of topics covered 100%. Database systems topics: Set theory and first order logic. Relational database systems theory: tuple, relational calculus, completeness and consistency. Algebras for set and relational operators. Queries: Conjunctive, negation, linearly recursive queries, query plans, NP-complete query optimization problems, I/O cost models. Transactions: schedules, serializability, recovery, locking, timestamp ordering, deadlock avoidance/resolution, indexing for append-only workloads, log-based algorithms. Algorithms and data structures for secondary storage: external merge sort, blocked binary search, B+-trees, extensible hashing, hash join, sort-merge join, multidimensional indexing. CAP Theorem, time complexity, I/O cost. Parallel and ditributed processing: speedup, scaleout, Amdahls Law, distributed vs one-node multicore architectures.

Applied math on large data sets topics (partially covered, depending on students interest, as time allows): large sparse matrix multiplication, large matrix factorization, non-convex optimization, stochastic gradient descent for large n, summarization of large data sets via sufficient statistics, central limit theorem, approximating functions with mixtures of distributions, extending and generalizing regression models to neural networks.

Grading

- 40%: 2 written exams: individual, in class. 1st exam: around 5th week of class. 2nd exam: around 10th week of class. Exams: based mostly on textbook chapters covered in class. Slides should be used as a guideline, but not as reference. Exams will not be based on slides.
- 20%: 2 written theory homeworks, developed by a team of 2 students. Deliverable: PDF with clean math notation. It is acceptable to use ChatGPT, disclosing how it was used.
- 40%: 2 programming assigments: (1) solving hard graph questions with SQL or Datalog queries; (2) simple algorithm on query or transaction evaluation in C++/Linux. Developed by a team of 2 students. Deliverables: working code on our Linux server.

Contact information

- Messages: messaging system chosen during 1st week of class (e.g. discord (most likely), whatsapp, telegram). Note: MS Teams will not be used because experience shows it does not scale well to hundreds of messages and it is not efficient for smartphones. Any programming or exam question should be posted and channels will be monitored by TAs and instructor.
- Preferred contact for academic and personal matters: in person after class or during office hours. Only if you can provide a justification of not coming to class, a short email message to firstname@uh.edu. Warning: do not write to the other Ordonez professor, in the Physics department.
- office hours: MW 2:30pm-3:30pm.

References

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases : The Logical Level*. Pearson Education POD, facsimile edition, 1994.

[2] H. Garcia-Molina, J.D. Ullman, and J. Widom. *Database Systems: The Complete Book*. Prentice Hall, 1st edition, 2001.