COSC6340: Data Systems Implementation

Catalog Description

To be updated.

Prerrequisites

The course is self-contained, assuming the student has a strong CS background. It is expected students have programming experience and basic knowledge on: optimization algorithms, relational database systems, operating systems and implementing programming languages (i.e. compilers theory).

Course contents

This is a graduate systems-level course on implementing data storage, data retrieval and data analysis systems, including structured files (e.g. CSV, JSON), relational database management systems (SQL, DBMSs), and data science Python libraries (e.g. Pandas, JSON). There is no textbook, but these references are helpful: [2], [3] and [1]. The course will require reading research papers and reading from development web sites.

Topics include the following. Review of relational database systems theory. Internal subsystems of most data systems: secondary storage principles, faster storage architectures, buffer and main memory management, indexing data structures, concurrency and version control, transaction processing, fault tolerance: for transaction processing (recovery) and for long query processing (parallel, incremental). Query and transaction processing: query optimizer, faster transaction processing (distributed, lockfree, relaxing consistency), Advanced SQL (SPJA queries, derived tables/view, pivoting, recursive queries). Reverse data model engineering, security mechanisms against hackers, pushing processing to main memory, blockchain support for distributed transactions, query languages and data manipulation libraries beyond SQL and Datalog.

Grading

- 80%: 1 programming project.
- 20%: Midterm exam (around 10th week, take home to be solved in 2 days, technically difficult).

Programming will be done with a combination of modern C++, classic C, ANSI SQL and Python libraries, depending on project. Programming project will be done in pairs (i.e. a team of 2 students). Programs will be carefully tested by TA and instructor for correctness and benchmarking, with correctness being the most important requirement, followed by a comparison against competing systems or libraries.

Contact information

- Messages: messaging system chosen on 1st week of class (e.g. whatsapp, viber, discord, telegram).
- contact for personal matters: via MS Teams.
- office hours: posted on CS department web pages.

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.

[3] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2006.