COSC6339: Big Data Analytics

Instructor: Carlos Ordonez

1 Short Description

Big data overview, Parallel DBMSs, Hadoop ecosystem. R and Python runtime systems.

2 Course Contents

This is a “systems” course, which will cover theory, algorithms, data structures and programming to analyze big data, covering both database, data science and non-database approaches. Past research is being expanded to analyze much bigger data sets, mixing structured and semi-structured content, with more complex statistical and machine learning models and exploiting parallel processing.

Topics: I. Big Data overview: 4 Vs, Data Lakes vs Data Warehouses vs Data Swamps, Data integration and cleaning, Parallel and Distributed Computing, Big Data Analytics problems: ML, graphs, streams. II. R and Python: language, functional constructs, runtime, data structures (data frames and matrices). III. Parallel DBMSs: parallel and distributed architectures, OLTP versus cubes, Data warehousing and denormalization, advanced SQL (pivoting, horagg, keyword search, recursive queries), UDFs, row/column storage, indexing versus ordering, ETL and pre-processing, data pre-processing and data cleaning, machine learning and graph algorithms. IV. Hadoop ecosystem: HDFS vs Posix, subsystems (Yarn, Storm, Zookeeper, Cassandra, Hive, SPARQL, Parquet), containers vs virtual machines, text versus numeric processing, Spark (including MapReduce), Graph systems, Search Engines (IR models, architecture, keywords, page rank, web spider).

The course will require reading important CS research papers, mainly from Big Data and Database systems, investigating latest research on DBLP and ACM libraries. There is no textbook, but [2] and [1] are recommended.

Pre-requisites: It is encouraged, but not required, that COSC6340 (Database Systems), COSC6373 (Parallel Computation), COSC6320 (Algorithms) and COSC6376 (Cloud Computing) were taken before (or concurrently). This is an advanced “systems” course, it is not a machine learning or data science course.

3 Grading

- 60%: 2 programming projects (same weight).
- 40%: Final exam.

The exam puts emphasis on parallel efficiency, memory management, I/O optimization, linear algebra, numerical methods and graph algorithms. Both programming projects are required to get C+. Programming projects must be done in pairs, where the team of 2 students is randomly assigned by the professor; the team partner cannot be chosen by students. The course will use R/Python, a parallel DBMS and the Hadoop/Spark system. Programs will be developed mostly in C++/C, SQL and Python or R. Students will analyze data sets from: science, documents and web data.
References
