COSC6339: Big Data Analytics

Instructor: Carlos Ordonez

1 Short Description

Big data overview, R system, Parallel DBMSs, Hadoop ecosystem.

2 Course Contents

This course will cover theory, algorithms, data structures and system programming to analyze big data. Data mining research is being revisited 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, Data integration, Parallelism, Big Data Analytics problems: ML, graphs, streams. II. R: language, functional programming, environment, runtime, data structures (vectors, matrices, lists, data frames), statistical and machine learning models, graphs. III. Parallel DBMSs: parallel architecture, OLTP versus cubes, Data warehousing and denormalization, query languages: SQL, datalog, AFL, UDFs, row/column/array storage, indexing versus ordering, ETL and pre-processing, advanced SQL (pivoting, horagg, keyword search, recursive queries), data pre-processing and data cleaning, integrating machine learning algorithms. IV. Hadoop ecosystem: HDFS vs Posix, subsystems (Yarn, Storm, Zookeeper, Cassandra, Hive, SPARQL), MapReduce, text versus numeric processing, Spark, Graph systems, Search Engines (IR models, architecture, keywords, page rank, spider); comparing big data analytics technologies (speedup, cost, scalability, fault tolerance, Java vs C++, programming ease).

The textbook is [2], complemented by [1]. The course will also require reading important CS research papers and investigating latest research on DBLP and ACM libraries.

Pre-requisites: It is encouraged, but not required, that COSC6340 (Database Systems) and COSC6342 (Machine Learning) and COSC6373 (Parallel Computation) were taken before (or concurrently). Courses like Data Mining, Numerical Analysis and Algorithms are also desirable.

3 Grading

- 80%: 2 programming projects (same weight).
- 20%: Midterm exam.

Both programming projects are required to get B-. Programming projects must be done in pairs (team of 2 students to be assigned by professor). The course will use R, an array DBMS (SciDB) and the Hadoop/Spark system. Programs will be developed in R, Java, C++ and Scala. Students will analyze data sets from: science, UCI machine learning repository, documents/files, web data.

References

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

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