COSC6339: Big Data

1 UH Catalog

- New Title: Big Data (instead of Big Data Analytics)
- Description: Algorithms, hardware and software infrastructure for storing and analyzing big data. Data pre-processing, data exploration, scalable machine learning, large graphs, parallel and distributed processing, parallel database systems.

2 Course Contents

2.1 Course approach

This is a ”systems” + ”data science” graduate course, covering theory, algorithms and system infrastructure to store and analyze big data. Topics include data science languages, parallel database systems, cloud computing, NoSQL and the Hadoop stack. Compared to other analytics courses, in this course we go deeper into how machine learning and data mining research was expanded to analyze big data, mixing structured and semi-structured content, with more complex statistical and machine learning models, exploiting parallel processing and larger main memory.

2.2 Topic Details

1. Big Data. Big Data Definition. Data repositories: data collection, data sources, file systems, databases, data lakes, polystores. Data management: storage, integration, data preprocessing, security.
2. Systems and infrastructure. Modern hardware (accelerators, large RAM, faster storage), large-scale file systems, cloud computing server architectures, Hadoop stack, parallel and distributed database systems, HPC libraries, Machine Learning libraries, graph analytic systems.
3. Large-scale exploratory data analysis: Querying, Data Quality, Statistics, Data mining, Visualization.
4. Advanced Big Data Analysis. Machine Learning, Linear algebra, Tensors, Neural Networks, Deep Learning, Graphs. Acceleration mechanisms (gradient descent, summarization, online algorithms).

The course will require reading CS research papers, from Big Data, Machine Learning and Database Systems conferences and journals.

3 Curriculum and Grading Information

3.1 Pre-requisites

This is a hands-on course combining ”systems”, ”machine learning” and ”data science”. Therefore, it is desirable students have background on these areas at the undergrad level.

- Courses: It is encouraged, that the equivalent of COSC2436 (basic data structures and algorithms), COSC3380 (Database Systems), COSC3360 (Operating Systems), COSC3337 (Data Science, introduction to Machine Learning), were taken before.
- Programming: Familiarity with Linux, Python, SQL and C++ development. All programming homeworks must work in Linux (read below).

3.2 Grading

- 50% Exams: There will be one exam (partial or final) and/or several quizzes covering theory, algorithms and system infrastructure (decided by instructor). Exam and quizzes will be in the classroom during official lecture time.
- 50%: Programming. Project involving a combination of languages and systems. Programming will be mostly Python, but some functions or steps will involve SQL, C++, or Java, depending on the target problem. A browser GUI in JavaScript will be optional as well. The delivered program should work on a Big Data System in the cloud. The project will require pre-processing, exploring, transforming big data to build a data set, suitable for machine learning model or graph analytics. The input will be combination of tables, plain files, documents. Optionally, there may be images. Required software will be open-source, installed on a cloud server, running in Linux. Target problems will be either: predictive machine learning models or graph analysis. Python interpreter: Python3. Compiler: GNU C++. Database System: relational database engine (SQL) or non-relational database (noSQL). Teams: The code will be developed in teams of 2 students, assigned by the instructor.
- Final grade: Exams and quizzes (if applicable) are mandatory. Students unable to take any exam must take an oral makeup exam and must provide a verifiable justification. The scale to assign letter grades will be standard (A is 90, B is 80, and so on). Programs must be submitted and produce correct output in order to get at least C+.