COSC6339: Big Data

1 UH Catalog

- Title: Big Data
- 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 data mining research was expanded to analyze big data, mixing structured and semi-structured content, with more complex statistical and machine learning models and exploiting parallel processing.

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. Large-scale exploratory data analysis. Querying, Statistics, Data mining, Visualization.
3. Advanced Big Data Analysis. Machine Learning, Linear algebra, Tensors, Neural Networks, Deep Learning, Graphs. Acceleration mechanisms (gradient descent, summarization, online algorithms).
4. Systems and infrastructure. Large-scale file systems, cloud computing analytic architectures, Hadoop stack, parallel and distributed database systems, HPC libraries, Machine Learning Systems, Graph analytic systems.

The course will require reading CS research papers, from Big Data, Machine Learning and Database Systems conferences and journals, available on DBLP, IEEE and ACM digital libraries.

3 Academic Information

3.1 Pre-requisites

This is a 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, intro 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%: one exam (partial or final) and/or several quizzes covering theory, algorithms and system infrastructure. Exam and quizzes will be in the classroom during official lecture time.
- 50%: Project involving a combination of languages and systems: C++, Python, SQL, Java and JavaScript. The program will work on some 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 on big data. The input will be combination of tables, plain files and 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. Compiler: GNU C++. Python interpreter: Python3, Database System: SQL engine or 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+.