EECS 284 – BIG DATA SYSTEMS AND ANALYTICS

Number of credits: 4 units

- Instructor: Florin Rusu
- Office: SE2-210
- Phone: 209-228-4286
- Email: frusu@ucmerced.edu
- Web: http://faculty.ucmerced.edu/frusu

Meeting time.

- Lecture: T,R 12-1:15PM; CLSSRM 276
- Lab: M 10:30AM-1:20PM; KOLLIG 202
- Office hours: T 11AM-12PM, R 11AM-12PM; SE2 210. Or by appointment.

Exam.

- Thursday, May 10, 8-11AM (CLSSRM 276)

Catalog description. This course aims to familiarize students with techniques for processing large amounts of data. Starting with the latest innovations in hardware, data processing architectures are presented as well as algorithms for managing large quantities of data. Although the main focus is data analytics, significant attention is dedicated to transactional processing.

Textbook and other required materials. There is no textbook required. Materials to be used during the course include *Readings in Database Systems, 5th Ed., MIT Press* by M. Stonebraker and J. Hellerstein and research papers from major data management/data mining/machine learning/systems conferences available online through the ACM and IEEE portals.

Course objectives/student learning outcomes. The goal of this course is to expose students to state-of-the-art techniques for managing and processing large amounts of data. At the end of the course, students will be able to assess the properties of high-performance computer architectures, apply parallelizing and data-partitioning methods to design scalable algorithms, and use the latest industry programming paradigms for data management. This will be achieved by presenting a series of data management architectures, algorithms, and programming paradigms, analyzing theoretically their properties, and assessing empirically their practical importance. In essence, the objective of this class is to introduce the latest technologies made available by the industry, such as multi-core processors, solid-state drives (SSD), etc., reflect on how they affect the existing data processing techniques, and design algorithms and methods that take full advantage of their characteristics.

Students will learn about the latest trends in the hardware industry and how they shape the evolution of the data processing techniques. They will get detailed exposure to the current research in data management through reading and providing critique for seminal papers in the field and direct experience with research prototype systems. These will both enhance their ability to read research literature as well as to understand how theoretical concepts are made practical and applied to real-life problems. The progress students make in assimilating the class material will be continuously tested through multiple oral presentations of a semester-long project in which the students are required to apply the learned concepts to their own research. The project report required at the end of the course is designed to enhance the technical writing skills of the students as well as their ability to provide constructive feedback for the work of their colleagues (students will be required to review the work of others in a peer review fashion). In summary, the students enrolled in this course will get exposure to the current research in data management and will experience the latest innovations from the industry. These will benefit both students more interested in research aspects as well as students looking for a more hands-on experience.

By taking this course, students will be able to (i.e., **student learning outcomes**):

- Analytically read research literature as well as understand how theoretical concepts are made practical and applied to real-life problems.
- Communicate course concepts through oral presentations.
- Demonstrate technical writing skills and an ability to provide constructive feedback for the work of their colleagues.
- Understand current research in data analytics and the latest innovations from the industry.

Program learning outcomes. The course relates to the following EECS program learning outcomes:

- Students are able to identify novel and significant open research questions in electrical engineering and computer science and are able to situate such questions in the contexts of current research literatures.
- Students are able to apply their knowledge of computing, mathematics, science, and engineering to the analysis of technological problems, as well as to the design and implementation of viable solutions to those problems.
- Students are able to design and conduct experiments and computational simulations for the purpose of evaluating and comparing proposed solutions on the basis of empirical evidence.
- Students possess the characteristics of lifelong learners; they are able to acquire and use new techniques, skills, and engineering and scientific tools for research and development in electrical engineering and computer science, as well as to develop new methods and make new discoveries.
- Students practice a high standard of professional ethics, including integrity in the conducting and writing of research.
- Students communicate effectively through oral, visual, and written means, effectively addressing a broad range of technical audiences.

Prerequisites by topic.

- Computer architecture understanding
- Computer system design concepts
- Algorithm fundamentals
- Or consent of instructor

Course policies. The course consists of 3 hour lectures per week, ”seminar style”. A research paper is presented by a student or the instructor in each lecture. Students are required to read the papers and write a summary to be handed to the instructor prior to the class. Each student is asked to write an individual research report on a commonly agreed topic with the instructor (it is desirable that the topic is related to the research area of the student). The progress students make with their reports is tested multiple times during the semester through oral presentations. In the lab, students will work on individual programming assignments related to the topics discussed in class or on the practical portion of their project (a high-end cluster is required for the projects).

Academic dishonesty statement. Each student in this course is expected to abide by the University of California, Merced's Academic Honesty Policy. Any work submitted by a student in this course for academic credit will be the student's own work.

Students are encouraged to study together and to discuss information and concepts covered in lectures. Students can provide/receive ”consulting” to/from other students. However, the permissible cooperation should never involve one student having possession of a copy of all or part of the work done by someone else, in the form of an email, an email attachment file, a storage device, or a hard copy. Should copying occur, both the student who copied work from another student and the student who gave material to be copied will receive zero credit for the corresponding assignment. Penalty for violation of this Policy can also be extended to include failure of the course and University disciplinary action.

During examinations, each student has to do only their own work. Talking or discussing is not permitted, nor students comparing their papers, copying from others, or collaborating in any way. Any collaborative behavior during examinations will result in failure of the exam and may lead to failure of the course and University disciplinary action.

Disability statement. Accommodations for students with disabilities: The University of California, Merced is committed to ensuring equal academic opportunities and inclusion for students with disabilities based on the principles of independent living, accessible universal design diversity. I am available to discuss appropriate academic accommodations that may be required for students with disabilities. Requests for academic accommodations are to be made during the first three weeks of the semester, except for unusual circumstances. Students are encouraged to register with the Disability Services Center to verify their eligibility for appropriate accommodations.

Topics. This offering of the course is focused around data management techniques in machine learning (ML). We study how ML is integrated in (big) data systems and how ML systems use data management and processing algorithms. Specifically, we focus on how to train and execute inference for large models and large data with scalable hardware and software. The content of the course evolves around scalable gradient descent optimization and includes the following topics:

- Parallel gradient descent algorithms
- Training data compression
- Out-of-core gradient descent algorithms
- GPU-based gradient descent algorithms
- Hardware accelerators
- Hyper-parameter tuning
- In-database ML
- Scalable inference
- ML over normalized data
- Scalable linear algebra
- Optimization method selection
- Distributed gradient descent with Parameter Server
- Training dataset generation

Assessment/Grading policy.

- Project: 65% (report 30%; code + demo 30%; review 5%)
- Presentations: 20% (2 X 10% for each presentation)
- Labs: 15% (3 X 5% for each lab)
- $\geq$ 900: A; $\geq$ 800: A-; $\geq$ 770: B+; $\geq$ 730: B; $\geq$ 700: B-; $\geq$ 670: C+; $\geq$ 630: C; $\geq$ 600: C-; $\geq$ 500: D; $<$ 500: F
- Curved grading may apply only in special situations.