**Big Data Systems & Analytics**

### Instructors

Ling Liu, School of CS, College of Computing

### Goal

This course teaches concepts, techniques, algorithms and optimizations of big data systems and data analytics, including big data processing systems, fundamental models and optimizations for data analytics and machine learning, and the real world deployments of big data systems and analytics.

### Course Description

This is a graduate-level course in big data systems and data analytics. The course teaches the basic design and optimization principles and algorithms for scalable and dependable big data systems and data analytics algorithms. The core of the course contains centralized and decentralized architectures, platforms and programming models for designing and developing big data systems, and for supporting large scale big data analytics in business and science-engineering domains.

### Learning Objectives

*By taking this course, students are expected to*

1. Become familiar with concepts, techniques and optimizations of big data systems and the most commonly used data analytics and machine learning models and algorithms;
2. Gain exposure to the applications of the techniques, models and optimizations of big data systems and analytics to real world problems;
3. Formulate big data systems or analytics problems, and identify the suitable big data systems or analytics models for them;
4. Design appropriate architectures, build software systems, or apply learning algorithms and optimization techniques to address them.

### Course Educational Outcomes

*Upon successful completion of this course, students should be able to*

1. Categorize, compare, and contrast various big data systems in terms of architecture, functionality, and optimization techniques; identify and articulate those that better suited for particular types of real-world big data problems than others with respect to functional and non-functional requirements;
2. Categorize, compare, and contrast various learning algorithms in terms of learning models and learning objectives, identify and articulate those that better suited for particular types of real-world big data problems than others with respect to learning objectives and learning quality metrics;
3. Design and implement a big data system for either big data storage and processing applications or big data analytics applications;
4. Identify and articulate various types of data analytics architectures and models for different types of machine learning problems (e.g., classification, clustering, frequent pattern mining, recommendation, summarization, …);
5. Discuss and critique research papers on these topics;
6. Design and carry out a big data system or analytics project within their area of interest, apply the learned theory and techniques to new types of big data problems within this area, and analyze the performance of the system and algorithms developed or deployed;
7. Understand and possibly identify open research questions in these areas.

### Recommended Textbook

Extensive course notes will be provided, in addition to reading materials from conferences and journals.

### Grading

20% Homework and 4-5 assignments

50% Projects

20% Final Exam

10% Class Discussion and Participation

### Prerequisites

Students are expected to have taken Introduction to database systems (CS 4400/6400 or equivalent). In addition, students are expected to have basic programming skills.

### Academic Integrity

Academic dishonesty will not be tolerated. This includes cheating, lying about course matters, plagiarism, or helping others commit a violation of the Honor Code. Plagiarism includes reproducing the words of others without both the use of quotation marks and citation. Students are reminded of the obligations and expectations associated with the Georgia Tech Academic Honor Code and Student Code of Conduct, available online at [www.honor.gatech.edu](http://www.honor.gatech.edu).

### Learning Accommodations

If needed, we will make classroom accommodations for students with documented disabilities. These accommodations must be arranged in advance and in accordance with the Office of Disability Services ([http://disabilityservices.gatech.edu](http://disabilityservices.gatech.edu/)).

Excused Absence Policy

### <http://www.catalog.gatech.edu/rules/4/>

### Outline of Topics

Data has been the No. 1 fast growing phenomenon on the Internet for the last decade. Big data systems demands both high performance computing and elastic utility driven computing. Big data analytics holds the potential to reveal deep insights such as social influence among customers by analyzing business transactions, user-generated feedback ratings, social and geographical data.

This course will cover concepts, techniques, algorithms and optimizations of big data systems and analytics, and explores big data opportunities from a variety of science and engineering applications, and examine various research problems and challenges that are critical for developing big data systems and big data applications. Main topics to be covered include but not limited to:

1. fundamentals of big data computing systems architectures and platforms
2. fundamentals of big data storage systems and optimizations,
3. fundamentals of big data processing systems and optimizations,
4. fundamentals of systems and software design for big data analytics
5. fundamentals of distributed big data computing, incl. cluster computing and distributed file systems.
6. fundamentals of geographically distributed data analytics systems.
7. fundamentals of general models for data analytics and knowledge discovery.
8. fundamentals of data summarization and frequent pattern mining models and algorithms.
9. fundamentals of unsupervised machine learning models and algorithms.
10. fundamentals of unsupervised machine learning models and algorithms.
11. fundamentals of ensemble learning models and algorithms
12. fundamentals of deep learning models and algorithms

We will also cover big data applications that pose new challenges to big data systems and analytics, such as healthcare, mobile commerce, social media, Internet of Things, software defined computing, cyber manufacturing, cyber-physical systems.

This course is designed to provide the fundamental training for big data scientists from high performance big data computing systems, to big data applications and big data analysis and management algorithms, and to look beyond the present status of the Big Data and conjecture what possible future technologies and applications will evolve.