5460. Big Data Scaling (2023 Spring)

Maizie (Xin) Zhou, PhD

Syllabus

Class Information

Class Hours: Monday and Wednesday 2:15 - 3:30 pm

Room: 19th & Grand G168

Office Hours: By appointment (maizie.zhou@vanderbilt.edu)

TAs: Can Luo (can.luo@vanderbilt.edu)

Mingxing Rao (mingxing.rao@vanderbilt.edu)

Yichen Liu (yichen.liu.1@vanderbilt.edu)

Grades: Grading will be based on the following elements:

Homework Assignments, In-class live coding participation, Midterm,

Final project (writeup and presentation) and Participation bonus.

Assessment:

• **Homework Assignments 35%:** There will be homework assignments nearly for every topic. You are allowed to work in groups on the homework, but you must write up your own solutions in your own words. ASSIGNMENTS ARE DUE AT **11:59 midnight** OF THE DUE DATE THROUGH BRIGHTSPACE.

• **In class live coding participation 15%:** There will be living coding participation nearly for every topic.

• **Midterm 25%:** in class, no internet

• **Final Project 25%:** Final projects can be done in groups of 1 - 3 people. We encourage you to form a group of 3 members, since groups of 3 usually lead to the best outcomes. We will talk about more details in class.

• **Participation Bonus 5%:** Students will receive bonus points by answering questions in class. Maximum bonus points account for extra 5%. Each question will account for 0.25%.

Late days Policy

Each student will have a total of 5 free late (calendar) days applicable to any assignment except the midterm and final project. Free late days can be used at any time, no questions asked. Each 24 hours or part thereof that a homework is late uses up one full late day. Once these late days are exhausted, any homework turned in late will be penalized 10% per late day.

Late days are never transferable between students, even students in the same group.

Class Announcements: All students are held responsible for all announcements made in the class and Slack channel.

Course Materials: The course relies on Jupyter notebooks, slides, and online resources (https://spark.apache.org/docs/latest/). The lecture slides and Jupyter notebooks for each week will be found in the BrightSpace as the course progresses. The textbook (Mining of Massive Dataset: http://www.mmds.org/) is useful and recommended, but not required. You can download it for free or purchase the hardcopy from Cambridge University Press.

All academic work at Vanderbilt is done under the Honor System.

The course will discuss data mining and machine learning algorithms, and the emphasis will be on MapReduce and Spark as tools for creating parallel algorithms that can process very large amounts of data.

Topics include Cloud Computing, Distributed File Systems, MapReduce, Apache Spark, SparkSQL, Regression, Tree Methods, Locality Sensitive Hashing (LSH), Clustering, Recommendation Systems, and Mining Data Streams.

Schedule:

Week	Topic / Contents
1	Intro to the Course and Cloud Computing Intro to Computing Clouds (ACREE, Linux) Friday Zoom Session by TAs (more intro for ACCRE and Slurm)
2	Parallel Processing in Python Big Data Processing Infrastructure (Distributed File Systems, Hadoop, MapReduce)
3	Hands-on tutorial for Google Clouds (Dataproc and cloud storage) Apache Spark (Spark Architecture, Resilient Distributed DataSets (RDDs), Transformations and actions)
4	Apache Spark (continue) Spark DataFrames
5	Spark DataFrames (continue) Spark SQL
6	Review (slides) Midterm
7	Spark MLlib and Regression
8	Locality-Sensitivity Hashing
9	Tree Methods
10	Clustering
11	Recommendation System
12	Mining Data Stream
13	Final Project Presentations

14	Project Writeup Due