# MAS DSE 230 Syllabus - Scalable Analytics

Spring 2021

**Instructor**
Mai Nguyen

**Teaching Assistant**
Sagar Hathwar

**Class Logistics**

| | |
|---|---|
| Lecture | Saturday 9am - 4pm on 4/03, 4/17, 5/01, 5/15, 5/29, and 6/05 |
| Canvas | Zoom links, Assignments |
| Piazza | Announcements, Q&A |
| Gradescope | Assignment submission, Grades |
| Office Hours | Will be announced on Piazza, and Zoom links available on Canvas |

**Course Description**
This course is designed to provide students with the skills and knowledge to perform analytics at scale. Topics cover both systems and analytics, and include basic principles of computer systems and parallelism; analytics process; analytics algorithms; scalable computing; and cloud-based analytics. Tools and techniques to perform analytics on large-scale data will be introduced. Students will get hands-on experience on distributed and cloud-based platforms to perform scalable analytics.

**Schedule**

| Session | Topic | Assignment | Points |
|---|---|---|---|
| 1 | Computer Systems & Parallelism | Spark | 5 |
| 2 | Big Data & Distributed Processing | Spark | 10 |
| 3 | Big Data Analytics | Spark<br>Project proposal presentations (in-class) | 15<br>10 |
| 4 | Big Data Analytics & Cloud Computing | Dask<br>AWS | 15<br>10 |

| | | | |
|---|---|---|---|
| 5 | AWS Analytics, Deep Learning, Other Topics | AWS | 10 |
| Finals | Project Presentations | Final project presentations (in-class) | 25 |

Session 1 – Basics of Computer Systems & Parallelism

- Big Data introduction
- Basics of computer hardware and software
- Memory hierarchy
- Parallelism principles
- Speedup

Session 2 – Big Data & Distributed Processing

- Big Data characteristics and challenges
- Distributed processing
- Hadoop
- Spark
- Analytics process

Session 3 – Big Data Analytics

- Spark core & libraries
- Analytics with Spark MLlib
- Model selection & evaluation
- Project overview presentations

Session 4 – Big Data Analytics & Cloud Computing

- Dask
- Cloud computing
- AWS basics

Session 5 – AWS Analytics, Deep Learning, & Other Topics

- AWS SageMaker
- Deep learning overview
- Other Topics

Finals - Presentations

- Project presentations

**Materials**

- Required
  - The Data Scientist's Guide to Apache Spark
  - Apache Spark
    - https://spark.apache.org/docs/latest/
  - Dask
    - https://dask.org/
  - AWS EMR
    - https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-overview.html
    - https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-spark.html
  - Amazon SageMaker
    - https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works.html
    - https://docs.aws.amazon.com/sagemaker/latest/dg/gs.html
- Recommended
  - *Spark: The Definitive Guide (1st edition)* by Chambers and Zaharia
  - *Learning Spark (2nd edition)* by Damji, Wenig, Das, & Lee
- Reference
  - *Introduction to Data Mining (2nd edition)* by Tan, Steinbach, Karpatne, & Kumar
  - *The Elements of Statistical Learning* by Hastie, Tibshirani, & Friedman
  - *Computer Organization and Design (5th edition)* by Patterson & Hennessy
  - *Operating Systems: Three Easy Pieces* by Remzi & Arpaci-Dusseau

**Grading**

- Grading components
  - Programming Assignments 65%
  - Project Proposal Presentation 10%
  - Final Project Presentation 25%
- Policy regarding assignments
  - Programming assignments are individual work only
  - PySpark or Dask will be used as specified for the programming assignments and project
  - Students can work in pairs on the project
    - We will provide a signup sheet. Please indicate your team information, or your decision to work individually, on this sheet before class meets for **Session 2**
  - You can post conceptual or high-level questions on Piazza. But *do not post any code* on Piazza.
  - Participation to promote understanding and thoughtful discussion of the course material is encouraged.

- Late policy
    - A late penalty of *20% per day* will be applied if an assignment is submitted after the due date. A late submission can be accepted *up to 3 days* after the due date.
    - Applies to programming assignments only, not to project. Project components must be submitted by the due date.
- Academic Integrity
    - If plagiarism is detected in your assignment or if cheating is detected during an exam, University authorities will be notified for appropriate disciplinary action to be taken. You will also get zero for that component of your grade.
    - The complete UCSD Policy on Integrity of Scholarship is available here: UCSD Policy on Integrity of Scholarship