Syllabus

General Information

CS1656 – Data Science and Big Data Technology

Instructor: Yong Zhao, 532N

Contact: yong.zhao@scupi.cn

Teaching Assistant: Yu Fu

Contact: fuyu05@stu.scu.edu.cn

Course Meeting Times

Lectures: One session / week, Friday 8:15AM – 11:00AM

Office Hours: One session / week, Tuesday 2:00AM – 5:00PM and by appointment

Overview

We are already in the Data Science and Big Data era, we have large scale data from almost all fields, such as from people, society, vehicles, devices, sensors, signals, finance, astrology, medicine, network, web sites etc. Almost all industries are embracing the challenges of Big Data and want to derive valuable information and gain insights from the data.

Because of the importance and broad impact of Big Data, new architecture, hardware and software tools and algorithms are quickly emerging. A data scientist needs to keep up with this ever-changing trend and be able to create state-of-the-art solutions for real-world challenges.

This course firstly gives an overview of the concept and development of data science, and the evolution of big data technology. Then, fundamental concepts and platforms will be covered, such as Cloud computing, data centers, distributed data storage, including structured and unstructured data storage, and then we will introduce batch and real time processing platforms such as Hadoop, Spark, Flink and other tools. The course will go on to introduce different analytical algorithms, and visualization and human computer interaction devices for big data. Moreover, students will learn introductory AI-related data technologies, such as Generative AI and Large Language Models. Students will have fundamental knowledge on Big Data technologies to handle various real-world applications and challenges.

Course Objectives

The goal of this course is for the students to understand the concepts and evolution of data science of big data technology, be familiar with big data storage and processing platform and systems; and get hands-on experience in big data analytics and algorithms.

This course can also position students to compete for research projects in the data science field and also be ready for industry jobs related to big data analytics and processing.

Prerequisite

Data Structure, math, and one or more programming languages: C, Java, Python

Class Schedule

Lecture slides will be available for copying or posted on Canvas.

1. Introduction to Data Science and Big Data Technology
2. The lifecycle of big data
3. Cloud Computing and data centers
4. Data collection and ETL
5. Distributed data storage and databases
6. Big data processing platforms
7. Batch data processing – the Hadoop ecology
8. Real time data processing – Spark and Flink
9. Big data analytics
10. Knowledge graph
11. Workflow
12. Visualization and Interaction
13. Security and privacy
14. GPU and CUDA
15. Generative AI and Large Language Model

Learning outcome

At the end of the course, the students should be able to:

- Understand the key concepts and technologies related to data science and big data
- Know the essential requirements to become a data scientist
- Have in-depth knowledge of big data lifecycle and big data processing
- Manage data with database management systems and cloud infrastructure
- Collect and analyze big data using big data platforms and systems
- Perform big data analysis using various machine learning algorithms

- Evaluate outcomes and make decisions based on data processing results
- Effectively represent and communicate results
- Have basic understanding of Generic AI and large language models

Grades

The course will be hands-on practices on big data processing, so the course projects will take a majority part of the grades.

Grades will be based on homework, course projects and final exam.

Homework and attendance: 30%

Course project (1 major project or a few small projects): 40%

Final Exam: 30%

Collaboration and Academic Honesty Policy

Individual work on all homework and examinations is required, Cheating and copying other students' homework/exam are strictly prohibited. Any violation of this policy will be treated severely.

Collaboration amongst students to understand the course material and to work on course projects is strongly encouraged, however each student should take on different/distinguishable responsibilities in the course projects.

Course Reading Material

This is no formal textbook used in this course, but the students can use the following books for reference.

1. Yong Zhao, Architecting *Big Data – Big Data Technology and Algorithms Explained*.
2. Ian Foster, Dennis Ganon, *Cloud Computing for Science and Engineering*
3. Balamurugan Balusamy et al., *Big Data: Concepts, Technology and Architecture*

Supplemental readings from selected papers may also be assigned throughout the semester.