ISOM3370: Big Data Technologies (Fall 2026)

Instructor: Dr. JIA, Jia (justinjia@ust.hk)
Office: LSK 5045
Office Hours: By appointment

Teaching Assistant: Mr. Samuel Lai (imsamuel@ust.hk)
Office: LSK 4065
Office Hours: By appointment

Class Schedule

Section	Date & Time	Venue
L1	Mon. 01:30PM - 02:50PM Fri. 09:00AM - 10:20AM	LSK 1001

Course Website: https://canvas.ust.hk. All course materials and announcements will be posted on this site. You are advised to check it regularly throughout the course.

1. COURSE DESCRIPTION

Over the decades there has been an explosion of data. With diversified data provisions, such as large Internet sites, sensor networks, scientific experiments, and government records, the volume of data that we create, and capture keeps increasing at an exponential rate. The off- the-shelf techniques and technologies that we already used to store and analyze data cannot work efficiently for large-scale data processing. The challenges arise especially in the context of data-intensive computing. We need to develop and create new techniques and technologies to excavate "Big Data" and benefit our specified purposes.

The emergence of large-distributed clusters enables data storage and computation to be distributed across thousands of commodity machines in data centers. One key breakthrough that makes this possible is the development of abstractions and frameworks that allow us to reason about computations at a massive scale, while hiding low-level details such as data movement, synchronization, and fault tolerance. Such disruptive technologies have become important data processing platforms for a variety of applications, and have transformed business, science, and many aspects of our society.

This course will introduce big data technologies, starting with MapReduce, which is the first of these datacenter-scale computation abstractions and whose Hadoop implementation lies at the core of an application stack that is gaining widespread adoption in both industry and academia. Because of the success of Hadoop, a large number of big data tools, with specialization ranging from cluster resource management to complex data analytics, were built on and around Hadoop, creating a complete big data application stack. We will then cover some of the tools in this stack, such as Hive and Spark. The course will cover some widely used distributed algorithms in academia and industry.

2. PREREQUISITES

(ISOM 3230 or ISOM 3320 or ISOM 3400) and ISOM 3360
Knowledge of Python programming, database and data mining is required.

3. LECTURE NOTES AND READINGS

All course materials (Lecture slides, assignments, and lab handouts) are available on the course website Canvas. Please check the course website frequently for updates.

4. ASSESSMENT AND GRADING

This course uses criterion-referencing assessment (i.e., grades will not be assigned on a curve). Assessment details, where applicable, are provided below:

Assessment Components	**Contribution to Overall Course Grade (%)**
A. Class Participation	5%
B. Lab and homework submissions	30%
C. Midterm Exam	30%
D. Final Exam	35%

A. Homework Assignments

There will be three individual homework assignments in total. Each assignment will include conceptual questions and hands-on tasks. Two of the assignments are lab-related, as described below. Completed assignments must be submitted through Canvas for grading. Detailed requirements will be announced during the semester.

Students are encouraged to submit their work early if there is any uncertainty about meeting the submission deadline. Submissions made up to 24 hours after the deadline will receive a 25% penalty; submissions made between 24 and 48 hours after the deadline will receive a 50% penalty. Submissions received more than 48 hours after the deadline will not be accepted for credit.

B. Lab Session

This is primarily a lecture-based course; however, lab participation is an essential part of the learning process. Students are expected to develop their understanding of big data technologies through hands-on practice. During lab sessions, you are expected to bring and use your own laptop, remain fully engaged, and follow the provided instructions. For each lab, you must complete and submit a report for grading. For the final two lab sessions, the lab report will take the form of a homework assignment.

C. Exams

This course includes two **closed-note, paper-based** exams. The midterm exam will assess topics covered in the first half of the course, while the final exam will cover material from the second half.

The midterm exam is tentatively scheduled to take place in class **on Oct. 26**. The final exam will be held during the final examination period, with the exact date to be announced later in the semester.

Arrangements for the Make-up Exam

Makeup opportunities for the exams will be considered only in cases of documented health or family emergencies, or for official, university-sanctioned activities. Students requesting a make-up exam for the final exam must submit the University's request form (Form EX-16) and appropriate supporting documentation, such as a medical certificate where applicable, to the course instructor by email for consideration.

Score Posting and Appeal Policy

Upon completion of grading, all scores will be posted on Canvas. Students are responsible for reviewing their scores and verifying their accuracy. Any appeal regarding a score must be submitted by email to imsamuel@ust.hk. Please note that score appeals will not be considered after the designated review and appeal period has elapsed (e.g., two working days after the scores are released).

5. POLICY ON COLLABORATION & AI USE

For Assessment B, students may seek assistance from other people, online resources, or generative AI tools, provided that both of the following conditions are met:

1. **Acknowledgement of assistance:** Students must clearly acknowledge any assistance received in their submission (e.g., in a short "Acknowledgements" section). They should state who or what provided the assistance (e.g., a person, website, or tool) and the nature of the assistance received. If generative AI was used, students must include the prompts used and a brief description of how the generated output was used.

2. **Understanding and explainability:** Students must understand everything they submit and be able to explain the reasoning and implementation, including any parts for which they received assistance.

Please note that failure to properly acknowledge assistance will result in no credit for the relevant submission.

6. TEXT AND REFERENCE MATERIALS

This course has no required textbook. Lecture notes, lab manuals, and additional exercises will be posted on the course website.

7. EMAIL POLICY

Please put **[ISOM3370]** at the beginning of the subject line of your email along with your email subject. Failure to do so may result in a longer response time.

As expected, there will be numerous emails when it is closer to the due dates. If you need any assistance, raise them as early as possible, and/or make good use of the labs and the office hours. Note that **neither the instructor nor the TAs will provide direct answers to the assignments**.

8. LEARNING ENVIRONMENT

We welcome feedback on our teaching throughout the semester. students are encouraged to contact me or the TA any time they have any questions, suggestions, concerns, or would like to ask for advice.

TENTATIVE LECTURE SCHEDULE

** Note both the lecture and lab schedules are tentative and subject to change without notice.*

Date	Topics	Remarks
Sep. 4	Course Introduction	
Sep. 7	Introduction to Hadoop and HDFS	
Sep. 11	Hadoop Distributed File System	
Sep. 14	**Lab**: Introduction to AWS	
Sep. 18	**Lab**: Hadoop Distributed File System	
Sep. 21	MapReduce	
Sep. 25	MapReduce Continued	
Sep. 28	**Lab**: Running Hadoop MapReduce Job	
Oct. 2	MapReduce for Web Search	
Oct. 5	MapReduce for PageRank	
Oct. 9	MapReduce for Data Warehousing	
Oct. 12	Hive: SQL on Hadoop	
Oct. 16	**Lab**: Hive	
Oct. 19	**Class Canceled (Public Holiday)**	
Oct. 23	Midterm Review	
Oct. 26	**Midterm Exam (In-class)**	
Oct. 30	Spark Introduction	
Nov. 2	Spark RDD Programming	
Nov. 6	**Lab**: Spark RDD Programming	
Nov. 9	Spark Programming Continued	
Nov. 13	Large-scale Machine Learning	
Nov. 16	Spark for Machine Learning MLlib	
Nov. 20	Spark for Machine Learning MLlib Continued	
Nov. 23	**Lab**: Spark MLlib	
Nov. 27	Big Data Computing in the Era of AI	
Nov. 30	Final Exam Review	