HKUST
BUSINESS SCHOOL
香港科大商學院

ISOM3370: Big Data Technologies

Spring 2023

Class Meetings	Tu/Th 03:00PM-04:20PM
Prerequisites	(ISOM 3230 or ISOM 3320 or ISOM 3400) and ISOM3360
Instructor	Yi Yang Email: imyiyang@ust.hk **Begin subject: [ISOM3370]…** Office Hours: By appointment
Teaching Assistant	Samuel Lai Email: imsamuel@ust.hk Office Hours: By appointment

1. Course Overview

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. Some basics of programming languages, such as Python, will also be covered to help you understand algorithms and run them on massive datasets.

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. Grading Policy

Your grades will be determined based on class and lab participation, homework assignments, the midterm and final exam, and group project.

Class and Lab Participation	10%
Lab and Homework Assignments	30%
Midterm Exam	25%
Final Exam	35%

Homework Assignment

There will be a total of **3 individual homework assignments**, each comprising questions to be answered and hands-on tasks. Completed assignments must be handed in via Canvas prior to the start of the class on the due date. Assignments will be graded and returned promptly.

Turn in your assignment early if there is any uncertainty about your ability to turn it in on the due date. Assignments up to 24 hours late will have their grade reduced by 25%; assignments up to one week late will have their grade reduced by 50%. After one week, late assignments will receive no credit.

Lab Session

This is primarily a lecture-based course, but lab participation is an essential part of the learning process in the form of active practice. You are NOT going to learn without practicing the big data technologies yourselves. During the lab session, I will expect you to be entirely devoted to the class by following the instructions. You will bring and use your own laptop to the class. For each lab, you need to finish and submit a report, even if you may not finish the lab in class. For the first 3 lab sessions, you are expected to submit a lab report. For the last 2 lab sessions, lab report is in the form of the homework assignment.

Exams

This course will have two exams. The midterm exam will test issues covered in the first half of the course. The final exam will cover the classes in the second half of the course. Review sessions will be scheduled to help you prepare for these examinations.

The midterm exam is tentatively scheduled on **Mar 23 in-class.** The final exam will be held during the final examination period; the date will be announced later in the semester.

Make-up exam policy: https://arr.ust.hk/reg/em/em_std_reg/reg_makeup.html
To quote, "If students wish the University to take into account illness or some other extenuating circumstances that have affected their performance in an examination, or ability to attend an examination, or to complete other assessment activities, they must report the circumstances of the case in writing and provide appropriate documentation to ARR, Academic Registry within one week of the scheduled date of the assessment activity. The Academic Registrar will review the case and make a recommendation to the relevant Dean, the Dean's designate or the Director of IPO."

Academic Integrity

Students at HKUST are expected to observe the Academic Honor Code at all times (see http://acadreg.ust.hk/generalreg.html for more information). Zero tolerance is shown to those who are caught cheating on any quiz or exam. In addition to receiving a zero mark on the quiz or exam involved, the final course grade will appear on your record with an X, to show that the grade resulted from cheating. This X grade will stay with your record until graduation. If you receive another X grade, you will be dismissed from HKUST.

Schedule of Lectures and Labs (subject to change)

Data	Topics	Remarks
Feb 7	Course Introduction	
Feb 9	Introduction to Hadoop and HDFS	
Feb 14	Hadoop Distributed File System	
Feb 16	**Lab**: Introduction to AWS	
Feb 21	**Lab**: Hadoop Distributed File System	
Feb 23	MapReduce	
Feb 28	MapReduce Continued	
Mar 2	**Lab**: Running Hadoop MapReduce Job	
Mar 7	MapReduce for Web Search	
Mar 9	Hive	
Mar 14	Hive Continued	
Mar 16	Midterm Review	
Mar 21	Midterm Q&A session	
Mar 23	Midterm Exam (in-class)	
Mar 28	Spark Introduction	
Mar 30	Spark RDD Programming	
Apr 4	**Lab**: Spark Programming	
Apr 6	[no class] Mid-term Break	
Apr 11	[no class] Mid-term Break	
Apr 13	Spark Programming Continued	
Apr 18	Large-scale Machine Learning	
Apr 20	Spark for Machine Learning MLlib	
Apr 25	Spark for Machine Learning MLlib Continued	
Apr 27	[no class] Public Holiday	
May 2	**Lab**: Spark MLlib	
May 4	Find Similar Items in Massive Data	
May 9	Final Exam Review	