City University of Hong Kong
Course Syllabus

offered by Department of Computer Science
with effect from Semester A 2022/23

Part I Course Overview

Course Title:	Big Data Algorithms and Techniques
Course Code:	CS5488
Course Duration:	1 Semester
Credit Units:	3 Credits
Level:	P5
Medium of Instruction:	English
Medium of Assessment:	English
Prerequisites: *(Course Code and Title)*	CS3402 Database Systems or CS5481 Data Engineering
Precursors: *(Course Code and Title)*	Nil
Equivalent Courses: *(Course Code and Title)*	Nil
Exclusive Courses: *(Course Code and Title)*	Nil

Part II Course Details

1. Abstract

This course is aimed at equipping students with the ability to manage very large data sets (Big Data) using a cluster of commodity machines with the main focus on the Hadoop ecosystem. It has three specific objectives: (1) to familiarize students with software systems and techniques for implementing distributed data-parallel programs, (2) to provide insight into internal mechanisms of large-scale data analytical systems, and (3) to acquaint students with big data solutions deployed in real-world settings. Students will also have the opportunity to analyse and to compare real-world big data solutions in a class project case study.

2. Course Intended Learning Outcomes (CILOs)

(CILOs state what the student is expected to be able to do at the end of the course according to a given standard of performance.)

No.	CILOs	Weighting (if applicable)	Discovery-enriched curriculum related learning outcomes (please tick where appropriate)		
			A1	*A2*	*A3*
1.	Identify data parallelism to be exploited in large-scale data processing problems.		✓	✓	
2.	Implement data parallel algorithms using techniques covered in the course.		✓	✓	
3.	Understand the internal mechanisms of the Hadoop framework.		✓		
4	Design scalable solutions to a real-world problem and sufficiently provide rationalizations to the design decisions.		✓	✓	✓
5	Analyse existing big data solutions deployed in real-world settings through case studies.		✓	✓	✓
		100%			

A1: Attitude
Develop an attitude of discovery/innovation/creativity, as demonstrated by students possessing a strong sense of curiosity, asking questions actively, challenging assumptions or engaging in inquiry together with teachers.

A2: Ability
Develop the ability/skill needed to discover/innovate/create, as demonstrated by students possessing critical thinking skills to assess ideas, acquiring research skills, synthesizing knowledge across disciplines or applying academic knowledge to self-life problems.

A3: *Accomplishments*
Demonstrate accomplishment of discovery/innovation/creativity through producing /constructing creative works/new artefacts, effective solutions to real-life problems or new processes.

3. Teaching and Learning Activities (TLAs)

(TLAs designed to facilitate students' achievement of the CILOs.)

TLA	Brief Description	CILO No.					Hours/week (if applicable)
		1	2	3	4	5	
1. Lecture	Lectures will cover (1) different types of data-parallel problems; (2) the APIs and tools for handling big data; (3) the internal mechanisms for job scheduling, failure handling, and task execution of the Hadoop framework; (4) case studies on real-world big data algorithms and solutions.	✓	✓	✓			2 hours/ week
2. Lab	Laboratory classes will provide the students with the opportunity to (1) familiarize themselves with different big data tools; (2) implement data-parallel algorithms; (3) design experimental studies.		✓	✓	✓		1 hour/ week
3. Class Project	For the class project, the students will work on a use case study on Hadoop-based solutions deployed in real-world settings.				✓	✓	

4. Assessment Tasks/Activities (ATs)

(ATs are designed to assess how well the students achieve the CILOs.)

Assessment Tasks/Activities	CILO No.					Weighting	Remarks
	1	2	3	4	5		
Continuous Assessment: 60%							
Class Project	✓	✓		✓	✓	40%	
Lab Sheets	✓	✓	✓	✓		5%	
Midterm Examination	✓	✓	✓			15%	
Examination^: 40% (duration: 2 hours)							
Final Examination	✓	✓	✓	✓		40%	
						100%	

^ For a student to pass the course, at least 30% of the maximum mark for the examination must be obtained.

5. Assessment Rubrics

(Grading of student achievements is based on student performance in assessment tasks/activities with the following rubrics.)

Applicable to students admitted in Semester A 2022/23 and thereafter

Assessment Task	Criterion	Excellent (A+, A, A-)	Good (B+, B)	Marginal (B-, C+, C)	Failure (F)
1. Group Project	1.1 Ability to identify challenges in various types of data parallel problems 1.2 Ability to critique existing big data solutions	High	Significant	Moderate to Basic	Inadequate
2. Lab Sheets	2.1 Ability to contribute to discussions on principle and concepts of scalable data processing	High	Significant	Moderate to Basic	Inadequate
3. Midterm Exam	3.1 Ability to demonstrate a good understanding of materials covered in the course	High	Significant	Moderate to Basic	Inadequate
4. Final Exam	4.1 Ability to demonstrate a good understanding of materials covered in the course	High	Significant	Moderate to Basic	Inadequate

Applicable to students admitted before Semester A 2022/23

Assessment Task	Criterion	Excellent (A+, A, A-)	Good (B+, B, B-)	Fair (C+, C, C-)	Marginal (D)	Failure (F)
1. Group Project	1.1 Ability to identify challenges in various types of data parallel problems 1.2 Ability to critique existing big data solutions	High	Significant	Moderate	Basic	Inadequate
2. Lab Sheets	2.1 Ability to contribute to discussions on principle and concepts of scalable data processing	High	Significant	Moderate	Basic	Inadequate
3. Midterm Exam	3.1, 4.1 Ability to demonstrate a good understanding of materials covered in the course	High	Significant	Moderate	Basic	Inadequate
4. Final Exam						

Part III Other Information (more details can be provided separately in the teaching plan)

Keyword Syllabus

(An indication of the key topics of the course.)

Big Data, Analytics, MapReduce, Distributed File System, Parallel Processing, Data-parallel Systems, RDBMS, NOSQL, Distributed Indexes, Key-value Stores, Query Languages, Data Manipulation Languages, Consistency, Reliability, Commodity Cluster, Failure Handling, In-memory Processing, Use Case Studies, Emerging Technologies for Big Data Computing (e.g. Hadoop and Spark).

2. Reading List

2.1 Compulsory Readings

(Compulsory readings can include books, book chapters, or journal/magazine articles. There are also collections of e-books, e-journals available from the CityU Library.)

1.	Tom White. *Hadoop: The Definitive Guide.* 4th edition.
2	Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia. *Learning Spark: Lightning-Fast Big Data Analysis.* 1st edition.

2.2 Additional Readings

(Additional references for students to learn to expand their knowledge about the subject.

1.	EMC Education Services. *Data Science and Big Data Analytics*. 1st edition.