City University of Hong Kong
Course Syllabus

offered by Department of Computer Science
with effect from Semester A 2019/20

Part I Course Overview

Course Title:	Data-Intensive Computing
Course Code:	CS4480
Course Duration:	1 semester
Credit Units:	3 credits
Level:	B4
Proposed Area: *(for GE courses only)*	☐ Arts and Humanities ☐ Study of Societies, Social and Business Organisations ☐ Science and Technology
Medium of Instruction:	English
Medium of Assessment:	English
Prerequisites: *(Course Code and Title)*	CS3402 Database Systems and CS3481 Fundamentals of Data Science
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

(A 150-word description about the course)

This course is aimed at equipping students with the ability to compute on large data sets using parallel and distributed programming on multiple computing units. Specifically, the main objective of this course is twofold: to familiarize students with software systems and techniques for designing and implementing parallel and distributed data computing programs; to provide insights into the internal mechanisms of scalable data processing systems. Students will also have the opportunity to work on a real-world data processing problem by implementing scalable data computing solutions using the techniques and software systems covered in this course and to deploy their solutions on multiple computing units.

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 the main characteristics of the parallel and distributed computing solutions to data processing		✓	✓	
2.	Design and implement the parallel and distributed computing algorithms for data processing		✓	✓	
3.	Understand the parallel and distributed computing theory behind scalable data processing		✓		
4.	Design scalable data computing solutions to a real-world data processing problem and sufficiently provide rationalizations to the design decisions.		✓	✓	
5.	Assess the performance of different scalable data processing solutions.		✓	✓	
		100%			

** If weighting is assigned to CILOs, they should add up to 100%.*

[#] Please specify the alignment of CILOs to the Gateway Education Programme Intended Learning outcomes (PILOs) in Section A of Annex.

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.)

Teaching pattern:

Suggested lecture/tutorial mix: 2 hrs. lecture; 1 hr. tutorial.

TLA	Brief Description	CILO No.					Hours/week (if applicable)
		1	2	3	4	5	
1. Lecture	Lectures will cover (1) different types of scalable data processing problems; (2) the parallel and distributed computing techniques for scalable data processing; (3) the parallel and distributed computing theory behind scalable data processing; (4) case studies on real-world big data algorithms and solutions.	✓	✓	✓			2 hours
2. In-Class Discussion	Tutorial classes will provide the students with the opportunity to (1) familiarize themselves with different data processing tools; (2) implement parallel and distributed algorithms for data processing; (3) design scalable data computing solutions.		✓	✓	✓		1 hour
3. Group Project	For the class project, the students will have the opportunity to work on a real-world data processing problem. Each group will be required to propose a scalable data processing solution to a real world problem. Each group will also submit a project report and conduct a project presentation.			✓	✓	✓	

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%							
Group Project	✓	✓		✓	✓	40%	
In-Class Discussion	✓	✓	✓	✓		5%	
Midterm Examination	✓	✓	✓			15%	
Examination^: 40% (duration: 2 hours)							
Final Examination	✓	✓	✓	✓		40%	
** The weightings should add up to 100%.*						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.)

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 computing	High	Significant	Moderate	Basic	Inadequate
	1.2 Ability to design and implement a scalable solution for a real-world data processing problem.	High	Significant	Moderate	Basic	Inadequate
	1.3 Ability to assess computing performance.	High	Significant	Moderate	Basic	Inadequate
2. In-Class Discussion	2.1 Ability to implement parallel and distributed data computing solutions.	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)

1. Keyword Syllabus

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

Big Data, Data Processing, MapReduce Concepts, Distributed Data Storage, Parallel and Distributed Computing Theory, Parallel and Distributed Data Processing, Scalable Data Computing System and Implementation Details, In-Memory Processing, Failure Handling, Emerging Technologies for Data Computing (e.g. Hadoop and Spark), Data-Intensive Computing Applications

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.