CS4480: DATA-INTENSIVE COMPUTING

Effective Term

Semester A 2022/23

Part I Course Overview

Course Title

Data-Intensive Computing

Subject Code

CS - Computer Science

Course Number

4480

Academic Unit

Computer Science (CS)

College/School

College of Engineering (EG)

Course Duration

One Semester

Credit Units

3

Level

B1, B2, B3, B4 - Bachelor's Degree

Medium of Instruction

English

Medium of Assessment

English

Prerequisites

CS3402 Database Systems
AND
(CS3481 Fundamentals of Data Science or
SDSC3002 Data Mining or
SDSC3006 Fundamentals of Machine Learning I)

Precursors

Nil

Equivalent Courses

Nil

Exclusive Courses

Nil

Part II Course Details

Abstract

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.

Course Intended Learning Outcomes (CILOs)

	CILOs	Weighting (if app.)	DEC-A1	DEC-A2	DEC-A3
1	Identify the main characteristics of the parallel and distributed computing solutions to data processing		x	x	
2	Design and implement the parallel and distributed computing algorithms for data processing		x	x	
3	Understand the parallel and distributed computing theory behind scalable data processing		x		
4	Design scalable data computing solutions to a real-world data processing problem and sufficiently provide rationalizations to the design decisions.		x	x	
5	Assess the performance of different scalable data processing solutions.		x	x	

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 real-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.

Teaching and Learning Activities (TLAs)

	TLAs	Brief Description	CILO No.	Hours/week (if applicable)
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.	1, 2, 3	3 hours/week
2	Tutorial	Tutorial classes will provide the students with the lab sheet 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.	2, 3, 4	8 hours / semester
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.	3, 4, 5	After class

Assessment Tasks / Activities (ATs)

	ATs	CILO No.	Weighting (%)	Remarks (e.g. Parameter for GenAI use)
1	Group Project	1, 2, 4, 5	40	
2	Lab Sheets	1, 2, 3, 4	5	
3	Midterm Examination	1, 2, 3	15	

Continuous Assessment (%)

60

Examination (%)

40

Examination Duration (Hours)

2

Additional Information for ATs

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

Assessment Rubrics (AR)

Assessment Task

Group Project

Criterion

1.1 Ability to identify challenges in various types of data computing

Excellent (A+, A, A-)

High

Good (B+, B, B-)

Significant

Fair (C+, C, C-)

Moderate

Marginal (D)

Basic

Failure (F)

Inadequate

Assessment Task

Group Project

Criterion

1.2 Ability to design and implement a scalable solution for a real-world data processing problem.

Excellent (A+, A, A-)

High

Good (B+, B, B-)

Significant

Fair (C+, C, C-)

Moderate

Marginal (D)

Basic

Failure (F)

Inadequate

Assessment Task

Group Project

Criterion

1.3 Ability to assess computing performance.

Excellent (A+, A, A-)

High

Good (B+, B, B-)

Significant

Fair (C+, C, C-)

Moderate

Marginal (D)

Basic

Failure (F)

Inadequate

Assessment Task

Lab Sheets

Criterion

2.1 Ability to implement parallel and distributed data computing solutions.

Excellent (A+, A, A-)

High

Good (B+, B, B-)

Significant

Fair (C+, C, C-)

Moderate

Marginal (D)

Basic

Failure (F)

Inadequate

Assessment Task

Midterm Exam

Criterion

3.1, 4.1 Ability to demonstrate a good understanding of materials covered in the course.

Excellent (A+, A, A-)
High

Good (B+, B, B-)
Significant

Fair (C+, C, C-)
Moderate

Marginal (D)
Basic

Failure (F)
Inadequate

Assessment Task
Final Exam

Criterion
3.1, 4.1 Ability to demonstrate a good understanding of materials covered in the course.

Excellent (A+, A, A-)
High

Good (B+, B, B-)
Significant

Fair (C+, C, C-)
Moderate

Marginal (D)
Basic

Failure (F)
Inadequate

Part III Other Information

Keyword Syllabus

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

Reading List

Compulsory Readings

	Title
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.

Additional Readings

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