CS5488: BIG DATA ALGORITHMS AND TECHNIQUES

Effective Term
Semester B 2024/25

Part I Course Overview

Course Title
Big Data Algorithms and Techniques

Subject Code
CS - Computer Science
Course Number
5488

Academic Unit
Computer Science (CS)

College/School
College of Computing (CC)

Course Duration
One Semester

Credit Units
3

Level
P5, P6 - Postgraduate Degree

Medium of Instruction
English

Medium of Assessment
English

Prerequisites
CS3402 Database Systems or
CS5481 Data Engineering

Precursors
Nil

Equivalent Courses
Nil

Exclusive Courses
Nil

Part II Course Details

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.

Course Intended Learning Outcomes (CILOs)

	CILOs	Weighting (if app.)	DEC-A1	DEC-A2	DEC-A3
1	Identify and explain data parallelism to be exploited in large-scale data processing problems.		x	x	
2	Implement data parallel algorithms using techniques covered in the course.		x	x	
3	Describe and explain the internal mechanisms of the Hadoop framework.		x		
4	Design scalable solutions to a real-world problem and sufficiently provide rationalizations to the design decisions.		x	x	x
5	Analyse existing big data solutions deployed in real-world settings through case studies.		x	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.

Learning and Teaching Activities (LTAs)

	LTAs	Brief Description	CILO No.	Hours/week (if applicable)
1	Lecture	Students will learn (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.	1, 2, 3	2 hours/ week
2	Lab	Students will have the opportunity to (1) familiarize themselves with different big data tools; (2) implement data-parallel algorithms; (3) design experimental studies.	2, 3, 4	1 hour/ week
3	Class Project	Students will work on a use case study on Hadoop-based solutions deployed in real-world settings.	4, 5	

Assessment Tasks / Activities (ATs)

	ATs	CILO No.	Weighting (%)	Remarks (e.g. Parameter for GenAI use)
1	Class Project	1, 2, 4, 5	40	
2	Lab Sheets	1, 2, 3, 4	10	

Continuous Assessment (%)

50

Examination (%)

50

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 (for students admitted before Semester A 2022/23 and in Semester A 2024/25 & thereafter)

Criterion

1.1 Ability to identify challenges in various types of data parallel problems

1.2 Ability to critique existing big data solutions

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 (for students admitted before Semester A 2022/23 and in Semester A 2024/25 & thereafter)

Criterion
2.1 Ability to contribute to discussions on principle and concepts of scalable data processing

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 (for students admitted before Semester A 2022/23 and in Semester A 2024/25 & thereafter)

Criterion
3.1 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
Group Project (for students admitted from Semester A 2022/23 to Summer Term 2024)

Criterion
1.1 Ability to identify challenges in various types of data parallel problems
1.2 Ability to critique existing big data solutions

Excellent
(A+, A, A-) High

Good
(B+, B) Significant

Marginal
(B-, C+, C) Moderate to Basic

Failure
(F) Inadequate

Assessment Task
Lab Sheets (for students admitted from Semester A 2022/23 to Summer Term 2024)

Criterion
2.1 Ability to contribute to discussions on principle and concepts of scalable data processing

Excellent
(A+, A, A-) High

Good
(B+, B) Significant

Marginal
(B-, C+, C) Moderate to Basic

Failure
(F) Inadequate

Assessment Task
Final Exam (for students admitted from Semester A 2022/23 to Summer Term 2024)

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

Excellent

(A+, A, A-) High

Good

(B+, B) Significant

Marginal

(B-, C+, C) Moderate to Basic

Failure

(F) Inadequate

Part III Other Information

Keyword Syllabus

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).

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.