SDSC3001: BIG DATA: THE ARTS AND SCIENCE OF SCALING

Effective Term

Semester A 2023/24

Part I Course Overview

Course Title

Big Data: The Arts and Science of Scaling

Subject Code

SDSC - School of Data Science

Course Number

3001

Academic Unit

School of Data Science (DS)

College/School

School of Data Science (DS)

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 system

Precursors

Nil

Equivalent Courses

Nil

Exclusive Courses

Nil

Part II Course Details

Abstract

This course aims at teaching students how to tame massive data which are intensively used in high-impact industrial applications. Students will learn two mainstream categories of technical solutions for big data, namely algorithmic approaches and systems approaches. For algorithm approaches, some popular stream algorithms such as heavy hitters and sketching algorithms used when we have a limited memory will be introduced. To deal with huge amount of data, the instructor will also teach sampling-based algorithms, such as approximate counting, that tame big data via sampling a representative small collection of data. For the system approaches, the instructor will introduce Spark, one of the most popular big data computing software nowadays, to the students. Topics in Spark include the MapReduce model, Spark RDDs, DataFrames, DataSets, Spark SQL and Spark ML.

Course Intended Learning Outcomes (CILOs)

	CILOs	Weighting (if app.)	DEC-A1	DEC-A2	DEC-A3
1	Understand that the scalability issue lies at the core of making data science practical.	10	x	x	
2	Understand basic stream algorithms and sampling algorithms. Be able to prove the effectiveness of these algorithms.	30	x	x	
3	Implement data processing algorithms using Spark.	30	x	x	x
4	Apply the algorithmic techniques and system techniques in solving scalability problems in real applications.	30		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	Lectures	Learning through teaching is primarily based on lectures. Mini-lectures and small-group exercises will be used to facilitate conceptual understanding and applications of various methods tools and techniques.	1, 2, 3, 4	39 hours/semester

2	Course Project	The team-based projects provide students with the opportunities to familiarize and apply the tools learnt during the lectures through practical problem solving.	3, 4	After class

Assessment Tasks / Activities (ATs)

	ATs	CILO No.	Weighting (%)	Remarks (e.g. Parameter for GenAI use)
1	Group projects	3, 4	30	
2	Assignments	1, 2, 3, 4	40	

Continuous Assessment (%)

70

Examination (%)

30

Examination Duration (Hours)

2

Additional Information for ATs

Note: To pass the course, apart from obtaining a minimum of 40% in the overall mark, a student must also obtain a minimum mark of 30% in both continuous assessment and examination components.

Assessment Rubrics (AR)

Assessment Task

Group projects

Criterion

The project is to evaluate the overall performance and the attitude of the students in understanding, utilizing, applying the methodologies, principles and skills. The teamwork and collaboration is also accessed.

Excellent (A+, A, A-)

High

Good (B+, B, B-)

Significant

Fair (C+, C, C-)

Moderate

Marginal (D)

Basic

Failure (F)

Not even reaching marginal levels

Assessment Task

Assignments

Criterion
Assess students’ understanding of computational methods and common techniques.

Excellent (A+, A, A-)
High

Good (B+, B, B-)
Significant

Fair (C+, C, C-)
Moderate

Marginal (D)
Basic

Failure (F)
Not even reaching marginal levels

Assessment Task
Examination

Criterion
Examination questions are designed to assess student’s level of achievement of the intended learning outcomes, with emphasis placed on understanding and correct application, mostly through clear explanation, and numerical calculation, of the various data processing techniques.

Excellent (A+, A, A-)
High

Good (B+, B, B-)
Significant

Fair (C+, C, C-)
Moderate

Marginal (D)
Basic

Failure (F)
Not even reaching marginal levels

Additional Information for AR
The midterm and tutorial exercises will be numerically-marked, while examination will be numerically-marked and grades-awarded accordingly.

Part III Other Information

Keyword Syllabus
Algorithmic approaches:
Stream algorithms: heavy hitters, distinct element counting, sketching algorithms, matrix sketching, graph sketching

Sampling algorithms: approximate counting, Chernoff bounds, Monte Carlo simulations, Markov Chain Monte Carlo, graph sampling
System approaches:
Spark basics: MapReduce, RDD, DataFrames, DataSets
Advanced features of Spark: Spark SQL, Spark Stream, Spark ML

Reading List

Compulsory Readings

	Title
1	Lecture notes

Additional Readings

	Title
1	Nil