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)

<table>
<thead>
<tr>
<th>CILOs</th>
<th>Weighting (if app.)</th>
<th>DEC-A1</th>
<th>DEC-A2</th>
<th>DEC-A3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify the main characteristics of the parallel and distributed computing solutions to data processing</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Design and implement the parallel and distributed computing algorithms for data processing</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Understand the parallel and distributed computing theory behind scalable data processing</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Design scalable data computing solutions to a real-world data processing problem and sufficiently provide rationalizations to the design decisions.</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Assess the performance of different scalable data processing solutions.</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>TLAs</th>
<th>Brief Description</th>
<th>CILO No.</th>
<th>Hours/week (if applicable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Lecture</td>
<td>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.</td>
<td>1, 2, 3</td>
<td>3 hours/week</td>
</tr>
<tr>
<td>2 Tutorial</td>
<td>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.</td>
<td>2, 3, 4</td>
<td>8 hours / semester</td>
</tr>
<tr>
<td>3 Group Project</td>
<td>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.</td>
<td>3, 4, 5</td>
<td>After class</td>
</tr>
</tbody>
</table>

### Assessment Tasks / Activities (ATs)

<table>
<thead>
<tr>
<th>ATs</th>
<th>CILO No.</th>
<th>Weighting (%)</th>
<th>Remarks (e.g. Parameter for GenAI use)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Group Project</td>
<td>1, 2, 4, 5</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>2 Lab Sheets</td>
<td>1, 2, 3, 4</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3 Midterm Examination</td>
<td>1, 2, 3</td>
<td>15</td>
<td></td>
</tr>
</tbody>
</table>

**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

<table>
<thead>
<tr>
<th>Title</th>
</tr>
</thead>
</table>
Additional Readings

<table>
<thead>
<tr>
<th>Title</th>
</tr>
</thead>
</table>