CS5481: DATA ENGINEERING

Effective Term
Semester B 2024/25

Part I Course Overview

Course Title
Data Engineering

Subject Code
CS - Computer Science
Course Number
5481

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
CS2312 Problem Solving and Programming or Equivalent computer programming courses

Precursors
Nil

Equivalent Courses
Nil

Exclusive Courses
Nil

Part II Course Details

Abstract
This course talks about the entire life cycle of data engineering process. First, it aims to enhance students' understanding of the whole data engineering process, including data acquiring, data cleaning and processing, data storage, data

management, and data applications. Second, it describes a number of advanced data engineering techniques throughout the process, including web crawler, database systems, data visualization, data processing algorithms, and data application examples. Finally, it discusses important issues about data management, such as data quality, security, privacy, and federated processing. All these are important in supporting sophisticated data engineering applications.

Course Intended Learning Outcomes (CILOs)

	CILOs	Weighting (if app.)	DEC-A1	DEC-A2	DEC-A3
1	Describe the lifecycle of data engineering process, such as data acquisition, data cleaning, data processing, data storage, data management, and data applications.	15		x	
2	Apply data engineering techniques to gather and process data.	35	x	x	x
3	Describe issues specific to data management, such as data quality, security, and privacy.	15		x	
4	Apply data engineering techniques for data application examples, such as recommendation, anomaly detection, and information retrieval.	35	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	Lectures	Students will engage in lectures explaining the concepts principles, and techniques in detail.	1, 2, 3, 4	2 hrs/wk
2	Tutorials	Students will apply knowledge learnt in the lectures to present and explain her/his solutions to given problems.	1, 2, 3, 4	1 hr/wk
3	Individual assignments	Students will independently work on two assignments. Each assignment contains questions designed to help students apply techniques/ algorithms to solve practical problems.	1, 2, 3, 4	

4	Group project	Students will create a new system design and implement appropriate data engineering applications. The students will apply the principles they have learnt from the course for their design.	1, 2, 3, 4	

Assessment Tasks / Activities (ATs)

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

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

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

Criterion

Ability to implement and assess data engineering techniques for data acquisition, data cleaning, data processing, data storage, data management, and data applications.

Excellent

(A+, A, A-) High

Good

(B+, B, B-) Significant

Fair

(C+, C, C-) Moderate

Marginal

(D) Basic

Failure

(F) Not even reaching marginal level

Assessment Task

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

Criterion

Ability and creativity in designing and implementing appropriate data engineering algorithms and techniques for innovative data engineering applications. Apply them with appropriate modification or design new solutions for different applications and evaluate their performances.

Excellent

(A+, A, A-) High

Good

(B+, B, B-) Significant

Fair

(C+, C, C-) Moderate

Marginal

(D) Basic

Failure

(F) Not even reaching marginal level

Assessment Task

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

Criterion

Ability to understand and apply data engineering techniques for data acquisition, data cleaning, data processing, data storage, data management, and data applications. Ability to analyse the performance of different data engineering 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 level

Assessment Task

Assignments (for students admitted from Semester A 2022/23 to Summer Term 2024)

Criterion

Ability to implement and assess data engineering techniques for data acquisition, data cleaning, data processing, data storage, data management, and data applications.

Excellent

(A+, A, A-) High

Good

(B+, B) Significant

Marginal

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

Failure

(F) Not even reaching marginal level

Assessment Task

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

Criterion

Ability and creativity in designing and implementing appropriate data engineering algorithms and techniques for innovative data engineering applications. Apply them with appropriate modification or design new solutions for different applications and evaluate their performances.

Excellent

(A+, A, A-) High

Good

(B+, B) Significant

Marginal

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

Failure

(F) Not even reaching marginal level

Assessment Task

Examination (for students admitted from Semester A 2022/23 to Summer Term 2024)

Criterion

Ability to understand and apply data engineering techniques for data acquisition, data cleaning, data processing, data storage, data management, and data applications. Ability to analyse the performance of different data engineering techniques.

Excellent

(A+, A, A-) High

Good

(B+, B) Significant

Marginal

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

Failure

(F) Not even reaching marginal level

Part III Other Information

Keyword Syllabus

Topics:
1. Data eco-system
Data sources and data format. Structured and unstructured data. Data engineering flow and data eco-system overview.
2. Data acquisition and data cleaning
Data types and acquisition methods. Web crawling operations and strategies. Politeness policy. Duplicate detection. Denoising. Outlier removing. Missing data.
3. Data preparation for analysis and storage
Data analysis technique selection and data preparation. Data sparsity. Data imbalance. Data storage technique selection. Structured and unstructured data preparation for storage.
4. Data visualization
Visualization analysis. Multidimensional data. Hierarchical data visualization. Graph data visualization. Temporal data visualization.
5. Data indexing
Dense/sparse primary/non-primary index. B+ tree. Hashing.
6. Data querying
Structured and unstructured queries. Querying languages. Querying algorithms. Querying optimizations. Personalization and contextualization.
7. Data applications
Recommendations. Information retrieval. Anomaly detection. Social network analysis.
8. Data management
Data quality. Data security. Data privacy. Federated learning.

Reading List

Compulsory Readings

	Title
1	Silberschatz A., Korth H.F. and Sudarshan S. Database System Concepts. 6th Ed. McGraw Hill (2011) (latest edition)
2	Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

Additional Readings

	Title
1	Nil