City University of Hong Kong
Course Syllabus

offered by the Department of Computer Science
with effect from Semester A 2022/23

Part I Course Overview

Course Title:	Data Engineering
Course Code:	CS5481
Course Duration:	One semester
Credit Units:	3 credits
Level:	P5
Medium of Instruction:	English
Medium of Assessment:	English
Prerequisites: *(Course Code and Title)*	CS2312 - Problem Solving and Programming or Equivalent computer programming courses
Precursors: *(Course Code and Title)*	Nil
Equivalent Courses: *(Course Code and Title)*	Nil
Exclusive Courses: *(Course Code and Title)*	Nil

Part II Course Details

1. 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.

2. Course Intended Learning Outcomes (CILOs)

(CILOs state what the student is expected to be able to do at the end of the course according to a given standard of performance.)

No.	CILOs	Weighting (if applicable)	Discovery-enriched curriculum related learning outcomes (please tick where appropriate)		
			A1	*A2*	*A3*
1.	Develop in-depth understanding of the lifecycle of data engineering process, such as data acquisition, data cleaning, data processing, data storage, data management, and data applications.	15%		✓	
2.	Apply data engineering techniques to gather and process data.	35%	✓	✓	✓
3.	Describe issues specific to data management, such as data quality, security, and privacy.	15%		✓	
4.	Apply data engineering techniques for data application examples, such as recommendation, anomaly detection, and information retrieval.	35%	✓	✓	✓
		100%			

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 self-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.

3. Teaching and Learning Activities (TLAs)

(TLAs designed to facilitate students' achievement of the CILOs.)

TLA	Brief Description	CILO No.				Hours/week (if applicable)
		1	2	3	4	
Lectures	Explain the concepts, principles, and techniques in detail.	✓	✓	✓	✓	2 hrs/wk
Tutorials	Require students to apply knowledge learnt in the lectures to present and explain her/his solutions to given problems.	✓	✓	✓	✓	1 hr/wk
Individual assignments	Require students to independently work on two assignments. Each assignment contains questions designed to help students apply techniques / algorithms to solve practical problems.	✓	✓	✓	✓	
Group project	Require students to 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.	✓	✓	✓	✓	

4. Assessment Tasks/Activities (ATs)

(ATs are designed to assess how well the students achieve the CILOs.)

Assessment Tasks/Activities	CILO No.						Weighting	Remarks
	1	2	3	4				
Continuous Assessment: 60%								
Assignments	✓	✓	✓	✓			30%	
Group project	✓	✓	✓	✓			30%	
Examination^: 40% (duration: 2 hours)								
							100%	

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

5. Assessment Rubrics

(Grading of student achievements is based on student performance in assessment tasks/activities with the following rubrics.)

Applicable to students admitted in Semester A 2022/23 and thereafter

Assessment Task	Criterion	Excellent (A+, A, A-)	Good (B+, B)	Marginal (B-, C+, C)	Failure (F)
Assignments	Ability to implement and assess data engineering techniques for data acquisition, data cleaning, data processing, data storage, data management, and data applications.	High	Significant	Moderate to Basic	Not even reaching marginal level
Group project	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.	High	Significant	Moderate to Basic	Not even reaching marginal level
Examination	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.	High	Significant	Moderate to Basic	Not even reaching marginal level

Applicable to students admitted before Semester A 2022/23

Assessment Task	Criterion	Excellent (A+, A, A-)	Good (B+, B, B-)	Fair (C+, C, C-)	Marginal (D)	Failure (F)
Assignments	Ability to implement and assess data engineering techniques for data acquisition, data cleaning, data processing, data storage, data management, and data applications.	High	Significant	Moderate	Basic	Not even reaching marginal level
Group project	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.	High	Significant	Moderate	Basic	Not even reaching marginal level
Examination	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.	High	Significant	Moderate	Basic	Not even reaching marginal level

Part III Other Information (more details can be provided separately in the teaching plan)

1. Keyword Syllabus

(An indication of the key topics of the course.)

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.

2. Reading List

2.1 Compulsory Readings

(Compulsory readings can include books, book chapters, or journal/magazine articles. There are also collections of e-books, e-journals available from the CityU Library.)

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.*

2.2 Additional Readings

(Additional references for students to learn to expand their knowledge about the subject.)