The Hong Kong University of Science and Technology

UG Course Syllabus (Spring 2025-26)

[Course Title] Cloud Computing and Big Data Systems

[Course Code] COMP4651

[No. of Credits] 3

[Any pre-/co-requisites] COMP2011 OR COMP2012H

Name: Wei Wang

Email: weiwa@cse.ust.hk

Course Description

Big data systems, including Cloud Computing and parallel data processing frameworks, emerge as enabling technologies in managing and mining the massive amount of data across hundreds or even thousands of commodity servers in datacentres. This course exposes students to both the theory and hands-on experience of this new technology. By walking through a number of hands-on labs and assignments, students are expected to gain first-hand experience programming on real world clusters in cloud.

List of Topics:

- Basic concepts of Cloud Computing and production Cloud services
- Virtualization: virtual machine and container
- Distributed Storage Systems: GFS, HDFS, and BigTable
- MapReduce: the de facto datacentre-scale programming abstraction and its open source implementation of Hadoop
- Big data processing: predictive analytics, descriptive analytics, graph analytics, text analytics. etc.
- Spark: a new generation parallel processing framework and its infrastructure, programming model, cluster deployment, tuning and debugging
- ML Systems for distributed training and inference
- The state-of-the-art research topics in Cloud systems, including workload management, resource allocation and scheduling.

Intended Learning Outcomes (ILOs)

By the end of this course, students should be able to:

ILO-1 : Describe the motivation, objectives, and architecture of cloud computing and big data systems.
ILO-2 : Understand the use of a production cloud computing platform.
ILO-3 : Understand the general architecture and the use of Hadoop Distributed File System (HDFS).
ILO-4 : Understand the general programming model of MapReduce and the use of Hadoop.
ILO-5 : Understand Resilient Distributed Dataset (RDD) and the use of Spark programming model based on RDD.
ILO-6 : Describe the major architecture difference between Hadoop and Spark.
ILO-7 : Write a MapReduce/Spark program with tens to hundreds of lines of code to solve common data

analytics problems at scale.
ILO-8 : Use software tools to develop and debug a program written in Hadoop and Spark.

Assessment and Grading

This course will be assessed using criterion-referencing and grades will not be assigned using a curve. Detailed rubrics for each assignment are provided below, outlining the criteria used for evaluation.

Assessment Task	Contribution to Overall Course grade (%)
In-class Quizzes and Participation	10%
Assignments and Project	45%
Final examination	45%

Assessments:

[List specific assessed tasks, exams, quizzes, their weightage, and due dates; perhaps, add a summary table as below, to precede the details for each assessment.]

Assessment Task	Contribution to Overall Course grade (%)	Due date
Coding Assignments*	30%	Every 2-3 weeks in the semester
Group Project	30%	18/05/2026
Final examination	40%	28/05/2026

* Assessment marks for individual assessed tasks will be released within two weeks of the due date.

Mapping of Course ILOs to Assessment Tasks

[add to/delete table as appropriate]

Assessed Task	Mapped ILOs	Explanation
Assignment-1	ILO-1, ILO-2	This task assesses students' ability to explain and apply cloud computing concepts to solve practical problems (ILO-1), evaluate their performance in commercial cloud platforms (ILO-2).
Assignment-2	ILO-3, ILO-4, ILO-7, ILO-8	This task assesses students' ability to deploy Hadoop Distributed File System (HDFS) and MapReduce frameworks in the cloud environment (ILO-3), program with MapReduce to solve big data counting problems (ILO-4 and ILO-7), and compare the performance of different implementations (ILO-8).
Assignment-3	ILO-5, ILO-6, ILO-7, ILO-8	This task assesses student's ability to use Spark RDD programming to solve real-world data analytics problems

		(ILO-5 and ILO-7), compare their performance with MapReduce implementations (ILO-6), and use software tools to develop and debug Spark programs (ILO-8).
Assignment-4	ILO-6, ILO-7, ILO-8	This task assesses student's ability to use Spark DataFrame programming to solve real-world data analytics problems (ILO-7), compare their performance with MapReduce implementations (ILO-6), and use software tools to develop and debug Spark programs (ILO-8).
Group Project	ILO-1, ILO-2, ILO-3, ILO-4, ILO-5, ILO-6, ILO-7, ILO-8	The course project is a term-long, open-ended team project involving 2-4 students. The project requires students to work on real-world cloud computing and data analytics problems and develop a full-stack system solution, from application development to algorithm design and to system deployment. Students need to submit a project report explaining the problem, the motivation, the solution, and evaluation results. This is a comprehensive hands-on project where all ILOs are examined.
Final Exam	ILO-1, ILO-2, ILO-3, ILO-4, ILO-5, ILO-6, ILO-7, ILO-8	The final exam thoroughly examines the students ability in understanding the concepts, motivations, and system designs and implementations of modern cloud computing and big data analytics technologies, including HDFS, Hadoop MapReduce, Spark, and other advanced analytics frameworks. The exam problems will cover all ILOs.

Grading Rubrics

Detailed rubrics for each assignment and course project will be provided. These rubrics clearly outline the criteria used for evaluation. Students can refer to these rubrics to understand how their work will be assessed.

Final Grade Descriptors:

[As appropriate to the course and aligned with university standards]

Grades	Short Description	Elaboration on subject grading description
A	Excellent Performance	Demonstrates a comprehensive grasp of subject matter, expertise in problem-solving, and significant creativity in thinking. Exhibits a high capacity for scholarship and collaboration, going beyond core requirements to achieve learning goals.

B	Good Performance	Shows good knowledge and understanding of the main subject matter, competence in problem-solving, and the ability to analyze and evaluate issues. Displays high motivation to learn and the ability to work effectively with others.
C	Satisfactory Performance	Possesses adequate knowledge of core subject matter, competence in dealing with familiar problems, and some capacity for analysis and critical thinking. Shows persistence and effort to achieve broadly defined learning goals.
D	Marginal Pass	Has threshold knowledge of core subject matter, potential to achieve key professional skills, and the ability to make basic judgments. Benefits from the course and has the potential to develop in the discipline.
F	Fail	Demonstrates insufficient understanding of the subject matter and lacks the necessary problem-solving skills. Shows limited ability to think critically or analytically and exhibits minimal effort towards achieving learning goals. Does not meet the threshold requirements for professional practice or development in the discipline.

Course AI Policy

[State the course policy on the use of generative artificial intelligence tools to complete assessment tasks.]

Students can use generative AI tools to polish the writing of the project report, debug programming assignments, and sketch out a skeleton code for their course project. The core code, however, must be hand-written. Students need to clearly state which part of the programs and reports are refined using generative AI tools and justify the necessity and benefits of doing this.

Communication and Feedback

Assessment marks for individual assessed tasks will be communicated via Canvas within two weeks of submission. Feedback on assignments will include strengths and areas for improvement. Students who have further questions about the feedback including marks should consult the instructor or the TAs within five working days after the feedback is received.

Resubmission Policy

Students can ask for an extension of up to 5 days when presented with unforeseeable challenges, e.g., accidents or healthy issues. Students must send in the extension requests at least one day before the submission deadline. Resubmission is not allowed.

Required Texts and Materials

Since Cloud computing and big data systems are emerging technologies under heavy development, there is no official textbook. The followings books are good references to learn Hadoop and Spark programming:

- T. White, "Hadoop: The Definitive Guide Links to an external site.," 4th Eds, O'Reilly, 2015.
- 2. B. Chambers and M. Zaharia, "Spark: The Definitive Guide -- Big Data Processing Made Simple Links to an external site.," O'Reilly, 2018.

Academic Integrity

Students are expected to adhere to the university's academic integrity policy. Students are expected to uphold HKUST's Academic Honor Code and to maintain the highest standards of academic integrity. The University has zero tolerance of academic misconduct. Please refer to Academic Integrity | HKUST – Academic Registry for the University's definition of plagiarism and ways to avoid cheating and plagiarism.

Additional Resources

In addition to the reference books, some course materials come from seminal papers published in recent years' top conferences, which will be released as the course develops.