The Hong Kong University of Science and Technology

UG Course Syllabus

[Course Title] Cloud Computing and Big Data Systems

[Course Code] COMP4651

[No. of Credits] 3-credit

[Any pre-/co-requisites]

Name: [Instructor(s) Name] Dr. WANG Wei

Email: [Your Email Address] 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 datacenters. This course exposes students to both the theory and hands-on experience of this new technology.

The course covers the following topics:

. Basic concepts of Cloud Computing and production Cloud services;

. MapReduce and Spark -- the de facto datacenter-scale programming abstraction and parallel processing frameworks -- and their infrastructure, programming model, cluster deployment, tuning and debugging, as well as a number of specialized data processing systems built on top of them.

. Serverless computing -- the next phase of cloud computing -- and its current implementation, Function-as-a-Service (FaaS)

. AI cloud, including the infrastructure and parallel computing frameworks

. Introduction to the state-of-the-art research topics in Cloud systems, including workload management, networking, resource allocation and scheduling.

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.

The course will cover the fundamentals of cloud computing and big data systems. Students are expected to:

. Describe the motivation, objectives, and architecture of cloud computing and big data systems.

. Understand the use of production cloud computing platform.

. Understand the general architecture and the use of Hadoop Distributed File System (HDFS).

. Understand the general programming model of MapReduce and the use of Hadoop.

. Understand Resilient Distributed Dataset (RDD) and the use of Spark programming model based on RDD.

. Describe the major architecture difference between Hadoop and Spark.

. Write a MapReduce/Spark program with tens to hundreds lines of code to solve common data analytics problems.

. Use software tools to develop and debug a program written in Hadoop and Spark.

. Understand the concept of serverless computing.

. Understand the infrastructure and software stack of AI cloud.

The syllabus of the course is as follows:

W1: Intro & Logistics

W2: Cloud fundamentals

W3: Service models & cloud challenges

W4: Virtualization

W5: Storage system: GFS & HDFS

W6-7: MapReduce & Hadoop

W8: MapReduce algorithm design

W9: From MapReduce to Spark

W10: Spark programming

W11: Spark DataFrame

W12: AI Cloud

W13: Serverless computing & advanced topics

Assessments:

[List specific assessed tasks, exams, quizzes, their weightage]

Assessment Task	Contribution to Overall Course grade (%)
Programming assignment x 4	30%
Open-ended course project	25%
Final Exam	45%
Total	100%

Required Texts and Materials

N/A

[Optional] Additional Resources

Open-ended course project