Course Code	Course Title
COMP 4651	**Cloud Computing and Big Data Systems**

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;
. Virtualization -- virtual machine and container;
. MapReduce -- the de facto datacenter-scale programming abstraction -- and its open source implementation of Hadoop;
. Apache Spark -- a new generation parallel processing framework -- and its infrastructure, programming model, cluster deployment, tuning and debugging, as well as a number of specialized data processing systems built on top of Spark;
. Introduction to the state-of-the-art research topics in Cloud systems, including workload management, 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.

- Prerequisites

The required background of this course includes

. Object-oriented programming (COMP2011 or equivalent)
. Data structures (COMP2012H or equivalent)
. Comfortable with Python/Java programming
. Comfortable with Unix/Linux

List of Topics

- Tentative Schedule

Week Topic
1 Logistics, Cloud concept and characteristics
2 Cloud fundamentals and Service models
3 Virtualization
4 Cloud Storage Systems
5 MapReduce
6 Hadoop
7 MapReduce Algorithm Design

8 RDD and Spark
9 Spark Programming I
10 Spark Programming II
11 Graph Analytics
12 Container Orchestration & Kubernetes
13 Serverless Computing

Textbook and References

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:

1. 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.

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.

Grading Scheme

Homework and Lab	30%
Course project	20%
Final Exam	50%
Total	100%

Course Objectives and Learning Outcomes

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.

Assessment Rubrics

N/A