MIE1628: Big Data Science

Prerequisites:

APS1070, MIE1624H, ECE1513H, CSC2515 (or equivalent) are strongly recommended but not required.

Given the wide range of programming languages deployed in data analytics, students will use Python as the main programming language to implement assignments in this course.

Course description:

This course covers Big Data fundamentals including an overview of Hadoop MapReduce and Spark. Covers Cloud fundamentals and Big Data Analytics on Cloud-based platforms including an introduction to a specific Cloud platform such as Microsoft Azure, Amazon Web Services, or Google Cloud Platform along with common practices for this platform. Covers Cloud technologies to store and process structured, unstructured and semi-structured data. Covers Cloud-based implementation of Real-time Analytics and Machine Learning.

Grading: Assignment/Exam	**Weight (%)**	**Due Date / Time**
Assignment 1	10	June 6 @ 24:00
Assignment 2	10	June 20 @ 24:00
Midterm	25	Week of June 21
Assignment 3	10	July 4 @ 24:00
Assignment 4	10	July 18 @ 24:00
Assignment 5	10	Aug 1 @ 24:00
Final Exam	25	Week of Aug 2

Assignment submissions will be online through *Github/Quercus*. It is the student's responsibility to verify that the assignments are submitted. Assignments submitted up to 48h late will be given a 20% penalty. Assignments that are submitted after 48h late will incur a mark of zero.

Academic honesty:
Do not submit code that you have not written yourself. Students suspected of plagiarism on a project, midterm or exam will be referred to the department for formal discipline for breaches of the Student Code of Conduct.

Student responsibilities:
It is the student's responsibility to attend lectures and ensure assignments are submitted on time.

Preliminary schedule of lecture topics:

No.	**Week**	**Lecture**	**Assignment**
1	May 10	Course Overview, Hadoop Framework	Self-Study
2	May 17	Hadoop in Detail	Assignment 1 (MapReduce)
3	May 24	Spark Framework	Assignment 1 (MapReduce)
4	May 31	Spark in Detail/Databricks	Assignment 2 (Spark)
5	June 7	Azure Cloud Fundamentals	Assignment 2 (Spark)
6	June 14	No Class	Assignment 3 (Cloud Fundamentals)
7	**June 21**	**Mid Term-In Lecture**	Self-Study
8	June 28	Azure Big Data Platform Overview and ETL process	Assignment 3 (Cloud Fundamentals)
9	July 5	Data warehousing in cloud	Assignment 4 (Machine Learning)
10	July 12	Azure SQL Database and Cosmos DB	Assignment 4 (Machine Learning)
11	July 19	Machine Learning/ Real-Stream Analytics in cloud	Assignment 5 (Real-Stream Analytics)
12	July 26	Revision using Big Data Architecture (End to End Use Case)	Assignment 5 (Real-Stream Analytics)
13	**Aug 2**	**Final Exam-In Lecture**	Self-Study

Assignments:
Assignment 1: Based on Kmeans Clustering using MapReduce
Assignment 2: Based on Recommender system using Spark
Assignment 3: Based on Cloud Data Platform
Assignment 4: EDA/Fraud Detection using Machine Learning in cloud
Assignment 5: Working with sensor data using Real Stream Analytics in cloud