MIE1628: Cloud-based Data Analytics

Course Overview:
MIE1628 is designed to equip students with fundamental concepts and practical skills in Cloud-based Data Analytics. Emphasizing hands-on experience, the course covers key aspects of Big Data technologies, including cloud fundamentals, real-time analytics, and machine learning on cloud platforms.

Prerequisites:
While not required, the following courses are strongly recommended:

- APS1070
- MIE1624H
- ECE1513H
- CSC2515

Students will primarily use **Python** and supplement their learning with **Java** for assignments.

Course Description:

This course provides an in-depth exploration of Big Data fundamentals, including an overview of technologies such as **Hadoop MapReduce** and **Spark**. Key topics will include Cloud fundamentals, Big Data Analytics on platforms like Microsoft Azure, Amazon Web Services, or Google Cloud Platform, along with common practices and technologies for storing and processing structured, unstructured, and semi-structured data. Students will engage with cloud-based implementations of real-time analytics and machine learning.

Key Topics

1. **Hadoop Framework:**
    - Overview of Hadoop architecture and components such as HDFS and MapReduce.
    - Practical implementation of MapReduce jobs for processing large datasets.
    - Understanding the scalability and fault-tolerance of the Hadoop ecosystem.
2. **Spark Framework:**
    - Introduction to Apache Spark and its advantages over Hadoop.
    - Concepts of RDDs (Resilient Distributed Datasets) and DataFrames.
    - Implementing data processing tasks using Spark SQL and Spark Streaming.
3. **Cloud Fundamentals:**
    - Understanding cloud computing models: IaaS, PaaS, and SaaS.
    - Exploration of leading cloud service providers like AWS, Azure, and GCP.
    - Overview of cloud storage options and best practices for data management.
4. **Big Data Analytics on Cloud Platforms:**
    - Techniques for analyzing large datasets using cloud-based tools and services.
    - Introduction to cloud-based machine learning services and their applications.
    - Integrating various data sources for comprehensive analysis using cloud technologies.
5. **Real-Time Analytics:**
    - Concepts and frameworks for implementing real-time data processing.
    - Utilizing technologies like Stream Analytics.
    - Case studies highlighting real-time analytics in various business scenarios.
6. **Machine Learning in the Cloud:**
    - Introduction to cloud-based machine learning tools for data analytics.

- Model training, evaluation, and deployment in cloud environments.
- Exploring ethical considerations and responsible AI practices.

Academic Integrity:

Maintaining academic integrity is crucial in this course. Students are expected to submit original work and are prohibited from sharing or using someone else's code. Any cases of suspected plagiarism will be reported in accordance with the University of Toronto's Student Code of Conduct. Breaches may result in formal disciplinary actions.

Grading Scheme:

Assignment/Exam	**Weight (%)**	**Due Date / Time**
Assignment 1	10	Jun 05 @ 24:00
Assignment 2	10	Jun 19 @ 24:00
Midterm	15	Jun 28 @ 18:00
Assignment 3	10	Jul 03 @ 24:00
Assignment 4	10	Jul 17 @ 24:00
Assignment 5	15	Aug 03 @ 24:00
Final Exam	30	Aug 09 @ 18:00

Late Submission Policy:

- Assignments submitted up to 48 hours late will incur a **20% penalty**.
- Submissions more than 48 hours late will receive a **zero**.

Assignments Overview:

- **Assignment 1:** Implementing K-means Clustering using MapReduce.
- **Assignment 2:** Developing a Recommender System using Spark.
- **Assignment 3:** Exploring the Cloud Data Platform.
- **Assignment 4:** Data Orchestration and SQL in the cloud.
- **Assignment 5:** Analyzing data using Real-Time Analytics and Machine Learning in the cloud.

Preliminary schedule of lecture topics:

No.	**Week**	**Lecture**	**Assignment**
1	May 12	Course Overview, Hadoop Framework	Self-Study
2	May 19	Hadoop in Detail	Assignment 1 (MapReduce)
3	May 26	Spark Framework	Assignment 1 (MapReduce)
4	Jun 02	Spark in Detail/Databricks	Assignment 2 (Spark)
5	Jun 09	Azure Cloud Fundamentals	Assignment 2 (Spark)
6	Jun 16	No class – Reading week	Assignment 3 (Cloud Fundamentals

7	**Jun 23**	**Mid Term**	Self-Study
8	Jun 30	Azure Big Data Platform Overview and ETL process	Assignment 3 (Cloud Fundamentals)
9	Jul 07	Data warehousing in cloud	Assignment 4 (Data Orchestration)
10	Jul 14	Azure SQL Database and Cosmos DB	Assignment 4 (Data Orchestration)
11	Jul 21	Machine Learning/ Real-Stream Analytics in cloud	Assignment 5 (Machine Learning)
12	Jul 28	Revision using Big Data Architecture (End to End Use Case)	Assignment 5 (Machine Learning)
13	**Aug 04**	**Final Exam**	Self-Study

Recommended Readings:

1. Big Data: Principles and Best Practices of Scalable Real-Time Data Systems by Nathan Marz and James Goebel, Link to Book
2. Hadoop: The Definitive Guide by Tom White
3. Spark: The Definitive Guide by Bill Chambers and Matei Zaharia
4. Microsoft Azure Essentials: Fundamentals of Azure by Michael S. McKeown
5. Microsoft Azure Documentation, Link to Azure Docs
6. Learning Azure with Microsoft Learn, Link to Microsoft Learn

---

Final Notes:

In this course, students will not only gain practical skills in cloud-based data analytics but also engage in critical discussions surrounding the ethical implications of data use and the technologies we develop. As we progress through the syllabus, I encourage students to share their insights and foster a collaborative learning environment where diverse perspectives are valued.