Winter 2023 CIS*6180

Welcome to CIS*6180. This course covers big data analytics.

Instructor:	Sohail Habib
Email:	shabib03@uoguelph.ca
Office location:	J.D. Maclachlan Building, Room 213
Office hours:	Mon 12:30pm-2:00pm
Lectures:	Mon 2:30pm - 5:20pm (Guelph, MCKN 224 (LEC))

Required Text

- None

Other Resources

- Balusamy, Balamurugan, Seifedine Kadry, and Amir H. Gandomi. *Big Data: Concepts, Technology, and Architecture*. John Wiley & Sons, 2021.
- Marr, Bernard. *Big data in practice: how 45 successful companies used big data analytics to deliver extraordinary results*. John Wiley & Sons, 2016.
- Chambers, Bill, and Matei Zaharia. *Spark: The definitive guide: Big data processing made simple*. " O'Reilly Media, Inc.", 2018.
- Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. Data Mining: Practical Machine Learning Tools and Techniques. Fourth Edition. Morgan Kaufmann, 2016.
- Sebastian Raschka and Vahid Mirjalili. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow. Second Edition. Packt Publishing, 2017.

Suggested Readings

- Divyakant Agrawal, Sudipto Das, and Amr El Abbadi. Data Management in the Cloud: Challenges and Opportunities. Synthesis Lectures on Data Management, Morgan and Claypool, 2012.
- M. Khan, X. Wu, X. Xu and W. Dou, "Big data challenges and opportunities in the hype of Industry 4.0," 2017 IEEE International Conference on Communications (ICC), 2017, pp. 1-6, doi: 10.1109/ICC.2017.7996801.
- Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S. Rellermeyer. 2020. A Survey on Distributed Machine Learning. ACM Comput. Surv. 53, 2, Article 30 (March 2021), 33 pages. https://doi-org.subzero.lib.uoguelph.ca/10.1145/3377454
- David J. DeWitt and Jim Gray. Parallel Database Systems: The Future of High Performance Database Systems. *Communications of the ACM 35(6), 1992*.
- Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The Google File System. *SOSP 2003*.

Outline

- Introduction to Big Data
- Big Data Characteristics and Challenges
- Big Data Life Cycle
- Scalable Computing
- Big Data Modeling and Management
- Big Data Integration and Processing
- Data Exploration
- Machine Learning with Big Data

Prerequisites

Python, Machine Learning , Statistics

Grading Policy

- Lab Assignment and Follow-up Quizzes (50%)
- Case Study/Paper Presentation (10%)
- Case Study/Paper Presentation Peer Review (5%)
- Final Project (35%)
    - Proposal 5%
    - Presentation 5%
    - Deliverables 25%

Reappraisal Policy

If you have an assignment that you would like to have reappraised, please follow the instructions given on CourseLink to submit your request. If you have an exam that you would like to have reappraised, please provide the course instructors with a written request on paper and your exam. In either case, include a justification for your claims. The appeals deadline is one week after the respective graded item is first made available. Note that for an exam the entire exam will be remarked, and the assigned grade may go up or down as a result. If your appeal is concerned with a simple calculation error, please see the instructor during their office hours.

Academic Misconduct

The University of Guelph takes a very serious view of Academic Misconduct. Included in this category are such activities as cheating on examinations, plagiarism, misrepresentation, and submitting the same material in two different courses without written permission. Students are expected to be familiar with the section on Academic Misconduct in the Graduate Calendar and should be aware that expulsion from the University is a possible penalty. If an instructor suspects that academic misconduct has occurred, that instructor has the right to examine students orally on the content or any other facet of submitted work. Moreover, it is expected that unless a student is explicitly given a collaborative project, all submitted work will have been done independently. For more details, see Sarah Brennan's video on Academic Integrity.

Course Schedule

Module	Sub-Module	Lecture Number	Date	Upcoming Deadlines
Big Data Introduction	History Characteristics Life Cycle	Lecture 1	Jan 9	
Scalable Computing	Distributed File Systems Parallel Computing Programming Models Hadoop Overview	Lecture 2	Jan 16	
Data Modeling and Management	Data Models Data stream and data lakes BDMS vs DBMS	Lecture 3	Jan 23	Lab 1 due
Data Integration and Processing	Postgres MongoDB	Lecture 4	Jan 30	
Data Integration and Processing	Aerospike Cassandra	Lecture 5	Feb 06	Project Proposal Due
Data Integration and Processing	Processing Pipelines Apache Spark	Lecture 6	Feb 13	Lab 2 due
Data Integration and Processing	Data Frames and SparkSQL	Lecture 8	Feb 27	
Machine Learning with Big Data	Overview Data Quality Feature Selection	Lecture 9	Mar 06	Lab 3 due
Machine Learning with Big Data	Feature Transformation Dimensionality reduction SparkML: Classification SparkML: Evaluation	Lecture 10	Mar 13	
Machine Learning with Big Data	SparkML: Regression SparkML: Clustering	Lecture 11	Mar 20	Lab 4 due
Project Presentation		Lecture 12	Mar 27	
Project Presentation		Lecture 13	Apr 03	