Course Outline

School:	Eng. Tech. & Applied Science
Department:	Information and Communication Engineering Technology (ICET)
Course Title:	Big Data Tools for Machine Learning
Course Code:	COMP 251
Course Hours/Credits:	56
Prerequisites:	COMP 247, COMP 254
Co-requisites:	N/A
Eligible for Prior Learning, Assessment and Recognition:	Yes
Originated by:	Mayy Habayeb
Current Semester:	Fall 2022
Approved by:	ppesikan c/o

Chairperson/Dean

Students are expected to review and understand all areas of the course outline.

Retain this course outline for future transfer credit applications. A fee may be charged for additional copies.

This course outline is available in alternative formats upon request.

Acknowledgement of Traditional Lands

Centennial is proud to be a part of a rich history of education in this province and in this city. We acknowledge that we are on the treaty lands and territory of the Mississaugas of the Credit First Nation and pay tribute to their legacy and the legacy of all First Peoples of Canada, as we strengthen ties with the communities we serve and build the future through learning and through our graduates. Today the traditional meeting place of Toronto is still home to many Indigenous People from across Turtle Island and we are grateful to have the opportunity to work in the communities that have grown in the treaty lands of the Mississaugas. We acknowledge that we are all treaty people and accept our responsibility to honor all our relations.

Course Description

In this course, students will be introduced to large scale learning: distributed learning. The concepts of distributed storage systems and parallel processing will be discussed. Storage types for big data (NoSQL) and big data tools (Hadoop eco system, YARN and Apache Spark) will be explained and students will gain hands-on experience by applying the big data tools in real world applications.

Program Outcomes

Successful completion of this and other courses in the program culminates in the achievement of the Vocational Learning Outcomes (program outcomes) set by the Ministry of Colleges and Universities in the Program Standard. The VLOs express the learning a student must reliably demonstrate before graduation. To ensure a meaningful learning experience and to better understand how this course and program prepare graduates for success, students are encouraged to review the Program Standard by visiting http://www.tcu.gov.on.ca/pepg/audiences/colleges/progstan/. For apprenticeship-based programs, visit http://www.collegeoftrades.ca/training-standards.

Course Learning Outcomes

The student will reliably demonstrate the ability to:

1. Explain and discuss the definition, drivers and challenges of ‘Big Data’.
2. Explain and discuss the main use cases of ’Big Data’.
3. Understand the concepts of distributed data storages, data clusters, data lakes and the variety of tools available for implementation.
4. Explain and discuss the various types of data stores available for structured, un-structured and semi-structured data.
5. Understand the concepts of parallel processing and variety of tools available for implementation (Map reduce, Spark,..).
6. Explain the fundamental concepts of Modern Data Warehousing and describe data mining building blocks and concepts
7. Practice and examine the elements of an analytics ‘Big Data’ pre-processing pipeline.
8. Understand the concept of graph structured data.
9. Discuss and explain the main concepts, advantages and challenges of data streaming.
10. Apply structured stream on a real time use case.
11. Design, code and test scripts to solve ‘Big Data’ business problems and create predictive learning models, using a variety of tools.

Essential Employability Skills (EES)

The student will reliably demonstrate the ability to*:

4. Apply a systematic approach to solve problems.
6. Locate, select, organize, and document information using appropriate technology and information systems.

**There are 11 Essential Employability Skills outcomes as per the Ministry Program Standard. Of these 11 outcomes, the following will be assessed in this course.*

Global Citizenship and Equity (GC&E) Outcomes

N/A

Methods of Instruction

Engaging and interactive lecture content.

Lab demonstrations and tutorials.

Hands on practical lab exercises.

Interactive discussion forms and boards.

Team project.

Text and other Instructional/Learning Materials

Text Book(s):

1.Frank Kane's Taming Big Data with Apache Spark and Python by Frank Kane Publisher Packt Publishing Ltd. ISBN :

2.Spark: The Definitive Guide by Bill Chambers, Matei Zaharia, 2018 Published by O'Reilly Media, Inc., ISBN: 9781491912218

3.Practical Data Science with Hapoop ®and Spark: Designing and building Effective Analytics at Scale by Ofer Mendelevitch, Casey Stella, Douglas Eadline Publisher: Addison-Wesley Professional ISBN: 9780134029733

4.Learning Spark- Lightning-Fast Data Analytics by Jules S. Damji, Brooke Wenig, Tathagata Das & Denny Lee Copyright c 2020 Databricks, Inc. Published by O'Reilly Media, Inc.,
See http://oreilly.com/catalog/errata.csp?isbn=9781492050049 for release details

Online Resource(s):

Safari IT Books online

Online documentation of : Spark, Hadoop

Material(s) required for completing this course:

Lecture and lab notes/PowerPoints/scripts and videos

Custom Courseware:

Vmware COMP251_image

Evaluation Scheme

- Test #1: Test covering materials of week 1 to 6
- Lab assignments: Lab assignment # 1: Datalakes
  Lab assignment # 2: Sharding
  Lab assignment # 3: Spark stand alone analytics
  Lab assignment # 4: Spark ML

- Quizzes: Four quizzes to examine the concepts of big data tools.
- Group research project #1: Students will work in groups to research the various big data platforms available. Each group will be assigned a platform to investigate and share the results with the class.
- Test #2: Test covering materials of weeks 7-13
- Online participation: Participation in online discussion boards.

Evaluation Name	CLO(s)	EES Outcome(s)	GCE Outcome(s)	Weight/100
Test #1	1, 2, 3, 4, 5, 6	4		20
Lab assignments	3, 5, 9, 10, 11	4, 6		20
Quizzes	1, 2, 3, 4, 5, 6, 8, 9	4		20
Group research project #1	3, 6, 7	6		15
Test #2	5, 9, 11	4, 6		20
Online participation	1, 2, 4, 5, 6, 8	4, 6		5
Total				100%

If students are unable to write a test they should immediately contact their professor or program Chair for advice. In exceptional and well documented circumstances (e.g. unforeseen family problems, serious illness, or death of a close family member), students may be able to write a make-up test.

All submitted work may be reviewed for authenticity and originality utilizing Turnitin®. Students who do not wish to have their work submitted to Turnitin® must, by the end of the second week of class, communicate this in writing to the instructor and make mutually agreeable alternate arrangements.

When writing tests, students must be able to produce official Centennial College photo identification or they may be refused the right to take the test or test results will be void.

Tests or assignments conducted remotely may require the use of online proctoring technology where the student's identification is verified and their activity is monitored and/or recorded, both audibly and visually through remote access to the student's computer and web camera. Students must communicate in writing to the instructor as soon as possible and prior to the test or assignment due date if they require an alternate assessment format to explore mutually agreeable alternatives.

Student Accommodation

The Centre for Accessible Learning and Counselling Services (CALCS) (http://centennialcollege.ca/calcs) provides programs and services which empower students in meeting their wellness goals, accommodation and disability-related needs. Our team of professional psychotherapists, social workers, educators, and staff offer brief, solution-focused psychotherapy, accommodation planning, health and wellness education, group counselling, psycho-educational workshops, adaptive technology, and peer support. Walk in for your first intake session at one of our service locations (Ashtonbee Room L1-04, Morningside Room 190, Progress Room C1-03, The Story Arts Centre Room 285, Downsview Room 105) or contact us at calcs@centennialcollege.ca, 416-289-5000 ext. 3850 to learn more about accessing CALCS services.

Use of Dictionaries

- Any dictionary (hard copy or electronic) may be used in regular class work.

Program or School Policies

N/A

Course Policies

N/A

College Policies

Students should familiarize themselves with all College Policies that cover academic matters and student conduct.

All students and employees have the right to study and work in an environment that is free from discrimination and harassment and promotes respect and equity. Centennial policies ensure all incidents of harassment, discrimination, bullying and violence will be addressed and responded to accordingly.

Academic Honesty

Academic honesty is integral to the learning process and a necessary ingredient of academic integrity. Forms of academic dishonesty include cheating, plagiarism, and impersonation, among others. Breaches of academic honesty may result in a failing grade on the assignment or course, suspension, or expulsion from the college. Students are bound to the College's AC100-11 Academic Honesty and Plagiarism policy.

To learn more, please visit the Libraries information page about Academic Integrity https://libraryguides.centennialcollege.ca/academicintegrity and review Centennial College's Academic Honesty Module: https://myappform.centennialcollege.ca/ecentennial/articulate/Centennial_College_Academic_Integrity_Module_%202/story.html

Use of Lecture/Course Materials

Materials used in Centennial College courses are subject to Intellectual Property and Copyright protection, and as such cannot be used and posted for public dissemination without prior permission from the original creator or copyright holder (e.g., student/professor/the College/or third-party source). This includes class/lecture recordings, course materials, and third-party copyright-protected materials (such as images, book chapters and articles). Copyright protections are automatic once an original work is created, and applies whether or not a copyright statement appears on the material. Students and employees are bound by College policies, including AC100-22 Intellectual Property, and SL100-02 Student Code of Conduct, and any student or employee found to be using or posting course materials or recordings for public dissemination without permission and/or inappropriately is in breach of these policies and may be sanctioned.

For more information on these and other policies, please visit www.centennialcollege.ca/about-centennial/college-overview/college-policies.

Students enrolled in a joint or collaborative program are subject to the partner institution's academic policies.

PLAR Process

This course is eligible for Prior Learning Assessment and Recognition (PLAR). PLAR is a process by which course credit may be granted for past learning acquired through work or other life experiences. The PLAR process involves completing an assessment (portfolio, test, assignment, etc.) that reliably demonstrates achievement of the course learning outcomes. Contact the academic school to obtain information on the PLAR process and the required assessment.

Semester:	Fall 2022	Professor Name:	See e-centennial course shell
Section Code:	All	Contact Information:	See e-centennial course shell
Meeting Time & Location:	See my centennial		

Topical Outline (subject to change):

Week	Topics	Readings/Materials	Weekly Learning Outcome(s)	Instructional Strategies	Evaluation Name and Weight	Evaluation Date
1	Course Overview. Introduction to Big Data. Use cases for big data. Data warehousing and Data Lake Stores. Data Lake Analytics components.	Microsoft tutorial Lecture notes.	Explain and discuss the definition of "Big data" and the five Vs'. Explain the definition of data warehouses and data lakes. Differentiate between data-lakes and Data warehouse. Discuss the business use cases for big data Create and configure a Data Lake Store resource. Create and configure a Data Lake Analytics workspace.	Lecture online content Lab online tutorial Videos In-class sessions for in-class program. Discussion boards		
2	Hadoop Eco system (HDFS, Yarn, pig, hive, ..etc.) Hadoop file system HDFS Distributed data storage concepts Parallel processing. Map Reduce.	Chapter 3 (orf, casey, Douglas) Lecture notes.	Discuss the Hadoop platform, its history, and its evolution. Discuss and explain the key components of the Hadoop Eco system. (HDFS, Yarn, Hive, Pig, Flume, Kafka, Storm, Scoop..etc.) Discuss and explain the concept of distribute file systems. Discuss and explain the concept of parallel processing and the Map Reduce process.	Lecture online content Lab online tutorial Videos Discussion boards In-class sessions for in-class program.	Lab assignment #1	
3	Structured un-structed and semi-structured data SQL / No SQL / New SQL	Chapter 17 (Deitel) Lecture notes on course shell.	Explain the major types of NoSQL databases: Key – value databases Document databases Columnar databases Graph databases Explain the concepts of NewSQL. Develop simple scripts to query data on a NoSQL datastore.	Lecture online content Lab online tutorial Videos Discussion boards In-class sessions for in-class program.	Quiz #1	week 3
4	Overview of Apache Spark components. Applications of spark. Spark's high-level	Chapters 4-10 (Bill & Matei) Chapter 5 (Frank) Lecture notes.	Explain the history and evolution of Spark. Explain the main components of Spark framework. (core, machine learning, sql..etc.) Explain and discuss the key actions and transformations available in Apache Spark	Lecture online content Lab online tutorial Videos Discussion boards In-class sessions for	Assignment #2	week 4

Week	Topics	Readings/Materials	Weekly Learning Outcome(s)	Instructional Strategies	Evaluation Name and Weight	Evaluation Date
	APIs.		SQL. Install and setup Spark on local machine. Design, develop Spark SQL scripts for creating, transforming and querying data forms.	in-class program.		
5	Fundamental DataFrame operations Columns Records &rows Dataframes Sampling Working with Different Types of Data Aggregation Joins Read and write from all different kinds of data sources	Chapters 5,6,7,8 (Bill & Matei) Chapter 2 (Frank) Lecture notes.	Create , manipulate dataframes using the high level api Read different types of files into spark List the main operations that include shuffling List the differences between parquet and csv files	Lecture online content Lab online tutorial Videos Discussion boards In-class sessions for in-class program.	Quiz #2	week 5
6	Spark sql Resilient Distributed Dataset (RDD) Spark's lower-level APIs DataBricks	Chapters 10,12,13,15,16,17,18 (Bill & Matei) Chapter 2 (Frank) Lecture notes.	Define Sparks RDDs. List and explain the main types and characteristics of RDDs. Discuss and explain the main scenarios when you need to use Sparks low level RDD Api s. Understand how to interoperate between DataFrames, Datasets, and RDDs. Discuss the main RDD transformations. Discuss the main RDD actions. Develop scripts to call Spark's lower-level APIs. Setup an account on the cloud and configure a cluste	Lecture online content Lab online tutorial Videos Discussion boards In-class sessions for in-class program.	Project #1 due	week 6
7	Review & tests	week 1 - 6	Week 1 -6	Test	Test 01 Assignment #3	week 7
8	Apache Spark ML/1 Pre-processing Supervised learning	Chapters 25,26,27 (Bill & Matei) Lecture notes	Discuss and explain the role of Transformers & Estimators for preprocessing in MLlib. Explain the use of the " VectorAssembler" tool to prepare the data. Explain the techniques to deal with	Lecture online content Lab online tutorial Videos Discussion boards In-class sessions for		week 8

Week	Topics	Readings/Materials	Weekly Learning Outcome(s)	Instructional Strategies	Evaluation Name and Weight	Evaluation Date
			continuous and categorical data in big data. Explain the techniques used for scaling and normalization in big data. Discuss and explain techniques to handle text data. (Tokenize…etc.) List available techniques for feature selection in big data. Apply preprocessing to a sample of big data using Apache Spark MLlib classes and methods. Design develop Apache spark scripts to address supervised learning business problems.	in-class program.		
9	Apache Spark ML/2 Unsupervised learning- clustering Recommendation	Chapters 28,29 (Bill & Matei) Chapter 2 (Frank)	Discuss and explain the “cold start” problem in recommender systems. Design, train, tune, and evaluate a recommender models using Apache Spark. Design develop Apache spark scripts to address unsupervised learning business problems.	Lecture online content Lab online tutorial Videos Discussion boards In-class sessions for in-class program.	quiz #3	week 9
10	Graph data	Chapters 30 (Bill & Matei)	Understand the concept of graph data. Explain the concept of motifs for pattern recognition in graph datasets. Understand how to use GraphFrames to perform graph analytics on Spark. Explore graph data using Apache Sprak. Explain the most popular graph analytics algorithms: PageRank, connected components, Breadth first search..etc. Apply graph algorithms on bike sharing dataset.	Lecture online content Lab online tutorial Videos Discussion boards In-class sessions for in-class program.	Assignment #4	
11	Data streaming concepts	Chapters 20 & 21 (Bill & Matei) Chapter 8 Learning Spark Lecture notes.	List the advantages and challenges of streaming. Differentiate between” continuous” versus “micro-batch” execution design concepts. Discuss and explain the Spark Structured Streaming API. Apply structured streaming in a development environment.	Lecture online content Lab online tutorial Videos Discussion boards In-class sessions for in-class program.		week11
12	Advanced structured streaming	Chapters 21 & 22 (Bill & Matei)	Discuss and explain the sources and sinks for the structured streaming API.	Lecture online content Lab online tutorial	quiz #4	week 12

Week	Topics	Readings/Materials	Weekly Learning Outcome(s)	Instructional Strategies	Evaluation Name and Weight	Evaluation Date
		Chapter 8 Learning Spark Lecture notes.	Apply structured streaming on different scenarios of sources and sinks. Differentiate between “Event Time” versus “Processing Time” design concepts. Apply structured stream on a real time use case working on the “Heterogeneity Human Activity Recognition Dataset”.	Videos Discussion boards In-class sessions for in-class program.		
13	Deep learning on “Big Data”	Chapters 31 (Bill & Matei) Lecture notes.	Discuss and explain several common approaches to using deep learning in Spark. List the deep learning libraries available for big data and their main use cases. Explain the deep learning pipelines and high-level APIs for scalable deep learning.	Lecture online content Videos Discussion boards	Assignment #5	
14	Final Test	Final Test	Final Test	Final Test	Final Test	week 14