ENSF 612: Fall 2021
Lecture 1. Introduction

Dr. Gias Uddin, Assistant Professor
Electrical and Software Engineering (ESE)
Schulich School of Engineering
University of Calgary
https://giasuddin.ca/

Topics

- Course objectives
- Topics
- Grading
- Labs
- Textbooks
- Course Website

Course Objectives

- Understand reasons for data increase, i.e., big data
- Understand benefits from big data analysis
- Understand challenges in analyzing large datasets
- Learn about some platforms for big data analysis
- Understand typical steps in big data analysis
- Learn common algorithms used in big data analysis
- Implement machine learning pipelines from big data

Course objectives – cont'd

- Learning outcomes
  - Familiarity with algorithms that can process large amounts of data
    - Large scale text processing
    - Social network analysis
    - Large scale linear algebra
  - Building predictive models from big data
    - Using large scale machine learning algorithms
  - Familiarity with the Apache Hadoop and Apache Spark platforms

Topics

- Introduction and motivation
  - Reasons for data proliferation
  - Benefits of big data analysis – real life case studies
  - Challenges in analyzing large datasets
  - Typical steps in big data analysis
- Introduction to Hadoop MapReduce framework
  - System architecture
  - Phases of a MapReduce job
  - Scalability and reliability of a MapReduce system
  - Algorithms that can benefit from MapReduce paradigm

Topics – Cont'd

- MapReduce programming using Hadoop streaming
- MapReduce algorithms – Text processing
  - Counting words in a large document corpus
  - Sorting large datasets
  - Searching in a large document corpus
  - Constructing N-grams
- MapReduce algorithms – Graph analysis
  - Large scale graph algorithms
  - Social network applications
- MapReduce algorithms – Linear algebra
  - Operations on large matrices

Topics – Cont'd

- Hadoop ecosystem
  - Tools/platforms built on top of Hadoop
- Limitations of Hadoop MapReduce
- In-memory big data clusters – Apache Spark
  - System architecture
  - Differences from Hadoop MapReduce
- Programming Spark using PySpark
- Large scale machine learning using Spark
  - Machine learning basics
  - Linear and Logistic regression

Grading

Component	Weight
2 Lab assignments – individual	10%
2 Quizzes – individual	10%
1 Midterm Exam – individual	25%
3 Project presentations – Group	15%
1 Project Final Report – Group	20%
1 Project Code and Data Quality – Group	20%

Need to obtain passing grade in project final report to pass course

Grading – Cont'd

50% of Grading – as an Individual: Assignments, Quiz, Midterm

- Lab assignments
  - Two assignments
  - Will involve writing Hadoop and Spark code
- Midterm
  - One midterm
  - Based on course lectures
- Quiz
  - Two quizzes
  - One before midterm – to practice
  - One after the last lecture of the course – to recap what we have learned in the course

Grading – Cont'd

50% of Grading – as a team: Course Project

- Project
  - To be completed in group of 3 students
  - Will involve coding in python and spark ML (Machine Learning)
- Project Grading involves three components
  - Presentations
  - Project Source Code and Data
  - Project Final Report

Grading – Cont'd

50% of Grading – as a team: Course Project

Item 1 - Project presentation (15%)

- To be done in a group of 3 students
- Based on progress on group project:
  - first on project idea,
  - second on project progress, and
  - third on project completion
- Each presentation is judged on of three items:
  - 2-minute video demo
  - 10-minute presentation
  - 5 minutes for Q&A

Grading – Cont'd

50% of Grading – as a team: Course Project

- Item 2 - Project data and source code (20%)
  - Each project will generate a curated dataset based on manual analysis and data preprocessing of several online data sources, such as Stack Overflow, GitHub
  - Each group will manually label the textual contents to some predefined categories
  - Each group will use the labeled data to experiment with a suite of Machine Learning classifiers in pyspark

Grading – Cont'd

50% of Grading – as a team: Course Project

- Item 3 - Project final report (20%)
  - A template will be provided to write the final report
  - Need to pass on the final report to pass on the course

Lab

There will be no labs

- Coding exercises and projects will be in python
  - Databricks community edition
    - For small data analysis and project demo
  - Hadoop Hortonworks sandbox
    - VM that you can run on your laptop
    - For code development and debugging using small data sets
  - U of C Hadoop cluster
    - For running code on medium/large sized data sets
  - Microsoft/Amazon/Google Cloud
    - For running code on medium/large sized data sets

Textbooks

- There is NO official textbook for this course
- Slides will be a major source of content
- However, they are not the only source
  - Take good notes during lectures
  - Refer to the recommended textbooks
- Recommended textbooks

Title	Hadoop: The Definitive Guide
Author(s)	Tom White
Edition, Year	Edition 3, 2012
Publisher	O' Reilly Publishers

Title	Data-Intensive Text Processing with MapReduce
Author(s)	Jimmy Lin and Chris Dyer
Edition, Year	Edition 1, 2012
Publisher	Morgan and Claypool Publishers

Textbooks – cont'd

- Recommended textbooks – cont'd

Title	Mining of Massive Datasets
Author(s)	Anand Rajaraman and Jeffrey David Ullman
Edition, Year	Edition 1, 2010
Publisher	Cambridge University Press

Title	Learning Spark: Ligthining-Fast Big Data Analysis
Author(s)	Holden Karau, Andy Kowinski, Patrick Wendell and Matei Zaharia
Edition, Year	Edition 1, 2015
Publisher	O' Reilly Publishers

Course Website

- D2L will be used for the course
- Will have lecture slides, assignments, and projects
- Follow the announcements carefully!

Topics

- Course objectives
- Topics
- Grading
- Labs
- Textbooks
- Course Website