SENG 5980-001: Big Data Engineering and Analytics

Syllabus as of 9/12/3017

Fall 2017 (3 credits)

Instructor: Donald Sawyer

Class Meetings

When: Tuesdays & Thursdays, 4:00 – 5:15 PM

Where: Akerman Hall 225

Teaching Assistant

None Assigned

Office Hours

Tuesdays 3-3:50 PM, and for 30 minutes after class each day, or by appointment. Email and IM (Google Talk) are effective ways of getting questions answered. If the class chooses, I will set up a Slack channel.

Office Location

Keller Hall 6-196

Course Objectives

This course is a robust discussion of many topics to learn the engineering concepts with regard to big data and analytics projects. Students will gain experience analyzing and implementing a big data storage solution, using Hadoop, and using Python for data analysis.

At the conclusion of the course, students will have experience with:

- Big data fundamentals
- Big data processing concepts
- Understanding different NoSQL storage solutions
    - Key/value data stores
    - Column data stores
    - Document data stores
    - Graph data stores
    - Message brokers and Kafka
    - ElasticSearch (ELK stack)
- Stream processing
- Analyzing/designing big data architectures
- Working within the Hadoop ecosystem
- Understanding the importance of CI/CD in big data projects

The course will be a mix of lecture, discussion, use case evaluation, and hands-on activities to cover the broad domain of big data engineering.

Prerequisites

Previous knowledge or coursework in the following are required:

- Data structures (CSCI 4041)
- Relational systems and/or SQL (preferred)

No previous experience with big data, statistics, or machine learning is required.

Textbooks

Big Data Fundamentals: Concepts, Drivers & Techniques
ISBN: 978-0-1342-9120-8
Authors: Erl, Khattak & Buhler

NoSQL For Mere Mortals
ISBN: 978-0-1342-1076-6
Author: Dan Sullivan

Recommended Texts:
Hadoop: The Definitive Guide, 4th Edition
ISBN: 978-1-4919-0163-2
Authors: Tom White
http://proquestcombo.safaribooksonline.com.ezp1.lib.umn.edu/9781491901687

Throughout the course, there will be various readings from free publications that can be downloaded from the internet or on University library resources.

Course Website

https://ay17.moodle.umn.edu/course/view.php?id=4182
Moodle will contain course/lecture materials, schedule, reading assignments, homework assignments, links to supplemental readings, schedules and contact information.

Assignments

There will be assignments/projects that will contribute to your grade, as well as a bonus Python assignment for extra credit.

The assignments will be completed in 2-4 person teams. Deliverables for each assignment are listed below.

Assignment 1: Storing and analyzing a big dataset using a NoSQL technology.

- Phase 1: Choose a dataset/technology to analyze (written deliverable)
- Phase 2: Load the data set and perform some data analysis. The deliverables will be a demo to the class and discussion on lessons learned, as well as a written deliverable.

Assignment 2: Hadoop Lab + Project

- Students will use a Hadoop virtual machine to perform multiple activities in Hadoop, such as interacting with HDFS, loading/query HBase, loading/querying Hive, Hive views, ETL with Pig, and other activities.
- Some class time will be given during class to work on the lab and ask questions. Access to a laptop by a group member will be necessary.
- Deliverable will be a write-up of lab solutions and answering some basic analytical questions.
- Use your dataset from Assignment 1 for Hadoop analysis.

Bonus Assignment: Python/R Data Analysis

- Use Python and Jupyter Notebooks to do data analysis.
- Deliverable is a markdown document with a completed data analysis study.

Quizzes/Exams

Reading Quizzes:

- There will be a 5-point Moodle quiz due at class start on the readings
- There will be no make-ups without prior approval

Attendance Pop Quizzes (bonus points):

- There will be 10 points available as pop quizzes during the semester
- Half the points are for attending class, half are for a completely correct answer

Midterm Exam: Tuesday, October 24

Final Exam: Tuesday, December 19 from 4-6 PM

Class Format

The class will be conducted in a lecture/discussion format. Lectures will introduce concepts, extend the reading assignments, and be used to explore concepts deeper.

Grading

The final grade will be determined on the delivery of the assignments and their group participation. Groups are required to submit all three assignments to pass the course.

Students will grade each other on their group participation and contributions. Each student is required to rate each member of their team by sending me an email within 48 hours of the due date rating their teammates with a score of 1-10. If a student doesn't provide a rating for the window, they will not receive any of their possible 10 participation points. *I will not send reminders about group ratings.*

Grading breakdown is as follows:
Assignment 1: 100 points (25 for phase 1, 75 for phase 2)
Assignment 2: 100 points (50 for lab, 50 for Hadoop project)
Group Participation: 20 points
Bonus: 10 (10 for Python Assignment, 10 for attendance)
Quizzes: ~220 points (some quizzes may be dropped)
Midterm: 100 points
Final Exam: 150 points

Assignments can be turned in late with a penalty. There will be a deduction of 50% of the points if the assignment is turned in within 48 hours of the due date. A grade of zero (0) is awarded thereafter.

The class will NOT be graded on a curve, and will use the following grade ranges:
A (>= 93), A- (90-92.99)
B+ (87-89.99), B (83-86.99), B- (80-82.99)
C+ (77-79.99), C (73-76.99), C- (70-72.99)
D+ (67-69.99), D (60-66.99)
F (< 60)

Collaboration

All assignments will be group assignments, with individual peer assessments. Group assignments will be assigned a single grade. In rare cases, consistent negative peer assessments will change an individual grade downward. Cheating will not be tolerated.

Integrity and Ethics

The policy of the university on scholarship and grades will be followed. Implicit in handing in homework, assignments, papers, and exams is that they represent the student's own work (or the result of sanctioned collaboration). Any exceptions should be explicitly noted. Representing someone else's work as one's own is grounds for failing the course.

Classroom Climate

All students are expected to behave as scholars at a leading institute of technology. This includes arriving on time, not talking during lecture (unless addressing the instructor), and not leaving the classroom before the end of the lecture. Disruptive students will be warned and potentially dismissed from the classroom.

Make-up	Make-ups for graded activities may be arranged if your absence is caused by documented illness or personal emergency. A written explanation (including supporting documentation) must be submitted to your instructor; if the explanation is acceptable, an alternative will be arranged. Whenever possible, make-up arrangements will be completed prior to the scheduled activity. A student not turning in an assignment will receive a score of 0. Alternative times for the final exam will be arranged only under university criteria for rescheduling a final exam.
Incompletes	The I grade indicates that the instructor has (1) reasonable expectations that the student can complete an unfinished course on her/his own no later than the end of the next quarter and (2) believes that legitimate reasons exist to justify extending the deadline for course completion. The only acceptable reasons will be documented illness or personal emergency. A written explanation (including supporting documentation) must be submitted to your instructor; if the explanation is acceptable, an Agreement for the Completion of Incomplete Work will be filled out as a contract between the student and the instructor.
Special Needs	It is University policy to provide, on a flexible and individual basis, reasonable accommodations to students that have disabilities that may affect their ability to participate in course activities or to meet course requirements. Students with disabilities are encouraged to contact their instructor early in the quarter to discuss their individual needs for accommodations.
Expected Workload	Students can expect to spend on average 4-6 hours per week outside of class on assignments, BEYOND the readings/review. During weeks when assignments are due, there may be a spike in increased time and effort.

THIS COURSE SYLLABUS IS SUBJECT TO CHANGE BY THE INSTRUCTOR.
