University *of* New Haven

TAGLIATELA COLLEGE OF ENGINEERING

<Department/Division>

DSCI 6007
Distributed and Scalable Data Engineering
Fall 2021
Meeting Times and Location(s): N/A – Online Asynchronous
Credit Hours: 3
Vahid Behzadan, Ph.D.
Faculty Contact Information:
Office Location: Maxy120A or Zoom
Phone: 203-479-4723 Email: vbehzadan@newhaven.edu
Office Hours: MW 12pm-1pm or by request
Department Chair: Dr. Ali Golbazi agolbazi@newhaven.edu

COURSE SYLLABUS

This syllabus is informational in nature and is not an express or implied contract. It is subject to change due to unforeseen circumstances, as a result of any circumstance outside the University's control, or as other needs arise. If, in the University's sole discretion, public health conditions or any other matter affecting the health, safety, upkeep or wellbeing of our campus community or operations requires the University to make any syllabus or course changes or move to remote teaching, alternative assignments may be provided so that the learning objectives for the course, as determined by the University, can still be met. The University does not guarantee that this syllabus will not change, nor does it guarantee specific in-person, on-campus classes, activities, opportunities, or services or any other particular format, timing, or location of education, classes, activities, or services.

If you are a student who believes that you require a reasonable accommodation related to camera usage for any assignment in this course, a representative of the Accessibility Resource Center (ARC) will review your request so that appropriate accommodations can be arranged. Students requesting accommodation(s) must be registered with the Accessibility Resource Center (ARC), and if not, must work with staff in the Accessibility Resource Center (ARC) to ensure that documentation of a disability is on file.

If you are a student with concerns regarding camera usage as a component in this course that does not rise to the level for a requested accommodation from the Accessibility Resource Center and/or your ability to participate in this course due to technology, a representative of the Dean of Students Office will review your request so that appropriate assistance can be arranged.

To request a reasonable accommodation or express concerns regarding the camera usage component of this course, please complete the online Camera Exemption Request Form.

The Accessibility Resource Center can be reached at (203) 932-7332 or by email at AccessibilityResCtr@newhaven.edu. For additional information, please refer to the Accessibility Resource Center (ARC) website at www.newhaven.edu/campusaccess. For additional assistance from the Dean of Students Office, please contact: deanofstudents@newhaven.edu. If you require assistance with the technology requirement, please visit the Student Technical Support page.

Course Description:
Advanced topics in "Big Data" infrastructure and architectures focusing on computing resources and programming environments to support the development of efficiently scalable high-volume distributed machine learning algorithms.

Required Text(s):
None. Data Engineering is a new and evolving field, and there is no standard book that covers it completely and is current. We will post readings for each day. They will be video tutorials, book chapters, and blog posts.

Optional References:

- The Data Engineering Cookbook by Andreas Kretz. Open-source work in progress.
- Designing Data-Intensive Applications (DDIA) by Martin Kleppmann. Clear, concise, and practical. Right now preview edition only, a game changer when finished.
- Big Data by Nathan Marz with James Warren. Much of the technology has changed since that book was written but the basic principles are the same.
- Learning Spark Spark is the new dominant analytics framework. This is an accessible introduction.
- Advanced Analytics with Spark Learn how to leverage Spark to solve Data Science problems through guided projects.
- The Manga Guide to Databases Learn databases without the tedium.

Other Materials/Supplies:
The class will be delivered online via Canvas. External tutorials, reading materials, and references will be provided. Class projects require access to a computer capable of running an Ubuntu 20.04 Virtual Machine on VirtualBox. Cloud-based exercises will be on Amazon AWS via a free AWS Academy account that will be provided to the students.

Course Structure/Course Format/Course Objectives:
This course is offered as an online asynchronous class: recorded lectures, external materials, and projects will be posted weekly on Friday afternoons. Several TA and instructor office hours will be held throughout the week to help with questions. The focus will be applying concepts to data through programming.

Course Objectives:
By the end of the course, you should be able to:

- Install and run a Linux virtual machine locally and in the cloud
- Utilize *NIX command line tools to manipulate and analyze data
- Deploy and manipulate data and working code in the cloud

- Write complex SQL queries
- Design a database that conforms to the third normal form (3NF)
- Design, create, and query NoSQL databases
- Identify embarrassingly parallelizable tasks and parallelize them
- Describe and apply the MapReduce algorithm
- Describe and apply Spark's Dataset abstraction
- Apply machine learning in a distributed architecture
- Analyze streaming data "real-time"
- Apply probabilistic data structures to handle high volume/velocity data
- Build and use an information retrieval (IR) or search engine
- *Build an end-to-end distributed data-pipeline*

Student Learning Outcomes:

Demonstrate achievement of course objectives in class discussion, lab assignments, and projects.

Course Requirements & Assessment:

Please see official University of New Haven Academic Policies located in the links below:
Graduate Grading System

Assignments/Projects

- All work must be turned in via Canvas, unless otherwised specified (some lab assignments will require submission on Github or AWS). Please turn in whatever you have for participation credit, even if incomplete.
- Pen-and-paper quizzes will also be required. However, submission will need to be in the form of scanned/photographed copies through Canvas.

Participation

- Active-learning techniques will be used, such as group discussions and “think-pair-share”, requiring students to work individually and/or with other students. Refusal to participate will be treated as absence from class and ultimately lead to dismissal from the class (see University Policies).

Midterm and Final Projects

- The midterm and final projects aim to evaluate the students’ ability to leverage the skills and materials covered in the lectures and labs in solving realistic problems in data engineering. The assessment of both midterm and final projects will be based on the outcome, demonstrated in a written technical report, as well as class presentations.

Grading:

Grades earned are based on your performance on class participation (including quizzes), lab assignments, and midterm + final projects.

Participation	%10
Labs	%25

Midterm Project	%15
Final Project	%50
Total**	100%

**Final Grades are assigned with the following scale:
Choose the scale applicable for your course. You may change the scale to the needs of the course/program.

Typical Undergraduate Scale

Grades Scored Between			Letter Equivalent
97	to	100	A+
94	to	Less than 97	A
90	to	Less than 94	A-
87	to	Less than 90	B+
84	to	Less than 87	B
80	to	Less than 84	B-
77	to	Less than 80	C+
74	to	Less than 77	C
70	to	Less than 74	C-
67	to	Less than 70	D+
63	to	Less than 67	D
60	to	Less than 63	D-
		Less than 60	F

Typical Graduate Scale

Grades Scored Between			Letter Equivalent
97	to	100	A+
94	to	Less than 97	A
90	to	Less than 94	A-
87	to	Less than 90	B+
84	to	Less than 87	B
80	to	Less than 84	B-
77	to	Less than 80	C+
74	to	Less than 77	C
70	to	Less than 74	C-
		Less than 70	F

The calculation of final grades is determined by the faculty member. The calculated grade in the total column in Canvas may or may not be reflective of your final grade.

Expectations:

Students should expect to spend at least 9 hours on academic studies per week on this course. There will be readings, simple questions/problems, lab assignments and projects. Students must work individually on assignments and projects unless specifically allowed to work in groups by the instructor. Any work taken from the internet must be cited properly (acceptance of code taken from elsewhere is at the discretion of the instructor) or will be considered plagiarism. Failure to adhere to this policy will result in penalties ranging from a zero on the assignment to a zero in the final grade. Students may also be subject to disciplinary action by the University of New Haven (see University Policies).

Course Outline/Schedule:

Day/Date	**Topic/Note**
Wk 1	Welcome to data engineering
Wk 2	Internet, HTTP, and HTML
Wk 3	Linux, Virtualization & the Cloud
Wk 4	Databases - Intro to NoSQL
Wk 5	Databases – Advanced SQL
Wk 6	Intro to Parallelization and MapReduce
Wk 7	Intro to Spark
Wk 8	More Spark – Midterm Project Announced

Wk 9	Even More Spark – Designing Big Data Systems – Midterm Project Due
Wk 10	Streaming – Introduction to Machine Learning in Sparks
Wk 11	Streaming - Continued
Wk 12	Introduction to Kubernetes – Final Project Topic Selection
Wk 13	Full-Stack Deep Learning – I
Wk 14	Full-Stack Deep Learning - II
Wk 15	Final Project Presentations
Finals Week	Final Project Report Due

Diversity Statement

The University of New Haven embraces diversity and recognizes our responsibility to foster a diverse, inclusive, and welcoming environment in which all members of the Charger community of all backgrounds and identities can learn, work, and live together. We benefit from the academic, social, and cultural developments that arise from a diverse campus that is committed to equity, inclusion, belonging, and accountability.

We have a responsibility as a community and as individuals to address and remove barriers, achieve success, and sustain a culture of inclusivity, empathy, kindness, and compassion. We encourage, welcome, and embrace participation in ongoing dialogue, engagement, and education to critically examine and thoughtfully respond to the changing realities of our community. Diversity, equity, inclusion, acceptance, and belonging enrich the Charger community and are instrumental to institutional success and fulfillment of the University mission.

Reporting Bias Incidents

At the University of New Haven, there is an expectation that all community members are committed to creating and supporting a climate which promotes civility, mutual respect, and open-mindedness. There also exists an understanding that with the freedom of expression comes the responsibility to support community members' right to live and work in an environment free from harassment and fear. It is expected that all members of the University community will engage in anti-bias behavior and refrain from actions that intimidate, humiliate, or demean persons or groups or that undermine their security or self-esteem.

If you have an immediate safety concern for yourself or others, and/or believe someone poses an immediate threat to themselves or others, please contact University Police at 203-932-7070 or call 911. Community members can report bias-motivated incidents by completing the form at **www.newhaven.edu/biasreporting**. Community members are encouraged to complete this form if they are the target of bias or harassing behaviors, witness such behaviors, or gain knowledge of these behaviors occurring within the University community. All matters concerning bias and harassment will be handled by the Dean of Students Office and Human Resources Office.

University-wide Academic Policies

A continually-updated list of University-wide academic policies and descriptions of key university student resources, can be found on Canvas. You can access them by simply clicking on the (?) help button.

The University-wide academic policies include (but are not limited to) the University's attendance policy, procedures for both adding / dropping a course and course withdrawals, an explanation for the sorts of circumstances where incomplete (INC) grades could be considered by the faculty, and the academic integrity policy (among others). Also in this location you will find information regarding the process for reporting bias and topics related to our maintaining a positive learning environment (including, but not limited to, discrimination and sexual misconduct).

The list of key university student resources to enable learning include (but are not limited to) the University's Center for Student Success, Writing Center, Center for Learning Resources, and the Accessibility Resource Center.

Course Delivery Options

For courses with a location of ONLI that also list time and day information, students should plan to be available during that time for synchronous online learning. Classes with no time listed are asynchronous, fully online, and coded ONLI. Due to social distancing requirements, some courses will be offered in Hybrid or Flex formats. ***Online learning may occur at any time, depending on how the University may determine classes are best taught under any circumstance.*** *All courses, including selected labs, practicums, project-based, and clinical courses may require limited, additional meeting times to accommodate accreditation, regulatory, and similar requirements and students will be advised about this at the beginning of those classes.*

Online asynchronous (ONLI): Fully online asynchronous course with **no live, required class sessions**.

University of New Haven

UNIVERSITY STUDENT SUPPORT SERVICES

The University recognizes that students can often use some help outside of class and offers academic assistance through several offices.

Accessibility Resources Center

Students with disabilities, chronic health-related conditions, or military service-connected disorders are encouraged to share, in confidence, information about course specific approved reasonable accommodations. The Accessibility Resources Center, located in Sheffield Hall, is responsible for and committed to providing supports and resources that serve to promote educational equity and ensure that students are able to participate in the opportunities available at the University of New Haven. Reasonable accommodations are not made without written documentation from the Accessibility Resources Center.

Center for Learning Resources (CLR)

The Center for Learning Resources (CLR), located in the Peterson Library, provides academic content support to the students of the University of New Haven using metacognitive strategies that help students become aware of and learn to apply optimal learning processes in the pursuit of creating independent learners. CLR tutors focus sessions on discussions of concepts and processes and typically use external examples to help students grasp and apply the material.

Center for Student Success (CSS)

The Center for Student Success provides students with a multitude of resources available on campus and assists students in fulfilling their educational, social and personal goals.

Counseling & Psychological Services (CAPS)

CAPS offers confidential, free services in order to support student mental health and wellbeing. The services include individual and group therapy, support groups, consultations, and 24/7 crisis support. We are available in person and remotely, and are in the office M-F, 8:30-4:30. Please call us to schedule an appointment or with any questions at 203-932-7333; you can also schedule online. If you experience a mental health crisis after hours, you can call our main number for support.

Myatt Center

The Myatt Center for Diversity and Inclusion is committed to creating a multicultural environment through intentional education, campus community engagement, and valuing the unique identities of each member of the Charger Community. Our commitment to diversity is driven by the core values of connection, belonging, inclusivity, equity, acceptance, and accountability. The Myatt Center's focus is to create a respectful and inclusive environment based our awareness and ability to engage with others who are different on many levels including ethnicity, race, sexual orientation, gender, military, religious belief, and life experiences.

Marvin K. Peterson Library

The Library provides access to online databases, e-books, e-journals, electronic U.S. Government Documents, print books, educational games, and audiovisual materials. A search can be conductedthrough all these resources at once by using the search box "Articles, Books, & More."

The Library provides three floors with individual quiet study space, collaborative group study space, study rooms with technology, whiteboards, Dell desktops, iMacs, scanners, and printers. The entire library is a wireless zone.

Librarians assist in locating relevant sources of information for research papers, thesis, honors thesis, and other projects. Librarians answer general reference questions and help with effectively evaluating sources of information. Help is available through a Chat Service, 24/7 Ask a Librarian Service, a Zoom Reference Service, and by E-Mail. Complete the Research Consultation Form to arrange a time convenient for you.

LibGuides are created to assist students with research. They contain an overview of resources available through the library, as well as tutorials, subject guides, and course specific guides.

University Writing Center

The mission of the Writing Center is to provide high-quality tutoring to undergraduate and graduate students as they write for a wide range of purposes and audiences. Tutors are undergraduate and graduate students who are majoring in a variety of fields across the University. We are here to work with you at any stage in the writing process; just bring in your assignment, your ideas, and any writing you've done so far. To make an appointment, you can register for an account with our scheduling site https://newhaven.mywconline.com or visit us in person at our desk on the first floor of Peterson Library (just to the left after you enter the library).