University at Buffalo
School of Management

MGS 655 : Distributed Computing and Big Data Technologies

3 credits

Section: F1S
Class Hours: M-W: 5:00-6:20 pm (Jacobs 112),
Office Hours: M: 3:50-4:50 pm

Instructor: Dr. Haimonti Dutta
Office: 339 Jacobs
Email: haimonti@buffalo.edu

Course will be available on UBLearns. Please check regularly for updates to schedule, class presentations, reading materials and other discussions. Occasionally email messages to the whole class will be sent through UBLearns. Missed classes (if applicable) will be made up based on the instructor's discretion.

Course Description

Large scale computing environments aggregate resources from networked components. Such components are organized in parallel or distributed architectures and are highly available and scalable. Examples include the internet, intranets, local and wide area networks, cloud computing infrastructure, and ad-hoc wireless sensor networks.

This course will introduce the techniques for creating functional, usable, high-performance distributed systems. It will discuss principles governing the design of large scale distributed systems focusing on architectures, communication protocols, synchronization, storage and file systems, consistency and data replication. Students will also learn how to work with Big Data in distributed environments. In particular, the Apache Hadoop framework for data intensive distributed applications will be discussed. It has two major components: (a) The Hadoop File System (HDFS) - A highly scalable and portable file system for storing the data and (b) Map-Reduce - A programming model for processing the data in parallel. This course will teach students how to use and manage these two technologies.

The class format will primarily be lectures given by the instructor. In addition, students will work in teams to gain hands-on experience in designing, implementing, and debugging large scale distributed systems. Evaluation will be based on quizzes occurring once a month and a semester long project.

Text Book and Related Reading Materials

- Andrew S. Tanenbaum and Maarten van Steen, "Distributed Systems – Principles and Paradigms", Second Edition, Prentice Hall (2007).
- Jimmy Lin and Chris Dyer, "Data Intensive Text Processing with MapReduce", Morgan and Claypool Publishers, 2000.
- Tom White, "Hadoop the Definitive Guide", O'Reilly. Available Online at `http://ce.sysu.edu.cn/hope/UploadFiles/Education/2011/10/201110221516245419.pdf`
- Arben Asslani, "Big Data Technologies for Business", Prospect Press, 2020.

Course Delivery

The course will be delivered in-person during class hours. There will be several (three-four) lectures, called lab sessions, during which the installation of Apache Hadoop will be demonstrated. Students will be notified by email when these lab sessions will be scheduled. The instructor is available during the office hours to discuss any issues pertaining to the course, materials and other pertinent details. You do not need to schedule an appointment to meet during office hours.

Pre-requisities

Prior background in programming and/or undergraduate level concepts in Operating Systems. Please contact the instructor if in doubt.

Course Requirements and Grading Policy

Component	Percentage
Quiz 1	30%
Quiz 2	30%
Project	35%
Individual Class Participation	5%

Final letter grades will be obtained by statistical method of grading on a curve (or bell curving). The quality points for each letter grade are shown in Table 1.

A grade of incomplete ("I") indicates that additional coursework is required to fulfill the requirements of a given course. Students may only be given an "I" grade if they have a passing average in coursework that has been completed and have well-defined parameters to complete the course requirements that could result in a grade better than the default grade. An "I" grade may not be assigned to a student who did not attend the course.

Prior to the end of the semester, students must initiate the request for an "I" grade and receive the instructor's approval. Assignment of an "I" grade is at the discretion of the instructor.

The instructor will specify a default letter grade at the time the "I" grade is submitted. A default grade is the letter grade the student will receive if no additional coursework is completed and/or a grade change form is not filed by the instructor. "I" grades must be completed within 12 months. The instructor may set shorter time limits for removing an incomplete than the 12-month time limit. Upon assigning an "I" grade, the instructor will provide the student specification, in writing or by electronic mail, of the requirements to be fulfilled, and shall file a copy with the appropriate departmental office. Please read the full Incomplete Policy: `https://www.buffalo.edu/grad/succeed/current-students/policy-library.html#i-grade`.

Grade	Quality Points
A	4.0
A-	3.67
B+	3.33
B	3.0
B-	2.67
C+	2.33
C	2.00
C-	1.67
D+	1.33
D	1.00
F	0.00

Table 1: Conversion of letter grades to quality points

Course Outline

1. Introduction to Large Scale Systems
    - Definition and types of Large Scale Systems
2. Architectures
    - Centralized, Distributed, Hybrid Architectures
    - Architecture vs Middlewares
3. Communication Protocols
    - Broadcast, Multicast and Gossip
4. Storage in Large Scale Systems
    - BigTable: A Distributed Storage System for Structured Data
5. Consistency and Replication
    - Basic Principles

- Protocols

6. Large Scale File Systems
    - Basic concepts and architecture
    - Distributed File System - Hadoop Distributed File System
    - Synchronization and Replication
7. Data Processing at Scale
    - Intro to MapReduce Programming
    - MapReduce Algorithm Design

Academic Integrity

Academic Integrity is critical to the learning process. It is your responsibility as a student to complete your work in an honest fashion, upholding the expectations your individual instructors have for you in this regard. The ultimate goal is to ensure that you learn the content in your courses in accordance with UB's academic integrity principles, regardless of whether instruction is in-person or remote. Thank you for upholding your own personal integrity and ensuring UB?s tradition of academic excellence. Please refer to the Office of Academic Integrity `https://www.buffalo.edu/academic-integrity.html` for further information.

Accessibility Resources

If you have any disability which requires reasonable accommodations to enable you to participate in this course, please contact the Office of Accessibility Resources, 60 Capen Hall, North Campus, 716-645-2608 or email: `stu-accessibility@buffalo.edu`, and also the instructor of this course during the first week of class. The office will provide you with information and review appropriate arrangements for reasonable accommodations, which can be found on the web at: `https://www.buffalo.edu/studentlife/who-we-are/departments/accessibility.html`.

Technology Recommendations

To effectively participate in this course, regardless of mode of instruction, the University recommends you have access to a Windows or Mac computer with webcam and broadband. Your best opportunity for success in the blended UB course delivery environment (in-person, hybrid, and remote) will require these minimum capabilities listed on the following website: `buffalo.edu/ubit/service-guides/hardware/getting-started-with-hardware/purchasing-or-using-an-existing-computer.html`. For this class, in addition, you will be required to download and install a free software – Apache Hadoop which is available from the following website: `https://hadoop.apache.org/releases.html`

Program Goal	Course Learning Outcomes	Assessment (Deliverable / Task)	Criteria for Success
Students will apply knowledge of management information systems to produce effective designs and solutions for specific problems	Learn about Distributed System Architectures and Communication Protocols	Quiz 1	Meets $\geq 85\%$ Marginally meets $75 - 84.9\%$ Fails to meet $< 75\%$
	Learn about Distributed Synchronization Consistency and Replication	Quiz 2	Meets $\geq 85\%$ Marginally meets $70 - 79.9\%$ Fails to meet $< 70\%$
Students will use new and emerging concepts and applications in proposing and creating IT solutions	Learn about Big Data Technologies Apache Hadoop MapReduce	Project	Meets $\geq 85\%$ Marginally meets $75 - 84.9\%$ to meet $< 75\%$

Table 2: Learning Outcomes

Course Schedule

Please see updated schedule on UBLearns every week. In general, course content covered every week is illustrated in Table 3.

Workload, Course Policies and Class Decorum

1. The project will comprise of 35% of the overall grade. This is intended to provide "hands-on" experience when dealing with real world problems. There will be two parts to the project - the first will deal with installation of Apache Hadoop and the second part testing out a basic program in the Hadoop framework.
2. There will be three quizzes (one each month) during the course - the best two will be used for grade estimation. Dates for the quizzes are fixed at the beginning of the semester and students are expected to make every effort to take them in-class. They will have multiple choice questions and/or short answer type questions or problems. Materials for quizzes will be non-inclusive – Quiz 2 will not include materials from Quiz 1 and so on. In the event a quiz is missed due to illness or other extraneous circumstances, please contact the instructor to make arrangements for missed work.
3. Students are expected to be well-versed in the student code of conduct available here: `http://www.buffalo.edu/news/key-issues/student-code-of-conduct.html`
4. There will be occasional labs during class (or in extra sessions, pre-scheduled). These labs will not be counted towards the final course grade, but are deemed helpful in doing the projects and becoming familiar with the Hadoop / MapReduce environments.
5. Plagiarism policy: We strictly abide by the policy as described in the UB handbook for undergraduate, MBA, and PhD students. No exceptions!

Week	Topics	Assignments
1	Course Overview, Fundamentals of Big Data	Instructor provided
2	Introduction to Distributed Systems	Tanenbaum, Ch. 1
3	Architectures	Tanenbaum, Ch. 2
4	Distributed File System	Tanenbaum, Ch. 11
5	Hadoop Distributed File System	Relevant Websites
6	MapReduce Programming	Tom White Ch. 2, Jimmy Lin - Intro
7	MapReduce Programming	Tom White Ch. 2, Jimmy Lin - Intro
8	MapReduce Algorithm Design	Jimmy Lin Ch 3
9	IR Algorithms for MapReduce	Jimmy Lin Ch 4
10	Dijkstra's Algorithm MapReduce	Jimmy Lin Ch 5
11	YARN	Websites
12	Apache Spark	Instructor provided
13	Apache Spark	Instructor provided
14	Apache Spark	Instructor provided
15	Apache Spark	Instructor provided

Table 3: Week-by-week breakdown of course content

(a) The UB plagiarism software on UBlearns will be used to verify your work – it has a high reliability 99% and students are expected to abide by the rules and write their own reports and solutions. If your work matches another student's or something on the internet, and there are no citations, you are liable to receiving a zero grade for the course.

(b) If you are found copying or resorting to unethical means during any quiz your paper will be taken and you will be provided a zero grade for the same.

(c) Policy on the use of Generative AI software (such as ChatGPT and related technologies) – Please acknowledge the use of Generative AI software in your project reports if you have made use of them. Please refer to the exact version used and demonstrate how you have made an attempt to integrate that information in your learning, and modified it for your purposes.

6. Dates for submission of project components (such as reports) and quizzes can be found from the schedule on UBlearns. All submissions are due at midnight on the date of submission. Late submissions are subject to 10% deductions in score.

7. To request for an extension on a deadline, please contact the instructor by email. However, the decision to grant such an extension is left to the discretion of the instructor.

8. **Weather Related Updates:** Please refer to the university's website for cancellations/delays due to weather or other unforeseen events: `http://emergency.buffalo.edu/`

Course Evaluation

Students are expected to provide feedback on the course at the end of the semester either directly on the UBCE portal or using the email notifications sent to this effect.

Course Materials Disclaimer

All materials prepared and/or assigned by me for this course are for the students' educational benefit. Other than for permitted collaborative work, students may not photograph, record, reproduce, transmit, distribute, upload, sell or exchange course materials, without my prior written permission. "Course Materials" include, but are not limited to, all instructor-prepared and assigned materials, such as lectures; lecture notes; discussion prompts; study aids; tests and assignments; and presentation materials such as PowerPoint slides, Prezi slides, or transparencies; and course packets or handouts. Public distribution of such materials may also constitute copyright infringement in violation of federal or state law. Violation of this policy may additionally subject a student to a finding of "academic dishonesty" under the Academic Integrity Policy and/or disciplinary charges under the Student Code of Conduct.

Public Health Compliance in Classroom Setting

As indicated in the Student Compliance Policy for COVID-19 Public Health Behavior Expectations (`https://www.buffalo.edu/studentlife/who-we-are/departments/conduct/coronavirus-student-compliance-policy.html`), in our classroom you are required to:

1. Obtain and wear masks/face coverings in campus public spaces, including campus outdoor spaces.
2. Maintain proper physical distancing in public spaces and must stay 6 feet apart from one another.
3. Stay home if you are sick.
4. Abide by New York State, federal and Center for Disease Control and Prevention (CDC) travel restrictions and precautionary quarantines.
5. Follow campus and public health directives for isolation or quarantine.
6. Should you need to miss class due to illness, isolation or quarantine, you are required to notify the course instructor and make arrangements to complete missed work.
7. You are responsible for following any additional directives in settings such as labs, clinical environments etc.

Students who are not complying with the public health behavior expectations will be asked to comply. Should the non-compliant behavior continue, course instructors are authorized to ask the student to leave the classroom. Non-compliant students may also be referred to the Office of Health Promotion to participate in an online public health class to better educate them on the importance of these public health directives for the entire community.

Diversity and Inclusiveness

Not respecting individual differences (for example culture, gender, sexual orientation, race, religion, disability status, or age) is perceived as intolerant. Not respecting others in general is a negative attribute and detrimental to the class. Examples of violations: not respecting or facilitating the input of individuals from different cultural backgrounds, making racist, sexist or otherwise insulting or derogatory comments about a class member; bullying a classmate or engaging in any actions that create a hostile learning environment. Please visit `http://www.buffalo.edu/inclusion/resources/IXResources.html`, which details resources, services, events and support related to equity and inclusion for students.

University Support Services

1. If in need for support for written work, students can access the University Support Services for e.g. the Center for Excellence in Writing. Several tutoring centers on campus provide academic success support and resources `http://www.buffalo.edu/writing/students/graduate.html`.

2. UB is committed to providing a safe learning environment free of all forms of discrimination and sexual harassment, including sexual assault, domestic and dating violence and stalking. If you have experienced gender-based violence (intimate partner violence, attempted or completed sexual assault, harassment, coercion, stalking, etc.), UB has resources to help. This includes academic accommodations, health and counseling services, housing accommodations, helping with legal protective orders, and assistance with reporting the incident to police or other UB officials if you so choose. Please contact UB's Title IX Coordinator at 716-645-2266 for more information. For confidential assistance, you may also contact a Crisis Services Campus Advocate at 716-796-4399.

3. Sexual Violence: UB is committed to providing a safe learning environment free of all forms of discrimination and sexual harassment, including sexual assault, domestic and dating violence and stalking. If you have experienced gender-based violence (intimate partner violence, attempted or completed sexual assault, harassment, coercion, stalking, etc.), UB has resources to help. This includes academic accommodations, health and counseling services, housing accommodations, helping with legal protective orders, and assistance with reporting the incident to police or other UB officials if you so choose. Please contact UB?s Title IX Coordinator at 716-645-2266 for more information. For confidential assistance, you may also contact a Crisis Services Campus Advocate at 716-796-4399.

   Please be aware UB faculty are mandated to report violence or harassment on the basis of sex or gender. This means that if you tell me about a situation, I will need to report it to the Office of Equity, Diversity and Inclusion. You will still have options about how the situation will be handled, including whether or not you wish to pursue a formal complaint. Please know that if you do not wish to have UB proceed with an investigation, your request will be honored unless UB's failure to act does not adequately mitigate the risk of harm to you or other members of the university community. You also have the option of speaking with trained counselors who can maintain complete confidentiality. UB?s Options for Confidentially Disclosing Sexual Violence provides a full explanation of the resources available, as well as contact information. You may call UB?s Office of Equity, Diversity and Inclusion at 716-645-2266 for more information, and you have the option of calling that office anonymously if you would prefer not to disclose your identity.

4. Mental Health: As a student you may experience a range of issues that can cause barriers to learning or reduce your ability to participate in daily activities. These might include strained relationships, anxiety, high levels of stress, alcohol/drug problems, feeling down, health concerns, or unwanted sexual experiences. Counseling, Health Services, and Health Promotion are here to help with these or other issues you may experience. You learn can more about these programs and services by visiting `https://www.buffalo.edu/studentlife/who-we-are/departments.html` or by contacting:

Service	Address	Phone No.
Counseling Services	120 Richmond Quad (Ellicott Complex, North Campus)	716-645-2720
Health Services	`https://www.buffalo.edu/studentlife/who-we-are/departments/health.html`	716-829-3316
Health Promotion	114 Student Union (North Campus)	716-645-2837

A list of significant publications that may be covered in the course

References

[1] J. Dean and S. Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters", *Proc. of Sixth Symposium on Operating System Design and Implementation(OSDI)*, 2004.

[2] S. Ghemawat, H. Gobioff, S.T. Leung. "The Google File System", SOSP 2003.

[3] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fiukes, and R.E. Gruber. "BigTable: A Distributed Storage System for Structured Data", OSDI, 2006.

[4] C. Chu, S. K. Kim, Y. Lin, Y.Y. Yu, G. Bradski, A. Y. Ng, K. Olukotun (Stanford). "Map-Reduce for Machine Learning on Multicore", NIPS 2006.

[5] Matei Zaharia, Andy Konwinski, Anthony D. Joseph, Randy Katz, and Ion Stoica, "Improving MapReduce Performance in Heterogeneous Environments", OSDI, 2008.