

# Large Scale Computing on the Cloud

2 Credits

BU.330.740.T2

Tuesday, 6:00 pm-9:00 pm

1/21/2025-3/11/2025

Spring 2025

Washington, DC

## Instructor

Minghong Xu, PhD

## Contact Information

[xu.minghong@jhu.edu](mailto:xu.minghong@jhu.edu)

## Office Hours

Mondays and Thursdays 11:30am - 1:30pm, and by appointment

## Teaching Assistant

Boxi Jiao [bjiao1@jh.edu](mailto:bjiao1@jh.edu)

Zibo Lin [zlin71@jh.edu](mailto:zlin71@jh.edu)

Biao Xiang [bxiang2@jh.edu](mailto:bxiang2@jh.edu)

## Required Texts & Learning Materials

There is *no required* textbook: all class materials will be available on our Canvas website. However, some books are very useful if you want to learn more about mining large-scaled datasets.

## Recommended Texts

* **Mining of Massive Datasets**, by Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman;

Publisher: Cambridge University Press (2011), ISBN-13: 978-1107015357

## Technology Requirements

* Access to AWS (Amazon Web Services) Academy will be provided by the instructor for large scale data applications. Instructions on AWS Academy accounts will be provided on Canvas course website.

## Course Description

Internet of Things (IoT) is connecting almost all the components together in every aspect of business and our daily life. As a result, huge amount of data is being generated. The term “big data” implies the large scale of data that cannot be stored on one single computer. The analyses of such large-scaled data usually require massively parallel software running on tens, hundreds, or even thousands of servers. Enterprise technology managers are often called upon to organize large-scale data repositories, to deploy AI/ML projects at scale on the cloud, and to support decision-making based on information that resides in distributed data sources.

This course prepares students with the fundamental concepts of distributed systems in the cloud and algorithms for mining massive datasets. It equips students with advanced techniques to extract the value from the large-scaled data generated and collected in everyday business life. The course uses a hands-on, learning-by-doing approach to practice on AWS platform. Topics include: MapReduce model, distributed file system (HDFS), advanced MapReduce (Spark), scaled machine learning project pipelines, recommender systems, computer vision, and natural language processing on the cloud.The focus is on creating awareness of the technologies, allowing some level of familiarity with them through assignments, and enabling some strategic thinking around the use of these in business.

## Prerequisite(s)

None

## Learning Objectives

By the end of this course, students will be able to:

1. Learn to manage large enterprise-wide data repositories to support corporation wide analytics efforts.
2. Conduct large-scale data analytics using scientific methods, make appropriate and powerful connections between techniques and real-world problems.
3. Demonstrate sophisticated understanding of the concepts and methods; know the exact scope and possible limitations of each method; show the capability of using large data warehouses, and distributed data repositories to provide constructive guidance in business decision making.
4. Use advanced techniques to conduct thorough and insightful analysis, drawing on distributed enterprise data repositories and interpret the results correctly to deliver useful information in support of decision-making
5. Develop the capabilities of making better business decisions by using advance techniques in large-scale data analytics.
6. Demonstrate substantial understanding of the real problems; link analytical methods to problem types; draw reasonable conclusions with sufficient explanation and elaboration.
7. Write an insightful and well-organized report for a real-world case study, including thorough and thoughtful details.

To view the complete list of the Carey Business School’s general learning goals and objectives, visit the [Carey website](https://carey.jhu.edu/faculty/resources/teaching-learning/learning-assessment).

Attendance  
Attendance and class participation are part of each student’s course grade. Students are expected to attend all scheduled class sessions. Failure to attend class will result in an inability to achieve the objectives of the course. Excessive absence will result in loss of points for participation. Regular attendance and active participation are required for students to successfully complete the course.

Class participation is an important part of learning. If you have a question, it’s likely that others do as well. I encourage *active* participation, and course grades will take into account students who make particularly strong contributions.

## Assignments

| **Assignment** | **Group or Individual** | **Learning Objectives** | **Weight** |
| --- | --- | --- | --- |
| Attendance and participation in class discussion | Individual | 1, 2, 3, 4, 5, 6 | 5% |
| Homework | Individual | 1, 2, 3, 4, 5, 6 | 40% |
| Project | Group | 1, 2, 3, 4, 5, 6, 7 | 30% |
| Final Exam | Individual | 1, 2, 3, 4, 5, 6 | 25% |
| Total |  |  | 100% |

*Homework*: there are four individual homework assignments. Please refer to the tentative schedule table below. All homework assignments should be submitted through the Canvas links.

*Group Projects:* 4 students form a group and work on the projects as a team. Students can identify a company or a scenario and its big data needs, and use techniques taught in class. Students are required to write a project report, and present in class using Power Point slides. Details and rubrics will be available on Canvas course website.

*Final Exam*: the final exam is individual exam. It will be administered via Respondus LockDown Browser (<https://download.respondus.com/lockdown/download.php?id=123533816>).

*Late submission* including assignments, projects and exams will *not* be accepted.

## Grading

The grade of A is reserved for those who demonstrate extraordinary performance as determined by the instructor. The grade of A- is awarded only for excellent performance. The grades of B+ and B are awarded for good performance. The grades of B-, C+, C, and C- are awarded for adequate but substandard performance. The grades of D+, D, and D- are not awarded at the graduate level. The grade of F indicates the student’s failure to satisfactorily complete the course work. For Core/Foundation courses, the grade point average of the class should not exceed 3.35. For Elective courses, the grade point average should not exceed 3.45.

## Tentative Course Calendar

Instructors reserve the right to alter course content and/or adjust the pace to accommodate class progress. Students are responsible for keeping up with all adjustments to the course calendar.

| **Week** | **Topic** | **Hands-on Learning** | **Due** |
| --- | --- | --- | --- |
| 1 | Course introduction  Overview of cloud platforms and AWS |  |  |
| 2 | Big Data framework: MapReduce  Frequent itemset mining  Optional: HDFS and YARN | Lab 1: AWS S3, EC2 and EMR | Project team due |
| 3 | Advanced Big Data framework: Spark  Data Engineering  Optional: Hive and Apache Pig | Lab 2: pySpark in Google Colab | HW 1 due |
| 4 | Scaled machine learning pipeline  Recommender system | Lab 3: AWS SageMaker | HW 2 due |
| 5 | Mining of massive datasets  Computer vision and business applications | Lab 4: AWS Rekognition | HW 3 due |
| 6 | Natural language processing on cloud  AI-assisted programming  Project preview and final review | Lab 5: AWS Q Developer | HW 4 due |
| 7 | Project presentation |  | Project report due |
| 8 | Final exam |  |  |



## Carey Business School Policies and General Information

Please note that failure to become acquainted with Carey policies will not excuse any student from adhering to these policies.

### Canvas Site

A Canvas course site is set up for this course. Each student is expected to check the site throughout the semester as Canvas will be the primary venue for outside classroom communications between the instructor and students. Students can access the course site at <https://canvas.jhu.edu/>.

### Technical Support

24/7 technical support for questions regarding Canvas, Zoom, and other technical issues is available. Please refer to Carey’s [Academic Resources webpage](https://carey.jhu.edu/student-experience/academic-resources) for contact information and other details.

### Students with Disabilities - Accommodations and Accessibility

Johns Hopkins University values diversity and inclusion. We are committed to providing welcoming, equitable, and accessible educational experiences for all students. Students with disabilities (including those with psychological conditions, medical conditions, and temporary disabilities) can request accommodations for this course by providing an Accommodation Letter issued by [Student Disability Services](https://carey.jhu.edu/student-experience/services-resources/student-disability-support-services). Please request accommodations for this course as early as possible to provide time for effective communication and arrangements. For further information or to start the process of requesting accommodations, please contact [Student Disability Services](mailto:carey.disability@jhu.edu) at the Carey Business School.

### Academic Ethics Policy

Carey expects graduates to be exemplary global citizens in addition to innovative business leaders. The Carey community believes that honesty, integrity, and community responsibility are qualities inherent in an exemplary citizen. The objective of the Academic Ethics Policy (AEP) is to create an environment of trust and respect among all members of the Carey academic community and hold Carey students accountable to the highest standards of academic integrity and excellence.

It is the responsibility of every Carey student, faculty member, and staff member to familiarize themselves with the AEP and its procedures. Failure to become acquainted with this information will not excuse any student, faculty, or staff member from the responsibility to abide by the AEP. Please contact the [Office of Student Affairs](mailto:carey.student@jhu.edu) if you have any questions. For the full policy, please visit the [Academic Ethics Policy webpage](https://carey.jhu.edu/student-experience/school-policies/academic-ethics-policy).

### Student Conduct Code

The fundamental purpose of the Johns Hopkins University’s regulation of student conduct is to promote and to protect the health, safety, welfare, property, and rights of all members of the University community as well as to promote the orderly operation of the University and to safeguard its property and facilities. Please contact the [Office of Student Affairs](mailto:carey.student@jhu.edu) if you have any questions regarding this policy. For the full policy, please visit the [Student Conduct Code webpage](https://studentaffairs.jhu.edu/policies-guidelines/student-code).

### Commitment to Respect

Respectful behavior creates an environment within the Carey Business School where all are valued and can be productive. Carey defines respectful behavior as conduct that, at a minimum, demonstrates consistent courtesy for others, including an effort to understand differences. As such, all in the community agree to the Carey Commitment to Respect, which states that we all strive to show that we value each other’s human dignity and our differences, and to choose behavior and language that demonstrates mutual respect. Please visit the [Commitment to Respect webpage](https://carey.jhu.edu/student-experience/school-policies/carey-business-school-community-commitment-respect) to learn more about the expectations and resources available.

### Classroom Policies for All On-Site and Remote-Live Classes

Carey is committed to maintaining the highest standards of excellence in all forms of instruction. To that end, we have developed [policies and procedures for all classes offered in on-site and remote-live formats](https://carey.jhu.edu/student-experience/school-policies/policies-procedures-on-site-remote-live-classes). These policies will govern all courses occurring in these formats, and all students are expected to familiarize themselves with and adhere to these policies.

### Student Success Center

The Student Success Center offers assistance in core writing and quantitative courses. For more information, visit the [Student Success Center webpage](https://carey.jhu.edu/student-experience/academic-support/student-success-center).

### Other Important Policies and Services

Students are encouraged to consult the [Student Handbook and Academic Catalog](https://carey.jhu.edu/student-experience/services-resources/student-handbook) and [Student Services and Resources](https://carey.jhu.edu/student-experience/services-resources) for information regarding other policies and services. For your convenience, there is a singular website students can visit to learn about all [JHU and Carey policies](https://carey.jhu.edu/student-experience/policies).

### Copyright Statement

Unless explicitly allowed by the instructor, course materials, class discussions, and examinations are created for and expected to be used by class participants only. The recording and rebroadcasting of such material, by any means, is forbidden. Violations are subject to sanctions under the [Academic Ethics Policy](https://carey.jhu.edu/student-experience/school-policies/academic-ethics-policy).