DSCI - 644

Software Engineering for Data Science

Instructor: Zimeng Lyu

Term: Spring 2025

Lecture: Tuesday/Thursday 12:30 - 1:45 pm
Location: Golisano 2590
Office Hour: Tuesday/Thursday 2 - 3 pm (Suggested, will finalize after add/drop deadline)
Zoom: https://rit.zoom.us/j/3395070767
Email: zimenglyu@mail.rit.edu
Slack: Invite link

Course Description

This course focuses on the software engineering challenges of building scalable and highly available big data software systems. Software design and development methodologies and available technologies addressing the major software aspects of a big data system including software architectures, application design patterns, different types of data models and data management, and deployment architectures will be covered in this course.

Prerequisite

SWEN-601 (Foundations of Data Science and Analytics) and DSCI-633 (Software Construction) or equivalent courses.

Course Topics

- Introduction to big data systems
- ETL (Extract, Transform, and Load) data pipelines
- Data cleaning and aggregation
- Relational and No-SQL database
- Data storage and retrieval
- Distributed data system

- Data replication
- Database schemas
- Batch processing and stream processing pipeline
- Security in data systems and other important topics

Course Outcomes

- Students will gain the ability to design and implement scalable ETL (Extract, Transform, and Load) data pipelines.
- Students will understand how to develop microservices-based architectures for data-intensive applications.
- Students will be able to implement distributed data processing systems, including data partitioning and distributed transaction management.
- Students will gain the ability to design and implement batch processing pipelines for online machine learning models, including data cleaning, transformation, and real-time visualization.
- Students will understand the software engineering challenges and best practices for building big data systems.

Textbook, Required Material and Other Resources

Textbooks:

- Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems by Martin Kleppmann.
- Rebuilding Reliable Data Pipelines Through Modern Tools by Ted Malaska
- Data Engineering with Python by Paul Crickard.
- Big Data: Principles and best practices of scalable realtime data systems by James Warren and Nathan Marz.
- Learning Spark by Jules S Damji, Brooke Wenig, Tathagata Das, and Denny Lee.

RIT Students can get electronic versions of some textbooks from O’Reilly Media, with RIT credentials.

Software and Hardware

Programming assignments will be done in the Python programming language, primarily using Spark family packages.

Method of Instruction

The course will be taught in person only.

Schedule and Time Commitment

Each course week runs from Monday to Sunday midnight each week. To succeed in this course plan to spend 9 to 12 hours each week:

- Reviewing lecture notes and reading the course text.
- Developing source code for the programming assignments.
- Working on individual projects.

Attendance Policy

Attendance is required for the on campus courses. All requests for excused absences are accepted only by email to your instructor before the beginning of class. Deductions to your final grade will be applied as follows:

Number of Missed Classes	Reduction to Final Grade
1	2%
2	5%
3	10%
4	20%
5	Five or more missed classes will result in a failing grade.

Grading

Assignments

The course grade will consist of four programming assignments, one midterm exam and one final exam. There will not be in-class quizzes.

Grade Breakdown

The grade will be calculated as follows:

Assignment	Percentage of Final Grade
4 Projects	17% each
Midterm and Final Exam	16% each

Grade Letter Determination

Grades are awarded as follows:

Grade	Percentage Range
A	>= 93
A-	>= 90 and < 93
B+	>= 87 and < 90
B	>= 83 and < 87
B-	>= 80 and < 83
C+	>= 77 and < 80
C	>= 73 and < 77
C-	>= 70 and < 73
D	>= 60 and < 70
F	< 60

Policies

Project Late Penalty

Deadlines are subject to change if communicated at least 3 days in advance.
Assignments must be submitted by the specified deadline. Late submissions will incur a penalty and be deducted from the assignment score. Assignments submitted more than 2 days after the deadline will not be accepted and will receive a score of zero.

Academic Integrity

As an institution of higher learning, RIT expects students to behave honestly and ethically at all times, especially when submitting work for evaluation in conjunction with any course or degree requirement. RIT Online encourages all students to become familiar with the RIT Honor Code and with RIT's Academic Integrity Policy.

Any copied text or figures without proper citation in and written or presented documents will be considered plagiarism and will be treated as such per RITs policies. A first offense will result in a

0 on the assignment, and a second offense will result in failure of the course and potential removal from the program.

Programming assignments are to be completed individually. Under no circumstances should code be shared between students. Any copied source code between students will be flagged as cheating, with those assignments receiving a 0. Depending on severity, further action may be taken, such as a referral to the student's department chair and advisor.

Further, any source code utilized in the programming assignments and projects which has been taken from an online source must be cited within code comments and a hyperlink provided to the website source. Failure to do so will be treated as plagiarism, with similar consequences as to copying source code as described above.

ChatGPT and other large language models are only allowed for assisting your studies but not intended for completing assignments for you. Students are fully responsible for the submitted assignments. Using such models to finish assignments will be considered plagiarism.

Reasonable Accommodations

RIT is committed to providing reasonable accommodations to students with disabilities. If you would like to request accommodations such as special seating or testing modifications due to a disability, please contact the Disability Services Office. It is located in the Student Alumni Union, Room 1150; the Web site is http://www.rit.edu/dso. After you receive accommodation approval, it is imperative that you see me during office hours so that we can work out whatever arrangement is necessary.

Use of copyrighted material

Certain materials used in this course are protected by copyright and may not be copied or distributed by students. You can find more information at http://www.rit.edu/academicaffairs/policiesmanual/sectionC/C3_2.html.

Sharing Protected Information

When sharing copyrighted content on the internet with your classmates, please make sure that you link to a legal source. Repeated access to illegal sources may cause you or your classmates to receive warnings through the Copyright Alert System, as well as possible downgrades in internet service. You can find out more about the Copyright Alert System at http://www.copyrightinformation.org/the-copyright-alert-system/what-is-a-copyright-alert/.

Emergencies

In the event of a University-wide emergency course requirements, classes, deadlines and grading schemes are subject to changes that may include alternative delivery methods, alternative methods of interaction with the instructor, class materials, and/or classmates, a revised attendance policy, and a revised semester calendar and/or grading scheme.

Student support availability

Student Learning, Support and Assessment offers a wide range of programs and services to support student success including the Academic Support Center, College Restoration Program, Disabilities Services, English Language Center, Higher Education Opportunity Program, Spectrum Support program, and TRiO Support Services. Students can find out about specific services and programs at www.rit.edu/slsa.