Spring 2023 DS-GA 1004 Big Data Syllabus

Vital information

- Instructors:
  - Brian McFee, brian.mcfee@nyu.edu
  - Pascal Wallisch, pascal.wallisch@nyu.edu
- Office hours:
  - Brian: Tuesdays from 9am-11am (NYC time) via zoom.
  - Pascal:
    In person: We 2.15-3.15 pm, Th 2-3 pm in Room 210 (60 FA)
    Remote: Fr: 11.30 *pm* - 12.30 *am*
- Class meetings: Monday, 18:45-20:25, GCASL-C95
- The final exam is on 2023-05-12 (Friday) at 6pm-7:50pm in GCASL C95

What is Big Data all about?

This class will introduce you to modern tools for working efficiently with large datasets. Specifically, we will cover:

- Relational databases and SQL
- Distributed storage
- Distributed computation frameworks
- Applications and algorithms

Course structure

Grading

Your grade for the class will be determined by the following break-down:

- 25% Lab programming assignments
- 25% Quizzes
- 25% Project
- 25% Final exam (Date TBA via Albert)

Additionally, the following policies will apply:

- Lab assignments must be completed by each student **individually**.
- For lab assignments, you will have two (2) slip days which can be used however you see fit over the entire semester. For example, you can turn in one assignment two days late, or two assignments one day late each, without penalty. If you want to use a slip day for a given lab, please email us about that desire.
- Beyond slip days, late assignments will incur a 20% point reduction for each additional day late. No assignment will be accepted more than 5 days after the due date.
    - Exceptions to the above may be granted in case of genuine emergencies, but contact the instructor as soon as possible.
- Quizzes will be administered remotely through Brightspace. Quizzes will be available for a 24-hour period, but you will have 1 hour to complete each quiz once you begin.
- Your lowest quiz grade will be dropped automatically.
- Quizzes are open-book and open note. **Quizzes must be completed individually and without collaboration.** Treat these as if they were in-class exams.

Course policies

If you need to get in touch with the instructional staff, there are many ways to do so.

- **Miscellaneous policy questions**: e.g., *when are quizzes? how do assignment due dates work?* etc. Please re-read this document first. If your question is still not addressed, please use the discussion forum.
- **Help with assignments or course topics**: post on the discussion forum, or sign up for office hours.
- **Anything sensitive or confidential**: e.g., health issues, emergencies, etc. Email the instructors.
- **Anything else**: We're happy to talk with students during office hours about various topics, related to the course or not.

Class environment

It is our job to make this course inclusive and equitable for all students. Please do your part by seeking to promote the success of others, and by treating each other in ways that respect and celebrate the diversity of talent that is drawn to this field. Here are a few specific things that you should know about my policies on creating an inclusive and equitable class environment (both in the classroom and on the course website/forum):

- **Preparation**: Students come to this class from a wide range of backgrounds, and greatly varying previous exposure to mathematics, programming, and data science more generally. I want to assure students who may feel out of place here that you are indeed prepared to succeed in this class! If you feel that there are gaps in your knowledge, please speak with the course staff and we will help you find additional materials as needed.
- **Classroom environment**: For some reason, it is common in technical or programming-oriented classes that some students ask "questions" that are not really questions so much as opportunities to demonstrate knowledge of jargon, or facts that are beyond the scope of the topic at hand. This can have discouraging effects on other students who may not be familiar with those terms, and worry that this indicates that they are less prepared to do well in the class. (Note: this is rarely the case: knowing terms outside the scope of the course is not a good predictor of success.) If you find yourself wanting to make such a question or comment in lecture, we encourage you to
consider whether office hours would be a better venue for exploring that topic. The course staff are more than happy to discuss tangentially related topics in office hours, when they would not distract from lecture or alienate other students.
- **Accessibility**: If you have any accessibility requirements, please present a letter from the Moses Center to us at your earliest convenience, so that we can ensure that materials and staff comply with your needs. We are always willing to do what it takes to support you, but we ask that you have requests submitted no later than 1 week prior to the assessment in question so that we have sufficient time to make any necessary arrangements.
- **Names and pronouns**: If you have a name and/or pronoun that doesn't match the class roster delivered from the registrar, please let us know and we will ensure that you are addressed correctly in our class. You are always welcome to use your preferred form of address on all class assignments and exams; just be sure to include your NYU netID number to ensure that we can link records properly.

Academic integrity and honesty

All students are expected to do their own work. Students may discuss assignments with each other, as well as with the course staff. Any discussion with others must be noted on a student's submitted assignment. Excessive collaboration (i.e., beyond discussing the assignment) will be considered a violation of academic integrity. Questions regarding acceptable collaboration should be directed to the class instructor prior to the collaboration. It is a violation of the honor code to copy or derive solutions from other students (or anyone at all), textbooks, previous instances of this course, or other courses covering the
same topics. Copying solutions from other students, or from students who previously took a similar course, is also clearly a violation of the honor code. Finally, a good point to keep in mind is that you must be able to explain and/or re-derive anything that you submit. This is particularly important if you should adapt solutions from online sources.

Please also refer to the general NYU academic integrity statement.

AI policy

We live in the age of viable generative AI. Banning these tools is neither realistic, nor desirable. In fact, learning to use these tools is an emerging skill. Note that AI tools do not always produce correct or accurate results. In addition, it is unwise to rely on them too much. There are situations where you won't have access to these tools, for instance during technical interviews. In addition, there are also skills someone with an advanced degree in Data Science is just expected to have on tap - without AI assistance or looking anything up. To integrate both considerations, you can use generative AI tools to do the assignments in this class, *except* the final exam (which will need to be done entirely without any electronic devices of any kind). If you use an AI to guide you in completing an assignment, you have to disclose which parts were generated by the AI.

Technology infrastructure

During this class, students are encouraged (but not required) to use Github Classroom as a part of course studies, and thus, will be required to agree to the Terms of Use (TOU) associated with
Marketplace Simulations. Github Classroom requires users to be over the age of 18. Personally identifiable information is required to create an account. This information includes name, email address, and IP address. This information will identify users to Github and companies with whom it shares data. Github Classroom is not an NYU service. Therefore, the user should not use their NYU login and password. Login and password information should be unique.

You should read carefully the Github Terms of Use and Privacy Policy regarding the impact on your privacy rights and intellectual property rights. If you have any questions or objections regarding those Terms of Use or the impact on the class, you are encouraged to speak to the instructor prior to enrollment.