Sittyba
Big Data (DS-GA 1004)
Spring Semester 2024

§0.0 Purpose, design & philosophy (PDP): Bigger datasets improve all aspects of data science, from significance testing to modeling to machine learning. The good news is that we are awash in data, and the quantity of data available is still growing rapidly. This abundance poses new challenges as to how to organize, curate and manage ever bigger datasets. In short, this class is designed to impart the skillsets that will allow you to help you meet these challenges and – in turn – to handle the data. This ability is essential to fulfill the core mission of the Center for Data Science as well as the promise of Data Science as a field: Being able to harness the power of big datasets.
Specifically, we will introduce a variety of commonly used tools, approaches and algorithms to store and process large datasets in a distributed fashion. We will also highlight some of the applications – from search to recommendation that are made possible by these methods. Finally, we will end on a discussion of social and ethical considerations brought about by big data.

§1.0 Instructor (and office hour information):

Pascal Wallisch, PhD
Tu: 2-3 pm (Walk-ins welcome, first come, first serve – if there is a line, take a fox stick and wait) We: 2.15-4.15 pm (Walk-ins welcome, first come, first serve – if there is a line, take a fox stick)

§1.1 Lab leaders (and their office hours – in person, remote or per Calendly – per link):

Andrew Deur: Tuesday 5 – 6 pm
Aman Singhal: Calendly
Shreemayi Sonti: Monday 9 – 10 am
Sharad Dargan: Thursday 3 – 4 pm

§1.2 Tutors:

Avinav Goel Calendly for one-on-one sessions · Email: nyuBigData@gmail.com

§1.3 Class email and discord: nyuBigData@gmail.com / https://discord.gg/KWf256P2Mh

§1.4 Lecture times: Monday 6:45 - 8:25 pm (100 minutes)

§1.5 Lecture space: GCASL, C95 (mirrored in https://nyu.zoom.us/j/93138358950)

§1.6 Lab time: Throughout the week (50 minutes - See Albert for your section)

§1.7 Lab space: See Albert for your section information

§1.8 Session types: There are 2 kinds of sessions each week. In the lecture, we will introduce new course content in a conceptual fashion. Throughout the week, there are smaller lab sections where we will implement and practice these concepts, often in code.
So if you want to follow along with the code, bring a laptop to the lab sessions.

§1.9 Readings: There are weekly readings of classical papers on Big Data concepts. You can find them in the suitable Brightspace folder (there will be a folder for every week).

§2.0 Course grading: The total grade is calculated based on the following components

1) After action appraisals (12)	1% each	12 % total
2) Final Interview & Skills Test (1, cumulative)	25%	25 % total
3) Capstone project (1, cumulative)	25%	25 % total
4) Homework assignments (5)	5% each	25 % total
5) Quizzes (6)	2% each	12 % total
6) Intake survey	0.5% each	0.5% total
7) Exit survey	0.5% each	0.5% total
	Total	100%

§2.1 Grade cutoffs:

A	95-100	B+	87-89.9	C+	77-79.9	D+	65-69.9	F	30-59.9
A-	90-94.9	B	83-86.9	C	73-76.9	D	60-64.9	I	0-29.9
		B-	80-82.9	C-	70-72.9				

§2.2 Attendance and Participation: Note that this class has been designed with live, interactive lectures in mind, so please attend the lectures live when you can. Historically, live attendance adds most value to student learning. The recordings and online delivery of lectures are only meant as a backup in cases where you are unable to attend due to illness or other issues.

In any case, you are responsible for the material covered in this course. Thus, consistent attendance is critical, as the final will focus on the material discussed during lecture and labs will be crucial to implement the subject material in code. Thus, we assign a participation grade with the AAA assignments. There are 14 lectures, and you need to complete 12 of these AAA assignments to get a full participation score.

§2.3 Quizzes: Delivered online via Brightspace, probe understanding of course content. Quizzes will be available for a 48-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.

§2.4 Homework assignments: These are more extensive assignments, and based on the content covered in the lab. You can do them in small groups of 1-3 students. This will also incentivize you to be a good team member, if you want to have colleagues for the capstone project.

Note that we will use Github classroom for the code part of this assignment (again, to resemble industry conditions), so you will need to sign up for a Github account.

§2.5 Final Interview & Skills test: The final is cumulative and comprehensive, meaning that it covers all aspects of class content. It will consist entirely of true/false questions that could come up in technical interviews. You are allowed to bring one sheet (double sided) of notes. Importantly, this will be in class, and no electronic devices of any kind will be allowed. Note: As this is an in-person final, and we only know that it will be during finals week but not yet when during finals week, **make travel arrangements accordingly**. If you miss it, you'll have to live with an incomplete as the class grade until we can resolve it.

§2.6 Capstone project: Is designed to tie together the skills you learned in this class. Like for the lab programming assignments, the capstone project will be done in small teams of 1-3 people and uses datasets we provide, in order to mimic real life industry conditions. The deliverable work product is a 5-10 page project report. We will release a spec sheet as to what we would like to see in this report towards the end of the semester. The datasets and the questions should spark joy. The project report is also something you will be able to use in your portfolio for job applications.

Like in the lab programming assignments, we will ask you to push your code via Github classroom.

§3.0 COURSE SCHEDULE

Week	Date	Topic	Assignment
1	01/22	Welcome to a big data world	Intake survey. Due: 02/05
2	01/29	Principles of relational databases: SQL	Homework 1. Due: 02/06
3	02/05	Map-Reduce	Quiz 1. Due: 02/11
4	02/12	HDFS	Homework 2. Due: 02/27
5	02/26	Big Data Infrastructure	Quiz 2. Due: 03/03
6	03/04	Spark	Homework 3. Due: 03/19
7	03/11	Column-oriented storage	Quiz 3. Due: 03/17
8	03/25	Dask	Homework 4. Due: 04/09
9	04/01	Search	Quiz 4. Due: 04/07
10	04/08	Guest lecture: Big data in the wild	Homework 5. Due: 04/23
11	04/15	Graph algorithms	Quiz 5. Due: 04/21
12	04/22	Recommender systems with big data	NA
13	04/29	Ethical and social implications	NA
14	05/06	Parallelization DAU: Of GPUs and clouds	NA

Note that Monday, February 19th is a University Holiday
Note that Spring Break is from March 18th to March 22nd

Capstone project – Release date: April 1st. Due date: May 6th.
Final: During finals week. Specifics will be announced once the registrar releases it

§4.0 Course policies

§4.1 Grace days: For homeworks, you will have an automatic grace day by which you can turn the assignment in late, without penalty. After that, late assignments will incur a 1% point reduction for each hour that it is late, as late assignments are not worthless, but worth less. Again, this is to mimic industry conditions. Time is of the essence.

§4.2 Communication: 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 on discord.
Help with assignments or course topics: sign up for office hours.
Anything sensitive or confidential: e.g., health issues, emergencies, etc.: Email the course email
Anything else: We're happy to talk with students during office hours about various topics, related to the course or not.

§4.3 General classroom protocol: Be excellent to each other. As this is a masters level class, the demeanor I expect from you (and will also aim to exhibit myself) is benevolent professionalism.

§4.4 Academic integrity and honesty:

All students are expected to do their own work in assignments that are flagged as such (quizzes, final, AAA). 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 if you have any doubts about any part of this statement .

§4.5 AI policy

We live in interesting times. Specifically, we now 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, as they are explicitly trained to say things that they think you want to hear. Hallucinations abound and are unlikely to be tamed with any of the currently popular approaches to generative AI, for a variety of reasons. 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 – particularly during emergencies. To integrate both considerations, you can use generative AI tools to do the assignments in this class, *except* the final (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.

§4.6 Technology infrastructure

During this class, students will 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 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, you are encouraged to speak to the instructor prior to enrollment.

§4.7 Prerequitisites

I do not presume that you know anything about high performance computing, cluster computing, cloud computing or anything like that. We'll start from scratch.

§4.8 Attendance

NYU is not an online school, so in-person attendance is expected. Students are responsible for anything that they might miss, so the recordings are intended to aid with that.