Syllabus

Course and contact information

Course: DATS 6450.13 (Applied Big Data Analytics) **Semester:** Spring, 2026 **Meeting time:** Thursdays 6:10-8:40pm **Location:** Philips 416

Instructor

Name: Abhijit Dasgupta, PhD
Phone: 240-813-8458
GW E-mail: abhijit.dasgupta@gwu.edu
Office hours:
In-person: Thursdays 5:10-6:10pm, Philips 416
Remote: By appointment (via link)
Communications: via Slack

> ! Communication policy
>
> All general class-related issues, including discussion of material, should be done on Slack. Any private issue, including absences, illness, issues around grades, etc. should be done via e-mail

---

Course prerequisites: DATS 6101, 6102, Python

Students completing the course will be able to:

- Build data pipelines and transformations using DuckDB and PySpark.
- Use concepts of parallelization, embarrasingly parallel problems, and applications in Python to process large data

- Compare computational efficiency, memory management, and scalability between single-node, cluster and distributed frameworks.
- Optimize queries, caching, and partitioning strategies across engines.
- Apply datashader or equivalent tools to visualize large datasets efficiently.
- Develop end-to-end reproducible analytics workflows integrating both systems.

Average amount of direct instruction or guided interaction with the instructor and average minimum amount of independent (out-of-class) learning expected per week:

This class will meet for 2.5 hours of in-person learning per week. It is expected that homework, project and independent study will take an average of 8-10 hours weekly. We note here that classes are in-person only and no remote option will be available during the semester. It will be the student's responsibility to make up work during an absence. Class attendance will be noted and class participation is encouraged to enable active learning of the material.

Required textbooks and/or other materials and recommended readings: There are no required textbooks for this class. Relevant readings will be assigned on a weekly basis

Class policy on the use of AI

Generative Artificial Intelligence (GAI) tools such as ChatGPT are becoming important resources in many fields and industries. Accordingly, you are permitted to use such tools to generate content submitted for evaluation in this course, including assignments, with the limitation that no more than 50% of your submitted code can be generated using AI tools. You remain responsible for all content you submit for evaluation, and to ensure that the material submitted runs and produces correct results. Material for final presentation and reports must be substantially your own work, with GAI tools available to help with brainstorming, editing, organization, polish.

You may use GAI tools to help generate ideas and brainstorm. However, you should note that the material generated by these tools may be inaccurate, incomplete, or otherwise problematic. Beware that use may also stifle your own independent thinking and creativity.

If you include content (e.g., ideas, text, code, images) that was generated, in whole or in part, by Generative Artificial Intelligence tools (including, but not limited to, ChatGPT and other large language models) in work submitted for evaluation in this course, you must document and credit your source. For example, text generated using ChatGPT-4 should include a citation such as: "ChatGPT-4. (YYYY, Month DD of query). 'Text of your query.' Generated using OpenAI. https://chat.openai.com/." Material generated using other tools should be cited accordingly. Failure to do so in this course constitutes failure to attribute under the George Washington University Code of Academic Integrity.

Note on code plagiarism

We will check for plagiarism of submitted code between students. We understand that there are common templates and structures in code that will naturally be common between students, but everyone does typically have their own coding style, naming conventions, comments and approaches. Any pair of submissions that appear to have more than 70% code in common will be subject to further review. Code that is over 90% common will be considered a potential violation of the Honor Code.

Schedule of topics:

Week	**Date**	**Topic**	**Lab**
1	2026-01-15	Course overview	Bash shell
3	2026-01-22	Parallel and distributed computing	Using multiprocesing and asyncio
2	2026-01-29	Introduction to Cloud Computing	AWS setup
4	2026-02-05	introduction to DuckDB and Polars	
5	2026-02-12	Advanced DuckDB operations	
6	2026-02-19	Python tools for larger data	pandas, dask, Ray, RAPIDS
7	2026-02-26	Developing workflows for big data (Midterm 1)	
8	2026-03-05	Apache Spark Fundamentals	Spark RDD
9	**2026-03-12**	**Spring Break**	
10	2026-03-19	Spark DataFrames and Spark SQL	
11	2026-03-26	Spark ML and Streaming	MLib
12	2026-04-02	Spark NLP	John Snow Labs
13	2026-04-09	Visualization and reporting large data (Midterm 2)	datashader
14	2026-04-16	Machine Learning at Scale	Tensorflow, TF probability, PyMC
15	2026-04-23	Final project presentations	

Scheduling of final examinations: There will be no final exams for this class. There will be a final group project which will be an implementation of a data science analytic pipeline for a large dataset that will run on the cloud.

Assignments and evaluations

Each week we will have several short assessment activities to build on class materials. These will include:

1. Short weekly homework that will be graded for correctness. You will have 7-10 days to complete each assignment. Assignments will be provided and submissions made via Github Classroom
2. Labs will be started in class and will need to be completed and submitted via Github Classroom. These will be graded for completion
3. You will be given weekly readings that introduce the material for the following week. There will be a short quiz based on the readings that must be completed before class each week.
4. There will be two midterm evaluations. These will be analytic and coding exercises to be done in class without the help of any generative AI or LLM coding assistance. These will typically be 45-60 minutes long and will also be submitted via Github Classroom.
5. You will form groups of 2-3 students and do an analytic group project on a Big dataset. This project will involve an end-to-end analytic pipeline and report based on either a data set you choose or a dataset provided to you (which is TBD). You will have required ungraded check-ins with the professor per schedule to show progress, and a presentation on the final day of class. The project report and code must be submitted per the calendar

Late policies

1. Quizzes, labs, midterms and projects cannot be late; a late submission will be entered as a 0.
    a. For quizzes and labs, the lowest score for each type will be removed from calculation of the final grade at the end of the semester
2. Homework will have the following late policy:
    a. There will be a penalty of 10% of the total points per day for each day late, for a maximum of 20% penalty
    b. Any homework can be submitted by the last day of class for a penalty of 40% of the total points for that homework

Grading

- homework (25%)
- lab completions (10%)
- quizzes (10%)

- midterm exams (15%)
- final group project (30%)
- class participation/attendance (10%)

Academic Integrity Code

Academic integrity is an essential part of the educational process, and all members of the GW community take these matters very seriously. As the instructor of record for this course, my role is to provide clear expectations and uphold them in all assessments. Violations of academic integrity occur when students fail to cite research sources properly, engage in unauthorized collaboration, falsify data, and otherwise violate the Code of Academic Integrity. If you have any questions about whether particular academic practices or resources are permitted, you should ask me for clarification. If you are reported for an academic integrity violation, you should contact Conflict Education and Student Accountability (CESA), formerly known as Student Rights and Responsibilities (SRR), to learn more about your rights and options in the process. Consequences can range from failure of assignment to expulsion from the University and may include a transcript notation. For more information, refer to the CESA website at students.gwu.edu/code-academic-integrity or contact CESA by email cesa@gwu.edu or phone 202-994-6757.

University Policy on Observance of Religious Holidays

Students must notify faculty during the first week of the semester in which they are enrolled in the course, or as early as possible, but no later than three weeks prior to the absence, of their intention to be absent from class on their day(s) of religious observance. If the holiday falls within the first three weeks of class, the student must inform faculty in the first week of the semester. For details and policy, see provost.gwu.edu/policies-procedures-and-guidelines.

Use of Electronic Course Materials and Class Recordings

Students are encouraged to use electronic course materials, including recorded class sessions, for private personal use in connection with their academic program of study. Electronic course materials and recorded class sessions should not be shared or used for non-course related purposes unless express permission has been granted by the instructor. Students who impermissibly share any electronic course materials are subject to discipline under the Student Code of Conduct. Contact the instructor if you have questions regarding what constitutes permissible or impermissible use of electronic course materials and/or recorded class sessions. Contact Disability Support Services at disabilitysupport.gwu.edu if you have questions or need assistance in accessing electronic course materials.

Academic support

Academic Commons

Academic Commons is the central location for academic support resources for GW students. To schedule a peer tutoring session for a variety of courses visit go.gwu.edu/tutoring. Visit academiccommons.gwu.edu for study skills tips, finding help with research, and connecting with other campus resources. For questions email academiccommons@gwu.edu.

GW Writing Center

GW Writing Center cultivates confident writers in the University community by facilitating collaborative, critical, and inclusive conversations at all stages of the writing process. Working alongside peer mentors, writers develop strategies to write independently in academic and public settings. Appointments can be booked online at gwu.mywconline.

Support For Students in and Outside the Classroom

Disability Support Services (DSS) 202-994-8250

Any student who may need an accommodation based on the potential impact of a disability should contact Disability Support Services at disabilitysupport.gwu.edu to establish eligibility and to coordinate reasonable accommodation.

Student Health Center 202-994-5300, 24/7

The Student Health Center (SHC) offers medical, counseling/psychological, and psychiatric services to GW students. More information about the SHC is available at healthcenter.gwu.edu. Students experiencing a medical or mental health emergency on campus should contact GW Emergency Services at 202-994-6111, or off campus at 911.

GW Campus Emergency Information

GW Emergency Services: 202-994-6111

For situation-specific instructions, refer to GW's Emergency Procedures guide.

GW Alert

GW Alert is an emergency notification system that sends alerts to the GW community. GW requests students, faculty, and staff maintain current contact information by logging on to alert.gwu.edu. Alerts are sent via email, text, social media, and other means, including the Guardian app. The Guardian app is a safety app that allows you to communicate quickly with GW Emergency Services, 911, and other resources. Learn more at safety.gwu.edu.

Protective Actions

GW prescribes four protective actions that can be issued by university officials depending on the type of emergency. All GW community members are expected to follow directions according to the specified protective action. The protective actions are Shelter, Evacuate, Secure, and Lockdown (details below). Learn more at safety.gwu.edu/gw-standard-emergency-statuses.

Shelter

- Protection from a specific hazard
- The hazard could be a tornado, earthquake, hazardous material spill, or other environmental emergency.
- Specific safety guidance will be shared on a case-by-case basis.

Action:

- Follow safety guidance for the hazard.

Evacuate

- Need to move people from one location to another.
- Students and staff should be prepared to follow specific instructions given by first responders and University officials.

Action:

- Evacuate to a designated location.
- Leave belongings behind.
- Follow additional instructions from first responders.

Secure

- Threat or hazard outside of buildings or around campus.
- ncreased security, secured building perimeter, increased situational awareness, and restricted access to entry doors.

Action:

- Go inside and stay inside.
- Activities inside may continue.

Lockdown

- Threat or hazard with the potential to impact individuals inside buildings.
- Room-based protocol that requires locking interior doors, turning off lights, and staying out of sight of corridor window.

Action:

- Locks, lights, out of sight
- Consider Run, Hide, Fight

Classroom Emergency Lockdown Buttons

Some classrooms have been equipped with classroom emergency lockdown buttons. If the button is pushed, GWorld Card access to the room will be disabled, and GW Dispatch will be alerted. The door must be manually closed if it is not closed when the button is pushed. Anyone in the classroom will be able to exit, but no one will be able to get in.