

DSAN 6000: Big Data and Cloud Computing

Section 01: M 3:30-6pm, Walsh 394

Section 02: Th 6:30-9pm, Walsh 394

Jeff Jacobs

jj1088@georgetown.edu

Fall 2026, Georgetown University

Welcome to the Fall 2026 version of **Big Data and Cloud Computing** at Georgetown University! Please note that the most up-to-date version of this syllabus will always be available at jjacobs.me/dsan6000

Course Staff and Office Hours

- Prof. Jeff Jacobs, jj1088@georgetown.edu: *Schedule office hours at jjacobs.me/meet*
 - Tues 4:30-7:00pm, Weds 3-4pm, Thurs 3-4pm
 - *(Please try to schedule at least 8 hours in advance, and let me know briefly in the “Additional Info” portion of the signup form what you’d like to discuss, so I have time to prepare!)*
- TA Samyu Vakkalanka, spv15@georgetown.edu
 - Office Hours: TBD
- TA Fangzhou Wang, fw256@georgetown.edu
 - Office Hours: TBD
- TA Siru Wu, sw1430@georgetown.edu
 - Office Hours: TBD

Course Description

Data is everywhere, and often it’s simply too large or complex for traditional tools to handle. This hands-on, workshop-style course introduces the principles and practice of big data analytics and cloud computing. Students learn how to work with distributed computing frameworks such as **Apache Spark**; modern tools like **DuckDB**, **Polars**, and **vector databases**; and cloud platforms like **AWS**. Key topics include parallelization and concurrency, data warehousing, scalable machine learning with Spark MLlib, streaming analytics, and serverless data engineering. The course emphasizes end-to-end workflows—from data ingestion and cleaning to analysis, modeling, and presentation—using **Python (PySpark)**, **SQL**, and **Git/GitHub**.

A major focus is on practical application: by the end of the course, students will be able to set up and manage cloud environments, process massive datasets, build scalable pipelines, and apply distributed machine learning methods. The course culminates in a final project where students execute a full big data workflow on real-world datasets using cloud resources.

Assignment Structure

On the basis of the guidelines we've developed for courses offered through DSAN, this course will have one in-class midterm but **no** final exam! Instead, you will work on a final project throughout the second half of the course. Each week you will have a (two-part) homework assignment, and final grades will be determined using the following weighting scheme:

Category	Percent of Final Grade
Homeworks	50%
In-Class Midterm	20%
Final Project	25%
Attendance	5%

Course Topics / Calendar

The following is a rough map of what we will work through together throughout the semester; given that **everyone learns at a different pace**, my aim is to leave us with a good amount of **flexibility** in terms of how much time we spend on each topic.

Nonetheless, if you are able to do the recommended readings listed below *before* the start of each week's lecture, this will be a good indication that you are "on pace", relative to the overall set of topics I aim to cover by the end of the semester.

If you find any discrepancies between this schedule and Georgetown's [official calendar](#), please let us know!

Unit	Week	Date	Topic
Unit 1: Core Concepts	1	Aug 31	Course Overview
	2	Sep 8	Cloud Computing
	3	Sep 14	Parallelization Concepts
Unit 2: Scaling Pandas	4	Sep 21	DuckDB
	5	Sep 28	Polars
	6	Oct 5	Data Engineering
Unit 3: Distributed Computing with Spark	7	Oct 19	Introduction to Spark
	8	Oct 26	ML with Spark
	9	Nov 2	NLP with Spark

Unit	Week	Date	Topic
Unit 4: ETL for Vector DBs	10	Nov 9	ETL Pipelines
	11	Nov 16	Vector Databases
		<i>Nov 23</i>	<i>No Class (Fall Break)</i>
Unit 5: Final Projects	12	Nov 30	Final Project Preparation
	13	Dec 7	Final Presentations

Assignment Distribution, Submission, and Grading

Each weekly programming assignment for the course will be distributed as a GitHub Repository Template, which will be linked to on Canvas. Since (nearly) all assignments will require you to use your AWS Academy account (to work on assignments with more computing power than is possible on your laptops), your standard workflow for assignments will be to

1. Click on the link to the GitHub Template in Canvas,
2. Click the green “**Use This Template**” button in the upper-right corner of the Template, then
3. Clone the resulting Repository to your “main” **EC2 Instance**

We will dive much more deeply into this workflow in Week 1 of the course!

Late Policy

After the due date, for each **homework** assignment, you will have a grace period of 24 hours to submit the assignment without a lateness penalty. After this 24 hour grace period, late penalties will be applied up until 66 hours after the due date. Specifically, late penalties will be applied based on the following scale (unless you obtain an excused lateness from one of the instructional staff!):

- **0 to 24 hours** after due date: no penalty
- **24 to 30 hours** after due date: 2.5% penalty
- **30 to 42 hours** after due date: 5% penalty
- **42 to 54 hours** after due date: 10% penalty
- **54 to 66 hours** after due date: 20% penalty
- **More than 66 hours** after due date: Assignment submissions no longer accepted (without instructor approval)

Excused lateness exceptions will be grounded in a policy where you can “pause” the class (due-date-wise, but also just, expectation-of-progress-wise more generally) to take the time you need to handle things happening in your life. Then, you can “resume” work once you feel more centered/once you find yourself back on stable ground. The “[Maslow’s hierarchy](#)” model from psychology (Figure 1) provides a straightforward way to think about this approach.

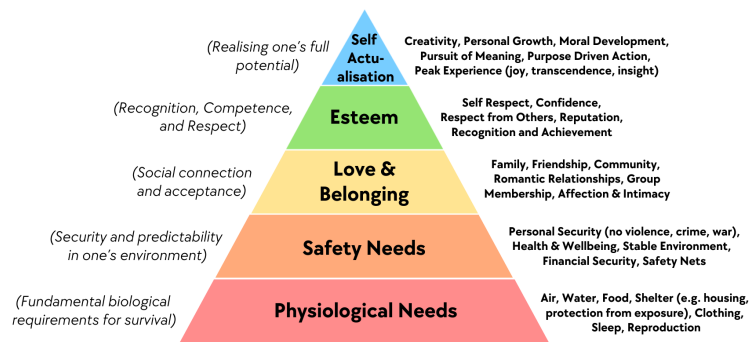


Figure 1: Maslow's Hierarchy of Needs, from [Wikimedia](#)

Learning about Big Data and Cloud Computing is way up there in the green and blue portions, so, it's important that you feel like the first four levels are solidified before you turn your focus back to classwork. And, it's a win-win, because it means that once you're back in action you can be more engaged and receptive to new topics, rather than being forced to "simulate" being in the blue triangle while in reality struggling in other spots!

Final Letter Grade Determination

Once all assignments have been graded, we will compute your final **numeric grade** according to the above weighting, rounded to two decimal places. The **letter grade** that we report to Georgetown on the basis of this numeric grade will then follow the DSAN letter grade policy (updated Fall 2026) as follows, where start and end points for each range are **inclusive**:

Range Start	Range End	Letter Grade
93.50	100.00	A
89.50	93.49	A-
88.00	89.49	B+
81.50	87.99	B
79.50	81.49	B-
60.00	79.49	C
0.00	59.99	F

Title IX/Sexual Misconduct Statement

Georgetown University and its faculty are committed to supporting survivors and those impacted by sexual misconduct, which includes sexual assault, sexual harassment, relationship violence, and stalking. Georgetown requires faculty members, unless otherwise designated as confidential,

to report all disclosures of sexual misconduct to the University Title IX Coordinator or a Deputy Title IX Coordinator.

If you disclose an incident of sexual misconduct to a professor in or outside of the classroom (with the exception of disclosures in papers), that faculty member must report the incident to the Title IX Coordinator, or Deputy Title IX Coordinator. The coordinator will, in turn, reach out to the student to provide support, resources, and the option to meet. [Please note that the student is not required to meet with the Title IX coordinator.]. More information about reporting options and resources can be found in the [Sexual Misconduct Resource Center](#).

If you would prefer to speak to someone confidentially, Georgetown has a number of fully confidential professional resources that can provide support and assistance. These resources include:

- *Health Education Services for Sexual Assault Response and Prevention*: Confidential email sarp@georgetown.edu
- *Counseling and Psychiatric Services (CAPS)*: 202-687-6985 [After hours you can call 833-960-3006 to reach Fonemed, a telehealth service, and ask for the on-call CAPS clinician]

College of Arts and Sciences Resources and Policies

You can find a collection of relevant resources and policies for students on the [GSAS website](#), and the Provost's policy on accommodating students' religious observances on the [Campus Ministry website](#).

You can also make use of the [Student Academic Resource Center](#). In particular, within the Resource Center there is a link to Georgetown's [Disability Support page](#). If you believe you have a disability, you can contact the Academic Resource Center (arc@georgetown.edu) for further information. The ARC is located in the Leavey Center, Suite 335 (202-687-8354), and it is the campus office responsible for reviewing documentation provided by students with disabilities and for determining reasonable accommodations in accordance with the Americans with Disabilities Act (ADA) and University policies.

References