

# Processing Big Data: Tools and Techniques

## CS 131

Fall 2025 Sections 02, 03 In Person 3 Unit(s) 08/20/2025 to 12/08/2025 Modified 08/19/2025

### Contact Information

---

#### Instructor: Jelena Gligorijevic

Email: [jelena.gligorijevic@sjsu.edu](mailto:jelena.gligorijevic@sjsu.edu)

Office: MacQuarrie Hall 211

Class Days: Monday and Wednesday

Section 02 Time: 7:30 - 8:45 AM

Section 03 Time: 9:00 - 10:15 AM

Classroom: Duncan Hall 415

#### Office Hours

Monday, 11:00 AM to 1:00 PM, MacQuarrie Hall 211

Please schedule your time here: <https://calendar.app.google/995u8uP4qdmTLBPf8>  
(<https://calendar.app.google/995u8uP4qdmTLBPf8>).

### Course Information

---

In-depth study of essential tools and techniques for processing big data over the UNIX operating system and/or other operating systems. On UNIX, it includes using grep, sed, awk, join, and programming advanced shell scripts for manipulating big data.

Prerequisite(s): CS 46B with a grade of "C-" or better; or CS 22B with a grade of "C-" or better AND graduate standing. Allowed Declared Majors: Computer Science BS, Data Science BS, MS Bioinformatics (MSBI).

Allowed Declared Majors: Computer Science BS, Data Science BS, Computer Science and Linguistics BS, and MS Bioinformatics (MS BI).

# \* Classroom Protocols

---

Students are expected to adhere to the [Student Conduct Code \(https://www.sjsu.edu/studentconduct/\)](https://www.sjsu.edu/studentconduct/).

This semester, we'll be using a private Discord server as our main space for class communication and collaboration.

Discord is where we can:

- Ask and answer course-related questions.
- Share ideas, resources, and helpful tips.
- Learn from one another's perspectives and approaches.
- Continue discussions beyond class time.
- Organize and coordinate with your project team.

Instead of sending most questions by email, please post them in the appropriate Discord channel so the whole class can see and benefit from the responses. Your classmates may have the same question, and sometimes, a peer's explanation will click for you even faster.

A direct link to join will be given in Canvas.

Private or sensitive matters should be addressed through direct messages or discussed during office hours.

Office hours are the best time to get help with assignments, conceptual questions, or technical troubleshooting.

## ☰ Program Information

---

Diversity Statement - At SJSU, it is important to create a safe learning environment where we can explore, learn, and grow together. We strive to build a diverse, equitable, inclusive culture that values, encourages, and supports students from all backgrounds and experiences.

## 📊 Course Learning Outcomes (CLOs)

---

By the end of the course, students will be able to:

- Analyze, manipulate and process large-scale data with the UNIX/Linux command line and other operating systems.
- Develop shell scripts for use in data-intensive applications.
- Build data analysis pipelines, automate tasks, make analyses reproducible and shareable.
- Compare data analysis on the command line with use of graphical user interface and web-based tools.
- Solve big data challenges with the UNIX/Linux shell and command-line tools.

- Apply data science solutions to datasets from example domains, such as biology, business, and finance.
- Perform big data analysis efficiently, document and reproduce analysis, use cloud computing for data intensive problems.

## Course Materials

---

Recommended reads

### UNIX Command Line: A Complete Introduction

**Author:** William Shotts Jr.

**Publisher:** No Starch Press

[Download it from the [author's webpage \(https://linuxcommand.org/tlcl.php\)](https://linuxcommand.org/tlcl.php)]

### Data Science at the Command Line, 2nd Edition

**Author:** Jeroen Janssens

**Publisher:** O'Reilly Media, Inc

**ISBN:** 978149208791

[You can read it free through SJSU library (<https://library.sjsu.edu/ebooks/safaribooks-online-o-reilly>), (<https://library.sjsu.edu/ebooks/safaribooks-online-o-reilly>)]

## Course Requirements and Assignments

---

This course is designed to be hands-on and cumulative, preparing students to process, manipulate, and scale data using real-world tools. Success in this course requires regular engagement, thoughtful experimentation, and a consistent time investment both inside and outside the classroom.

#### \* Homework Assignments (15%) \*

Weekly homework assignments will reinforce topics covered in class. All homework is individual and submitted via Canvas.

Collaboration is encouraged in concept, but your code must be your own. Academic integrity violations will be referred to the Office of Student Conduct.

#### \* Quizzes (5%) \*

Short quizzes will be administered in class 1-2 times per week. These are designed to assess retention of recent material and promote consistent engagement.

I understand that life happens. Students are allowed up to two make-up quizzes/homework during the semester, provided they notify the instructor in advance (when possible) or within a reasonable time after the missed class. Make-up quizzes/homework assignments must be completed within one week of the original date.

### **\* Project Assignments (40%) \***

As part of your CS 131 experience, you will join the SJSU BigData Lab - a simulated, hands-on startup environment where you will work in rotating roles on a data-driven team. This project is designed to give you real-world exposure to big data engineering, team collaboration, agile workflows, and data storytelling.

You won't just learn about tools - you'll build something meaningful with them.

#### **The setup**

You are part of the SJSU BigData Lab, a fictive company where your professor serves as the CEO.

You will work in teams of 4-5 students. Each team becomes a Product Team responsible for one dataset-driven product of their own choosing. You will choose a dataset and theme of interest (e.g., sports analytics, transportation, music, education, social trends, etc.)

You will rotate roles every two-three weeks to gain experience across product, engineering, and storytelling functions, and gain better understanding of your preference in the types of role you like most.

#### **Team Roles (Rotated Biweekly)**

##### **1. Product Manager (PM)**

- Represents the team in meetings with the CEO (professor) during sprint planning (biweekly).
- Translates strategic goals into actionable tasks.
- Coordinates team responsibilities.
- Owns the vision for the product and helps scope the sprint.

##### **2. Big Data Engineers (2-3 per sprint)**

- Implement the core technical work using course tools.
- Follow Git best practices (branches, pull requests, peer review).
- Attend weekly stand-up meetings with the CEO to share progress and raise blockers.

### 3. Big Data Storyteller

- Creates the biweekly project report.
- Translates technical output into a clear narrative.
- Connects CEO goals → PM tasks → engineering work → final insights.

#### Final Deliverables

- A cleaned, well-documented Git repo with 5-6 sprints of development.
- 5-6 professional reports from the storyteller (you'll rotate through this role).
- A final 10-minute team presentation.
- Peer Feedback Form

#### \* Exams: Midterms and Final (40%) \*

There will be two midterm exams and a comprehensive final exam.

- **Midterm 1 (10%)** – Covers Command-Line Fundamentals
- **Midterm 2 (10%)** – Covers Shell Scripting and Text Processing
- **Final Exam (20%)** – Cumulative, includes Distributed Computing: MapReduce, Spark, Containerization, Workflows, Cloud Platforms

Exams will test:

- Conceptual understanding of data processing tools
- Practical problem-solving
- Interpretation and optimization of code snippets

Make-up exams will only be considered for emergencies with proper documentation.

## ✓ Grading Information

---

Grading breakdown:

- Homework Assignments 15%
- Project Assignments 40%
- Quizzes 5%
- Midterm 1 10%

- Midterm 2 10%
- Final 20%

Final grades:

Grade	Points
A plus	> 96
A	93 - 95.99
A minus	90 - 92.99
B plus	86 - 89.99
B	83 - 85.99
B minus	80 - 82.99
C plus	76 - 79.99
C	73 - 75.99
C minus	70 - 72.99
D plus	66 - 69.99
D	63 - 65.99
D minus	60 - 62.99
F	< 60

## University Policies

---

Per [University Policy S16-9 \(PDF\)](http://www.sjsu.edu/senate/docs/S16-9.pdf) (<http://www.sjsu.edu/senate/docs/S16-9.pdf>), relevant university policy concerning all courses, such as student responsibilities, academic integrity, accommodations, dropping and adding, consent for recording of class, etc. and available student services (e.g. learning assistance, counseling, and other resources) are listed on the [Syllabus Information](https://www.sjsu.edu/curriculum/courses/syllabus-info.php) (<https://www.sjsu.edu/curriculum/courses/syllabus-info.php>) web page. Make sure to visit this page to review and be aware of these university policies and resources.

## Course Schedule

---

The course schedule is subject to change with fair notice. Changes will be announced on Canvas.

Week	Day	Date	New
W1	W	August 20	Course Introduction. Big Data Overview.
W2	M	August 25	Intro to UNIX Commands, Git, and Remote access
W2	W	August 27	Basic UNIX Commands
W3	M	September 1	Labor day, no classes
W3	W	September 3	Job Control, Background Jobs, and Terminal Tools
W4	M	September 8	Regular expressions, piping and redirection
W4	W	September 10	Regular expressions, piping and redirection
W5	M	September 15	Filtering and Matching (grep -E, wc, sort, uniq, cut)
W5	W	September 17	Midterm 1 prep
W6	M	September 22	Midterm 1.
W6	W	September 24	Editing and Transforming Streams (sed, awk)
W7	M	September 29	Editing and Transforming Streams (sed, awk)
W7	W	October 1	Shell Scripting for Automation
W8	M	October 6	Shell Scripting for Automation
W8	W	October 8	Midterm 2 prep
W9	M	October 13	Midterm 2
W9	W	October 15	When to Move Beyond Bash.

W10	M	October 20	Handling Larger-Than-Memory Data on One Machine
W10	W	October 22	Introduction to Cloud Environments & Scalable setups
W11	M	October 27	Distributed File Systems and MapReduce Concept
W11	W	October 29	Hadoop Ecosystem and Hive
W12	M	November 3	From MapReduce to Spark
W12	W	November 5	Spark Programming Continued
W13	M	November 10	Spark Algorithmic Techniques
W13	W	November 12	Optimizing and Tuning Spark Jobs
W14	M	November 17	Building Data Pipelines & Airflow
W14	W	November 19	Containerization with Docker
W15	M	November 24	Deploying Big Data Applications on Cloud
W15	W	November 26	Non Instructional Day, no classes
W16	M	December 1	Deploying Big Data Applications on Cloud. Final exam prep.
W16	W	December 3	Project presentations
W17	M	December 8	Project presentations