

Processing Big Data: Tools and Techniques

CS 131

Spring 2026 Section 01 In Person 3 Unit(s) 01/22/2026 to 05/11/2026 Modified 01/22/2026

Contact Information

Instructor: Dr. Jelena Gligorijevic

Email: jelena.gligorijevic@sjsu.edu

Office: MacQuarrie Hall 211

Office Hours

Tuesday, 12:30 PM to 3:30 AM, MacQuarrie Hall 211

Please schedule your time here: <https://calendar.app.google/Apm1tfq4oUxz1ZHa9>
(<https://calendar.app.google/Apm1tfq4oUxz1ZHa9>).

Please check my office-hours calendar before coming, since other students may already be scheduled during that time. I will keep the calendar updated and add additional availability when possible/needed. Booking through the calendar helps ensure you don't wait and that we can use the time efficiently.

Course Information

Section 01

Tuesday, Thursday, 7:30 AM to 8:45 AM, Duncan Hall 415

Section 02

Tuesday, Thursday, 9:00 AM to 10:15 AM, Duncan Hall 415

Section 03

Tuesday, Thursday, 10:30 AM to 11:45 AM, Duncan Hall 415

Course Description and Requisites

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).

Grading: Letter Graded

* Classroom Protocols

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

Communication:

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.

AI Use Policy

AI tools (e.g., ChatGPT, Copilot, Gemini) can be helpful for learning, brainstorming, and debugging. In this course, AI use is permitted in limited ways, but your submitted work must reflect your own understanding and original effort.

You may use AI tools to get ideas, debug, or get unstuck. **Any AI use must be disclosed in your submission** (tool + how you used it).

You may **not** copy/paste AI output into your reports or submit AI-generated solutions (text or code) as your own. All homework and project submissions must be your original work, and you must be able to explain your answers and code.

If you cannot sufficiently explain your submitted work, the penalty may include zero points on the assignment and a report to the Office of Student Conduct and Ethical Development.

Example

AI Use Disclosure: I used ChatGPT to help debug a Spark join issue and to brainstorm an approach for handling missing values. I did not copy code or text from the tool; I used the suggestions to guide my own implementation and wrote the final solution myself.

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

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

[The book is available for free through the SJSU library

<https://learning.oreilly.com/library/view/data-science-at/9781492087908/>

[\(https://learning.oreilly.com/library/view/data-science-at/9781492087908/\)](https://learning.oreilly.com/library/view/data-science-at/9781492087908/) .]

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 (4%) ***

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.

* Participation (1%) *

✓ Grading Information

Breakdown

- Homework Assignments 15%

- Project Assignments 40%
- Quizzes 4%
- Midterm 1 10%
- Midterm 2 10%
- Final 20%
- Participation 1%

Criteria

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 in Canvas.

Class #	Week	Day	Date	Topic
1	W1	Thursday	January 22	Course Introduction. Big Data Overview.
2	W2	Tuesday	January 27	Intro to UNIX Commands, Git, and Remote access
3	W2	Thursday	January 29	Basic UNIX Commands.
4	W3	Tuesday	February 3	Basic UNIX Commands.
5	W3	Thursday	February 5	I/O, redirection, pipelines
6	W4	Tuesday	February 10	Filtering and Matching
7	W4	Thursday	February 12	wc/sort/uniq/cut
8	W5	Tuesday	February 17	Regular expressions, piping and redirection
9	W5	Thursday	February 19	Midterm 1 prep
10	W6	Tuesday	February 24	Midterm 1.
11	W6	Thursday	February 26	Processes, job control basics
12	W7	Tuesday	March 3	Editing and Transforming Streams (sed)

13	W7	Thursday	March 5	AWK functions.
14	W8	Tuesday	March 10	AWK functions.
15	W8	Thursday	March 12	Shell Scripting for Automation
16	W9	Tuesday	March 17	Shell Scripting for Automation
17	W9	Thursday	March 19	Shell Scripting for Automation
18	W10	Tuesday	March 24	Midterm 2 prep
19	W10	Thursday	March 26	Midterm 2
20	W11	Tuesday	April 7	When to move beyond bash. Handling Larger-Than-Memory Data on One Machine
21	W11	Thursday	April 9	Introduction to Cloud Environments.
22	W12	Tuesday	April 14	Introduction to Cloud Environments. Scalable setups.
23	W12	Thursday	April 16	Distributed File Systems and MapReduce Concept
24	W13	Tuesday	April 21	PySpark
25	W13	Thursday	April 23	PySpark
26	W14	Tuesday	April 28	Orchestration and Containerization
27	W14	Thursday	April 30	Final exam review
28	W15	Tuesday	May 5	Final project presentations
29	W15	Thursday	May 7	Final project presentations

Final exams:

Section 1: Thursday, May 14 8:30-10:30 AM

Section 2: Tuesday, May 19 8:30-10:30 AM

Section 3: Thursday, May 14 10:45 AM -12:45 PM