SJSU SAN JOSÉ STATE UNIVERSITY

College of Science · Computer Science

Processing Big Data: Tools and Techniques

CS 131

Fall 2026 Section 02 In Person 3 Unit(s) 08/19/2026 to 12/07/2026 Modified 08/26/2026

Contact Information

Instructor: Dr. Jelena Gligorijevic

Email: jelena.gligorijevic@sjsu.edu
Office: MacQuarrie Hall 211

Office Hours

Tuesday, 12:30 PM to 2:30 PM, MacQuarrie Hall 211

Please schedule your time here: https://calendar.app.google/B34F7MkEbNAPLESg9 (https://calendar.app.google/B34F7MkEbNAPLESg9)

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. If you need more time, please book more than one time slot.

Course Information

Lecture

Tuesday, Thursday, 10:30 AM to 11:45 AM, Science Building 311

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/).

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)]

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/) .]

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.

*** Hands-on Assignments (10%) ***

Hands-on assignments will reinforce topics covered in class. All hands-on 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, typically once a 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/hands-on 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.). Your dataset must be public.

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

The five sprints

The project runs as five back-to-back sprints. Each sprint uses the tools taught in the weeks it spans, and each sprint report ends with a short stakeholder update.

Final Deliverables

- A cleaned, well-documented Git repo with 5 sprints of development.
- 5 professional reports from the storyteller (you'll rotate through this role).
- A final 10-minute team presentation (in class, December 1 and 3).
- Final report due December 1; Peer Feedback Form due Friday, December 4.

* Exams: Midterms (40%) *

There will be two midterm exams.

- Midterm 1 (20%)
- Midterm 2 (20%)

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.

*** Individual Project Walkthrough (5%) ***

Towards the end of the semester, each student schedules a short (~10 minute) 1-on-1 session with the professor through the office-hours calendar. You will walk me through parts of your team's work (a pipeline, a query, a design decision, or a benchmark result) and explain how and why it works. This is your opportunity to show your individual understanding of the work your team submitted.

*** Participation (1%) ***

✓ Grading Information

Breakdown

- Project Assignments 40%
- Midterm 1 20%
- Midterm 2 20%
- Hands-on Assignments 10%
- Individual Project Walkthrough 5%
- Quizzes 4%
- Participation 1%

Criteria

Final grades:

Grade	Points
A plus	≥ 97
A	93 - 96.99

Grade	Points
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), 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) 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	August 20	Course Introduction. Big Data Overview.
2	W2	Tuesday	August 25	Intro to UNIX Commands and the Terminal.
3	W2	Thursday	August 27	Basic UNIX Commands. Viewing and Editing Files.
4	W3	Tuesday	September 1	Git, GitHub, and Remote access. *(Sprint 1 assigned)*
5	W3	Thursday	September 3	I/O, redirection, pipelines

6	W4	Tuesday	September 8	wc/sort/uniq/cut/join
7	W4	Thursday	September 10	Filtering and Matching. Regular expressions.
8	W5	Tuesday	September 15	find, xargs, wildcards, permissions *(Sprint 1 due; Sprint 2 assigned)*
9	W5	Thursday	September 17	Processes, job control basics
10	W6	Tuesday	September 22	Cloud Shell and Cloud Storage.
11	W6	Thursday	September 24	Editing and Transforming Streams (sed, awk)
12	W7	Tuesday	September 29	Shell Scripting for Automation *(Sprint 2 due)*
13	W7	Thursday	October 1	Midterm 1 prep
14	W8	Tuesday	October 6	**Midterm 1.** *(Sprint 3 assigned)*
15	W8	Thursday	October 8	SQL for Big Data (DuckDB)
16	W9	Tuesday	October 13	SQL for Big Data: joins and window functions
17	W9	Thursday	October 15	Columnar formats: Parquet
18	W10	Tuesday	October 20	shell → DuckDB → Cloud Storage → BigQuery *(Sprint 3 due; Sprint 4 assigned)*
19	W10	Thursday	October 22	Distributed File Systems and MapReduce Concept. Cloud environments.
20	W11	Tuesday	October 27	From MapReduce to Spark
21	W11	Thursday	October 29	PySpark
22	W12	Tuesday	November 3	PySpark: algorithmic techniques
23	W12	Thursday	November 5	Spark tuning. Dataproc Serverless.
24	W13	Tuesday	November 10	BigQuery at scale *(Sprint 4 due; Sprint 5 assigned)*
25	W13	Thursday	November 12	Scaling lab: Dataproc Serverless runs
26	W14	Tuesday	November 17	Midterm 2 prep
27	W14	Thursday	November 19	**Midterm 2.**
28	W15	Tuesday	November 24	Containerization (Docker)
—	W15	Thursday	November 26	Thanksgiving - no class
29	W16	Tuesday	December 1	Final project presentations *(final report due)*
30	W16	Thursday	December 3	Final project presentations