SJSU SAN JOSÉ STATE UNIVERSITY

College of Science · Computer Science

Processing Big Data: Tools and Techniques

CS 131

Fall 2026 In Person 3 Unit(s) 08/19/2026 to 12/07/2026 Modified 08/20/2026

Contact Information

Instructor: Jelena Segan
Email: jelena.segan@sjsu.edu
Office: DH282

Office Hours

Tuesday, Thursday, 10:30 AM to 11:30 AM, DH 282
Please schedule your time here: https://calendar.app.google/aqFQSNQqX7Ny62zX9 (https://calendar.app.google/aqFQSNQqX7Ny62zX9)

Course Information

Lecture

Tuesday, Thursday, 12:00 PM to 1:15 PM, Dudley Moorhead Hall 231

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 (15%) ***

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 (5%) *

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 (35%) *

As part of your CS 131 experience, you will work in a small team on a semester-long Big Data Project. The goal is to apply the tools and techniques learned throughout the course to a real-world data-processing problem.

You will work in teams of approximately 3 - 4 students. Each team will choose a publicly available dataset and a topic of interest, such as sports, transportation, music, education, social trends, finance, or another approved area.

Your dataset should be sufficiently large or complex to demonstrate the value of scalable data-processing techniques. All datasets and project topics must be approved by the instructor.

The project is cumulative: as we learn new tools and techniques throughout the semester, you will have opportunities to apply them to your project.

- Project Proposal 3%

Each team will submit a short proposal describing:

- The topic and problem/question you plan to investigate
- The dataset and its source
- The approximate size and structure of the data
- Why the dataset is appropriate for the course
- Your initial plan for processing and analyzing the data

The instructor will provide feedback and approve the project before the team proceeds.

- Project Progress Check 7%

Approximately midway through the project, each team will submit a progress update demonstrating that meaningful work has begun. This should include:

- A working code repository
- Evidence of data loading, cleaning, or preprocessing
- Data-processing techniques and course tools used so far
- Preliminary results, when applicable
- Challenges encountered
- A brief plan for completing the project

This checkpoint is intended to provide feedback and help teams address problems before the final submission.

- Final Project Presentation 20%

At the end of the semester, each team will submit:

- A clean and well-documented code repository
- Final data-processing pipeline/code
- Analysis and results
- A short written report or README explaining the project
- A brief 8 - 10 minute in-class presentation

The final project will be evaluated based on the quality and correctness of the technical work, appropriate use of course concepts and tools, analysis and interpretation of results, documentation, and presentation.

- Individual Contribution 5%

Although the project is completed as a team, all students are expected to contribute meaningfully. During the semester, each student will have a brief in-person conversation with the instructor about their individual contribution to the project, what they worked on, and what they learned.

Students should be prepared to explain their own work and answer a few questions about the project. Individual contribution will be considered when determining each student's project grade.

* Exams: Midterms (44%) *

There will be two midterm exams.

- Midterm 1 (22%)
- Midterm 2 (22%)

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

- Project Assignments 35%
- Midterm 1 22%
- Midterm 2 22%
- Hands-on Assignments 15%
- Quizzes 5%
- Participation 1%

Criteria

Final grades:

Grade	Points
A plus	≥ 97
A	93 - 96.99
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

Grade	Points
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.
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
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
13	W7	Thursday	October 1	Midterm 1 prep
14	W8	Tuesday	October 6	**Midterm 1.**
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
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
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