

Processing Big Data: Tools and Techniques

CS 131

Summer 2026 Section 01 In Person 3 Unit(s) 06/01/2026 to 08/07/2026 Modified 05/26/2026

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

Communication with the instructor

Students are requested to use the Canvas message function to contact the instructor. Private messages sent to the instructor's email address gets lost due to the large volume of emails received.

The instructor does not write messages after normal business hours, on weekends or holidays.

Reviewing code for the homework and technical trouble-shooting should be done during the office hours.

Never send your entire code for an assignment to the instructor. The instructor will not fix all the bugs in your code.

AI Use Policy

Students are permitted to use AI tools in this course, provided all usage is fully and clearly disclosed within each submission (specify the tool used, the purpose, and the sections of the submission where it was applied). Regardless of AI involvement, every submission must be verified and understood by the student. At any point, students may be asked without prior notice to explain any part of their work. Inability to provide a consistent and coherent explanation will be treated as inappropriate AI use, resulting in grade deductions or, in egregious cases, a report of academic dishonesty.

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)

Upon successful completion of this 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

Textbook:

- UNIX Command Line: A Complete Introduction. William Shotts Jr. [[Download it \(https://linuxcommand.org/tlcl.php\)](https://linuxcommand.org/tlcl.php) from the author's page]

Other references:

- Data Science at the Command Line, 2nd Edition. Jeroen Janssens, Publisher(s): O'Reilly Media, Inc. ISBN: 978149208791. [You can read it for free [through SJSU library \(https://library.sjsu.edu/ebooks/safari-books-online-o-reilly\)](https://library.sjsu.edu/ebooks/safari-books-online-o-reilly).]
- Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python, 2nd Edition. Peter Bruce, Andrew Bruce, and Peter Gedeck, Publisher(s): O'Reilly Media, Inc. ISBN: 149207294X. [You can read it for free [through SJSU library \(https://library.sjsu.edu/ebooks/safari-books-online-o-reilly\)](https://library.sjsu.edu/ebooks/safari-books-online-o-reilly).]
- Linux Journey. <https://linuxjourney.com/>

Technology:

- Practice of command-line operations will be done on Google Cloud. Instructions to subscribe to a free student account will be provided.
- Some assignments and worksheet tasks need to be submitted through GitHub. Details will be given in the first assignment and worksheet instructions.

- Docker Engine <https://docs.docker.com/engine/> (<https://docs.docker.com/engine/>).

✓ Grading Information

Assignment	Grade Weight
Exam 1	30%
Exam 2	30%
Project Proposal Doc	5%
Project Progress Doc	5%
Final Project	20%
Worksheets 1-5 (2% each)	10%

Extra-credits and Reworks

No extra-credit assignments or rework opportunities will be given.

Late Submission

Late submissions within 24 hours will be deducted 10% of its final grade. Submissions over 24 hours late will have 20% grade deduction. Late submissions over 2 days will not be accepted.

Missed Assignments or Exams

When students need to miss an assignment deadline or exam due to health conditions or any other emergency, it should be reported within ONE week after the due date.

Final Grade Table

Total Grade	Letter Grade
97% and above	A plus
93% to 96%	A
90% to 92%	A minus
87% to 89%	B plus
83% to 86%	B
80% to 82%	B minus
77% to 79%	C plus

73% to 76%	C
70% to 72%	C minus
67% to 69%	D plus
65% to 66%	D
60% to 64%	D minus
59% and below	F

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

Date	Topic	Assignment
6/1	Course Introduction (00), Google Cloud Setup (01)	
6/3	Git, Linux Command Basics (02)	
6/8	Text, Redirections, and regex (03)	ws1
6/10	Permission and processes (04)	
6/15	The Environment and Utility Commands (05)	ws2
6/17	Hands-on lab	
6/22	Exam review 1	ws3
6/24	Exam 1	
6/29	shell scripting (06)	
7/1	sed (07), awk (08)	ws4

7/6	Spark (09)	
7/8	Spark, Project discussion	Project proposal due
7/13	Spark	
7/15	Spark, Project discussion	ws5
7/20	Docker (10)	
7/22	Docker, REST API	Project progress due
7/27	Exam review 2	
7/29	Exam 2	
8/3	project presentation	Project due
8/5	project presentation	