SJSU SAN JOSÉ STATE UNIVERSITY

College of Science · Computer Science

Processing Big Data: Tools and Techniques Section 01

CS 131

Spring 2025 In Person 3 Unit(s) 01/23/2025 to 05/12/2025 Modified 01/20/2025

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 or BIOL 123B with a grade of "C-" or better. Allowed Declared Majors: Computer Science BS, Data Science BS, MS Bioinformatics (MS BI).

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.

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) from the author's page]

Other good readings:

- Data Science at the Command Line, 2nd Edition. Jeroen Janssens, Publisher(s): O'Reilly Media, Inc. ISBN: 978149208791. [You can read it free through SJSU library (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 free through SJSU library (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 and Amazon AWS. Instructions to subscribe for a free student account will be provided.
- Some assignments and worksheet tasks need to be submitted through Github. Details will be given in first assignment and worksheet instructions.

Course Requirements and Assignments

Exams

Three exams will be conducted during the regular class hours. A tentative schedule will be given in the course schedule below.

The exams will contain multiple choice questions, true/false and short answer questions. Exams are open book, open notes, and comprehensive. No make-up exams except in case of verifiable emergency circumstances.

Hands-On Worksheets

You will have ten hands-on worksheets (ws1 - ws10). A worksheet will be due weekly. Please refer to Canvas for detailed instructions and deadlines. You need to submit the worksheets by their closing time on the due date. There will be no makeup on worksheets.

No worksheet will be re-opened after its closing date. As this is a fast-paced course, it is essential that you submit the worksheets in a timely fashion in order to keep up.

The purpose of the hands-on worksheets is to develop your understanding of the material and skills in using the command-line tools. The hands-on worksheets will involve learning how to use command line tools for analyzing and manipulating datasets from various domains, such as biology, business, finance.

We will take time at the beginning of each class to discuss any difficulties students have in completing the worksheets from previous classes.

Homework assignments

There will be five assignments in total (a1 - a5). The evaluation of a2 and a5 includes oral presentations in class. Please check the tentative schedule below.

All assignment solutions that you submit must be completely your own work (i.e., your solution cannot be copied from another source, such as other students, the internet, etc.). While it is fine to discuss the worksheet/assignment solutions with other students, solutions submitted on Canvas should reflect your own efforts. Oral examination might be requested. All homework should be submitted on Canvas, not by e-mail.

✓ Grading Information

Assignment	Grade Weight
Exam 1	15 %
Exam 2	15 %
Exam 3	15 %
Assignment 1	10 %
Assignment 2 (+ oral presentation)	10 %
Assignment 3	10 %

Assignment 4	10 %
Assignment 5 (+ oral presentation)	10 %
Worksheets	5 %

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 deducted. 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), 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

Date	Topic	Note
1/27	Course Introduction	
1/29	Intro to UNIX Commands, Git, and SSH	Due: ws1 prep
2/3	Basic UNIX Commands	
2/5	Redirection, regex, and vim	Due: ws2
2/10	Redirection, regex, and vim	
2/12	Permission and processes	Due: ws3
2/17	Monitoring and .bashrc	Due: a1
2/19	Exam review 1	
2/24	Exam 1	
2/26	shell scripting	
3/3	shell scripting	
3/5	sed	Due: ws4
3/10	awk	
3/12	a2 presentation	Due: a2
3/17	a2 presentation	Due: ws5
3/19	awk	
3/24	Exam review 2	Due: ws6

3/26	Exam 2	
3/31	Spring Recess (No class)	
4/2	Spring Recess (No class)	
4/7	Spark	Due: a3 Due: ws7
4/9	Spark	
4/14	Spark ML regression?	Due: ws8
4/16	Spark	
4/21	a5 discussion	Due: a4
4/23	make	Due: ws9
4/28	Airflow	
4/30	Docker	
5/5	Exam review 3	Due: ws10
5/7	Exam 3	
5/12	a5 presentation	Due: a5