SJSU SAN JOSÉ STATE UNIVERSITY

College of Science · Computer Science

Processing Big Data: Tools and Techniques

CS 131

Fall 2026 Section 01 In Person 3 Unit(s) 08/19/2026 to 12/07/2026 Modified 08/19/2026

Contact Information

Instructor(s): William B Andreopoulos

Office Location: MacQuarrie Hall 416

Telephone: (408) 924 5085

Email: william.andreopoulos@sjsu.edu

Office Hours: Friday 9-11am

Class Days/Time: Monday and Wednesday, 6:00pm-7:15pm

Classroom: Dudley Moorhead Hall 166 (in-person)

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

As this is an in-person section, course-related communication should preferably be done in-person during the regular class meeting time (in-person or via Zoom) or office hours. For online communication, students should use the course Discord channel. Rather than emailing redundant questions to the teaching staff, students should post questions on the course Discord channel where the entire class can read and benefit

from the responses. The system is catered to getting students help efficiently from classmates, the TA, embedded tutor, and the instructor. *Private messages sent to the instructor's other email addresses may get lost due to the large volume of emails received.*

The professor responds primarily to the Discord channel. The professor will re-post questions that are of general interest (e.g. about homework) or discuss them in class. Students are responsible for everything said in class. It is students' responsibility to keep up with what is said in class and not re-post the same questions repeatedly.

When students use the course Discord channel, they are expected to be identifiable through their names. Anonymous postings are unacceptable. Students who use fake nicknames may be removed from the Discord channel.
The instructor does not write messages after normal business hours, on weekends or holidays.

Technical trouble-shooting should be done during the office hours.

Never email your entire code for an assignment to the instructor. Limit the code you post to 20 lines or less.

Announcements that concern everyone, such as reminders about due dates or class policy, will be posted.

Class Attendance

Attendance (in-person or via Zoom) is highly recommended. Classes will be recorded as Zoom screencasts and posted on Canvas. Students are responsible for all material presented in class.

The polling questions in the slides are in the form of multiple-choice and true-false questions. Students should participate and follow the polling questions, either via Zoom polling or Zoom chat or ask in class.

Regrading Procedure

Grades assigned are final, unless there was an error in the grading. Special requests (e.g. grade changes or deadline extensions) should be done in-person; such special requests sent via electronic messages to the teaching staff will not be honored, since this is an in-person section. To request a higher grade, students should first submit the Canvas "Regrade request" form so there is a record of the request. After submitting a regrade request, please speak with the professor during office hours. Grades may be reevaluated at anytime and may go down as a result of a regrade.

At the end of the semester grade roundups (e.g. 89.95% to 90%) will only be considered if a student has pursued any extra credit opportunities offered and completed the SOTE evaluation.

Classroom Protocol

Students on Zoom should be muted when not speaking, and must be dressed appropriately when their camera is on.

Course material developed by the instructor is the intellectual property of the instructor. Students can not publicly share or upload instructor generated material for this course such as exam questions, lecture notes, hands-on exercises or homework solutions without instructor permission.

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:

1. Analyze, manipulate and process large-scale data with the UNIX/Linux command line and other operating systems.
2. Develop shell scripts for use in data-intensive applications.
3. Build data analysis pipelines, automate tasks, make analyses reproducible and shareable.
4. Compare data analysis on the command line with use of graphical user interface and web-based tools.
5. Solve big data challenges with the UNIX/Linux shell and command-line tools.
6. Apply data science solutions to datasets from example domains, such as biology, business, finance.
7. Perform big data analyses efficiently, document and reproduce analyses, use cloud computing for data-intensive problems.

Course Materials

Recommended Texts/Readings

Textbook

Beginner: UNIX Command Line: A Complete Introduction. William Shotts Jr.

Moderate: Linux Command Line and Shell Scripting Bible. Blum and Bresnahan

Advanced: UNIX Power Tools. Jerry Peek, Tim O'Reilly, and Mike Loukides.

Other good readings:

Advanced Programming in the UNIX Environment. W. Richard Stevens, Stephen A. Rago. 3rd Edition, 2013, Addison-Wesley.

Introduction to UNIX and Linux. John Muster.

Data Science at the Command Line, 2nd Edition, by Jeroen Janssens,
Released August 2021, Publisher(s): O'Reilly Media, Inc.
ISBN: 9781492087915

https://www.datascienceatthecommandline.com/2e/

-

A copy of my slides will be available to the students enrolled in the class.

Additional handouts will be provided through Canvas.

Other technology requirements / equipment / material

Practice of command-line operations will be done on IBM's computing cloud, Google Cloud and Amazon AWS. Instructions to subscribe for a free student account will be provided.

Course Requirements and Assignments

SJSU classes are designed such that in order to be successful, it is expected that students will spend a minimum of forty-five hours for each unit of credit (normally three hours per unit per week), including preparing for class, participating in course activities, completing assignments, and so on.

Reading assignments: Readings will regularly be assigned for the next class (see schedule). Slides will be posted under the Canvas modules before the next class.

Hands-On Worksheets:
We will have a number of hands-on worksheets. A worksheet submission is due approximately every week. Please refer to Canvas for detailed instructions and deadlines. The worksheet submission page on Canvas closes after it is due. You need to submit the worksheets by their closing time on the due date. A worksheet will not be re-opened after its closing date. Late worksheets will not be accepted. As this is a fast-paced course, it is essential that you submit your worksheet homework 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. Students will use IBM's computing cloud and Amazon AWS for practice. 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: Programming assignments will be assigned. More information will be given at the time of the first programming assignment.

All homework solutions that students submit must be completely their own work. While it is fine to discuss the worksheet/assignment solutions with other students, solutions submitted on Canvas should reflect a student's own efforts. *Do not write the code for anyone else. Never copy any code you find on another source, such as a website. Canvas automatically checks submissions for plagiarism from multiple online sources.* Oral examination might be requested.

All homework should be submitted online. Homework sent via an email or message will not be graded. Homework cannot be graded after it has been reviewed in class or a solution has been posted. All homework is due on the last day of class.

Late policy: Late penalty is 2% per day up to 14 days. After 14 days (or after the last day of classes if it is sooner) the submission page will be closed and no submission will be accepted. A homework submission page will not be re-opened after its closing date.

Participation: 1% of the grade consists of participation points, given for answering at least 30 Zoom polling questions during class throughout the semester. **Polling questions:** The polling questions are in the form of multiple-choice and true-false questions. It is not necessary to get the correct answer in a polling question to get participation credit.

Examinations

Midterm exams: There will be two Midterm exams during the semester.

Final exam: One final cumulative exam.

The exams will contain multiple choice questions, true/false and short answer questions. Exams are closed book, closed notes, and comprehensive. Exams are in-person. The exams should be done individually. No make-up exams except in case of verifiable medical reasons.

Extra credit opportunities

Extra credit of 1% is given to a student who volunteers to review his/her code solution for an entire assignment or a worksheet in-class (either via Zoom or in person). A code review lasts for 10 minutes max. These will take the form of code reviews, where the student walks us through his/her code solution for an assignment or a worksheet, we discuss the proposed solution and if there are better ways to solve the problem. Students have to add their name to a code review worksheet to reserve a code review timeslot. An assignment or worksheet can only be reviewed once. A student may reserve one timeslot at a time. If, after presenting, there are other timeslots available, a student may reserve another timeslot.

Graders/TAs

TBD

Use of generative AI tools

All assignments and worksheets submitted are expected to be the students' own original work. The instructor may, at any time, ask a student to explain the meaning of any part of any answer that they submit. If the student can't explain the answer to a question sufficiently, the penalty for such incidents will be zero points on the homework and a report to the Office of Student and Ethical Conduct.

✓ Grading Information

The course grade is based on:

50% Assignments

20% Midterms (10% each)

20% Final

9% Worksheets

1% Participation (Polling questions)

Grade	*Points*	*Percentage*
A plus	*960 to 1000*	*96 to 100%*
A	*930 to 959*	*93 to 95%*
A minus	*900 to 929*	*90 to 92%*
B plus	*860 to 899*	*86 to 89 %*
B	*830 to 859*	*83 to 85%*
B minus	*800 to 829*	*80 to 82%*
C plus	*760 to 799*	*76 to 79%*
C	*730 to 759*	*73 to 75%*
C minus	*700 to 729*	*70 to 72%*
D plus	*660 to 699*	*66 to 69%*
D	*630 to 659*	*63 to 65%*
D minus	*600 to 629*	*60 to 62%*

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

Week	Topic
08/24-08/26	Introduction to the Bash shell command line, passwords, ssh/sftp/scp with keys, git
08/31-09/02	Shell interpretation of user input, wildcards, aliases, editing, pagers, which, tar/zip, wc, uniq, grep, sort, history
09/07-09/09	Home directories, terminal setup and environment variables, shell prompt setup, pathnames, permissions, sudo
09/14-09/16	Processes, Job control, finding files (-exec), dealing with many files, data pre-processing, task automation, crontabs, top/htop, input/output redirection
09/21-09/23	File systems, directories, permissions, move, rsync, copy, symbolic and hard links, counting inodes and files
09/28-09/30	Saving and restoring work with screen and tmux. **Midterm 1**
10/05-10/07	Pipes and pipeline concept for data analytics tasks, jobs vs. processes, curl, gnu parallel, inter-process communication, sockets, signals, profiling, job priorities
10/12-10/14	Awk, sed, grep, join, diff, with bioinformatics/data analytics examples
10/19-10/21	Awk, sed, grep, join, cut, paste, tr, regular expressions with bioinformatics/data analytics examples

10/26-10/28	Shell scripting, quotas, disk space
11/02-11/04	Shell scripting, nslookup, traceroute. **Midterm 2**
11/09-11/11	Reproducible data processing with containers (Docker, Singularity). Workflow tools (Snakemake, Airflow, Nextflow, Clara Parabricks, Luigi, WDL, CWL, Galaxy), Hadoop, Spark. A case study. Amazon Cloud
11/16-11/18	Amazon Cloud: Data science and Machine Learning with AWS SageMaker
11/23-11/25	Google Cloud: Big Data/Analytics, BigTable, BigQuery, AI/ML
11/30-12/02	Google Cloud: Graph Databases (neo4j, JanusGraph). Workflow managers in HPC clusters (Slurm, Torque) to process large amounts of data.
12/07	*Final exam review*
	Final exam. Wed, December 9, 5:30-7:30 PM

The schedule is subject to change with fair notice.