San José State University
Department of Computer Science
CS131: Processing Big Data – Tools and Techniques, Fall 2022

Course and Contact Information

Instructor(s): William B Andreopoulos
Office Location: MacQuarrie Hall 416
Telephone: (408) 924 5085
Email: william.andreopoulos@sjsu.edu
Office Hours: Wednesday 13:30-14:30 (MH 416), Friday 14:00-15:30 (Zoom)
Class Days/Time: Tuesday and Thursday, 15:00-16:15
Classroom: Online via Zoom

Course Description

An 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.

Course Format

This course adopts an online classroom delivery format.

Faculty Web Page and MYSJSU Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found on Canvas Learning Management System course login website at http://sjsu.instructure.com. You are responsible for regularly checking with the course messaging system to learn of any updates. You should modify the Canvas settings for notifications of announcements and Slack messages to be sent to you.

Prerequisites

This course is offered to MS Bioinformatics or undergraduate Computer Science (BSCS) or Data Science (BSDS) students at San Jose State University. MS-BI students must have completed BIOL 123B with a grade of C- or better. BSCS and BSDS students must have completed CS 46B with a grade of C- or better.

Course Learning Outcomes (CLO)

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.

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 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. 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: Assignments will be assigned for each module of the course. The assignments will be similar to worksheets. All assignments should be submitted on the corresponding assignment page in Canvas by 11:59 P.M. on the due date. The programming assignments cumulatively will be worth 50% of your grade.

More information will be given at the time of the first assignment. There will be a penalty for late submission 2% for every day up to 15 days; after 15 days no submission will be accepted and the submission page will be closed. Homework sent by email will not be graded, students need to upload them to Canvas.

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.

Participation during class via Zoom: The polling questions are in the form of multiple-choice and true-false questions. All students are expected to participate with Zoom polling. Credit is given based on participation and it is not necessary to get the correct answer in polls to get credit. Please contact eCampus at ecampus@sjsu.edu with any questions or issues with the Zoom technology.

Examinations

Midterm Exam One: Thursday, September 29, 2022.
Midterm Exam Two: Thursday, November 3, 2022.
Final Exam: Friday, December 9, 2022.
The midterm exams are each one hour and fifteen minutes long. The final exam is two hours and fifteen minutes long.

The exams will contain multiple choice questions, true/false and short answer questions. Exams are *open book, open notes*, and comprehensive. The exams should be done individually and are not group work. No make-up exams except in case of verifiable emergency circumstances.

Grading Information

The course grade is based on:

50% Five Assignments
9% Weekly Worksheets
1% Participation (Zoom attendance)
20% Two midterms
20% Final

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%*

Communication with the instructor

Students should follow the correct channels for communication. Questions should preferably be done during the regular class meeting time via Zoom or office hours. For course-related electronic communication students should use the Discord channel:

1) We will be using the course Discord channel for class discussion. The system is catered to getting you help efficiently from classmates, the TA, and the instructor. Rather than emailing redundant questions to the teaching staff, students should post questions on the Discord channel where the entire class can read and benefit from the responses. The professor may re-post questions that are of general interest to the general channel or discuss them in class. The professor may ask students to reveal their real name if they are making special requests on Discord (e.g. deadline extensions) to prevent abuse.

2) Students are invited to join the office hours.

Private messages sent to the instructor's other email addresses get 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 email your entire code for an assignment to the instructor. The instructor will not fix all the bugs in your code. 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

Regular class attendance (via Zoom) is expected. Classes will be recorded as Zoom screencasts and posted on Canvas. Students are responsible for all material presented in all classes.

Regrading Procedure

Grades assigned are final, unless there was an error in the grading. If a student wants to request a regrade of a homework or test, please follow instructions on the "Regrade request" page on Canvas. A request for a regrade is not a technique to drum up a few more points. If the course instructor thinks a component was scored too generously the first time, it may be lowered in a regrade. Thus, regrading may result in a lower grade.

Classroom Protocol

Students 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.

Add/Drop Policy

For those wishing to add this course, the waitlist ends on August 27, 2022. After this date a permission number is required to add the class. The last day to drop a course without a "W" grade and to add classes via MySJSU is

September 15, 2022. To drop after this date, a Late Drop petition will be required. According to University and Department guidelines, dropping after September 15, 2022, requires a serious and compelling reason to drop a course. Grades alone do not constitute a reason to drop a course. Students who stop attending without officially dropping will be issued a "U" at the end of the semester, which is counted as an F in calculations of GPA.

Students are responsible for understanding the policies and procedures about add/drop, grade forgiveness, etc. Refer to the current semester's Catalog Policies section at http://info.sjsu.edu/static/catalog/policies.html . Add/drop deadlines can be found on the current academic year calendars document on the Academic Calendars webpage at http://www.sjsu.edu/provost/services/academic_calendars/ . The Late Drop Policy is available at http://www.sjsu.edu/aars/policies/latedrops/policy/ . Students should be aware of the current deadlines and penalties for dropping classes. Information about the latest changes and news is available at the Advising Hub at http://www.sjsu.edu/advising/.

Consent for Recording of Class and Public Sharing of Instructor Material

University Policy S12-7, http://www.sjsu.edu/senate/docs/S12-7.pdf , requires students to obtain instructor's permission to record the course. Common courtesy and professional behavior dictate that you notify someone when you are recording him/her. You must obtain the instructor's permission to make audio or video recordings in this class. Such permission allows the recordings to be used for your private, study purposes only. The recordings are the intellectual property of the instructor; you have not been given any rights to reproduce or distribute the material.

Course material developed by the instructor is the intellectual property of the instructor and cannot be shared publicly without his/her approval. You may 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 consent.

Academic Integrity

Your commitment as a student to learning is evidenced by your enrollment at San Jose State University. The University Academic Integrity Policy S07-2 at http://www.sjsu.edu/senate/docs/S07-2.pdf requires you to be honest in all your academic course work. Faculty members are required to report all infractions to the office of Student Conduct and Ethical Development. The Student Conduct and Ethical Development website is available at http://www.sjsu.edu/studentconduct/ . Instances of academic dishonesty will not be tolerated. Cheating on exams or plagiarism (presenting the work of another as your own, or the use of another person's ideas without giving proper credit) will result in a failing grade and sanctions by the University. For this class, all assignments are to be completed by the individual student unless otherwise specified. If you would like to include your assignment or any material you have submitted, or plan to submit for another class, please note that SJSU's Academic Integrity Policy S07-2 requires approval of instructors.

- Anyone caught cheating (including sharing answers with others during exams) in the class will receive a failing grade on the exam or assignment, in addition to other sanctions that are permitted by the University, including but not limited to the filing of a report with the Dean of Student Services and expulsion from the University.

Campus Policy in Compliance with the American Disabilities Act

If you need course adaptations or accommodations because of a disability, or if you need to make special arrangements in case the building must be evacuated, please make an appointment with me as soon as possible, or see me during office hours. Presidential Directive 97-03 at http://www.sjsu.edu/president/docs/directives/PD_1997-03.pdf requires that students with disabilities

requesting accommodations must register with the Accessible Education Center (AEC) at http://www.sjsu.edu/aec to establish a record of their disability.

In 2013, the Disability Resource Center changed its name to be known as the Accessible Education Center, to incorporate a philosophy of accessible education for students with disabilities. The new name change reflects the broad scope of attention and support to SJSU students with disabilities and the University's continued advocacy and commitment to increasing accessibility and inclusivity on campus.

University Policies

Per University Policy S16-9, university-wide policy information relevant to all courses, such as academic integrity, accommodations, etc. will be available on Office of Graduate and Undergraduate Programs' Syllabus Information web page at http://www.sjsu.edu/gup/syllabusinfo/

CS131: Processing Big Data – Tools and Techniques

The schedule is subject to change with fair notice.

Course Schedule

Week	Topic
08/22	Introduction to the Bash shell command line, passwords, ssh/sftp/scp with keys, git
08/29	Shell interpretation of user input, wildcards, aliases, editing, pagers, which, tar/zip, wc, uniq, grep, sort, history
09/05	Home directories, terminal setup and environment variables, shell prompt setup, pathnames, permissions, sudo
09/12	Processes, Job control, finding files (-exec), dealing with many files, data pre-processing, task automation, crontabs, top/htop, input/output redirection
09/19	File systems, directories, permissions, move, rsync, copy, symbolic and hard links, counting inodes and files
09/26	Saving and restoring work with screen and tmux. **Midterm 1**
10/03	Pipes and pipeline concept for data analytics tasks, jobs vs. processes, curl, gnu parallel, inter-process communication, sockets, signals, profiling, job priorities
10/10	Awk, sed, grep, join, diff, with bioinformatics/data analytics examples
10/17	Awk, sed, grep, join, cut, paste, tr, regular expressions with bioinformatics/data analytics examples
10/24	Shell scripting, quotas, disk space
10/31	Shell scripting, nslookup, traceroute. **Midterm 2**
11/07	Reproducible data processing with containers (Docker, Singularity). Workflow tools (Snakemake, Airflow, Nextflow, Clara Parabricks, Luigi, WDL, CWL, Galaxy), Hadoop, Spark. A case study of a comprehensive data pipeline workflow for RNA-Seq and amplicon analysis. Amazon Cloud
11/14	Amazon Cloud: Data science and Machine Learning with AWS SageMaker
11/21	Google Cloud: Big Data/Analytics, BigTable, BigQuery, AI/ML
11/28	Google Cloud: Graph Databases (neo4j, JanusGraph). Workflow managers in HPC clusters (Slurm, Torque) to process large amounts of data. *Final exam review*
	Final exam. Friday, December 9 2:45-5:00 PM