SJSU SAN JOSÉ STATE UNIVERSITY

College of Science · Computer Science

Processing Big Data: Tools and Techniques Section 80

CS 131

Fall 2023 3 Unit(s) 08/21/2023 to 12/06/2023 Modified 08/26/2023

Contact Information

Instructor(s): William B Andreopoulos

Office Location: MacQuarrie Hall 416

Telephone: (408) 924 5085

Email: william.andreopoulos@sjsu.edu

Office Hours: Friday 12:00-14:00 (Zoom)

Class Days/Time: Tuesday and Thursday, 15:00-16:15

Classroom: Online via Zoom

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

This course adopts an online classroom delivery format. 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.

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.

Communication with the instructor

Students should use the correct channels for course-related communication. Questions can be done during the regular class meeting time or office hours (via Zoom). For online 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 on Discord.

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.

Regrading Procedure

Grades assigned are final, unless there was an error in the grading. There will be no grade change through sending electronic messages to the teaching staff. If a student wants to request a higher grade for homework, they must follow instructions on the "Regrade request" page on Canvas. After submitting a regrade request, please speak with the professor during office hours or after class. 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 overall.

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: 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 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 on Canvas. Homework sent via an email or message will not be graded.

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 open book, open notes, and comprehensive. The exams should be done individually. 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%*

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
08/21	Introduction to the Bash shell command line, passwords, ssh/sftp/scp with keys, git
08/28	Shell interpretation of user input, wildcards, aliases, editing, pagers, which, tar/zip, wc, uniq, grep, sort, history
09/04	Home directories, terminal setup and environment variables, shell prompt setup, pathnames, permissions, sudo
09/11	Processes, Job control, finding files (-exec), dealing with many files, data pre-processing, task automation, crontabs, top/htop, input/output redirection
09/18	File systems, directories, permissions, move, rsync, copy, symbolic and hard links, counting inodes and files
09/25	Saving and restoring work with screen and tmux.**Midterm 1**
10/02	Pipes and pipeline concept for data analytics tasks, jobs vs. processes, curl, gnu parallel, inter-process communication, sockets, signals, profiling, job priorities
10/09	Awk, sed, grep, join, diff, with bioinformatics/data analytics examples
10/16	Awk, sed, grep, join, cut, paste, tr, regular expressions with bioinformatics/data analytics examples
10/23	Shell scripting, quotas, disk space
10/30	Shell scripting, nslookup, traceroute. **Midterm 2**
11/06	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
11/13	Amazon Cloud: Data science and Machine Learning with AWS SageMaker
11/20	Google Cloud: Big Data/Analytics, BigTable, BigQuery, AI/ML

11/27	Google Cloud: Graph Databases (neo4j, JanusGraph). Workflow managers in HPC clusters (Slurm, Torque) to process large amounts of data. *Final exam review*
	Final exam. Monday, December 11, 2:45-5:00 PM

The schedule is subject to change with fair notice.