College of Science · Computer Science

Processing Big Data: Tools and Techniques Section 01

CS 131

Spring 2023 3 Unit(s) 01/25/2023 to 05/15/2023 Modified 01/22/2023

Contact Information

Dr. Genya Ishigaki

Email: genya.ishigaki@sjsu.edu
Office: MH 215
Phone: (408) 924-5076
Website: https://sjsu-interconnect.github.io/ (https://sjsu-interconnect.github.io/)

Office Hours

Monday, Wednesday, 3:00 PM to 4:00 PM, MH 215

You don't need to make an appointment for these office hours. You can simply stop by my office.

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, finance.
- Perform big data analyses efficiently, document and reproduce analyses, use cloud computing for data- intensive problems.

Course Materials

Textbooks:

- 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, Jeroen Janssens, Released August 2021, Publisher(s): O'Reilly Media, Inc. ISBN: 978149208791

Technology:

- 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.
- 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.

Homework assignments

An assignment will be assigned for each module of the course (Five assignments in total). The assignments will be similar to worksheets.

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.

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 fastpaced 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.

✔ Grading Information

Assignment	Grade Weight
Exam 1	15 %
Exam 2	15 %
Exam 3	15 %
Assignment 1	10 %
Assignment 2	10 %
Assignment 3	10 %
Assignment 4	10 %
Assignment 5	10 %
Worksheets	5 %

Worksheet Grading
You will receive a Pass or Fail grade for each Worksheet. To receive the full credits from the worksheets, you need to pass 90 % of all Worksheet assignments throughout the semester.

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
92% to 96%	A
90% to 91%	A minus
87% to 89%	B plus
82% to 86%	B

80% to 81%	B minus
77% to 79%	C plus
72% to 76%	C
70% to 71%	C minus
67% to 69%	D plus
62% to 66%	D
60% to 61%	D minus
59% and below	F

University Policies

Per University Policy S16-9 (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 Syllabus Information web page (https://www.sjsu.edu/curriculum/courses/syllabus-info.php) (https://www.sjsu.edu/curriculum/courses/syllabus-info.php). Make sure to visit this page to review and be aware of these university policies and resources.

Course Schedule

Date	Topic
1/25	Course intro
1/30	Introduction to the Bash shell command line, passwords, ssh/sftp/scp with keys, git
2/1	cont.
2/6	Shell interpretation of user input, wildcards, aliases, editing, pagers, which, tar/zip, wc, uniq, grep, sort, history
2/8	cont.
2/13	Home directories, terminal setup and environment variables, shell prompt setup, pathnames, permissions, sudo
2/15	cont.
2/20	Processes, Job control, finding files (-exec), dealing with many files, data pre-processing, task automation, crontabs, top/htop, input/output redirection
2/22	cont.
2/27	File systems, directories, permissions, move, rsync, copy, symbolic and hard links, counting inodes and files
3/1	Saving and restoring work with screen and tmux.
3/6	Exam Review
3/8	Exam 1
3/13	Pipes and pipeline concept for data analytics tasks, jobs vs. processes, curl, gnu parallel, inter-process communication, sockets, signals, profiling, job priorities
3/15	cont.

3/20	Awk, sed, grep, join, diff, with data analytics examples
3/22	cont.
3/27	Spring recess; No class
3/29	Spring recess; No class
4/3	Awk, sed, grep, join, cut, paste, tr, regular expressions with data analytics examples
4/5	cont.
4/10	Shell scripting, quotas, disk space
4/12	Shell scripting, nslookup, traceroute.
4/17	Exam Review
4/19	Exam 2
4/24	Reproducible data processing with containers; Workflow tools
4/26	cont.
5/1	AI/ML on Cloud
5/3	cont.
5/8	cont.
5/10	Exam Review
5/15	Exam 3