Compute Servers for Teaching Big Data

Clark Fitzgerald

January 7, 2025

Abstract

I've taught an upper division statistics "Big Data" course at UC Davis and CSU Sacramento and have gained some perspective on what works well for this course. My goals with the course are to get students comfortable with using remote machines and to do realistic analysis of data sets that don't fit in memory, typically on the order of 100GB or so.

I've tried what feels like an excessive number of ways to run the server for the students to get the 'server experience': Campus supported Linux cluster (SLURM) AWS EC2 - student accounts and Jupyter Notebooks Google colab NSF Jetstream cloud (Access) Physical server in the corner of my office (current solution!)