Lecture 1 - Introduction

DSE 512

Drew Schmidt
2022-01-25

About Me

- Taught Math/Stats many years ago
- Worked in supercomputing for 10 years
- Supported hundreds of projects at large scale
- Currently work at small startup
- Very active R developer

Software

- pbdR https://pbdr.org
- fml Project
  - Code https://github.com/fml-fam
  - Blog https://fml-fam.github.io/blog/
- HPCRAN https://hpcran.org/
- GitHub https://github.com/wrathematics

Publications

- Schmidt, D., 2020, November. A Survey of Singular Value Decomposition Methods for Distributed Tall/Skinny Data. In 2020 IEEE/ACM 11th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems (ScalA) (pp. 27-34). IEEE.
- Hasan, S.S., Schmidt, D., Kannan, R. and Imam, N., 2019, December. A Scalable Graph Analytics Framework for Programming with Big Data in R (pbdR). In 2019 IEEE International Conference on Big Data (Big Data) (pp. 4783-4792). IEEE.
- Schmidt, D., Chen, W.C., Matheson, M.A. and Ostrouchov, G., 2017. Programming with BIG data in R: Scaling analytics from one to thousands of nodes. Big Data Research, 8, pp.1-11.

Course Files

- Syllabus, slides, assignments posted on Canvas
- Assignment 1 is live (it's short)
- Slides will be posted after presentation
- 2021's files including recordings (different instructor) are also available

About the Course

- Title: Introduction to Data Science and Computing II
- Description: Topics include: platforms for scalable computing including Map Reduce, Hadoop, Spark, and HPC, setting up computing in cloud, and modern data science work flows.

About the Course

The computer science of data science.

Course Content

- HPC and the Cloud
  - Remote computing
  - Shared file systems
  - Batch programming
- Software Management and Running Jobs
  - Modules
  - Containers
- Performance Acceleration and Parallelism
  - GPGPU
  - Distributed Programming

Technologies

- Assume you're comfortable with
  - a HLL (probably R or Python)
  - basic shell stuff
- Homeworks can use any *reasonable* programming language (I strongly recommend R or Python)
- Some things we will use in various capacities:
  - ssh, sftp/scp, parallel file systems, job schedulers, ...
  - AWS EC2 and ISAAC
  - Containers: Docker and Singularity
  - Some basic GPGPU
  - MPI

Course Structure

Class Hours

- Lecture component
- Interactive component

Outside of Class

- A homework assignment every few weeks
- Maybe a "super homework" (a project)

Next Time

- Overview of HPC and the Cloud
- Assignment 1 - due this Thursday 11:59pm

Questions?