DSA 8430
Parallel Computing for Data Analytics

Course Welcome

&

Overview

Course Information

Instructor

Grant Scott

GrantScott@missouri.edu

Canvas Site

https://umsystem.instructure.com/courses/94406

- Grade Book
- Everything else in Jupyter

Course Dates

- Feb 12 – March 19
- Break: March 20-April 2
- April 3 – April 22

Course Synopsis

- Develop an understanding of parallel computing, especially in distributed systems

- Use of parallel computing systems for data search, analytics, and machine learning

- Use a variety of systems for learning activities to force concepts to stick instead of "button-ology"

- Tutorial and Case Study designed learning activities

- A goal is for you to be able to leverage activities and technologies learned for research or profession

Course Learning Activities Design

- Course Learning Activities are going to be:
    - Platform Tutorial, Engineering, and Expert Videos
    - A few Jupyter-based programming
    - Mostly Cloud / Cluster construction and usage
- Labs: Set up and preparation
- Practices: Tutorial Style Walk Throughs & Case Studies
- Exercises: Extension or Repeat of Practices with alternative data

Course Modules

1. Introduction/Evolution of Parallel Computing Architectures

2. Divide and Conquer Algorithms

3. Parallel Data Search

4. Parallel Data Analytics

5. Cloud-based Distributed Computing with Kubernetes

6. Advanced Distributed Computing

7. Emerging Big Data Machine Learning Ecosystems

8. Final Project

Cloud and Cluster Systems

- GCP – Google Cloud Platform

- Rich set of web tools and services

- AWS – Amazon Web Services

- Largest commercial cloud

- NRP – National Research Platform

- Global scale Kubernetes cluster

(aka Nautilus)

- NRP – Jupyter Lab

- Jupyter Lab instance running on NRP

Getting Help

Slack

#parallel_analytics_8430

https://datagoggles.slack.com/archives/C02RU15UXGS

Graduate Teaching Assistants

- Tanmoy – Informatics PhD
- Anes – Computer Science PhD

Online Cohort Office Hours

- Thursday's 6:30 - 7:30 PM

https://umsystem.zoom.us/j/97887572840?pwd=WGxydzd3bkhtbnVramxYYTJnVkppdz09

Campus Cohort Tag-up & Office Hours

- Friday's 2:30 - 3:30 PM
- 240 Naka Hall
  - Required for F-1 Visa student residency guidelines

Getting Started

Setup (Practices)

For this course, we will be working on a variety of distributed systems (clusters, grids, clouds) three remote environments. In each of the practices, you will bring an artifact of your project to each module points.

- Accessing the NSF Nautilus Cluster
- Setting up your AWS Account
- Setting up your GCP Account
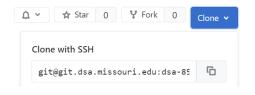
Cloning onto the Nautilus System

In the practice you set up your Nautilus Jupyter Hub connection and created a public key. Additionally, you stored that public key into the DSA Gitlab

In this exercise, you will build upon that to manually clone the repository from DSA systems into Nautilus.

Part A: Login to Nautilus JH

The Mizzou Jupyter Hub instance on Nautilus is located at:

https://mizzou.nrp-nautilus.io/

Part B: Login to DSA Gitlab

The Mizzou DSA Gitlab instance is located at:

https://git.dsa.missouri.edu/

Part C: Navigate to your course project in Mizzou Gitlab, get Clone link

🔔 ∨   ☆ Star 0   🍴 Fork 0   Clone ∨

Clone with SSH

git@git.dsa.missouri.edu:dsa-85 [📋]

Part D: Within Nautilus JH, Clone down the repository

1. Ensure you have your environments/accounts set up and ready to go for next week

2. Work on cloning a second copy of the course into one of the Nautilus Jupyter Lab environment