UNIVERSITY of ROCHESTER | SIMON BUSINESS SCHOOL

CIRC Workshop for Simon Ph.D. Students

Date and Location: see email for zoom link, Aug 22, 2020

The workshop material is available here

Instructor:

Shengyu Zhu
Marketing PhD Student

Email: shengyu.zhu@simon.rochester.edu

Training Information

This workshop aims to cover how to use the High-Performance Computing (HPC) for data-driven research projects involving big data sets. Specifically, I will cover the HPC resources freely available for all Simon Ph.D. students, the BlueHive cluster at the Center for Integrated Research Computing at the University of Rochester. It would be most useful if your current project involves big data set (>1/2 of your desktop memory, but <100G) that cannot be loaded at one time due to memory limitation of your own desktop/laptop.

1. **Some software must be installed for usage of CIRC/BlueHive, please install them before the workshop:**

Off-campus connection

- VPN is needed if you want to access CIRC off-campus. http://tech.rochester.edu/services/remote-access-vpn/

CIRC account and SimonX node access

- You need to request a CIRC account to use Bluehive, please register one if you haven’t done yet (Sue send an email about this, please take a look at that email for details):

Two-Factor Authentication (Duo) is needed each time you log in to CIRC, please set up this if you haven’t already:

- https://tech.rochester.edu/services/two-factor-authentication/

CIRC related software, exact download link cannot be given publicly, please go to CIRC website -> Info.CIRC->download (you need to use university VPN to access this website)

- Windows: Install FastX (GUI interface with BlueHive), WinSCP (Transferring files between Windows and BlueHive)
- Mac: Install FastX, fetch(Transferring files between Mac and BlueHive)

2. Workshop topics:

Overview of CIRC

- *What is BlueHive*
- *Why/When do you need to use BlueHive*
- *Available software on BlueHive*
- *BlueHive Storage*

GUI *(graphical user interface)* connection to BlueHive

- *Using FastX to connect to BlueHive*
- *How to use GUI software like Rstudio, Stata, Matlab on BlueHive*
- *How to transfer files between BlueHive and your own desktop*
- Using JupyterHub to interact with BlueHive

CLI (command-line interface) connection to BlueHive: using SSH

- Linux basics, *GUI (graphical user interface) vs. CLI (command-line interface)*
- *Connect to BlueHive using ssh in Terminal*
- *Login Node V.S. Compute Node*

- *Using the command line to run R and Python*
- *General Linux Command Syntax*
- *Using the command line to navigate files and directories in Linux*
- *Files and directory organization*
- *Use Python for automation to create folders and moving files*

Access Rstudio server/Jupyter notebook on BlueHive by using local web

Batch Mode

- *Run R/Python script in non-interactive mode*
- *Using R markdown to track the R script output and write research log*
- *Make Bash/R script executable*

SLURM (Simple Linux Utility for Resource Management) job scheduler

- *Submit/ track status/cancel jobs*
- *How to see how much computation resources (of SimonX node) are not occupied and available for usage in real time*
- *Job arrays/pipeline of jobs with dependencies*

Version control

- *Why version control?*
- *Git, Bitbucket, and SourceTree*

Toy project demonstration (*I will use a toy project to demonstrate my own workflow with BlueHive*)

- *Use Python to execute all the code of your research project from the beginning to the end.*
    - *When you clean data using R, run a regression using Stata, writing a paper using Latex, and making slides using Beamer.*
- *Using Rmarkdown for research log*

3. Other software recommended for a better experience of using BlueHive:

The following software will be used during this workshop, you would have a better experience of using BlueHive if you finish the installation and setup in your laptop before you come.

- Version Control Software
    - I will demonstrate bitbucket with git, you can skip this part if you are already familiar with the usage of GitHub. But Dropbox is not a good tool for version control.
    - Creating an account of Bitbucket: https://bitbucket.org/account/signup/
    - Install SourceTree: https://www.sourcetreeapp.com/
- Python and R
    - I use python 3 to automate the execution of all the code of my research project. https://www.anaconda.com/download/
    - I use R and Rstudio for cleaning/analyzing data. https://www.rstudio.com/products/rstudio/download/preview/

- Console Emulator for Windows
    - http://cmder.net/ (Download the full version). This software is easier to use than cmd (command prompt) and PowerShell.
    - Mac users don’t need similar software.

- Sublime package
    - It can enable you to test your local code directly with the cluster, without copy and paste code every time.
    - Install Sublime text editor first https://www.sublimetext.com/
    - Install package control in the sublime text editor. https://packagecontrol.io/installation