Python for Geospatial Big Data and Data Science Using the FASRC

Command Cheat Sheet

# Web Resources

FASRC Documentation & tutorials

<https://docs.rc.fas.harvard.edu/>

<https://docs.rc.fas.harvard.edu/kb/running-jobs/> – use SLURM

The FASRC web interface

<https://rcood.rc.fas.harvard.edu/>

This provides access to the webapps, such as Jupyter

Monitor currently running SLURM jobs

<https://portal.rc.fas.harvard.edu/login/?next=/jobs/>

Browser-Shell

<https://rcood.rc.fas.harvard.edu/pun/sys/shell/ssh/COMPUTE_NODE.rc.fas.harvard.edu>

This opens a terminal in your browser to connect to a compute node. Replace “COMPUTE\_NODE” with an actual ID, such as, for example, “holy7c24603”. This works only if you have an (interactive) session running on the node you want to connect to.

GitHub repo with all workshop files

<https://github.com/cga-harvard/python-workshop-gis-big-data>

Google Drive repo with all datasets we’ll work with

<https://drive.google.com/drive/folders/1k6G9hu8IiaZmMt3MgDw-3vBTtLTpyqdK?usp=sharing>

# Connect to the Cluster

Open a SSH shell to the login node

ssh USERNAME@login.rc.fas.harvard.edu

Request a compute node and start an interactive session

srun --pty -p test --mem 1000 -t 0-01:00 /bin/bash

This gives you 1h on a test partition with 1GB or memory

# Create, start and use Python environments

Load the Python module

module load Mambaforge/23.3.1-fasrc01

Create a python environment and install packages

mamba create -n workshop python=3.9 --file requirements.txt

This command expects a file “requirements.txt” in the same directory you call this command from. This file lists all packages one by one per line.

Activate this python environment

mamba activate workshop

Run a Python script

Python script.py

To run a Jupyter notebook on the command line, convert it to a python script first

jupytext --to py jupyter\_notebook.ipynb

This creates a “jupyter\_notebook.py” file in the same directory.

# Monitor jobs and processes

List running SLURM jobs

squeue

Monitor the load on a system

htop

# Working with Data

All datasets we’ll work with are accessible from all nodes and all users at

/n/holyscratch01/cga/python-workshop-gis-big-data

To copy data to the cluster, you can use SCP

scp -r ./\* USER@login.rc.fas.harvard.edu:/PATH/TO/DESTINATION/

Replace USER and /PATH/TO/DESTINATION/

Similarly, to copy data from the cluster to your local machine, you can use

scp -r USER@login.rc.fas.harvard.edu:/PATH/TO/SOURCE ./

Replace USER and /PATH/TO/SOURCE/

# Create a sbatch file to submit as a SLURM job

The following is a template for a sbatch file to run a python script as a SLURM job. Save the following as, for example, “project\_sbatch\_job.sh”, and adjust the parameters:

#!/bin/bash

# https://docs.rc.fas.harvard.edu/kb/running-jobs/#articleTOC\_8

#SBATCH -c 2 # Number of cores (-c)

#SBATCH -t 0-00:10 # Runtime in D-HH:MM, minimum of 10 minutes

#SBATCH -p test # Partition to submit to

#SBATCH --mem=16000 # Memory pool for all cores

#SBATCH -o /n/home01/USER/job\_stdout\_%j.out # log STDOUT to file

#SBATCH -o /n/home01/USER/job\_errout\_%j.err # log errors to file

# load modules

module load Mambaforge/23.3.1-fasrc01

# set python environmant

mamba activate workshop

# run code

python global\_precipitation\_sentiment.py

You can then simply run sbatch project\_sbatch\_job.sh to submit the job to SLURM. Caution: ensure your script saves results to a destination that is accessible to you after the job finished.