USC Bootcamp: Data Management

Asya Shklyar
May 2021

Where is your data?

Your laptop/desktop

- External drive
  - USB2
  - USB3
  - External RAID/SCSI/JBOD

USC HPC

- Scratch
- Project (Home?)

Cloud

- AWS S3
- GCP
- Azure

Your previous institution

Collaborator

Snowball

Backups

Dropbox

Box

Google Drive

OneDrive

GitHub?

Official publishing repos (Driad etc)

Where do you want your data to be?

Stay where it is for a year

Actively working on it

Need to access from time to time

Multiple people

Combining datasets

Visualization

Versioning

Tools (transfer)

scp - no resuming (also sftp)

rsync - can be slow single-threaded

- Parallel RSync
  - Multiple options
  - Preserving permissions

rclone

- Setup via GUI or CLI

Globus

- Personal
- Institutional Enterprise Endpoint

DTN (hpc-transfer1/2)

File system - project may be faster

Aspera? UDP vs TCP

Considerations

Checksum before and after

Encryption/Decryption (manage keys)

Resuming

Throughput/Time

One system or many (include parallel fs discussion)

Small files vs large files

- Tar
- Gzip (not always efficient eg sequence data)

Duplicated data

Symlinks

Permissions

- Groups
- AD vs local (UIDs)

Network

- Wired
  - 1 GB
  - 10 GB
  - 100 GB
- Wireless

Latency/MPI

Data Formats more complex than txt of Excel

HDF5 https://www.hdfgroup.org/solutions/hdf5/

GIS https://gisgeography.com/gis-formats/

iRods and other workflow tools

Pegasus - no GUI; can be used for any discipline

iRods is pretty involved and is mostly for moving data between layers

Snakemake popular in Bioinformatics

Other workflow tools for Bioinformatics:

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008622

Containers

Containers do not store data inside them, if you want your data save it to a location that is more permanent

If you modified the container significantly, save it or a copy

Troubleshooting is also more involved (logs inside the container)

Use “time” command for benchmarking your transfer speed

Installed as a module, just do “module load time”

Example: “time scp yourfile yourusername@remotesystem”

If you want to save the output (gets lost if you disconnect from terminal):

screen or tmux

Redirect to a file - add: “> output.txt &” and “bg”