Research Computing at Mines Workshop Data Management

September 5, 2024

Presented by:
Nicholas A. Danes, PhD
Computational Scientist
Research Computing Group (IT)

Goals

- Data Policies
- How to check data usage on HPC
- Job Data Best Practices
- Archiving data
- Ways to transfer data in/out HPC

Data Policy on HPC's

- Every HPC/Cyberinfrastructure group will have a different data policy!
  - Check their websites for specific policy

Wendian

Data Policy

There are a few data solutions we offer at Mines for research data:

- Orebits - High capacity research storage *not* connected to HPC

The following are only on Wendian & Mio:

- `/scratch` - Active research data, subject to 180 day data purge
- `/projects` - Scratch research data used within a research project that shared with multiple users, also subject to 180 day data purge. The PI must request a projects directory with the list of authorized users.
- `/sets` - Long term data storage available on HPC. The PI must request a sets directory with a list of authorized users.

Note that all data on these directories have no redudancy, so please keep up with your own backups of active research data.

The table below breaks down the purge policy and associated costs of each the data solutions:

Type	Purge Policy	Cost	Redunancy
Scratch (`/scratch`)	>180 days	Free	No
Projects (`/projects`)	>180 days	Free	No
Wendian Long-Term Storage (`/sets`)	None	Free	No
Orebits	None	$2/TB/month	Yes

Alpine (CU Boulder)

All users are allocated space on the `/home` and `/projects` filesystems. In addition, separate `scratch` directories are visible from Alpine and Blanca. These scratch directories are hosted on separate, high-performance filesystems designed to support intensive, parallel I/O operations.

Please note that the use of `/home` or `/projects` for high-performance I/O may negatively affect the environment for all users. As a result, all compute jobs should write to the appropriate `scratch` filesystem. **Users performing intensive I/O on the `/home` or `/projects` filesystems will have their jobs terminated and may have their accounts temporarily disabled.**

The Home Filesystem

Every user is allocated 2 GB of space on the `/home` filesystem in a subdirectory corresponding to their user name (e.g., `/home/janedoe`). Home directories are backed up frequently and are intended for the use of their owner only; sharing the contents of home directories with other users is strongly discouraged. Your `/home` directory is a good place to store source code, small compiled programs, and job scripts.

The Projects Filesystem

Each user has access to a 250 GB of space in their subdirectory of `/projects` (e.g., `/projects/janedoe`). As with the `/home` system, these directories are visible from all Research Computing nodes and are regularly backed up. The projects directory is intended to store software builds and smaller data sets.

Scratch Filesystems

Alpine users are provided a subdirectory on `/scratch/alpine`, the high-performance parallel scratch filesystem meant for I/O from jobs running on that system (e.g., `/scratch/alpine/janedoe`). By default, each user is limited to a quota of 10 TB worth of storage space and 20M files and directories. If you need these limits increased, see our Scratch Quota Increases policy. Blanca users should write to `/rc_scratch/janedoe` instead of `/scratch/alpine`.

Scratch space should be used for all compute jobs run on Alpine or Blanca. These high-performance scratch directories are **not backed up**, and are not appropriate for long-term storage. Data may be purged at any time subject to overall system needs. Files are automatically removed 90 days after their initial creation.

Checking data usage on Wendian

- For individual files:
  - Use Wendian OnDemand's files interface
    - Can see individual file sizes
    - Will warn you when your home directory is close to limit (20 gb)
    - Not ideal for finding the largest files in your directories
  - Use an FTP client
    - FTP clients like Filezilla, WinSCP can give you insight on how much storage a directory contains
      - Filezilla (Windows/macOS/Linux): https://filezilla-project.org/
      - WinSCP (Windows only): https://winscp.net/eng/index.php
      - Wendian FTP information (requires GlobalProtect VPN):
        - Host: sftp://wendian.mines.edu
        - Username: Mines username
        - Password: Mines SSO password
        - Port: 22

Checking data usage on Wendian

- For most use cases, like looking for the largest folders, use the Linux terminal
  - du command can provide ways to look for large files in your directories.
    - https://www.geeksforgeeks.org/du-command-linux-examples/
    - We will explore this in the lab/demo!

Job Data Best Practices

- When running jobs through Slurm, you will end up with lots of output files:
  - Standard output
  - Standard error output
  - Data files from your model (csv, hdf5, txt, etc)
- Best to organize this with the job script, not after the fact!
- OPTIONAL: Use git to manage input files, source code

Job Data Best Practices

- General Recommendations
  - Label output files using SLURM JOB ID
    - You can access this in the job script using ${SLURM_JOBID} environment variable
  - Create output directories using SLURM JOB ID
    - Typically this involves copying the input files to a new folder called ${SLURM_JOBID} and then running the code from there
    - We did this with the GROMACS lab!
  - Periodically check in on old run data
    - Archive it using tar or zip
      - Makes the file smaller and easier to transfer

Job Transfer Options

- Once you have data to move off of Wendian, how to do it?
  - Command-line options
    - scp – You used this in the day 1 lab!
      - Easy to use, but if the file transfers are large it's not ideal if you need to disconnect the machine doing the transfer
    - rsync – Can be used incrementally to transfer a file, but also would prefer something that you don't have to check on
  - Open OnDemand – use Download/Upload feature in the Files tab
    - Works with directories
  - Globus – Provides a web interface for high throughput data transfers
    - Website: https://app.globus.org/
    - Also provides a means to transfer data between institutions
      - Requires Globus license on both ends

Lab #1: Hands-on Lab for Data Transfer

- Check how much storage we've used with du
- Archive and transfer data from our GROMACS runs 3 different ways:
  - Command line using scp/rsync
  - Using FTP via Filezilla
  - Globus