Visual Data Management at NERSC

Lisa Gerhardt
National Energy Research Scientific Computing Center
Berkeley Lab
Berkeley, California, United States
lgerhardt@lbl.gov

Annette Greiner
National Energy Research Scientific Computing Center
Berkeley Lab
Berkeley, California, United States
amgreiner@lbl.gov

Zelong Li
Department of Computer Science
Iowa State University
Ames, Iowa, United States
zelongl@iastate.edu

Harrison Horton
Department of Computer Sciences and Information Technology
University of Saint Mary
Leavenworth, Kansas, United States
harrisonhor10@gmail.com

Keywords—data management, visualization, web, national user facility

I. Extended Abstract

Wrangling data at a scientific computing center can be a major challenge for users, particularly when quotas may impact their ability to utilize resources. In such an environment, a task as simple as listing space usage for one's files can take hours. The National Energy Research Scientific Computing Center (NERSC) has roughly 60 PBs of shared storage utilizing more than 2.2 billion inodes, and a 150 PB high-performance tape archive, all accessible from the Cori supercomputer[1]. As data volumes increase exponentially[2], managing data is becoming a larger burden on scientists. To ease the pain, we have designed and built a "Data Dashboard" (Figure 1). Here, in a web-enabled visual application, our 7,000 users can easily review their usage against quotas, discover patterns, and identify candidate files for archiving or deletion.

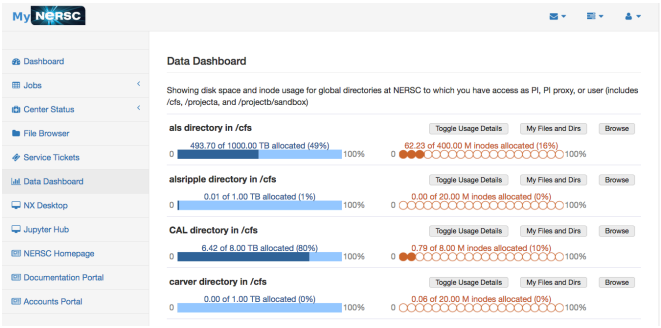

Fig. 1. The Data Dashboard at page load within the MyNERSC user portal. Each row shows usage against quota for space (left colum) and inodes (right column) for one project directory.

We are also developing a "PI Toolbox" (Figure 2) to allow scientists to directly control the permissions of their files and directories. In the last year, scientists at NERSC moved 260 PB of data between centers. As these numbers grow, we are also gaining new users with less experience in transferring large files efficiently. We are therefore working on a "PB Data Portal" to facilitate sharing these large volumes of scientific data. Existing commercial options lack the access management we need to release to many users, and noncommercial solutions have thus far been customized to narrow user segments or specific storage systems. We have therefore built a new, generalizable framework on tools that come with common file system software, like Spectrum Scale[3], Lustre[4], and HPSS[5]. We describe the process for developing our tools, the framework supporting them, and the challenges for such a framework moving into the exascale age.

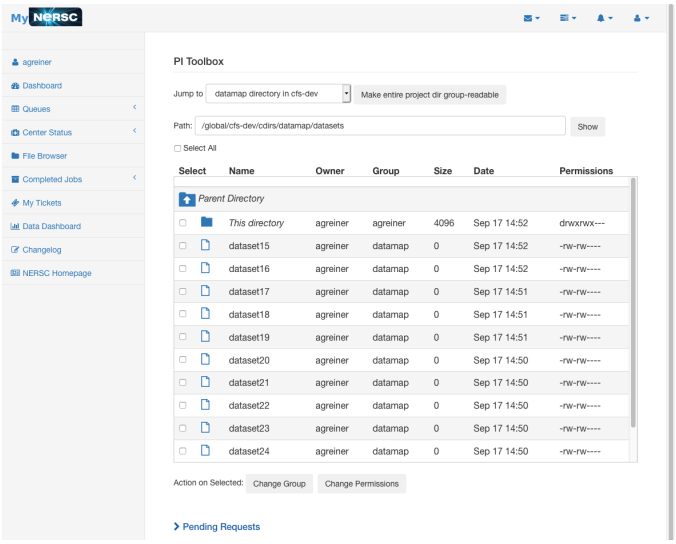

Fig. 2. The PI Toolbox after user selection of a project directory from the "Jump to" menu. Users can reset group ownership and permissions for the entire directory, or change the group ownership or the group permissions of individual files and directories.

References

[1] http://www.nersc.gov/systems/, 2020.
[2] S. Habib, et al., ASCR/HEP requirements review report (2016). ArXiv:1603.09302v2.
[3] http://www-03.ibm.com/systems/storage/spectrum/scale/, 2020
[4] http://lustre.org/, 2020
[5] http://hpss-collaboration.org, 2020