Integration of Clowder Research Data Framework with NCSA Labs Workbench

Maxwell Burnette, Rob Kooper, Michael Lambert
National Center for Supercomputing Applications
University of Illinois
Urbana, IL, USA
{mburnet2, kooper, lambert8}@illinois.edu

ABSTRACT

NCSA has two open-source applications focused on research data management and accessibility: Clowder [1] is a scalable data repository with extensive metadata search capabilities and support for automated extraction of metadata from uploaded files, and Labs Workbench [2] is an application catalog capable of registering and running instances of containerized research environments in the cloud. Both of these applications are designed around making research data available and interactable for a broad set of communities with minimal effort from the user. In the summer and fall of 2021, we are integrating these two applications together in order to enable users to seamlessly move data between Clowder instances where files are organized and Workbench applications where files can be examined and processed, and outputs can be shared back into the Clowder environment.

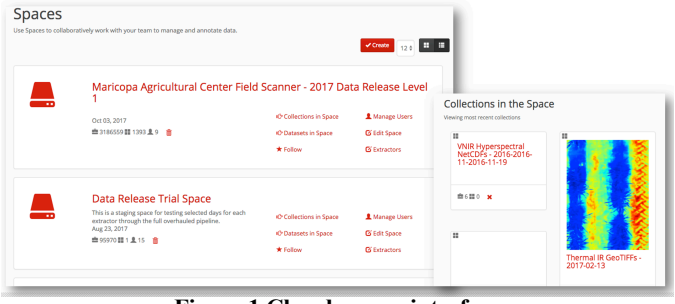

Figure 1 Clowder user interface

As scientific research has increasingly moved towards cloud computing, big data and dense software dependency trees, it has also become increasingly difficult for researchers to perfectly replicate the environments necessary to house and analyze these datasets. Big files are slow to move around, individual laptops may not have the necessary storage or processing power, and differences between operating systems cause inconsistencies. Clowder has a user-friendly GUI for uploading files and datasets, tagging and searching for them, and submitting them to extractors for processing. Labs Workbench provides a cloud management environment for containers running applications in browsers, such as Jupyter notebooks or GIS interfaces. We recognized an opportunity to move these complementary feature sets together to build a complete platform of data storage, sharing, analysis and discovery.

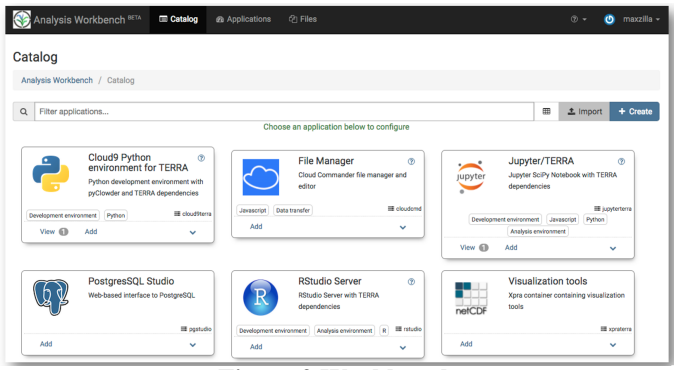

Figure 2 Workbench

The goal of our integration is to provide a direct path for datasets to move between Clowder instances and Workbench applications easily and seamlessly, particularly in environments where data storage can be mounted on co-located Clowder and Workbench virtual machines. Users will see options in the Clowder interface to send their data to their chosen Workbench instance, while in Workbench the files will land in a shared home directory accessible between all of the user's applications. Behind the scenes, we have leveraged Clowder's existing interface for submitting datasets to extractors, which are containers running prepared scripts that do a single task like running text-to-speech on an audio file or creating a face recognition mask on a photograph. This extractor framework is widely used, and we wanted to provide simple ways for researchers to develop new extractors and share them back to the community. Workbench was a great opportunity to provide a simple path: researchers can move a sample subset of data from Clowder to Workbench, develop an algorithm to process meaningful metadata from it, and deploy that algorithm as a Clowder extractor in an identical container environment to process the full dataset. In cases where data are co-located on a shared mount, large datasets can be moved and processed instantly [3].

Keywords cloud computing; research platforms; containerization; data repositories

REFERENCES

[1] Luigi Marini, Indira Gutierrez-Polo, Rob Kooper, Sandeep Puthanveetil Satheesan, Maxwell Burnette, Jong Lee, Todd Nicholson, Yan Zhao, and Kenton McHenry. 2018. Clowder: Open Source Data Management for Long Tail Data. In Proceedings of the Practice and Experience on Advanced Research Computing (PEARC '18). ACM, New York, NY, USA, Article 40, 8 pages. DOI: https://doi.org/10.1145/3219104.3219159

[2] Craig Willis, Mike Lambert, Kenton McHenry, and Christine Kirkpatrick. 2017. Container-based Analysis Environments for Low-Barrier Access to Research Data. In Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact (PEARC17). Association for Computing Machinery, New York, NY, USA, Article 58, 1–4. DOI:https://doi.org/10.1145/3093338.3104164

[3] Maxwell Burnette, Rob Kooper, J. D. Maloney, Gareth S. Rohde, Jeffrey A. Terstriep, Craig Willis, Noah Fahlgren, Todd Mockler, Maria Newcomb, Vasit Sagan, Pedro Andrade-Sanchez, Nadia Shakoor, Paheding Sidike, Rick Ward, and David LeBauer. 2018. TERRA-REF Data Processing Infrastructure. In Proceedings of the Practice and Experience on Advanced Research Computing (PEARC '18). Association for Computing Machinery, New York, NY, USA, Article 27, 1–7. DOI:https://doi.org/10.1145/3219104.3219152M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.