Container-based Analysis Environments for Low-Barrier Access to Research Data

Craig Willis[1,3], Mike Lambert[1,3], Kenton McHenry[1,3], and Christine Kirkpatrick[2,3]
[1]National Center for Supercomputing Applications
[2]San Diego Supercomputer Center
[3]National Data Service
{willis8, lambert8, mchenry}@illinois.edu, christine@sdsc.edu

ABSTRACT

The growing size of high-value sensor-born or computationally derived scientific datasets are pushing the boundaries of traditional models of data access and discovery. Due to their size, these datasets are often only accessible through the systems on which they were created. Access for scientific exploration and reproducibility is limited to file transfer or by applying for access to the systems used to store or generate the original data, which is often infeasible. There is a growing trend toward providing access to large-scale research datasets in-place via container-based analysis environments. This paper describes the National Data Service (NDS) Labs Workbench platform and DataDNS initiative. The Labs Workbench platform is designed to provide scalable and low-barrier access to research data via container-based services. The DataDNS effort is a new initiative designed to enable discovery, access, and in-place analysis for large-scale data, providing a suite of interoperable services to enable researchers, as well as the tools they are most familiar with, to access and analyze these datasets where they reside.

Categories and Subject Descriptors

H.3.3 []

Keywords

1. INTRODUCTION

Sensor-based, research-computing, and high-performance computing systems now produce massive amounts of data. Traditional data publishing services, such as community and institutional repositories, are not equipped to handle these very large researrch datasets. Increasingly, researchers are leaving their data on research computing infrastructure or turning to cloud-based services to facilitate sharing. However, research-computing and HPC centers are typically not prepared to support long-term storage and access, as is often required by publishers, further disconnecting these research products from traditional discovery methods. New models of in-place publishing are needed to connect research computing and research data publishing infrastructure to support discovery and access for re-use and reproducibility.

Institutional and community repositories are beginning to support "remote" data publishing. With this approach, researchers are responsible for arranging data storage with best-effort preservation and datasets are effectively published in-place. Researchers provide the repository with descriptive metadata about the datasets including methods of access and are assigned a digital object identifier (DOI). Under this model, users can discover these datasets and information about how to access them via traditional discovery mechanisms.

Existing approaches to providing access to these hosted datasets are inefficient and often ineffective. Typically, research-computing infrastructure has supported two basic models of data access: transfer and direct access to the hosting system. Transfer services, such as Globus Online, enable users to access these datasets by transferring them to local systems via GridFTP. However, in many cases, the datasets are too large to move. Researchers can apply for access to the systems used to store and generate the original data, but access is often restricted and the application process involved and time consuming.

Container-based analysis environments are emerging as a mechanism to provide low-barrier access to research data in-place. Using container technology such as Docker, projects provide access to large datasets through custom analysis environments, such as Jupyter notebooks or Rstudio. In these systems, remote users are able to register for an account via web-based interfaces that allow them to launch specialized, resource-constrained analysis environments to explore data in-place. Examples includes yt.Hub, SciServer, Galaxy Portal, and the ARPAE TERRA-REF project.

This paper describes two initiatives of the National Data Service (NDS) to facilitate in-place access to large research datasets. The Labs Workbench platform, used by the ARPAE TERRA-REF project, is designed to support exposing research data via customizable container-based analysis environments. The DataDNS initiative is intended to connect various container-based analysis systems to the traditional publishing and discovery platforms to enable discovery, access, and analysis of datasets where they reside.

This paper is organized as follows. Section 2 provides a description of container-based analysis environments. In Section 3 we describe the NDS Labs Workbench system and specifically how it's being used for the ARPAE TERRA-

PEARC '2017 New Orleans

REF project, followed by a description of the NDS DataDNS initiative in Section 4 and next steps in Section 5.

2. RESEARCH IN THE CLOUD

Space permitting, we can talk about the cloud-computing trend and services like Amazon Public Datasets.

3. CONTAINER-BASED ANALYSIS ENVIRONMENTS

Container technology is increasingly used in research computing as an alternative to hypervisor-based virtual machines for the packaging, deployment, and execution of software. Containers are considered to be more resource-efficient and provide a clear and light-weight abstraction for packaging[1]. Because of this, container management platforms such as Docker[2] and rkt[3] have achieved widespread adoption.

Container-based analysis environments refers to systems that leverage container technology for the packaging and execution of interactive scientific research software. Containers are currently used in a variety of systems providing interactive and non-interactive access to research data. Systems such as yt.hub, SciServer, Galaxy Portal, Whole Tale, and the TERRA-REF Analysis Workbench provide interactive containerized environments based on software including Jupyter, RStudio, and MatLab. Other systems, such as Cyverse and SciPort[3], leverage containers for the management and execution of scientific workflows, which we'll refer to as *container-based execution environments*. These are more likely to fit into batch-scheduling infrastructure.

Todo: mention something about containers for preservation.

4. LABS WORKBENCH

Labs Workbench is an open-source platform developed by the National Data Service Consortium to facilitate sharing, discovery, evaluation, and development of research data management and analysis tools. The platform leverages Docker containers and the Kubernetes container orchestration framework to provide turn-key access to community developed tools. The Labs Workbench system includes the following features:

1. Scalable container orchestration via Kubernetes
2. Catalog of community-contributed tools
3. Ability to import custom tools and personal catalogs
4. Authentication
5. Logging and monitoring
6. Shared filesystem via GlusterFS

Labs Workbench has been effectively used in classroom and workshop environments to provide consistent, web-based access to applications. In early 2017, the Labs Workbench was deployed by the ARPAE TERRA-REF project to provide customized interactive analysis environments for the TERRA-REF reference dataset.

[1]The reader can refer to more detailed analysis of containers in [4, 2, 1]
[2]http://www.docker.com
[3]https://coreos.com/rkt

5. USE CASE: TERRA-REF

The ARPA-E TERRA program is focused on cutting-edge techniques for the improvement of biofuel crops in part through the creation and publication of a large public reference dataset, TERRA-REF, and associated compute pipeline. The TERRA-REF data storage and computing system will provide researches with access to over 2PB of raw sensor and derived data hosted on the NCSA ROGER system and made available via Globus, Clowder, BETYdb, and the Labs Workbench. Researchers can also apply for access to the ROGER system directly for access via virtual machines and batch-processing.

In early 2017, the Labs Workbench platform was used to host a workshop, enabling 50 participants to explore the TERRA-REF data using customized Jupyter, Rstudio and other interactive environments. Shortly after, an instance of Labs Workbench, named the TERRA-REF Analysis Workbench, was deployed on the NCSA Nebula OpenStack system to provide ongoing access to the TERRA-REF data for project contributors. Custom containers were created based on Jupyter, Rstudio, and Wetty to include dependencies required to access and use the TERRA-REF data including NetCDF/NCO and OpenCV. Additional containers were created to support system development, including a Python IDE for Clowder extractors. PostgresStudio was added to provide access to the BETYdb PostGIS database.

The Labs Workbench platform was later used to support the two-week PI4 Computational Mathematics Bootcamp, an NSF funded program for graduate students to gain computational skills needed in scientific and engineering research labs. Students were provided with access to the TERRA-REF containers and data for a series of tutorials. This demonstrates that a system used to provide access to research data for analysis and re-use can also be an effective system for training and education.

6. DATADNS

The Labs Workbench is just one example of an emerging trend to provide access to research data via cloud and container-based environments. Other examples include yt.hub, an innovator in this area, that uses Girder and Jupyter to enable access to access to analysis environments for a number of fields including astronomy and cosmology. The yt.hub architecture is currently being adapted to support the Renaissance Simulations Laboratory (RSL), providing access to the Renaissance Simulations generated on BlueWaters, and the Whole Tale project. Similarly, the SciServer Compute system provides access to astronomy, materials science, turbulence, and earth science datasets via Python, Rstudio, MatLab and Jupyter containers. SciServer also offers a container-based job scheduling component not available in these other systems.

DataDNS is an initiative to support discovery, access, and in-place analysis for large-scale data based on the availability of these and related services. DataDNS is intended to connect traditional research data publishing infrastructure to research computing infrastructure. The goal is to enable users to not only discover these datasets, but also to enable access for in-place analysis. DataDNS will provide an active registry and resolution service to connect researchers to locations with access, analysis, and compute capabilities.

Through DataDNS, researchers can register datasets to

enable access in-place. Datasets will include information about associated capabilities, such as transfer, container-based, cloud-based or batch compute. Researchers will not only be able to find the data, but also access it and perform analysis, when possible.

7. CONCLUSION

8. ACKNOWLEDGMENTS

This work was supported in part by X. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the X.

9. REFERENCES

[1] D. Bernstein. Containers and cloud: From lxc to docker to kubernetes. *IEEE Cloud Computing*, 1(3):81–84, Sept 2014.

[2] W. Felter, A. Ferreira, R. Rajamony, and J. Rubio. An updated performance comparison of virtual machines and linux containers. In *2015 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 171–172, March 2015.

[3] W. Gerlach, W. Tang, K. Keegan, T. Harrison, A. Wilke, J. Bischof, M. D'Souza, S. Devoid, D. Murphy-Olson, N. Desai, and F. Meyer. Skyport: Container-based execution environment management for multi-cloud scientific workflows. In *Proceedings of the 5th International Workshop on Data-Intensive Computing in the Clouds*, DataCloud '14, pages 25–32, Piscataway, NJ, USA, 2014. IEEE Press.

[4] S. Soltesz, H. Pötzl, M. E. Fiuczynski, A. Bavier, and L. Peterson. Container-based operating system virtualization: A scalable, high-performance alternative to hypervisors. *SIGOPS Oper. Syst. Rev.*, 41(3):275–287, Mar. 2007.