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 accessible only 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 [**Information Systems**]: Computing Platforms

Keywords

Research data; Container-based analysis environments

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 very large research datasets. Increasingly, researchers are leaving their data on research computing infrastructure or turning to cloud-based services to facilitate sharing and re-use. 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 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 for data storage with best-effort preservation and datasets are effectively published in-place. Researchers provide the repository with descriptive metadata 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. However, it is up to the researcher or site where the data is hosted to support methods of access.

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. Data 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 or copy. In these cases, researchers can apply for access to the systems used to generate and store the original data, but access is often restricted and, if available, 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 server. Typically, 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 [15], SciServer [9], and the TERRA-REF Analysis Workbench [14].

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

2. 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, provide a clear and light-weight abstraction for packaging, and are increasingly seen as a possible solution for the preservation of research software [10]. Because of this, container management platforms such as Docker[1] and rkt[2] have achieved widespread adoption. This section focuses on container use for access to research data. For more information about container technology, the reader is directed to more detailed analysis in [1, 4, 13].

Container-based analysis environments refer 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 [15], SciServer [9], Whole Tale [7], and the TERRA-REF Analysis Workbench [14] provide interactive containerized environments based on software including Jupyter [6], RStudio [12], and MATLAB [8]. Other systems, such as Cyverse[2] and SciPort[5], leverage containers for the management and execution of scientific workflows. We refer to these non-interactive environments as *container-based execution environments*, which are more likely to fit into workflow systems and batch-scheduling infrastructure.

Increasingly, research systems developers are seeing the need to support the computational research lifecycle from preliminary exploration through development and execution either on dedicated cloud-based or HPC systems. Container-based analysis environments support the early stages of exploratory analysis and development, while container-based execution environments serve as a bridge to traditional batch-compute systems.

3. LABS WORKBENCH

Labs Workbench is an open-source platform developed by the National Data Service Consortium (NDSC) 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.

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.

3.1 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[3]. The TERRA-REF data storage and computing system will provide researchers with access to over 2PB of raw sensor and derived data hosted on the NSF ROGER system and made available via Globus, Clowder, BETYdb, and the Labs Workbench. Researchers can also apply for accounts on the ROGER system directly for access via batch-processing. The TERRA-REF team is working with the Dryad data repository[3] to publish the remotely-hosted reference dataset.

In early 2017, the TERRA-REF team used the Labs Workbench platform to host a workshop, enabling 50 participants to explore the reference dataset using domain-specific 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 other services to include dependencies required to access and use the TERRA-REF data such as NetCDF/NCO and OpenCV libraries. Additional containers were created to support system development, including a Python IDE for Clowder extractors and PostgresSQL Studio 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.

The TERRA-REF Analysis Workbench demonstrates that the same system can be used to provide access to data products for scientific exploration and reuse; collaborative system development; as well as education and training.

3.2 System architecture

The Labs Workbench platform is based on the Kubernetes[4] container orchestration framework, deployed using the OpenStack[5] virtualization platform. Figure 1 illustrates the basic architecture. First, a set of virtual machines and volumes is provisioned via the OpenStack API. Second, Kubernetes core services (controller, kubelet, and etcd) are deployed on each node. Third, Labs Workbench core and system services are deployed via Kubernetes. Labs Workbench core services include the Nginx ingress load-balancer (nginx-ilb), which provides authenticated access to running containers in the cluster, as well as a thin REST API server (ndslabs-apiserver) an Angular user interface (ndslabs-webui). System services include standard logging, monitoring and alerting (LMA) tools such as Kibana/ElasticSearch, Grafana/Heapster, and Nagios. The system includes a private Docker registry for image caching. Via Labs Workbench, users can launch a variety of Docker-based applications including but not limited to Jupyter and

---

RStudio environments. The service catalog itself is a set of customizable Javascript Object Notation (JSON) objects currently stored in Github[6]. The Labs Workbench platform uses Gluster[7] to provide shared storage between nodes and containers. Additional shared network volumes can also be mounted via other network services, such as NFS.
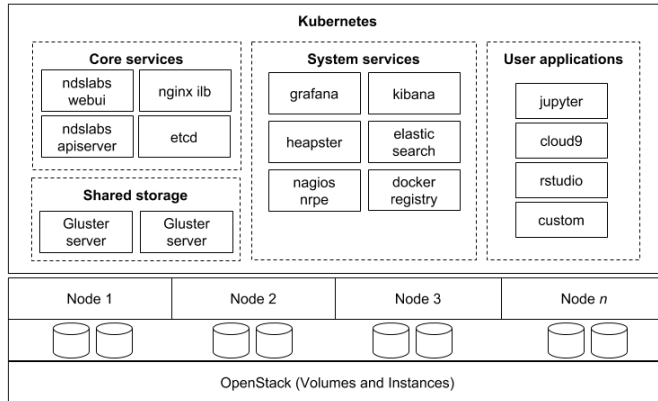

Figure 1: Labs Workbench systems architecture

The Labs Workbench system includes the following features:

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

The Labs Workbench user interface is presented in Figure 2. The platform focused primarily on interactive, web-based applications including complex applications with multiple dependencies. Support is also available for command-line applications via web-based IDEs or terminal emulators.

4. 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 [15], an innovator in this area, that uses the Girder data management platform[8] and custom Jupyter analysis environments to enable access to data for a number of fields including astronomy and cosmology. The yt Hub architecture is currently being adapted to support the Whole Tale project [7] and the Renaissance Simulations Laboratory (RSL), providing access to the Renaissance Simulations data[11]. Similarly, the SciServer Compute system provides access to astronomy, materials science, turbulence, and earth science datasets via domain-specific Rstudio, MATLAB and Jupyter containers [9].

While container-based analysis environments enable new forms of access to research data, they do not solve the discovery problem. The DataDNS initiative will leverage

---

[6]https://www.github.com/nds-org/ndslabs-specs
[7]https://www.gluster.org/
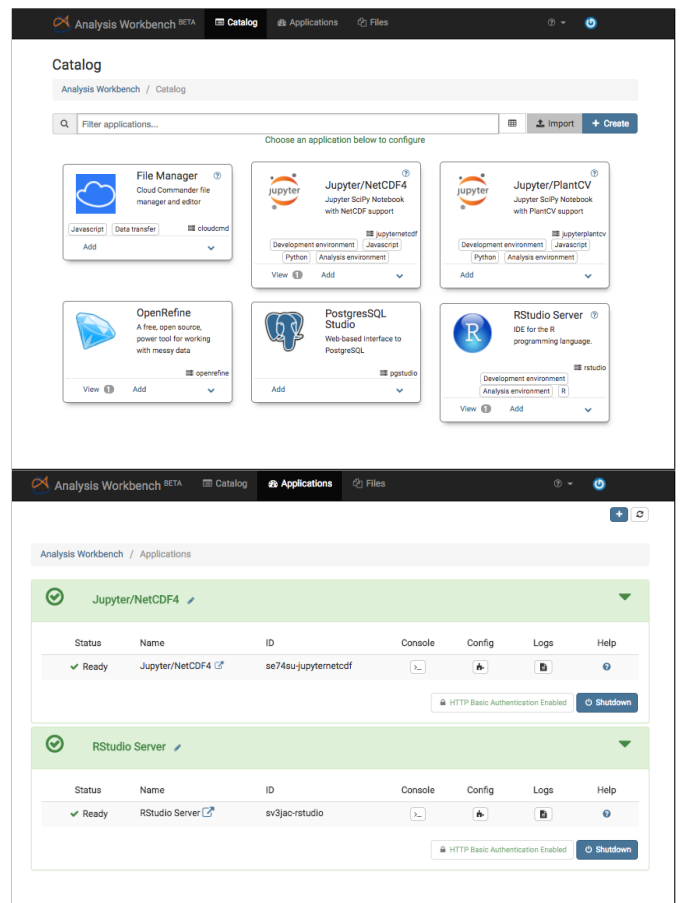[8]https://github.com/girder/girder

Figure 2: TERRA-REF Analysis Workbench catalog and application dashboard

these and other capabilities of research computing platforms to enable discovery, access, and in-place analysis of large-scale data. DataDNS is connects traditional research data publishing infrastructure to research computing infrastructure through an active registry and resolution service that tracks the locations and capabilities associated with remotely hosted datasets. Capabilities include file transfer or container-based, cloud-based or batch compute. Through DataDNS, researchers will not only be able to discover these datasets, but also access and launch remote analysis, when possible.

5. CONCLUSIONS AND NEXT STEPS

Traditional data publishing service, such as community and institutional repositories, are not equipped to handle publication of or access to very large research datasets. Container-based analysis environments, such as Labs Workbench, provide a scalable, resource constrained, and low-barrier approach to providing access to data hosted on research computing and HPC systems. These services open up new opportunities to support access to large-scale data. DataDNS will connect these and other compute-centric resources to traditional discovery methods, such as institutional and community repositories, to further enable re-use of remotely hosted data.

Development of the Labs Workbench platform and

DataDNS system are ongoing. Upcoming features include support for single sign-on, simplified installation processes, and support for commercial cloud services. We are also working on improved support for mounting external data into the system and exploring approaches to supporting job scheduling and execution.

6. ACKNOWLEDGEMENTS

7. REFERENCES

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

[2] U. K. Devisetty, K. Kennedy, P. Sarando, N. Merchant, and E. Lyons. Bringing your tools to cyverse discovery environment using docker. *F1000Research*, 5, 2016.

[3] DOE Advanced Research Project Agency - Energy. Transportation energy resources from renewable agriculture (TERRA). `https://arpa-e.energy.gov/?q=arpa-e-programs/terra`.

[4] 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)*, March 2015.

[5] 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. IEEE Press, 2014.

[6] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, et al. Jupyter notebooks – a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 2016.

[7] B. Ludaescher, K. Chard, N. Gaffney, M. B. Jones, J. Nabrzyski, V. Stodden, and M. Turk. Capturing the" whole tale" of computational research: Reproducibility in computing environments. *arXiv preprint arXiv:1610.09958*, 2016.

[8] MathWorks. Matlab. `https://www.mathworks.com/`.

[9] D. Medvedev, G. Lemson, and M. Rippin. Sciserver compute: Bringing analysis close to the data. In *Proceedings of the 28th International Conference on Scientific and Statistical Database Management*, SSDBM '16. ACM, 2016.

[10] H. Meng, R. Kommineni, Q. Pham, R. Gardner, T. Malik, and D. Thain. An invariant framework for conducting reproducible computational science. *Journal of Computational Science*, 9, 2015. Computational Science at the Gates of Nature.

[11] B. W. O'Shea, J. H. Wise, H. Xu, and M. L. Norman. Probing the ultraviolet luminosity function of the earliest galaxies with the renaissance simulations. *The Astrophysical Journal Letters*, 807(1), 2015.

[12] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., 2015.

[13] 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), Mar. 2007.

[14] C. Willis, D. LeBauer, M. Lambert, and M. Burnette. TERRA-REF Analysis Workbench: Container-based Environments for Low-Barrier Access to Research Data, May 2017.

[15] J. ZuHone and K. Kowalik. The galaxy cluster merger catalog: An online repository of mock observations from simulated galaxy cluster mergers. *arXiv preprint arXiv:1609.04121*, 2016.