

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

Craig Willis, Mike Lambert, Kenton McHenry, and Christine Kirkpatrick

School of Information Sciences
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
San Diego Supercomputer Center
{willis8, lambert8, mchenry}@illinois.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 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 datasets from traditional discovery models. New models of in-place publishing are needed to connect research computing and research data publishing infrastructure to support discovery and access.

Institutional and community repositories are beginning to support “remote” data publishing. The data is effectively published in-place with best-effort preservation and researchers provide 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 three 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. Projects provide access to large datasets through custom container-based analysis environments, such as Jupyter notebooks and Rstudio. In these systems, remote users are able to register for an account on a web-based system that allows them to launch specialized, resource-constrained analysis environments to explore data in-place. Examples include 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 leverage systems such as Labs Workbench to enable discovery, access, and in-place analysis of datasets where they reside.

This paper is organized as follows...

2. CONTAINER-BASED ANALYSIS ENVIRONMENTS

What are container-based analysis environments. Exam-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

PEARC '2017 New Orleans

Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$15.00.

ples of systems using container-based interfaces: SciServer, Cyverse, yt.hub, etc. Also include science cases Renaissance Simulations, DarkSky, TERRA-REF.

Should we say something about Amazon Public Datasets and similar trends? Describe AWS public datasets, what are available, how you get there. Also, describe the process for accessing. (Not containers, but...)

3. LABS WORKBENCH

What is Labs workbench and how does it solve these problems.

3.1 Use Case: TERRA-REF

Detailed description of TERRA-REF and it's use of Labs Workbench The ARPA-E TERRA program is focused on the development of cutting-edge techniques for the improvement of biofuel crops in part through the creation and publication of a large public reference dataset, called TERRA-REF, and associated computational pipeline. The TERRA-REF data-storage and computing system will provide researchers with access to 2PB of raw sensor and derived data hosted in the NCSA ROGER system and made available via Globus, Clowder, and the NDS Labs Workbench.

4. DATADNS

What is DataDNS and how does it solve these problems.

5. CONCLUSION

6. 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.