

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 poster 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

A few background points:

1. Sensor-based, research-computing and high-performance computing (HPC) systems produce massive amounts of data.
2. Traditional models of access and discovery, such as community and institutional repositories, are not equipped to handle these very large datasets.

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.

3. Increasing researchers are leaving their data on research computing infrastructure or turning to cloud-based services enable access and reuse.
4. Since research-computing and HPC centers cannot support long-term storage of very large datasets or guarantee access, as is often required by publishers, these datasets are further disconnected from traditional discovery models.
5. As computing center are turning to lower-cost cloud-based storage options, researchers are left with best-effort preservation solutions.
6. There are lots of examples of low-barrier access being provided via Docker containers (yt.hub, SciServer, Whole Tale, Galaxy Portal, TERRA-REF, etc).
7. Amazon public datasets

Central argument:

1. Labs Workbench is a platform that enables computing centers to provide low-barrier access to research data in-place (example case: TERRA-REF)

Access is restricted due to resources to transfer and compute on the data; the requirement to apply for access to the associated system; the absence of methods of discovery in general.

This paper is organized as follows...

2. CONTAINER-BASED ANALYSIS ENVIRONMENTS

What are container-based analysis environments. Examples 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 its 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.