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

Sensor-based, research- and high-performance computing (HPC) systems produce massive amounts of data. Cosmology, climate, and turbulence simulations can produce petabytes of data per run. Similarly, sensor-based systems can produce petabytes of data. These data products take much longer to analyze and retain their scientific value for years. Additionally, their richness often supports multiple lines of investigation.

Traditional models of

PEARC '2017 New Orleans

2. CONTAINER-BASED ANALYSIS ENVIRONMENTS

Examples of systems using container-based interfaces: SciServer, Cyverse, yt.hub, etc.

Examples include the Renaissance Simulations, DarkSky, TERRA-REF,

3. LABS WORKBENCH

4. CASE: TERRA-REF

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.

4.1 Workshops and training

5. RELATED WORK

6. LABS WORKBENCH

7. DATADNS

8. CONCLUSION

9. 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.