SeedMe2: Data sharing building blocks

Amit Chourasia*, David R. Nadeau, John Moreland, Dmitry Mishin and Michael L. Norman
San Diego Supercomputer Center, University of California, San Diego
* email: amit@sdsc.edu

Abstract: *The need for data sharing and rapid data access has become central with the rise of collaborative research in many disciplines. Several data sharing approaches have emerged for consumer use cases that primarily need an easy way to share files using web browsers. However, these approaches are not well-suited to the particular demands of large-scale data sharing for computational research. Whereas consumer approaches primarily support manual user interfaces to add and remove files, the huge number of files that can be generated during and after a large-scale computation job make manual data sharing interfaces impractical. Instead, these tasks require mechanisms that integrate into computation workflows to automatically post files during and after computation jobs. Furthermore, scientific data sharing requires additional metadata and descriptive information that characterizes shared data to record job and compute platform characteristics, input data, job parameters, job completion status, and other record-keeping required to document the trajectory of computational research. Without these features, consumer data sharing approaches are not well suited for computational science.*

In this work we describe SeedMe2 (Stream, Encode, Explore and Disseminate My Experiments) as a data sharing platform that caters to unique needs of computational scientists to support data sharing, context descriptions, discussion, light visualization of supported microformats and easy workflow integration.

1. Introduction

Collaborative research depends on data sharing and timely access to data. This is especially true in computational science research, where experiments are conducted on disparate compute resources from laptops to High Performance Computing (HPC) clusters. The ability to share relevant job data, parameters, and results quickly and easily is essential for efficient collaboration.

Presented at Gateways 2016, San Diego Supercomputer Center, La Jolla, CA, November 2-3, 2016.
https://gateways2016.figshare.com/

2. SeedMe2 platform

The SeedMe2 (Stream, Encode, Explore and Disseminate My Experiments) project is developing web-based building blocks and cyberinfrastructure to enable easy sharing and streaming of transient data and preliminary results from computing resources to a variety of platforms, from mobile devices to workstations, making it possible to quickly and conveniently view and assess results and provide an essential missing component in High Performance Computing (HPC) and cloud computing infrastructure. This work is an evolution of the SeedMe project [1, 2] that to provide modular and flexible data sharing building blocks to the computation community.

SeedMe2 has a modular design and will include several building blocks that may be included as desired. The following blocks are envisioned:

- Data sharing with access controls.
- Virtual file system: Provides three configurations to control permission at a) every folder b) top folder c) hierarchical.
- Data description and discussion.
- Federated authentication using CILogon [3].
- Search and index of supported data formats.
- Microformats and corresponding visualization: Supported microformats include CSV, table and graph in ASCII or JSON format.
- REST API [4].
- Clients in Python, JAVA and command line.

3. Architecture

SeedMe2 architecture (Fig. 1) is based on the Drupal8 [5] web site content management system, which requires a compatible webserver (typically Apache [6]) and a database (typically MySQL[7]). SeedMe2 building blocks are structured as modules that can be plugged into the Drupal8 and customized as desired.

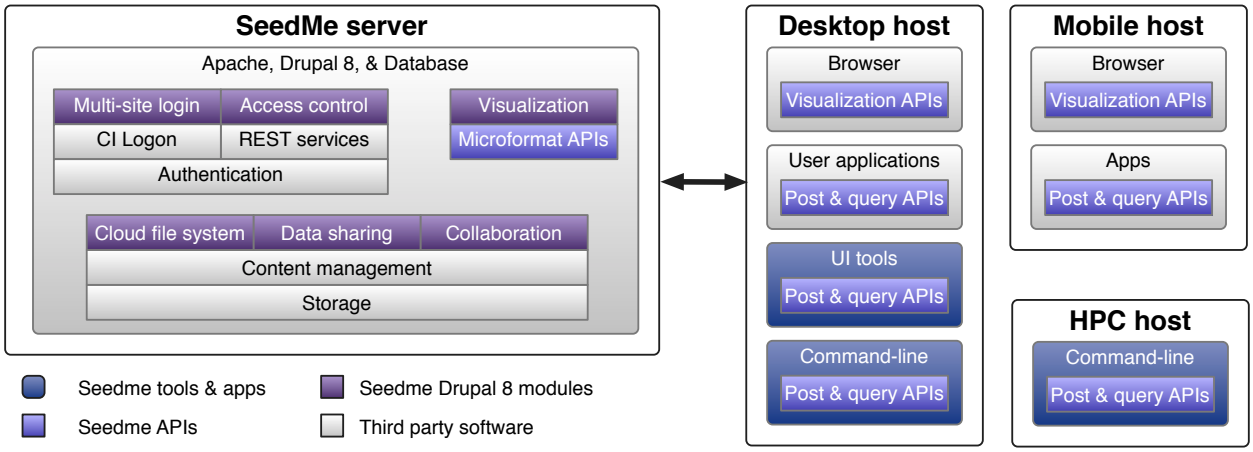

Fig. 1. SeedMe modules (purple) and APIs (blue), for interaction from mobile, desktop and HPC hosts.

4. Deployment

We anticipate three deployment scenarios for the SeedMe2 platform, including:

- Platform as-a-service: A central service is which is open to all researchers.
- Do it yourself (DIY) cloud: A project customizes and configures a branded instance.
- Provider cloud: Computing centers or other vendors run an instance for their users.

5. Applications

We envision a variety of applications that may use the SeedMe2 platform, including:

- Adhoc data sharing by computational users.
- Application integration: Third party applications add support to share data directly with central service.
- Gateway integration: Science gateway either uses the central service or their own instance to enable data sharing for their users.

6. Acknowledgments

This work is supported by the National Science Foundation under Grant No. ACI-1443083. "Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF."

7. References

[1] SeedMe. 2016. SeedMe (Stream Encode, Explore and Disseminate My Experiments) Retrieved Sep 2, 2016 from https://www.seedme.org

[2] Amit Chourasia, Mona Wong-Barnum, Dmitry Mishin, David R. Nadeau and Michael L. Norman. Presented at the XSEDE 2016 conference. Miami, FL Jul 17-21, 2016. DOI: http://dx.doi.org/10.1145/2949550.2949590G.

[3] Basney, J. Fleury, T. and Gaynor, J. 2014 "CILogon: A Federated X.509 Certification Authority for CyberInfrastructure Logon," Concurrency and Computation: Practice and Experience, Volume 26, Issue 13, pages 2225- 2239, September 2014. http://dx.doi.org/10.1002/cpe.3265

[4] Fielding, R. T. and Taylor, R. N. 2002. "Principled Design of the Modern Web Architecture", *ACM Transactions on Internet Technology (TOIT)* (New York: Association for Computing Machinery) **2** (2): 115–150, May 2002

[5] Drupal. 2016. *Drupal – Open Source CMS*. Retrieved Sep 2, 2016 from http://drupal.org/

[6] Apache. 2016. *The Apache HTTP Server Project*. Retrieved Sep 2, 2016 from http://httpd.apache.org/

[7] MySQL. 2016. *MySQL*. Retrieved Sep 2, 2016 from http://www.mysql.com/