Plan Overview

A Data Management Plan created using DMPTool

Title: CSSI Elements: Exosphere, A Researcher-Friendly Interface to the Cloud

Creator: David LeBauer

Affiliation: University of Arizona

Funder: National Science Foundation (NSF)

Funding opportunity number: 56703

Template: NSF-CISE: Computer and Information Science and Engineering

Project abstract:
Cloud computing supports fundamentally new advances in science and engineering by providing new methods of storing, analysing, visualizing, and sharing data. Research clouds deliver on-demand and persistent resources with elastic scaling, fully-configurable software environments, and interactivity beyond what HPC systems can provide. This enables rapid development of analytical models and long-lived applications (Science Gateways) to facilitate data sharing and reuse. NSF has invested in research cloud systems, and these systems continue to add new capabilities. For example, Jetstream 2 delivers persistent access to GPUs for massively parallel computation and accelerated visualization. Modern computational techniques are unevenly adopted across science domains. For research clouds to achieve the greatest impact, they must be easy for researchers to use and easy for cloud operators to deploy. Previous efforts to build such interfaces have not met this challenge: the default user interface for OpenStack research clouds, named Horizon, was designed for use by Information Technology professionals. CyVerse Atmosphere, another research cloud interface, is user-friendly but does not support persistent workloads. Atmosphere is also difficult to deploy and costly to support. This leaves an accessibility gap between research and the power of research clouds. We close this accessibility gap with Exosphere: a new client interface which brings advanced capabilities of research clouds within reach of non-advanced users. Exosphere is an easy-to-use web-based application that is also easy to deploy. It delivers both command-line and graphical desktop access to cloud resources with one click, and displays workload resource usage with real-time plots. NSF-funded projects already use Exosphere and
feedback has been positive. A three-year CSSI award will realize
Exosphere's potential for transformational science applications on
research clouds. In the first project year we will improve the maturity
and user experience of the Exosphere graphical interface, and expose
more advanced capabilities of research clouds. These capabilities
include elastic scaling, iterative exploration environments, secure data
science workbench sharing, and displaying vital signs of cloud
resources in a dashboard. Exosphere will deliver GPU-accelerated
streaming desktop environments in the researcher's web browser to
enable high-performance rendering and computation in scientific
desktop software, a fundamentally new capability in the research cloud
ecosystem. In the second year, we will support one-click launch of data
science workbenches, e.g. JupyterLab and RStudio Server, and
support the Scientific Filesystem to encourage reproducibility. We will
also transform distributed science education workshops with powerful
features for facilitators. In the third year, we will build support for at
least one commercial cloud platform, Amazon Web Services, Google
Cloud Platform, or Microsoft Azure, to support efficient use of these
expensive resources, and migration of workloads between research
and commercial clouds. Intellectual Merit: Exosphere multiplies the
force of research clouds by removing barriers to adoption and enabling
transformational science applications. Modern computing methods and
data science techniques enable fundamentally new and more powerful
approaches. Exosphere will empower more scientists to leverage these
approaches with large datasets and complex analyses. This will enable
fundamentally new discoveries, creating sustained and sustainable
impacts. Broader Impacts: Exosphere's focus on usability and access
to research clouds, including those that NSF provides, will benefit
under-served groups, including researchers at minority-serving and
primarily-undergraduate institutions who otherwise lack training to use
modern, cloud-enabled research techniques. Easier access to shared
cloud infrastructure will accelerate scientific discovery by reducing
bottlenecks of personal computers, and reducing reliance on expensive
workstations. Thus, Exosphere will help broaden participation in
research and STEM fields. Further, Exosphere increases the value of
existing research clouds in the national CI ecosystem by lowering
barriers to adoption and adding new capabilities. Finally, Exosphere
has applications outside of research computing. Many industrial, not-
for-profit, and educational groups operate OpenStack cloud
infrastructure. Exosphere could bring the first user-friendly, easy-to-
deploy cloud computing interface to these communities.

Last modified: 02-02-2021

Copyright information:
The above plan creator(s) have agreed that others may use as much of
the text of this plan as they would like in their own plans, and customize
it as necessary. You do not need to credit the creator(s) as the source
of the language used, but using any of the plan’s text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal.
CSSI Elements: Exosphere, A Researcher-Friendly Interface to the Cloud

Types of data

The Data Management Plan should describe the types of data, metadata, scripts used to generate the data or metadata, experimental results, samples, physical collections, software, curriculum materials, or other materials to be produced in the course of the project.

This project will neither collect samples nor generate scientific data. Data produced by the project will be related to the development, use, and distribution of software. The project will generate source code (text files), compiled software (binary files), and documentation (structured text markup files and compiled HTML). These may be accompanied by media files (images, diagrams, and videos). Anonymous user interviews and surveys will also be conducted, and this will generate response data.

Data and metadata standards

The Data Management Plan should address the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies). It should also cover any other types of information that would be maintained and shared regarding data, e.g. the means by which it was generated, detailed analytical and procedural information required to reproduce experimental results, and other metadata.

Documentation: user-facing and internal documentation will be written in a structured markup format such as Markdown.

File formats: all text files, using git for version control; GitLab for code hosting and continuous integration; Zenodo for archiving of two or more versioned releases per year that will including the addition of features proposed in this study.

Source code is written in a mix of programming languages, primarily Elm and JavaScript. Some code that is deployed to cloud instances will use Bash, Python, or Ansible (YAML). In any case, the project will use industry-standard formatting tools (e.g. elm-format, ESLint, PyLint) to ensure that source code complies with language-specific conventions.

We will provide metadata related to software and its use in standard formats designed for both humans and machines. Software metadata will be stored in the root directory of the repository - in JSON files following the Codemeta and Zenodo schemas to facilitate discovery and reuse. Furthermore, we will make relevant information available to contributors by maintaining a README.md, LICENSE, CONTRIBUTING.md, and related files in the root directory of the repository. A Dockerfile will describe dependencies and enable easy deployment. Continuous integration metrics will inform users of testing coverage and failures.

User survey response data will be stored in a standard format; this project's user survey may be integrated with the Jetstream Cloud user survey, in which case the data format will be determined by Jetstream personnel.
Policies for access, sharing, and privacy

The Data Management Plan should address the policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements. It should cover any factors that limit the ability to manage and share data, e.g. legal and ethical restrictions on access to human subject data.

All products of this work, and as much design and discussion activity as possible will be conducted in public forums including GitLab. All data and code generated within the scope of this project will be available in public repositories and databases. We will develop, test, and release code in GitLab. Gitlab repository contents and binaries will be automatically archived and assigned a DOI by Zenodo with each tagged release. Publications are budgeted to be open-access, and we will deposit publications and slides on pre-print servers and other public venues. Anonymous user survey data will be published or shared only in aggregate form, and only after approval from the applicable institutional review boards (IRBs) to ensure the privacy of participants.

Policies for re-use, re-distribution, derivatives

The Data Management Plan should address the policies and provision for re-use, re-distribution, and the production of derivatives.

We are committed to enhancing the value of research and furthering the advancement of public knowledge. The Exosphere software products are released under the permissive BSD 3-Clause License. All other resources developed during the course of the project will be made available to the scientific community with licenses that support further reuse, modification, and redistribution such as the BSD-3-Clause, CC-By, CC-0 Public Domain, or similar.

Plans for archiving and preservation

The Data Management Plan should address the plans for archiving data, samples, and other research products, and for the preservation of access to them. It should cover the period of time the data will be retained and shared; how data are to be managed, maintained, and disseminated; and mechanisms and formats for storing data and making them accessible to others, which may include third party facilities and repositories.

Code will be maintained on GitLab, with a mirror on GitHub, archives on Zenodo, and copies on no fewer than four personal computers. All code changes will be documented and stored using Git version control. Each release will be tagged in the git repository and archived on Zenodo.
Additional Guidance on Selecting or Evaluating a Repository:
The following questions are intended to assist PIs and panel members to prepare
Data Management Plans and to evaluate them during merit review, respectively.
The questions are sequential, that is, if (1) applies, then the remaining questions
are irrelevant unless (2) also applies or the PI chooses to deposit the data or
software in multiple repositories. The more detailed questions, (4)-(6), apply if (1)
and (2) do not.

1. Does the solicitation specify a repository for the data or software?
2. Does the PI's home institution have an institutional repository that
   mandates local deposit of the data/software?
3. Is there a discipline-relevant repository used by the research community
   either as the expected repository for data/software or as the expected
   repository for discovering and reusing data/software?
4. Is the repository sustainable? And if not, are there contingency plans?
5. Does the repository require at least minimal identification and description of
   the data product sufficient to enable discovery, access, and retrieval? For
   purposes of data citation, NSF requires a persistent identifier and some level
   of metadata including acknowledgment of the creator/author and federal
   support.
6. Has the PI made any contingency plans in the event a designated repository
   becomes unavailable?

Question not answered.

Roles and responsibilities

The Data Management Plan should clearly articulate how the PI and co-PIs plan to
manage and disseminate data generated by the project. The plan should outline
the rights and obligations of all parties as to their roles and responsibilities in the
management and retention of research data, and consider changes that would
occur should a PI or co-PI leave the institution or project. It should describe how
the research team plans to deposit data into any relevant and appropriate
disciplinary repositories that are appropriately managed and that are likely to
maintain the metadata necessary for future use and discovery. Any costs
associated with implementing the DMP should be explained in the Budget
Justification.

PI LeBauer will be responsible for ensuring adherence to this plan. LeBauer and
Exosphere developers Julian Pistorius and Chris Martin have extensive experience with
source code maintenance and software distribution and have been using the approach
described in this plan for many years, including with Exosphere. If the PI leaves the
project or institution, responsibilities will be transferred to collaborator Blake Joyce, PhD,
Assistant Director of Research Computing at UA HPC center. All contributions to
publications, software code, design, or feedback will be recognized appropriately
through co-authorship and acknowledgements.