XSEDE Capability Delivery Plan
DA-2 Data Preparation
Last revised 2016-10-05

Background

Use cases describe community needs, requirements, and recommendations for improvements to cyberinfrastructure “CI” resources and services. A Capability Delivery Plan “CDP” is an executive summary of use case support gaps, of plans to fill those gaps with new or enhanced capabilities, and of existing operational components that already support aspects of a use case.

Use Case Summary

Use case DA-2 describes how users engage in data preparation, which may include selecting, cleaning, supplementing, integrating, formatting, and modeling data. More specifically: users should be able to collect data from local and remote resources; stage data from archival resources to compute resources; and execute a broad array of data preparation processes. XSEDE needs to 1) provide tools for creating, sharing, and discovering metadata, 2) libraries for reading standard file formats such as NetCDF and HDF5, 3) the ability to access remote data over the internet in unconstrained ways.

Use case document(s): http://hdl.handle.net/2142/45702

CDP Summary

The functionality described in this use case is partially supported by the operational components listed below.

Gap(s) that we currently plan to address:

- Metadata management tools
- Consistent data library and tools information
- Information about remote data access from SP resources
- Review/enhance XSEDE portal Data Analytics page

Gap(s) that will not be addressed at this time:

- None

Time and effort summary:

- Provide metadata tools: 12 FTE-weeks
- Consistent data library and tools information: 2 FTE-weeks (affects all SPs)
- Information about remote data access from SP resources: 0.5 FTE-weeks

- Review/enhance XSEDE portal Data Analytics page: 0.5 FTE-weeks

Functionality Gaps

1. Metadata management tools (suggested priority: high)

XSEDE doens't offer metadata creation, sharing, and discovery tools.

Plans: Identity and integrate tools for creating and sharing metadata to enable data discovering. (xci-?)

2. Consistent data library and tools information (suggested priority: medium)

XSEDE information about the data libraries and tools available on various SP resources may be incomplete or inconsistent.

Plans: Explore what information is currently available about data libraries and tools on SP resources, and standardize the information, including module names, to make it easier for users to discover and access those data tools. (xci-?)

3. Information about remote data access from SP resources (suggested priority: high)

XSEDE doesn't offer SP HPC/HTC/Storage/Visualization resource specific information on remote data access over the internet .

Plans: Provide users with HPC/HTC/Storage/Visualization resource specific information on remote data access over the internet, including connectivity constraints, bandwidth constraints, and other useful information (https://software.xsede.org/view/xci-27)

4. Review/enhance XSEDE portal Data Analytics page (suggested priority: high)

Enhance the XSEDE portal Data Analytics page with information about the tools described in this use case.

Plans: Prepare/enhance a Data Analytics XSEDE portal home page that includes information about metadata tools, data libraries, data manipulation tools, and resource specific remote data access constraints (https://software.xsede.org/view/xci-1)

2. Quality attributes

Verifying quality attributes requires significant one-time and ongoing testing. XSEDE has decided that the costs of this testing would not bring sufficient benefit. Instead XSEDE will monitor user satisfaction, usage, and available performance metrics and address quality issues when raised by users. *There are no plans to address this verification gap.*

System Components That Support This Use Case

The following XSEDE operational components currently support this use case:

(Hyperlink the component <Name> to the XCSR Component Description Repository)

Component	Supported Functionality
XSEDE User Portal (XUP)	The front-end user interface to the XSEDE system where end users register with XSEDE, manage their user profile information, request allocations to use XSEDE SP resources, and access documentation about XSEDE resources, software, services, and more.
RDR	Manages compute, storage, and visualization resource descriptions
Software/Service Descriptions and Availability	Manages software and service descriptions and availability information
XSEDE Pub/Sub	Enables SPs to publish software and service availability information