Transforming the Way Scientists Share Data

Christine Kirkpatrick

June 2019
Overview

Components of Research Cyberinfrastructure

• Infrastructure: Open Storage Network
• Applications/Glue
  • National Data Service
  • Working examples
• Data: (GO) FAIR
• People networks: Big Data Hubs
Open Storage Network
US Research Cyber-Infrastructure Today

**Computation**
- Shared Resource (XSEDE, PRAC)
- Standardized
- NSF-Funded

**Networking**
- Over 200 universities with 40/100Gb Connectivity
- Standardized
- NSF-Funded

**Storage**
- Largely Balkanized
- No Standards Requirement
- No CI Funding
Six Prototype Deployment Sites

- Johns Hopkins University
- Massachusetts Green HPC Center
- Northwestern University (Starlight)
- University of CA San Diego (SDSC)
- University of Illinois (NCSA)
- University NC Chapel Hill (RENCI)

Supported by the Schmidt Foundation and NSF grants #1747552, 1747493, 1747507, 1747490, 1747483
# Example Use Cases

<table>
<thead>
<tr>
<th>Project</th>
<th>Average size of data entities</th>
<th>Total data volume</th>
<th>Storage problem being solved</th>
<th>Use case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Zone Observatories</td>
<td>10 MB</td>
<td>50 TB</td>
<td>Provide storage space and access to CZO datasets and community-generated data</td>
<td>Community long-tail data</td>
</tr>
<tr>
<td>TerraFusion</td>
<td>10 GB</td>
<td>1 PB</td>
<td>Transport datasets across the US at high speed, obtain data slices with high probability of reutilization</td>
<td>Experiment-to-site, Slice-and-compute</td>
</tr>
<tr>
<td>HathiTrust Research Center collection</td>
<td>200 MB</td>
<td>500 TB</td>
<td>Provide storage space and access to the HTRC dataset and further community-generated derivatives</td>
<td>Common resource access</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>10 GB</td>
<td>1 PB</td>
<td>Make available a well-curated dataset for testing machine learning algorithms</td>
<td>Dataset-as-benchmark</td>
</tr>
<tr>
<td>Large Synoptic Survey Telescope</td>
<td>2 TB</td>
<td>100 PB</td>
<td>Transport datasets across the US at high speed, obtain data slices with high probability of reutilization</td>
<td>Experiment-to-site, Slice-and-compute, Workflow staging space</td>
</tr>
<tr>
<td>Combined Array for Research in Millimeter Astronomy</td>
<td>50 MB</td>
<td>50 TB</td>
<td>Transport datasets across the US at high speed, obtain data slices with high probability of reutilization</td>
<td>Experiment-to-site, Slice-and-compute</td>
</tr>
</tbody>
</table>
Experiment-to-site

Data origin

A subset of curated instrument data, or derived datasets from it is setup for fan-out

OSN

The data is copied to OSN until delivered to the sites with a TTL

OSN determines best replication and transport based on geographical proximity and infrastructure

Data is delivered to every site

All sites signal OSN completion of transfers

OSN eliminates all replicas after TTL expires
A parameterized data query is created and sent to the data origin outside of OSN.

The query is parsed, and the subset is generated outside of OSN.

OSN stores the dataset with a persistence indicator and a TTL, returns an object identifier.

The object identifier is recorded and sent to the requesting site using non-OSN means.

The site receives the object identifier and downloads the subset from OSN.

OSN eliminates all replicas after TTL expires.
# OSN / S3 Cache

<table>
<thead>
<tr>
<th>Open Storage Network</th>
<th>User's Amazon Web Services Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSN Pod</td>
<td>AWS Cloud</td>
</tr>
<tr>
<td></td>
<td>Amazon Simple Storage Service (S3)</td>
</tr>
<tr>
<td></td>
<td>Amazon EC2</td>
</tr>
<tr>
<td></td>
<td>Amazon EMR</td>
</tr>
<tr>
<td></td>
<td>TensorFlow on AWS</td>
</tr>
</tbody>
</table>
AWS Snowball

Planning to also evaluate AWS Outposts
What’s Next

• Refine governance and shared administration
• Progress on policies, procedures around data lifecycle
• Partnership with OSG, NRP/PRP
• Outreach and user-focused workshop, October 2019 at TACC (Austin)
• See us at:
  • PEARC 19 BOF
  • SC 19 BOF, exhibit booth
US National Data Service
### Where We Are

<table>
<thead>
<tr>
<th>Data</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Indexing</td>
<td>Curating</td>
<td>Publishing</td>
<td>Software</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Datanet</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Archives</td>
<td>Services</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hardware</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hosts</td>
<td>Routing</td>
<td>Computation</td>
<td>Storage</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Applications</td>
<td>Workflows</td>
<td>Analysis / Visualization</td>
<td>Browsers</td>
<td></td>
</tr>
</tbody>
</table>
U.S. National Data Service

National effort to bring together infrastructure supporting the publication, discovery, and reuse of data

→ From the Internet to the “Datanet”

1. Large-scale Data Service Interoperability
   • Distributed cloud and compute
   • Innovation in the gaps: services, software, integration

2. Incubator of Data Projects & Pilots
   • Quick start sandbox
   • Choose services based on features (not time to install)

3. Training Platform
NDS Labs Workbench: Tools-centric

https://www.workbench.nationaldataservice.org
Use case: TERRA-REF

David LeBauer (University of Arizona)

- High-throughput indoor and outdoor sensor platforms, UAV and field data, large-scale genome sequencing
- Petabyte scale data storage and computing pipeline
  - Data processing workflows
  - Raw and derived data
  - Data sharing and re-use
- Nationwide, multi-institution collaboration
  - Researchers, data scientists, and software developers
- Labs Workbench for remote, interactive access to data
CHEESE Technologies/Architecture

Justin Yang & Rajesh Kalyanam (Purdue), Craig Willis (NCSA)

http://www.cheesehub.org

UI
API
crd

Jupyter
Postgres

SQL Injection

Hacker
(Jupyter)

Client
(VNC)

Server
(Apache)

Arpspoof/SSL Strip

NGINX ILB

Monitoring
Logging

User namespace

System namespace

Kubernetes 1.1x (RBAC, Weave overlay network)

Ubuntu 18.04 VMs

OpenStack: Elastic Compute and Storage

Workbench

Work supported by the National Science Foundation under NSF grant # 1820573.
Metabolomics Workbench Cloud on AWS

PI: Shankar Subramaniam, UC San Diego

AWS Cloud

Region - us-west-2 (Oregon)

Elastic Load Balancing (ELB)

Amazon Elastic Container Service for Kubernetes

Availability Zone - us-west-2a

Auto Scaling

Spot Instance

M5 Instance

Availability Zone - us-west-2b

Auto Scaling

Spot Instance

M5 Instance

Availability Zone - us-west-2c

Auto Scaling

Spot Instance

M5 Instance

AWS Lambda

Amazon Elastic File System

https://www.metabolomicsworkbench.org/

Thank you, NIH STRIDES!

https://datascience.nih.gov/strides

Funded by NIH #1U2CDK119886-01
SDSC’s Phylogenetics Science Gateway Awarded NSF/Internet2 Grant

Popular CIPRES resource to give researchers AWS Cloud access

The widely used CyberInfrastructure for Phylogenetic REsearch (CIPRES) science gateway, based at the San Diego Supercomputer Center (SDSC) at UC San Diego, has been awarded a one-year Internet2 grant funded by the National Science Foundation (NSF). The grant will let gateway users take advantage of more powerful compute processors available from a commercial cloud provider to accelerate their scientific discoveries.
## Registered Data Sets

### Renaissance Simulations

**C'OShea, Brian (cshea@msu.edu); Wise, John; Xu, Hao; Norman, Michael**

*More Information*


### Dark Sky Simulations

**Warren, Michael; Friedland, Alexander; Holz, Daniel; Skillman, Samuel; Sutter, Paul; Turk, Matthew (mjturk@illinois.edu); Wechsler, Risa**

*More Information*


### Magnetohydrodynamic Turbulence Simulations

**Mösta, Philipp (pmoesta@berkeley.edu)**

*More Information*

(Research) Data Process Methodologies / Data Lifecycles
FAIR

Findable
Accessible
Interoperable
Reproducible

Key efforts:

• **FAIR data in Earth, Space, & Enviro Science:** [www.copdess.org/enabling-fair-data-project/](http://www.copdess.org/enabling-fair-data-project/)
• **GO FAIR:** [go-fair.org/](http://go-fair.org/)
• **FAIRsFAIR:** [www.fairsfair.eu](http://www.fairsfair.eu)
• **FAIRplus:** [fairplus-project.eu](http://fairplus-project.eu)
It Takes a Village to be FAIR

Faulty Assumptions

• Curationists will make data FAIR (on their own).
• Security people do all the security work.
• Webmasters make all material accessible.

Partners in FAIR data stewardship:

• Research computing
• Libraries
• Research labs (researchers, postdocs)
• Administrators
GO FAIR US: Pillars, Implementation Networks

GO CHANGE
Culture

GO TRAIN
Training

GO BUILD
Technology

GO FAIR International Support and Coordination Office

Sea Data Cloud
(GAIA) System Terre
Training Curriculum
Training Frameworks
Vaccine IS

Annotation
ASTRON
BiodiFAIRse
Biodiversities
C2CAMP
CBS (Economics)
IN Matrix: Converging on metadata/data formats, terminologies
GO FAIR US Office

- Train FAIR Data Stewards
  - Train the trainers
- Partnership with Phortos Consultants
  - Training and consulting for local industry
  - Assist with FAIR Data Stewardship Plans
  - Assist organizations/companies to GO FAIR
- Create and harden FAIR tooling
- Extend Implementation Networks (IN) into US
PresQT Data and Software Preservation Quality Tool Project

Contributors: John Wang, Sandra Gesing, Rick Johnson, Natalie Meyers, Jeffrey R. Spies, David Minor, Markus Krusche

Affiliated institutions: Center For Open Science, University of Notre Dame

Date created: 2016-05-30 05:09 PM | Last Updated: 2018-12-20 07:49 AM

Identifiers: DOI 10.17605/OSF.IO/D3JX7 | ARK c7605/osf.io/d3jx7

Category: Project

Description: The goal is to collaboratively design interoperable and repository agnostic data and software preservation quality tools.

License: CC-By Attribution 4.0 International 🌐

Start managing your projects on the OSF today.
NSF Big Data Hubs
Four hubs, one mission

(Just renewed!)
Big Data Hubs: What We Do

Build and strengthen partnerships across industry, academia, nonprofits, government

• to address societal + scientific challenges,
• spur economic development,
• and accelerate innovation in the big data ecosystem
Stay Connected: All Hub Infrastructure Working Group

- Virtual meetings every first Friday at 3 PM ET
  - Anyone can attend, no registration or affiliation needed (Karl Gustafson kgustafs@renci.org)
  - Come listen in on what everyone is doing
  - Bring a problem, or an idea
  - Or give a lighting talk in the future

  *Come for the Lightning Talks, leave with some CI*

- PEARC 2019 - July 29 CI WG Affinity Group reception
- YouTube playlist – ~40 presentations
  [http://tinyurl.com/bdhub-videos](http://tinyurl.com/bdhub-videos)
“...to make the fruits of research and scholarship better and available to all who need or want them.” Berman et al.

christine@sdsc.edu
@SuperChristineK

Christine Kirkpatrick
Division Director, Research Data Services, SDSC
Executive Director, National Data Service

Some content courtesy of Kevin Coakley, Niall Gaffney, Meredith Lee, Kenton McHenry, Alainna White, Craig Willis