Research Data Challenges

- Storage
- Everything else!!
  - The bytes are not enough on their own
    \[00110100 \ 00110010\]
  - Metadata, curation tools, indexes, storage abstraction, replication, data transfer, authentication, access control, transformation, analysis tools, ...

\[\rightarrow \text{Software}\]
Architectural Vision for Research Cyberinfrastructure

https://dibbs17.org
Tools – Convertors & Extractors

PEcAn#ED_convert.R

```r
#!/usr/bin/Rscript
#PEcAn
data
#pecan.zip
#ed.zip
.libPaths("/home/polyglot/R/library")
sink(stdout(), type="message")

# global variables
overwrite <- TRUE
verbose <- TRUE

# get command line arguments
args <- commandArgs(trailingOnly = TRUE)
usage <- function(msg) {
  print(msg)
  print(paste0("Usage: ", args[0], ""))
  print(paste0("Example: ", args[0], "
...
```

wordcount.py

```python
import subprocess

def wordcount(input_file):
    result = subprocess.check_output(['wc', input_file], stderr=subprocess.STDOUT)
    (lines, words, characters, _) = result.split()
    metadata = {
        'lines': lines,
        'words': words,
        'characters': characters
    }

    result = {
        'metadata': metadata
    }

    return result
```
Welcome to the Brown Dog Tools Catalog, a community resource for publishing data transformation tools and making them available for use by the scientific community. To add a new tool click "Contribute".

Tools:

Filter by tool types: All Types  Filter by tool levels: All Levels

Displaying 1 to 24 of totally 42 available:

**The Predictive Ecosystem Analyzer**
Convert data sources to formats required by community built ecological models
Tool type: Converter
Tool level: Level 4
Deployments: bd-api, bde-api-dev
Submitter: mchenry
Last modified: Nov 23, 2016

**GI Detector**
Identify GI types in satellite images
Tool type: Extractor
Tool level: Level 4
Deployments: bde-api-dev
Submitter: mchenry
Last modified: Nov 22, 2016

**FFmpeg**
Convert between various video formats
Tool type: Converter
Tool level: Level 4
Submitter: mchenry
Last modified: Nov 20, 2016

**OpenJPEG**
Convert JPEG2000 images
Tool type: Converter
Tool level: Level 4
Submitter: mchenry
Last modified: Nov 20, 2016

**Daffodil**
Convert the contents of arbitrary files into an XML representation
Tool type: Converter
Tool level: Level 4
Deployments: bd-api, bde-api-dev
Submitter: mchenry
Last modified: Nov 20, 2016

**Ghostscript**
Convert between various document formats
Tool type: Converter
Tool level: Level 4
Deployments: bd-api, bde-api-dev
Submitter: mchenry
Last modified: Nov 20, 2016

**Kebecis**
Brown Dog API
Documentation for the Brown Dog API

Created by Brown Dog Support
See more at https://opensource.ncsa.illinois.edu/confluence/questions/topics/41156618/brown-dog
Contact the developer
University of Illinois/NCSA Open Source License

<table>
<thead>
<tr>
<th>Authorization</th>
<th>Operation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>POST /keys</td>
<td></td>
<td>Create a new api key</td>
</tr>
<tr>
<td>POST /keys/{key}/tokens</td>
<td></td>
<td>Create a new access token from api key</td>
</tr>
<tr>
<td>DELETE /keys/{key}</td>
<td></td>
<td>Delete api key and all related access tokens</td>
</tr>
<tr>
<td>DELETE /tokens/{token}</td>
<td></td>
<td>Delete token</td>
</tr>
<tr>
<td>GET /tokens/{token}</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conversions

GET /dap/outputs
List all output formats that can be reached.
<table>
<thead>
<tr>
<th>Method</th>
<th>URL</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>POST</td>
<td>/keys/{key}/tokens</td>
<td>Create a new access token from api key</td>
</tr>
<tr>
<td>DELETE</td>
<td>/keys/{key}</td>
<td>Delete api key and all related access tokens</td>
</tr>
<tr>
<td>GET</td>
<td>/tokens/{token}</td>
<td>Get info about token</td>
</tr>
</tbody>
</table>

**Conversions**

<table>
<thead>
<tr>
<th>Method</th>
<th>URL</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GET</td>
<td>/dap/outputs</td>
<td>List all output formats that can be reached</td>
</tr>
<tr>
<td>GET</td>
<td>/dap/inputs</td>
<td>List all input formats that can be accepted. Limit the number of hops allowed by setting chain parameter, e.g. ?chain=2</td>
</tr>
<tr>
<td>GET</td>
<td>/dap/inputs/{input_format}</td>
<td>List all output formats that can reach the specified input format</td>
</tr>
<tr>
<td>GET</td>
<td>/dap/convert</td>
<td>List all output formats that can be reached</td>
</tr>
<tr>
<td>GET</td>
<td>/dap/convert/{output_format}</td>
<td>List all output formats that can reach the specified output format</td>
</tr>
<tr>
<td>POST</td>
<td>/dap/convert/{output_format}</td>
<td>Convert the uploaded file to the requested output format</td>
</tr>
<tr>
<td>GET</td>
<td>/dap/convert/{output_format}/{file_url}</td>
<td>Convert the specified file to the requested output format</td>
</tr>
<tr>
<td>GET</td>
<td>/dap/path/{output_format}/{input_format}</td>
<td>Return the conversion path that would be used for this conversion</td>
</tr>
<tr>
<td>GET</td>
<td>/dap/software</td>
<td>List all available conversion software</td>
</tr>
<tr>
<td>GET</td>
<td>/dap/software/{software}</td>
<td>List all output formats available for this software</td>
</tr>
<tr>
<td>GET</td>
<td>/dap/software/{software}/{output_format}</td>
<td>List all input formats available for this software</td>
</tr>
<tr>
<td>POST</td>
<td>/dap/software/{software}/{output_format}</td>
<td>Convert the uploaded file to the requested output format using the specified software</td>
</tr>
<tr>
<td>GET</td>
<td>/dap/software/{software}/{output_format}/{file_url}</td>
<td>Convert the specified file to the requested output format using the specified software</td>
</tr>
</tbody>
</table>
A Science Driven Data Transformation Service

BD Fiddle
BD Catalog
Clients

Sign Up!

63,029,921 files/datasets transformed
# Make sure that that path to the input file is correct.

```r
library(BrownDog)

output_file <- convert_file("https://bd-api.ncsa.illinois.edu", "IMG0008.PCD", "png", ".", "e941ba85-8020-4115-9512-09ebb7ab156")
```

# Make sure that that path to the input file is correct.

```matlab
bd_funcs = bd_client;

output_file = bd_funcs.convert('https://bd-api.ncsa.illinois.edu', 'IMG0008.PCD', 'png', '.', '/converted_file.png', 'e941ba85-8020-4115-9512-09ebb7ab156');
```

The `bd.js` library requires URLs.
2) Select Transformation:

Automatic

Manual

Convert:
- Clowder
- GDAL
- GeoExpress
- ImageMagick
- OpenEXR

Extract:
- gi_detector
- ncsa.cv.caitech101
- ncsa.cv.closeups
- ncsa.cv.eyes
- ncsa.cv.faces
- ncsa.cv.meangrey
- ncsa.cv.profiles
- ncsa.cv.image.metadata
- ncsa.cv.image.ocr
- ncsa.cv.image.preview

Output:

Email Output
Generate workflow
Big Data
Submit
2) Select Transformation:

**Automatic**
- Convert:
  - ascii
  - azw3
  - bl.zip
  - bmp
  - clim
  - csv
  - dalec
  - dib
  - doc
  - docx

**Manual**
- Extract:
  - image/jpeg

3) Code Snippets:

```json
"Human Preference": {
  "Human Preference": 4,
  "Human Preference Definition": "An index [1,5] from a GI model that rates human preference for Green Infrastructure",
},
"Green Index": {
  "Green Index Definition": "The green index is the estimated percentage of green pixels within an image",
  "Green Index": "53.8"
}

@context: {
  "https://clowder.ncsa.illinois.edu/contexts/metadata.jsonld",
  "eyes": "http://clowder.ncsa.illinois.edu/ncsa.csv.eyes@eyes"
}
```
# Make sure that the path to the input file is correct.
library(BrownDog)

metadata <- extract_file("https://bd-api-dev.ncsa.illinois.edu", "IMG_0997.jpg", "e6dab924-04c8-45c0-94aa-f0608c3c1a45")

% Make sure that the path to the input file is correct.
bd_funcs = bd_client;
metadata = bd_funcs.extract(vals[1], 'IMG_0997.jpg', 'e6dab924-04c8-45c0-94aa-f0608c3c1a45');

The bd.js library requires URLs.
2) Select Transformation:

Automatic

- Manual

Convert:
- Clowder
- GDAL
- GeoExpress
- ImageMagick
- OpenJPEG

Extract:
- gi_detector
- ncsa.cv.caltech101
- ncsa.cv.closeups
- ncsa.cv.eyes
- ncsa.cv.faces
- ncsa.cv.meangrey
- ncsa.cv.profiles
- ncsa.im.metadat
- ncsa.im.scr

Output:

3) Code Snippets:
### General Software

#### Excel Green Route Index

1. Provide an Access Token: 81de62b9-430f-4d69-a64_

2. Select cells in the sheet as a 2 column list of latitude, longitude pairs along a path.

3. Submit to Brown Dog service. A new sheet will be added to the workbook with the results once they are ready.

   - Submit Selection

Submitted 6 rows and 2 columns.

Adding metadata to sheet

- Original File Metadata
  - Derived File 0 metadata
  - Derived File 1 metadata
  - Derived File 2 metadata
  - Derived File 3 metadata
  - Derived File 4 metadata
  - Derived File 5 metadata
  - Derived File 6 metadata
  - Derived File 7 metadata
  - Derived File 8 metadata
  - Derived File 9 metadata
  - Derived File 10 metadata
  - Derived File 11 metadata

"powered by BROWN DOG"

"powered by NCSA"
BD Command Line Interface

kmchenry@kgm-d3:~/git/bdcli$ bd -o png examples/photo_cd/IMG001.PCD | bd -v
BD Bookmarklet

Brown Dog Bookmarklet

Drag icon to bookmark toolbar to install

bdbookmarklet.js (beta)
<table>
<thead>
<tr>
<th>Name</th>
<th>Last modified</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent Directory</td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>0005.jp2</td>
<td>2014-07-03 13:49</td>
<td>633K</td>
<td></td>
</tr>
<tr>
<td>0094.jp2</td>
<td>2014-07-03 15:15</td>
<td>7.6M</td>
<td></td>
</tr>
<tr>
<td>105_1.sld</td>
<td>2014-07-03 13:47</td>
<td>3.7M</td>
<td></td>
</tr>
<tr>
<td>aac aif aiff ascii asf avi azw3 bmp clam dalee dib doc docx ed.zip eps epsi epub</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>City_Limits.7z</td>
<td>2014-10-01 12:43</td>
<td>763</td>
<td></td>
</tr>
<tr>
<td>Human_knownGene_chr22_1-51304566.bed</td>
<td>2015-04-14 13:57</td>
<td>260K</td>
<td></td>
</tr>
<tr>
<td>IMG_0997.jpg</td>
<td>2016-07-03 23:22</td>
<td>2.1M</td>
<td></td>
</tr>
<tr>
<td>alice.txt</td>
<td>2014-12-27 23:06</td>
<td>141K</td>
<td></td>
</tr>
</tbody>
</table>

Conversion menus added to links
Press Ctrl+F to index and search data on page
Lines of Code: 273

Other files:
- Model
- Image Sample

Dependencies:
- numpy, argparse, glob, cv2, cPickle, random, h5py, skimage, sklearn, scipy

Difficulties:
- Install OpenCV (cv2)
Lines of Code 47
Other files None
Dependencies bd, requests, os, glob, argparse, time, json, PIL
Difficulties

Total Code from 1 File
Welcome to Clowder

Welcome to Clowder, a scalable data repository where you can share, organize and analyze data. This is a demo instance to try the system out. Please do not use this instance to store real data. We delete the content of this instance when we need to and it does not have very much disk space available. Thank you.

Powered by Clowder (1.3.2+18 branch: master sha1:7a7a6f1e8).
Active Curation

- Active curation (AC) involves recording data and metadata as close to the source as practical and driving that acquisition through the deployment of capabilities that help data producers manage their research.

- Social Curation (SC) drives this economic analysis further, looking at ways that crossgroup interactions can further motivate best practices.

http://sead-data.net/sites/default/files/pubs/ActiveandSocialCurationKeystoDataServiceSustainability.pdf
Broad Data & Tool Support
Broad Data & Tool Support
# Metadata Definitions

The following metadata definitions will be available throughout Clowder.

<table>
<thead>
<tr>
<th>Label</th>
<th>URI</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td><a href="http://purl.org/dc/terms/abstract">http://purl.org/dc/terms/abstract</a></td>
<td>String</td>
</tr>
<tr>
<td>Alternative Title</td>
<td><a href="http://purl.org/dc/terms/alternative">http://purl.org/dc/terms/alternative</a></td>
<td>String</td>
</tr>
<tr>
<td>Audience</td>
<td><a href="http://purl.org/dc/terms/audience">http://purl.org/dc/terms/audience</a></td>
<td>String</td>
</tr>
<tr>
<td>CSDMS Standard Name</td>
<td><a href="http://csdms.colorado.edu/wiki/CSN_Searchable_List">http://csdms.colorado.edu/wiki/CSN_Searchable_List</a></td>
<td>List</td>
</tr>
<tr>
<td>ODM2 Variable Name</td>
<td><a href="http://vocabulary.odm2.org/variablename">http://vocabulary.odm2.org/variablename</a></td>
<td>List</td>
</tr>
<tr>
<td>References</td>
<td><a href="http://purl.org/dc/terms/references">http://purl.org/dc/terms/references</a></td>
<td>String</td>
</tr>
<tr>
<td>SAS Spatial Geocode</td>
<td><a href="http://ecgs.ncsa.illinois.edu/gsis/sas/geocode">http://ecgs.ncsa.illinois.edu/gsis/sas/geocode</a></td>
<td>Location</td>
</tr>
<tr>
<td>SAS Variable Name</td>
<td><a href="http://ecgs.ncsa.illinois.edu/gsis/sas/vars">http://ecgs.ncsa.illinois.edu/gsis/sas/vars</a></td>
<td>Queryable List</td>
</tr>
</tbody>
</table>

Metadata form with ODM2 Variable Name dropdown and added by Luigi.

Choose a dataset... what's that?

Existing Datasets

New Dataset

Choose a dataset template:
- polyvinyl nanoparticles

Choose a name for your dataset:

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Custom Community Dashboards
<table>
<thead>
<tr>
<th>XSEDE Image Analysis of Brain Anatomy</th>
<th>NSF</th>
<th>XSEDI Large Scale Video Analytics</th>
<th>NSF</th>
<th>XSEDI Real Stories of Bad Kids</th>
<th>NSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector-Base Database</td>
<td>CDC</td>
<td>NCIC</td>
<td>NSF</td>
<td>Sigma</td>
<td>NSF</td>
</tr>
<tr>
<td>Southern Methodist University</td>
<td>Syngenta</td>
<td>Chinese Academy of Sciences</td>
<td>NSF</td>
<td>LinkCGI</td>
<td>NSF</td>
</tr>
<tr>
<td>Taiwan MCHC</td>
<td>T2C</td>
<td>Europe Commission</td>
<td>NSF</td>
<td>European Commission</td>
<td>NSF</td>
</tr>
<tr>
<td>IMI-CHI</td>
<td>NIST</td>
<td>LinkCGI</td>
<td>NSF</td>
<td>European Commission</td>
<td>NSF</td>
</tr>
<tr>
<td>UNC-Chapel Hill</td>
<td>NSF</td>
<td>LinkCGI</td>
<td>NSF</td>
<td>European Commission</td>
<td>NSF</td>
</tr>
<tr>
<td>University of Michigan</td>
<td>NSF</td>
<td>LinkCGI</td>
<td>NSF</td>
<td>European Commission</td>
<td>NSF</td>
</tr>
</tbody>
</table>

- **Biology**
- **Civil Engineering**
- **Comp. and Inf. Science**
- **Education**
- **Geoscience**
- **Humanities**
- **Industry**
- **Materials Science**
- **Medicine**
- **Social Science**
https://www.workbench.nationaldataservice.org
Labs Workbench Catalog

Clowder
A scalable data repository where you can share, organize and analyze data.

DSpace
Turnkey software used for creating open access digital repositories.

ElasticSearch
A distributed search and analytics engine.

File Manager
Cloud Commander file manager and editor.

Girder
Web-based data management platform.

HTTP Tunnel
NDS Labs Utility for accessing TCP-based services via secure tunnel

iQvoc
SKOS-based vocabulary management system

iRODS Cloudbrowser
Web-based interface to any iRODS grid

iRODS iCAT
Open source data management

Jenkins CI
An open-source automation server

Docker
Docker command-line environment for NDS Labs

Dataverse
A web application to share, preserve, cite, explore and analyze research data.
<table>
<thead>
<tr>
<th>Status</th>
<th>Name</th>
<th>ID</th>
<th>Console</th>
<th>Edit</th>
<th>Remove</th>
<th>Help</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stopped</td>
<td>Dataverse</td>
<td>s6n4g2-dataverse</td>
<td>❯</td>
<td>❯</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stopped</td>
<td>PostgreSQL</td>
<td>s6n4g2-postgres</td>
<td>❯</td>
<td>❯</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stopped</td>
<td>Solr</td>
<td>s6n4g2-solr</td>
<td>❯</td>
<td>❯</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stopped</td>
<td>Rserve</td>
<td>s6n4g2-rserve</td>
<td>❯</td>
<td>❯</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For DSpace:

<table>
<thead>
<tr>
<th>Status</th>
<th>Name</th>
<th>ID</th>
<th>Console</th>
<th>Edit</th>
<th>Remove</th>
<th>Help</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stopped</td>
<td>DSpace</td>
<td>szca0w-dspace</td>
<td>❯</td>
<td>❯</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stopped</td>
<td>PostgreSQL</td>
<td>szca0w-postgres</td>
<td>❯</td>
<td>❯</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Here are all the dataverses, datasets, and files you have access to. You can filter through them by publication status and roles.

Search my data...  Find

Dataverses
Datasets
Files

Publication Status
Published
Unpublished
Draft
In Review
Deaccessioned

Roles
Admin
File Downloader
Dataverse + Dataset Creator
Dataverse Creator
Dataset Creator
Contributor
Curator
Member

Sorry, nothing was found for these roles: Admin, File Downloader, Dataverse + Dataset Creator, Dataverse Creator, Dataset Creator, Contributor, Curator, Member
Configuration for Dataverse

ID: s6n4g2-dataverse

Docker

<table>
<thead>
<tr>
<th>Registry</th>
<th>Image Name</th>
<th>Image Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>docker.io</td>
<td>ndslabs/dataverse</td>
<td>Default (latest)</td>
</tr>
</tbody>
</table>

Environment

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
<th>Default</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADMIN_PASSWORD</td>
<td>vuW994bZdR</td>
<td></td>
</tr>
<tr>
<td>MAIL_SERVER</td>
<td>smtp.ncsa.illinois.edu</td>
<td>smtp.ncsa.illinois.edu</td>
</tr>
<tr>
<td>POSTGRES_DB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>POSTGRES_PASSWORD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>POSTGRES_USER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RSERVER_PASSWORD</td>
<td>rserve</td>
<td>rserve</td>
</tr>
<tr>
<td>RSERVER_USER</td>
<td>rserve</td>
<td>rserve</td>
</tr>
<tr>
<td>TOOLMANAGER_URL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>THORAVENS_URL</td>
<td>https://$(NDSLABS_STACK)-tworavens.$(NDSLABS_DOMAIN)</td>
<td></td>
</tr>
</tbody>
</table>

Endpoints

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Internal Address</th>
<th>External Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTTP</td>
<td>10.254.79.87:8080</td>
<td>s6n4g2-dataverse.workbench.nationaldataverse.org</td>
</tr>
<tr>
<td>Status</td>
<td>Name</td>
<td>ID</td>
</tr>
<tr>
<td>------------</td>
<td>-----------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Stopped</td>
<td>Clowder</td>
<td>so150m-clowder</td>
</tr>
<tr>
<td>Stopped</td>
<td>MongoDB</td>
<td>so150m-mongo</td>
</tr>
<tr>
<td>Stopped</td>
<td>RabbitMQ</td>
<td>so150m-rabbitmq</td>
</tr>
<tr>
<td>Stopped</td>
<td>PlantCV extractor</td>
<td>so150m-plantcv</td>
</tr>
</tbody>
</table>

**Dataverse**

**DSpace**
Further/Custom Analysis
129,471,866 datasets/files transformed
Cost of the ”Cloud”

Big Data Hubs
Stimulate regional grassroots partnerships focused on Big Data

CC*
Coordinated campus-level cyberinfrastructure components of data, networking, and computing infrastructure across 200+ universities
## OSN Nodes

<table>
<thead>
<tr>
<th></th>
<th>88-Disk OSN Node</th>
<th>24 Disk OSN Node (128GB)</th>
<th>24 Disk OSN Node (256GB)</th>
<th>36 Disk OSN Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Sockets</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Per-CPU Cores (HT)</td>
<td>6 (12)</td>
<td>12 (24)</td>
<td>12 (24)</td>
<td>18 (36)</td>
</tr>
<tr>
<td>Total CPU Cores (HT)</td>
<td>12 (24)</td>
<td>12 (24)</td>
<td>12 (24)</td>
<td>18 (36)</td>
</tr>
<tr>
<td>CPU Speed</td>
<td>3.4 GHz</td>
<td>2.1 GHz</td>
<td>2.1 GHz</td>
<td>2.3 GHz</td>
</tr>
<tr>
<td>Total GFLOPS</td>
<td>652.8</td>
<td>403.2</td>
<td>403.2</td>
<td>662.4</td>
</tr>
<tr>
<td>Total RAM (GB)</td>
<td>256</td>
<td>128</td>
<td>256</td>
<td></td>
</tr>
<tr>
<td>RAM Speed (MHz)</td>
<td>2400</td>
<td>2400</td>
<td>2400</td>
<td>2666</td>
</tr>
<tr>
<td>Drive Count (AKA OSD)</td>
<td>88</td>
<td>24</td>
<td>24</td>
<td>36</td>
</tr>
<tr>
<td>Per-Node Data Storage (TB)</td>
<td>704</td>
<td>192</td>
<td>192</td>
<td>288</td>
</tr>
<tr>
<td>~1.5PB Node Count</td>
<td>2</td>
<td>8</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Total Data Storage</td>
<td>1408</td>
<td>1536</td>
<td>1536</td>
<td>1440</td>
</tr>
<tr>
<td>Node Cost</td>
<td>$51,492.00</td>
<td>$15,528.00</td>
<td>$16,672.00</td>
<td>$17,845.00</td>
</tr>
<tr>
<td>Total Cost</td>
<td>$102,984.00</td>
<td>$124,224.00</td>
<td>$133,376.00</td>
<td>$89,225.00</td>
</tr>
<tr>
<td>$ per TB</td>
<td>$73.14</td>
<td>$80.88</td>
<td>$86.83</td>
<td>$61.96</td>
</tr>
<tr>
<td>Utilized Rack Space (RU)</td>
<td>24</td>
<td>36</td>
<td>36</td>
<td>24</td>
</tr>
<tr>
<td>Max Power Consumption (Watts)</td>
<td>7600</td>
<td>9600</td>
<td>9600</td>
<td>6800</td>
</tr>
<tr>
<td>Ceph RAM:Storage Cap Ratio (Nearest to 1 is best)</td>
<td>0.36</td>
<td>0.67</td>
<td>1.33</td>
<td>0.89</td>
</tr>
<tr>
<td>Ceph CPU Thread: OSD Ratio (Nearest to 1 is best)</td>
<td>0.27</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Open Storage Network

Midwest Big Data Hub
Accelerating the Big Data Innovation Ecosystem

Hubs based on Census Regions of the United States

Alaska & Hawaii are part of the West Region
US Territories can participate in any region
## 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, facilitate inter-site data processing</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>
Architecture

Leverage NSF Data Software Investments e.g. DataNet, DIBBs, CSSI

OSN Pod

<table>
<thead>
<tr>
<th>Globus</th>
<th>Clowder</th>
<th>WholeTale</th>
<th>iRODS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>S3 API</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ceph + radosgw</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Remote Operations, Boot, Monitoring, Management</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OSN Hardware</td>
</tr>
</tbody>
</table>

Leverage NSF XSEDE Capabilities

AMIE  XRAS
OSN Clowder Interface

- IML-CZO use case data