Using Open Science Data Cloud
An open-source, cloud-based infrastructure that provides the scientific community with resources for managing terabyte and petabyte-scale scientific datasets.
Storing data
- 10 PB of storage across all resources
- Public and protected infrastructures for sensitive data
- 1 PB Public Data Commons of popular scientific datasets

Analyzing data
- About 9000 compute cores
- Both OpenStack and Hadoop-based clouds

Sharing data and analysis tools
- Access to shared group storage for collaborations
- Access to a public pool of virtual machine snapshots
- Authenticated log-in using University/Institution credentials
- Connected to high-performance networks for data transfer
Cloud Resources by cluster size / type

Hadoop Clusters
- Skidmore (900 cores)
- OCC-Y (1100 cores)

General Compute (OpenStack)
- Sullivan (1200 cores)
- Adler (300 cores)
- Goldberg (1800 cores)

Protected Compute (health data)
- Bionimbus (3000 cores)
- Atwood (360 cores)

Public Access Datasets
- Public Data Commons (1 PB storage)
The OSDC is an interdisciplinary hub.

- Over 700+ total unique user accounts
- Users from 140+ different institutions
- Each month, on average
  - About 200 unique users
  - 1.8 million core hours
  - 800 TB user data stored

- Researchers in
  - Biology
  - Medicine
  - Computer Science
  - Mathematics
  - Earth Science
  - Social Science
  - Urban Science
  - Digitized Humanities
Protected data cloud **Bionimbus-PDC** (PDC) for analyzing human genomic data

- Collaboration with Institute for Genomics and Systems Biology (IGSB) at UChicago
- Allows users authorized by NIH to compute over human genomic data from dbGaP in a secure and compliant fashion
- Contains data from The Cancer Genome Atlas (TCGA)

**bionimbus-pdc.openscienceidatacloud.org**

*Featured Projects: Biology*
Bookworm

Web-based tool hosted by OSDC for visualizing trends in repositories of digitized texts

- Open Library books
- arXiV science publications
- Chronicling America historical newspapers
- US Congress bills
- Social Science Research Network paper abstracts
- Create your own bookworm

bookworm.culturomics.org
Collaboration with NASA to develop open source technology for cloud-based processing of satellite imagery to support earth sciences.

The OSDC is used to process Earth Observing 1 (EO-1) satellite imagery from the Advanced Land Imager (ALI) and the Hyperion instruments and to make this data available to interested users.

- Namibia flood dashboard
- Hadoop-based Matsu “Wheel” system for processing all data

matsu.opensciencedatacloud.org
User view of OSDC “Tukey” web portal log-in
Users log-in to web portal with University credentials
User view of OSDC “Tukey” web portal

Quota Summary
Used 2 of 42 Available Instances

Used 16 of 48 Available vCPUs

Used 32,768 MB of 132,072 MB Available RAM

Select a month to query its usage:

November 2014

This Month's Cloud Core Hours:
This Month's Cloud Disk Usage (GB):
This Month's Hadoop Disk Usage (GB):
This Month's Hadoop Job Hours:

Download CSV Summary
User view of OSDC “Tukey” web portal

<table>
<thead>
<tr>
<th>Instance Name</th>
<th>IP Address</th>
<th>Size</th>
<th>Keypair</th>
<th>Status</th>
<th>Task</th>
<th>Power State</th>
<th>Cloud</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeoServer_20141112</td>
<td>172.16.1.81</td>
<td>m1.xlarge</td>
<td>16GB RAM</td>
<td>8 VCPU</td>
<td>20GB Disk</td>
<td>mpatterson</td>
<td>Active</td>
<td>None</td>
</tr>
<tr>
<td>atwood_lme4_mcmcB</td>
<td>172.16.1.20</td>
<td>m1.xlarge</td>
<td>16GB RAM</td>
<td>8 VCPU</td>
<td>20GB Disk</td>
<td>Active</td>
<td>None</td>
<td>Running</td>
</tr>
</tbody>
</table>
### User view - launching a VM

#### Images

<table>
<thead>
<tr>
<th>Image Name</th>
<th>Type</th>
<th>Status</th>
<th>Public</th>
<th>Format</th>
<th>Cloud</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>new-cube-20141030</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>QCOW2</td>
<td>Sullivan</td>
<td>Launch</td>
</tr>
<tr>
<td>Ubuntu-14.04-LTS-v1.0</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>QCOW2</td>
<td>Sullivan</td>
<td>Launch</td>
</tr>
<tr>
<td>CentOS-5.10-v1.0</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>QCOW2</td>
<td>Sullivan</td>
<td>Launch</td>
</tr>
<tr>
<td>CentOS-6.0-v1.0</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>QCOW2</td>
<td>Sullivan</td>
<td>Launch</td>
</tr>
<tr>
<td>Ubuntu-12.04-LTS-v1.3</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>QCOW2</td>
<td>Sullivan</td>
<td>Launch</td>
</tr>
<tr>
<td>Ubuntu-14.04-LTS-v1.0</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>QCOW2</td>
<td>Atwood</td>
<td>Launch</td>
</tr>
<tr>
<td>Ubuntu-12.04-LTS-v1.6</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>QCOW2</td>
<td>Atwood</td>
<td>Launch</td>
</tr>
<tr>
<td>Ubuntu-12.04-LTS-v1.5_x64v4</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>QCOW2</td>
<td>Atwood</td>
<td>Launch</td>
</tr>
<tr>
<td>Ubuntu-12.04-LTS-v1.5</td>
<td>Image</td>
<td>Active</td>
<td>Yes</td>
<td>QCOW2</td>
<td>Atwood</td>
<td>Launch</td>
</tr>
</tbody>
</table>

#### User Snapshots

<table>
<thead>
<tr>
<th>Image Name</th>
<th>Type</th>
<th>Status</th>
<th>Public</th>
<th>Format</th>
<th>Cloud</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>pireweb2</td>
<td>Snapshot</td>
<td>Active</td>
<td>No</td>
<td>QCOW2</td>
<td>Sullivan</td>
<td>Launch</td>
</tr>
<tr>
<td>pireweb</td>
<td>Snapshot</td>
<td>Active</td>
<td>No</td>
<td>QCOW2</td>
<td>Sullivan</td>
<td>Launch</td>
</tr>
<tr>
<td>OSDC_PiREtools</td>
<td>Snapshot</td>
<td>Active</td>
<td>Yes</td>
<td>QCOW2</td>
<td>Sullivan</td>
<td>Launch</td>
</tr>
<tr>
<td>OSDC_ToolExplorer_UDR</td>
<td>Snapshot</td>
<td>Active</td>
<td>Yes</td>
<td>QCOW2</td>
<td>Sullivan</td>
<td>Launch</td>
</tr>
</tbody>
</table>
User view – log on to VM, install tools, ready for analysis
Tips

- Public datasets are available and automatically mounted to your virtual machines at /glusterfs/osdc_public_data/

- Your home directory is in /glusterfs/users/username
  - Anything in there is accessible from all virtual machines
  - Anything you store anywhere else will go away when your VM is terminated

- Create a keypair under “Access and Security” and choose “All resources”.

- Remember to make two ‘hops’ and carry your key on both hops to get to your virtual machine
  - ssh-add yourkeypair.pem
  - ssh -A username@sullivan.opensciencedatacloud.org
  - ssh -A ubuntu@ipaddress

- Check the status page: www.opensciencedatacloud.org/status/
- Read the docs: www.opensciencedatacloud.org/support/
ID and Metadata Services for Data Commons
Challenges

Significant concerns for the scientific and research community in storing data long-term in an accessible and usable manner

- Preserving data provenance
  - What are these data, how were they produced?
- Maintaining scientific reproducibility and workflows
  - Can I access these data the same way I used to?

Challenges:

- Data moves, location changes.
  - Hardware dies/changes, no home for data.
- Identifiers
  - Variety of identifiers- DOI, ARK, UUID, etc

Identification:

Hello
My name is:

ark:/31807/12345
aka EO1H12015LLTG
aka ‘mydata’
ID and metadata services

ID services goal:

- Flexibility to support data in multiple locations/access points
- Flexibility to support multiple identifiers

Metadata services goal:

- For all data (cross-discipline), support simple core metadata
- Support or allow discipline-specific metadata for search capabilities as needed per field
ID service user demo with EO-1 data: Signpost