Workshop session:
Using Pangeo JupyterHubs to work with large public datasets

Scott Henderson, Amanda Tan, Anthony Arendt (UW eScience)
Andrew Pawloski (Element 84)
Session take-aways

1) Pangeo is not a single platform, but a coordination of open source software with distributed community support

2) Goal: facilitate scientific computational workflows on Cloud or HPC infrastructure and overcome major data management bottlenecks

3) There are various ways to use Pangeo JupyterHubs currently and contributions are welcome!
Community Tools for Analysis of NASA Earth Observation System Data in the Cloud

Project Overview

Proposal (project) Number: 17-ACCESS17-0003
Co-Operative Agreement Number(s): 80NSSC18M0157, 80NSSC18M0158, 80NSSC18M0159
PIs: Anthony Arendt (1), Ethan Gutmann (2), Daniel Pilone (3)
Institutions: (1) University Of Washington, Seattle, (2) University Corporation For Atmospheric Research, (3) Element 84, Inc.
THE BIG DATA GEOSCIENCE ERA IS HERE

From Remote Sensing Platforms:
- New sensors / platforms
- Continuous observations
- Multiple versions of derived datasets

Projected NASA Cloud Storage

From Earth System Models:
- Higher resolution
- More process representation
- Larger ensembles
- On track for exabytes by CMIP7
PANGEO OVERVIEW

A community platform for Big Data geoscience

WHAT IS PANGEO?

Pangeo is first and foremost a community promoting open, reproducible, and scalable science. This community provides documentation, develops and maintains software, and deploys computing infrastructure to make scientific research and programming easier. The Pangeo software ecosystem involves open source tools such as xarray, iris, dask, jupyter, and many other packages. There is no single software package called “pangeo”; rather, the Pangeo project serves as a coordination point between scientists, software, and computing infrastructure. On this website, scientists can find guides for accessing data and performing analysis using these tools (read the Guide for Scientists, browse the Geoscience Use Cases, and learn about the Packages). Those interested in building infrastructure can find instructions for deploying Pangeo environments on HPC or cloud clusters (learn about the Technical Architecture or read the Deployment Setup Guides). For more general information, read About Pangeo, see the Collaborators, or read the Frequently Asked Questions. Welcome to the Pangeo community!

http://pangeo.io
https://github.com/pangeo-data/pangeo
The Pangeo Architecture

Analysis Ready Data
Stored and cataloged on
globally-available distributed
storage (e.g. S3, GCS)

Distributed storage

Cloud/HPC

Parallel computing system built
on top of Kubernetes or HPC.
Dask tells the nodes what to do.

Jupyter for interactive access
on remote systems

Xarray provides data structures
and intuitive interface for
interacting with datasets
SCIENTIFIC PYTHON FOR DATA SCIENCE

Credit: Stephan Hoyer, Jake Vanderplas (SciPy 2015)
PANGEO COLLABORATORS

And many more...
# BUILD YOUR OWN PANGEO

<table>
<thead>
<tr>
<th>Storage Formats</th>
<th>HDF</th>
<th>OPeNDAP</th>
<th>Cloud Optimized COG/Zarr/Parquet/etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ND-Arrays</td>
<td>NumPy</td>
<td>DASK</td>
<td>More coming...</td>
</tr>
<tr>
<td>Data Models</td>
<td>xarray</td>
<td>Iris</td>
<td>pandas</td>
</tr>
<tr>
<td>Processing Mode</td>
<td>Jupyter</td>
<td>Interactive</td>
<td>Batch</td>
</tr>
<tr>
<td>Compute Platform</td>
<td>HPC</td>
<td>Cloud</td>
<td>Local</td>
</tr>
</tbody>
</table>
NASA ACCESS Project Overview

Main Goal:
Facilitate the Geoscience community's transition into cloud computing by building on top of the growing Pangeo ecosystem.
1. Deploy a scalable cloud-based JupyterHub on AWS for community use
2. Integrate existing NASA data discovery tools with cloud based data access protocols
3. Create an advanced, cloud-native framework for custom analysis of large remote data archives
4. Demonstrate scientific use-cases with GRACE, Sentinel-1, and Hydrologic Models
Project Team

Anthony Arendt
PI, University of Washington

Ethan Gutmann
PI, NCAR/RAL

Matt Rocklin
Former PI, Anaconda Inc.

Dan Pilone
PI, Element 84

Rob Fatland
Co-I, University of Washington

Joe Hamman
Co-I, NCAR/CGD

Scott Henderson
Co-I, University of Washington

Andrew Pawlowski
Element 84

Amanda Tan
University of Washington
Develop Python based APIs to NASA metadata repositories

**Goal:** Provide high-level advanced search, discovery, and ingestion capabilities to NASA datasets

**Approach:**
- CMR is the fundamental building block for search and discovery
- CMR’s Python API is too fine-grained to put into analysis code
- Build a data ingestion API on top of CMR

**Status:**
- Developed examples to understand usage patterns
- Generated a design document to bring together CMR, STAC, and Intake
- Development of Stac-Intake package is beginning now
Develop Python based API to access cloud hosted NASA datasets

```python
import xarray as xr
fname = "https://landsat-pds.s3.amazonaws.com/L8/001/002/LC80010022016230LGN00/LC80010022016230LGN00_B1.TIF"
ds = xr.open_rasterio(fname, chunks={'x': 256, 'y': 256})
ds
```
Develop Python based API to access cloud hosted NASA datasets

Bringing it all together

- Python already has the building blocks for accessing data stored in the Cloud
- We’re largely focusing our effort on integrating and extending existing tools

https://medium.com/pangeo

Cloud Native Geoprocessing of Earth Observation Satellite Data with Pangeo
Integrate Python APIs with Pangeo Ecosystem

What’s going on under the hood?

• Intake provides a data catalog and API
• Intake uses Xarray and Dask
• Xarray uses RasterIO for access to COGs stored on AWS
• RasterIO uses GDAL
Configure and deploy JupyterHub on cloud resources

Initial deployment of Pangeo Jupyterhub on AWS:

- Live at: http://nasa.pangeo.io/
- Currently improving kubernetes integration and user authentication methods

Continued development of Pangeo BinderHub:

- Live at: https://binder.pangeo.io/
- BinderHub builds on top of JupyterHub and provides a useful tool for sharing reproducible scientific workflows
Develop documentation and real-world examples

Tutorials
- AGU tutorial and ESIP-tech-dive are two examples

Online Documentation
- Improvements to Pangeo, Xarray, and Dask user guides

Science Use Cases
- Sentinel-1 regional time series (displacements from slow-slip earthquakes, landslide monitoring, annual snowpack properties).
- Getting Ready for NISAR (GRFN) Cloud-native analysis

github.com/pangeo-data/pangeo-tutorial-agu-2018
Hackweeks to support community training

[Image of GeoHackWeek 2019]
https://geohackweek.github.io

[Image of OceanHackWeek 2019]
https://oceanhackweek.github.io/

[Image of Cryospheric Science with ICESat-2 HackWeek 2019]
ICESat-2 Medium Blog Post
Demo Links

- Introductory materials (Google Cloud)  

- Landsat 8 analysis (AWS)  
Thanks for attending!

bit.ly/pangeo-workshop-survey