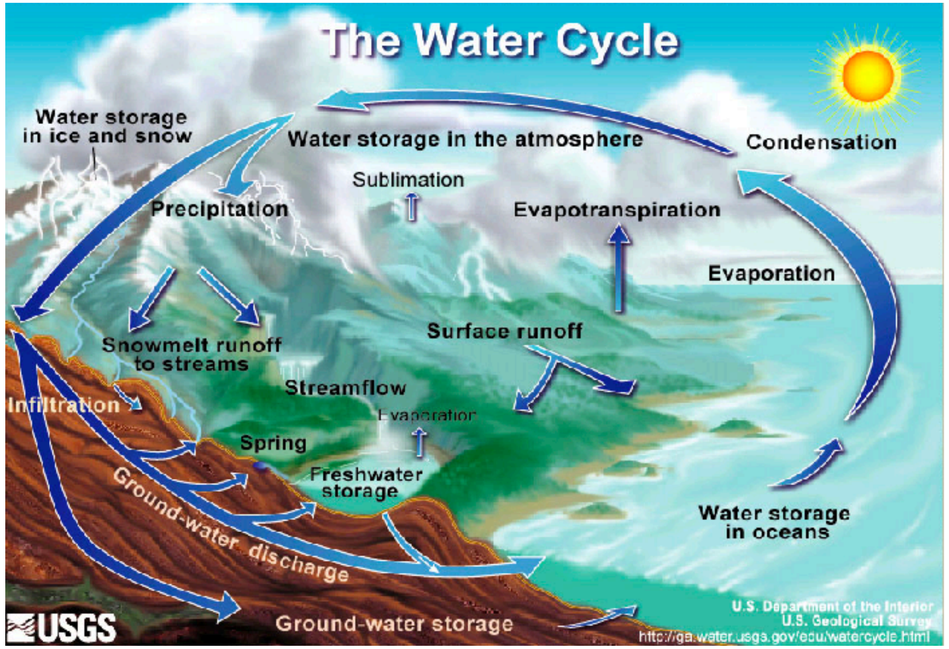
PANGEO

A COMMUNITY-DRIVEN EFFORT FOR
BIG DATA GEOSCIENCE

HELLO!

- Who am I?
  - Joe Hamman, Ph.D., P.E.
  - I am a scientist at the National Center for Atmospheric Research
  - I study the impacts of climate change on the water cycle.
  - I am a core developer of Xarray
  - I am a founding member of the Pangeo project



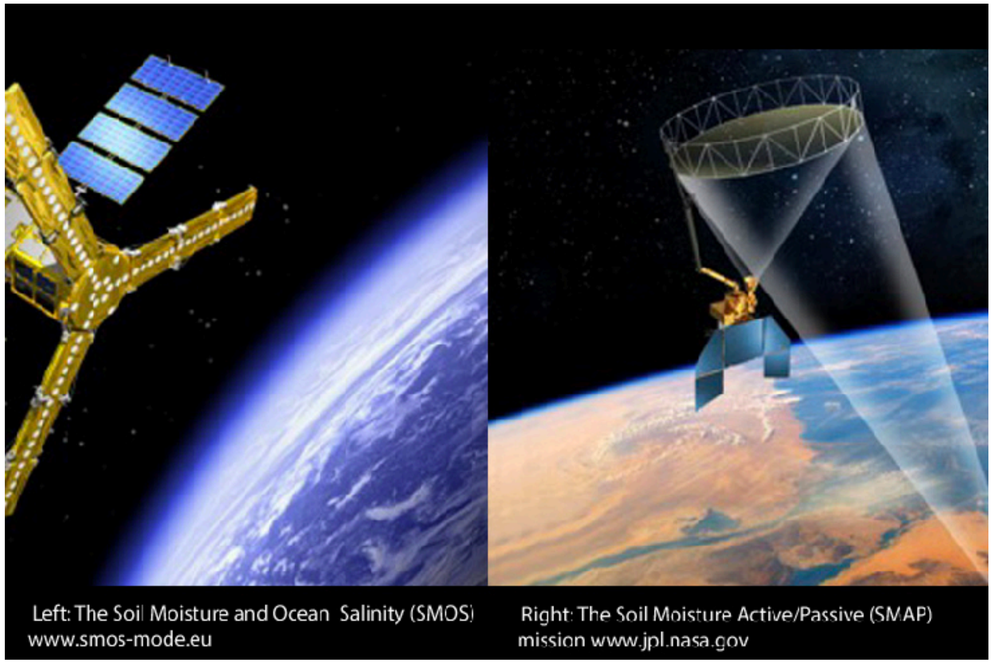
Github: @jhamman
Twitter: @HammanHydro
Web: joehamman.com

PANGEO

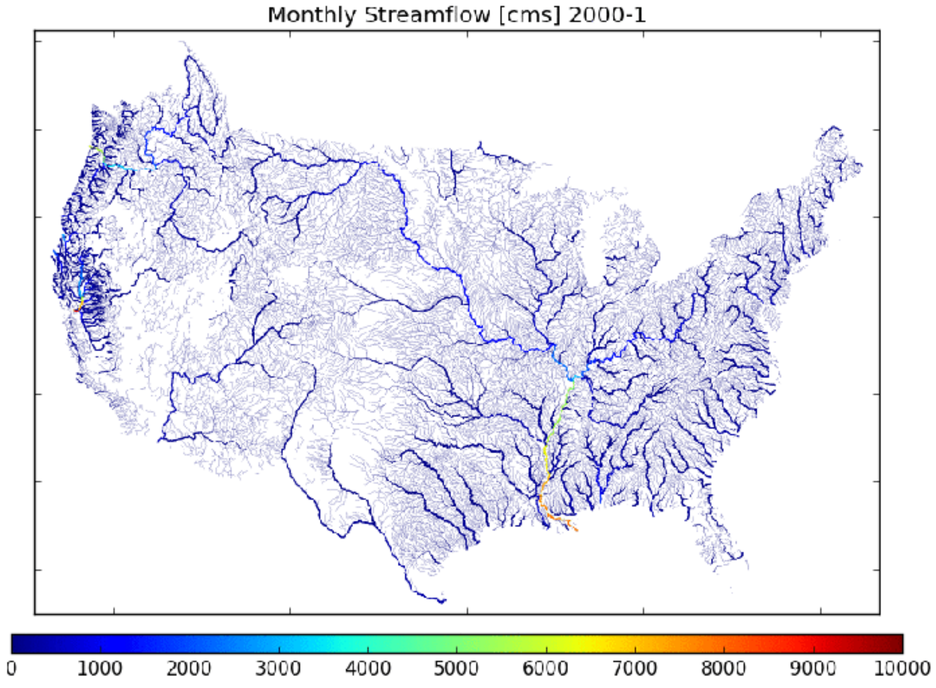
BIG DATA IN THE GEOSCIENCES

We use our observations to test our models… and our models to test our observations

Observations



Simulations

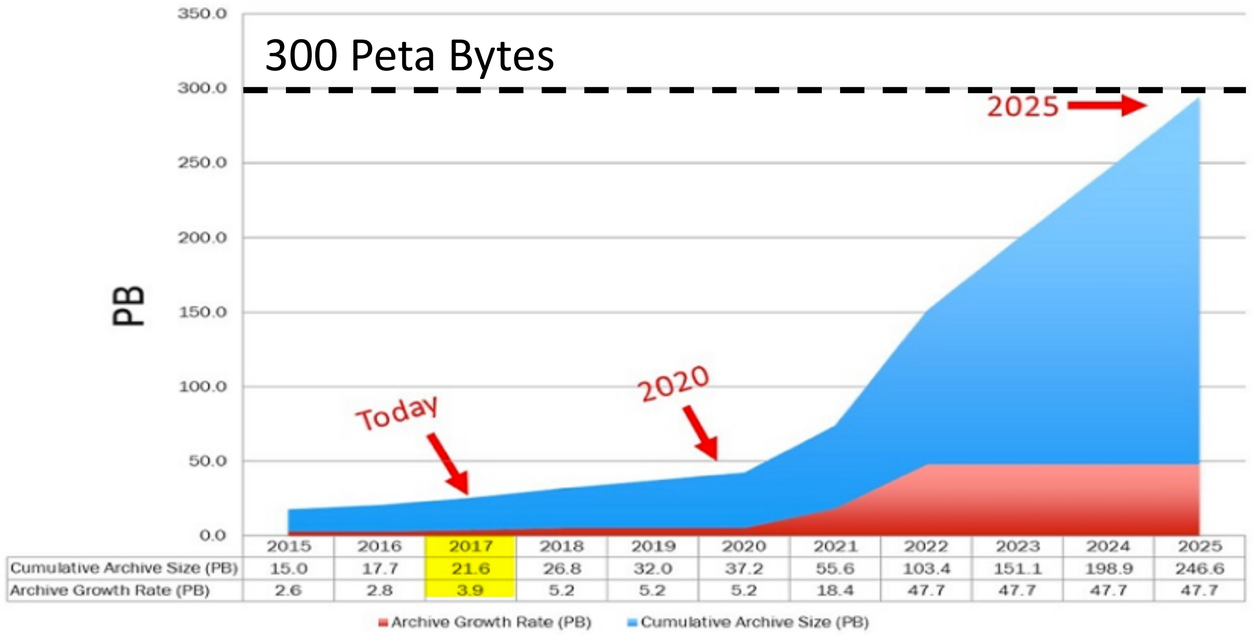


BIG DATA IN THE GEOSCIENCES

Growth in observed datasets

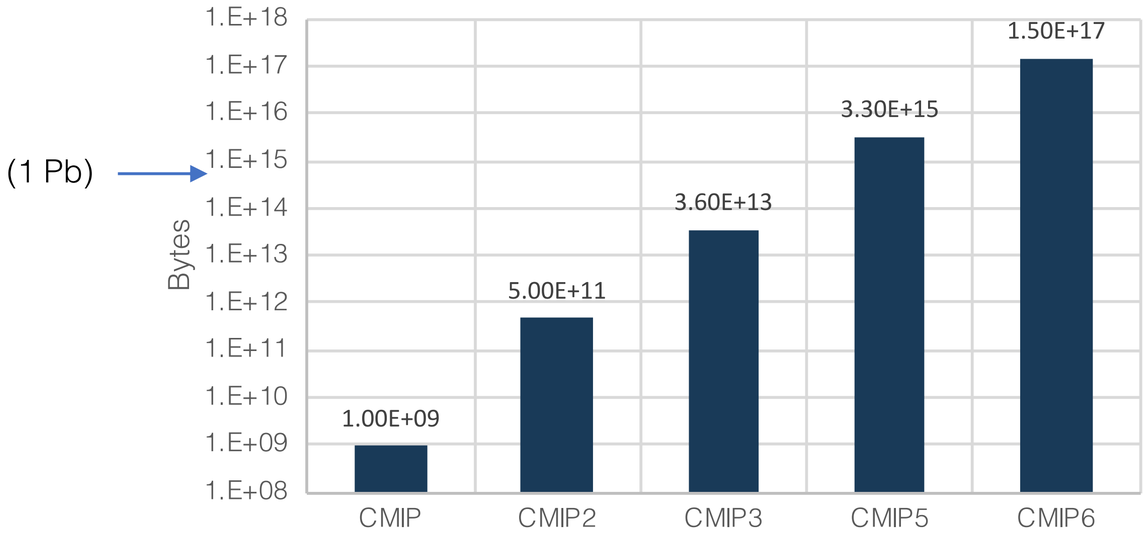
- New sensors / platforms
- Continuous observations
- Multiple versions of derived datasets

Projected NASA Cloud Storage



	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Cumulative Archive Size (PB)	15.0	17.7	21.6	26.8	32.0	37.2	55.6	103.4	151.1	198.9	246.6
Archive Growth Rate (PB)	2.6	2.8	3.9	5.2	5.2	5.2	18.4	47.7	47.7	47.7	47.7

Size of CMIP Archives



Growth in model generated datasets

- Higher resolution
- More process representation
- Larger ensembles
- On track for exabytes by CMIP7

WHAT IS PANGEO?

Pangeo is a community working to develop software and infrastructure to enable big-data geoscience.

- **Mission:** To cultivate an ecosystem in which the next generation of open-source analysis tools for the big-data geosciences can be developed, distributed, and sustained.

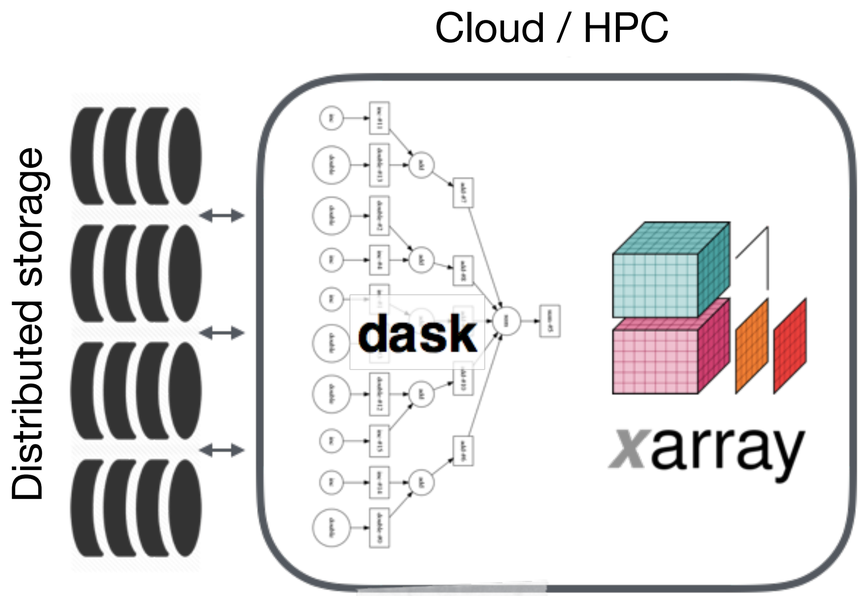
- **Vision:**
  - Open and collaborative development
  - Tools for scaling computations from small to very large datasets
  - Frameworks for moving scientific analysis to the data
  - Welcoming and inclusive development culture

SCIENTIFIC PYTHON FOR GEOSCIENCE

Credit: Stephan Hoyer, Jake Vanderplas (SciPy 2015)

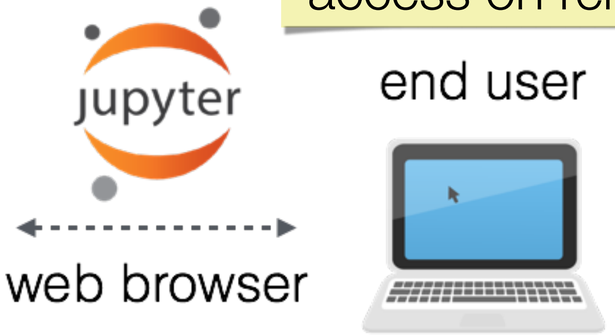
PANGEO ARCHITECTURE

"Analysis Ready Data" stored on globally-available distributed storage.



Parallel computing system allows users deploy clusters of compute nodes for data processing.

Dask tells the nodes what to do.

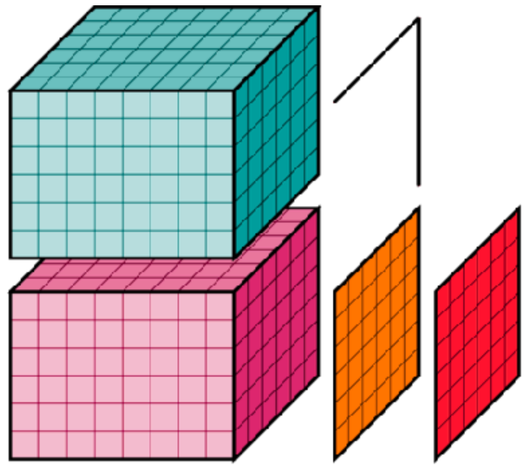


Jupyter for interactive access on remote systems

Xarray provides data structures and intuitive interface for interacting with datasets

XARRAY

http://xarray.pydata.org



xarray

- label-based indexing and arithmetic
- interoperability with the core scientific Python packages (e.g., pandas, NumPy, Matplotlib)
- out-of-core computation on datasets that don't fit into memory (thanks dask!)
- wide range of input/output (I/O) options: netCDF, HDF, geoTIFF, zarr
- advanced multi-dimensional data manipulation tools such as group-by and resampling

PANGEO

DASK

http://dask.pydata.org

Dask is a flexible parallel computing library for analytic computing

	8	8	8
5	('x', 0, 0)	('x', 0, 1)	('x', 0, 2)
5	('x', 1, 0)	('x', 1, 1)	('x', 1, 2)
5	('x', 2, 0)	('x', 2, 1)	('x', 2, 2)
5	('x', 3, 0)	('x', 3, 1)	('x', 3, 2)

ND-Arrays are split into chunks that comfortably fit in memory

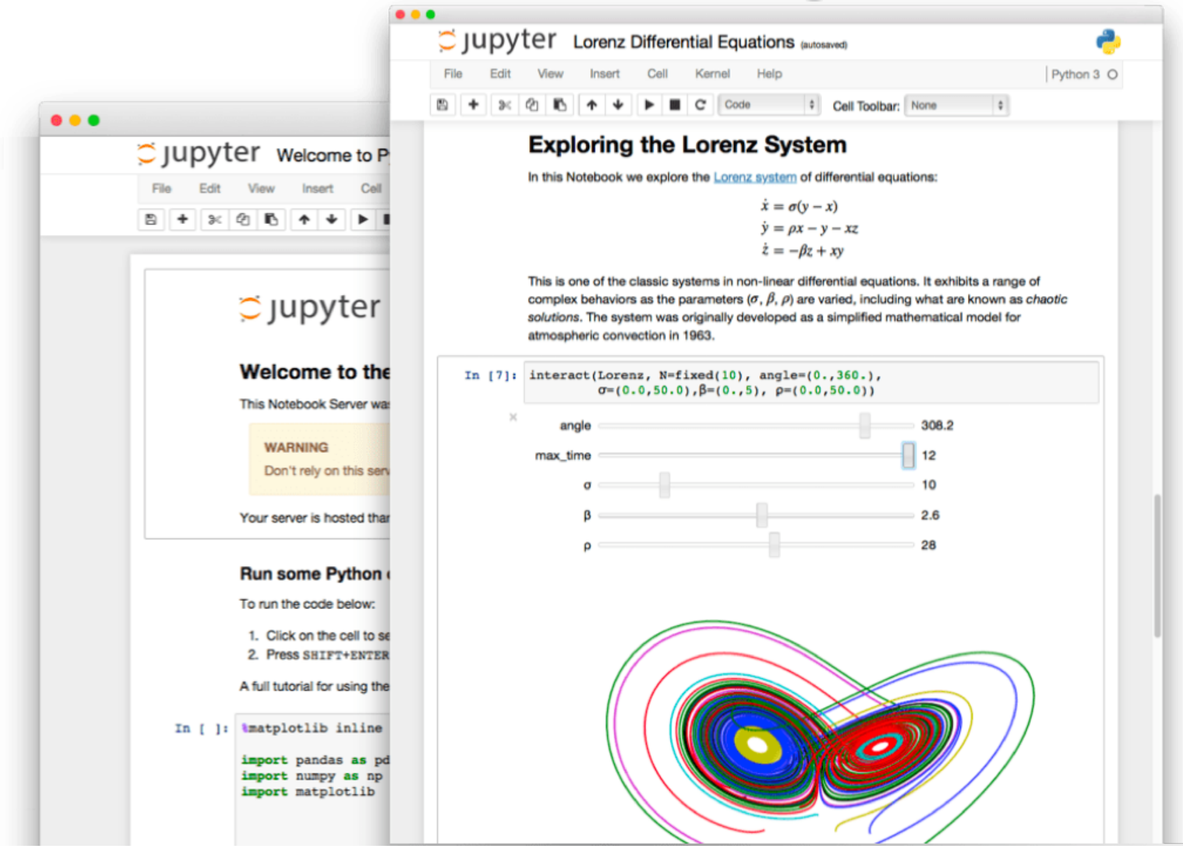
Complex computations represented as a graph of individual tasks.

Scheduler optimizes execution of graph.

JUPYTER

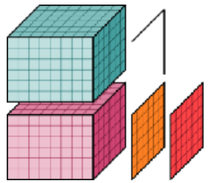
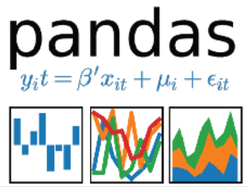
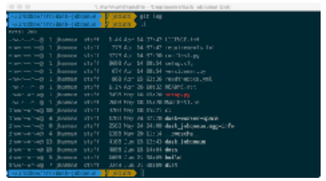
http://jupyter.org

- Jupyter Notebooks are web applications that facilitate interactive data analysis
- Core Python support but also may use kernels from over 40 other languages
- JupyterHub is a multi-user gateway Jupyter notebooks
- Jupyter Notebooks / JupyterHubs can be readily deployed in a range of computing environments (local/HPC/Cloud)



PANGEO

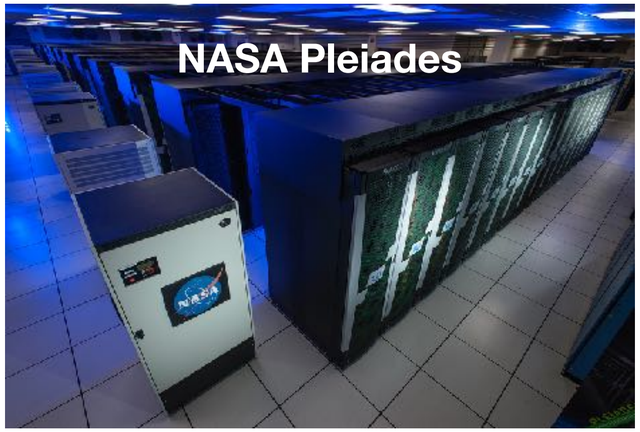
BUILD YOUR OWN PANGEO

Storage Formats	Binary/ File-based 	Data services 	Cloud Optimized COG/Zarr/Parquet/etc.
ND-Arrays			More coming…
Data Models	 		
Processing Mode	 Interactive	Batch 	Serverless 
Compute Platform	HPC 	Cloud 	Local 

Interoperability is key!

PANGEO DEPLOYMENTS

HTTP://PANGEO-DATA.ORG/DEPLOYMENTS.HTML



NCAR Cheyenne

(SCALE USING JOB QUEUE SYSTEM)

PANGEO.PYDATA.ORG

Google Cloud Platform

Over 500 unique users since March!

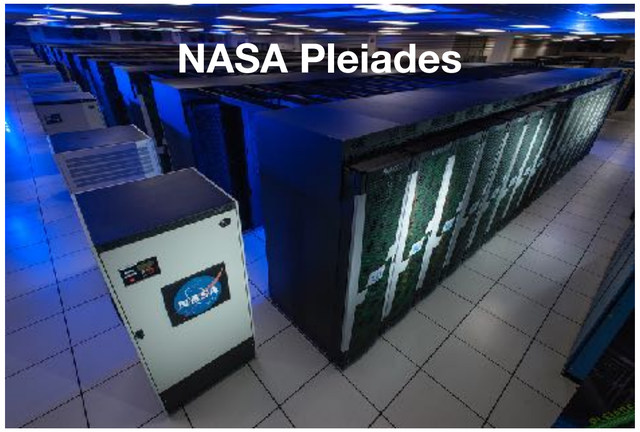
Microsoft Azure

aws

(SCALE USING KUBERNETES)

PANGEO DEPLOYMENTS

HTTP://PANGEO-DATA.ORG/DEPLOYMENTS.HTML



NCAR Cheyenne

(SCALE USING JOB QUEUE SYSTEM)

- Dask-Jobqueue helps us deploy/scale Dask on HPC systems
  - Easily deploy Dask on job queuing systems like PBS, Slurm, MOAB, and SGE.
  - High-level interactive Python interface

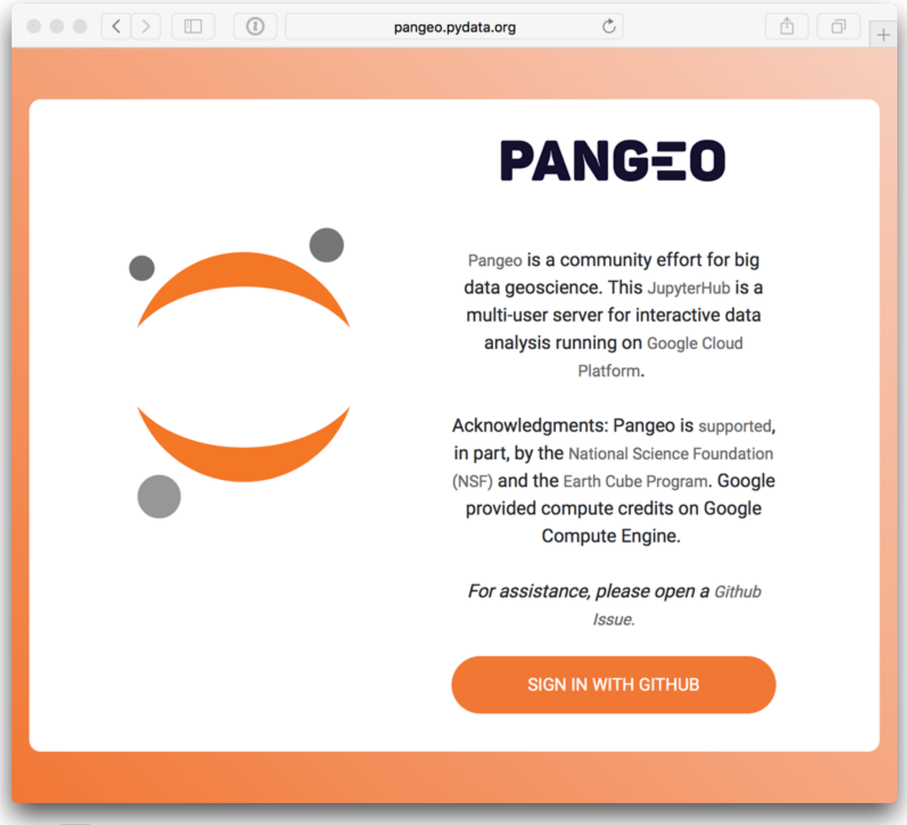
```python
from dask_jobqueue import PBSCluster
cluster = PBSCluster()
cluster.scale(10)          # Ask for ten jobs

from dask.distributed import Client
client = Client(cluster)  # Connect this local process to remote workers

# wait for jobs to arrive, depending on the queue, this may take some time

import dask.array as da
x = ...                   # Dask commands now use these distributed resources
```

PANGEO.PYDATA.ORG



- **What is pangeo.pydata.org?**
  - Multi-user JupyterHub running on Google Cloud Platform
  - Zero-to-jupyterhub deployment using Kubernetes
  - Dask scales using “Dask-Dubernetes”
- **Why the cloud?**
  - Highly scalable (storage, compute, user access)
  - Easy to customize
  - Cost effective

Google Cloud Platform

CURRENT CHALLENGES

1. STORING/SHARING N-DIMENSIONAL DATA IN THE CLOUD

- **Too big to move**: assume data is to be used but not copied
- **Self-describing**: data and metadata packaged together
- **On-demand**: data can be read/used in its current form from anywhere
- **Analysis-ready**: no pre-processing required
- **Discoverable**: easy to use/maintain data catalogs

CURRENT CHALLENGES

2. MANAGING MULTIPLE DEPLOYMENTS

- **Maintenance overhead**: multiple data-proximate deployments = multiple things to maintain
- **User/admin controls**: need to establish better ways of managing user environments and controls

3. CULTURAL CHALLENGES

- **Inertia**: there are many well-established workflows that already exist
- **Understanding costs**: need new way of thinking about paying for compute/storage

PANGEO

HOW TO GET INVOLVED

HTTP://PANGEO-DATA.ORG

- Access and existing Pangeo deployment on an HPC cluster, or cloud resources (eg. pangeo.pydata.org)
- Adapt Pangeo elements to meet your projects needs (data portals, etc.) and give feedback via github: github.com/pangeo-data/pangeo
- Participate in open-source software development!
- Come to the Pangeo BOF at 1 PM

EARTHCUBE AWARD TEAM

EARTHCUBE

Google Cloud Platform

Lamont-Doherty Earth Observatory
COLUMBIA UNIVERSITY | EARTH INSTITUTE

Ryan Abernathey, Chiara Lepore, Michael Tippet, Naomi Henderson, Richard Seager

Kevin Paul, Joe Hamman, Ryan May, Davide Del Vento

Matthew Rocklin

OTHER CONTRIBUTORS

Jacob Tomlinson, Niall Roberts, Alberto Arribas

Developing and operating Pangeo environment to support analysis of UK Met office products

Rich Signell

Deploying Pangeo on AWS to support analysis of coastal ocean modeling

Justin Simcock

Operating Pangeo in the cloud to support Climate Impact Lab research and analysis

Supporting Pangeo via SWOT mission and recently funded ACCESS award to UW / NCAR 🎉

Yuvi Panda, Chris Holdgraf

Spending lots of time helping us make things work on the cloud

PANGEO

PANGEO-DATA.ORG

PANGEO.PYDATA.ORG