PANGEO FORGE

A cloud-native data repository for ocean, weather, and climate science.

RYAN ABERNATHEY

WHO AM I?

Physical Oceanographer

Ph.D. From MIT, 2012

Associate Prof. at Columbia / LDEO

https://ocean-transport.github.io/

Core developer of *Xarray*

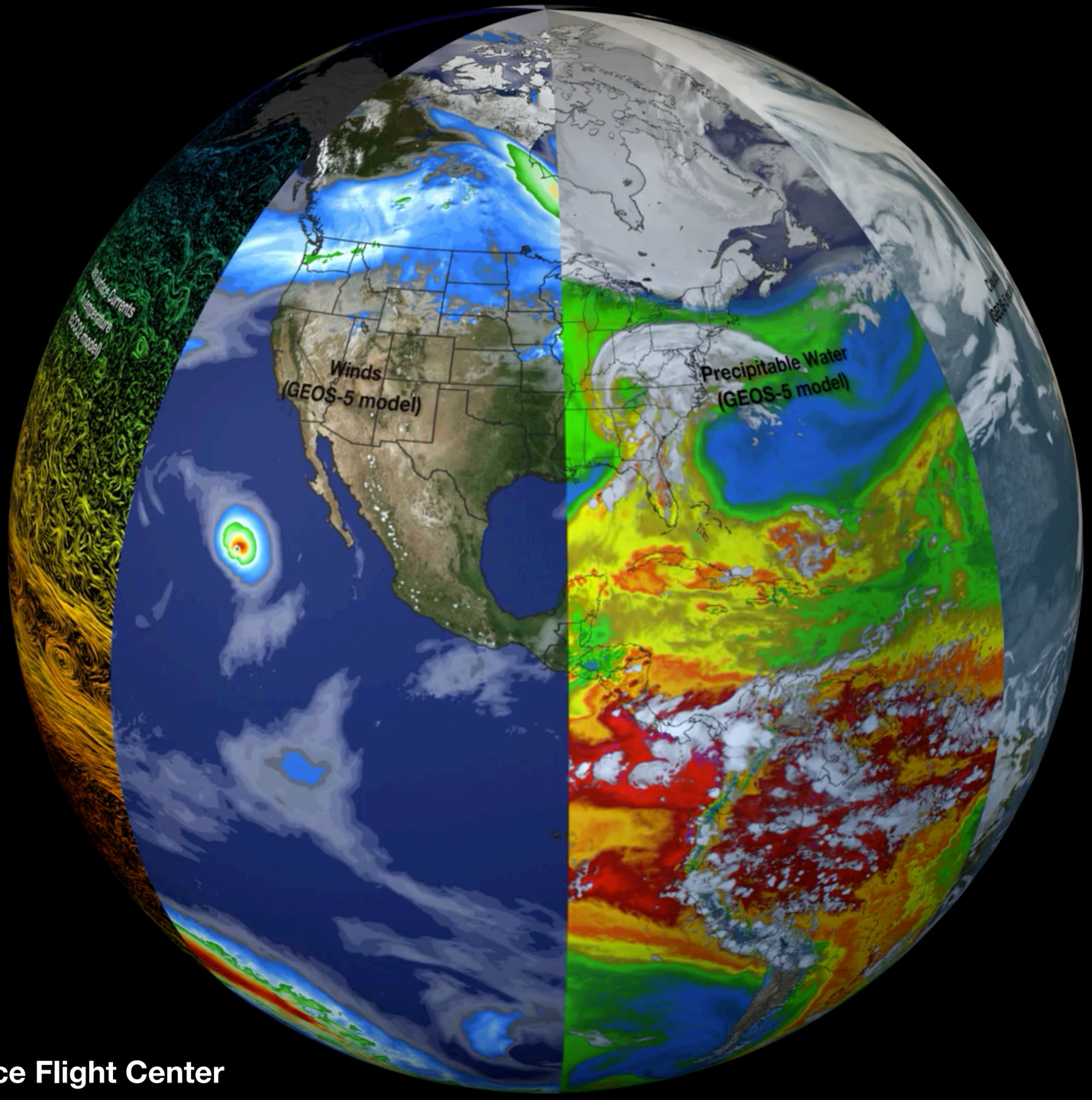
Core developer of *Zarr*

Co-founder of *Pangeo*

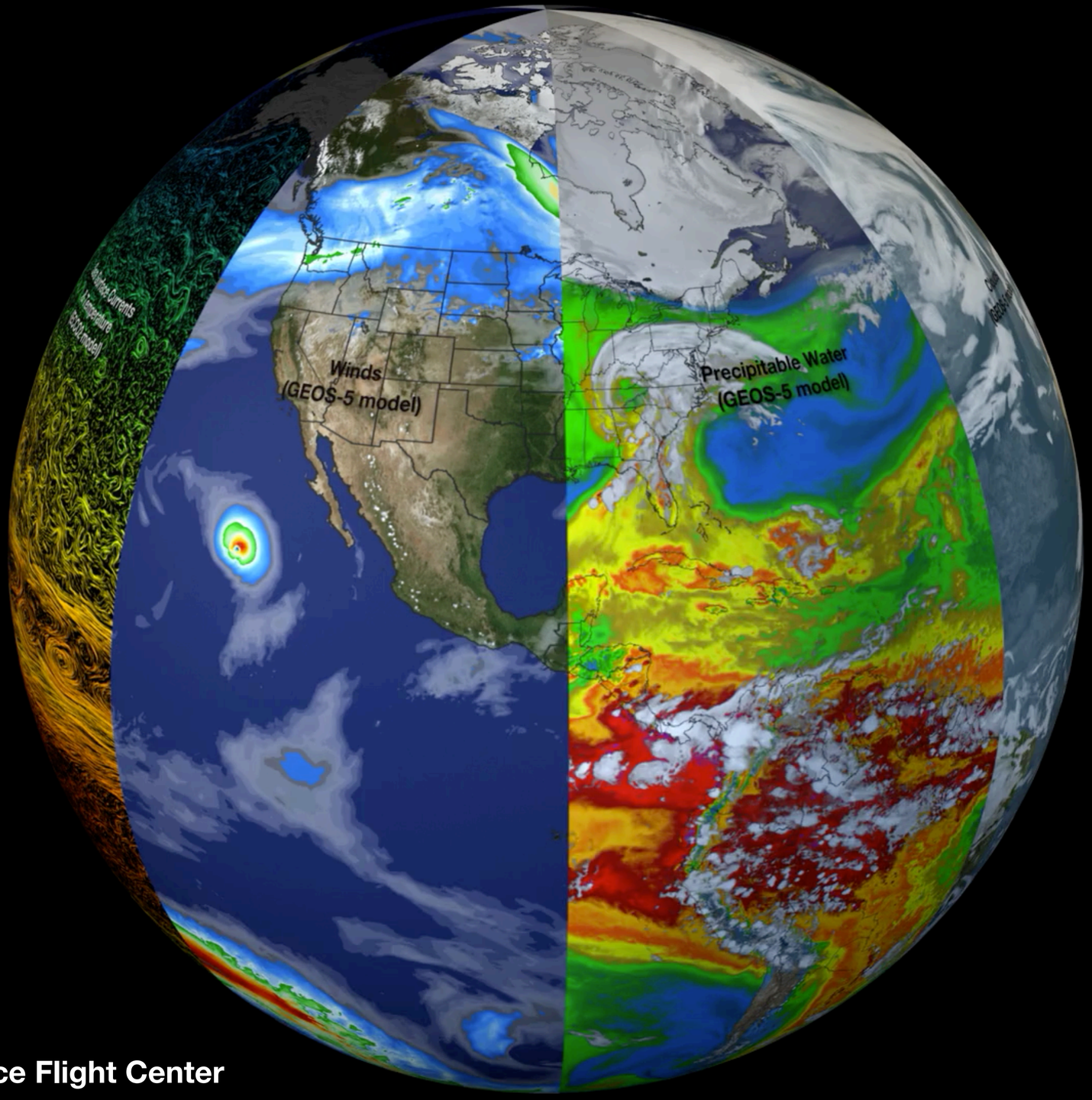
Open Source Advocate

THIS TALK

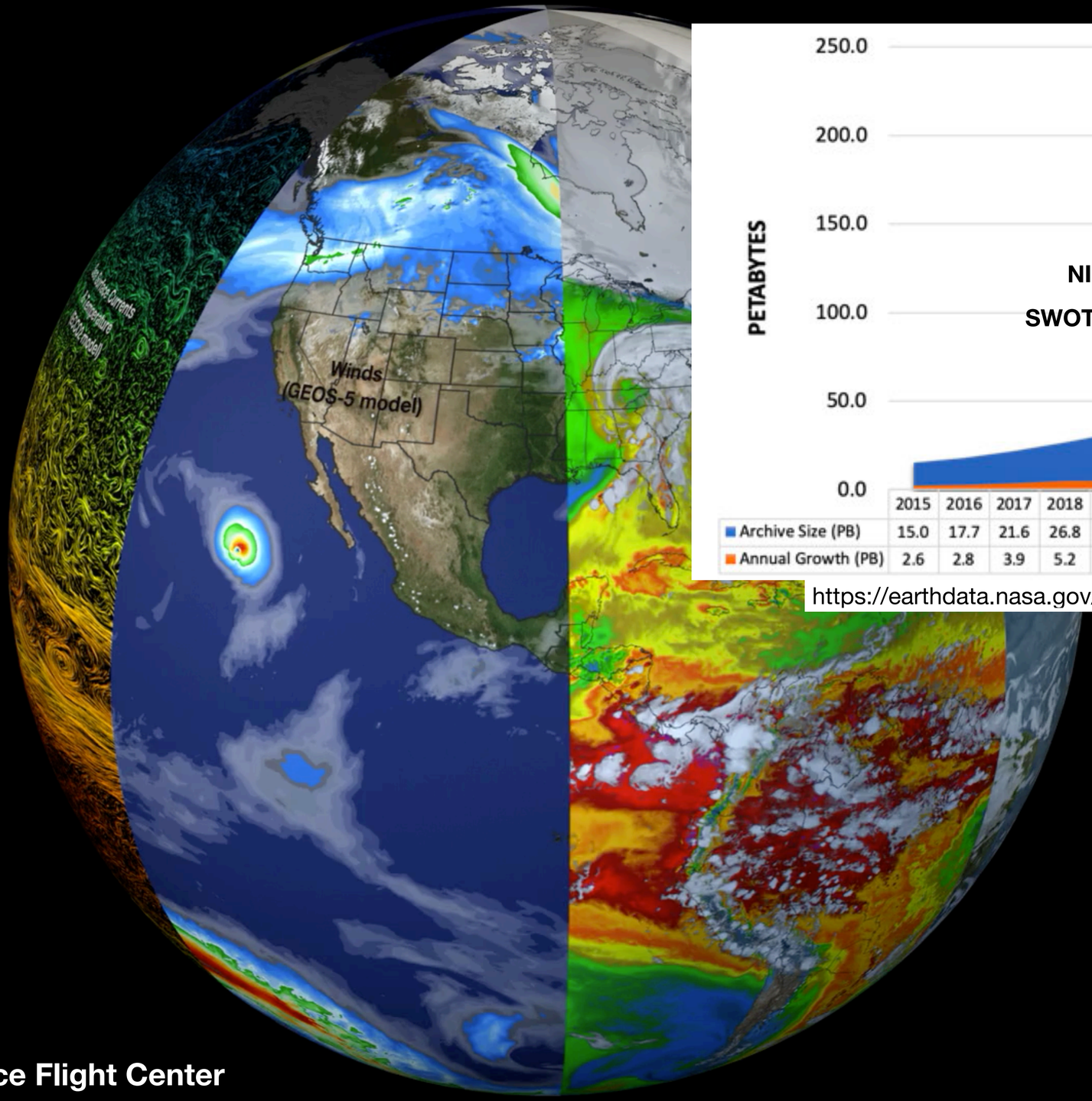
- *Problem:* Weather and climate data is large and complex 🤯
  - *Solution:* Data-proximate computing in the cloud 😎
- *Problem:* Analysis-ready, cloud-optimized data is scarce 🤔
  - *Solution:* **Pangeo Forge** 💪

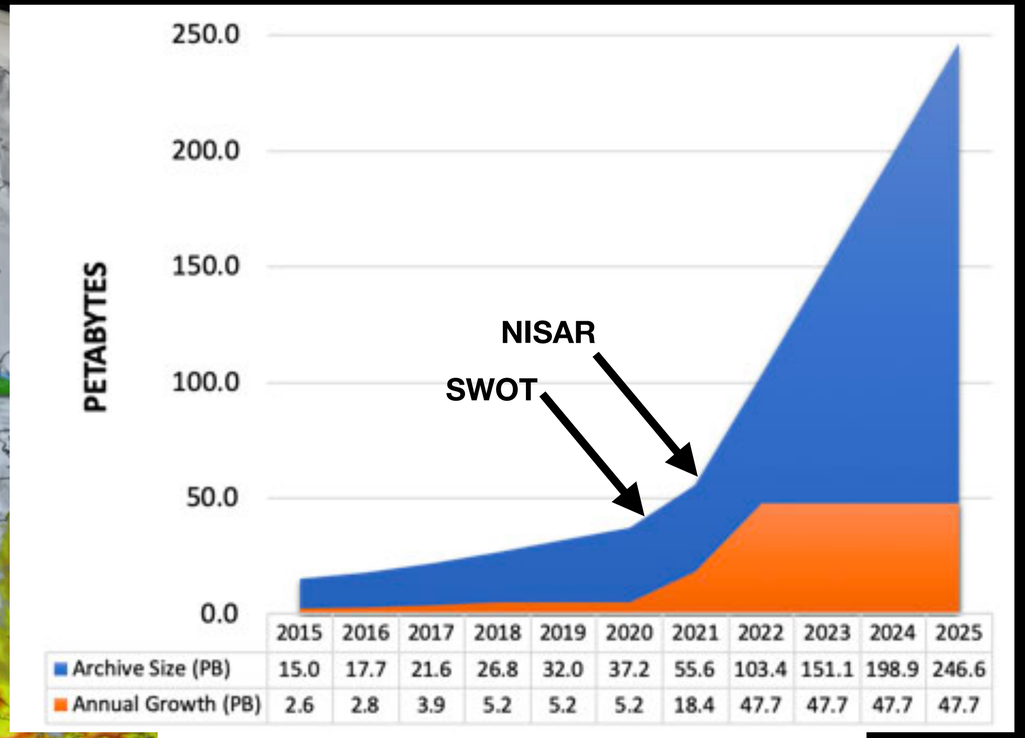


Credit: NASA's Goddard Space Flight Center



Credit: NASA's Goddard Space Flight Center





	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
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
Annual Growth (PB)	2.6	2.8	3.9	5.2	5.2	5.2	18.4	47.7	47.7	47.7	47.7

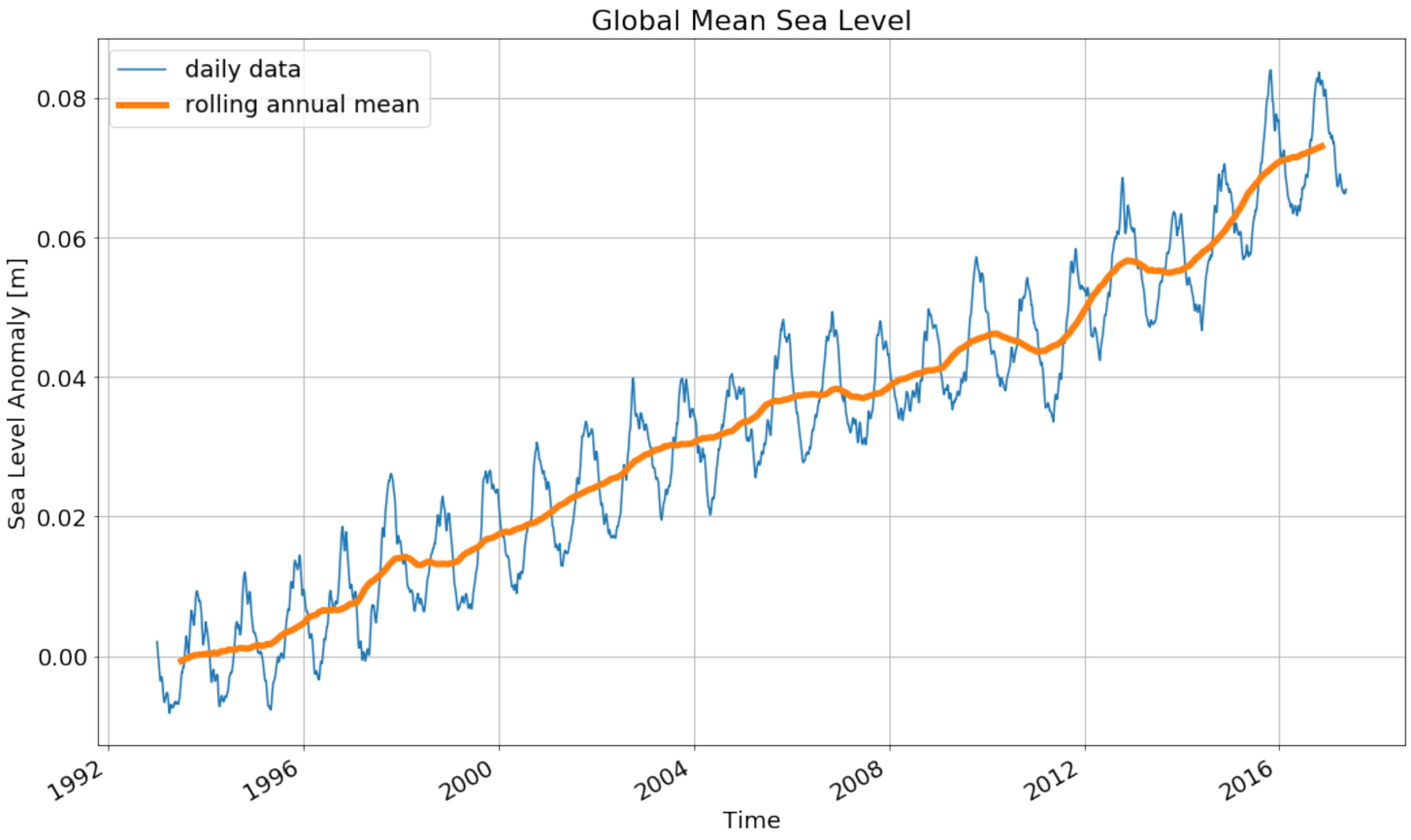
https://earthdata.nasa.gov/eosdis/cloud-evolution

Credit: NASA's Goddard Space Flight Center

WHAT SCIENCE DO WE WANT TO DO WITH ALL THIS DATA?

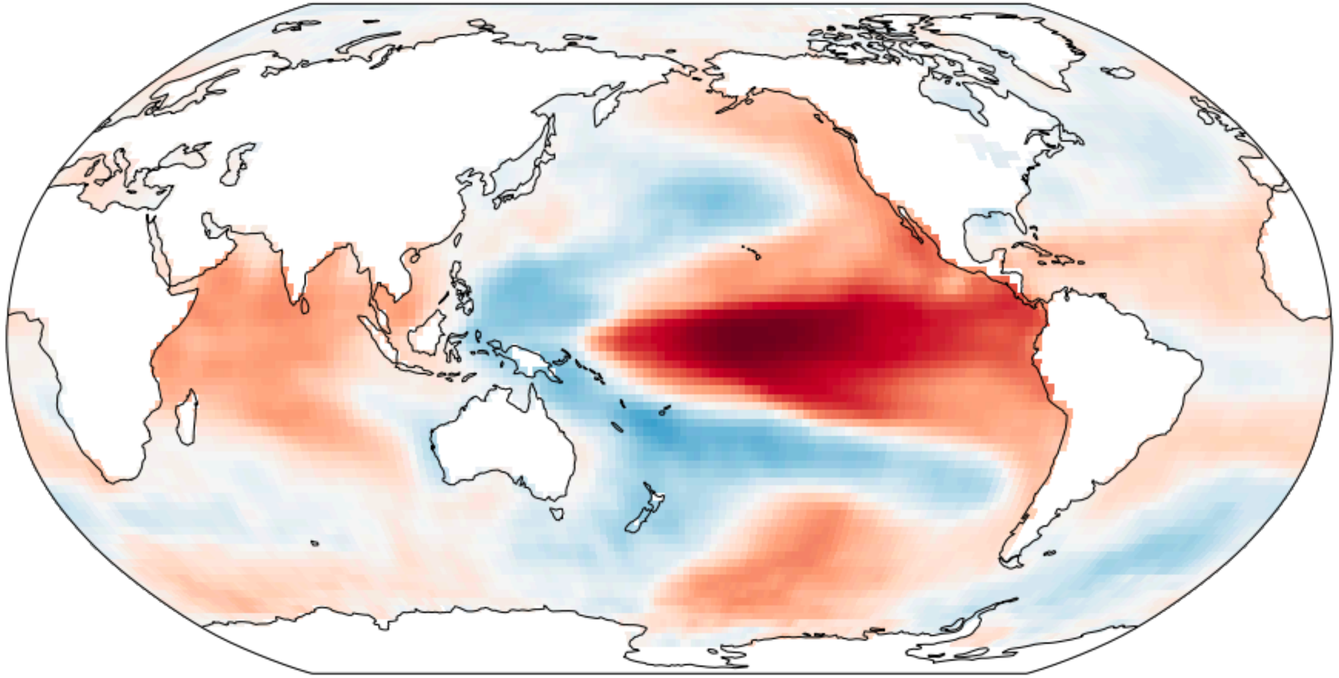
WHAT SCIENCE DO WE WANT TO DO WITH ALL THIS DATA?

Take the mean!



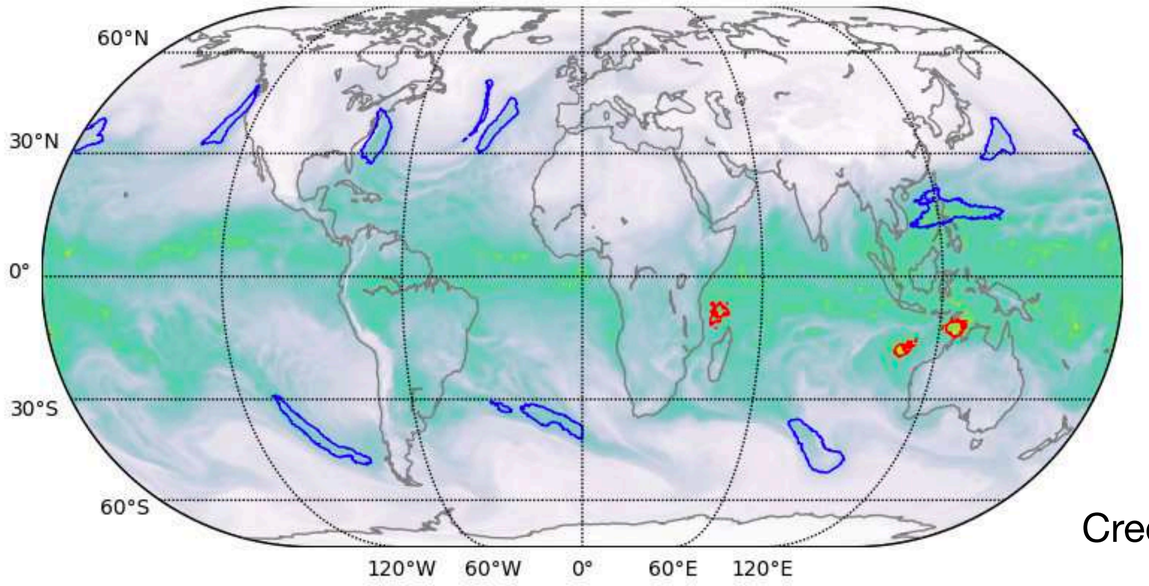
WHAT SCIENCE DO WE WANT TO DO WITH ALL THIS DATA?

Analyze spatiotemporal variability

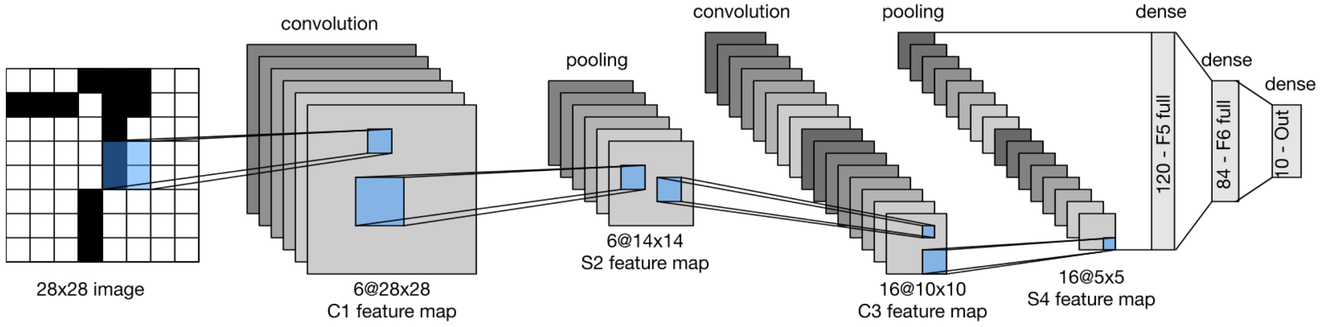


WHAT SCIENCE DO WE WANT TO DO WITH ALL THIS DATA?

Machine learning!

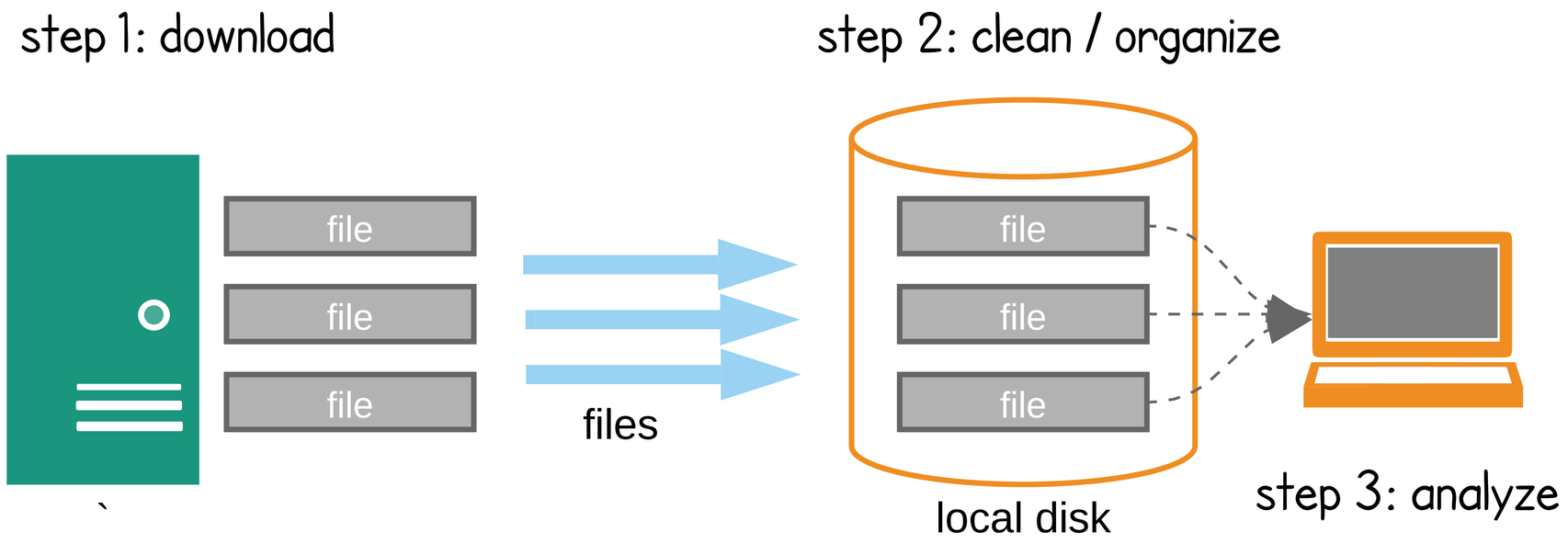


Credit: Berkeley Lab

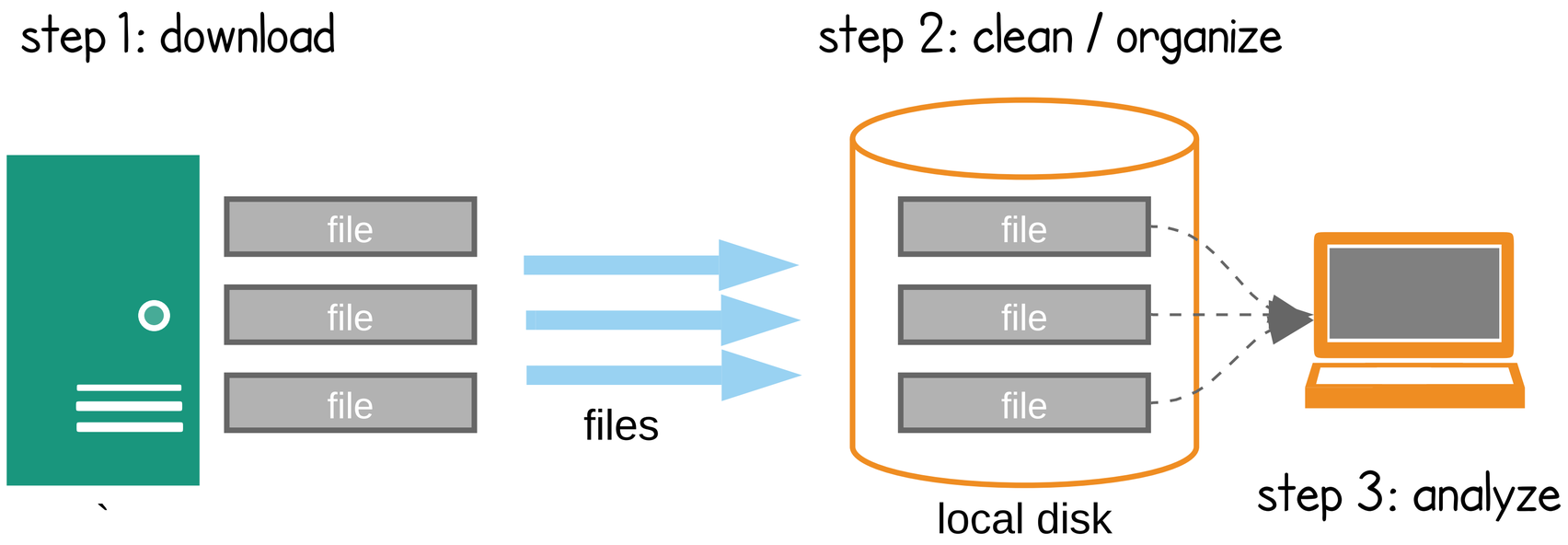


HOW?

THE "DOWNLOAD" MODEL

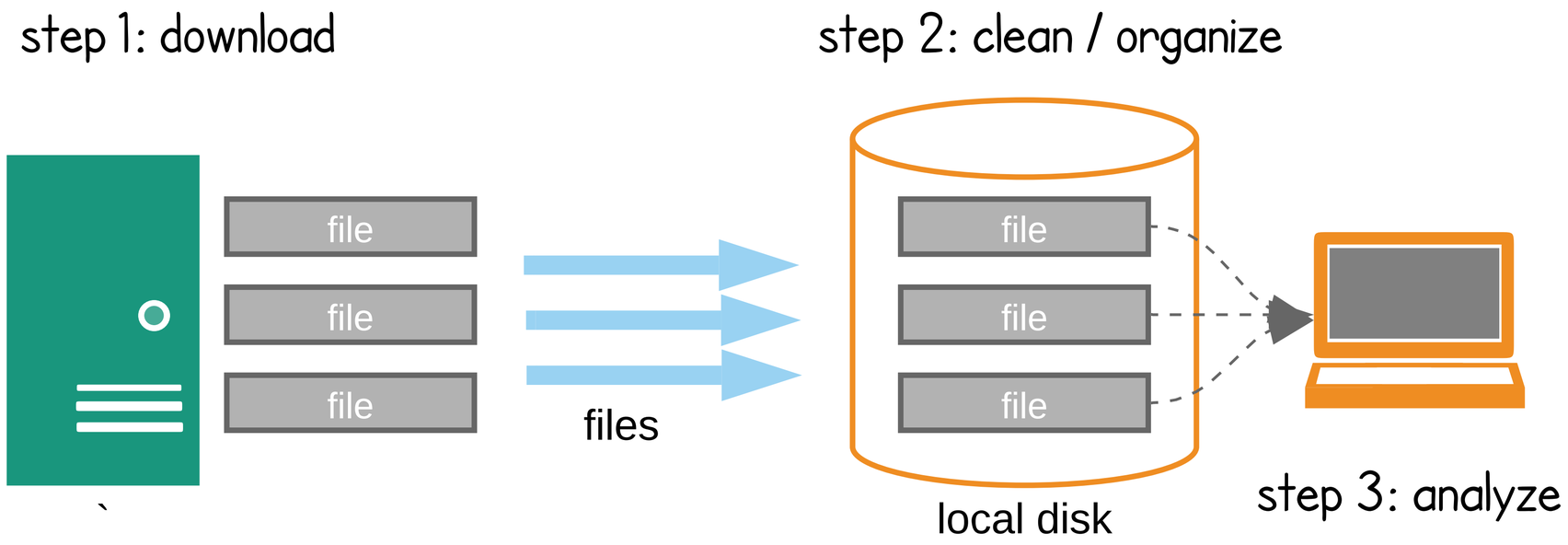


THE "DOWNLOAD" MODEL



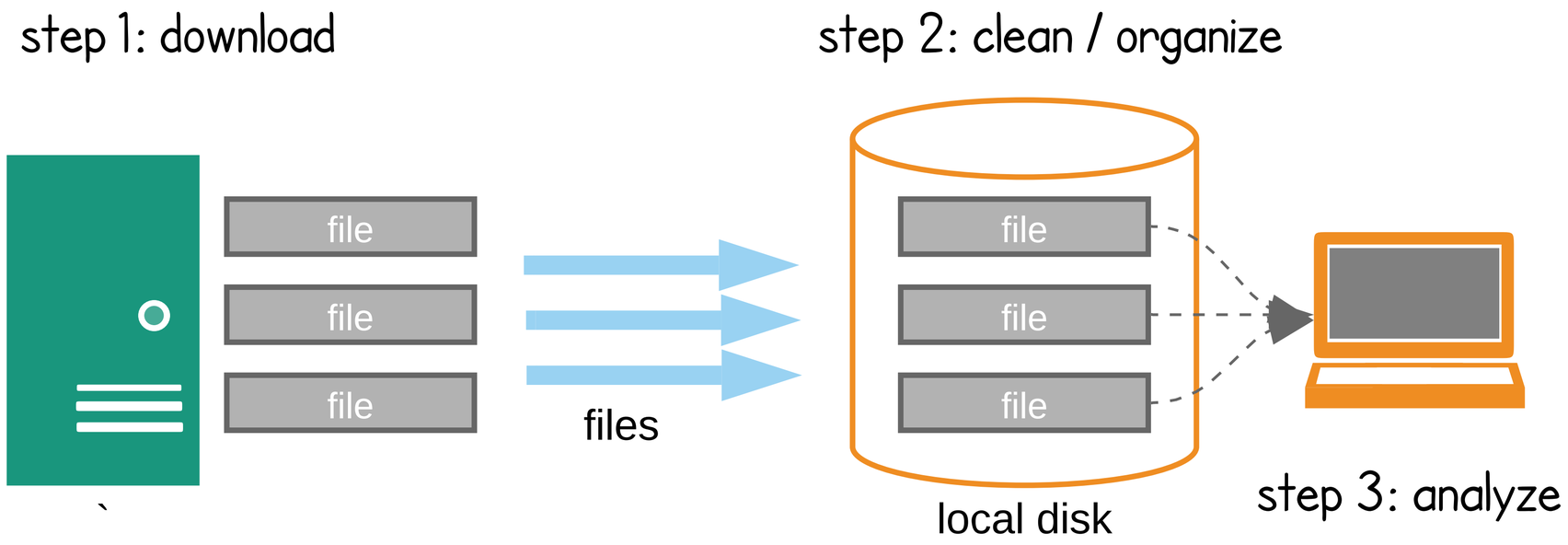
MB 

THE "DOWNLOAD" MODEL



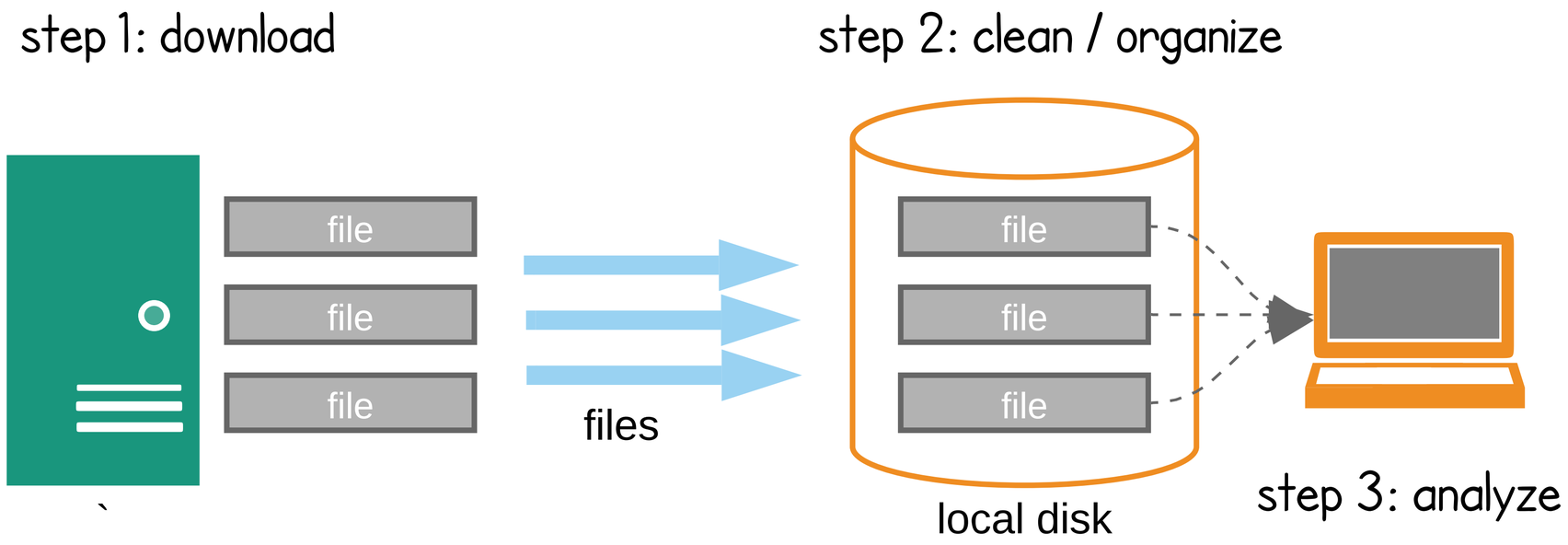
GB 😐

THE "DOWNLOAD" MODEL



TB 

THE "DOWNLOAD" MODEL



PB 😱

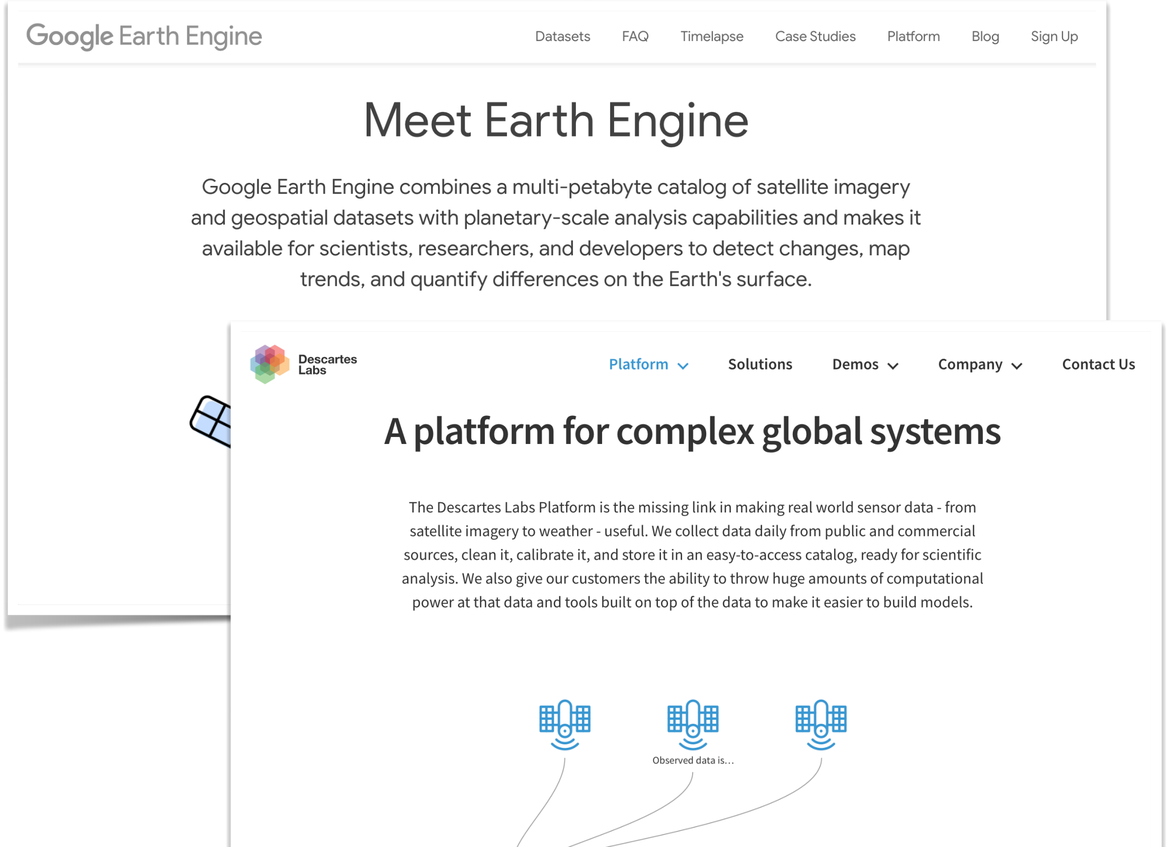
NEVER MIND…

Let’s “bring the compute to the data”!

HOW?

CLOUD-BASED ANALYSIS SOLUTIONS

Vertically Integrated Proprietary Platforms

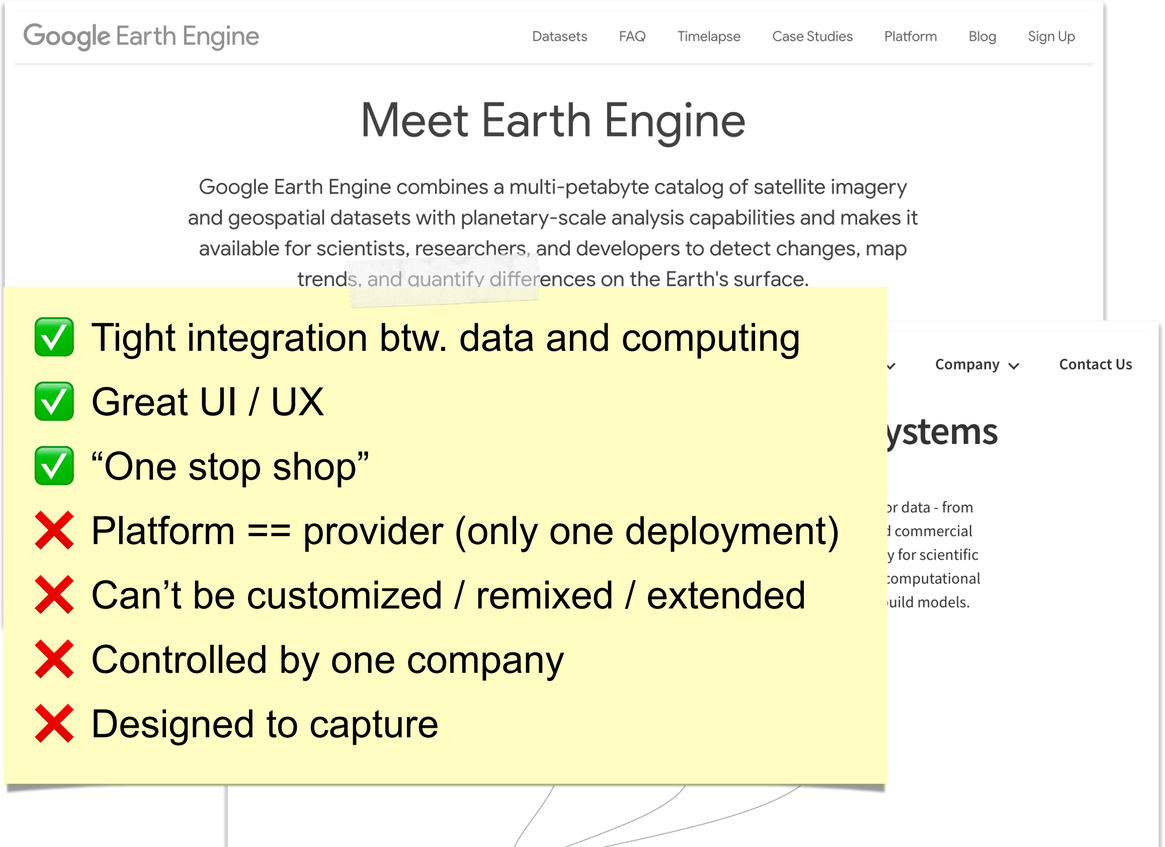


Open Source Modular Architecture

...etc.

CLOUD-BASED ANALYSIS SOLUTIONS

Vertically Integrated Proprietary Platforms



- ✅ Tight integration btw. data and computing
- ✅ Great UI / UX
- ✅ “One stop shop”
- ❌ Platform == provider (only one deployment)
- ❌ Can’t be customized / remixed / extended
- ❌ Controlled by one company
- ❌ Designed to capture

Open Source Modular Architecture

- ✅ Free software
- ✅ Community-driven development
- ✅ Deploy anywhere
- ✅ Interoperable
- ❌ Have to manage your own deployment
- ❌ Integration of data / computing can be hard
- ❌ UI / UX is less polished

PILLARS OF CLOUD NATIVE

1. Analysis-Ready, Cloud-Optimized Data

xarray.Dataset

▸ Dimensions: (**latitude**: 720, **longitude**: 1440, **nv**: 2, **time**: 8901)

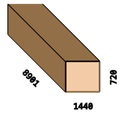
▾ Coordinates:

crs	()	int32	...
lat_bnds	(time, latitude, nv)	float32	dask.array<chunksize=(…
latitude	(latitude)	float32	-89.875 -89.625 ... 89.…
lon_bnds	(longitude, nv)	float32	dask.array<chunksize=(…
longitude	(longitude)	float32	0.125 0.375 ... 359.625…
nv	(nv)	int32	0 1
time	(time)	datetime64[ns]	1993-01-01 ... 2017-05…
axis :	T		
long_name :	Time		
standard_na…	time		

▾ Data variables:

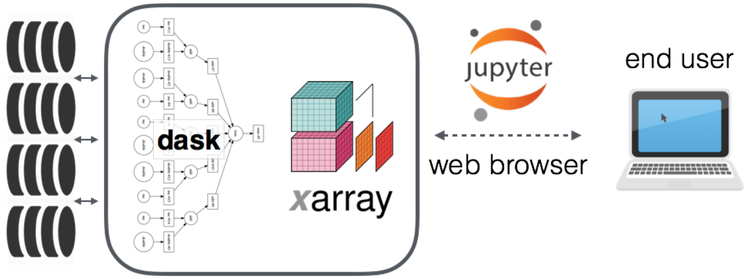
adt	(time, latitude, longitude)	float64	dask.array<chunksize=(…

	Array	Chunk
Bytes	73.83 GB	41.47 MB
Shape	(8901, 720, 1440)	(5, 720, 1440)
Count	1782 Tasks	1781 Chunks
Type	float64	numpy.ndarray

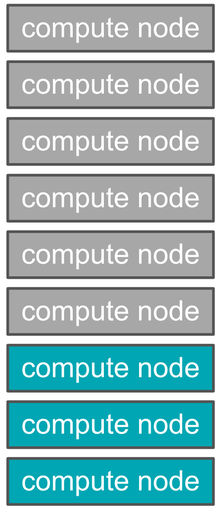


err	(time, latitude, longitude)	float64	dask.array<chunksize=(…
sla	(time, latitude, longitude)	float64	dask.array<chunksize=(…
ugos	(time, latitude, longitude)	float64	dask.array<chunksize=(…
ugosa	(time, latitude, longitude)	float64	dask.array<chunksize=(…
vgos	(time, latitude, longitude)	float64	dask.array<chunksize=(…
vgosa	(time, latitude, longitude)	float64	dask.array<chunksize=(…

2. Data-Proximate Computing



3. Elastic Distributed Processing



PANGEO CLOUD STACK

xarray — High-level API for analysis of multidimensional labelled arrays.

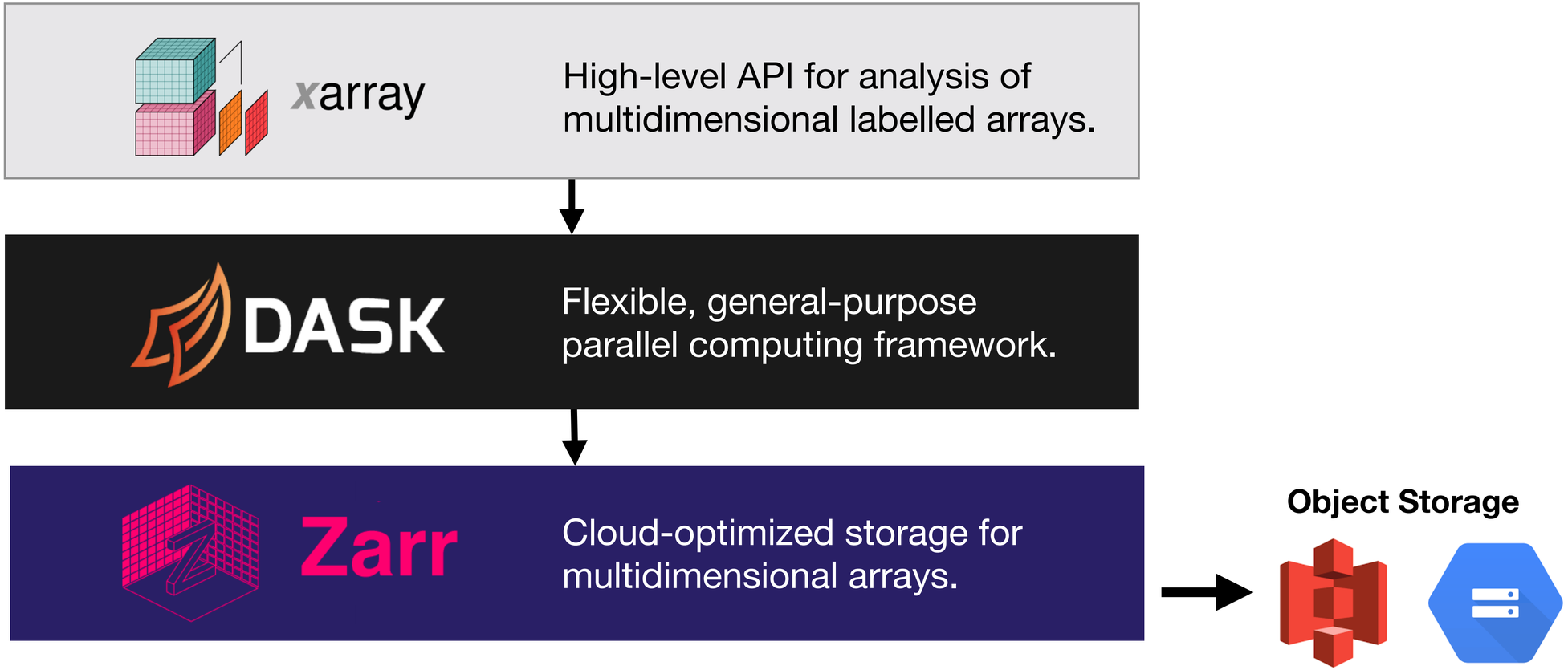
↓

DASK — Flexible, general-purpose parallel computing framework.

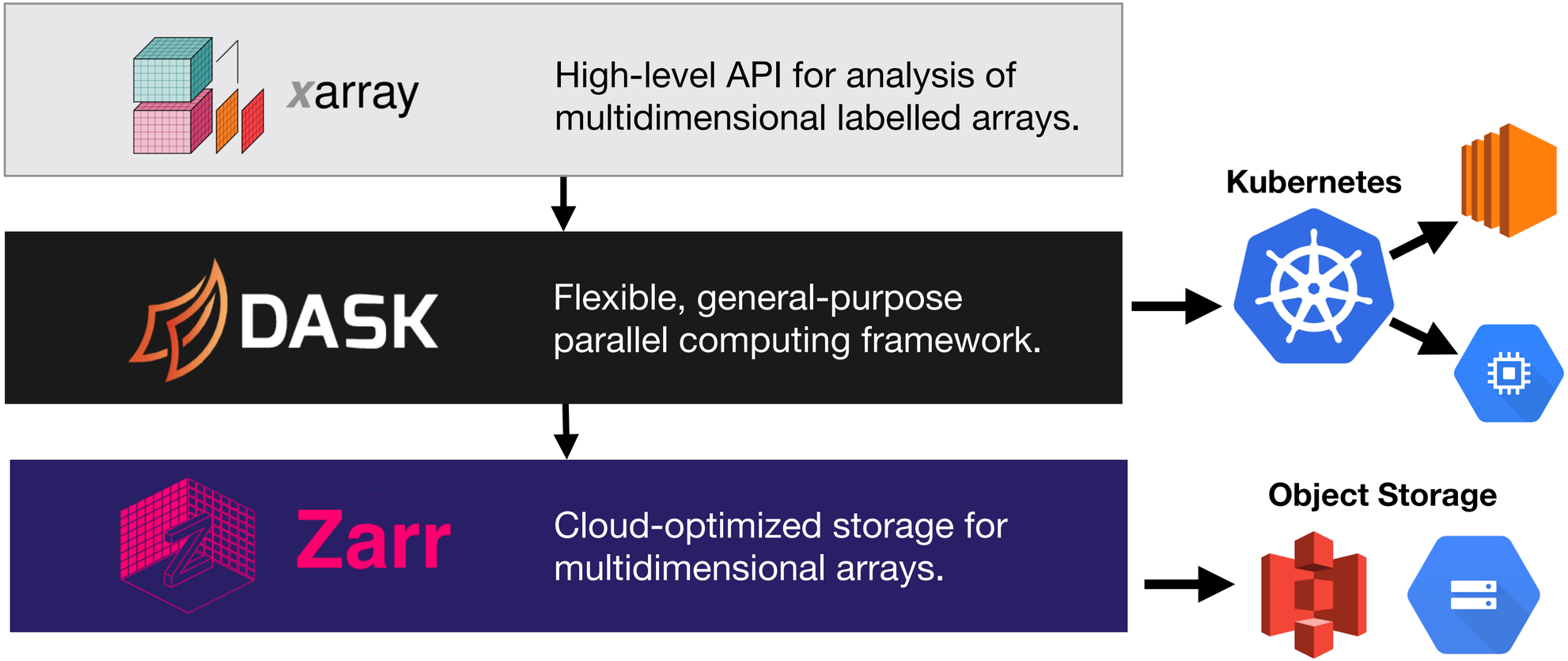
↓

Zarr — Cloud-optimized storage for multidimensional arrays.

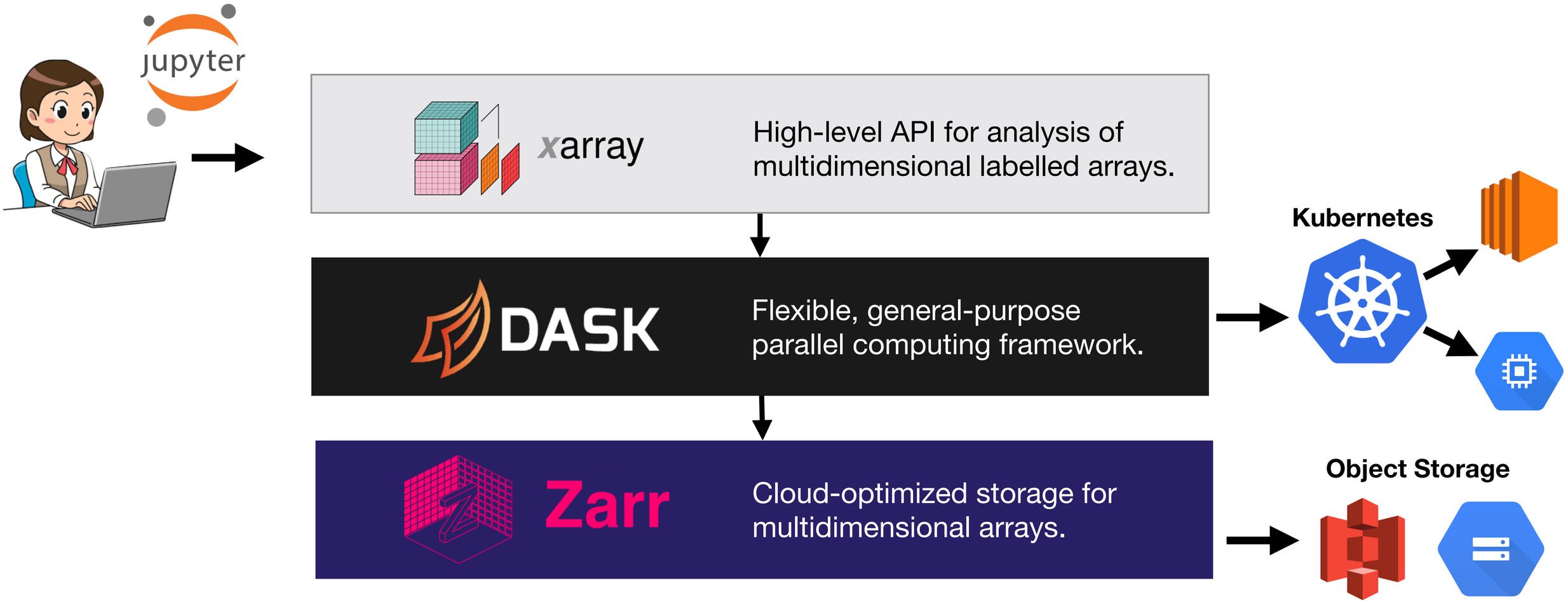
PANGEO CLOUD STACK



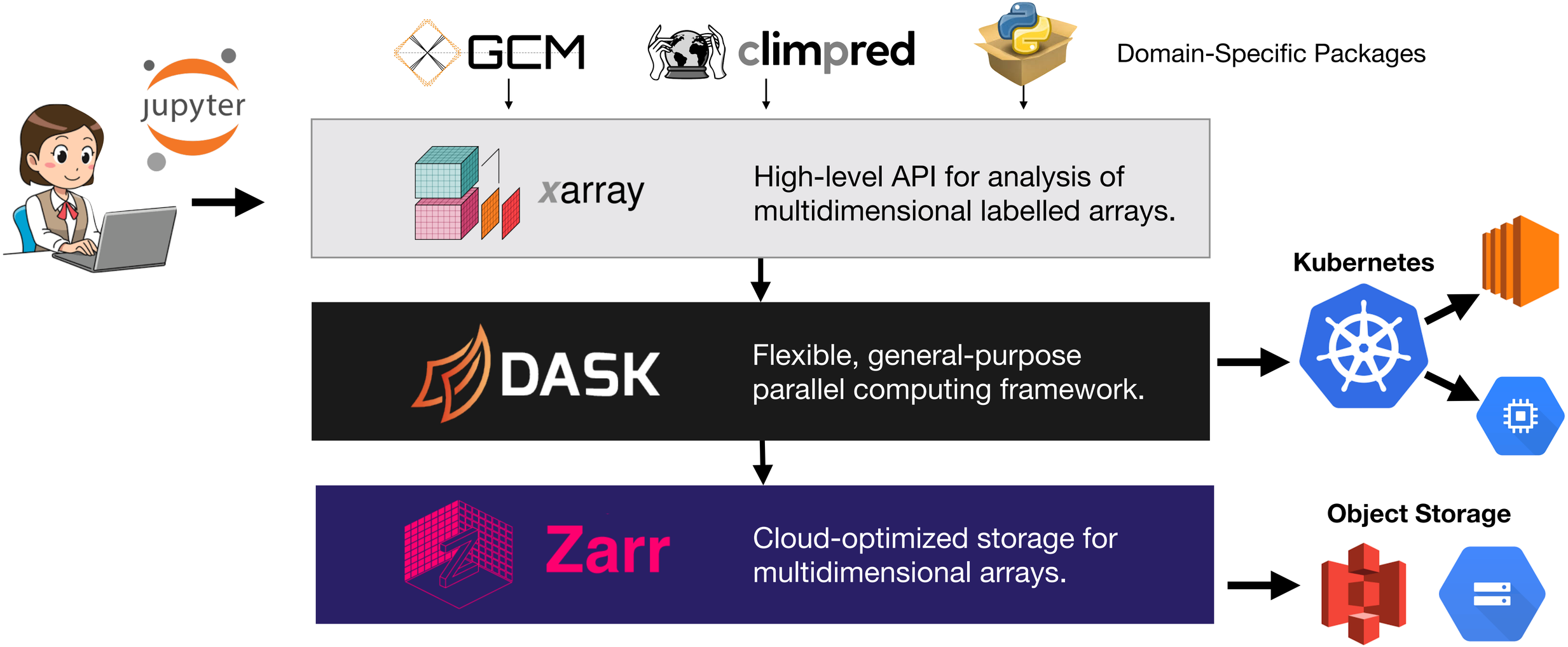
PANGEO CLOUD STACK



PANGEO CLOUD STACK

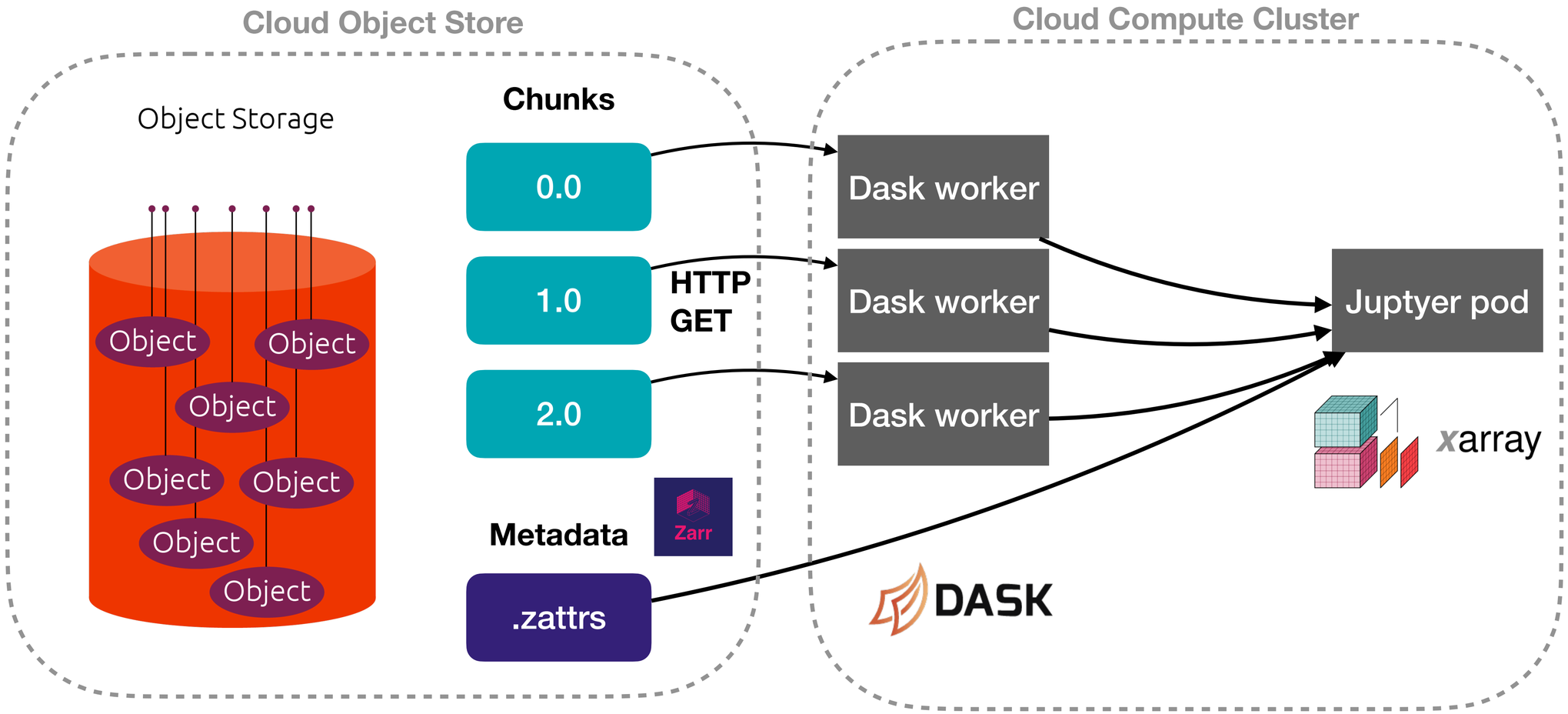


PANGEO CLOUD STACK



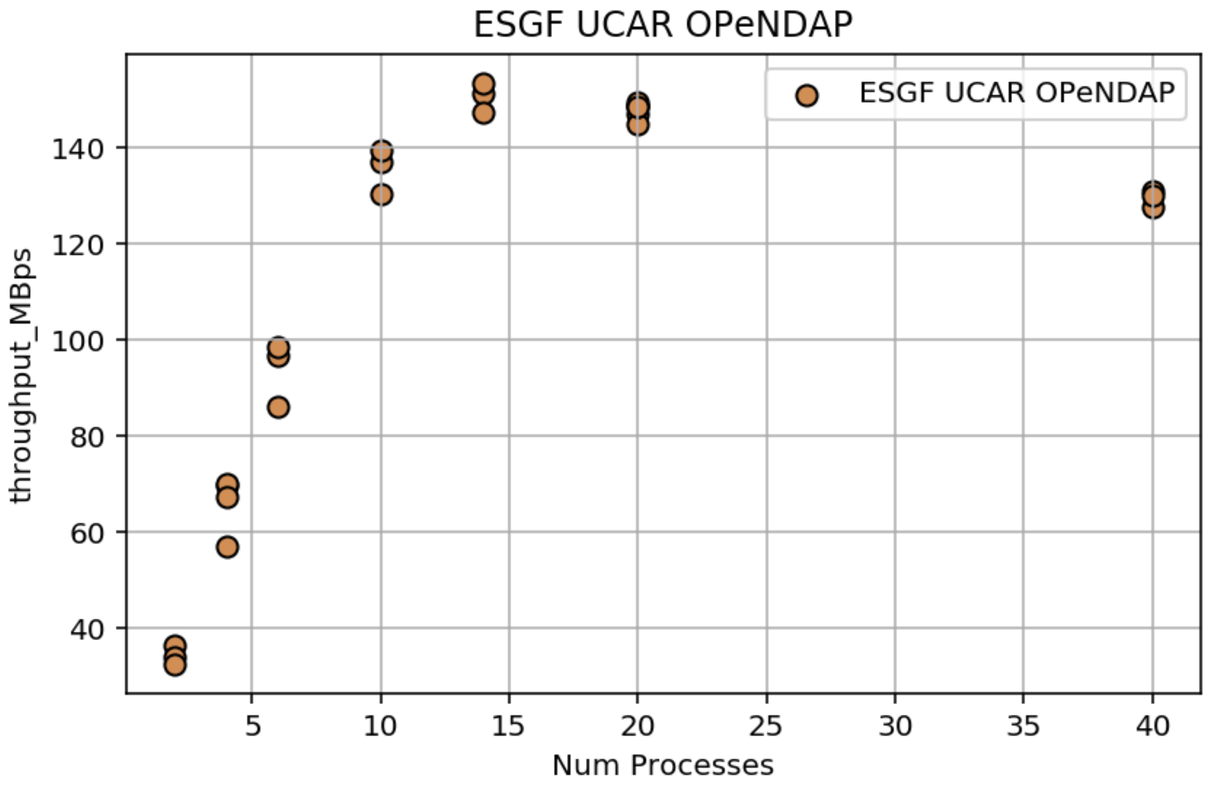
THE PANGEO CLOUD STACK

http://pangeo.io/cloud.html



CLOUD OPTIMIZED SCALES!

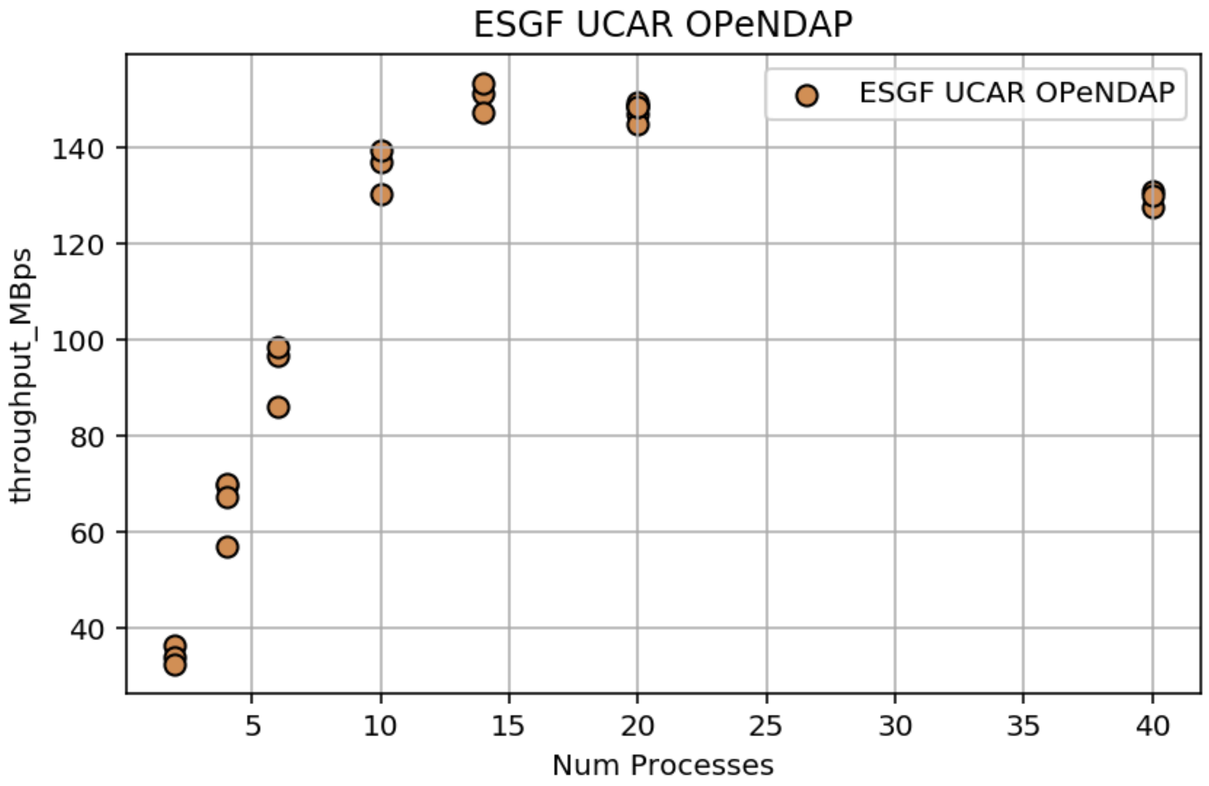
Legacy Server



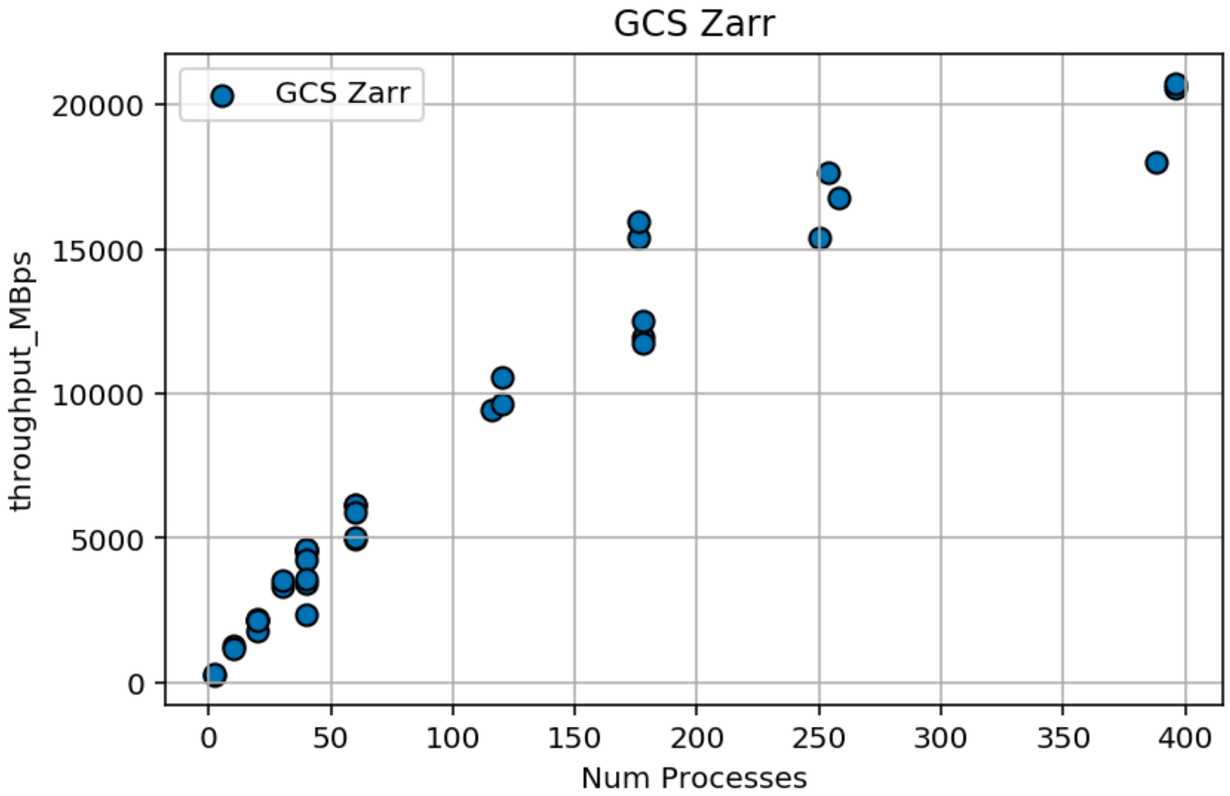
Xarray + Dask + Zarr

CLOUD OPTIMIZED SCALES!

Legacy Server



Xarray + Dask + Zarr

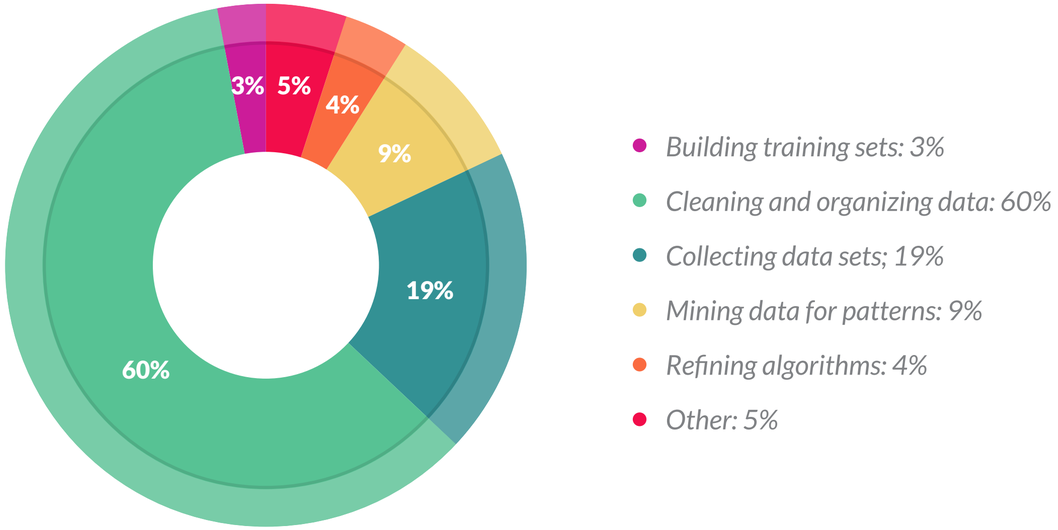


ARCO DATA

Analysis **R**eady, **C**loud **O**ptimzed

What is "Analysis Ready"?

- Think in "Datasets" not "data files"
- No need for tedious homogenizing / cleaning steps
- Curated and cataloged



How do data scientists spend their time?
Crowdflower Data Science Report (2016)

PANGEO

EXAMPLE OF ARCO DATA

Everything in one dataset object

Rich metadata

xarray.Dataset

▸ Dimensions: (**latitude**: 720, **longitude**: 1440, **nv**: 2, **time**: 8901)

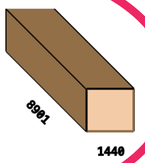
▾ Coordinates:

crs	()	int32	...
lat_bnds	(time, latitude, nv)	float32	dask.array<chunksize=(…
latitude	(latitude)	float32	-89.875 -89.625 ... 89.…
lon_bnds	(longitude, nv)	float32	dask.array<chunksize=(…
longitude	(longitude)	float32	0.125 0.375 ... 359.625…
nv	(nv)	int32	0 1
time	(time)	datetime64[ns]	1993-01-01 ... 2017-05…
axis :	T		
long_name :	Time		
standard_na…	time		

▾ Data variables:

adt	(time, latitude, longitude)	float64	dask.array<chunksize=(…

	Array	Chunk
Bytes	73.83 GB	41.47 MB
Shape	(8901, 720, 1440)	(5, 720, 1440)
Count	1782 Tasks	1781 Chunks
Type	float64	numpy.ndarray



err	(time, latitude, longitude)	float64	dask.array<chunksize=(…
sla	(time, latitude, longitude)	float64	dask.array<chunksize=(…
ugos	(time, latitude, longitude)	float64	dask.array<chunksize=(…
ugosa	(time, latitude, longitude)	float64	dask.array<chunksize=(…
vgos	(time, latitude, longitude)	float64	dask.array<chunksize=(…
vgosa	(time, latitude, longitude)	float64	dask.array<chunksize=(…

Chunked appropriately for analysis

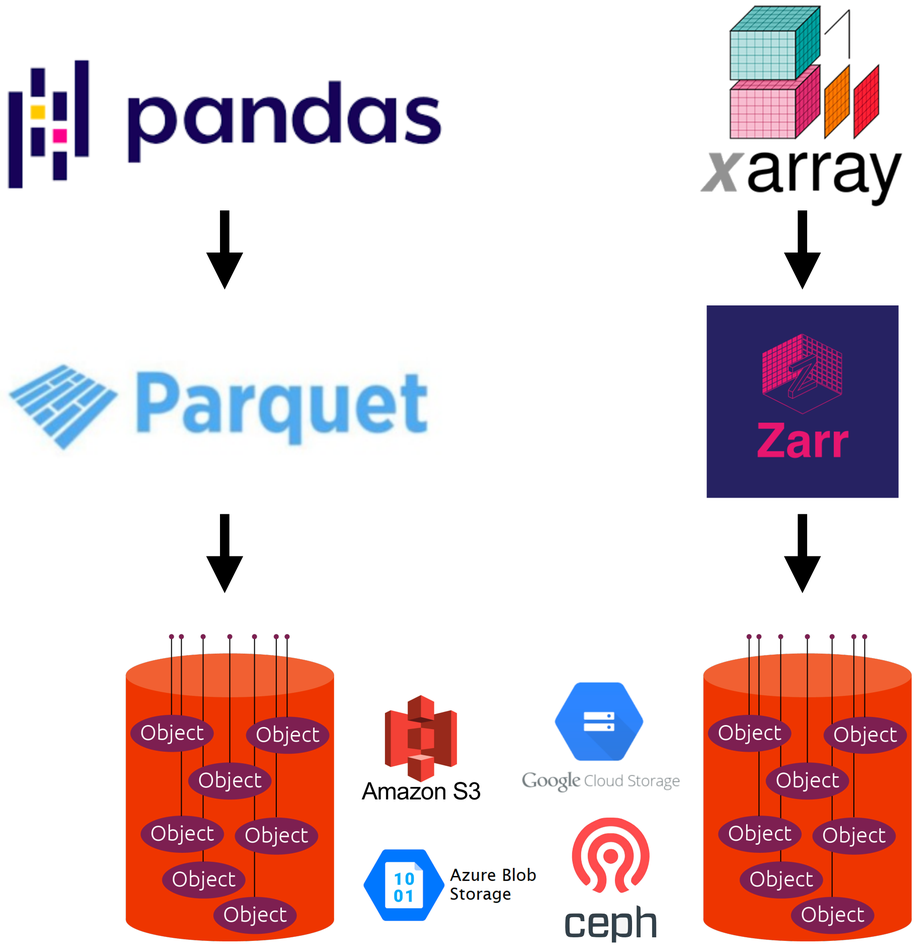
https://catalog.pangeo.io/browse/master/ocean/sea_surface_height/

ARCO DATA

Analysis **R**eady, **C**loud **O**ptimzed

What is "Cloud Optimized"?

- Compatible with object storage (access via HTTP)
- Supports lazy access and intelligent subsetting
- Integrates with high-level analysis libraries and distributed frameworks

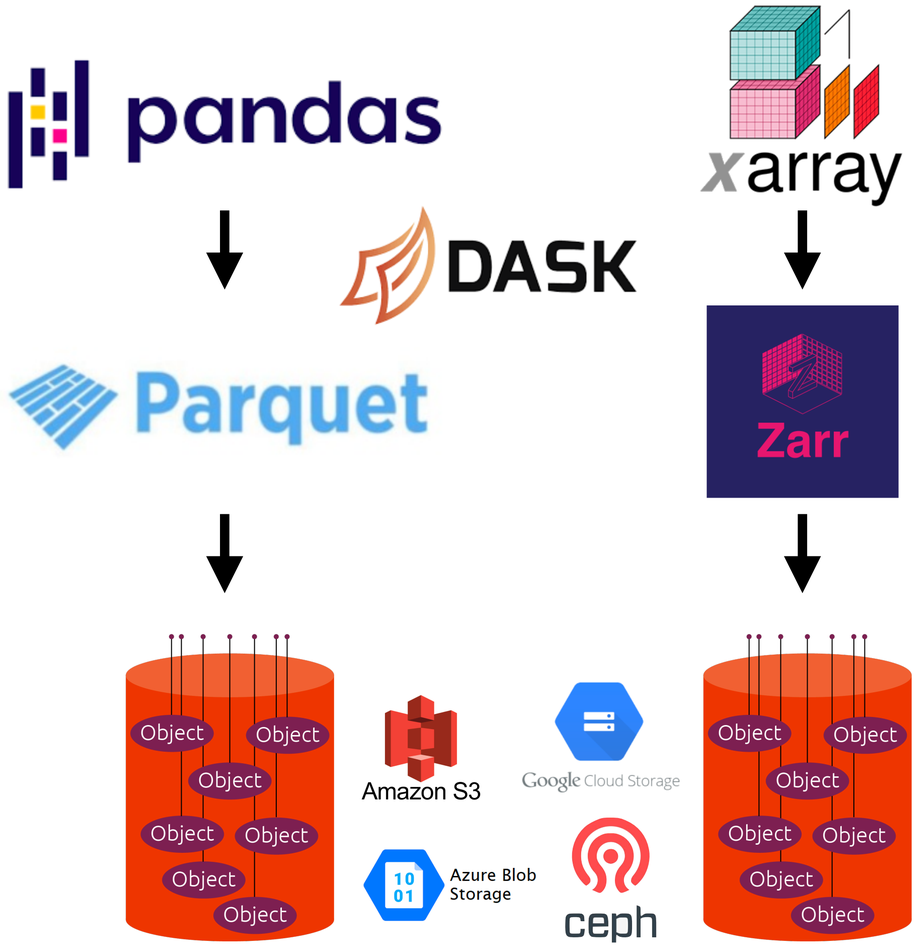


ARCO DATA

Analysis **R**eady, **C**loud **O**ptimzed

What is "Cloud Optimized"?

- Compatible with object storage (access via HTTP)
- Supports lazy access and intelligent subsetting
- Integrates with high-level analysis libraries and distributed frameworks



PANGEO CLOUD DATA CATALOG

CATALOG.PANGEO.IO

PANGEO CATALOG Blog Forum

PANGEO CATALOG

master

MASTER

Pangeo Master Data Catalog

```
https://raw.githubusercontent.com/pangeo-data/pangeo-datastore/master/intake-catalogs/master.yaml
```

Child Catalogs

ocean
Pangeo Oceanography Dataset Catalog

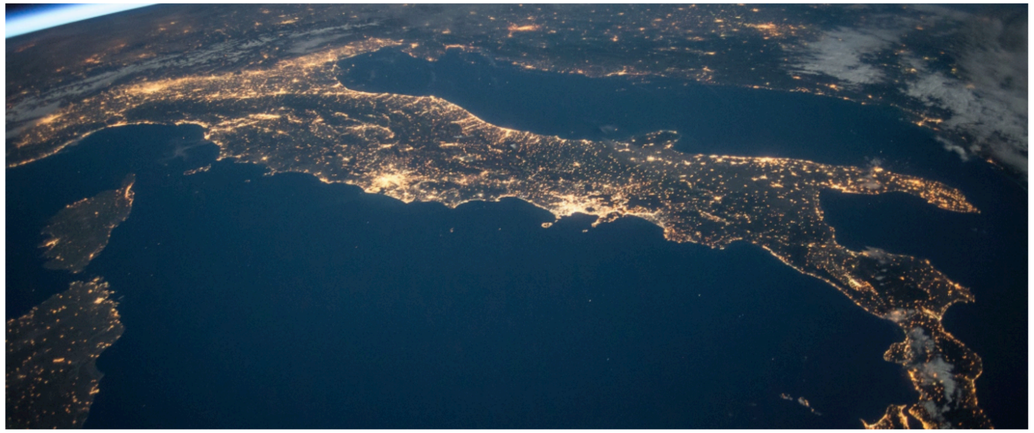
atmosphere
Pangeo Atmospheric Science Dataset Catalog

climate
Pangeo Climate Dataset Catalog. Include model ensembles such as CMIP6 and LENS.

hydro
Pangeo Hydrology Dataset Catalog

DATA ANALYTICS

New climate model data now in Google Public Datasets



Shane Glass
Program Manager, Google Cloud Public Dataset Program

December 9, 2019

Exploring public datasets is an important aspect of modern data analytics, and all this gathered data can help us understand our world. At Google Cloud, we maintain a collection of public datasets, and we're pleased to collaborate with the Lamont-Doherty Earth Observatory (LDEO) of Columbia University and the Pangeo Project to host the latest climate simulation data in the cloud.

PROBLEM:

Making ARCO Data is Hard!

To produce useful ARCO data, you must have:

Data Scientist

PROBLEM:

Making ARCO Data is Hard!

😩

To produce useful ARCO data, you must have:

Domain Expertise:
How to find, clean, and homogenize data

Data Scientist

PROBLEM:

Making ARCO Data is Hard!

To produce useful ARCO data, you must have:

Domain Expertise:
How to find, clean, and homogenize data

Tech Knowledge:
How to efficiently produce cloud-optimized formats

Data Scientist

PROBLEM:

Making ARCO Data is Hard!

To produce useful ARCO data, you must have:

Domain Expertise:
How to find, clean, and homogenize data

Tech Knowledge:
How to efficiently produce cloud-optimized formats

Compute Resources:
A place where to stage and upload the ARCO data

Data Scientist

PROBLEM:

Making ARCO Data is Hard!

To produce useful ARCO data, you must have:

Domain Expertise:
How to find, clean, and homogenize data

Tech Knowledge:
How to efficiently produce cloud-optimized formats

Compute Resources:
A place where to stage and upload the ARCO data

Communication Skills:
To explain to others how to use the data

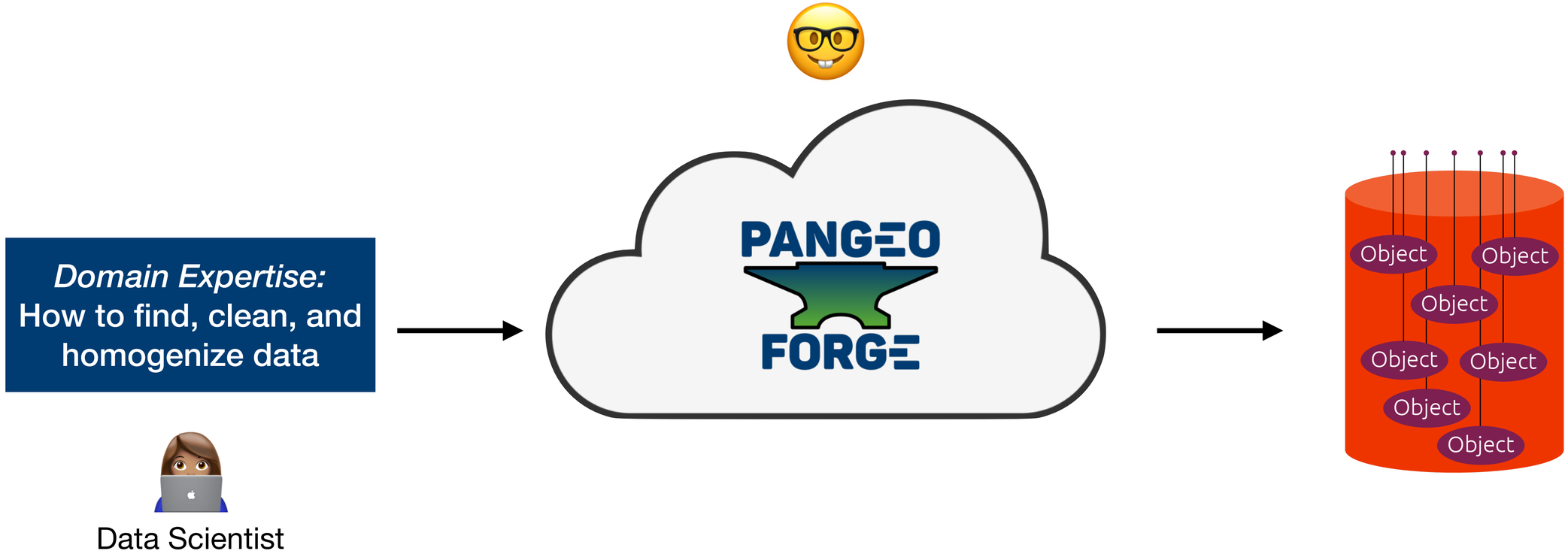
Data Scientist

PANGEO FORGE

Let’s democratize the production of ARCO data!

PANGEO FORGE

Let’s democratize the production of ARCO data!



PANGEO FORGE RECIPES

https://pangeo-forge.readthedocs.io/

Recipe defines:

- How to get the *inputs* (e.g. 14000 daily netCDF files)
- How to combine the inputs
- Target format (e.g. Zarr)

```
input_url_pattern = (
    "https://www.ncei.noaa.gov/data/sea-surface-temperature-optimum-interpola
    "/v2.1/access/avhrr/{yyyymm}/oisst-avhrr-v02r01.{yyyymmdd}.nc"
)
```

```
recipe = NetCDFtoZarrSequentialRecipe(
    input_urls=input_urls,
    sequence_dim="time",
    inputs_per_chunk=20
)
recipe
```

PANGEO

PANGEO FORGE BAKERIES

Recipes

https://github.com/pangeo-forge/

terraclimate-feedstock

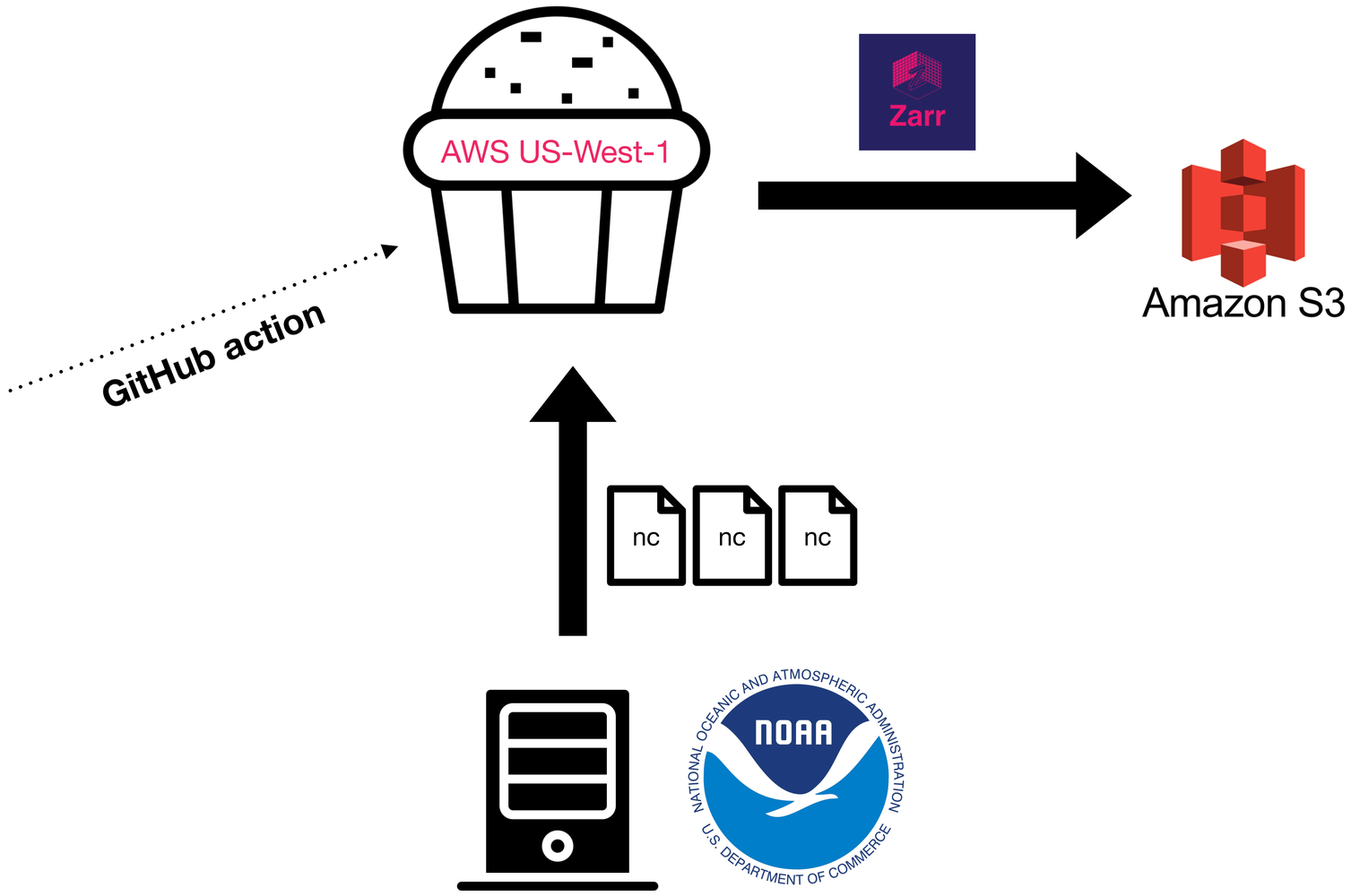
A pangeo-smithy repository for the terraclimate dataset.

Python Apache-2.0 3 2 1 3 Updated on Jan 1

noaa-oisst-avhrr-feedstock

Python Apache-2.0 2 1 0 4 Updated on Jan 1

GitHub



PANGEO FORGE BAKERIES

Recipes

https://github.com/pangeo-forge/

terraclimate-feedstock
A pangeo-smithy repository for the terraclimate dataset.
Python Apache-2.0 3 2 1 3 Updated on Jan 1

noaa-oisst-avhrr-feedstock
Python Apache-2.0 2 1 0 4 Updated on Jan 1

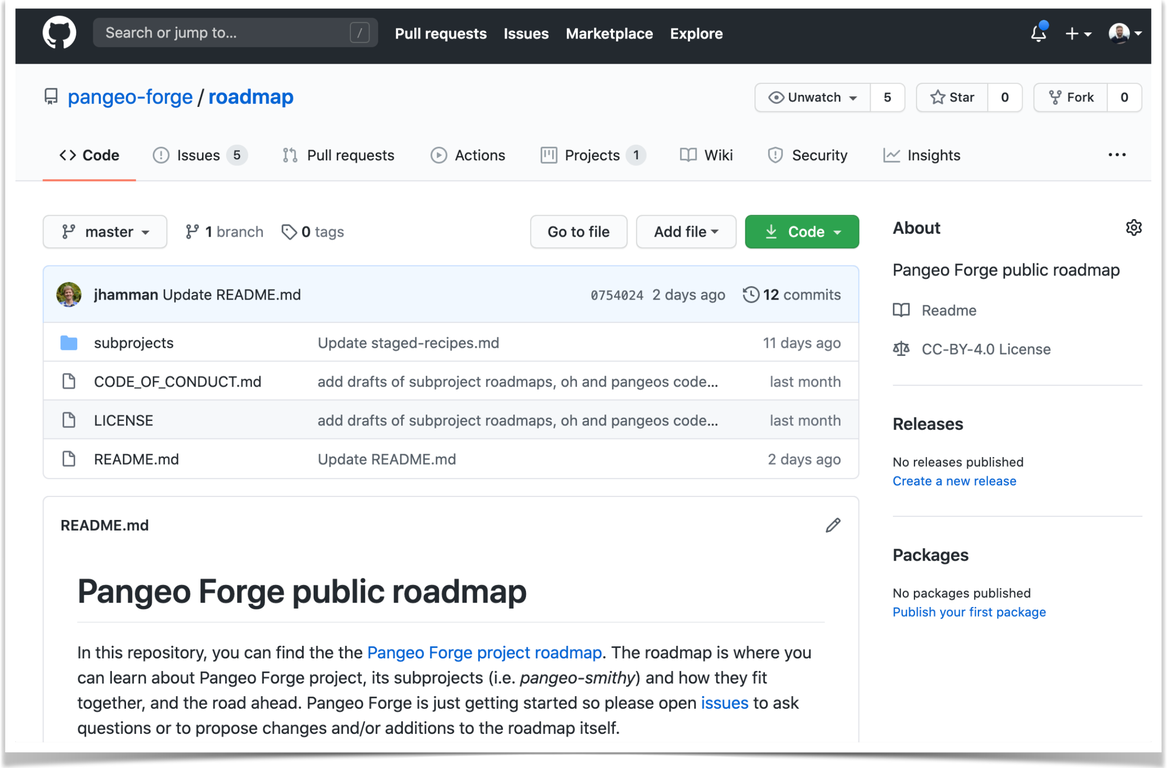
GitHub

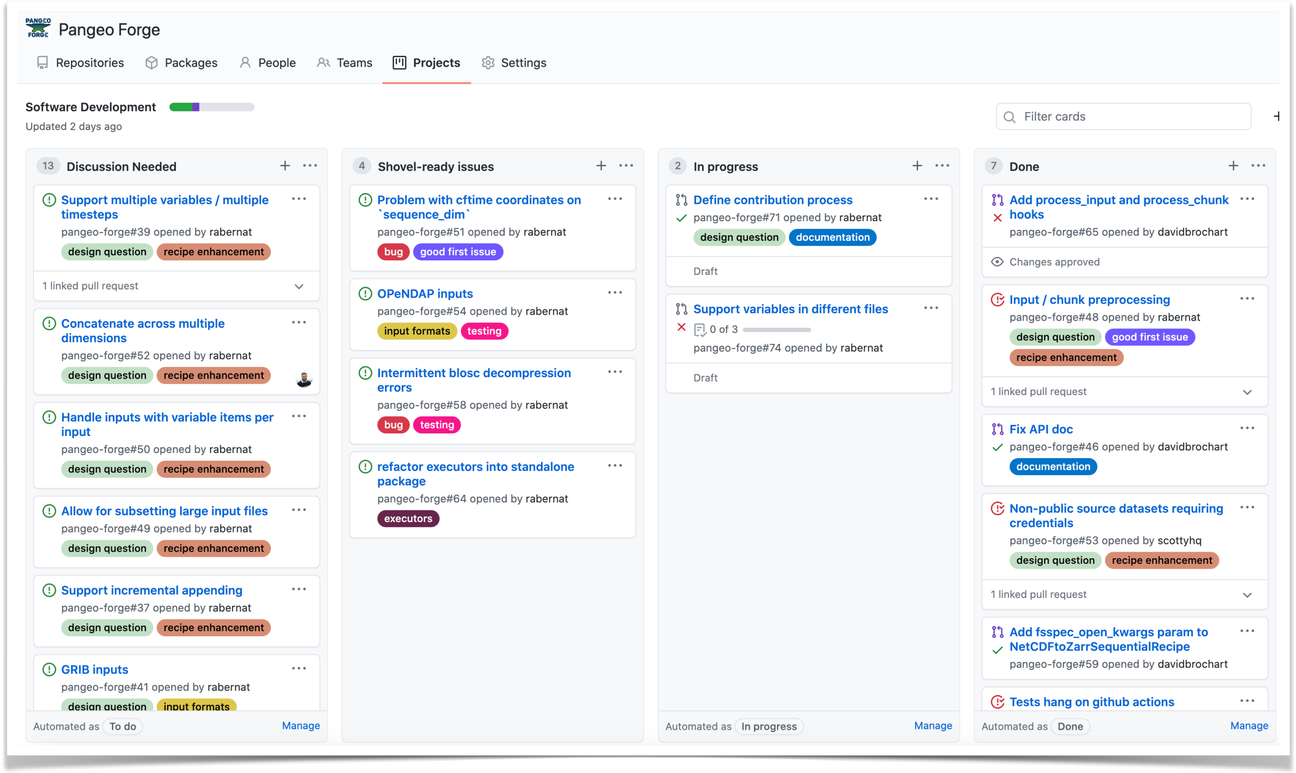
GitHub action

Bakeries are “Franchisable”
Will operate in a federation

PANGEO FORGE DEVELOPMENT

This is a 💯% open project!





https://github.com/pangeo-forge/roadmap

SUMMARY

- *Problem:* scientific data is large and complex 🤯
  - *Solution:* data-proximate computing in the cloud 😎
- *Problem:* analysis-ready, cloud-optimized data is scarce 🤔
  - *Solution:* **Pangeo Forge** 💪

We need your help!

https://github.com/pangeo-forge/roadmap

LEARN MORE

- "Cloud-Native Repositories for Big Scientific Data" Abernathey et al. 2021. *Computing in Science and Engineering* https://doi.org/10.22541/au.160443768.88917719/v2

- "Opening new horizons: How to migrate the Copernicus Global Land Service to a Cloud environment". Abernathey et al., 2021. *Publications Office of the European Union.* http://dx.doi.org/10.2760/668980

- "Closed Platforms vs. Open Architectures for Cloud-Native Earth System Analytics". Abernathey & Hamman, 2020. https://medium.com/pangeo/closed-platforms-vs-open-architectures-for-cloud-native-earth-system-analytics-1ad88708ebb6