Community-driven cloud initiatives leveraging FAIR principles

Aparna Radhakrishnan

Session: Why Analysis-Ready-Cloud-Optimized (ARCO) Data for Scalable Cloud Computing and Data Analytics to Support Open Science?

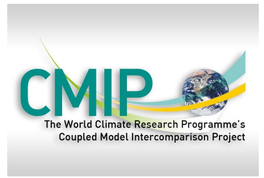
NOAA Environmental Data Management Workshop 2022

Community-driven effort

Special thanks to Pangeo/ESGF Cloud Data Working Group

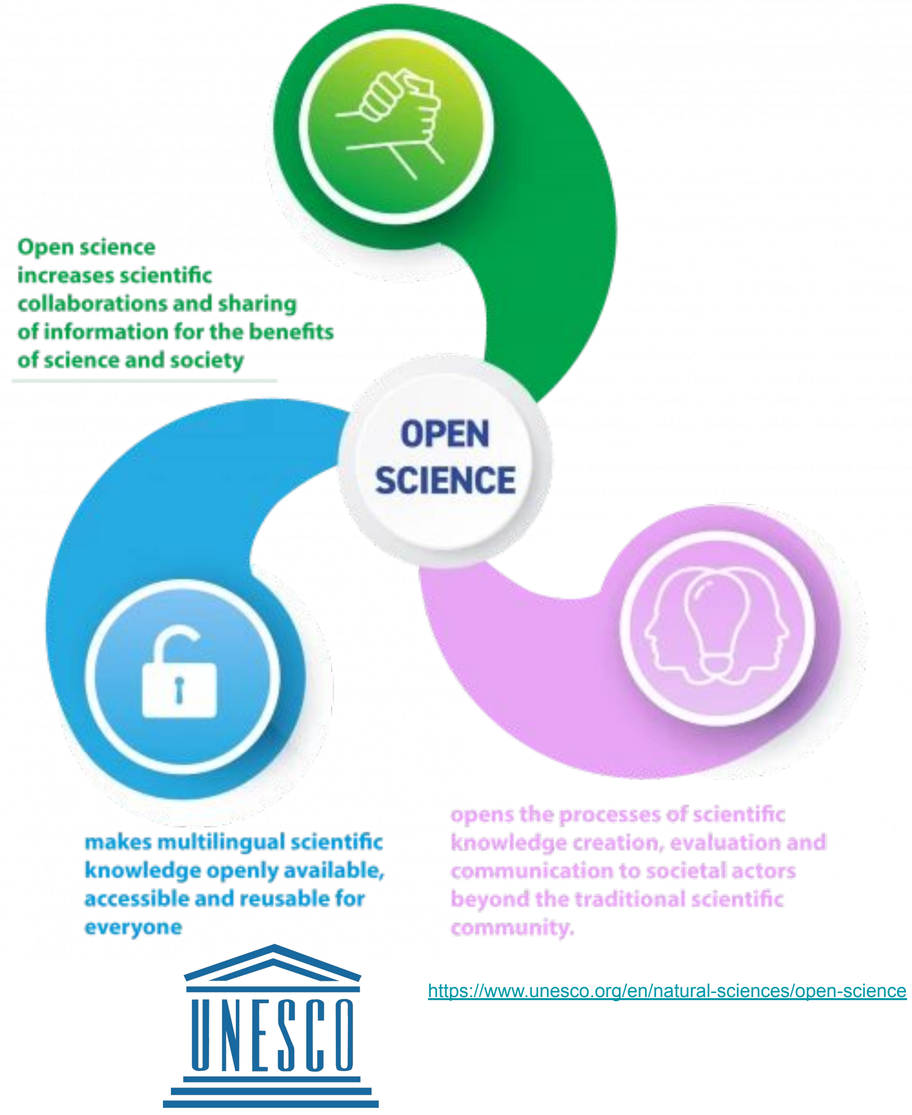
Content contributors: Ryan Abernathey, Philip Kershaw, Sasha Ames

Other contributors: V. Balaji, Julius Busecke, Ag Stephens, Serguei Nikonov, Kristopher Rand, Charles Stern, Hans Vahlenkamp, Sasha Ames , Ana Privette , Naomi Naik, Mackenzie Blanusa, Chris Blanton, Nkeh Perry Boh, Richard Smith, Rhys Evans, Zac Flamig, Diana Gergel, Thomas Jackson, Rebecca Monge, Natalie O'Leary, Zouberou Sayibou.



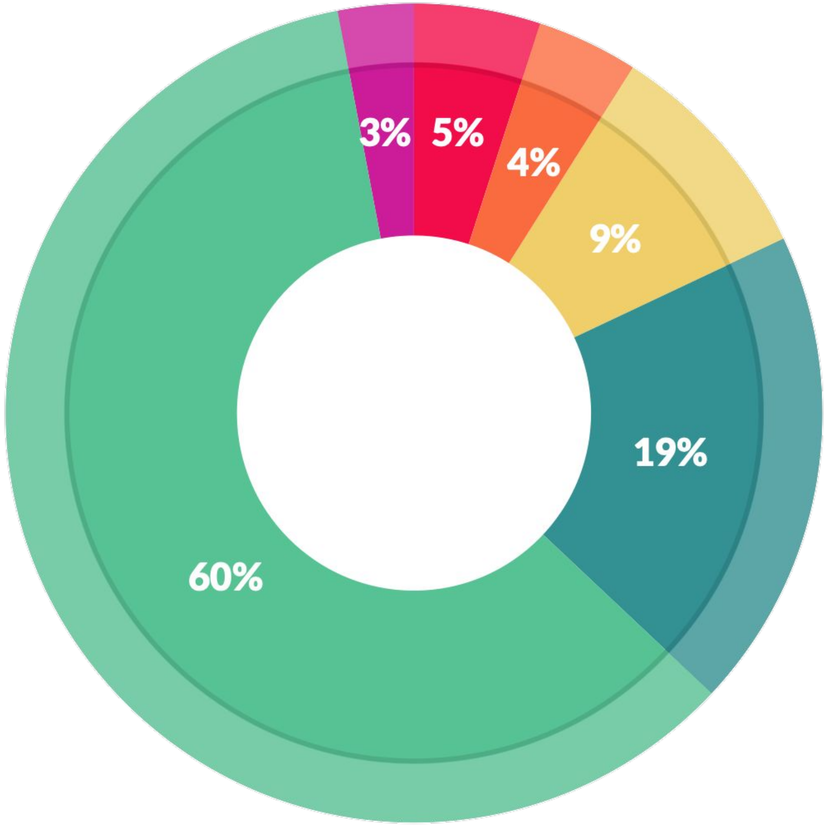
Open Science

Image credit



https://www.unesco.org/en/natural-sciences/open-science

Need Analysis-Ready datasets (ARD)



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

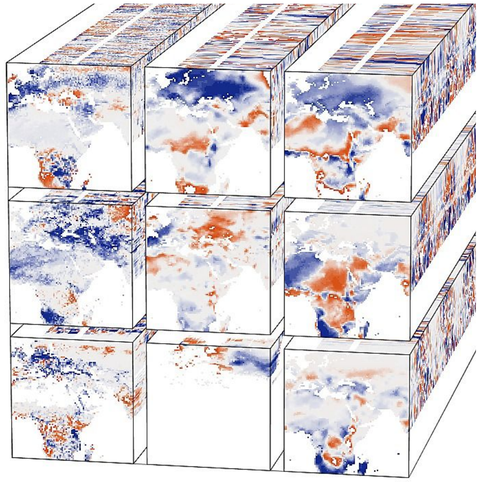
ARD is readily findable and usable!

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

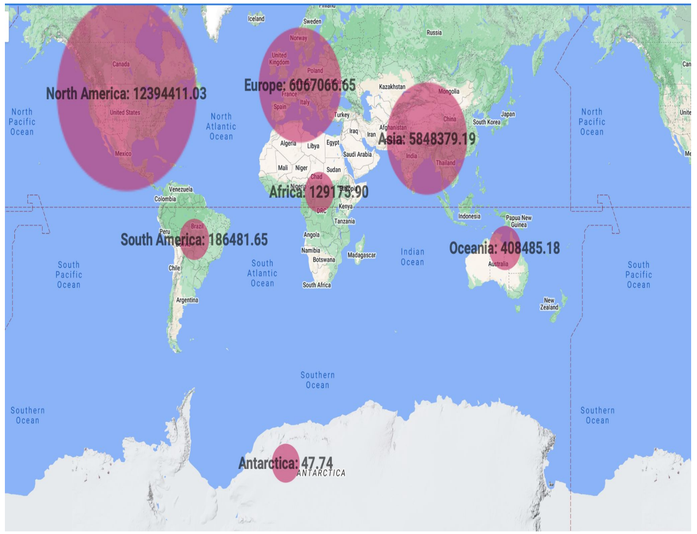
Need Analysis-Ready Cloud-Optimized Datasets (ARCO)

Need results NOW

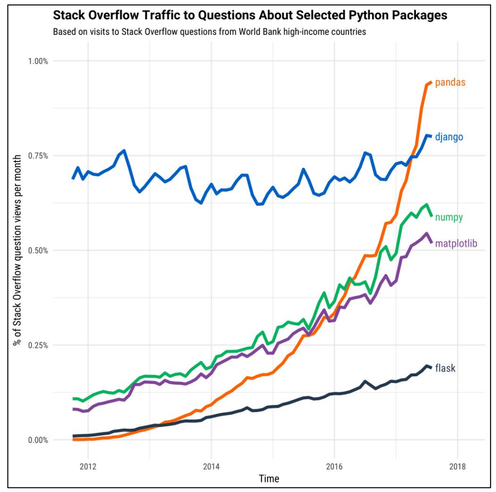
Appreciate multi-dasking!



Research constrained by resources limits participation and hence results



Democratize access to climate model output



Analysis methods to blend in with existing libraries

All in one dataset object

Rich metadata

chunking

xarray.Dataset

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

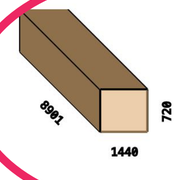
▾ Coordinates:

	()	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)	datetime64[ns]	1993-01-01 ... 2017-05…

▾ Data variables:

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



ARCO dataset example

Cloud-Native Scientific Data Analytics

1. Analysis-Ready, Cloud-Optimized Data

xarray.Dataset

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

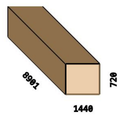
▾ Coordinates:

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:

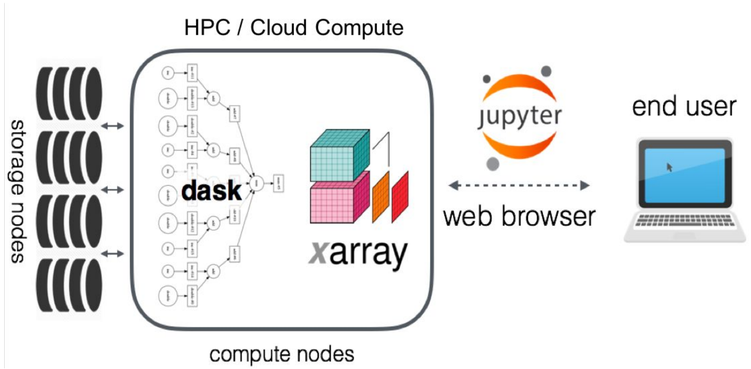
adt	(time, latitude, longitude)	float64	dask.array<chunksize=(…

	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



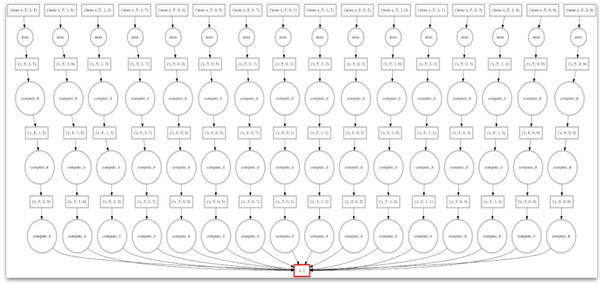
err	(time, latitude, longitude)	float64	dask.array<chunksize=(…
sla	(time, latitude, longitude)	float64	dask.array<chunksize=(…
ugos	(time, latitude, longitude)	float64	dask.array<chunksize=(…
ugosa	(time, latitude, longitude)	float64	dask.array<chunksize=(…
vgos	(time, latitude, longitude)	float64	dask.array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

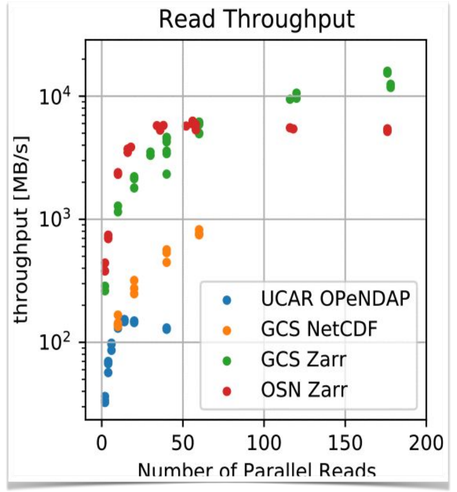
2. Data-Proximate Computing

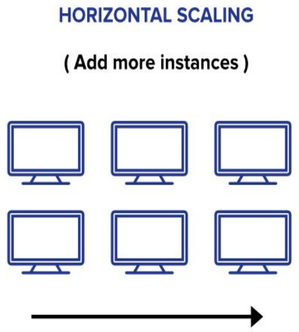


R. P. Abernathey et al., "Cloud-Native Repositories for Big Scientific Data," in Computing in Science & Engineering, vol. 23, no. 2, pp. 26-35, 1 March-April 2021, doi: 10.1109/MCSE.2021.3059437.

3. Elastic Distributed Processing





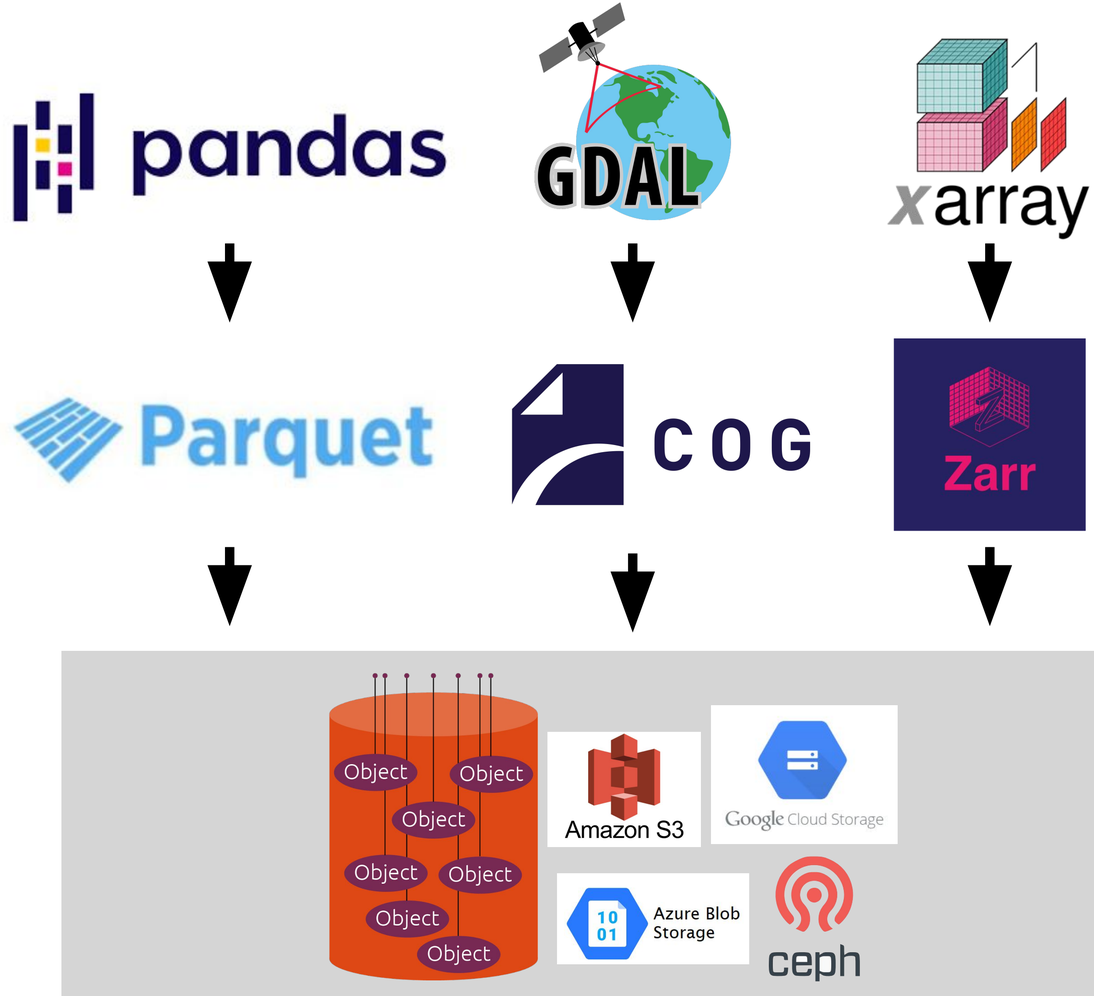


ARCO Data

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

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



R. P. Abernathey et al., "Cloud-Native Repositories for Big Scientific Data," in Computing in Science & Engineering, vol. 23, no. 2, pp. 26-35, 1 March-April 2021, doi: 10.1109/MCSE.2021.3059437.

Slide credit: Ryan Abernathey, Pangeo-forge FOSS4G 2022

CMIP6 ARD and ARCO data in the cloud
How?

Democratizing access to climate model output

CMIP6 S3 bucket

netCDF

arn:aws:s3:::esgf-world

AWS CLI Access aws s3 ls s3://esgf-world/ --no-sign-request

https://esgf-world.s3.amazonaws.com/index.html

Intake-esm catalog
https://cmip6-nc.s3.amazonaws.com/esgf-world.csv.gz
THREDDS catalog
https://aws-cloudnode.esgf.io/thredds/catalog/catalog.html
SpatioTemporal Asset Catalogs (STAC) underway

Zarr

arn:aws:s3:::cmip6-pds

AWS CLI Access aws s3 ls s3://cmip6-pds --no-sign-request

https://cmip6-pds.s3.amazonaws.com/index.html#CMIP6/

Intake-esm catalog
https://cmip6-pds.s3.amazonaws.com/pangeo-cmip6.csv
STAC catalogs underway

Checkout the CMIP6 registry in AWS to read more information, including CMIP6 data citations.
Ack: Amazon Sustainability Data Initiative (ASDI)

Earth System Grid Federation(ESGF)- A globally distributed archive of climate data

35
Projects

179
Countries

24
Data nodes

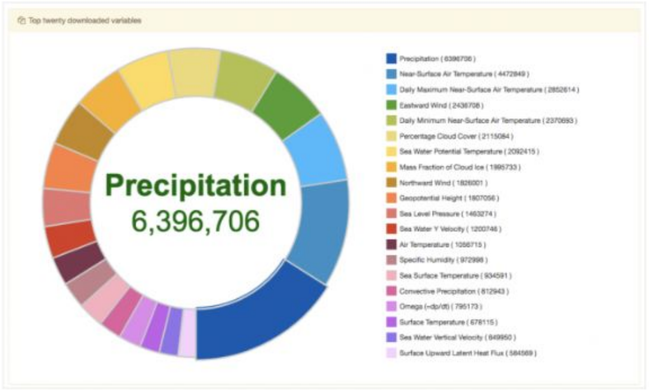
13,547,449
published datasets

1,104,293,736
files downloaded

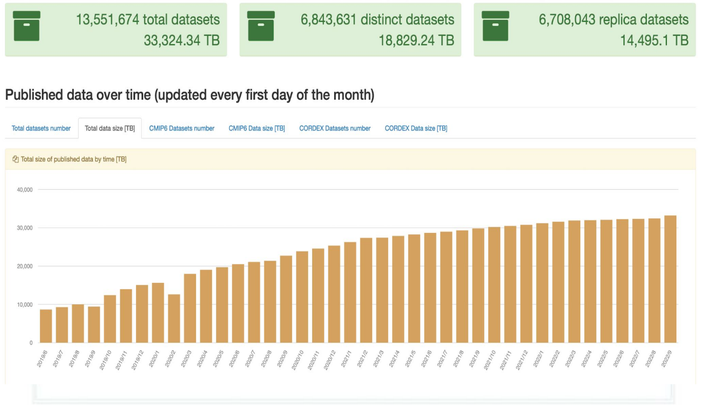
32.88
PB downloaded

ESGF Federation

Data usage



Data publication

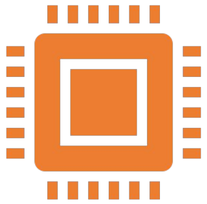


ESGF Dashboard: http://esgf-ui.cmcc.it

Accelerate research ->Collaborate-> Diversify the audience

Towards a more accessible, discoverable data holding

ESGF Future architecture: Few focus areas



Platforms and systems administration

Modular, scalable architecture: Containers, Kubernetes

Embrace infrastructure-as-code approach

Search services

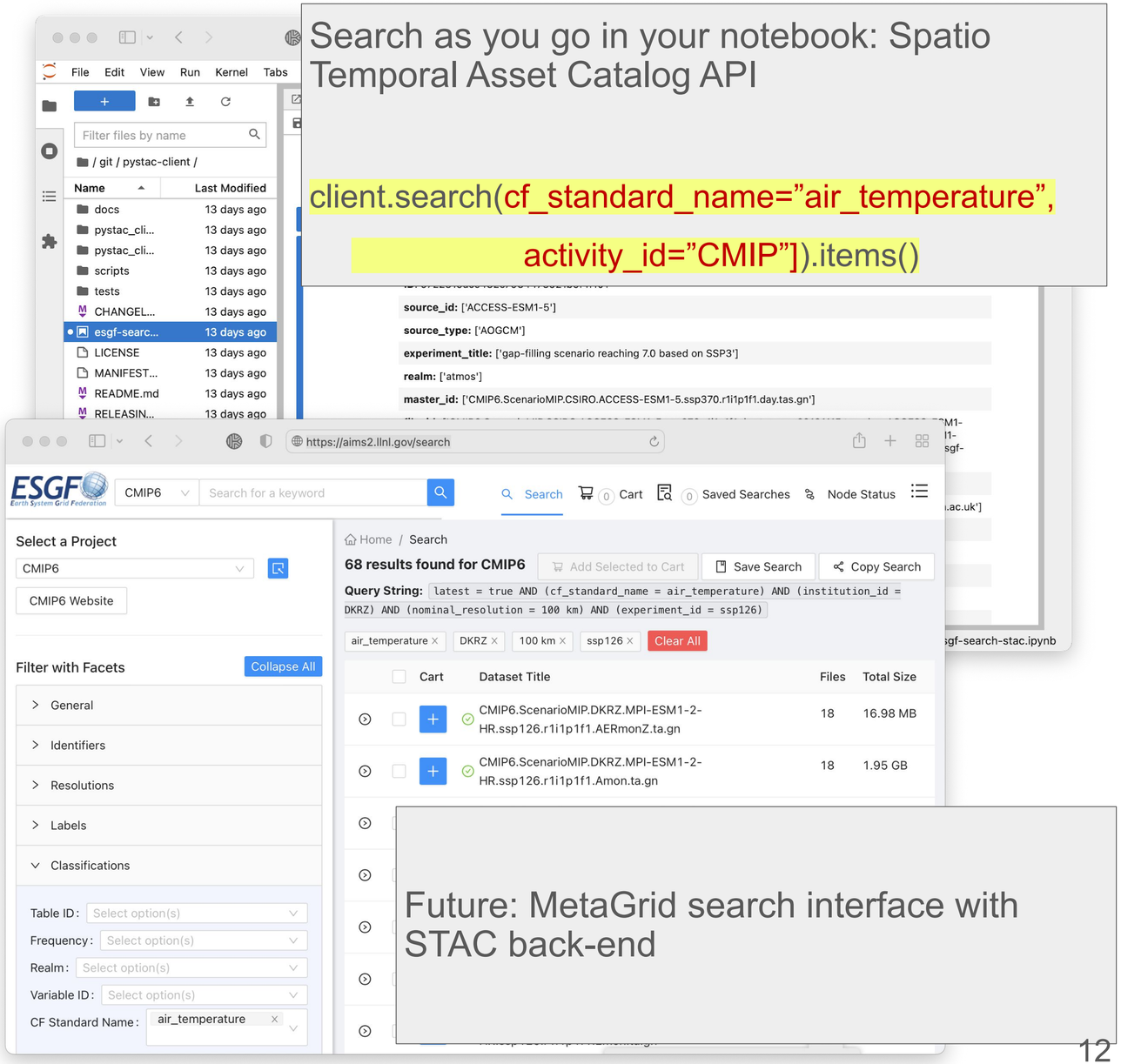
Modernise, centralise and simplify

Use community standards: STAC

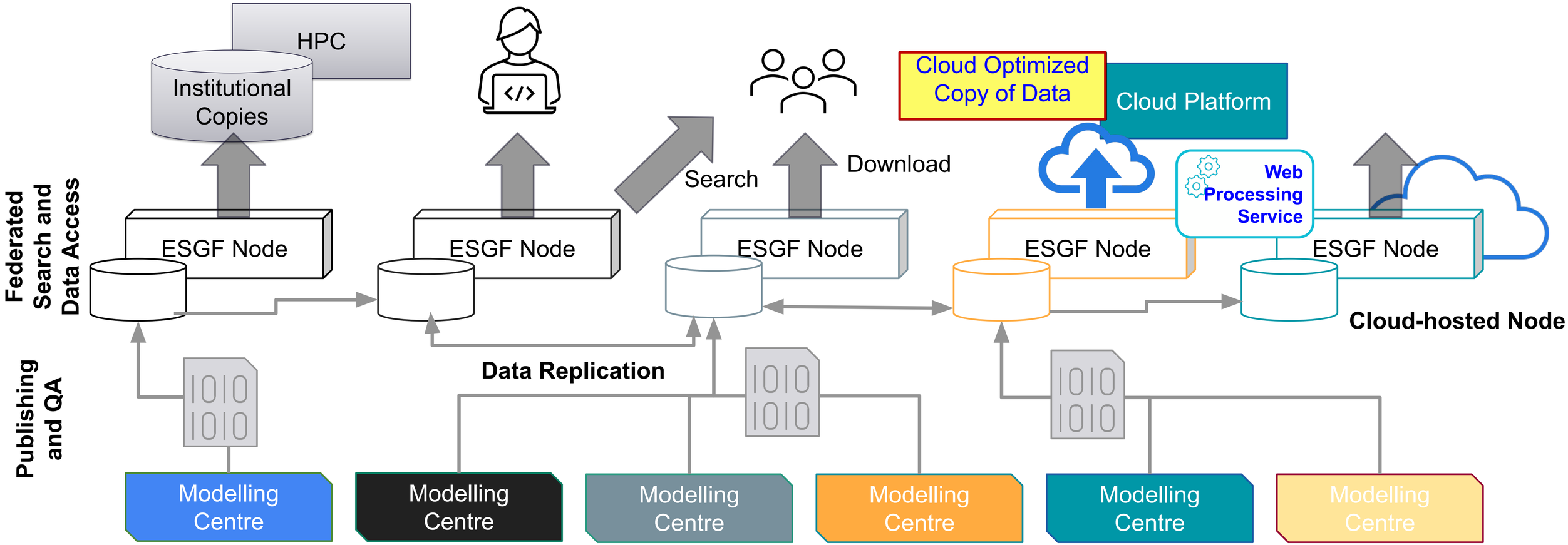
IS-ENES3 has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084

Making data collection pleasant

- Metadata, metadata..
- Leverage community data standards.
  - E.g. Climate Forecast conventions, CMIP Data Reference Syntax and controlled vocabulary
- Extend to different data formats



Future inclusive architecture of ESGF



Adapted from: Philip Kershaw, ESGF - CPCMW 2022

A reproducible pathway to ARCO data

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

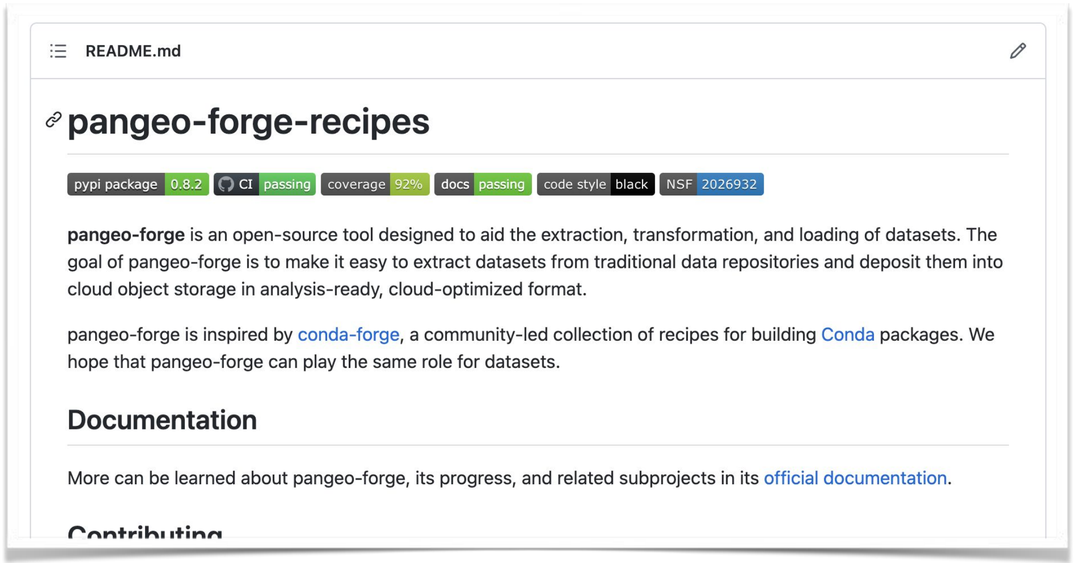
Analysis Skills:
To validate and make use of the ARCO data.

Data Scientist

Slide credit: Ryan Abernathey, Pangeo-forge FOSS4G 2022

Pangeo Forge Recipes

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



Open source python package for describing and running data pipelines ("recipes")

inspired by: conda-forge

Pangeo Forge Cloud

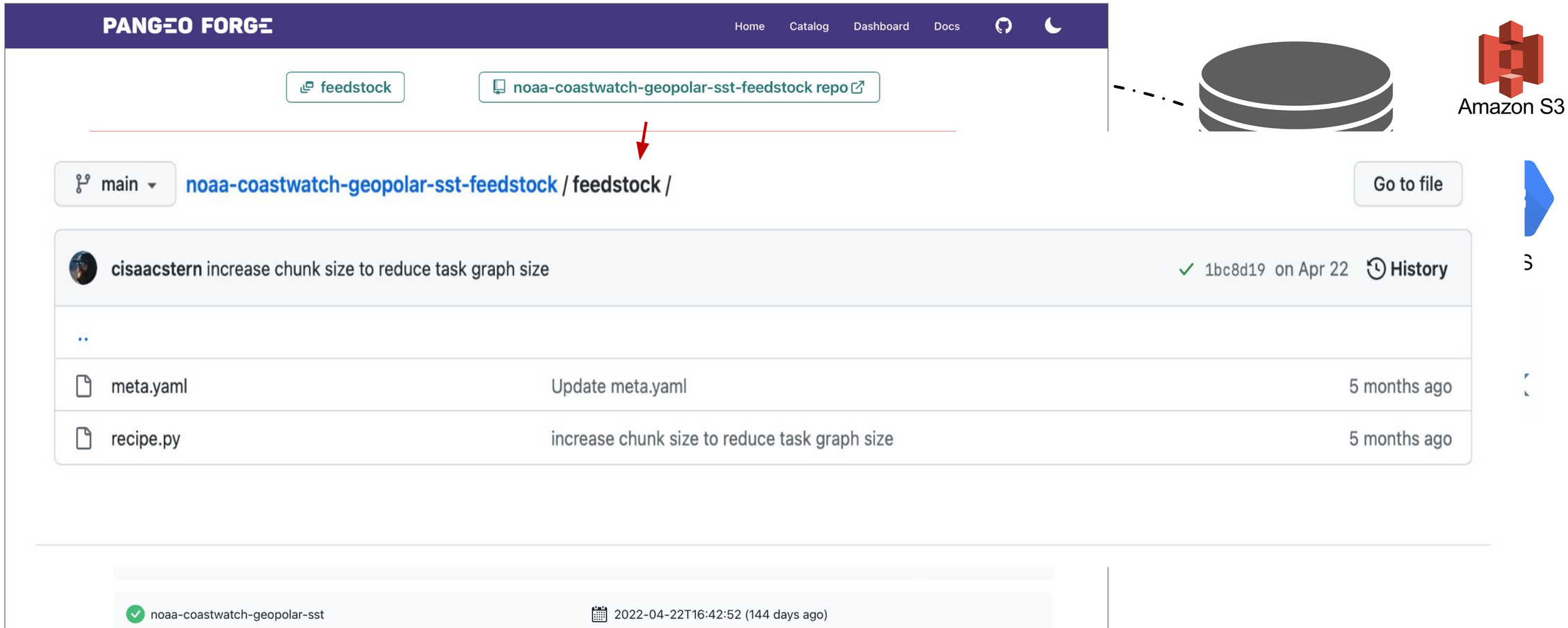
https://pangeo-forge.org/

Cloud platform for automatically executing recipes stored in GitHub repos.

Slide credit: Ryan Abernathey, Pangeo-forge FOSS4G 2022

Pangeo Forge Cloud

https://pangeo-forge.org/



Pangeo Forge Cloud

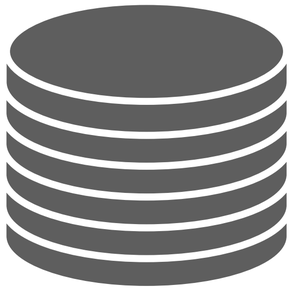
https://pangeo-forge.org/

Contains the code and metadata for one or more Recipes

Feedstock

Bakery

Runs the recipes in the cloud using elastic scaling clusters



Storage

Amazon S3

GCS

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

open storage network

https://www.openstoragenetwork.org

Slide credit: Ryan Abernathey, Pangeo-forge FOSS4G 2022

Acknowledgement

- **Charles Stern** (Columbia / LDEO)
- Joe Hamman (CarbonPlan)
- Anderson Banihirwe (CarbonPlan)
- Rachel Wegener (U. Maryland)
- Chiara Lepore (GRO Intelligence)

Funding: NSF Earthcube Program

- Sean Harkins (Development Seed)
- Alex Merose (Google Research)
- Tom Augspurger (Microsoft)
- Martin Durant (Anaconda)
- Many recipe contributors

Funding: NSF Earthcube Program

This is a 💯% open project! Join us!

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

Lessons learned

- Prioritization
  - F in FAIR may also mean Fundability to sustain
- Research acceleration
  - Extend, not always build from scratch.
- Recognition [and traceability]
  - Think beyond citations as an impact metric!