



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?

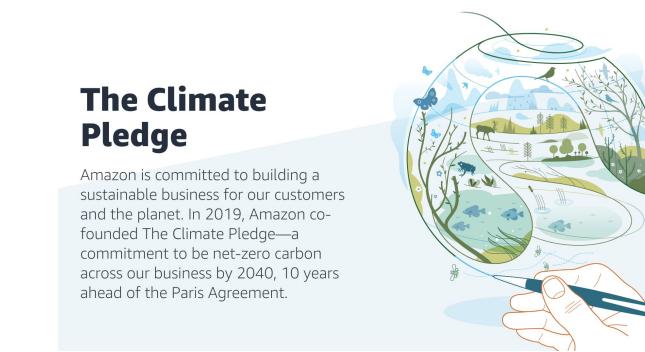
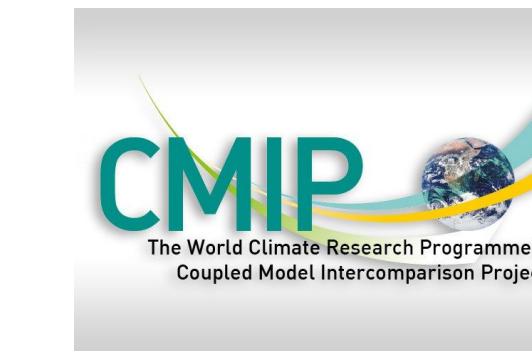
NOAA Environmental Data Management Workshop 2022

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 Nikonorov, 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.

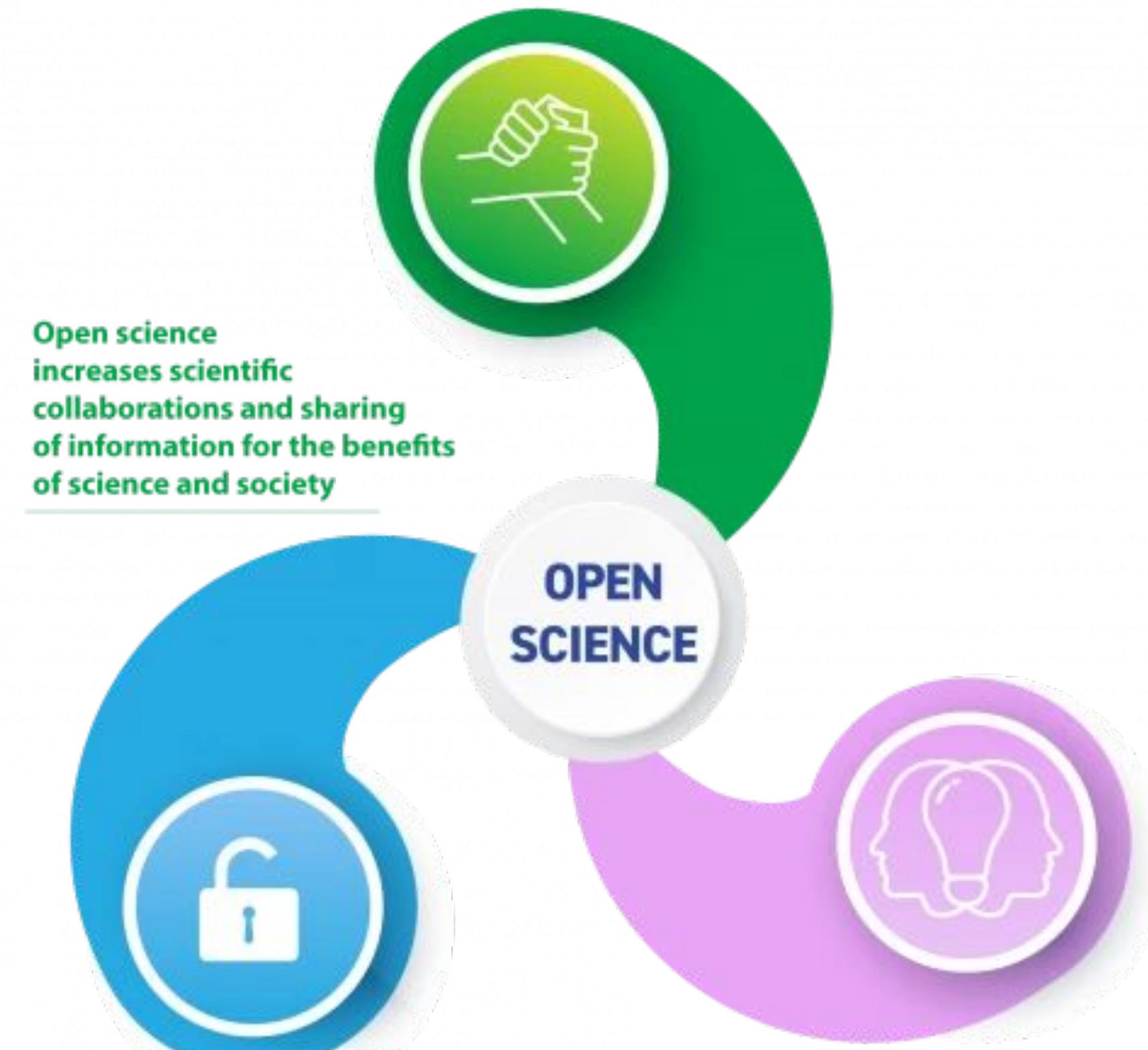


Open Science



designed by freepik

[Image credit](#)



Open science
increases scientific
collaborations and sharing
of information for the benefits
of science and society

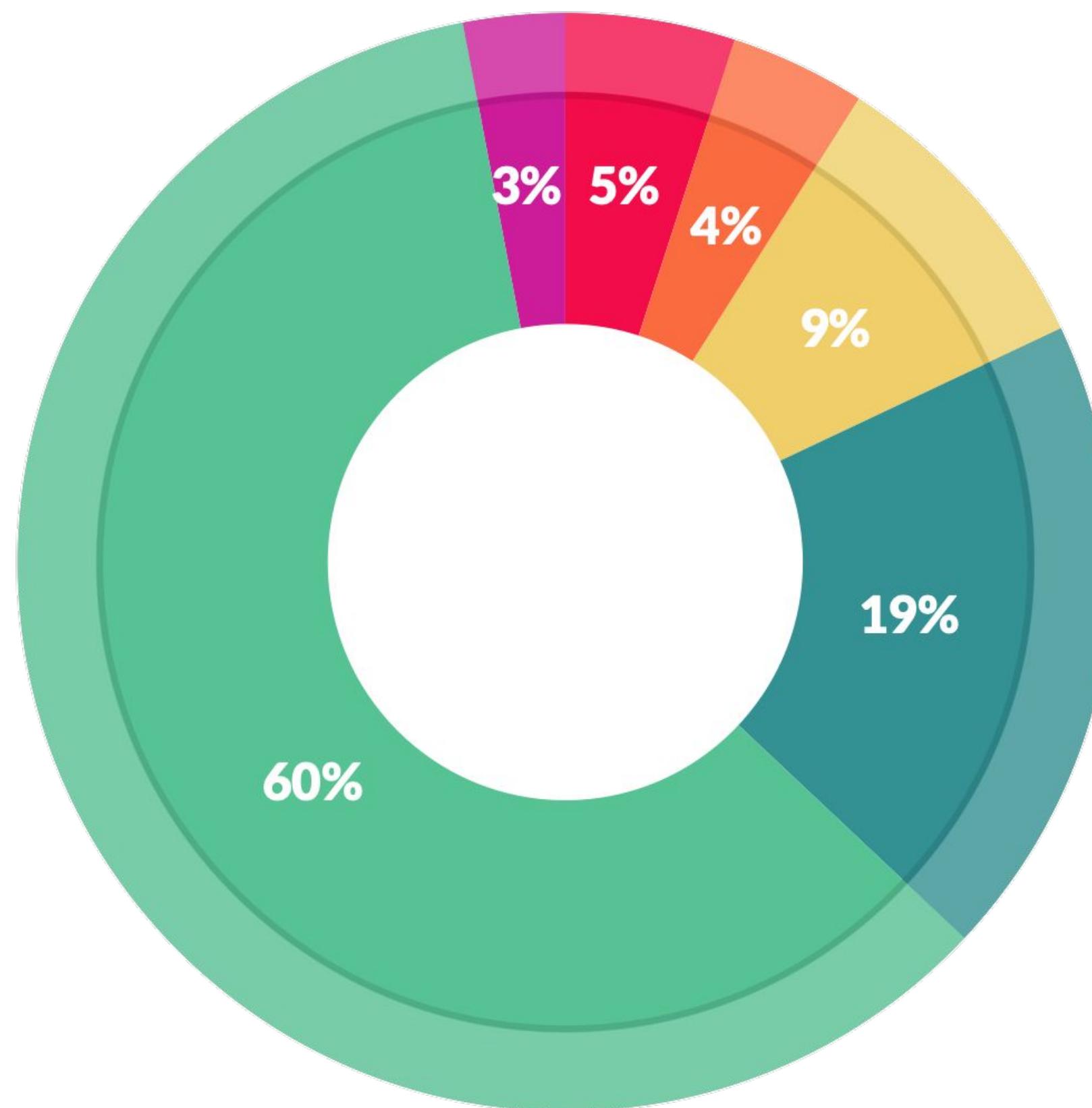
makes multilingual scientific
knowledge openly available,
accessible and reusable for
everyone

opens the processes of scientific
knowledge creation, evaluation and
communication to societal actors
beyond the traditional scientific
community.



<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!

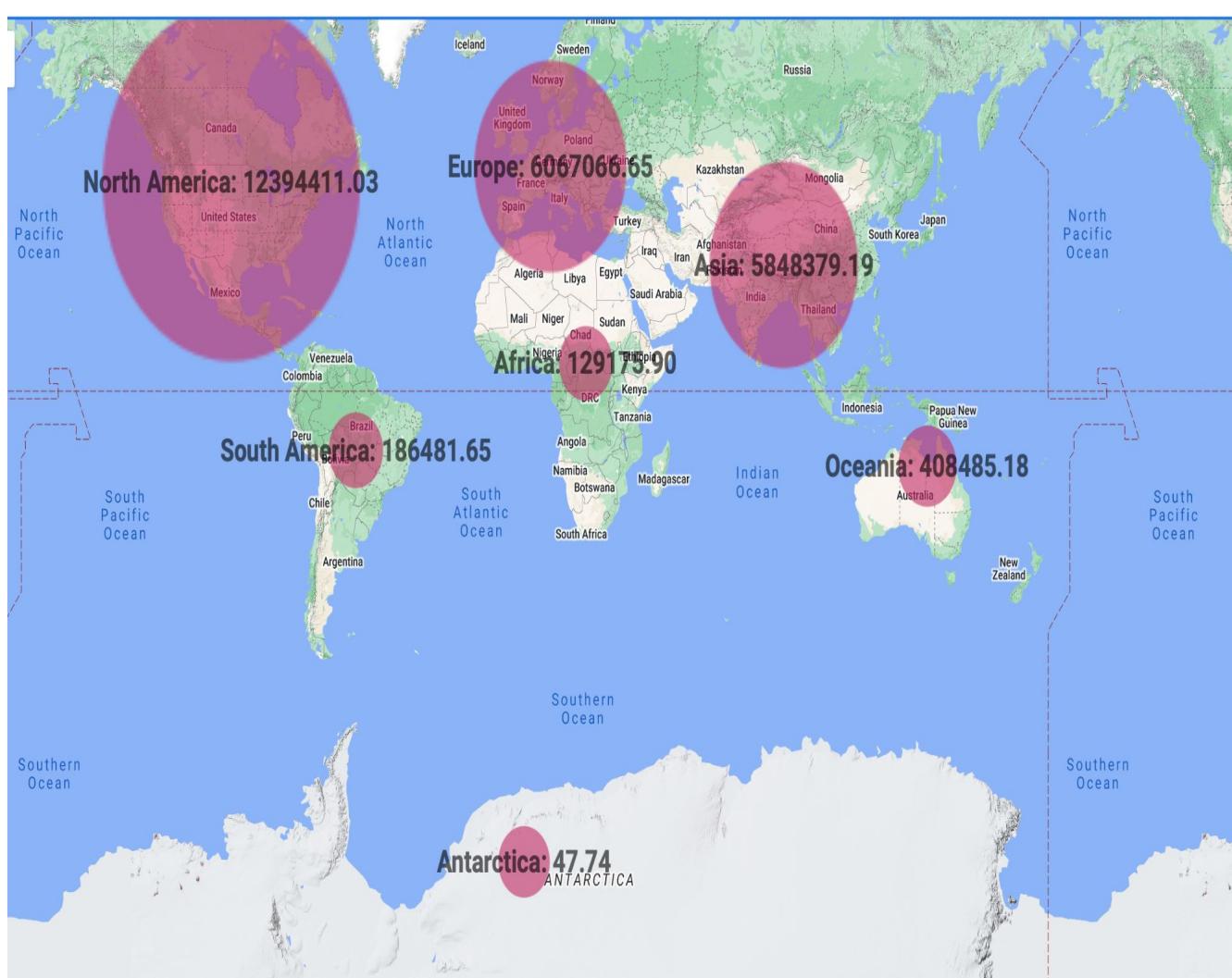
How do data scientists spend their time?
Crowdflower Data Science Report (2016)

Need Analysis-Ready Cloud-Optimized Datasets (ARCO)

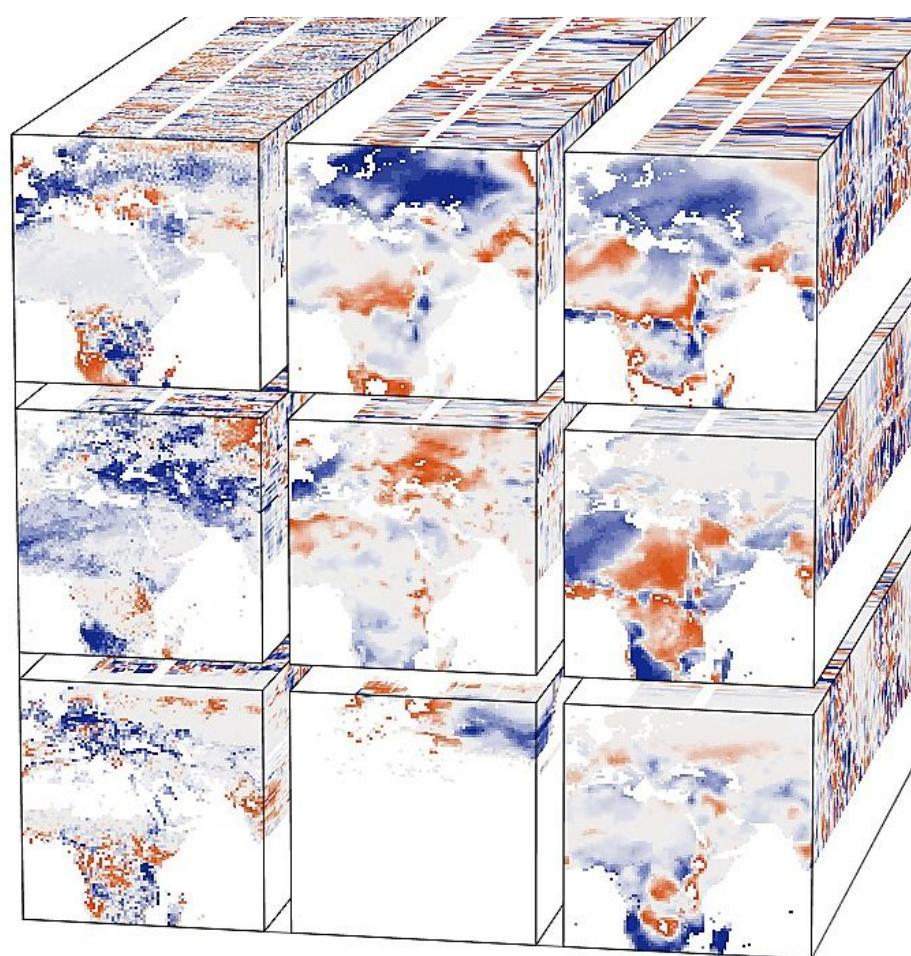


Need results NOW

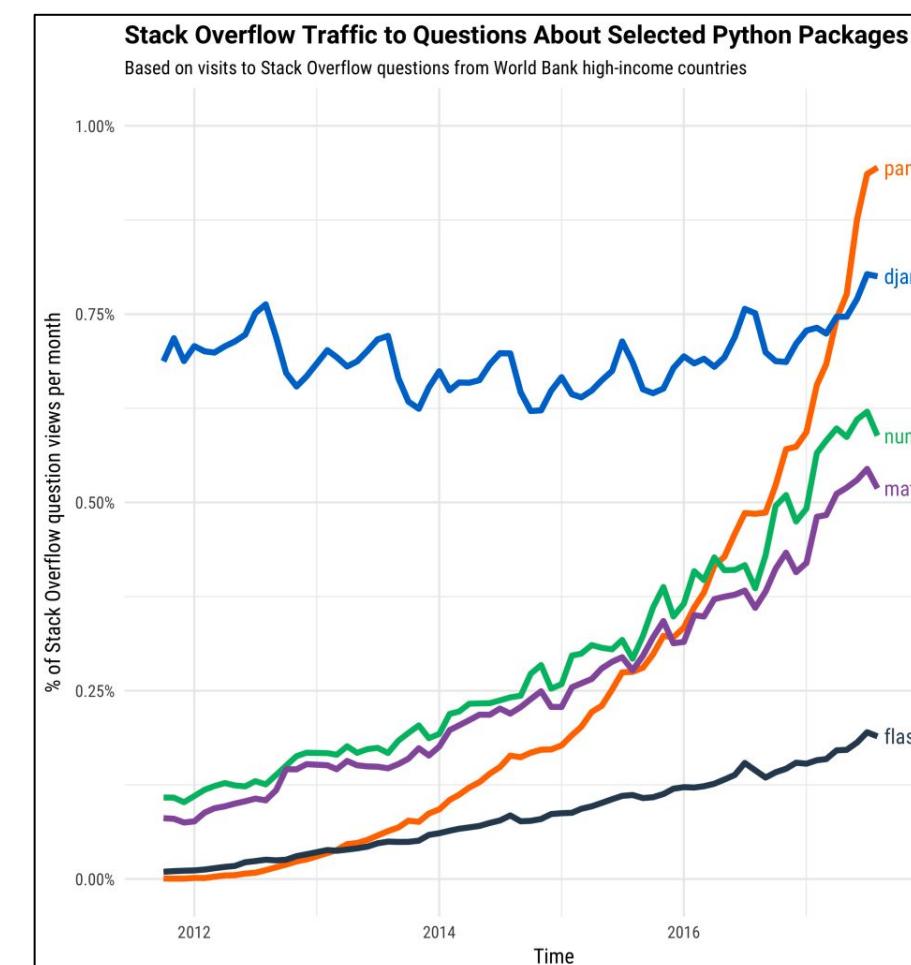
Appreciate multi-dasking!



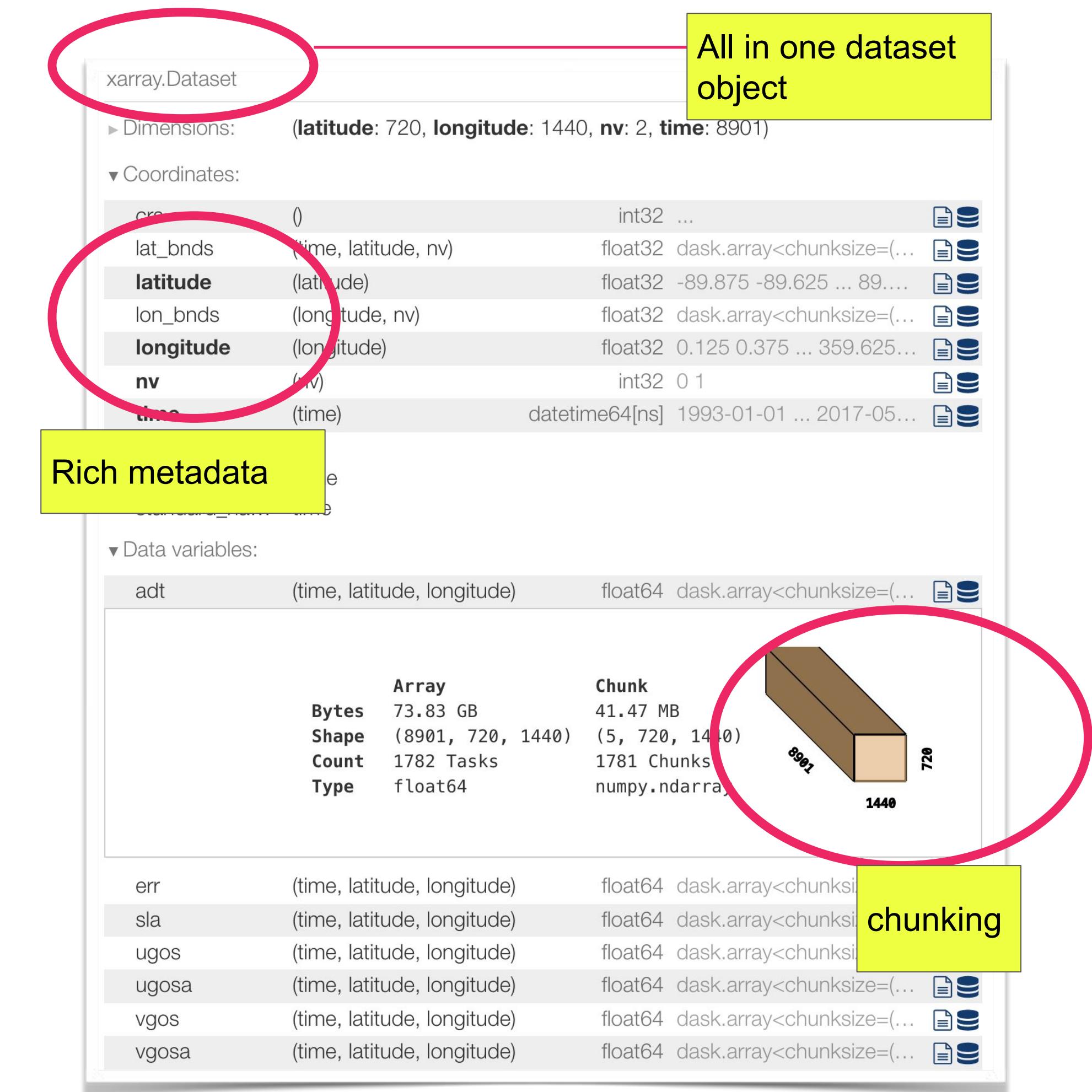
Democratize access to climate model output



Research constrained by resources limits participation and hence results



Analysis methods to blend in with existing libraries



ARCO dataset example

Cloud-Native Scientific Data Analytics

1. Analysis-Ready, Cloud-Optimized Data

The screenshot shows the xarray.Dataset interface. It displays the following information:

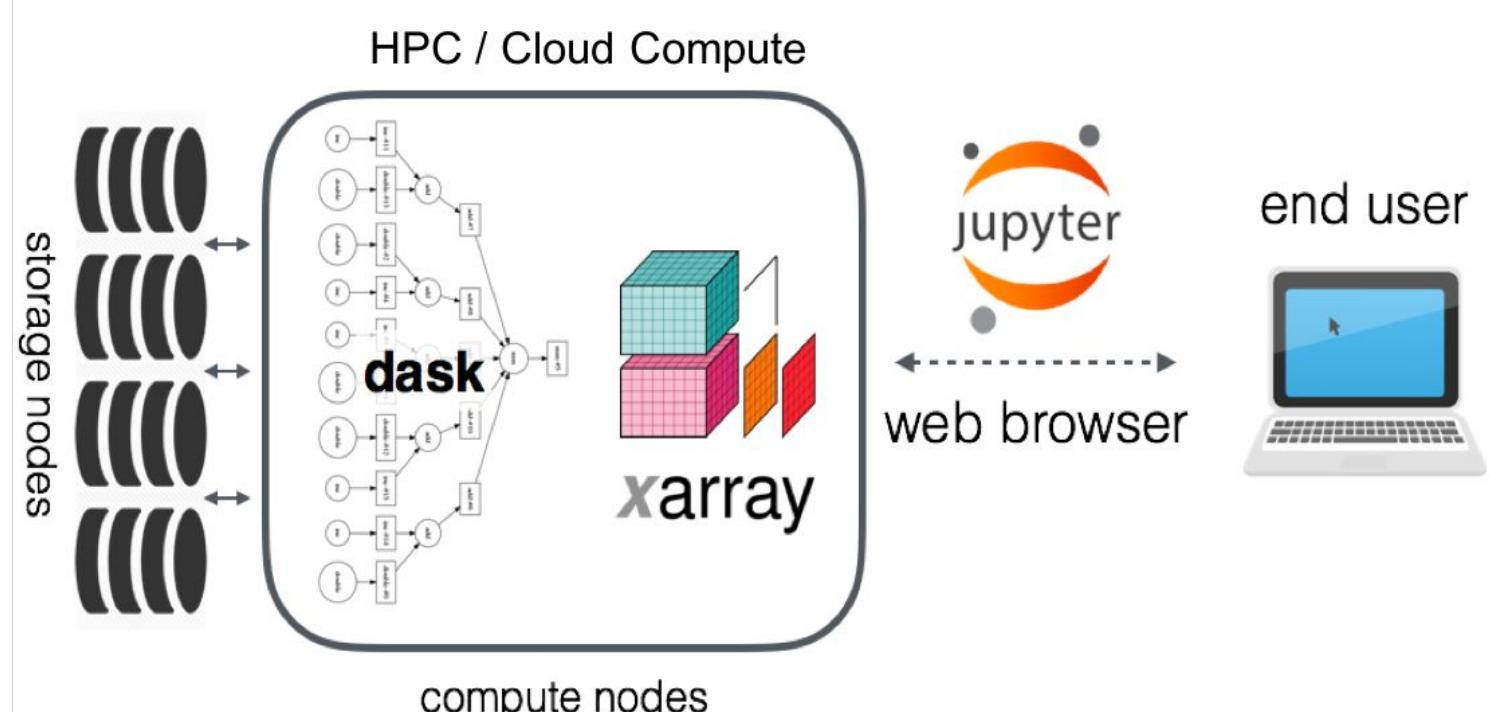
- Dimensions:** latitude: 720, longitude: 1440, nv: 2, time: 8901
- Coordinates:** crs, lat_bnds, latitude, lon_bnds, longitude, nv, time, axis, long_name, standard_name.
- Data variables:** adt, err, sla, ugos, ugosa, vgosa, vgos, vgosa.

For the 'adt' variable, detailed information is provided:

	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

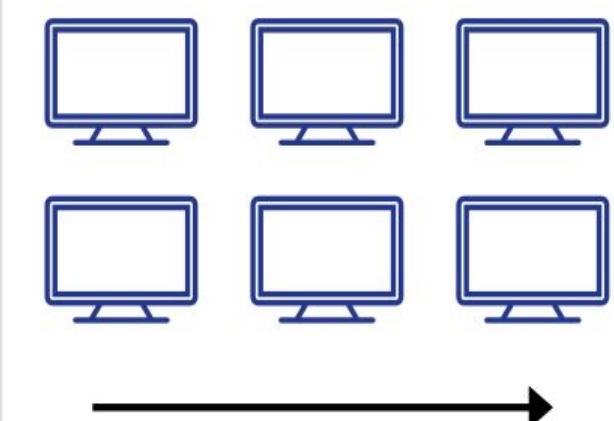
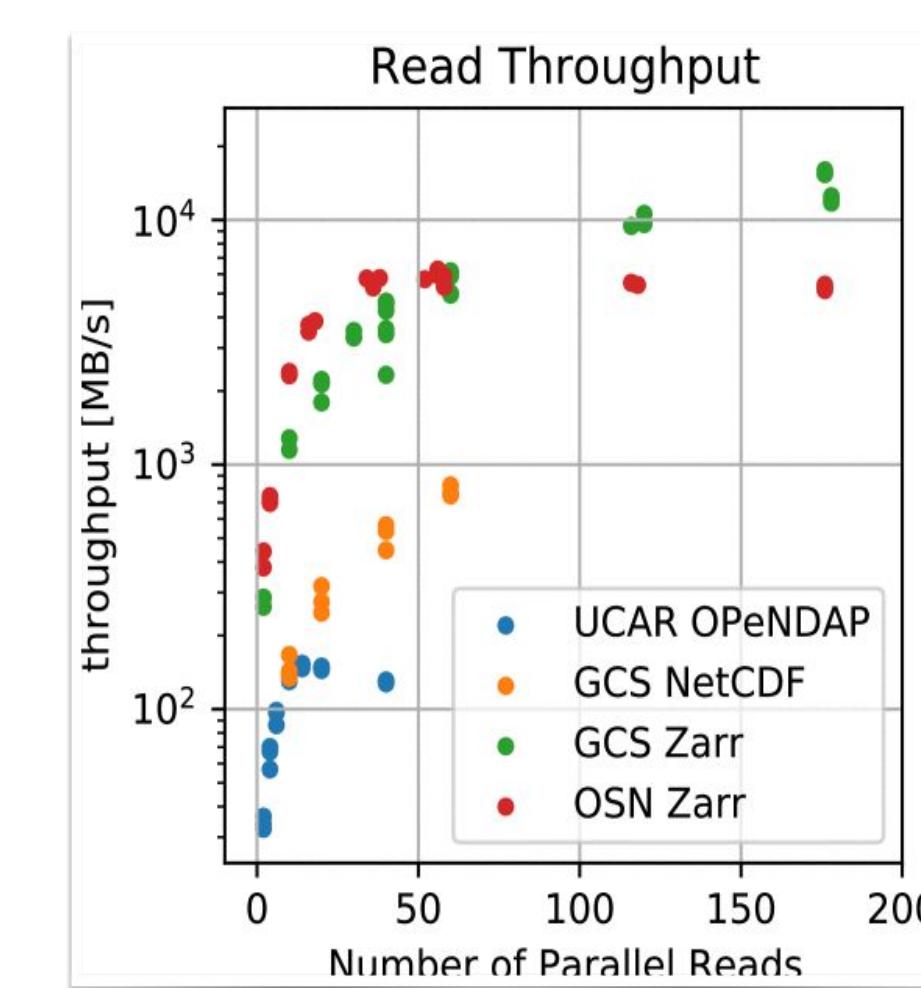
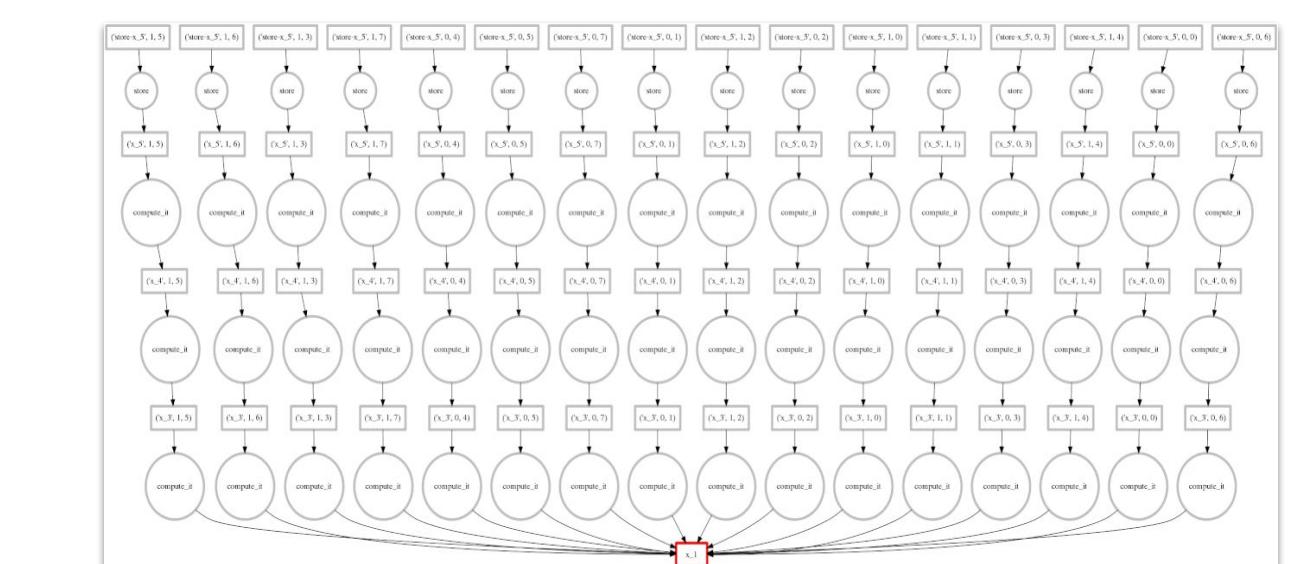
A 3D bar chart visualizes the data structure with dimensions latitude (720), longitude (1440), and time (8901).

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.

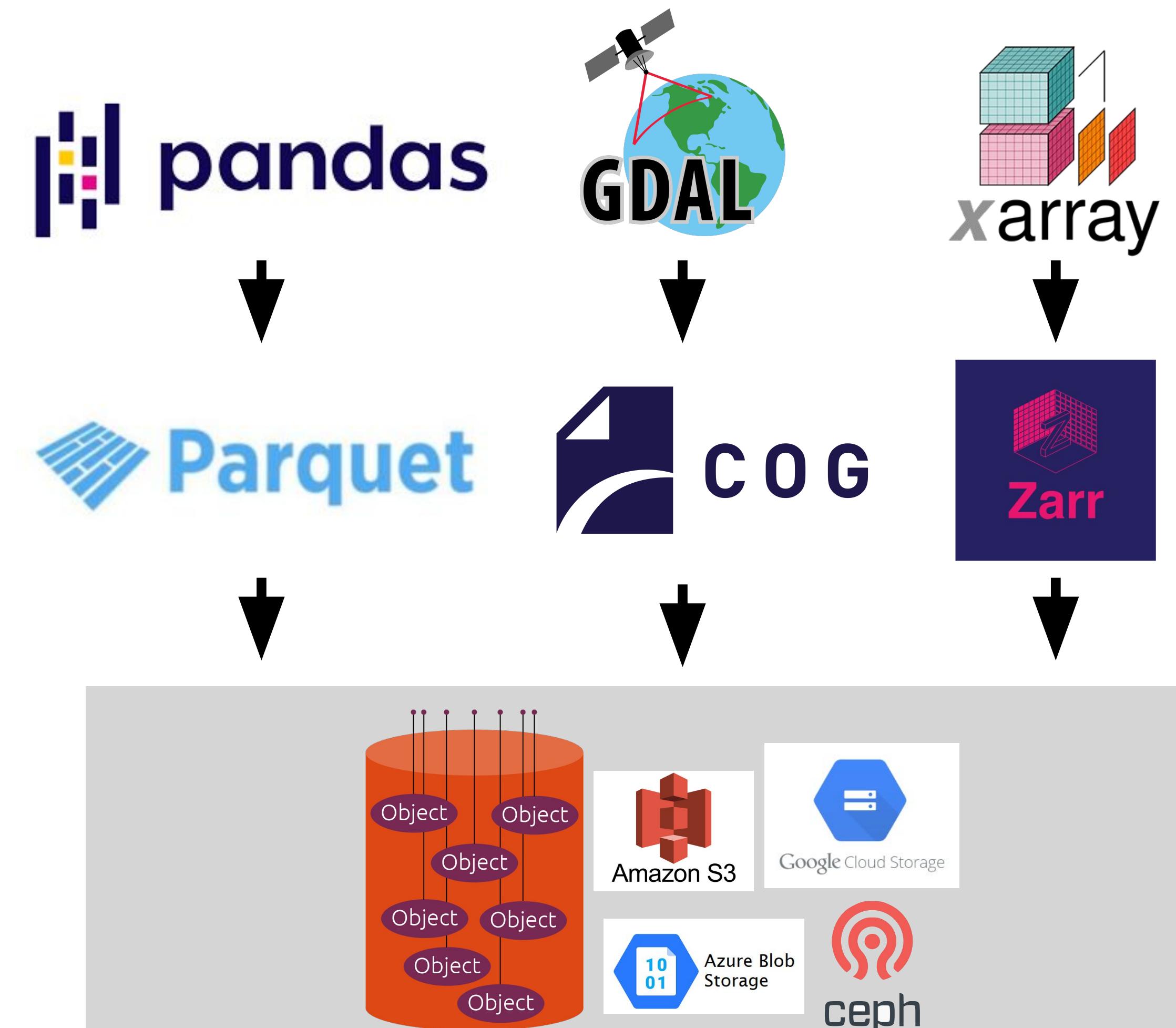
3. Elastic Distributed Processing



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

Analysis Ready, Cloud Optimized



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

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>



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-nc.s3.amazonaws.com/esgf-world.csv.gz>

THREDDS catalog

<https://aws-cloudnode.esgf.io/thredds/catalog/catalog.html>

SpatioTemporal Asset Catalogs (STAC)
underway

Intake-esm catalog

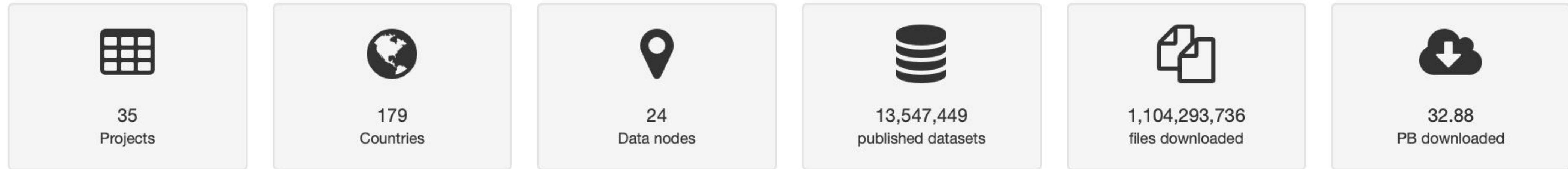
<https://cmip6-pds.s3.amazonaws.com/pangeo-cmip6.csv>

STAC catalogs underway

Checkout the [CMIP6 registry in AWS](#) to read more information, including CMIP6 data citations.

Ack: Amazon Sustainability Data Initiative (ASDI)

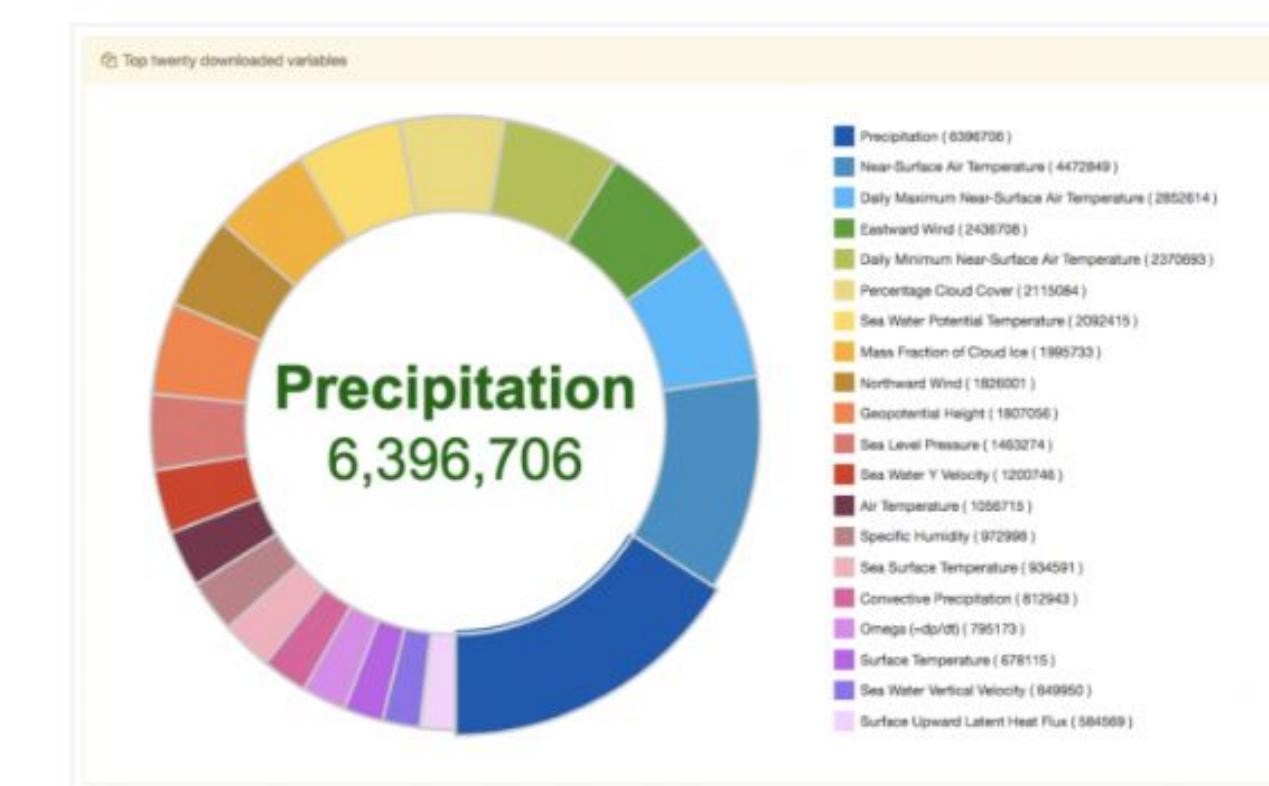
Earth System Grid Federation(ESGF)- A globally distributed archive of climate data



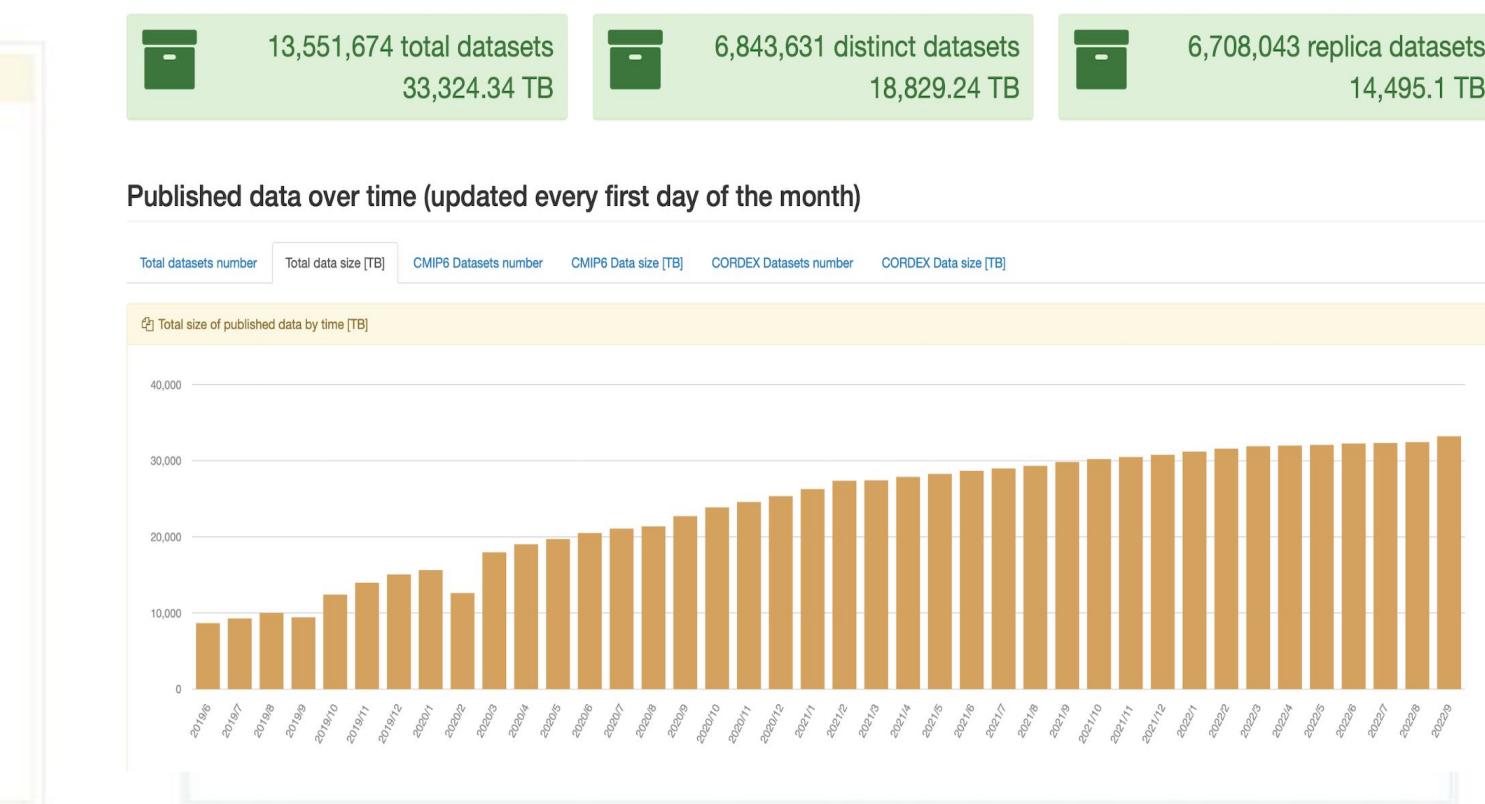
ESGF Federation



Data usage



Data publication

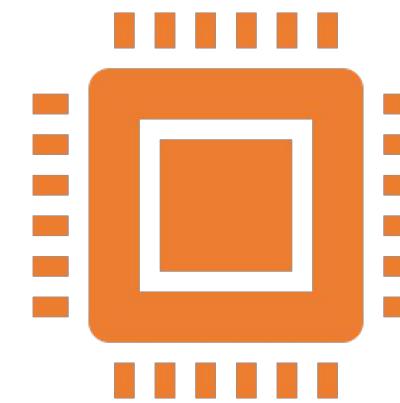


ESGF Dashboard: <http://esgf-ui.cmcc.it>

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



Philip Kershaw, ESGF - IS-ENES 2021

11



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

Search as you go in your notebook: Spatio Temporal Asset Catalog API

```
client.search(cf_standard_name="air_temperature", activity_id="CMIP").items()
```

source_id: ['ACCESS-ESM1-5']
source_type: ['AOGCM']
experiment_title: ['gap-filling scenario reaching 7.0 based on SSP3']
realm: ['atmos']
master_id: ['CMIP6.ScenarioMIP.CSIRO.ACCESS-ESM1-5.ssp370.r1i1p1f1.day.tas.grn']

client.esgf-search-stac.ipynb

ESGF Earth System Grid Federation CMIP6 Search for a keyword Search Cart Saved Searches Node Status

Select a Project CMIP6 Website

Filter with Facets General Identifiers Resolutions Labels Classifications

Table ID: Select option(s) Frequency: Select option(s) Realm: Select option(s) Variable ID: Select option(s) CF Standard Name: air_temperature

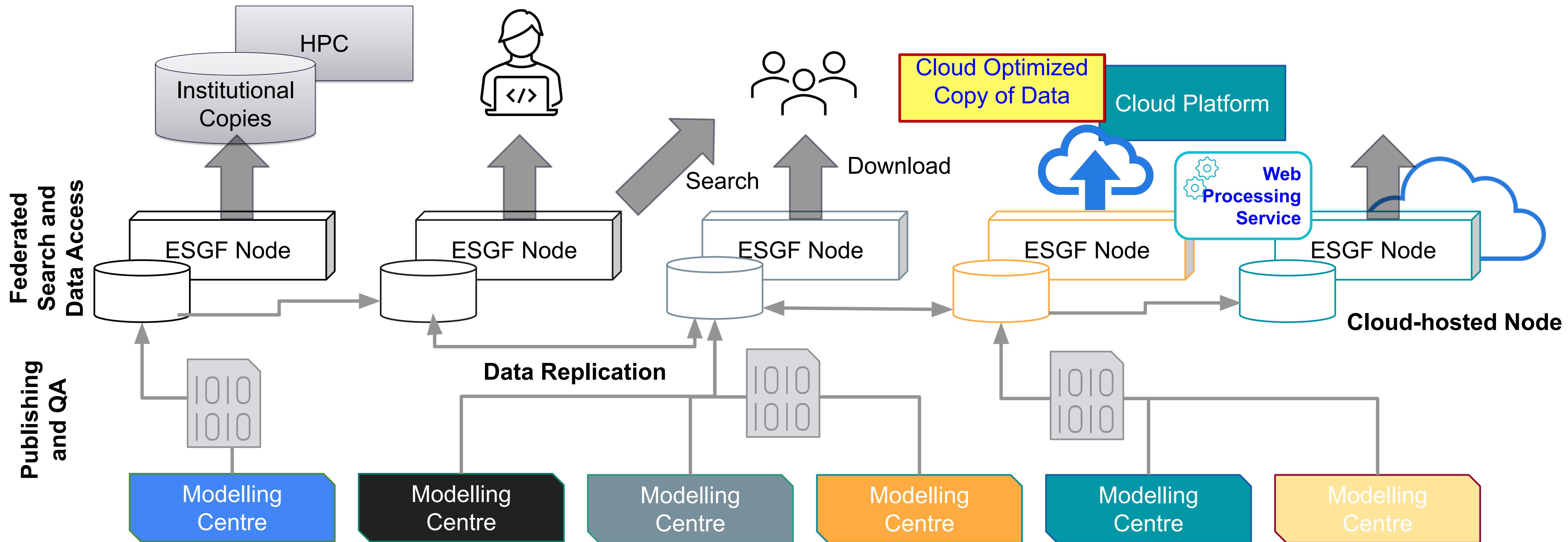
68 results found for CMIP6 Add Selected to Cart Save Search Copy Search

Query String: latest = true AND (cf_standard_name = air_temperature) AND (institution_id = DKRZ) AND (nominal_resolution = 100 km) AND (experiment_id = ssp126)

Cart	Dataset Title	Files	Total Size
<input type="checkbox"/>	CMIP6.ScenarioMIP.DKRZ.MPI-ESM1-2-HR.ssp126.r1i1p1f1.AERmonZ.ta.grn	18	16.98 MB
<input type="checkbox"/>	CMIP6.ScenarioMIP.DKRZ.MPI-ESM1-2-HR.ssp126.r1i1p1f1.Amon.ta.grn	18	1.95 GB

Future: MetaGrid search interface with STAC back-end

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.



Pangeo Forge Recipes

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

A screenshot of a GitHub repository page for "pangeo-forge-recipes". The page includes a README.md file, a project description, a "Documentation" section, and a "Contributing" section. The project description highlights that pangeo-forge is an open-source tool designed to aid the extraction, transformation, and loading of datasets, inspired by conda-forge.

Open source python package for
describing and running data pipelines
("recipes") inspired by: conda-forge

Pangeo Forge Cloud

<https://pangeo-forge.org/>

A screenshot of the Pangeo Forge Cloud homepage. It features the Pangeo Forge logo and a dark purple header with navigation links for Home, Catalog, Dashboard, Docs, and GitHub. Below the header, there's a summary of project metrics: 1 feedstock, 4 recipe runs, and 0 datasets.

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

Home Catalog Dashboard Docs

feedstock noaa-coastwatch-geopolar-sst-feedstock repo

main noaa-coastwatch-geopolar-sst-feedstock / feedstock / Go to file

cisaacstern increase chunk size to reduce task graph size ✓ 1bc8d19 on Apr 22 History

..

meta.yaml Update meta.yaml 5 months ago

recipe.py increase chunk size to reduce task graph size 5 months ago

Amazon S3

Google Cloud Storage

noaa-coastwatch-geopolar-sst

2022-04-22T16:42:52 (144 days ago)

Pangeo Forge Cloud

<https://pangeo-forge.org/>

Contains the code
and metadata for one
or more Recipes



Feedstock



[terraclimate-feedstock](#)

A pangeo-smithy repository for the terraclimate dataset.

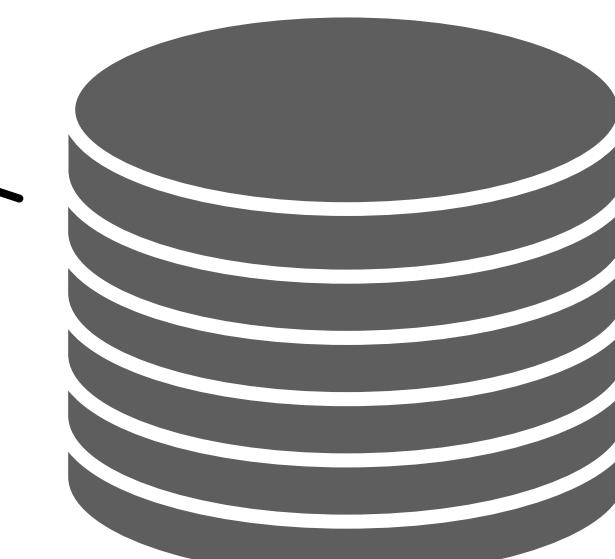
● Python Apache-2.0 3 2 1 3 Updated on Jan 1

[noaa-oisst-avhrr-feedstock](#)

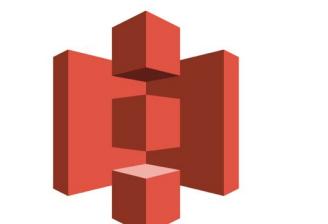
● Python Apache-2.0 2 1 0 4 Updated on Jan 1



Runs the recipes in the
cloud using elastic
scaling clusters



Storage



Amazon S3



GCS



[open storage network](https://www.openstoragenetwork.org)

<https://www.openstoragenetwork.org>

Acknowledgement

- **Charles Stern** (Columbia / LDEO)
- Joe Hamman (CarbonPlan)
- Anderson Banhirwe (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 100% 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!

