BEYOND FAIR

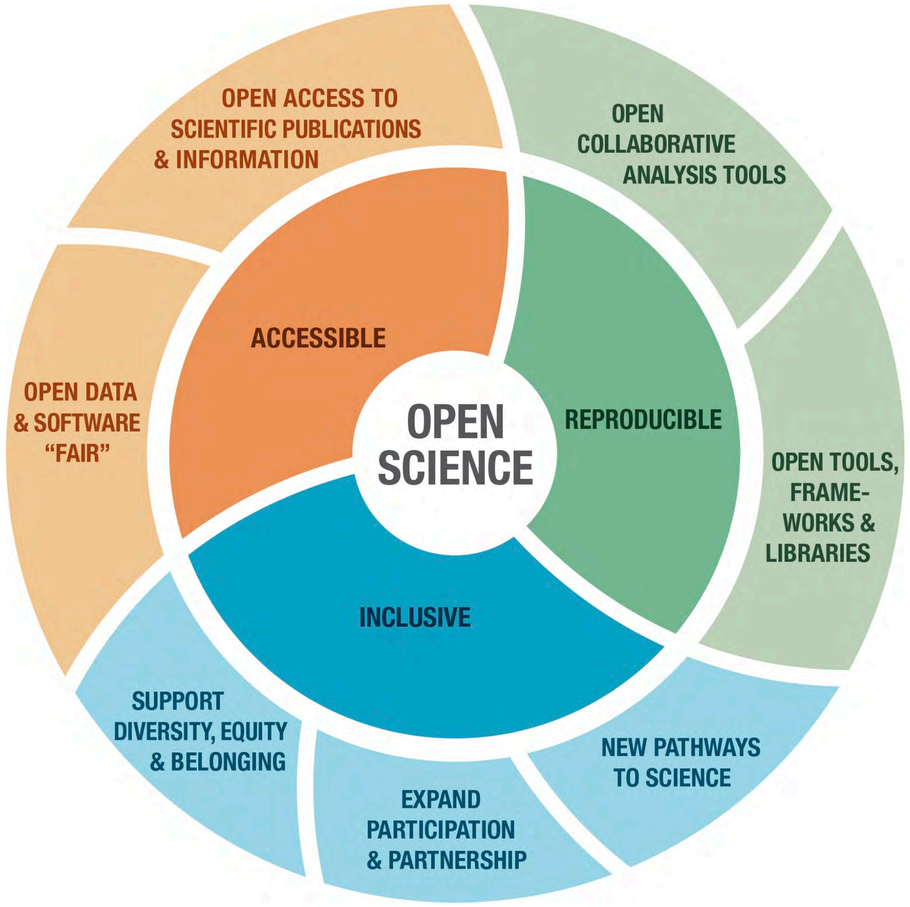
What data infrastructure does open science need?

Ryan Abernathey

PANGEO

OOIFB Data Systems Committee Meeting, 2022

THE OPEN SCIENCE VISION



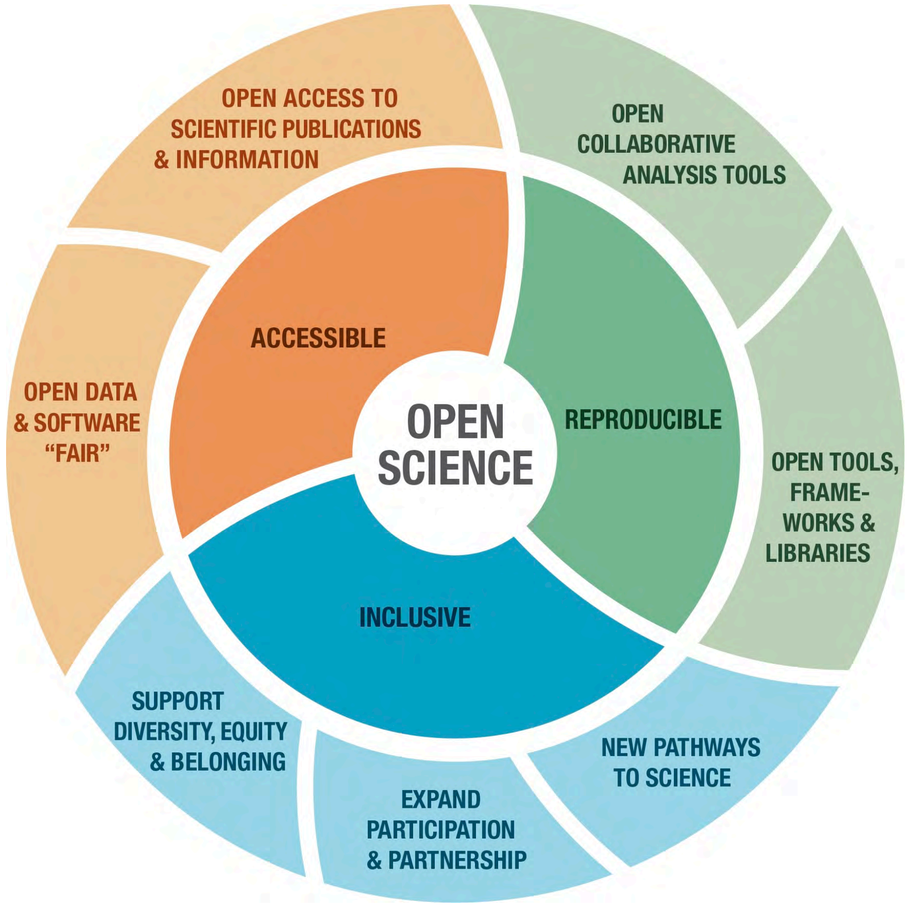
https://earthdata.nasa.gov/esds/open-science

```
for 👩🏾‍🔬 in everyone_in_the_world:
    for 📄 in all_scientific_knowledge:
        👩🏾‍🔬.verify(📄)
        discovery = 👩🏾‍🔬.extend(📄)
```

This would transform the 🌎 by allowing all of humanity to participate in the scientific process.

What are the barriers to realizing this vision?

THE OPEN SCIENCE VISION



https://earthdata.nasa.gov/esds/open-science

```
for 👩🏾‍🔬 in everyone_in_the_world:
    for 👨‍🔬 in everyone_else_in_the_world:
        📄 = (👩🏾‍🔬 + 👨‍🔬).collaborate()
```

This would transform the 🌎 by allowing all of humanity to participate in the scientific process.

What are the barriers to realizing this vision?

❤️🥰FAIR🥰❤️

FAIR = Findable, Accessible, Interoperable, Reusable

FAIR is great. Nobody disagrees with FAIR.

But making data-intensive scientific workflows FAIR is easier said than done. FAIR does not specify the protocols, technologies, or infrastructure.

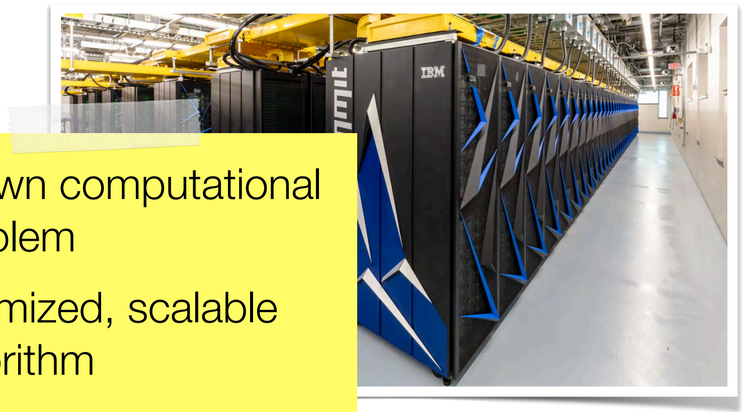
FAIR is not a platform.

PANGEO

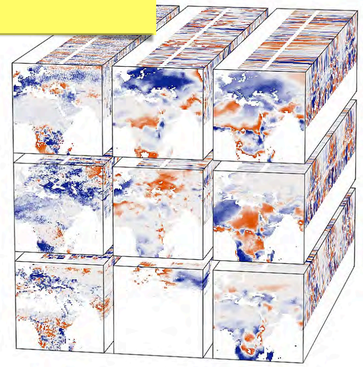
Simulation vs. Data-Intensive Science

$$\rho \frac{\mathrm{D}\mathbf{u}}{\mathrm{D}t} = -\nabla p + \nabla \cdot \boldsymbol{\tau} + \rho \mathbf{g}$$

Equations



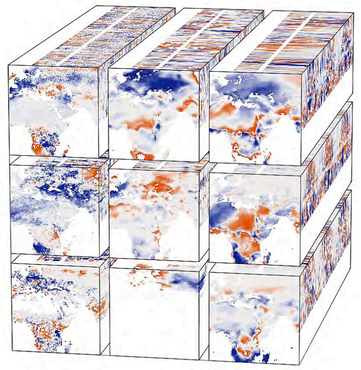
- known computational problem
- optimized, scalable algorithm
- standard architecture (HPC / MPI)



Big Data

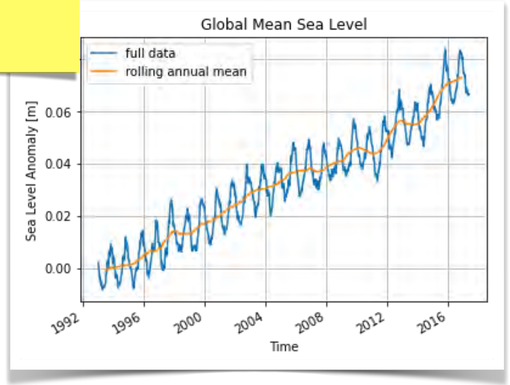
https://figshare.com/articles/figure/Earth_Data_Cube/4822930/2

- open-ended problem
- exploratory analysis
- “human in the loop”
- ML model development & training
- visualization needed
- highly varied computational patterns / algorithms



Big Data

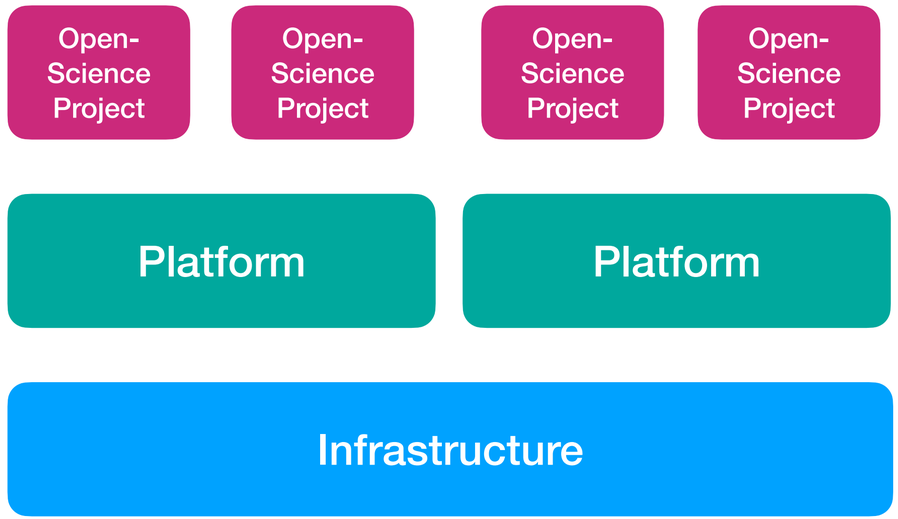
?



- Insights
- Understanding
- Predictions

CLAIM: OPEN SCIENCE NEEDS A PLATFORM

- The word “platform” is terribly overloaded.
- A platform is something you can build on—specifically, new scientific discoveries and new translational applications. Let’s call these *projects*. 📄
- For open science to take off at a global scale, *everyone in the world needs access to the platform* (like Facebook)
- This is why we are excited about cloud, but cloud as-is (e.g. AWS) is not itself an open-science platform.
- Does the open science platform need to be open? 🤔



PANGEO

OUTLINE

10 mins	The status quo of data-intensive scientific infrastructure
10 mins	Cloud computing and Pangeo
10 mins	From Software to SaaS: Pangeo Forge and Earthmover
10 mins	Where are things headed?

PART I: THE STATUS QUO

DATA-INTENSIVE SCIENCE INFRASTRUCTURE: THE STATUS QUO*

Personal Laptop

Group Server

Department Cluster

Agency Supercomputer

more storage, more CPU, more security, more constraints

STATUS QUO: WHAT INFRASTRUCTURE CAN WE RELY ON?

✅ UNIX operating system 💾

✅ Files / POSIX filesystems 🗂️

✅ Programming languages: C, FORTRAN, Python, R, Julia

✅ Terminal access >_

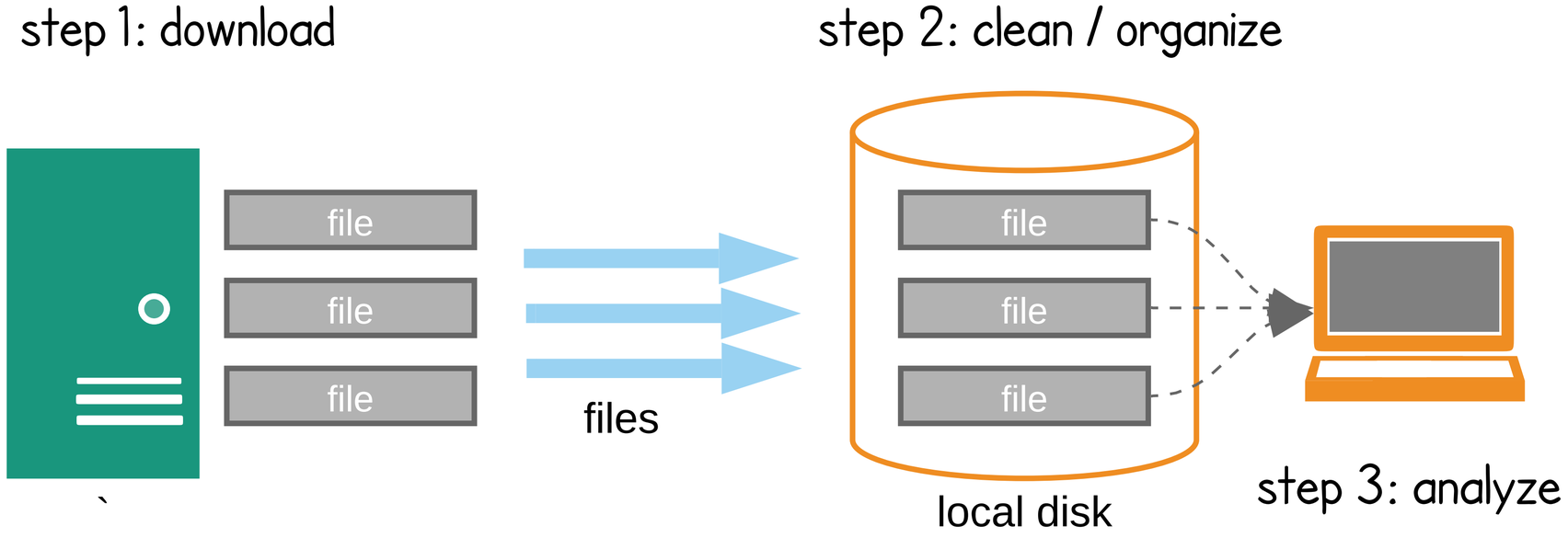
⚠️ Batch queuing system ⏬
HPC only

⚠️ The internet 🌐
Not on HPC nodes!

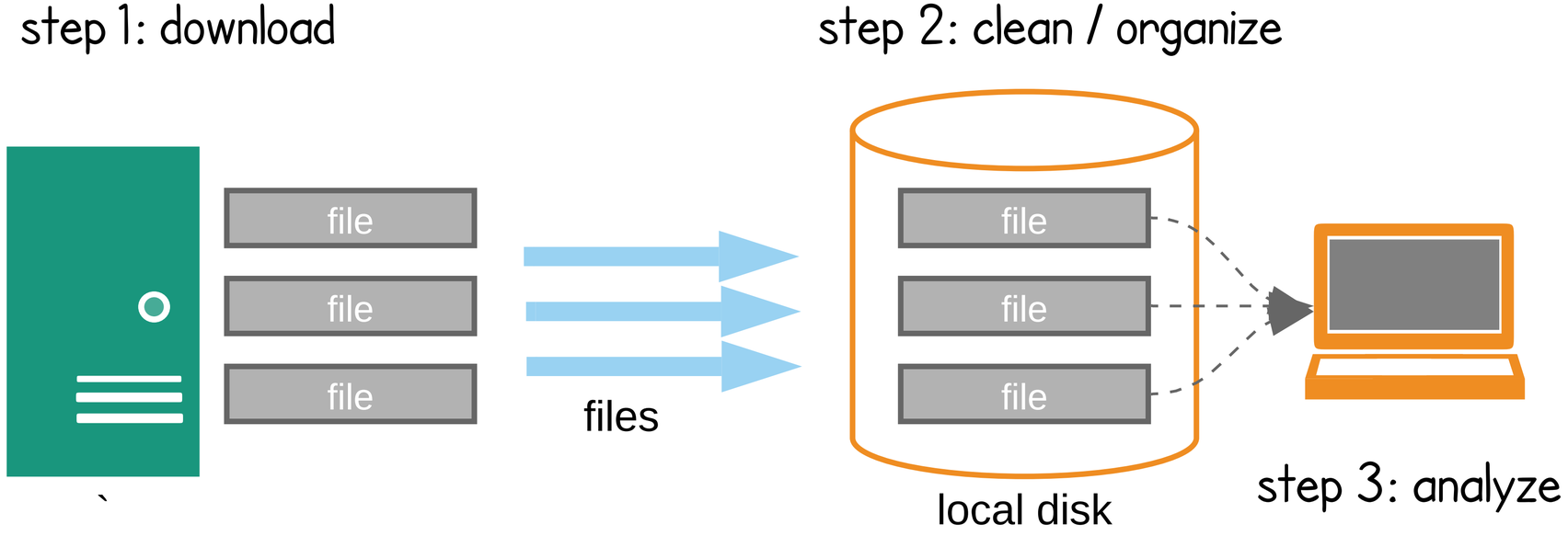
⚠️ Globus for file transfer 🔄
Not supported everywhere

❌ High level data services, APIs, etc.
Virtually unknown in my world

THE "DOWNLOAD" MODEL

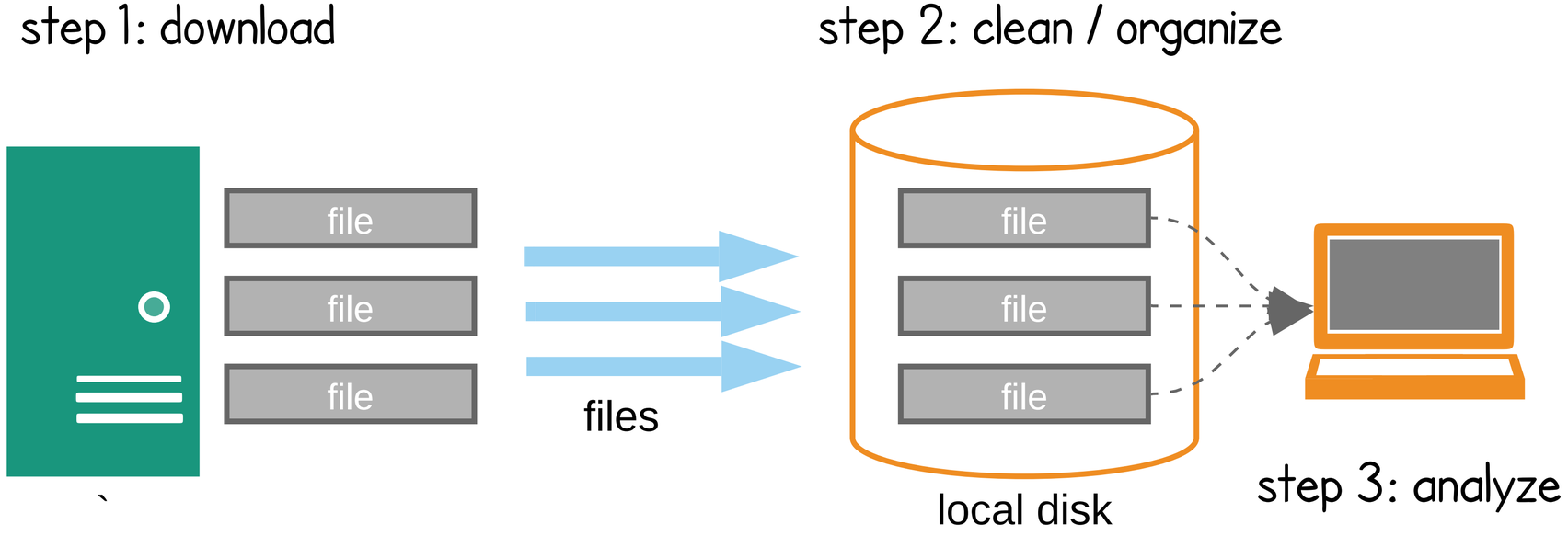


THE "DOWNLOAD" MODEL



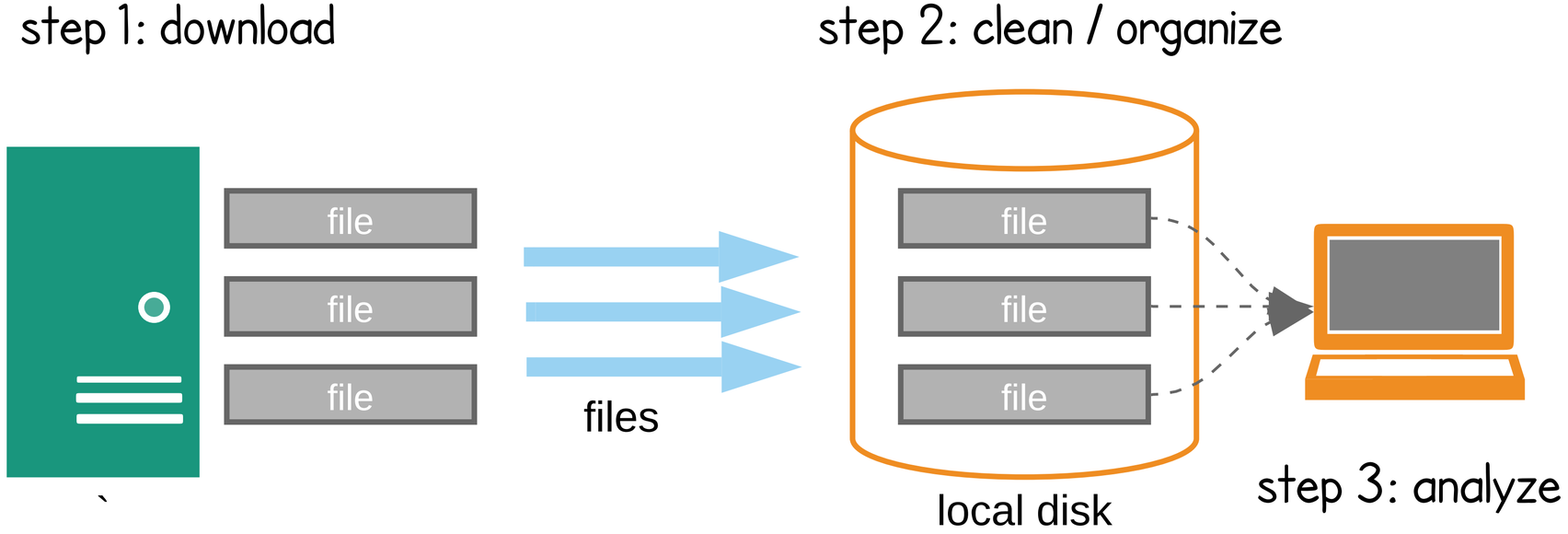
MB

THE "DOWNLOAD" MODEL



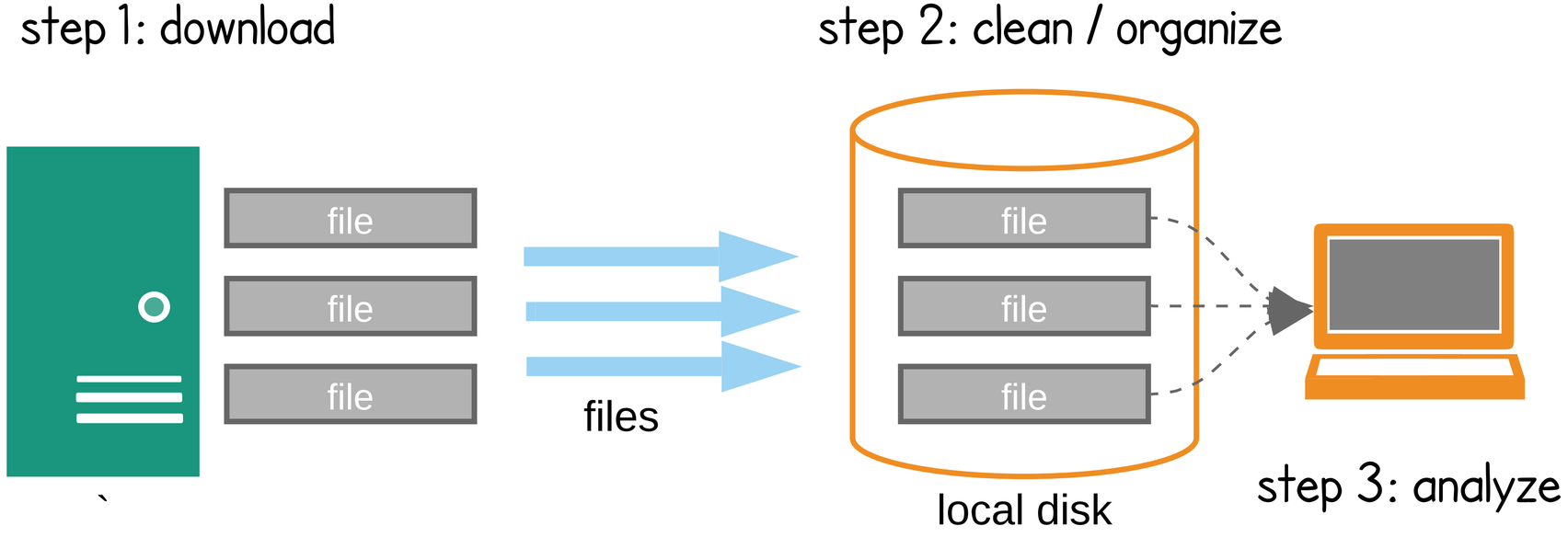
GB

THE "DOWNLOAD" MODEL



TB

THE "DOWNLOAD" MODEL



PB

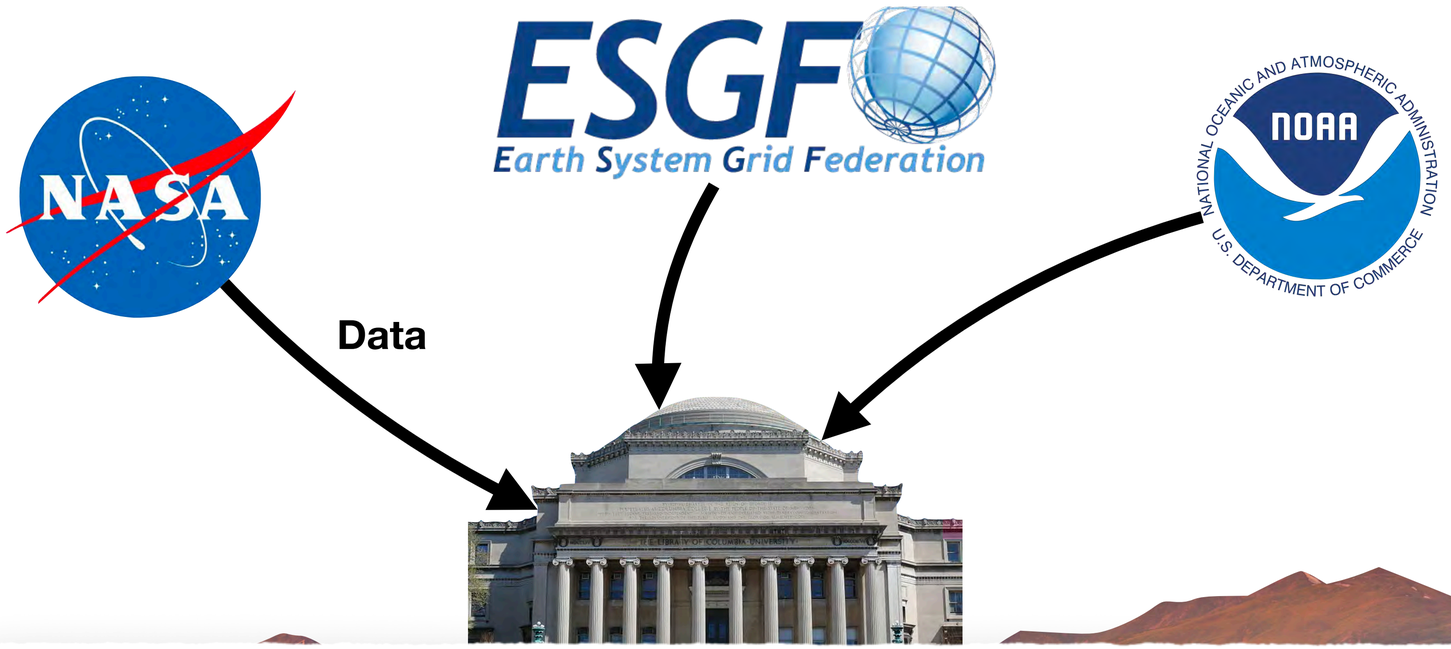
PANGEO

PRIVILEGED INSTITUTIONS CREATE "DATA FORTRESSES*"

Image credit: Moahim, CC BY-SA 4.0, via Wikimedia Commons

*Coined by Chelle Gentemann

PRIVILEGED INSTITUTIONS CREATE "DATA FORTRESSES*"




```
# step 1: open data
open("/some/random/files/on/my/cluster")
# step 2: do mind-blowing AI!
```

Image credit: Moahim, CC BY-SA 4.0, via Wikimedia Commons

*Coined by Chelle Gentemann

PROBLEMS WITH THE STATUS QUO

🗂️ Emphasis on *files* as a medium of data exchange creates lots of work for individual scientists (downloading, organizing, cleaning). Most file-based datasets are a mess—even simulation output.

😫 Yet the hard work of data wrangling is rarely collaborative. Outputs are not reusable and can't be shared.

💰 Doing data-intensive science requires either expensive local infrastructure or access to a big agency supercomputer. *This really limits participation.*

🏰 Data intensive science is locked inside *data fortresses*. Limiting access to outsiders is a feature, not a bug. *Restricts collaboration and reproducibility!*

❄️ Each fortress is a special snowflake. Code developed in one will not run inside another.

PART II: CLOUD COMPUTING AND PANGEO

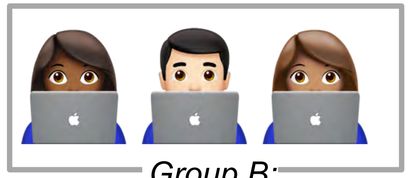
PANGEO

WHAT ABOUT CLOUD?

Can we create a “data watering hole”* instead of a fortress?

Research

Group A:

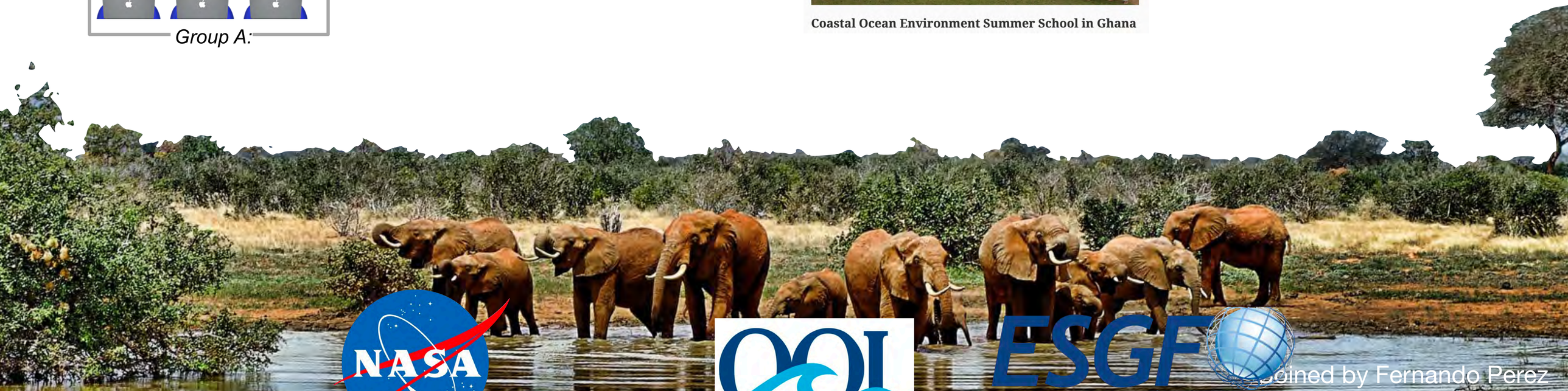


Group B:

Education & Outreach

Coastal Ocean Environment Summer School in Ghana

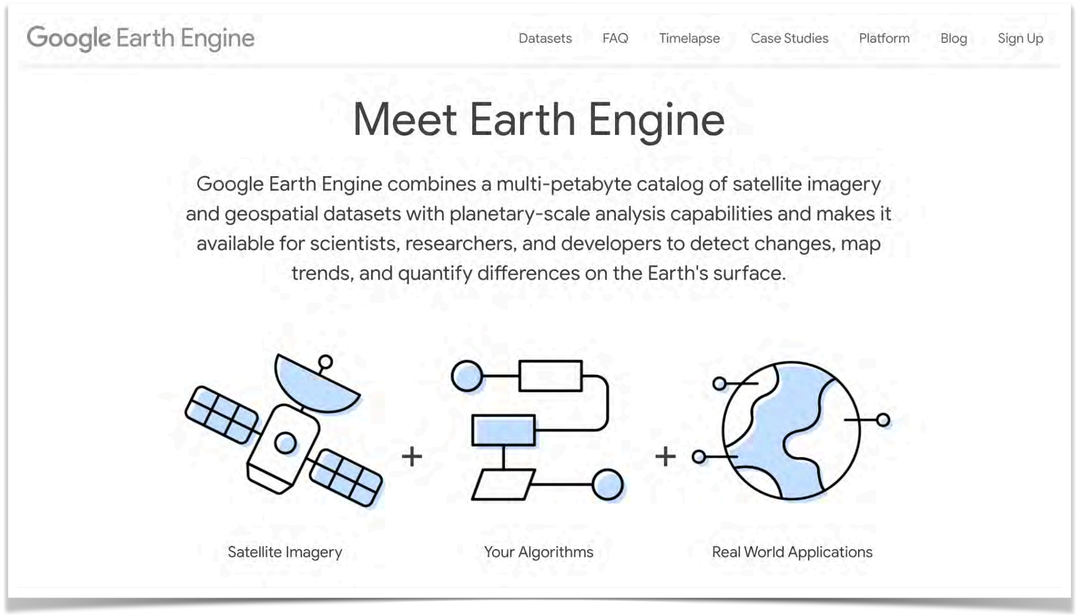
Industry Partners

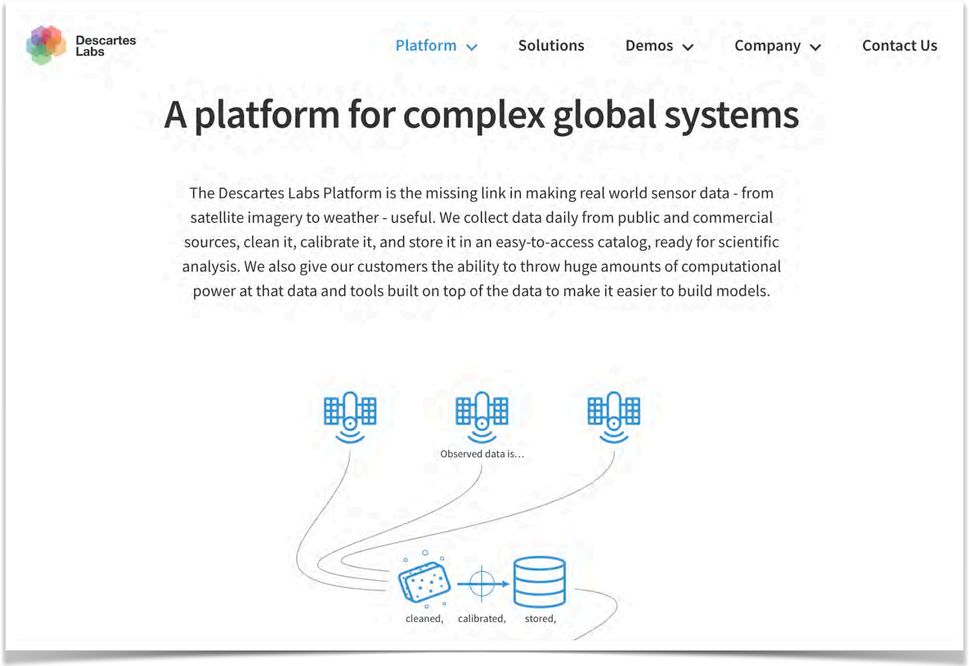


*Coined by Fernando Perez

PANGEO

OPTION A: VERTICALLY INTEGRATED PLATFORM





All the data
All the compute

PANGEO

OPTION B: INTEROPERABLE CLOUD-NATIVE DATA, SOFTWARE, AND SERVICES

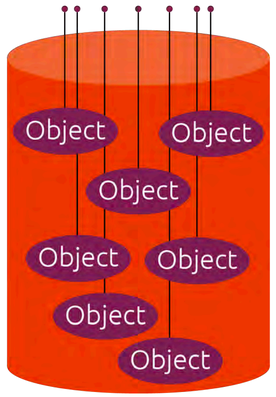
Community-Maintained ARCO Data Lake[s]

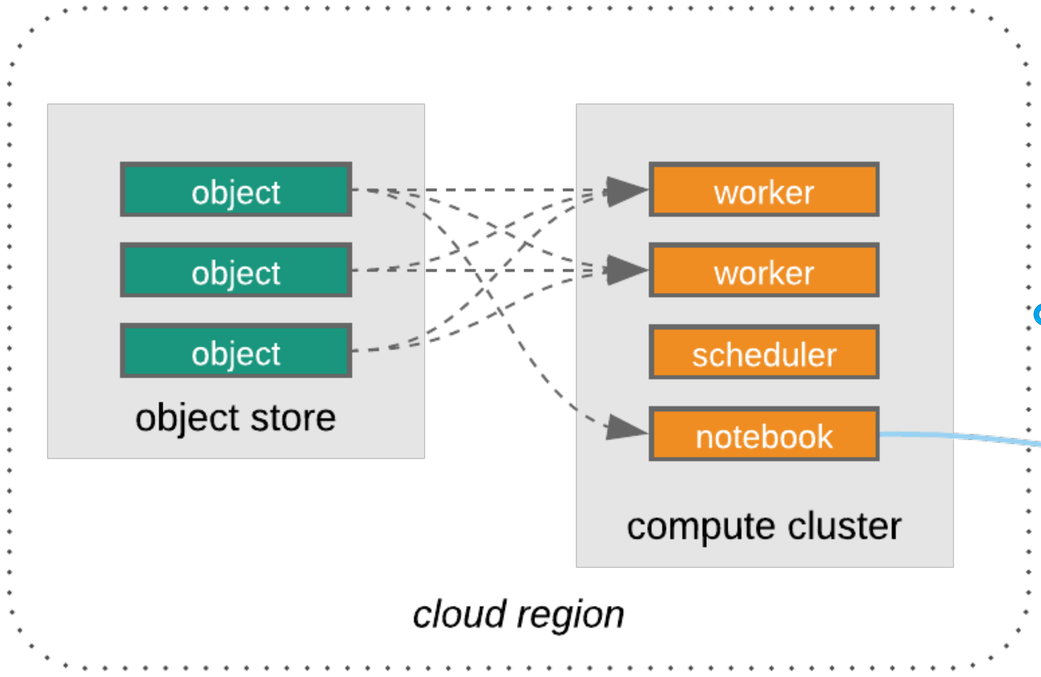
Parquet

COG CLOUD OPTIMIZED GEOTIFF

Zarr

[tile]DB





Interactive Computing

jupyter julia R

Distributed Processing

RAY DASK APACHE Spark

Data Provider's Resources

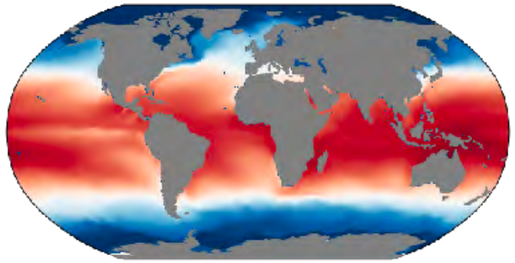
Google Cloud Platform aws Microsoft Azure

Data Consumer's Resources

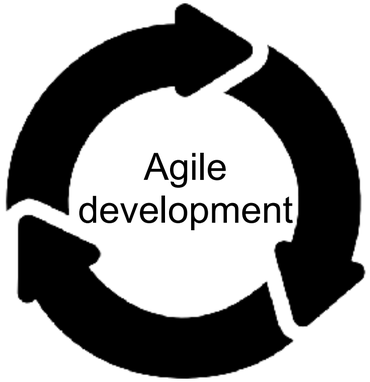
PANGEO

THE PANGEO COMMUNITY PROCESS

Scientific users / use cases



- Define science questions
- Use software / infrastructure
- Identify bugs / bottlenecks
- Provide feedback to developers



Open-source software libraries

jupyter zarr. xarray DASK

- Contribute widely the the open source scientific python ecosystem
- Maintain / extend existing libraries, start new ones reluctantly
- Solve integration challenges

GitHub

HPC and cloud infrastructure

- Deploy interactive analysis environments
- Curate analysis-ready datasets
- Platform agnostic

aws

ACCIDENTAL PIVOT TO CLOUD

Uz, Baris M <bmuz@nsf.gov> Mon, Jun 26, 2017, 7:01 PM

to me, kpaul@ucar.edu

Dear Ryan and Kevin,
I am delighted to give you good news about your recent EarthCube proposal. Based on supportive reviews and our excitement about the proposal, we are prepared to make a funding recommendation for it. The total budget in your proposal was for $1.46M. Unfortunately, it looks like we are only going to be able to allocate $1.2M to this project. This would be about an 18% reduction in the total budget. I would like to know how much of an impact this would have in the scope and/or chance of success of your project.

Ryan Abernathey <rpa@ldeo.columbia.edu> Mon, Jun 26, 2017, 8:13 PM

to Baris, kpaul@ucar.edu

Hi Mete,

Thanks for this excellent news. We would definitely like to rebudget and accept the $1.2M allocation. I will confer with my co-I's about the best way to accomplish this.

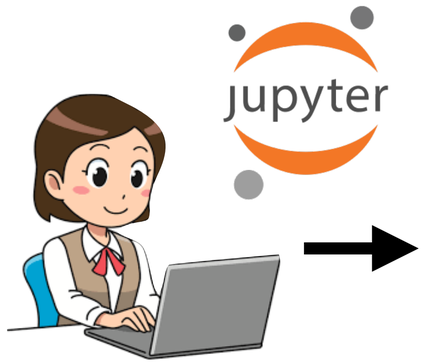
In the meantime, I have a few questions:
- What is the time frame for us to submit the revised budget?
- Do you have any specific suggestions about how to rebduget, perhaps informed by feedback from the reviews?
- Is there any chance of us getting access to commercial cloud computing resources via NSF (as with the BIGDATA program https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767)? That could help reduce our computing budget.

Thanks,

Ryan Abernathey

PANGEO

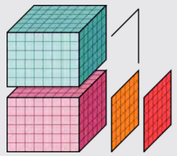
THE PANGEO CLOUD-NATIVE STACK



Rich interactive computing environment in the web browser.

Domain specific packages

xgcm | xrft | xhistogram | climpred | gcm-filters | Etc.



xarray

High-level API for analysis of multidimensional labelled arrays.

DASK

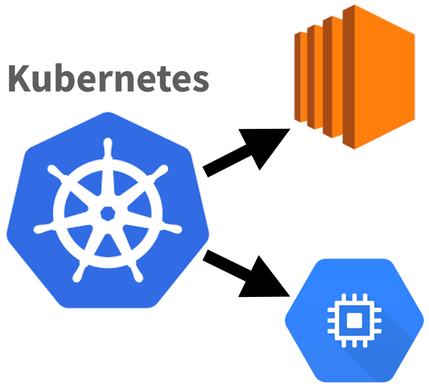
Flexible, general-purpose parallel computing framework.

Zarr

Cloud-optimized storage for multidimensional arrays.

Cloud Infrastructure

Kubernetes



Object Storage

PANGEO

ANALYSIS-READY, CLOUD OPTIMIZED: ARCO DATA

THEME ARTICLE: JUPYTER IN COMPUTATIONAL SCIENCE

Cloud-Native Repositories for Big Scientific Data

Ryan P. Abernathey, Charles C. Blackmon-Luca, Timothy J. Crone, Naomi Henderson, Chiara Lepore, *Lamont–Doherty Earth Observatory of Columbia University, Palisades, NY, 10964, USA*

Tom Augspurger, *Anaconda, Des Moines, IA, 50313, USA*

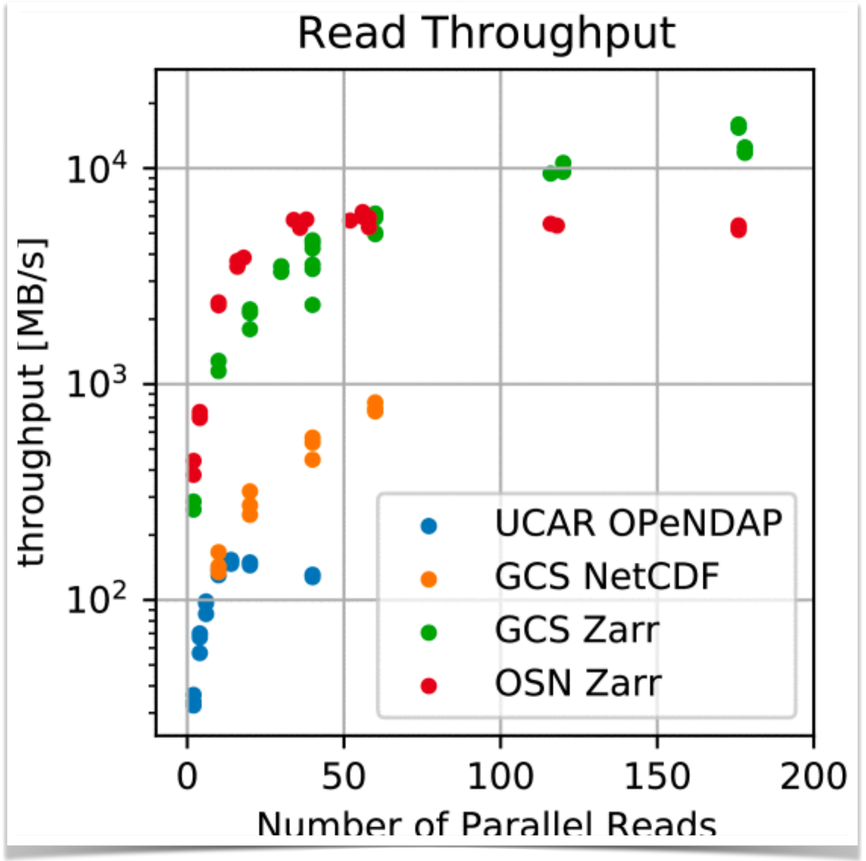
Anderson Banihirwe and Joseph J. Hamman, *National Center for Atmospheric Research, Boulder, CO, 80305, USA*

Chelle L. Gentemann, *Farallon Institute, Petaluma, CA, 94952, USA*

Theo A. McCaie and Niall H. Robinson, *Met Office, University of Exeter, Exeter EX4 4PY, U.K.*

Richard P. Signell, *US Geological Survey, Woods Hole, MA, 02543, USA*

https://doi.org/10.1109/MCSE.2021.3059437



This also demonstrates the potential of the “hybrid cloud” model with OSN.

PANGEO

PANGEO HAS BROAD ADOPTION

pangeo Retweeted

Andrew Pauling
@andrewp109

#cmip6hack is just wrapping up, and has changed the way I will think about, and hopefully do, climate model analysis in the future. The @pangeo_data infrastructure makes it all so easy.

5:29 PM · Oct 18, 2019 · Twitter Web App

6 Retweets **24** Likes

Erin Dougherty
@edougherty_

My #dayofscience: paper revisions 🙃 and using @pangeo_data to analyze massive amounts of high-res climate data to understand floods in a current and future climate over the U.S. When not doing this, I ❤️ observing #wx directly via field work and watching storms. ⛈️

2:38 PM · Oct 15, 2019 · Twitter for iPhone

5 Retweets **51** Likes

October 20, 2021 (v1) | Presentation | Open Access

Scalable Sustainability with the Planetary Computer

Microsoft

View

Augspurger, Tom; Morris, Dan; Emanuele, Rob; McFarland, Matt; Sanchez-Andrade Nuno, Bruno;

We'll see how Microsoft's Planetary Computer team has embraced the ideas and techniques pioneered by pangeo, putting them to work for sustainability. We'll see the use of STAC for cataloging and querying data, analysis-ready data in blob storage, and JupyterHub-based environments for sca

Uploaded on October 25, 2021

November 3, 2021 (v1) | Presentation | Open Access

xcube - Python package for Earth Observation data cubes

esa
European Space Agency

View

Brandt, Gunnar; Norman Fomferra;

xcube is a software package and toolkit for generation and exploitation of analysis-ready data cubes from Earth Observation and other geographical data. It is based on Python's popular data science technology stack, particularly on xarray, dask, and zarr, and is freely available as open-source s

Uploaded on November 11, 2021

LIMITATIONS OF THE PANGEO APPROACH

- Pangeo *software* can be deployed as a *platform*:
  JupyterHub in the cloud with Xarray, Dask, etc., connected to ARCO data sources

- But there are *many distinct deployments of this platform* - dozens of similar yet distinct JupyterHubs with different configurations, environments, capabilities, etc
  ➡️ Sharing projects between these hubs is still very hard

- Deploying hubs generally requires DevOps work (billed a developer time or contractor services). There is no “Pangeo as a Service”

- Getting data into the cloud in ARCO format is hard and full of toil

CHALLENGES WITH CLOUD IN GENERAL

- Non-agency scientists have many barriers to adopting cloud:
  Overhead policies, purchasing challenges, lack of IT support, etc.

- Cloud is too complicated! The services offered are not useful to scientists:
  An extra layer of science-oriented services must be developed

- Europe basically forbids scientists from using US-based cloud providers

- Not much has changed for university scientists since 2017

PART III: FROM SOFTWARE TO SAAS

PANGEO FORGE & EARTHMOVER

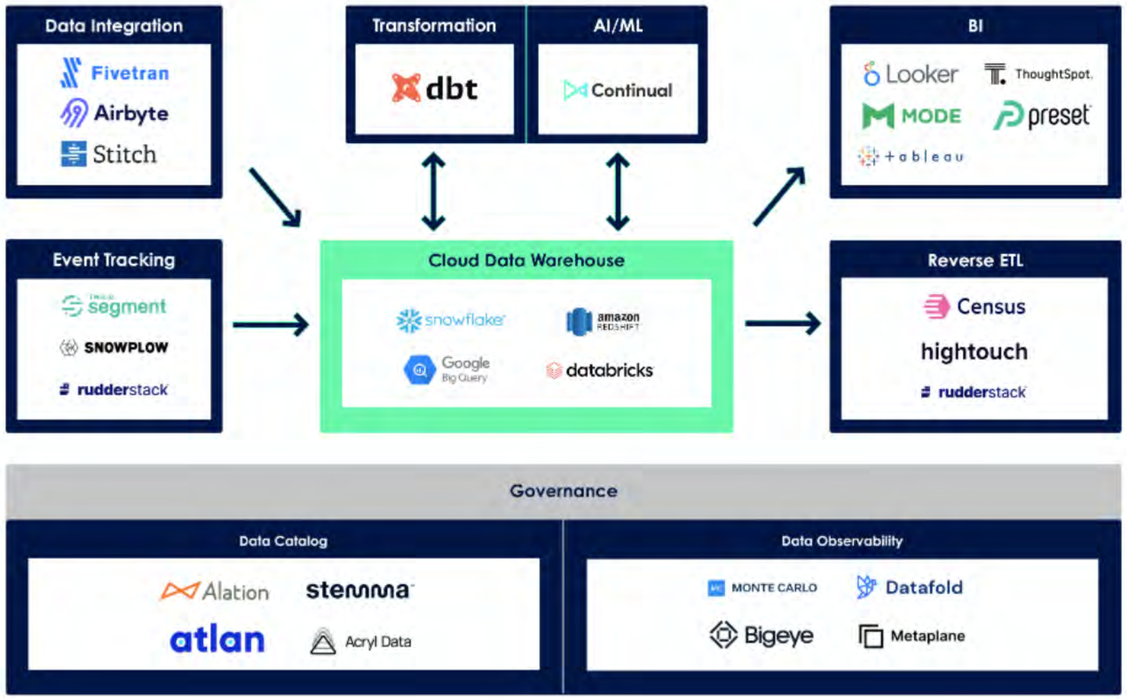
TOOLS FOR COLLABORATION

Some of the most impactful services used in open science….

These are all proprietary SaaS (Software as a Service) applications.
They may use open standards, but they are *not* open source.

We (or our institutions) have no problem paying for them.

THE MODERN DATA STACK



https://continual.ai/post/the-modern-data-stack-ecosystem-fall-2021-edition

- In the past 5 years, a platform has emerged for *enterprise data science* called the *Modern Data Stack*
- The MDS is centered around a “data lake” or “data warehouse”
- Different platform elements are provided by different SaaS companies; integration through standards and APIs
- No one in science uses any of this stuff

HOW CAN WE DELIVER AN OPEN SCIENCE PLATFORM IN A SCALABLE, SUSTAINABLE WAY?

- Embrace commercial SaaS: a Modern Data Stack for Science
- Cultivate community-operated SaaS: e.g. Wikipedia, Conda Forge, Binder, 2i2c Hubs, Pangeo Forge
- *We probably need a mix of both*

PANGEO FORGE

Community-operated SaaS for ETL (Extract / Transform / Load) of ARCO Data

EARTHMOVER
THE CLOUD DATA STACK FOR SCIENCE

Our new startup. Building a commercial **cloud data lake platform** for scientific data.

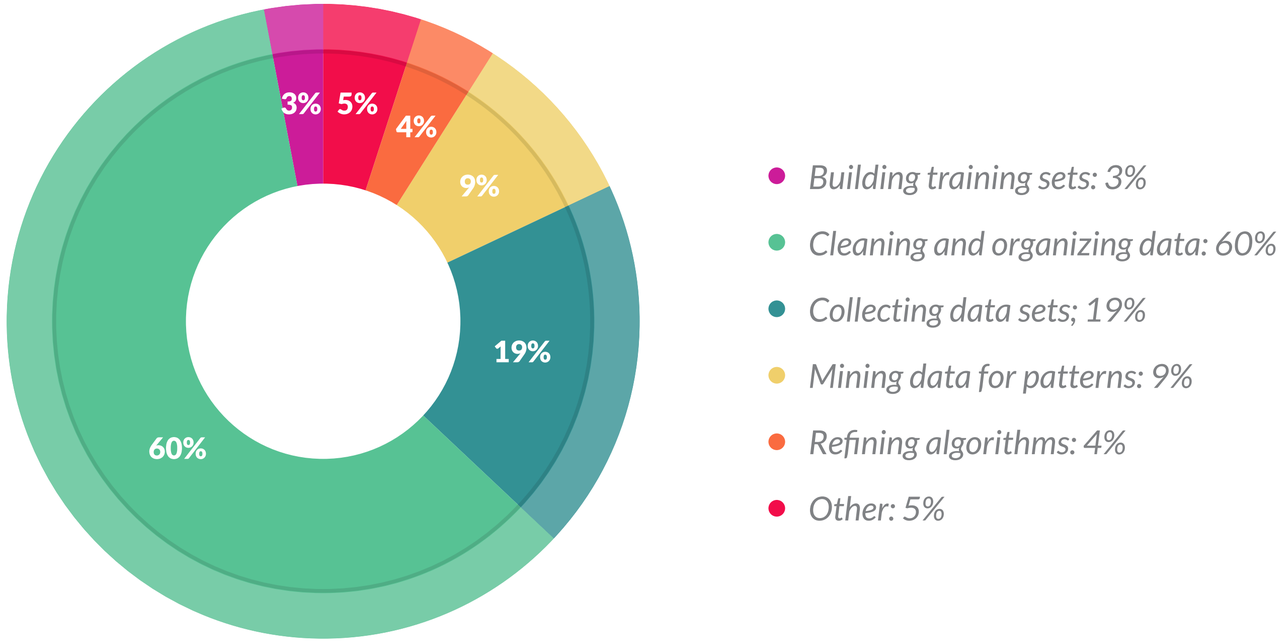
PANGEO

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



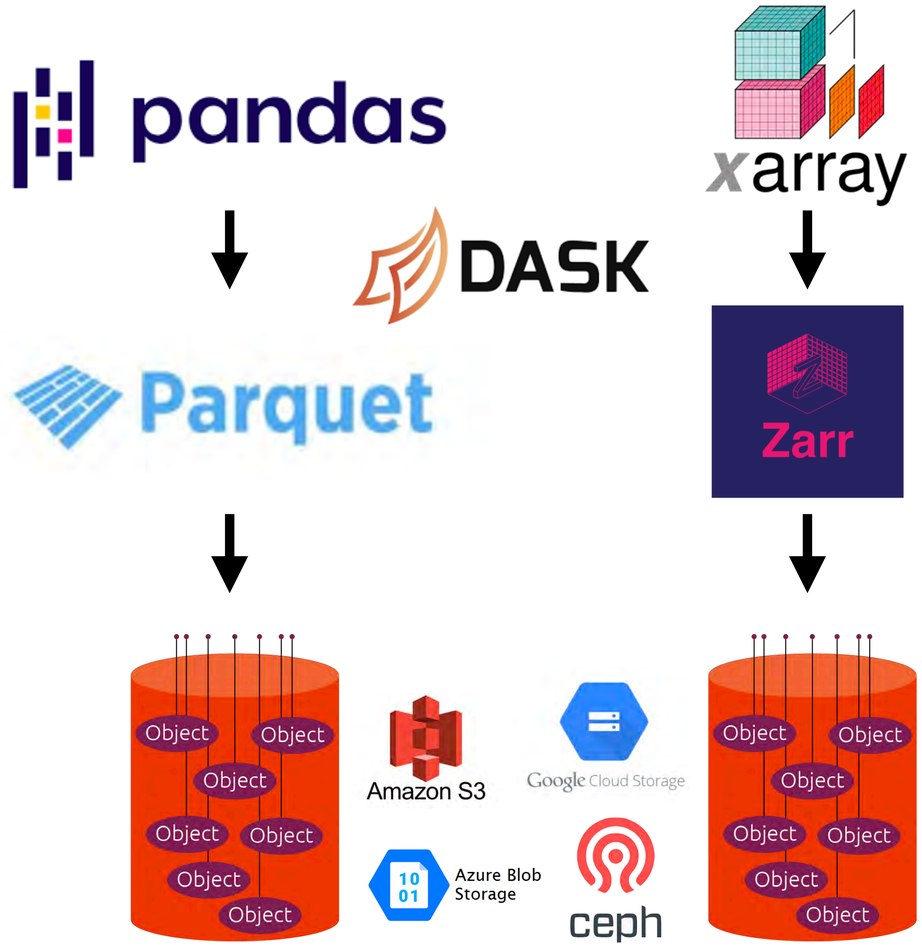
How do data scientists spend their time?
Crowdflower Data Science Report (2016)

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

ARCO DATA IS FAST!

THEME ARTICLE: JUPYTER IN COMPUTATIONAL SCIENCE

Cloud-Native Repositories for Big Scientific Data

Ryan P. Abernathey, Charles C. Blackmon-Luca, Timothy J. Crone, Naomi Henderson, Chiara Lepore, *Lamont–Doherty Earth Observatory of Columbia University, Palisades, NY, 10964, USA*

Tom Augspurger, *Anaconda, Des Moines, IA, 50313, USA*

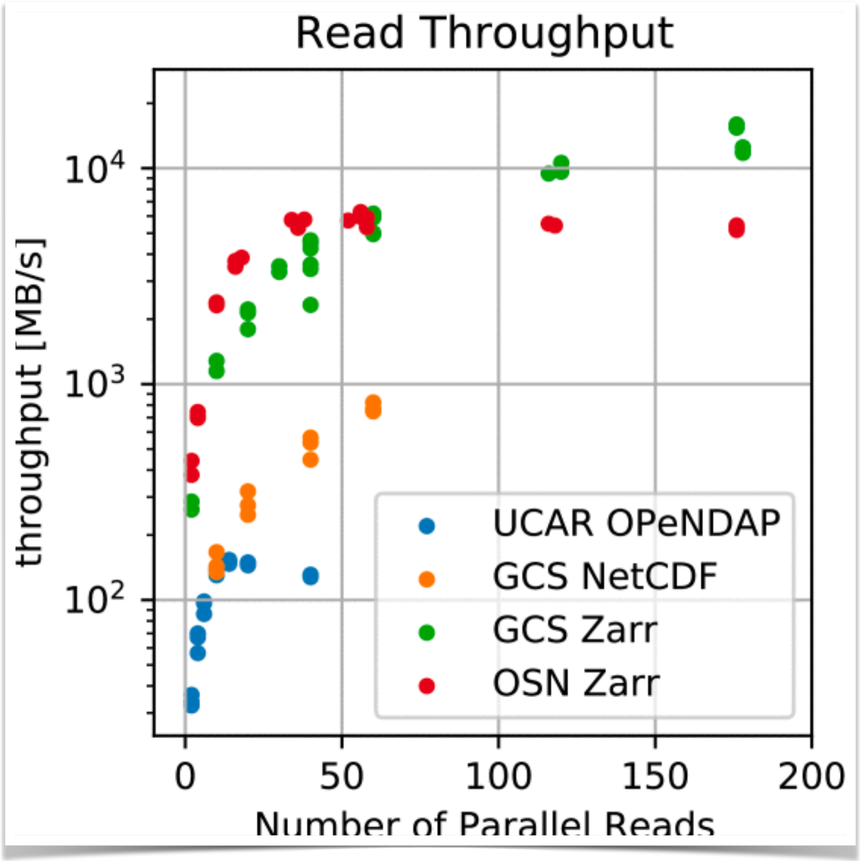
Anderson Banihirwe and Joseph J. Hamman, *National Center for Atmospheric Research, Boulder, CO, 80305, USA*

Chelle L. Gentemann, *Farallon Institute, Petaluma, CA, 94952, USA*

Theo A. McCaie and Niall H. Robinson, *Met Office, University of Exeter, Exeter EX4 4PY, U.K.*

Richard P. Signell, *US Geological Survey, Woods Hole, MA, 02543, USA*

https://doi.org/10.1109/MCSE.2021.3059437



This also demonstrates the potential of the “hybrid cloud” model with OSN.

PANGEO

PROBLEM:

Making ARCO Data is Hard!

To produce useful ARCO data, you must have:

Domain Expertise:	*Tech Knowledge:*	*Compute Resources:*	*Analysis Skills:*
How to find, clean, and homogenize data	How to efficiently produce cloud-optimized formats	A place where to stage and upload the ARCO data	To validate and make use of the ARCO data.

Data Scientist

WHOSE JOB IS IT TO MAKE ARCO DATA?

Data providers are concerned with preservation and archival quality.

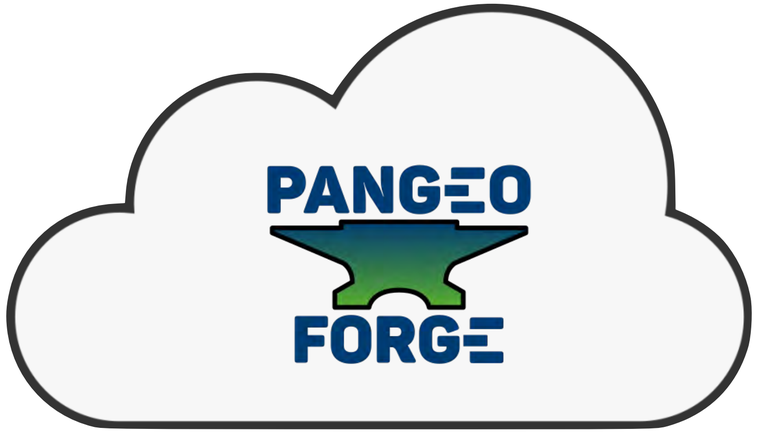
Scientists users know what they need to make the data analysis-ready.

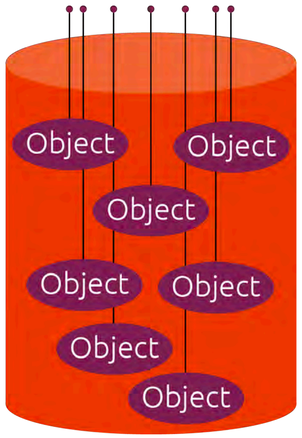
PANGEO FORGE

Let’s democratize the production of ARCO data!

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

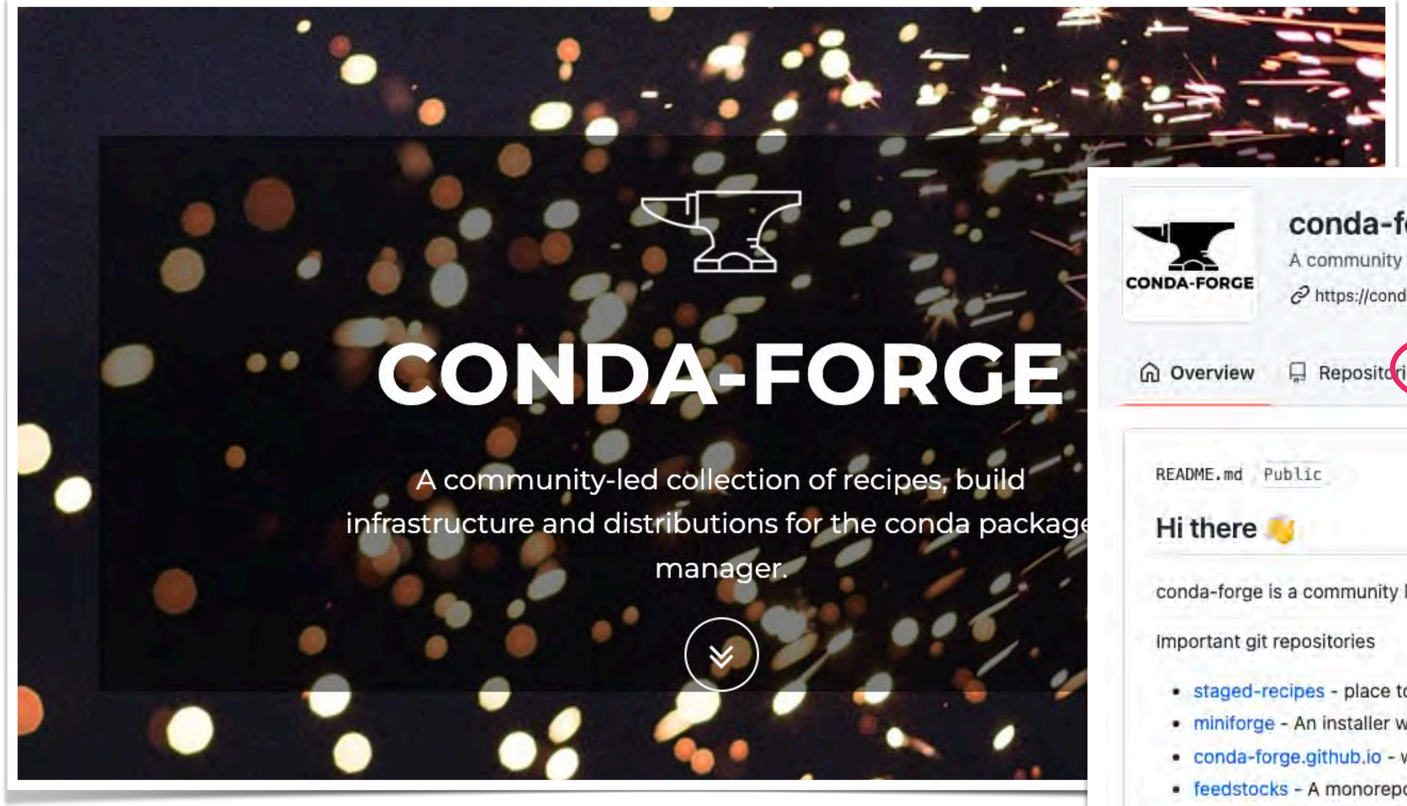
Data Scientist

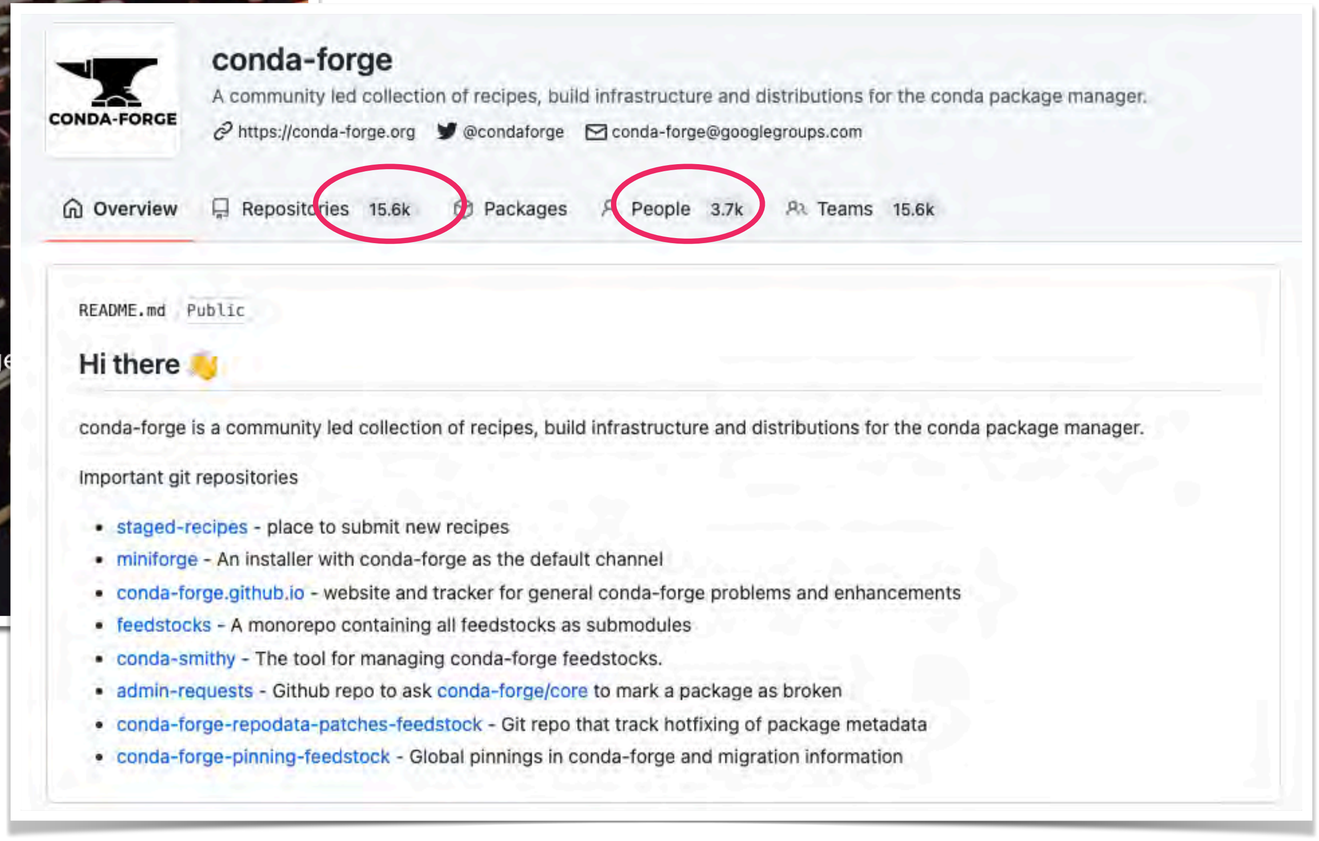




PANGEO

INSPIRATION: CONDA FORGE

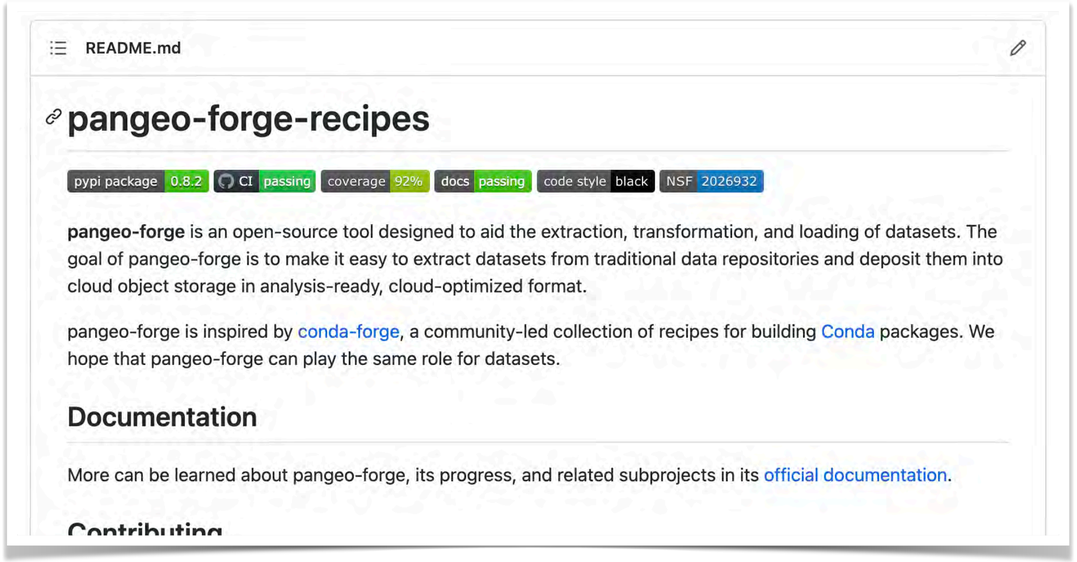




PANGEO FORGE

Pangeo Forge Recipes

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



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

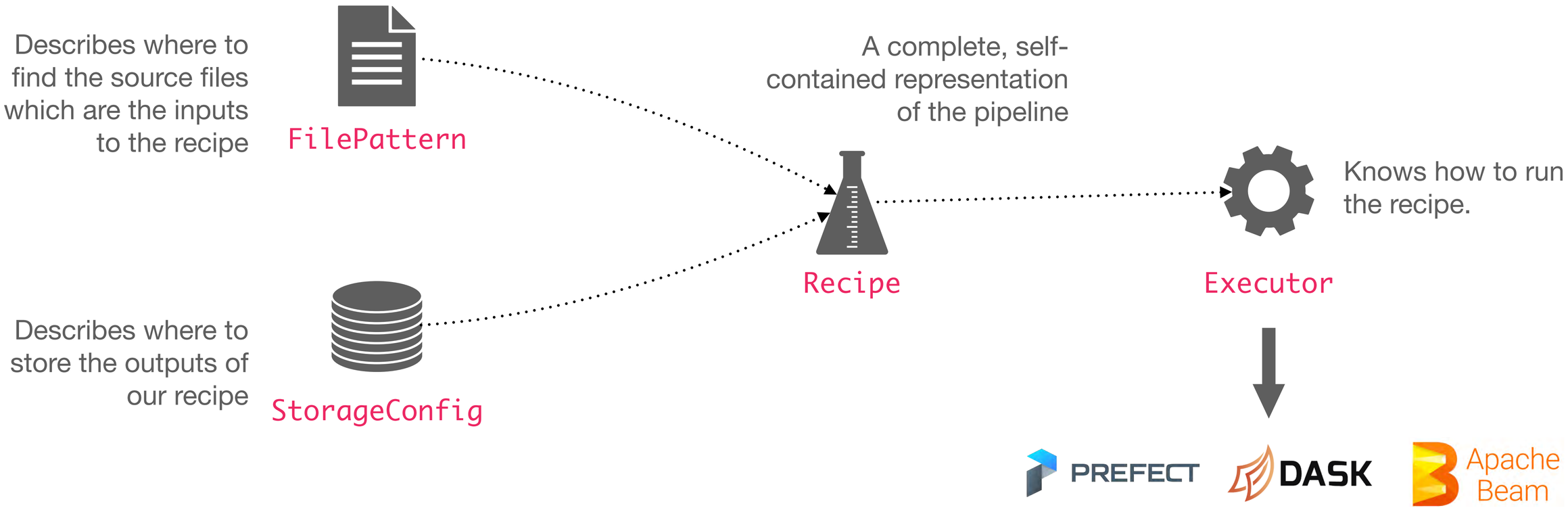
Pangeo Forge Cloud

https://pangeo-forge.org/

Cloud platform for automatically executing recipes stored in GitHub repos.

PANGEO FORGE RECIPES

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



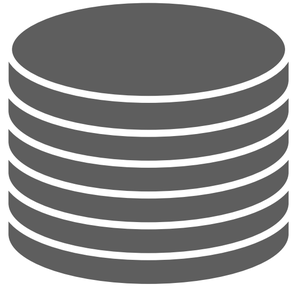
PANGEO FORGE CLOUD

https://pangeo-forge.org/

Contains the code and metadata for one or more Recipes

Feedstock

Runs the recipes in the cloud using elastic scaling clusters



Storage

Amazon S3

GCS

GitHub

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

VISION: COLLABORATIVE DATA CURATION

Feedstock

GitHub

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

Data User

These data look weird…

…Oh the metadata need an update.

Data Producer

Data Manager

Ok I'll make a PR to the recipe.

Charles Stern

Oceanographers building a full-stack cloud SaaS automation platform.

https://twitter.com/Colinoscopy/status/1255890780641689601

CURRENT STATUS

🙌 Pangeo Forge Cloud is live and open for business!
pangeo-forge.org

💣 Our recipes often contain 10000+ tasks. We are hitting the limits on Prefect as a workflow engine. Currently refactoring to move to Apache Beam.

😫 Data has *lots of edge cases!* This is really hard.

🌎 But we remain very excited about the potential of Pangeo Forge to transform how scientists interact with data.

EARTHMOVER

A Public Benefit Corporation

Mission: *To empower people to use scientific data to solve humanity's greatest challenges*

Founders:

Ryan Abernathey

Joe Hamman

Product: ArrayLake

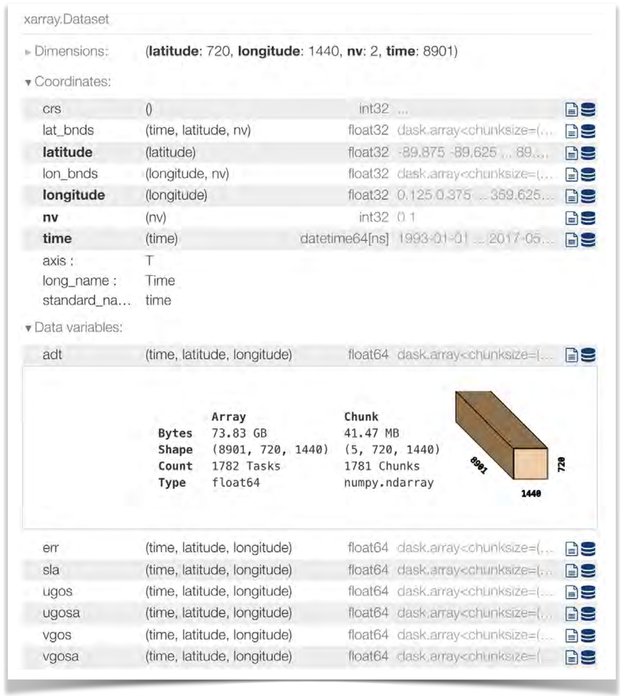
- High performance for analytics (based on Zarr data model)
- Ingest and index data from archival formats (NetCDF, HDF, GRIB, etc.)
- Automatic background optimizations
- Versioning / snapshots / time travel
- Data Governance
- Compare to Databricks, Snowflake, Dremio

PART IV: WHERE ARE WE HEADING?

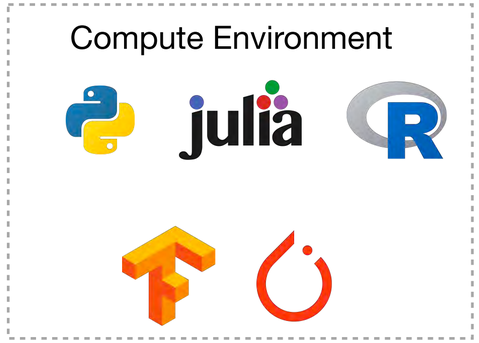
PANGEO

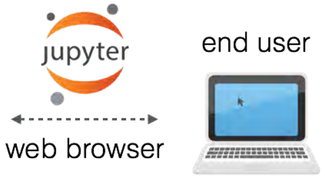
PILLARS OF CLOUD NATIVE SCIENTIFIC DATA ANALYTICS

1. Analysis-Ready, Cloud-Optimized Data

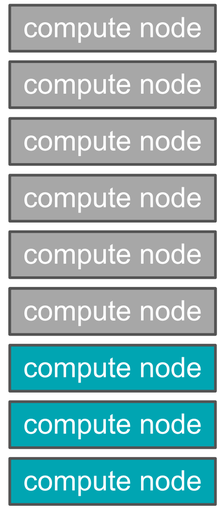


2. Data-Proximate Computing





3. Elastic Distributed Processing



RAY DASK APACHE Spark

PANGEO

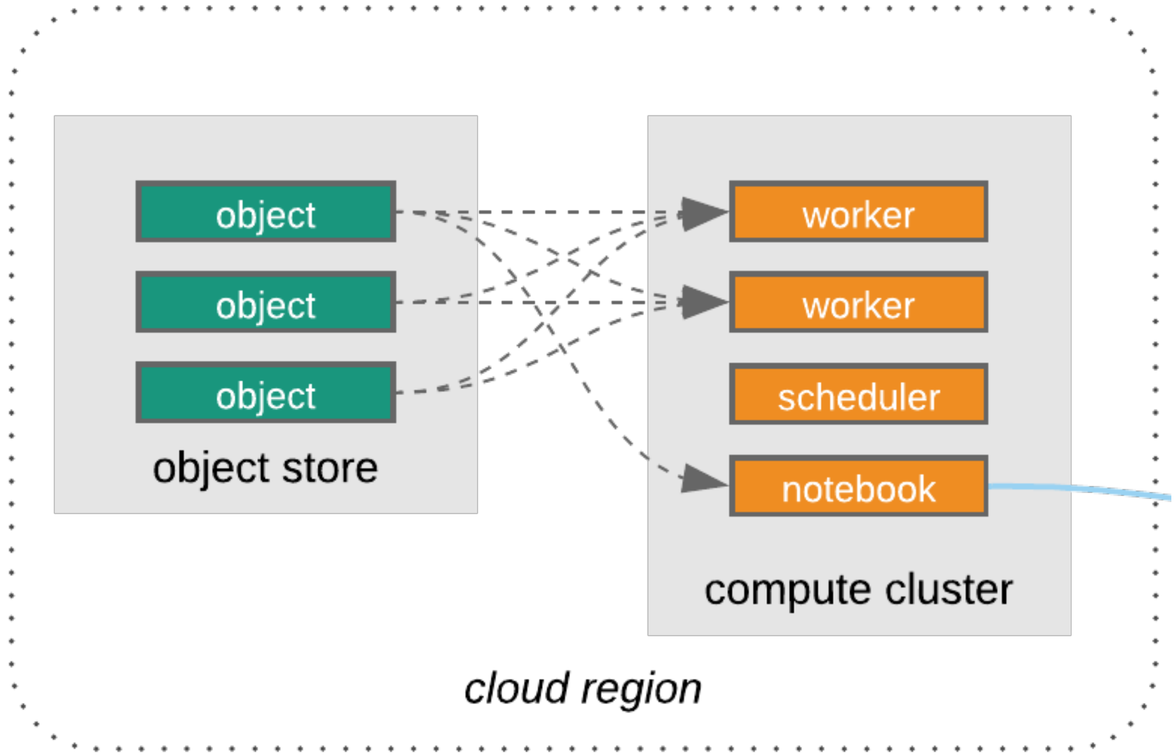
SEPARATION OF STORAGE AND COMPUTE

Storage costs are steady.

Data provider pays for storage costs.

May be subsidized by cloud provider.
(*Thanks AWS, GCP, Azure!*

Or can live outside the cloud (e.g. Wasabi, OSN)



Compute costs for interactive data analysis are bursty.

Can take advantage of *spot pricing*

Multi-tenancy:
We can all use the same stack, but each institution pays for its own users.

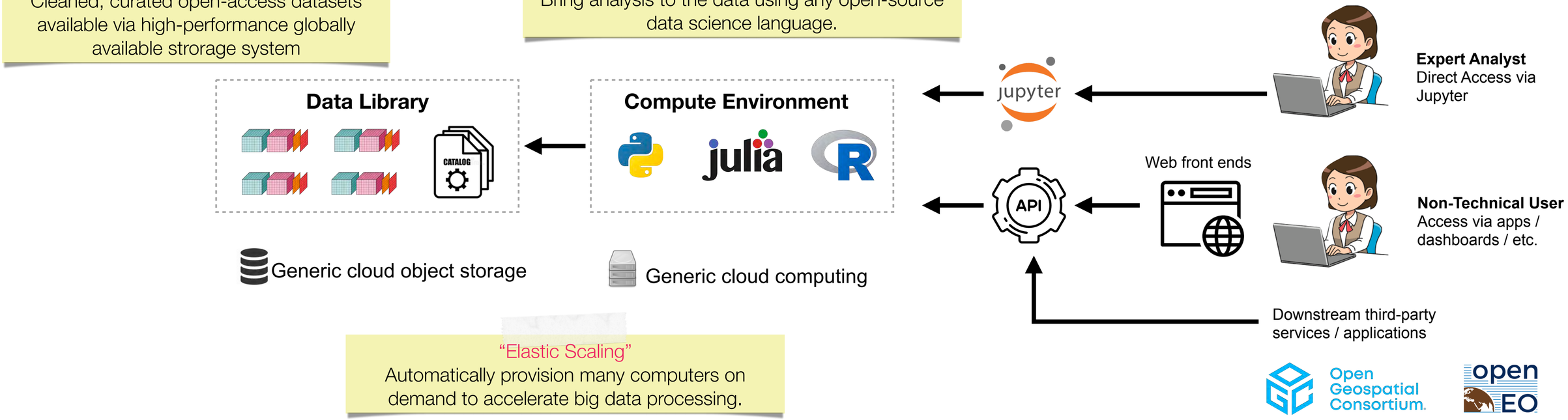
This is completely different from the status quo infrastructure!

PANGEO

OPEN SCIENCE PLATFORM

"Analysis Ready, Cloud Optimized Data"
Cleaned, curated open-access datasets available via high-performance globally available strorage system

"Data Proximate Computing"
Bring analysis to the data using any open-source data science language.



"Elastic Scaling"
Automatically provision many computers on demand to accelerate big data processing.

Open Geospatial Consortium

open EO

Google Cloud Platform

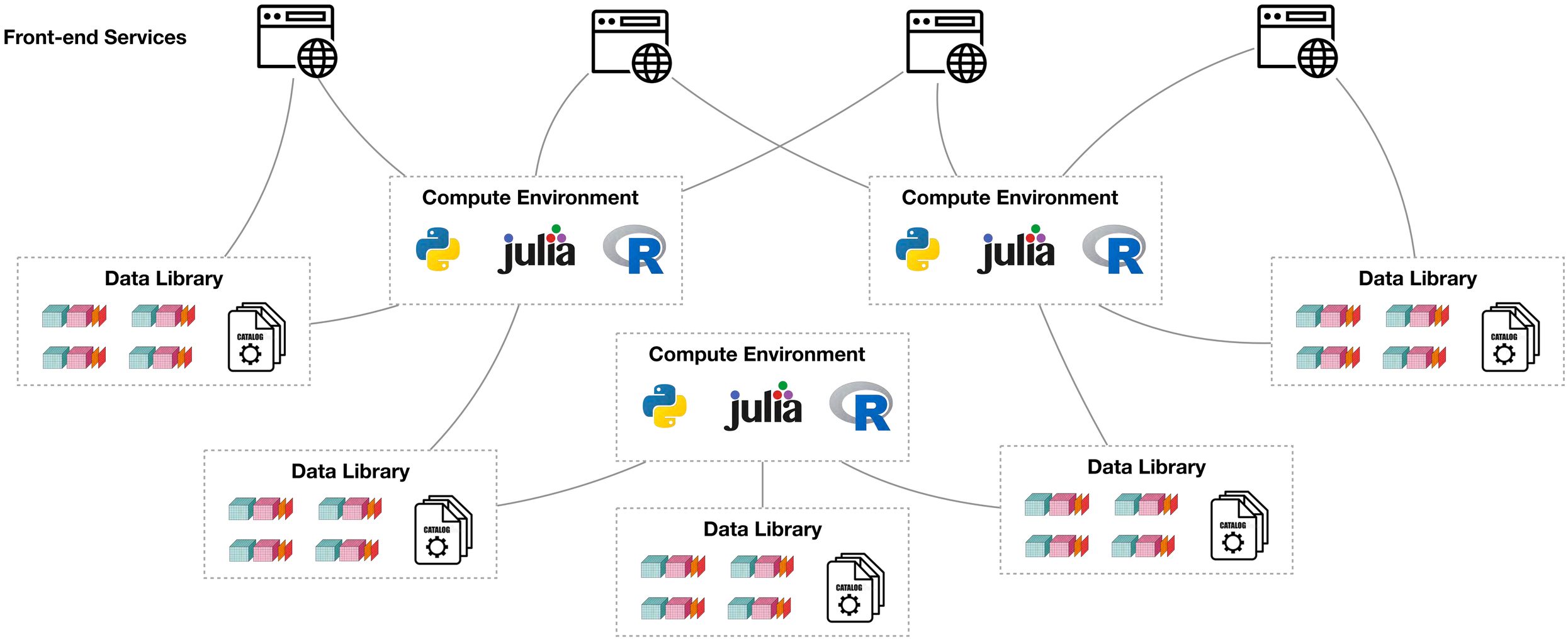
Microsoft Azure

aws

OVHcloud

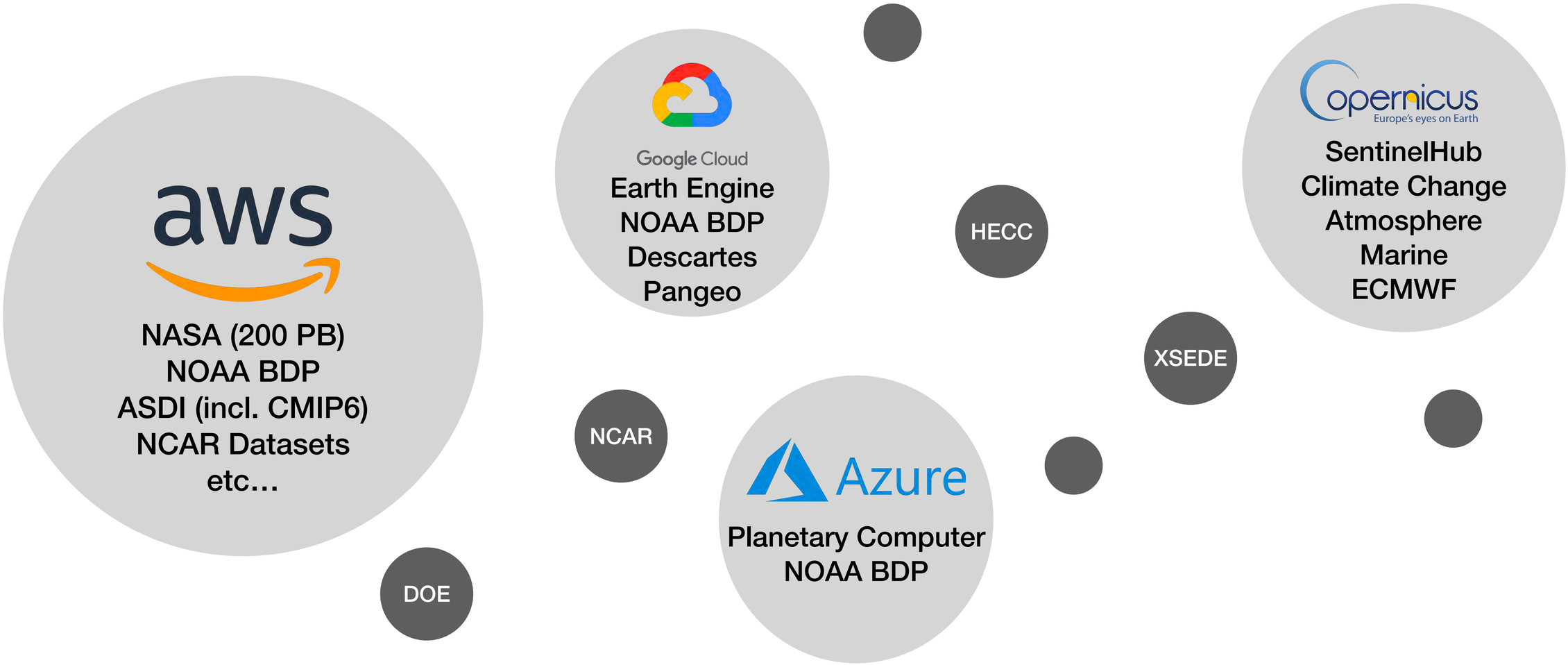
Runs on any modern cloud-like platform or on premises data center

FEDERATED, EXTENSIBLE MODEL



DATA GRAVITY

“Data gravity is the ability of a body of data to attract applications, services and other data." - Dave McCrory

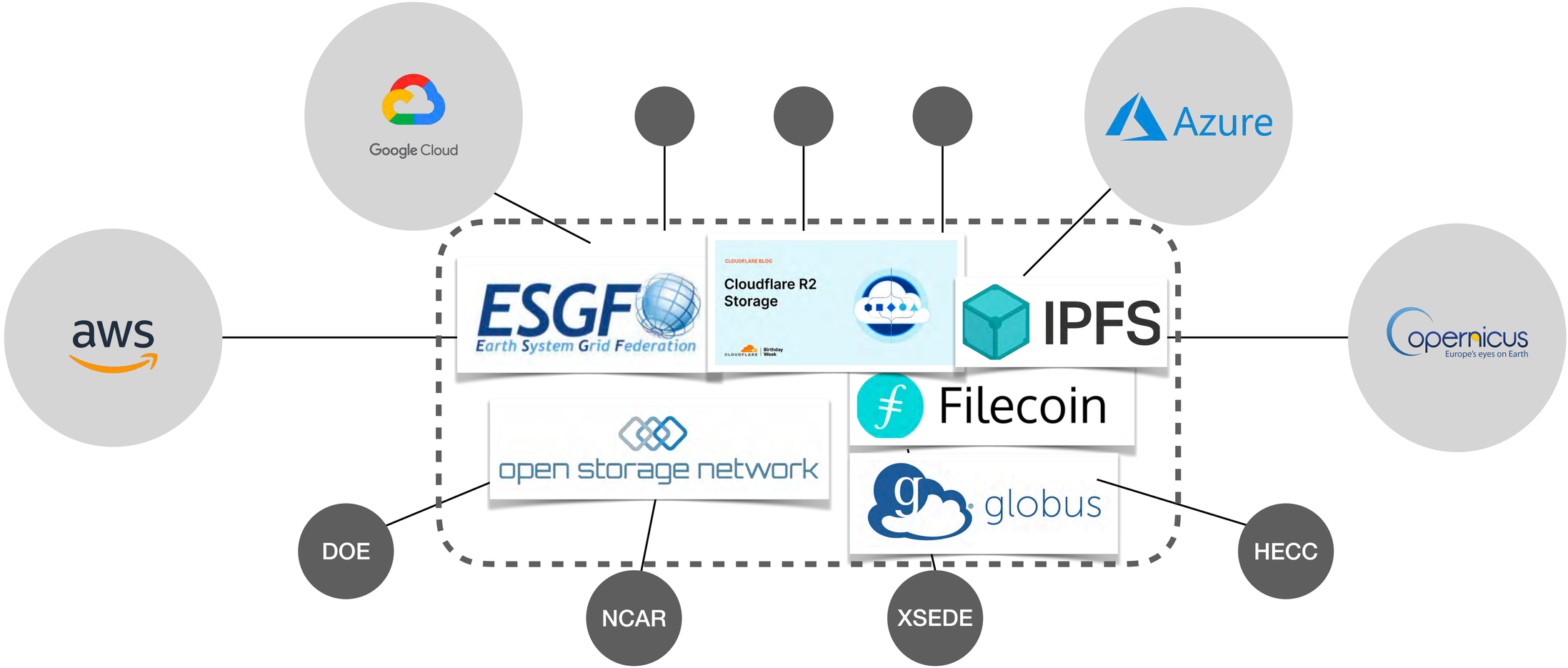


DATA GRAVITY

What is the stable steady-state solution?

WE NEED A GLOBAL SCIENTIFIC DATA COMMONS

Need to be exploring: edge storage, decentralized web, web3



SHOUT OUTS

globus labs

2i2c
the international interactive computing collaboration

SGCI Science Gateways Community Institute

Parsl

FOUNDRY
DATA, MODELS, SCIENCE

SciServer

funcX

jupyter

CYVERSE

binder

Jetstream2

open storage network

ESGF
Earth System Grid Federation

unidata

ESIP

DISCUSSION QUESTIONS

- What's the right model to deliver data and computing services to the research community? Commercial vendors? Co-ops?
- How can we avoid recreating existing silos in the cloud?
- Who should we pay for cloud infrastructure for the science community? University? Agency? PI?
- How can we make cloud interoperate more with HPC and on-premises computing resources?

LEARN MORE

http://pangeo.io

https://discourse.pangeo.io/

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

https://medium.com/pangeo

@pangeo_data