



the future of fusion data infrastructure

UKAEA | 2026

Tom Nicholas

Forward Engineer, Earthmover PBC





MY JOURNEY



Plasma Physicist at CCFE

PhD in **SOL turbulence modelling** (with Fulvio Militello & Ben Dudson)

Spent years wrangling BOUT++ data using Xarray



Open Source Hacker

Became a maintainer of **Xarray**, later `xarray.DataTree`, **Zarr**, **VirtualiZarr**, **Icechunk**

Met Earthmover co-founders Ryan and Joe through shared interest in OSS for science



Land-locked Oceanographer

Pivoted to **geoscience data** (weather / climate / satellite images)

Research Software Engineer in academic + non-profit Oceanography research



Real Software Engineer

Building **Earthmover** with a team of like-minded ex-researchers + industry veterans

Bringing this stack **back to fusion??**



earthmover

🕒 Our Mission

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

- Founded in 2022 as a **public benefit corporation** by **Ryan Abernathey** and **Joe Hamman** (co-founders of **Pangeo**).
- 14 people spread across US, Europe, and South America.
- "Open core" business model - investing in **community open source** and providing a **managed cloud data platform** for institutions.
- Serving clients in **private sector** (energy, commodities, insurance, AI startups), **non-profits, academia, and government**.



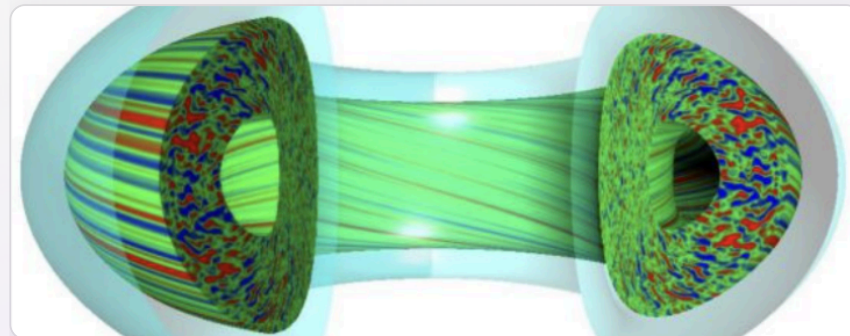
WHY THIS TRANSFERS

fusion data == geo data

Same kinds of data, with **same challenges**:

- **Big multidimensional arrays** — (time, x, y, z)
- **Simulation ensembles**, parameter scans
- **Instrument archives** — sensors and imagery: high rate, heterogeneous, append-only
- **Legacy formats** — e.g. millions of HDF5 files
- **AI/ML surrogates** trained on the *whole* archive

Geoscience community has been working on this for years, so **Fusion doesn't need to start over.**



fusion — gyrokinetic turbulence in a tokamak



geoscience — mesoscale eddies in a km-scale ocean model



IMPORTANT TRENDS IN FUSION DATA IN 2026



High-Fidelity Simulation

Gyrokinetic, MHD, and whole-device models are reaching higher and higher spatiotemporal resolution — and generating archives to match.



AI/ML Workloads

AI has a **voracious appetite** for training data. Surrogate models and control policies need **high throughput access** to **entire archives** for training and validation.



Agent-based Workflows

Query data with **natural language**, not code.

Deploy agents to explore shot databases and **perform independent research**.



Cloud Computing

Cloud computing enables any organization to access **large-scale storage and compute** without huge up-front capital cost. **Object storage** over filesystem storage.



Private Sector Participation

Commonwealth
Fusion Systems

Tokamak
Energy

First Light
Fusion

Proxima
Fusion

TAE
Technologies

Helion



Open Science

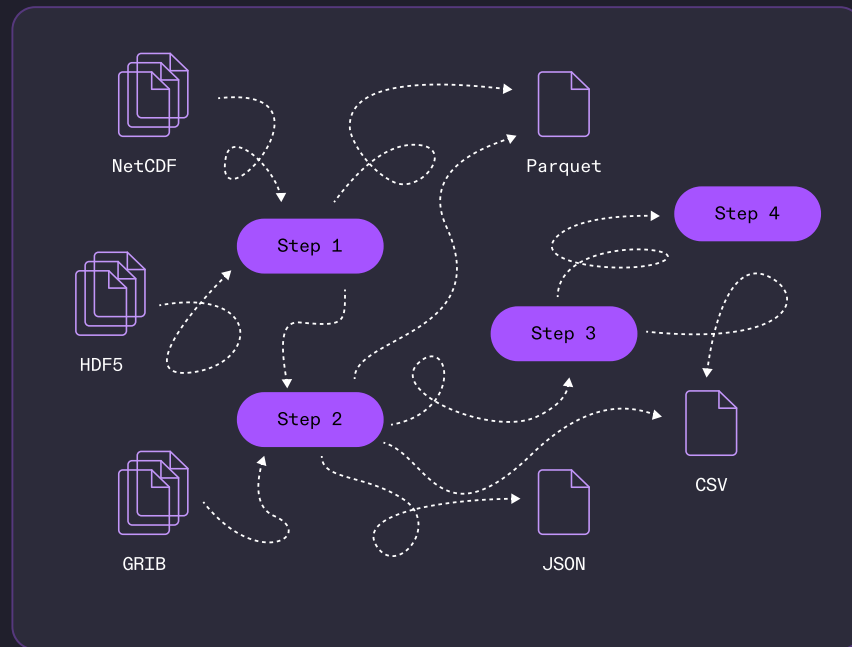
Public programs and journals increasingly value **open sharing of code and data**, even as private programs keep their own data close.



THE PROBLEM

scientific data is stuck in files

- Millions of netCDF and HDF5 files spread across HPC scratch, FTP servers, and laptops
- "Download first, analyze later" stops working once the data outgrows your disk
- No versioning and no shared source of truth:
`final_v2 REALLY final.nc`
- Every collaborator rebuilds the same pipeline from scratch





THE ALTERNATIVE

from files to arco data cubes

Humans and agents don't want to deal with thousands of individual files. They want ARCO [1] data.

Analysis Ready

All timesteps, variables, ensemble members, etc. are organized into a single logical object (e.g. a data cube).

Cloud Optimized

Applications can efficiently stream data directly from cloud object storage with low latency and high throughput.

Zarr is the most widely used ARCO data format for array data (across geoscience, bioscience, etc.).

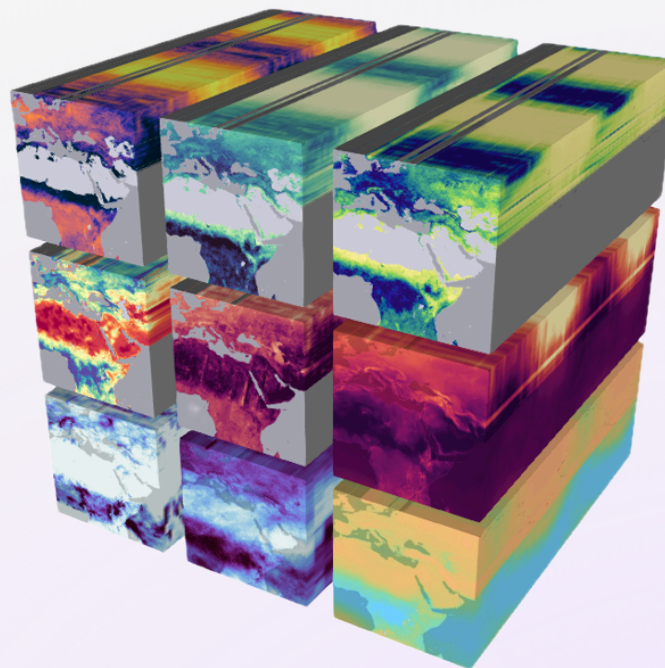


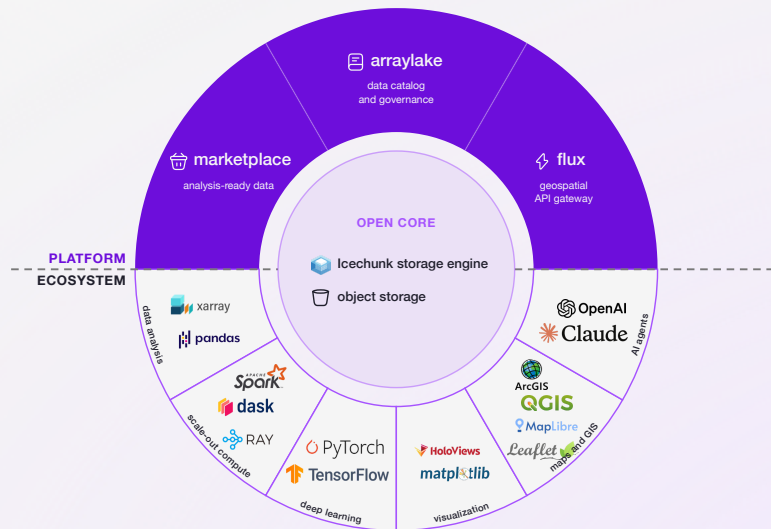
Image credit: Mahecha et al. 2020. <https://doi.org/10.5194/esd-11-201-2020>

[1] "ARCO" was coined by our CEO Ryan Abernathey in [Abernathey et al. 2021](#), "Cloud-Native Repositories for Big Scientific Data", and has since become a widely-used term.



WHAT WE'RE BUILDING

one platform, from first commit to operational service



Icechunk storage engine → Arraylake catalog → Flux maps & APIs → Marketplace



EVOLVING DATA STORAGE

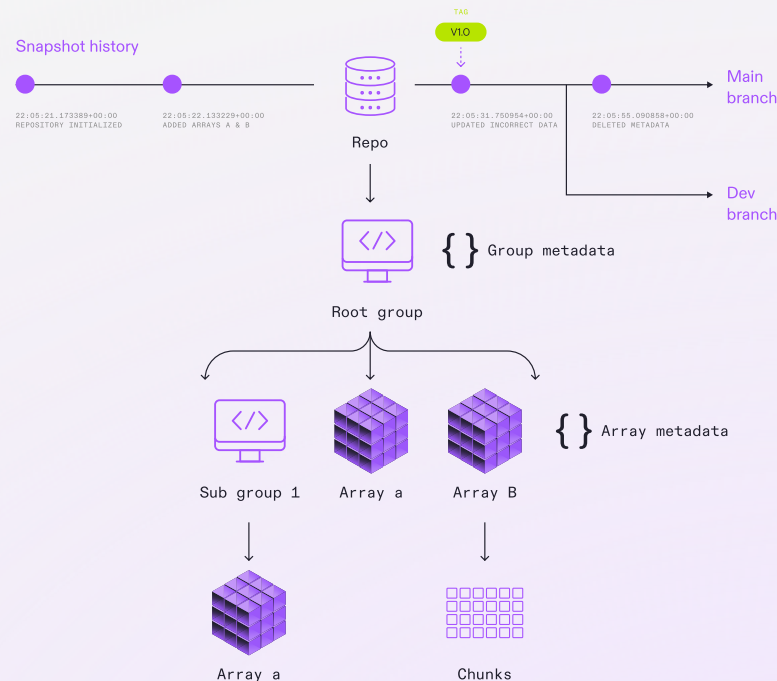
icechunk: zarr for production

A specialized **Zarr Store** which supports:

- **ACID transactions** in a distributed context
- **Version history** (snapshots, branches, tags) with optional expiration and garbage collection
- **Cheap metadata operations** (move, shift, reindex)
- **Virtual chunks** - pointers to blobs of numerical data in external files (HDF5, NetCDF, custom binary formats)
- **Open source** (Apache 2.0), **serverless**

Implemented in Rust with a **high-performance I/O layer**.

<https://icechunk.io>





EVOLVING DATA STORAGE

icechunk: zarr for production

A specialized **Zarr Store** which supports:

- **ACID transactions** in a distributed context
- **Version history** (snapshots, branches, tags) with optional expiration and garbage collection
- **Cheap metadata operations** (move, shift, reindex)
- **Virtual chunks** - pointers to blobs of numerical data in external files (HDF5, NetCDF, custom binary formats)
- **Open source** (Apache 2.0), **serverless**

Implemented in Rust with a **high-performance I/O layer**.

<https://icechunk.io>



Single-node throughput benchmarks (15 Gbps NIC)

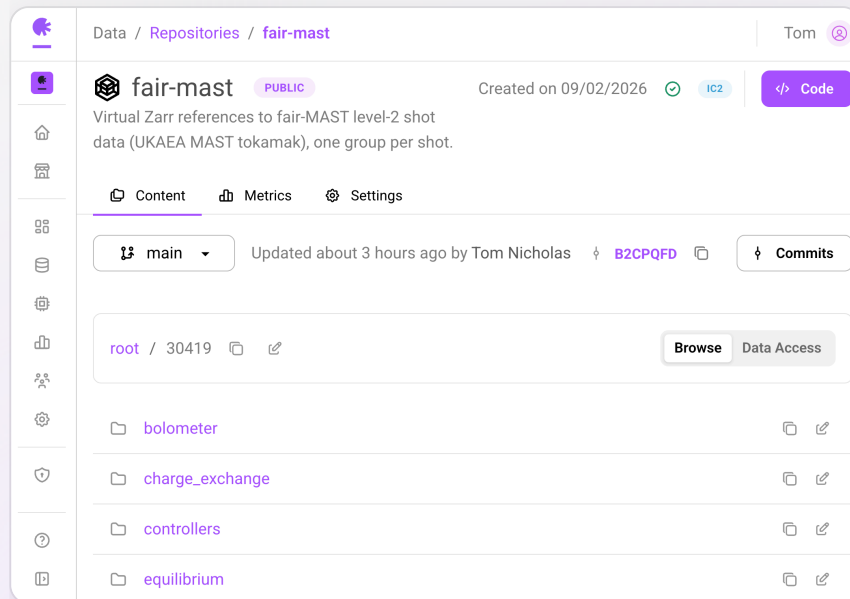
<https://www.earthmover.io/blog/i-o-maxing-tensors-in-the-cloud/>



ICECHUNK

virtual chunks: bring the archives you already have

- A virtual chunk is a lightweight pointer (url, offset, length) into an existing binary file (e.g. HDF5)
- You get Icechunk versioning and cloud performance without copying or converting a single byte
- `earthmover-demos/fair-mast` is a **zero-copy reference** to the public **FAIR-MAST** Zarr stores — MAST tokamak data, catalogued and versioned, without duplicating a byte
- Existing archives can be **indexed in place** (note reads go through the archive's original storage, so performance depends on it; hot subsets can be materialized later)



app.earthmover.io/earthmover-demos/fair-mast

NATIVE READER BILL

\$0

< 1 s

0 B downloaded

VIRTUAL CHUNKS BILL

\$0

< 1 s

0 B downloaded

Archival (NetCDF / GRIB)



NATIVE NETCDF / GRIB READER

Your laptop

DECODE

RESULT ARRAY

ICECHUNK VIRTUAL DATASET

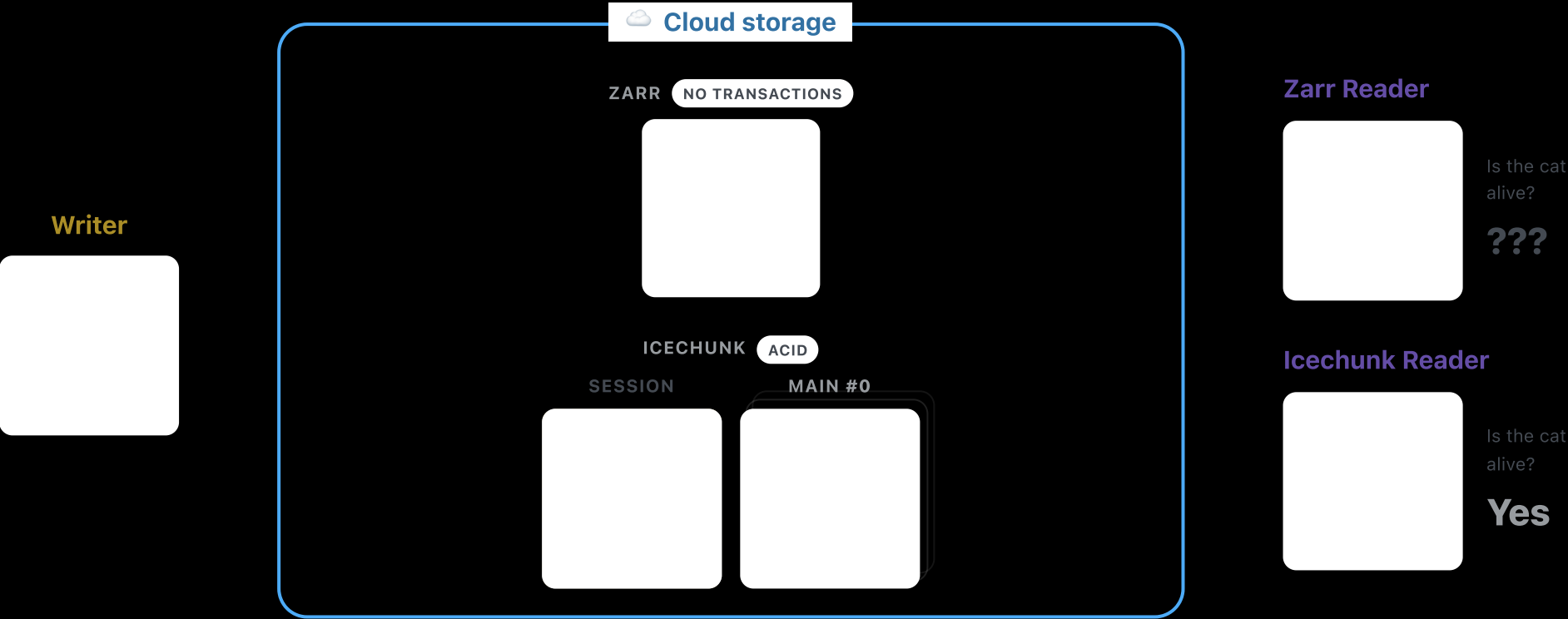
Icechunk repo



Your laptop

RESULT ARRAY

[← Icechunk, illustrated](#)



STORAGE USED

0 KB

Latest snapshot

FULL-COPY ALTERNATIVE

0 KB

SAVED



Object storage



EVOLVING COMPUTE

zax: pure-rust tensor compute engine

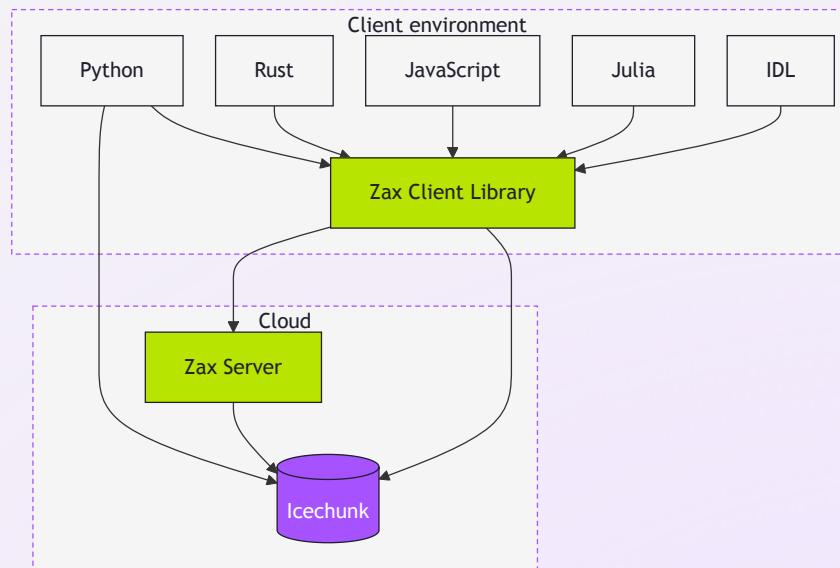
Inspired by:

- **Xarray** - Same NetCDF-based data model and user-friendly API
- **Polars** - Database-style planner / optimizer / executor architecture with full-stack awareness

Key Features

- **No Dask!** - End-to-end execution in Rust.
- Bindings to **all major languages**.
- **Client / server mode** — Run compute next to data
- **Open Source** (Apache 2.0)

data flow



(Public release Fall 2026)



EVOLVING COMPUTE

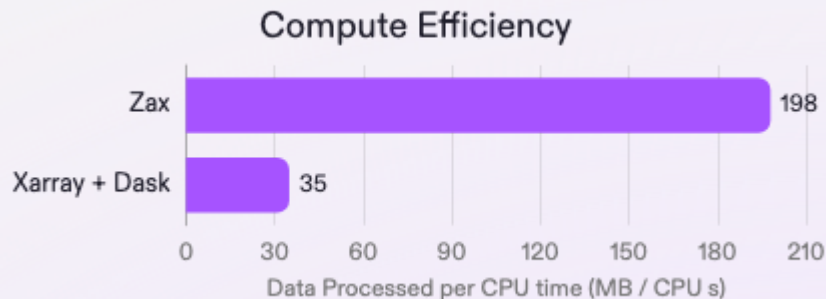
zax: pure-rust tensor compute engine

Inspired by:

- **Xarray** - Same NetCDF-based data model and user-friendly API
- **Polars** - Database-style planner / optimizer / executor architecture with full-stack awareness

Key Features

- **No Dask!** - End-to-end execution in Rust.
- Bindings to **all major languages**.
- **Client / server mode** — Run compute next to data
- **Open Source** (Apache 2.0)



Important

Zax is > 5x more efficient at processing data than the dominant Python / Xarray / Dask stack.

Faster science and **lower energy usage** per calculation.



PLATFORM

the earthmover stack

platform

Flux

Scalable Tensor Compute Services

Tiles

EDR

DAP

Zax Server

Zax SQL

OpenEO

Arraylake

Data Management and Governance

Data Catalog

Credential Vending

Access Controls

Data Marketplace

open source

Icechunk

Tensor Storage Engine

ACID Transactions

Data Versioning

Best-in-class I/O Performance

Virtualization of Legacy Formats

cloud

BYOB

Bring your own bucket (or use ours)



JASMIN

generic S3-compatible



PLATFORM

the earthmover stack

platform

Flux

Scalable Tensor Compute Services

Tiles

EDR

DAP

Zax Server

Zax SQL

OpenEO

Tokamak data queries?

Tokamak data visualization?

open source

Arraylake

Data Management and Governance

Data Catalog

Credential Vending

Access Controls

Data Marketplace

cloud

Icechunk

Tensor Storage Engine

ACID Transactions

Data Versioning

Best-in-class I/O Performance

Virtualization of Legacy Formats

BYOB

Bring your own bucket (or use ours)



generic S3-compatible



TRACTION

earthmover public sector adoption



Providing core data infrastructure for the **Tipping Points** project, integrating 26 independent research groups.



Helping **Integrated Modeling Virtual Institute** unlock value from model data in the cloud. (MERRA-21C, ECCO, etc.)



Building **NWS CIRRUS** with Booz Allen Hamilton. Pilots ongoing within OAR (PMEL, GFDL).



Building the Destination Earth Data Lake **Data Cube** project, in partnership with EODC, DevSeed, etc.



Scoping pilot for operational / research forecast data delivery system.



Scoping pilot for BARRA model delivery.



TRACTION

earthmover public sector adoption

ARIA is a UK public R&D agency making bold bets on transformative science — **similar to UKAEA**.

Advanced
Research
Innovation
Agency

ARIA

Providing core data infrastructure for the **Tipping Points** project, integrating 26 independent research groups.

NASA

Helping **Integrated Modeling Virtual Institute** unlock value from model data in the cloud. (MERRA-21C, ECCO, etc.)



Building **NWS CIRRUS** with Booz Allen Hamilton. Pilots ongoing within OAR (PMEL, GFDL).



Building the Destination Earth Data Lake **Data Cube** project, in partnership with EODC, DevSeed, etc.



Scoping pilot for operational / research forecast data delivery system.



The Bureau
of Meteorology

Scoping pilot for BARRA model delivery.



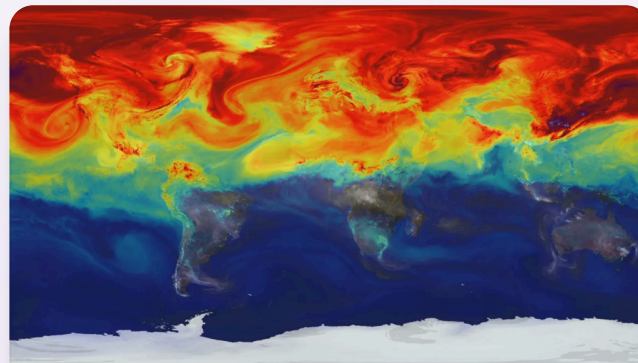
CASE STUDY

ARIA needed a Data Commons ASAP

UK R&D funding agency **ARIA** needed a Data Commons for its **Forecasting Climate Tipping Points** programme, for **26 independent research teams** generating diverse data to share and build on.

Requirements

- **Federated storage** - JASMIN on-prem + cloud
- **Data access controls** for private and embargoed data
- **Tensor-aware** data lake — tabular data lakes won't work
- **Infrastructure maintenance** for the life of the programme
- **Turn-key solution**, not **months to years** of custom development



**Earthmover Selected to Power ARIA's
"Forecasting Tipping Points" Catalogue**

February 9, 2026



ARRAYLAKE

a catalog for your team

- Organizations, repositories, branches and commits, all browsable in a web app
- Searchable metadata instead of README archaeology
- Access control per repository, plus API tokens for scripts and apps
- Data stays readable after the programme ends

Dataset Logical Size: 10.7 TB

DIMENSIONS (5)

climate_scenario : 4 prediction_year : 4 y : 151552 x : 163840 rp : 3

COORDINATES (5)

Name	Dimensions	Dtype	Shape	Chunks	Size	Attrs	Encoding	Preview
climate_scenario	climate_scenario	variablelengthutf8	4	4	--			rcp45 rcp26 rcp85
prediction_year	prediction_year	int64	4	4	32 B			2040 2026 2050 203
rp	rp	int64	3	3	24 B			20 100 200
x	x	float64	163840	163840	1.31 MB			2.5e+6 2.5e+6 2.5e
y	y	float64	151552	151552	1.21 MB			5.4e+6 5.4e+6 5.4e

DATA VARIABLES (7)

Name	Dimensions	Dtype	Shape	Chunks	Size	Attrs	Encoding
aai	climate_scenario prediction_year y x	float32	4, 4, 151552, 163840	1, 1, 512, 512	1.59 TB		
beta	climate_scenario prediction_year y x	float32	4, 4, 151552, 163840	1, 1, 512, 512	1.59 TB		
hazard_index	climate_scenario prediction_year y x	int8	4, 4, 151552, 163840	1, 1, 512, 512	397 GB		
intensity	rp climate_scenario prediction_year y x	float32	3, 4, 4, 151552, 163840	3, 1, 1, 512, 512	4.77 TB		
mu	climate_scenario prediction_year y x	float32	4, 4, 151552, 163840	1, 1, 512, 512	1.59 TB		
rp0	climate_scenario prediction_year y x	int16	4, 4, 151552, 163840	1, 1, 512, 512	795 GB		

A 10.7 TB climate risk dataset in the Arraylake catalog



ARRAYLAKE

one login, any cloud

No more juggling S3 keys and tokens per storage system. Log in once; Arraylake vends short-lived, scoped credentials for wherever the data lives:

```
import arraylake as al
import xarray as xr

client = al.Client()
client.login() # once

# open data from wherever you have internet
ssp126 = client.get_repo("earthmover-aria/CMIP6....ssp126...ta")
ssp245 = client.get_repo("earthmover-aria/CMIP6....ssp245...ta")
ssp585 = client.get_repo("earthmover-aria/CMIP6....ssp585...ta")
tipmip = client.get_repo("earthmover-aria/tipmip-demo")

for repo in [ssp126, ssp245, ssp585, tipmip]:
    ds = xr.open_zarr(repo.readonly_session("main").store)
    ds.ta.isel(plev=0).mean(["lon", "lat"]).plot()
```

scenario	repo	data lives on	format
SSP126	JASMIN	Google Cloud	Zarr (virtual)
SSP245	JASMIN	AWS	netCDF (virtual)
SSP585	JASMIN	AWS	Zarr (virtual)
TIPMIP	JASMIN	JASMIN	Icechunk

One plot, four repos, three clouds, three formats.

Under 1 minute on 32 cores in eu-west-2.

you

arraylake auth login
once



Arraylake

checks your permissions, vends short-lived credentials



JASMIN

aws



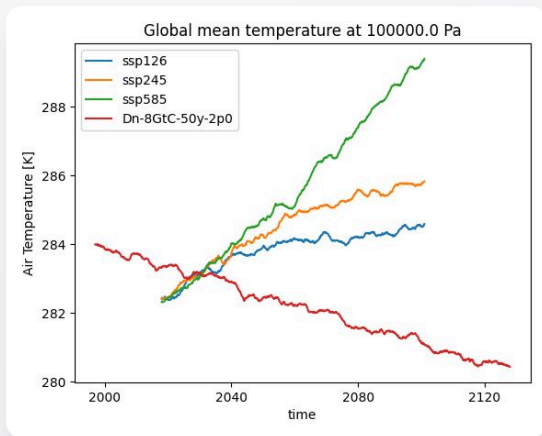
CLOUDFLARE



ARRAYLAKE

one login, any cloud

No more juggling S3 keys and tokens per storage system. Log in once; Arraylake vends short-lived, scoped credentials for wherever the data lives:



scenario	repo	data lives on	format
SSP126	JASMIN	Google Cloud	Zarr (virtual)
SSP245	JASMIN	AWS	netCDF (virtual)
SSP585	JASMIN	AWS	Zarr (virtual)
TIPMIP	JASMIN	JASMIN	Icechunk

One plot, four repos, three clouds, three formats.

Under 1 minute on 32 cores in eu-west-2.

you

arraylake auth login
once



Arraylake

checks your permissions, vends short-lived credentials



JASMIN

aws



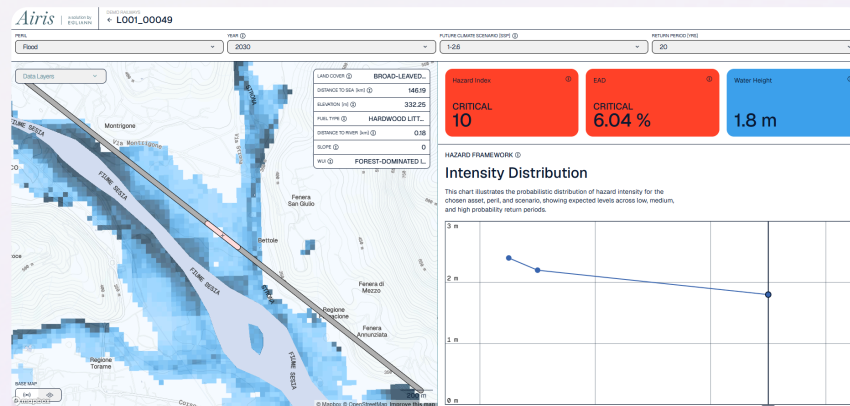
CLOUDFLARE



ARRAYLAKE

flux: from data to visualization or query api in minutes

- Any dataset becomes a standard web service, where we run the server for you, scalably
- Domain agnostic standards: **OpenDAP** (inc. return as netCDF), **SQL**
- Supports various Open Geospatial Consortium API standards (EDR, WMS, Tiles)
 - Works with real model grids: e.g. curvilinear, tripole, unstructured meshes
- Could we support similar **fusion API standards**?



A flood-risk platform whose map tiles come straight from Flux (from our blog)

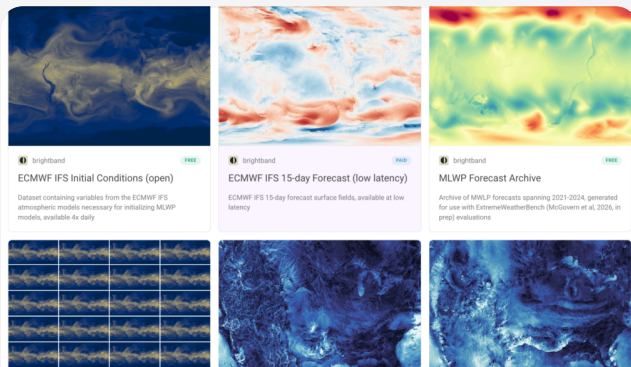


MARKETPLACE

tailored subscriptions, updated automatically

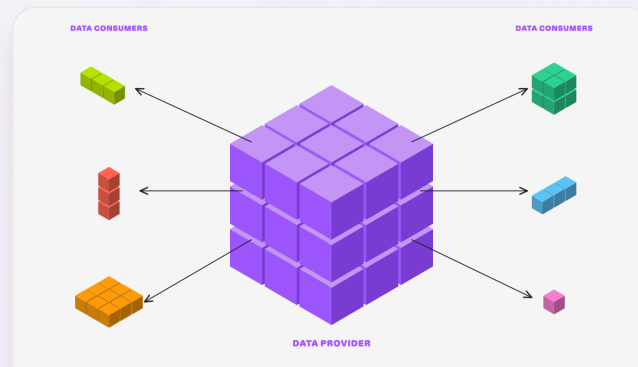
Subscribe to **data feeds across institutions**, with new data **instantly available**, as a versioned **commit history**.

Subscription Sync allows **embargoed data**: e.g. new data visible only to collaborators, but public once embargo lifts.



Announcing the Earthmover Data Marketplace

January 22, 2026



Filtered Subscriptions: Fine-Grained Cloud-Native Data Sharing

April 15, 2026



WHAT THIS ENABLES

agent-assisted science becomes easy



Nikolay Koldunov • 1st

Climate scientist. Senior scientist at Alfred Wegener Institute
4mo • Edited •

Our team built a demo agent application for working with ERA5 data based on **Earthmover** Arraylake. Fast data access lets us pull the right slices of ERA5 quickly, iterate, and visualize, without turning every question into a coffee break.

What Eurus does:

- Natural language → precise retrieval + analysis for ERA5 workflows
- Multiple interfaces: interactive CLI, web chat UI, and an MCP server for tool-using assistants/IDEs
- Built-in Python REPL (pandas/xarray/matplotlib/cartopy)

See an example of the workflow in the video below.

Repo: <https://lnkd.in/deQMCNmH>

Live Hugging Face deployment: <https://lnkd.in/dAbhnKSr>

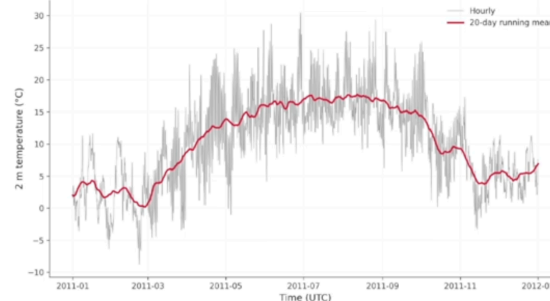
⚠ To run the demo (locally or on HF), you'll need:

- an OpenAI API key
- an Arraylake API key



Overlay added: a **20-day running mean** (computed as a centered **480-hour** rolling mean) plotted on top of the hourly Bremen 2 m temperature series for 2011.

Bremen area (53.0-53.2°N, 8.7-9.0°E): Hourly 2 m temperature with 20-day running mean (2011, ERA5)



Enlarge

Download

Show Code

plot temperature in Helsinki for the same time in h

Send

Enter to send, Shift+Enter for new line



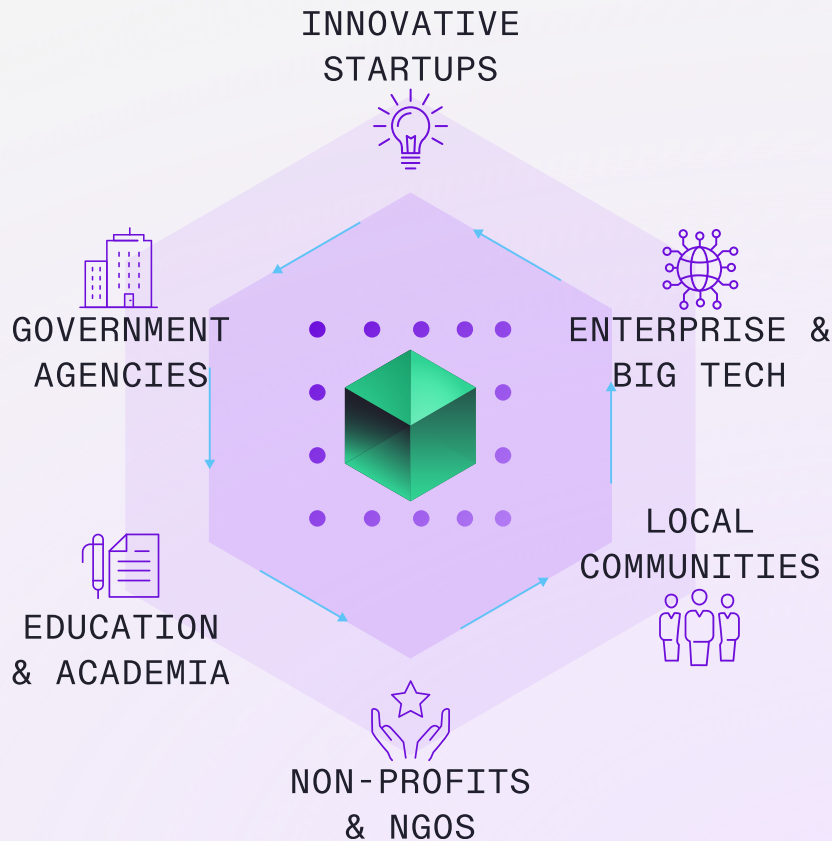
OUR VISION

a connected ecosystem

We envision a world in which scientific data can be **explored, visualized, analyzed, shared, and built-upon effortlessly.**

Our technology provides a foundation for people and organizations to create new scientific knowledge and data-driven products which will help humans thrive in sustainable harmony with each other and with our planet.

Our company becomes a key trusted partner to industry, government, and academia while helping to sustain a multi-stakeholder open-source ecosystem.





NEXT STEPS

let's work together

1

What open-source tooling do you need?

Interface layers (e.g. for MDSplus, IMAS, custom binary formats)?

Support existing **fusion API standards?**

Feature additions to Xarray / Zarr / Icechunk / Arraylake?

2

Which archive would you want to try this on first?

Let's pick an important data archive and **show we can make it work.**

Could be simulation data, raw diagnostics, inferred analysis...

3

Can we define the reference architecture together?

This vision gets most powerful when it's shared across institutions.

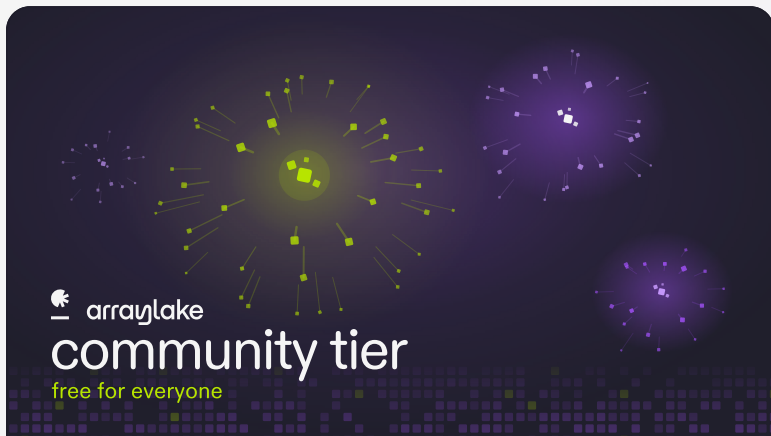
Co-authoring a white paper together would be a great way to align.



GET STARTED

come and try it

TRY IT YOURSELF — FREE



Announcing the Arraylake Community Tier

Your own repos · bring your own bucket or 10 GB managed storage · 50 compute credits / month

STAY IN TOUCH



Scan to join the mailing list · earthmover.io