

AMS Annual Meeting 2026

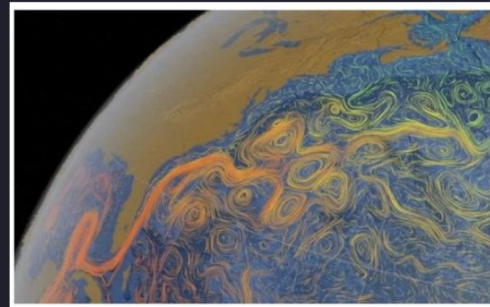
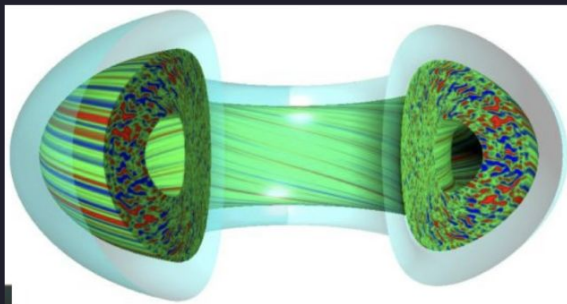
Shortcutting cloud migrations with VirtualiZarr, Icechunk, and Earthmover

About me



Tom Nicholas, PhD
FORWARD ENGINEER

- Plasma Physicist
- Pivoted to geoscience data
- Open source maintainer



MISSION STATEMENT

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

Three Trends Causing Disruption



AI/ML Workloads

AI has a **voracious appetite** for training data. Teams need access to **entire archives**.



Cloud Computing

Access data directly from **cloud object storage**.

Access **scalable compute** instantly without owning infrastructure.



Open Science

Funders and journals increasingly mandating the sharing of **open data**.

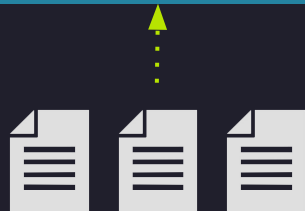
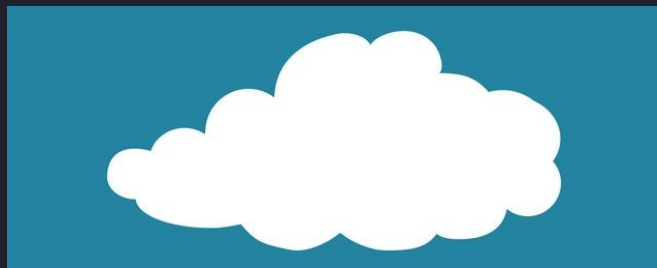
Old-School Data Distribution Systems: Struggling to Keep Up



Provides **download access** for GRIB and NetCDF files.

- On-prem servers struggle to fulfill throughput demands.
- GRIB and NetCDF files give **poor performance** in the cloud...

Pressure to move to the cloud...



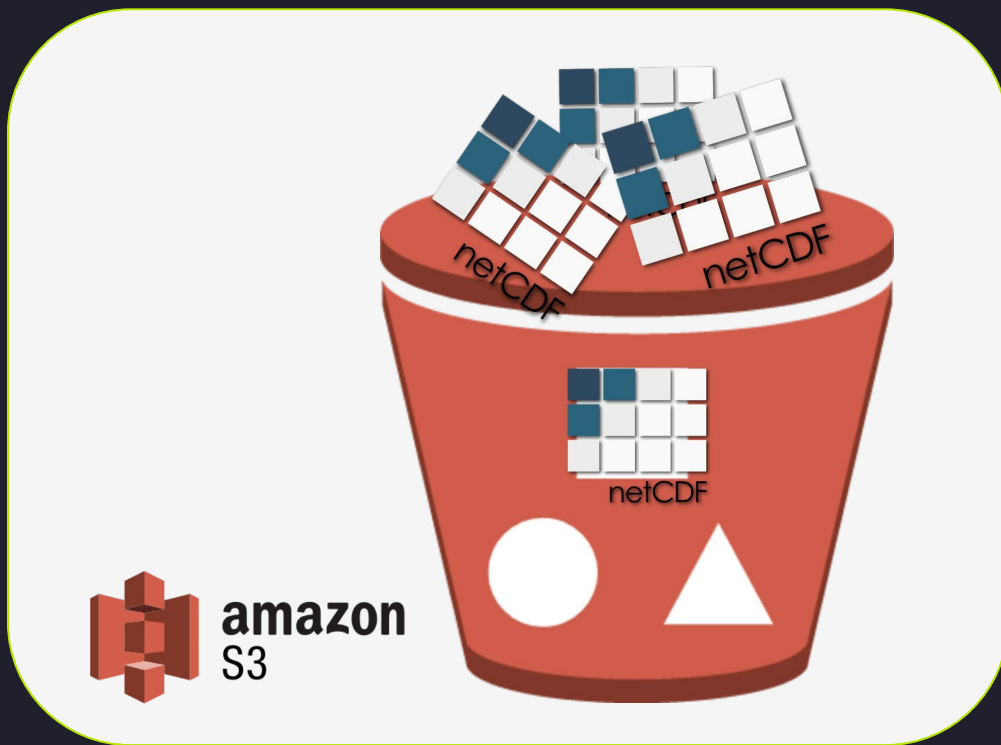
GRIB / NETCDF FILES



- Top-down mandates
- Amongst shrinking budgets
- Lack of cloud expertise

Naive “lift and shift”

Result: Pile of files

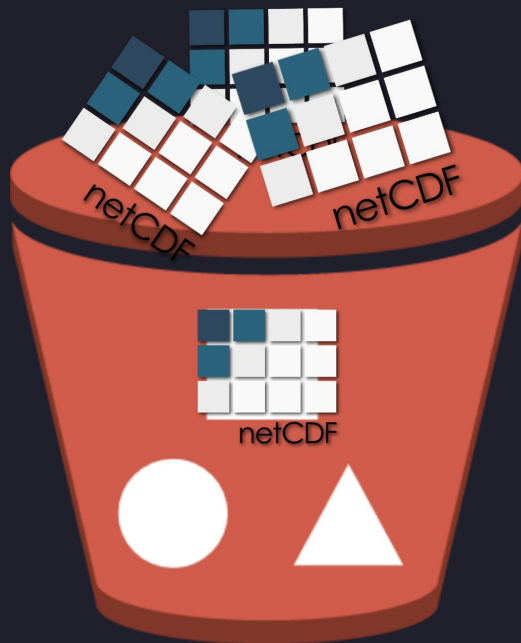


Problems:

- Not organised ✗
- Poor performance ✗
- Not discoverable ✗

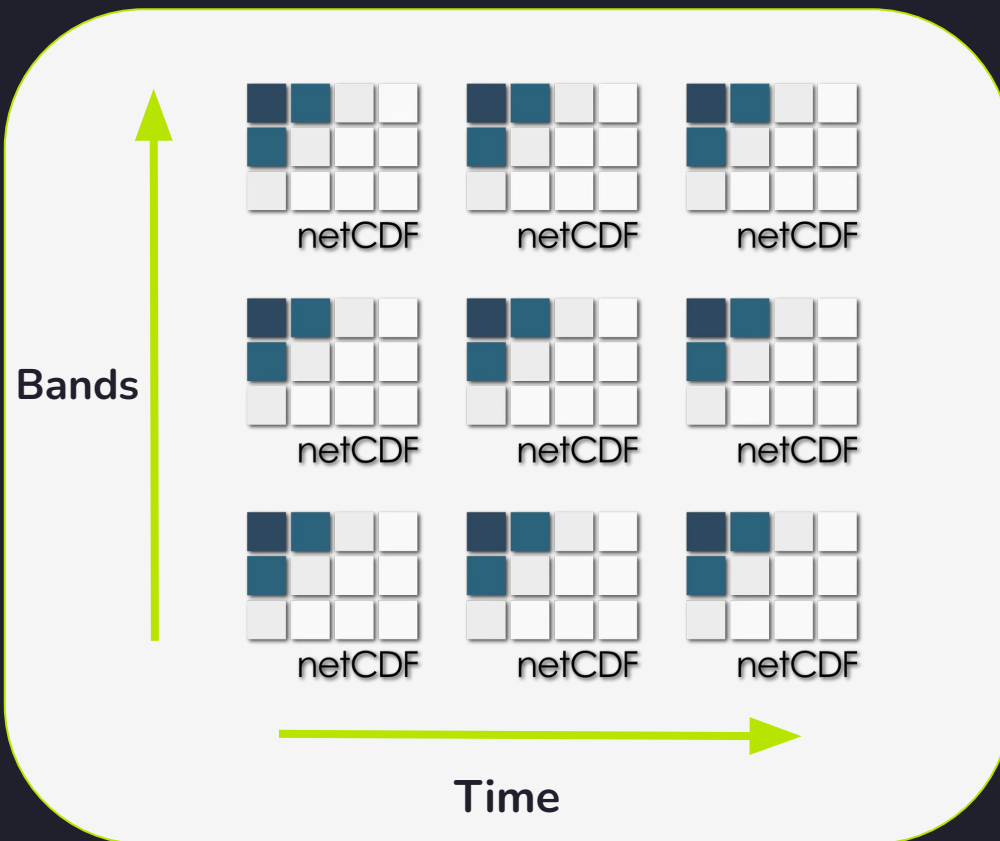
VirtualiZarr, Icechunk,
and the Earthmover
Platform **solve these
issues!**

Example: GOES-16



- In AWS Open data bucket ✓
- Not cloud-optimized (NetCDF) ✗
- Impossible to download everything ✗
 - 300,000 files for one product
 - 66PB for one product
- Structure not apparent ✗
- Only sort of discoverable... ✗

But the data does have structure!

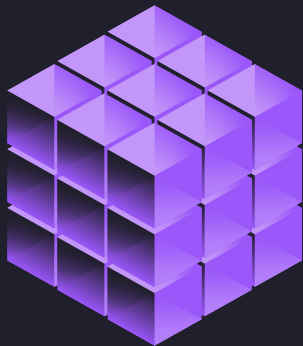


The Emerging Standard: Cloud-optimized Datacubes

Analysis Ready

Cloud Optimized

- ✓ All variables
- ✓ All timesteps
- ✓ Any query



Zarr



xarray

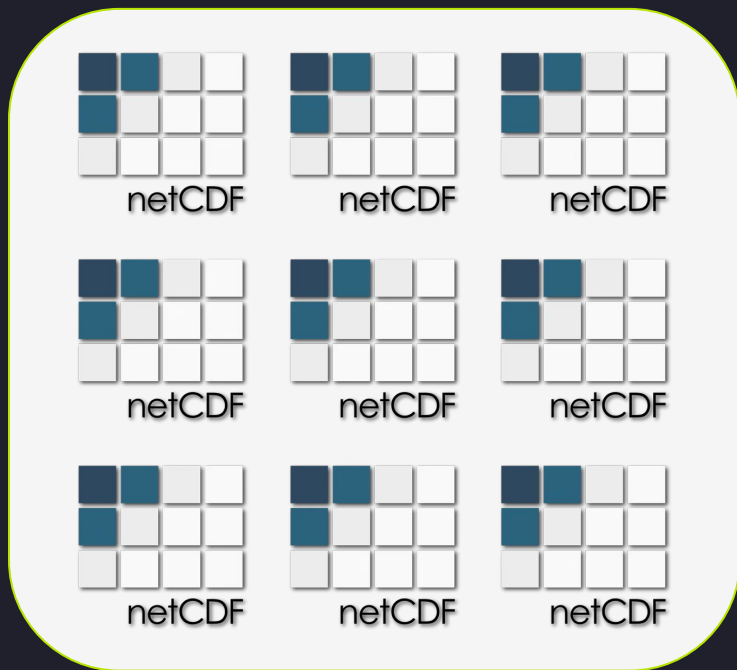


PyTorch

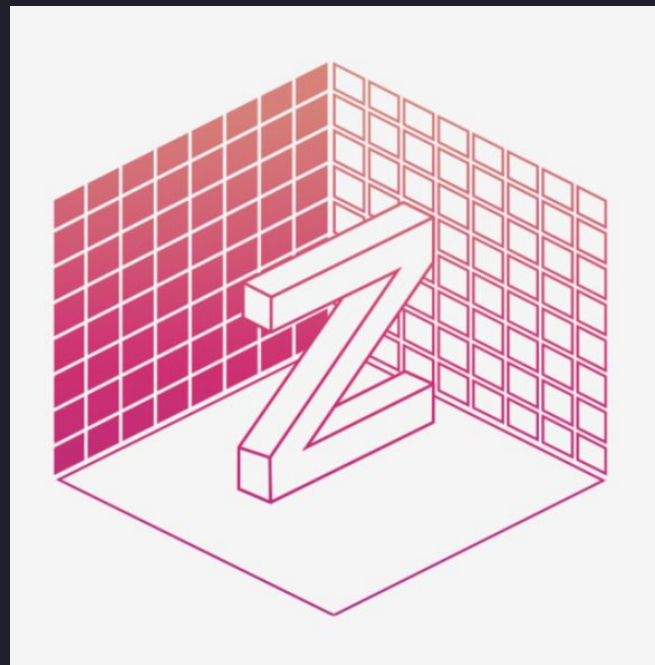
Cloud Optimized - Data are stored in cloud object storage using a cloud-native file format (Zarr), enabling high throughput and low latency queries. Icechunk enables ACID transactions for Zarr.

Datacubes - Data are organized into hypercubes which allow arbitrary slicing across forecast time, forecast step, space, and ensemble dimensions. *No files to think about!*

Rewrite as Zarr?



66PB



+66PB

What if it just looked like Zarr...



Icechunk stores “Virtual Chunks”

Transactional,
cloud-native storage
engine for Zarr

Works with Zarr-Python
and Xarray

Supports **virtual chunks**



Core implemented in
Rust; thin Python
wrapper

100% open source
(Apache 2.0)

VirtualiZarr extracts “Virtual Chunks”

```
import virtualizarr as vz  
vz.open_virtual_dataset(
```



“Chunk Manifest”

```
{  
  "0.0.0": {"path": "s3://bucket/foo.nc", "offset": 100, "length": 100},  
  "0.0.1": {"path": "s3://bucket/foo.nc", "offset": 200, "length": 100},  
  "0.1.0": {"path": "s3://bucket/foo.nc", "offset": 300, "length": 100},  
  "0.1.1": {"path": "s3://bucket/foo.nc", "offset": 400, "length": 100},  
}
```



Discoverability through Arraylake

 earthmover / goes-16

 goes-16

Content

Metrics

SI

main Updated about

root

Description

Multiple reflectance and emissive bands.

Dataset

DIMENSIONS (6)

t : 4790 band : 16 y : 1

COORDINATES (7)

Name	Dimensions
band	band
t	t

Repository Code Snippets

Python (Zarr)

Python (Xarray)

```
from arraylake import Client
import xarray as xr

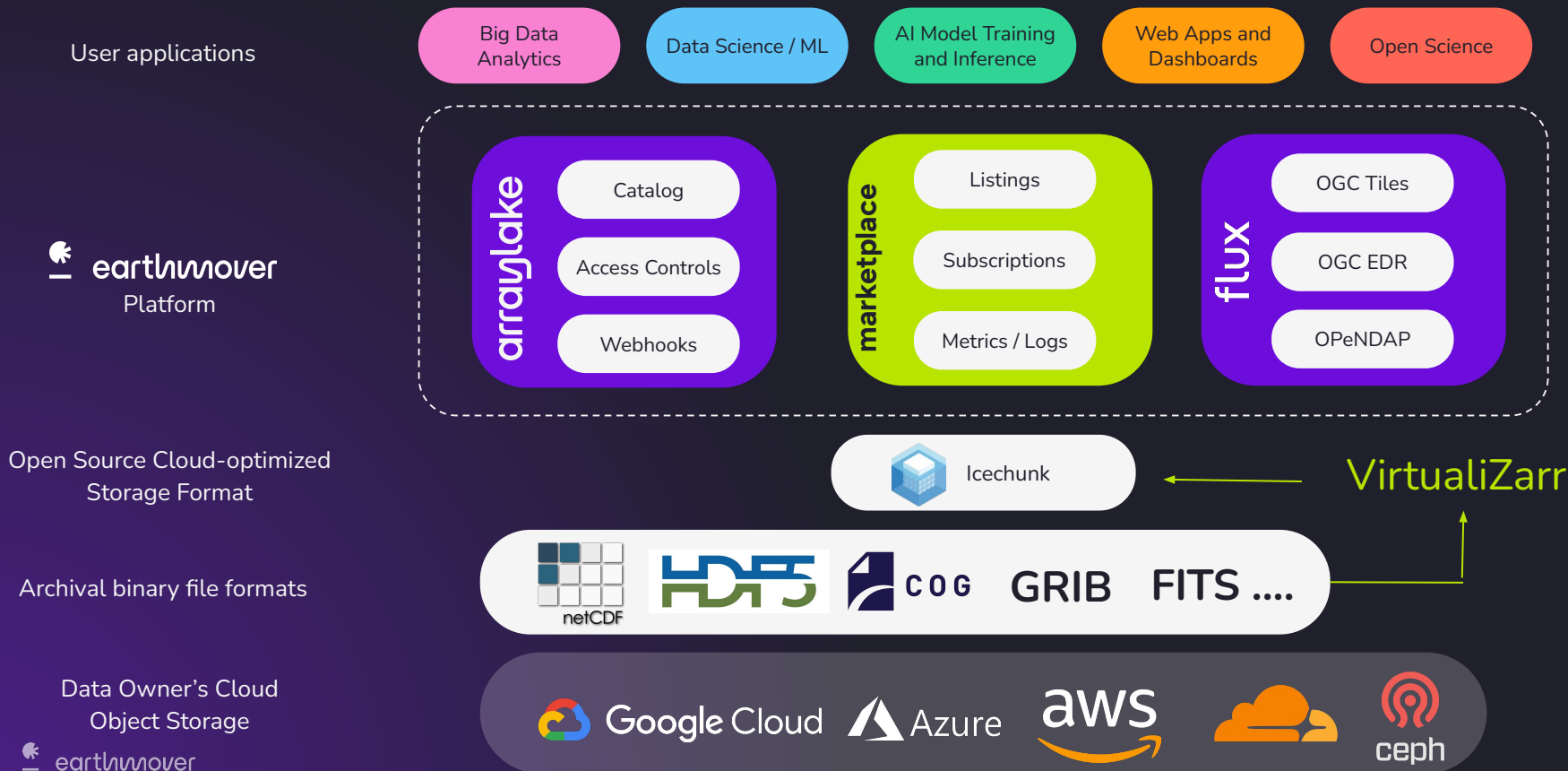
client = Client()
repo = client.get_repo(
    "earthmover/goes-16",
    authorize_virtual_chunk_access={
        "s3://noaa-goes16/": "noaa-goes16",
    },
)
session = repo.readonly_session(branch="main")
ds = xr.open_zarr(session.store, zarr_format=3)
```

float64 4790 20000 2023

GOES-16 True Color - April 8, 2024 18:00 UTC



Earthmover: The ARCO Data Platform



Earthmover Data Marketplace

ARCO Data from Trusted Providers



<https://app.earthmover.io/marketplace>

Earthmover's Vision

Frictionless global scientific data exchange

Come find us @ **booth 711** for
a Marketplace demo!

