earthmover

mobilizing Earth System data and AI to confront the challenges of climate change

- cloud-native data repository for data providers
- private data lake for research and analytics teams
- feature store for AI / ML training
- Production-grade backend for data-driven apps
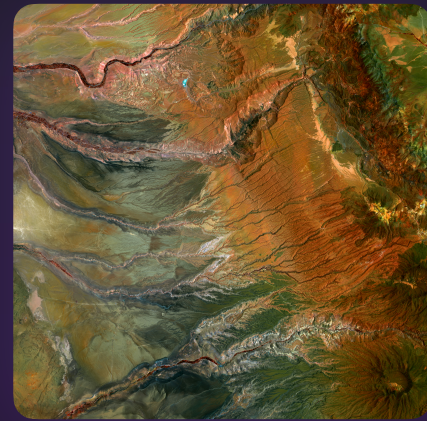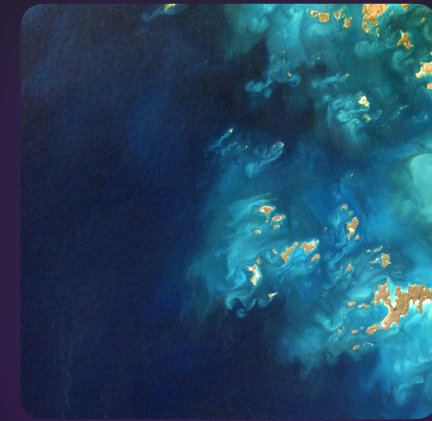
problems

progress is bottlenecked on inefficient data practices

✗ Earth system data exchange is based on files; datasets are sharded arbitrarily into files / granules

✗ data scientists waste huge amount of time writing data pipelines rather than doing interesting science

✗ foundation model progress limited by data throughput

✗ organizations are struggling to migrate to the cloud

mainstream cloud data tools don't speak NetCDF

✓ we have a pretty great universal data model for Earth-system data (NetCDF + CF Conventions)

✗ but popular cloud data lakehouse / warehouse platforms expect data to conform to the tabular / relational model

→ every team has to build their own bespoke cloud data management solution
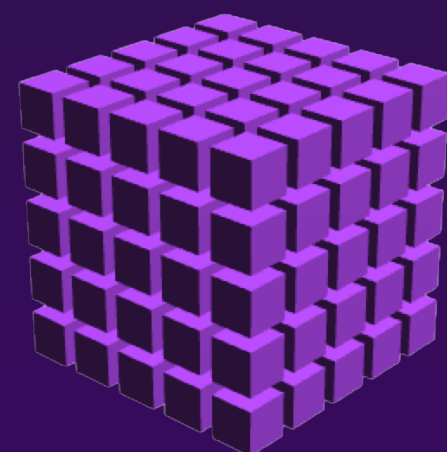
solution

arraylake

**ARCO**   **AI-ready**

a cloud-native database for the Earth System

✓ high-performance array / tensor storage API (Zarr)

✓ rich web data catalog

✓ full-metadata search

✓ works with all the file formats (Zarr, NetCDF, HDF, GRIB, TIFF)

✓ access controls

✓ ACID transactions

✓ version control for data
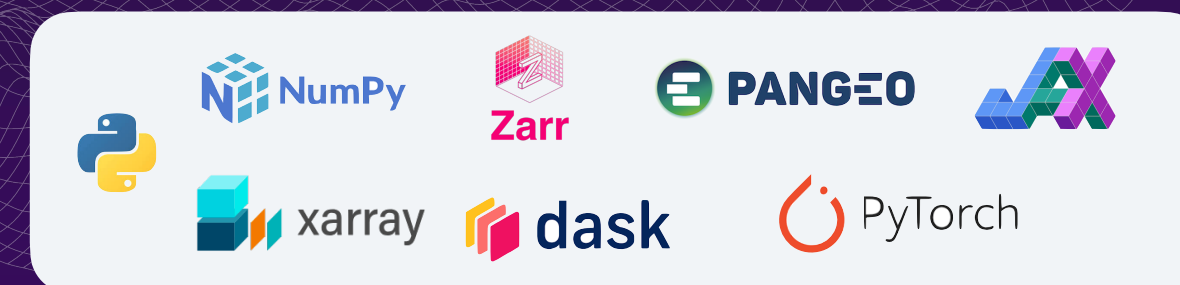
✓ metrics, logging, audit logs

compatible with all cloud storage

universal array / tensor data model
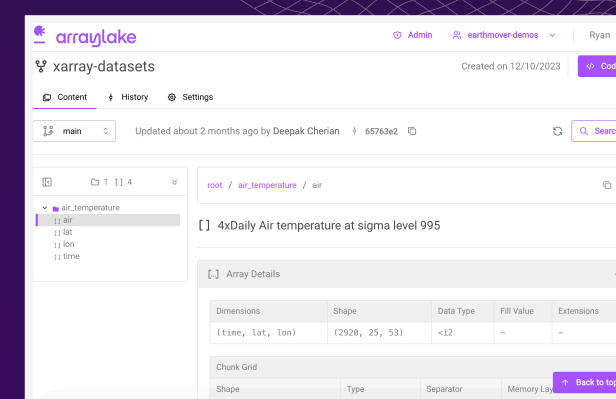
web-based data catalog

built for the Python data science and ML ecosystem