

Open Standards for Earth Science Data Platforms

Theo McCaie - UK Met Office, Infrastructure and Data Engineering
Research Lead



TLDR;

**An open standards
approach to data
platforms will accelerate
earth science.**

PANGEO

A community platform for Big Data geoscience

PANGEO

Strengths:

- Multi-cloud
- Open Source
- Vibrant Community
- Elastic scalability
- Emerging consensus

Still developing:

- Billing opacity and control
- Maintenance overhead
- Ease of deployment
- No SLAs etc

Meanwhile...

Cloud vendor data science platforms

Strengths:

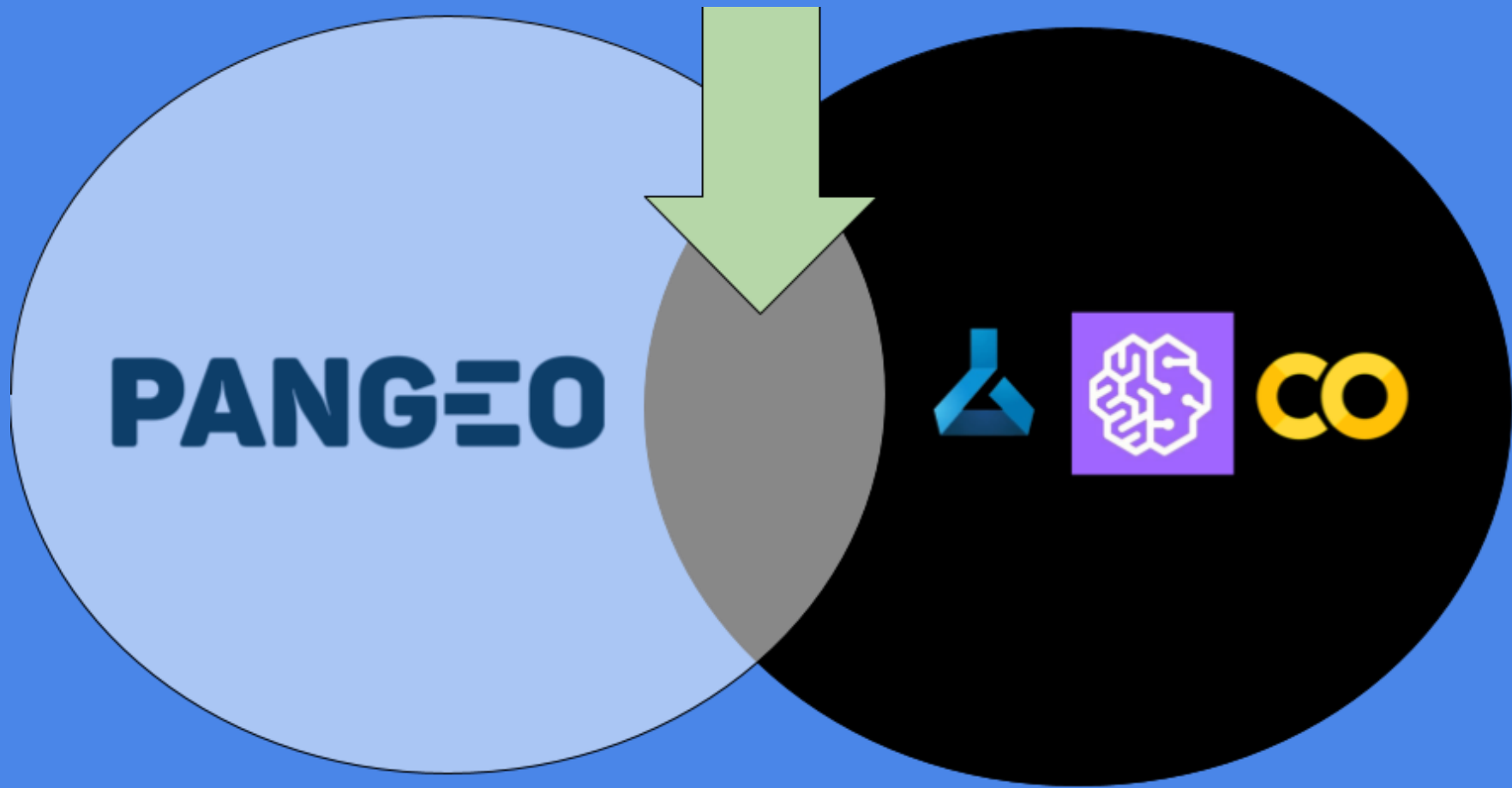
- Cost management
- SLAs
- Easy deployment
- TCO

At a cost of:

- Vendor lock in
- Difficulty working with datasets across multiple vendors
- Closed source



Brace for it...



Open standards

- An agreed way of doing things that can be implemented by anyone
- Mutually beneficial for customers and suppliers
- Enable interoperability
- Examples:
 - Petrol pumps
 - Domestic electrics
 - The internet
 - Data science platforms???



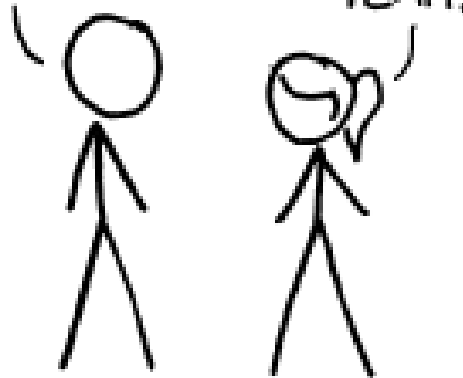
Credit: <https://www.bbc.co.uk/news/uk-27390466>

HOW STANDARDS PROLIFERATE:

(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)

SITUATION:
THERE ARE
14 COMPETING
STANDARDS.

14?! RIDICULOUS!
WE NEED TO DEVELOP
ONE UNIVERSAL STANDARD
THAT COVERS EVERYONE'S
USE CASES.



SOON:

SITUATION:
THERE ARE
15 COMPETING
STANDARDS.

Credit: <http://xkcd.com/927>

Capability

Entry point and user orchestration

This capability gives users a URL from which to access and manage their resources.

Authentication

The ability to securely identify users

Compute instance

An interactive compute environment exposed through a web browser.

Compute cluster

Being able to provision and manage many nodes simultaneously to perform analyses

Discovery and Analysis Ready Data

Being able to find and quickly load data such that it's ready to process immediately.

Contract

JupyterHub

The main user orchestration layer in Pangeo. Jupyter Hub provides consistent user Experience across a range of systems and targets.

OAuth2

Industry-wide, arguably the current standard API for authentication. Supported by a wide range of services.

JupyterHub Spawner

Any interactive compute service that is spawnable and manageable via a JupyterHub Spawner is considered as fulfilling the contract.

Dask Cloud Provider

In the Pangeo Community Dask is becoming the dominant distributed compute tool. Any distributed compute service that can be managed through dask-cloud-provider fulfils the contract.

Intake

Intake is growing in popularity as a way of exposing datasets, distributing datasets and loading data such that it's immediately ready for analysis. Supplying datasets as an Intake catalogue and supporting any necessary drivers is required to meet this capability.

Dominant design

- Identify core capabilities required in the platform
- Find where there is consensus on how to achieve this
- Led by Pangeo community

Still fermenting

- For some capabilities it's unclear what the dominant design will be
- In other cases we don't know that the capability exists yet
- Keep watching this space and bring into the open standard fold as appropriate

Capability

Environment management

The ability to manage and install the necessary software and consistently reproduce a software environment.

Homespace

Where users store personal files, configurations, notes etc (but not significant volumes of data). Crucially needs to be accessible across platforms and services.

Experiment orchestration

The ability to run many hundreds or thousands of identical experiments with different input parameters and collect and compare the results. (Hyperparameter tuning is one example)

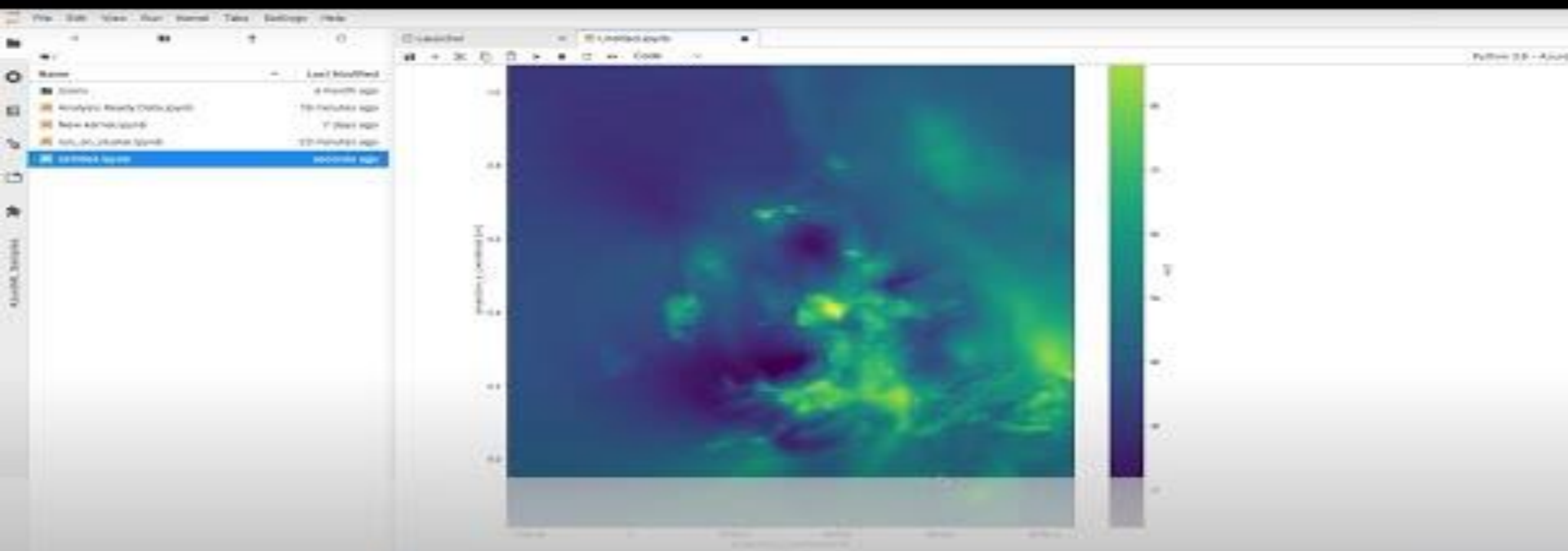
Contract

Conda is perhaps the main solution in this space however there are still challenges in quickly and easily ensuring a container/VM is spun up with an equivalent ecosystem to some other, especially when considering multiple architectures. Docker is also worth considering in this space but again comes with different challenges particularly when 'mixing' elements from different environments

Whilst there is a myriad of file storage solutions from Dropbox to Azure NetApp Files there isn't a standard API/interface/mounting solution that makes these compatible with other services and available on any VM/container/platform/etc.

There are a range of tools to do this task and most cloud providers offer a solution but there is no clearly emerging standard by which this can be easily rendered interoperable between services.

From theory to
practice...



Analysis ready data allows interrogation in the language of the subject domain...

Conclusions - open standards could:

- Democratise data access (you don't need a platforms team to engage meaningfully)
- Speed up scientific delivery (focus on science not on tools, improve collaboration, consistent UX)
- Provide a common platform for research and operations (faster science to service?)
- Encourage healthy competition, dissuade aggressive lock in tactics
- But:
 - Requires buy-in from the vendors and community

Thanks, links, further reading...

- [Grey new world](#) - an article exploring the 'grey space' between open and closed source and the advantages of embracing this may offer to the geo-science community.
- [Analysis ready data](#) - a post explaining "Analysis Ready Data", it's importance and how despite rigorous data standards many institutions fail to produce it.
- [What do we want from a dream data platform as a Service?](#) - a series of user stories, personas and explanation describing what a "data science platform-as a Service" should look like.
- [What should a geospatial data service look like?](#) - a more detailed look at the challenges and solutions for powerful geospatial data services.
- [Homespace, the missing as a service?](#) - an argument that 'homespace' is a service that is missing from the current cloud ecosystem.
- [aml-jupyterhub on GitHub](#) - a JupyterHub spawner that targets Azure Machine Learning service as the Compute Instance Provider. As demoed in this poster.
- [Using TileDB and Pangeo to Provide Access to Thousands of NetCDF Files as Analysis-Ready Data](#) - a poster detailing on method for creating analysis ready data.

These slides:



shorturl.at/dkHK0