Pangeo for Plasma

Lessons for plasma software from the climate data analytics community

Thomas Nicholas
(Columbia University / Lamont-Doherty Earth Observatory)

thomas.nicholas@columbia.edu
Who am I?
Who am I?

PhD with Ben Dudson, Fulvio Militello, BOUT++
Who am I?

PhD with Ben Dudson, Fulvio Militello, BOUT++

RSE with Ryan Abernathey, various projects
What do I do now?
What I hope to convince you of

- Our computational infrastructure **needs to change a lot**
- Can use **solutions from climate science** community
- **Modular approach** makes everyone's work easier
- **Opportunities exist** for plasma coders...
The White House announces

The Federal Year of Open Science

NASA ✦ NSF ✦ NOAA ✦ DOA ✦ DOC ✦ DOE ✦ GSA ✦ NEH ✦ NIH ✦ NIST ✦ USDA ✦ USGS

Along with other organizations, including CENDI group, voluntary collaboration among Federal managers, and HELIOS, a coalition of 80+ universities

A multi-agency initiative across the federal government to spark change and inspire open science engagement through events and activities that will advance adoption of open science.

Website: https://open.science.gov/
WH: https://www.whitehouse.gov/ostp/news-updates/
Nature: https://doi.org/10.1038/d41586-023-00019-y
Climate Science == Plasma Physics

- **Multidimensional** (often fluid turbulent)
- **Large** (bigger than local RAM)
- On regular but warped **grids**
- Often pulled from **central** servers
- From multiple sources but with **common structure** (e.g. experimental and simulation data for same device).
Climate Science == Plasma Physics

- **Multidimensional** (often fluid turbulent)
- **Large** (bigger than local RAM)
- On regular but warped **grids**
- Often pulled from **central** servers
- From multiple sources but with **common structure** (e.g. experimental and simulation data for same device).
Typical scientific workflow
Typical scientific workflow

step 1: download

files

FTP / OPeNDAP / etc.

step 2: analyze

local disk

step 3: debug

Because you likely rolled-your-own code...
Problem 1: Code not reused
Problem 1: Code not reused

Modern data science libraries 🚀

Me as PhD student, circa 2017
Problem 2: Data accessibility
Problem 2: Data accessibility

PRIVILEGED INSTITUTIONS CREATE “DATA FORTRESSES*”
Problem 2: Data accessibility

PRIVILEGED INSTITUTIONS CREATE “DATA FORTRESSES*”
Problem 2: Data accessibility

PRIVILEGED INSTITUTIONS CREATE "DATA FORTRESSES * "

Image credit: Moahim, CC BY-SA 4.0, via Wikimedia Commons
Problem 3: Scale

“Brb, let me just go download the data to my laptop…”
Problem 3: Scale

“Brb, let me just go download the data to my laptop…”
Problem 3: Scale

“Brb, let me just go download the data to my laptop…”
Problem 3: Scale

“Brb, let me just go download the data to my laptop…”
Problem 3: Scale

“Brb, let me just go download the data to my laptop…”
Problem 3: Scale

ITER Scientific Data Centre

HOW TO MANAGE 2 PETABYTES OF NEW DATA EVERY DAY
Geoscientists’ solution:
Geoscientists’ solution:

A community platform for Big Data geoscience
Solution 1: Modular, open ecosystem
Solution 1: Modular, open ecosystem

ECOSYSTEM

- Science projects
- Domain-specific packages
- Domain-agnostic libraries
- General-purpose tools
Solution 1: Modular, open ecosystem

PANGEO COMMUNITY

HTTP://PANGEO.IO
Solution 2: Cloud Computing
Solution 2: Cloud Computing
Solution 2: Cloud Computing

**PANGEO ARCHITECTURE**

- **Dask** tells the nodes what to do.
- **Xarray** provides data structures and intuitive interface for interacting with datasets.
- **Jupyter** for interactive access remote systems.
- Distributed storage
- Web browser

“Analysis Ready Data” stored on globally-available distributed storage.

Parallel computing system allows users deploy clusters of compute nodes for data processing.
Solution 2: Cloud Computing

PANGEO DEPLOYMENTS

NASA Pleiades

OCEAN.PANGEO.IO

Google Cloud Platform

NCAR Cheyenne

HTTP://PANGEO.IO/DEPLOYMENTS.HTML
Solution 2: Cloud Computing

PANGEO CLOUD DATA CATALOG

CATALOG.PANGEO.IO
Solution 3: Parallel computing frameworks

Collections (create task graphs)
- Dask Array
- Dask DataFrame
- Dask Bag
- Dask Delayed
- Futures

Task Graph

Schedulers (execute task graphs)
- Single-machine (threads, processes, synchronous)
- Distributed

Dask
Apache Spark
Numba
Solution 3: Parallel computing frameworks
How might this work for plasma?
How might this work for plasma?

Fusion plasma projects

Domain-agnostic libraries

General-purpose tools
How might this work for plasma?

**ECOSYSTEM**

- Fusion plasma projects
- Domain-agnostic libraries
- General-purpose tools
- tokamak-specific packages
How might this work for plasma?

PANGEA ARCHITECTURE

“Analysis Ready Data” stored on globally-available distributed storage.

Parallel computing system allows users deploy clusters of compute nodes for data processing.

Dask tells the nodes what to do.

Xarray provides data structures and intuitive interface for interacting with datasets.

Jupyter for interactive access remote systems.

End user through web browser.
How might this work for plasma?

Blog post: https://hackmd.io/@TomNicholas/rkyERwcoO#
Other bonuses of joining this ecosystem

- Parallel and out-of-core analysis
- Labelled dimensions
- Unit-aware arithmetic
- Easier reproducibility
- Plotting flexibility
- Machine Learning integration
Summary

● Geoscience has same problems as plasma physics 🌍🤝🌞

● Being solved using:
  ○ Modular community software ecosystem 🔧
  ○ Cloud computing ☁️
  ○ Parallel execution frameworks 🚀

● It's working for them - it could work for us! 📷
LEARN MORE

http://pangeo.io

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

https://medium.com/pangeo

@pangeo_data