Community-supported software components

Pangeo Environment

commercial cloud / large HPC

High performance file / object storage

compute

My Laptop

Raw data

scientist's compute nodes

data_chunk.0000

Jupyter pod

data_chunk.0001

Dask pod

data_chunk.0002

Dask pod

data_chunk.0003

Dask pod

kubernetes   slurm   DASK

xarray

jupyter

```
import dask.dataframe as ddf
df = ddf.read_parquet('s3://bckt')
df.foo.value_counts()
```

```
import xarray as xr
ds = xr.open_zar('gs://my/bucket')
ds.mean(dim=['time', 'lon'])
```

Try **https://tinyurl.com/pangeo-cmip6** right now on the cloud!
Figure courtesy Ryan Abernathey, Columbia.

WCRP CMIP6
World Climate Research Programme

Cloud considerations

- Commercial cloud vendors usually have a "Public Datasets" program for "publicly available high-value cloud-optimized datasets" for users seeking to "democratize access to data by making it available for analysis" (from Amazon Public Datasets webpage).
- For public datasets, typically ingress (upload to cloud) and storage are free, egress (download from cloud) is not. Academic users can apply for free cloud research credits.
- Caveats: programs can be ended at any time, have a time limit (e.g 2 years at AWS): negotiable.
- Many funding agencies evaluating the relative cost of moving to cloud vs purchasing or leasing on-premises.
- Major advantage is democratization: large data volumes available for analysis from anywhere in the world without replication.
- See discussion in *Science*, 8 February 2019: Government data, commercial cloud: Will public access suffer?
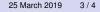
Computational and storage costs of CMIP6

- The CPMIP Project attempts to measure the computational and storage costs, and energy footprint, of CMIP6. Primavera doing the same.
- IS-ENES3 is taking a leading role in collecting this information, led by our hosts, the BSC team! (contact: Mario.Acosta@bsc.es).
- ES-DOC information request on models includes speed (SYPD), cost (CHSY), energy cost (JPSY), data intensity (GB/CH).
- First look (highly preliminary results!):

	Model Dev (kSY)	CMIP6 (kSY)	C Footprint (tons)
GFDL	53	(planned) 21	2750
IPSL	100	50	650

(1 transatlantic flight = 0.5T carbon per passenger).

Numbers courtesy Marie-Alice Foujols and Casimir Delavergne, IPSL, Alistair Adcroft and Aparna Radhakrishnan, GFDL. No implied endorsement of results by NOAA or IPSL.

WCRP CMIP6
World Climate Research Programme

A longer term view

- Despite growing international investment in climate modeling infrastructure, it remains fragile: single points of failure need to be addressed because they can lead to enormous disruptions
  - Some individuals are irreplaceable
  - Some software is not well documented
- ESGF has become essential to the climate research community:
  - CMIP, input4MIPs, obs4MIPs, etc.
  - Modeling and analysis groups have invested in it
- Given resource constraints, we should treat ESGF as part of an operational climate research enterprise; it must be reliable and robust
- Underlying data technologies are in flux: ESGF will adopt promising technical evolutions as they mature, and operationalize them for a broad community.