Cloud Data Demo

Rich Signell
USGS, Woods Hole, MA
…and the Pangeo Community

Coastal Coupling Community of Practice
2021-02-22

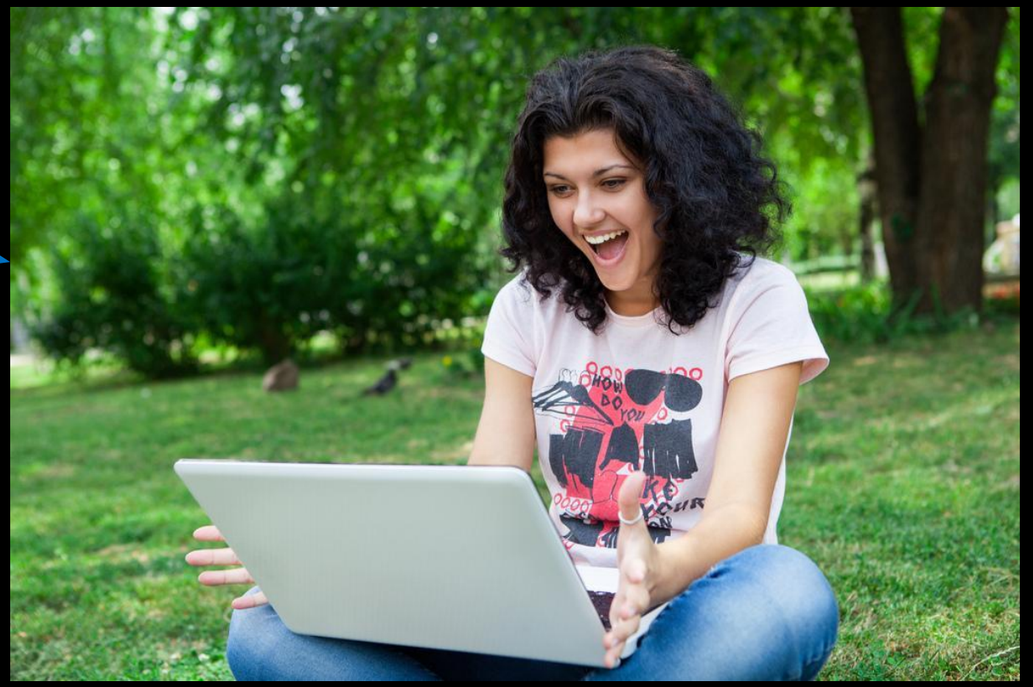
Open Architectures for Cloud-Native Earth System Analytics

Rich Signell
USGS, Woods Hole, MA
…and the Pangeo Community

Coastal Coupling Community of Practice
2021-02-22

Traditional Model Data Analysis

Model Data Analysis on the Cloud



Closed Platforms vs. Open Architectures

Closed platforms aim to bring all the data into a central location and provide tools for users to perform analysis on the data.

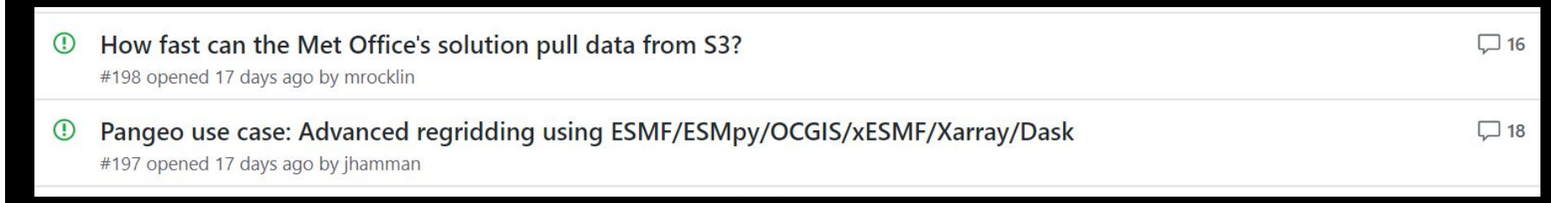
Examples: Google Earth Engine Descartes Labs Platform, Copernicus, DesignSafe-CI

Open architectures assume data will be distributed and seek interoperability between different data catalogs and computational tools.

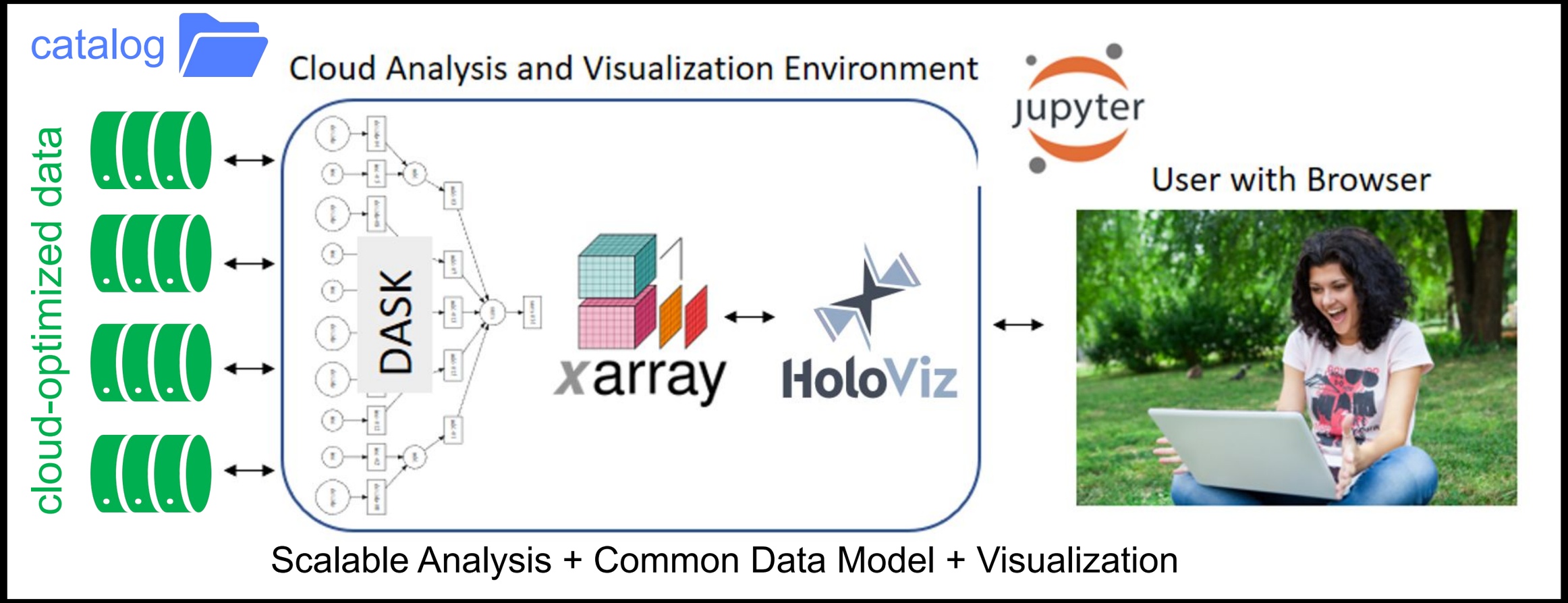
Examples: Pangeo, Open Data Cube

Medium Blog Post on Closed Platforms vs. Open Architectures by Ryan Abernathy and Joe Hamman

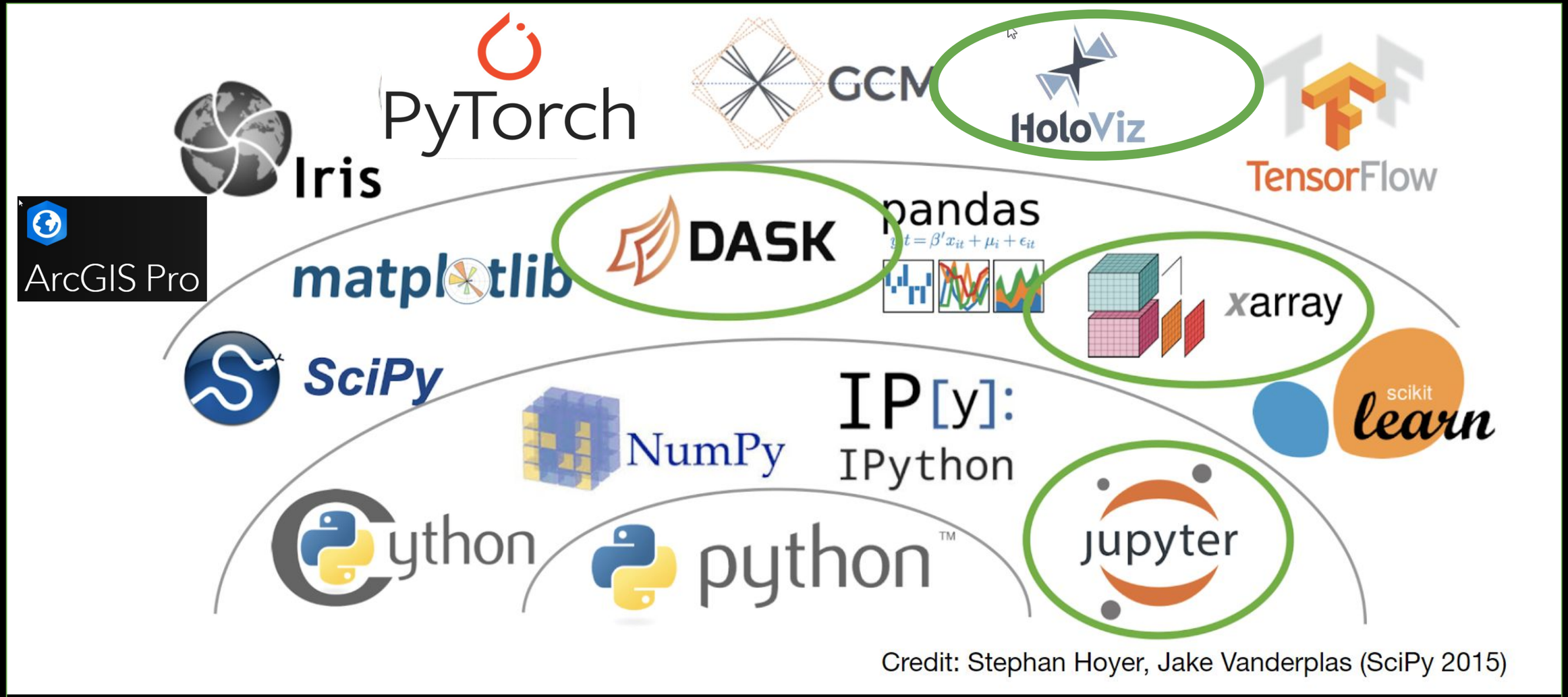
Pangeo is a Global Community



Pangeo Cloud Architecture



Pangeo is a Flexible Open-Source Framework



Cloud-optimized data

To overcome latency of object storage and take advantage of multiple cpus:

- Metadata can be read with one or just a few reads
- Data is written in chunks with compression
- Chunks are big enough so that latency is not a significant fraction of the read time (e.g. ~100mb)
- Chunk shape should be chosen based on expected use cases (e.g. not always 1 in the time dimension)

Cloud-optimized data formats

- Zarr: ndarray data (e.g. met/ocean/hydro model output, time stacks of remote sensing data)
- Cloud-optimized GeoTIFF (COG): geospatial image data
- Parquet: tabular data
- Entwine Point Tile: point cloud data

Note: other formats like NetCDF4/HDF5, GRIB2 typically don’t perform well with native readers, but there is a way...

It’s possible to use other formats effectively on the cloud also...

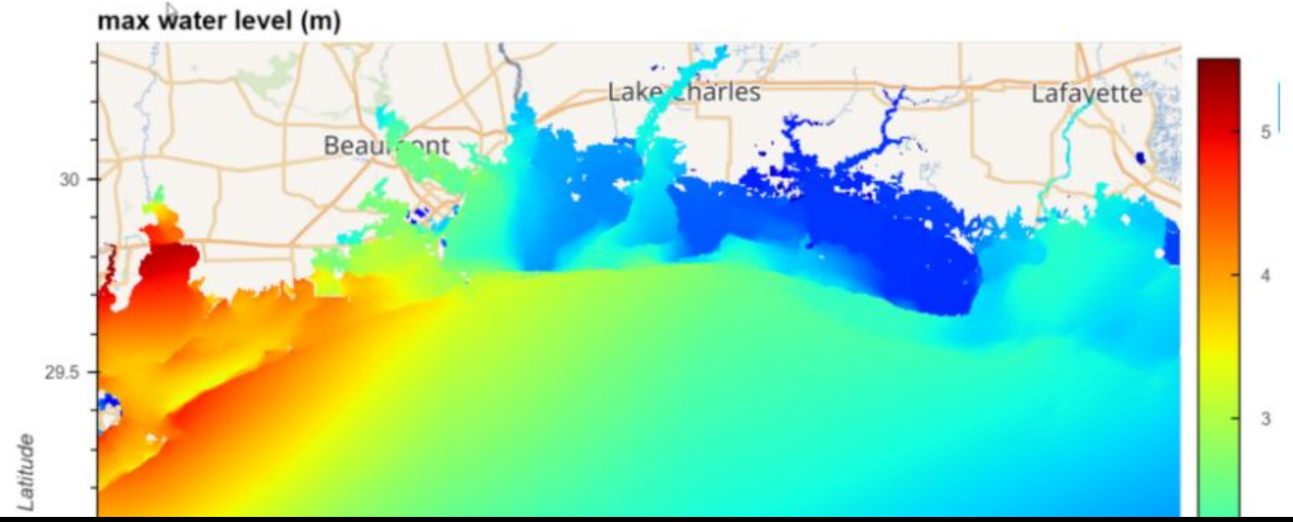
ABOUT PANGEO | WEBSITE

Medium

Cloud-Performant NetCDF4/HDF5 with Zarr, Fsspec, and Intake

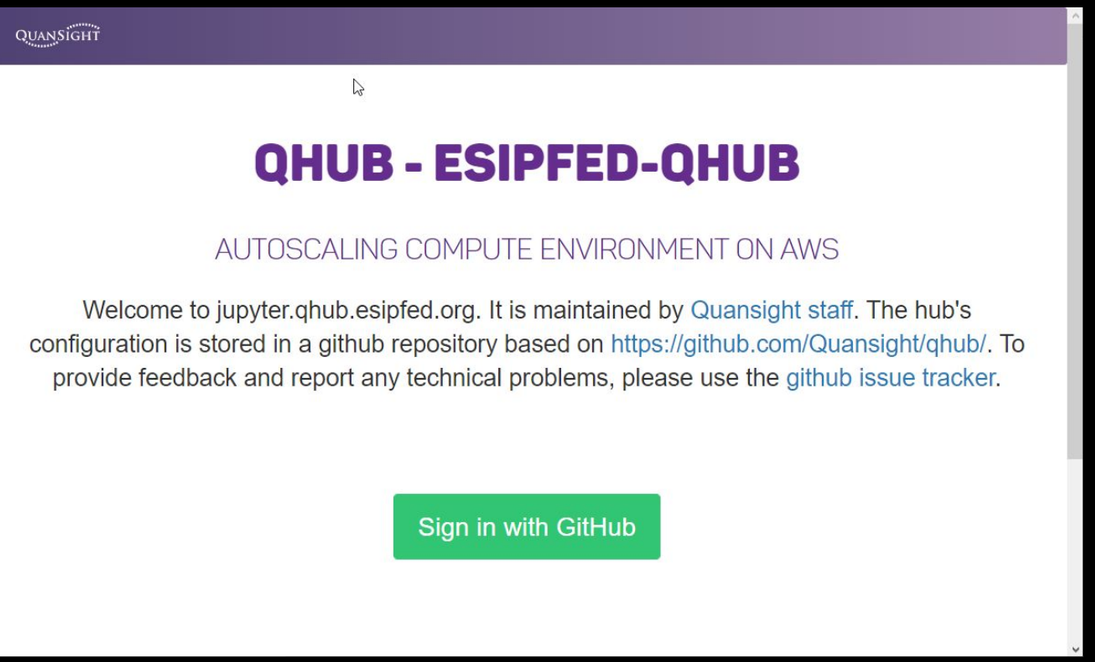
Richard Signell Follow

Dec 14, 2020 · 6 min read



Cloud Data Pilot Project

thanks to ESIP, AWS and Qhub

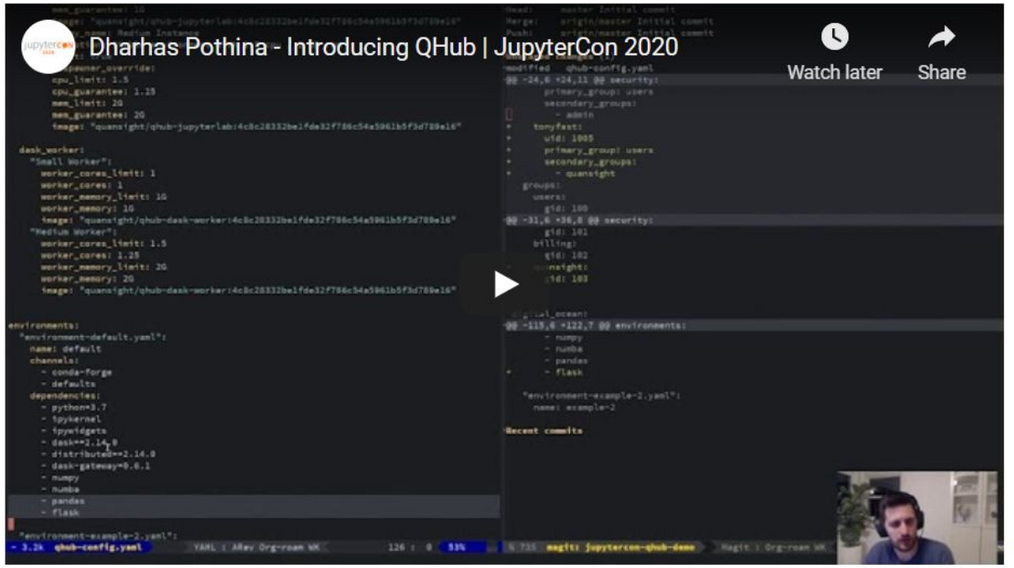


Announcing QHub

Updated: Oct 23, 2020

Today, we are announcing the release of QHub, a new open source project from Quansight that enables teams to build and maintain a cost-effective and scalable compute/data science platform in the cloud or on-premises. QHub can be deployed with minimal in-house DevOps experience.

See the demonstration here:



Conclusion: Open Cloud Architecture advantages

- No downloading to local machines
- No local compute infrastructure
- Pay only for what you use
- Unmatched compute power
- Builds Community
- Reproducible
- Inclusive