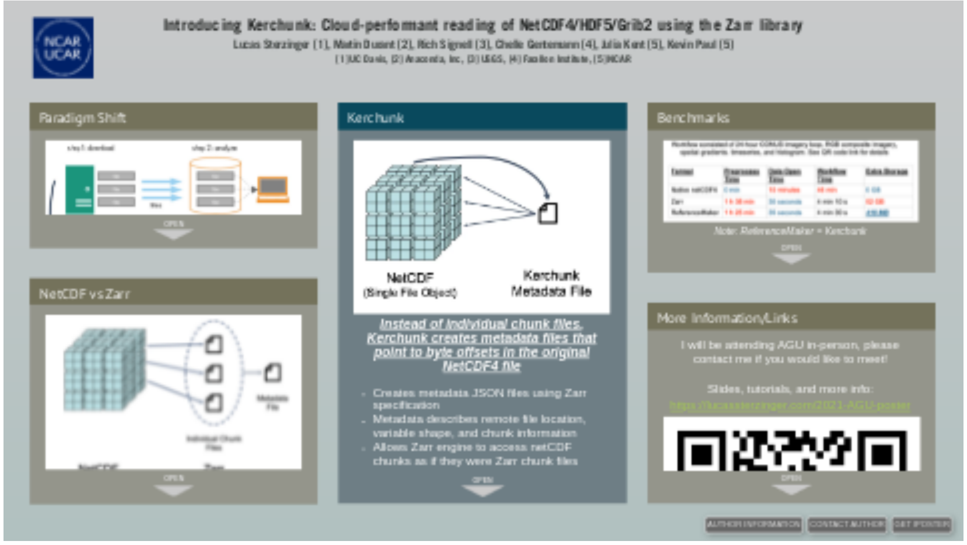
Introducing Kerchunk: Cloud-performant reading of NetCDF4/HDF5/Grib2 using the Zarr library



Introducing Kerchunk: Cloud-performant reading of NetCDF4/HDF5/Grib2 using the Zarr library

Lucas Sterzinger [1], Martin Durant [2], Rich Signell [3], Chelle Gentemann [4], Julia Kent [5], Kevin Paul [5]

[1] UC Davis, [2] Anaconda, Inc, [3] USGS, [4] Farallon Institute, [5] NCAR

Paradigm Shift

OPEN

NetCDF vs Zarr

OPEN

Kerchunk

Instead of individual chunk files, Kerchunk creates metadata files that point to byte offsets in the original NetCDF4 file

- Creates metadata JSON files using Zarr specification
- Metadata describes remote file location, variable shape, and chunk information
- Allows Zarr engine to access netCDF chunks as if they were Zarr chunk files

OPEN

Benchmarks

[illegible]

Note: ReferenceMaker = Kerchunk

OPEN

More Information/Links

I will be attending AGU in-person, please contact me if you would like to meet!

Slides, tutorials, and more info:
https://lucassterzinger.com/2021-AGU-poster

OPEN

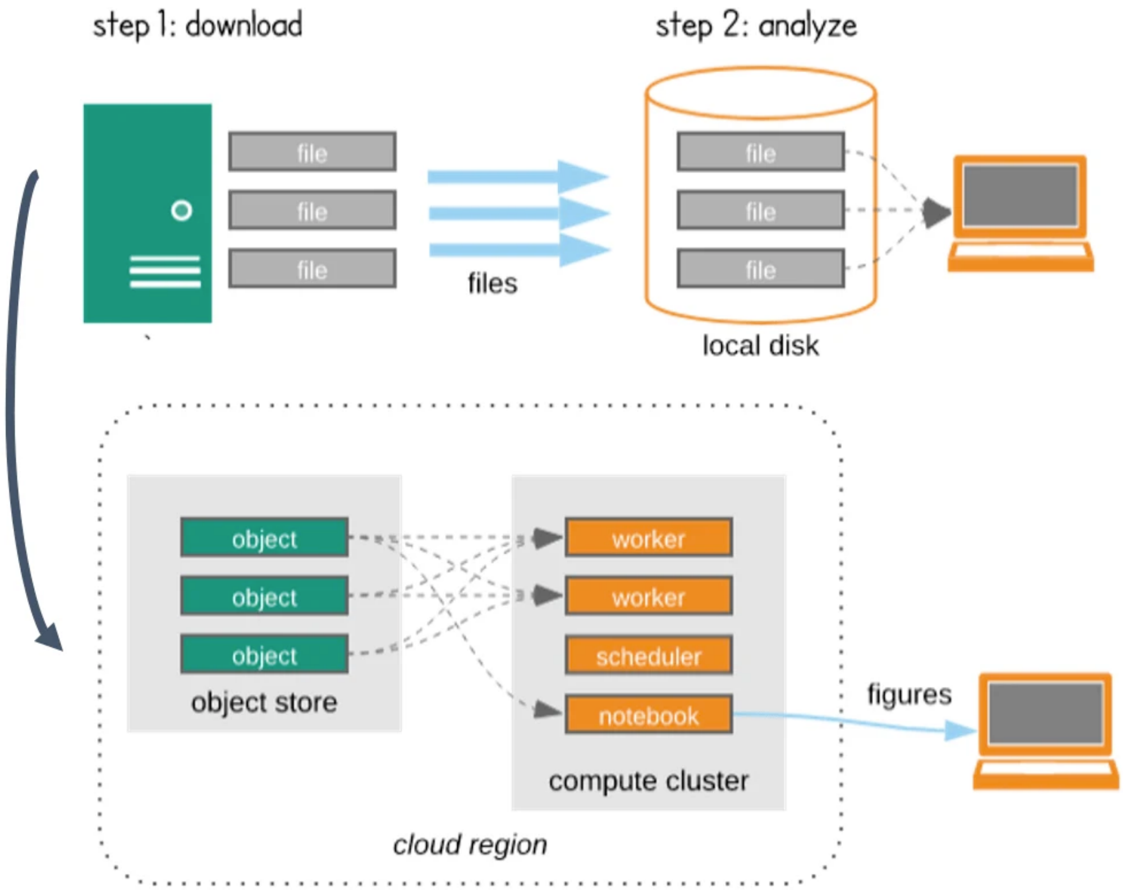
AUTHOR INFORMATION | CONTACT AUTHOR | GET IPOSTER

Lucas Sterzinger [1], Martin Durant [2], Rich Signell [3], Chelle Gentemann [4], Julia Kent [5], Kevin Paul [5]

[1] UC Davis, [2] Anaconda, Inc, [3] USGS, [4] Farallon Institute, [5] NCAR

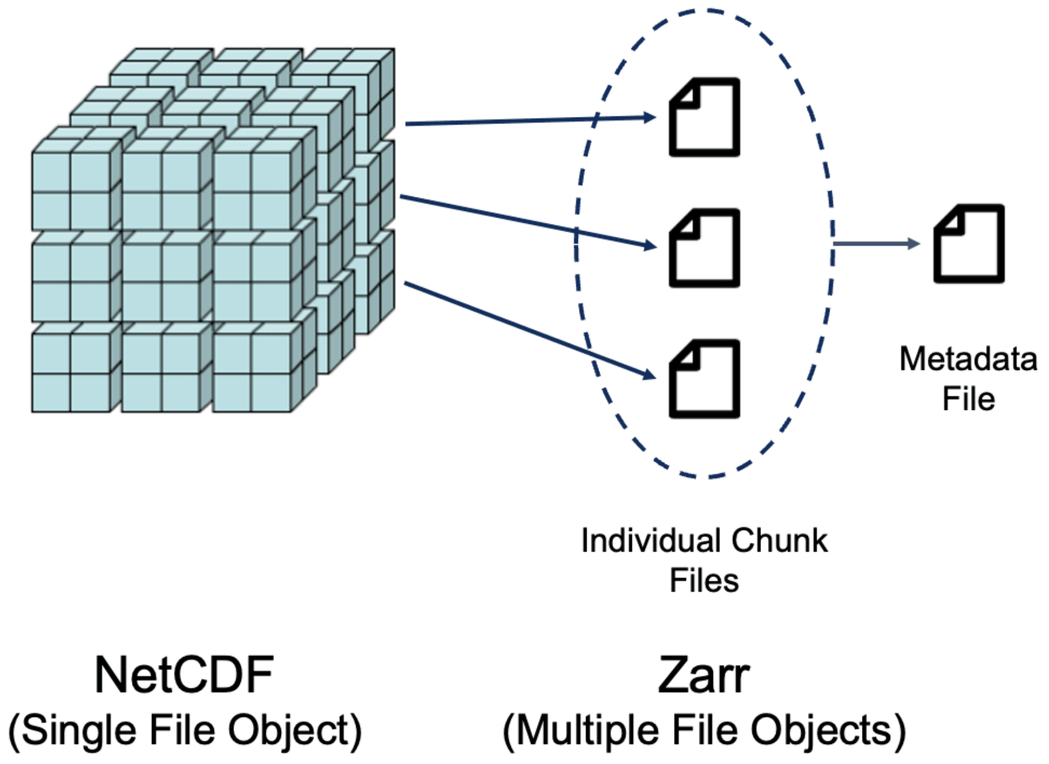
PRESENTED AT:

PARADIGM SHIFT



With the increasing size of scientific datasets, data analysis workflows are moving from downloading data to a local machine for analysis to doing the analysis on cloud computing services within the same region as the data being hosted.

NETCDF VS ZARR



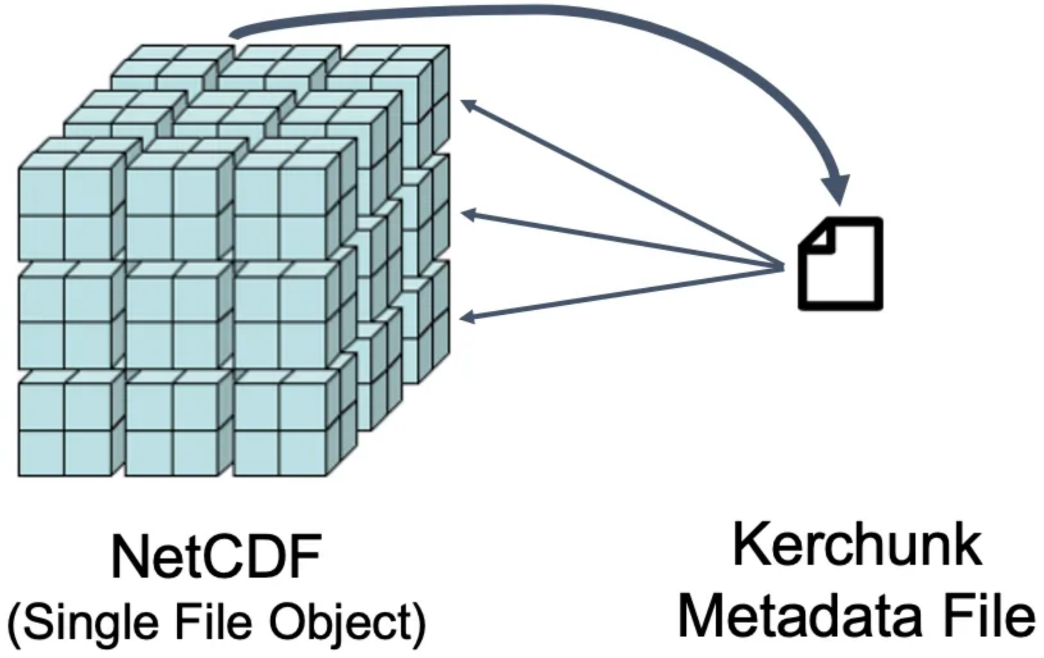
NetCDF4 files are stored as a single file object, often comprised of several data chunks. Metadata is scattered throughout the file, requiring many small reads. In a cloud environment, building metadata in this way is not efficient.

Zarr breaks up data chunks into individual files, with metadata describing them stored in plaintext. This format is gaining popularity for cloud-hosted data since metadata can be read quickly and individual chunks are available to be downloaded in parallel.

However, many cloud-hosted datasets are still uploaded in their original NetCDF format, so a way to more quickly access these files is needed.

KERCHUNK

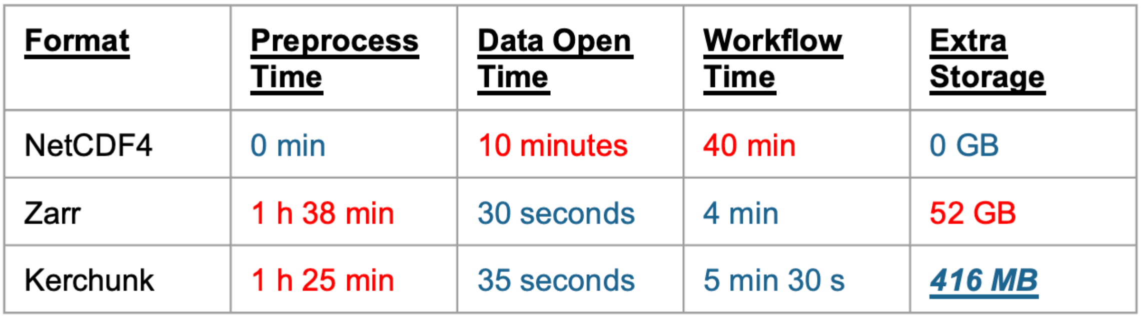
[VIDEO] https://res.cloudinary.com/amuze-interactive/video/upload/vc_auto/v1639596604/agu-fm2021/D8-E7-AB-FD-34-05-5D-2E-97-90-DA-10-DC-8C-83-7E/Video/agu-kerchunk-recorded-smaller_kf6how.mp4



<u>Instead of individual chunk files, Kerchunk creates metadata files that point to byte offsets in the original NetCDF4 file</u>

- Creates metadata JSON files using Zarr specification
- Metadata describes remote file location, variable shape, and chunk information
- Allows Zarr engine to access netCDF chunks as if they were Zarr chunk files
- Metadata file is only a few MB per data file and easily shareable
- Can be generated/hosted by 3rd parties

BENCHMARKS



Format	**Preprocess Time**	**Data Open Time**	**Workflow Time**	**Extra Storage**
NetCDF4	0 min	10 minutes	40 min	0 GB
Zarr	1 h 38 min	30 seconds	4 min	52 GB
Kerchunk	1 h 25 min	35 seconds	5 min 30 s	***416 MB***

Workflow consisted of GOES-16 24-hour CONUS imagery loop, RGB composite, spatial gradients, timeseries, and histogram. See link in section below for more details

Huge speed boost with low storage cost!

- Harness the cloud optimization of Zarr without needing to convert any data
- Reference files can be also created and hosted by third parties

MORE INFORMATION/LINKS

I will be attending AGU in-person, please contact me if you would like to meet!

See Kerchunk in action! Access slides, tutorials, and more info:

https://lucassterzinger.com/2021-AGU-poster (https://lucassterzinger.com/2021-AGU-poster)

[VIDEO] https://res.cloudinary.com/amuze-interactive/image/upload/f_auto,q_auto/v1638473113/agu-fm2021/d8-e7-ab-fd-34-05-5d-2e-97-90-da-10-dc-8c-83-7e/image/qr_ialpbo.mp4

AUTHOR INFORMATION

Lucas Sterzinger - UC Davis

lsterzinger@ucdavis.edu

@lucassterzinger (https://twitter.com/lucassterzinger)