Maintaining netCDF:
Updating Java Tutorial Code and Performance Testing in Python

UCAR | COMMUNITY PROGRAMS

Isabelle Pfander
Mentors: Hailey Johnson, Sean Arms
Supervisor: Ryan May

NSF

unidata

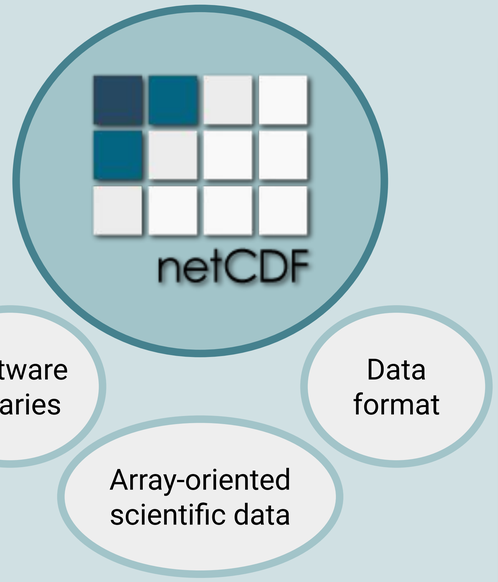
1. Updating netCDF Tutorial Code

Network Common Data Form (netCDF) is a combination of software libraries and APIs describing a data model for scientific multidimensional arrays.

Maintaining this codebase requires documentation, user support, testing, and adapting to new technologies. I maintained netCDF-Java documentation by updating Java tutorial code, testing code snippets, and modernizing tutorial texts to improve user understanding.

Creating tutorial classes → Updating & writing code → Testing

Scan the QR code to view online tutorials



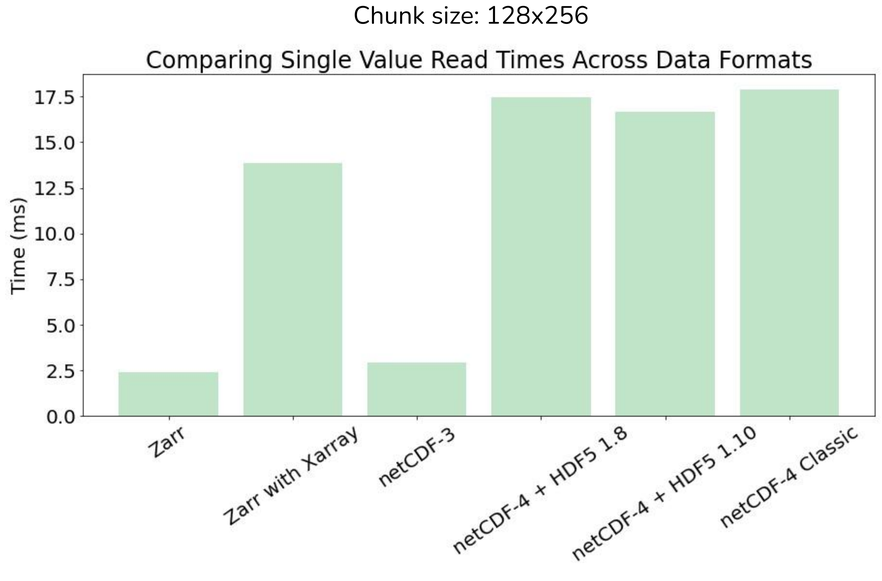
2. Comparing Data Formats

HDF5 is a file format used by netCDF-4 providing compression and chunking to the netCDF data model. Zarr is a Python-based storage format, which has support in netCDF-C and will soon in netCDF-Java.

This project compares read times for the data formats below:

- netCDF-3: no compression or chunking
- netCDF-4: zlib compression
- netCDF-4 Classic: zlib compression
- Zarr: Blosc compression
- Zarr & netCDF read with Xarray package

3. Benchmarking Results

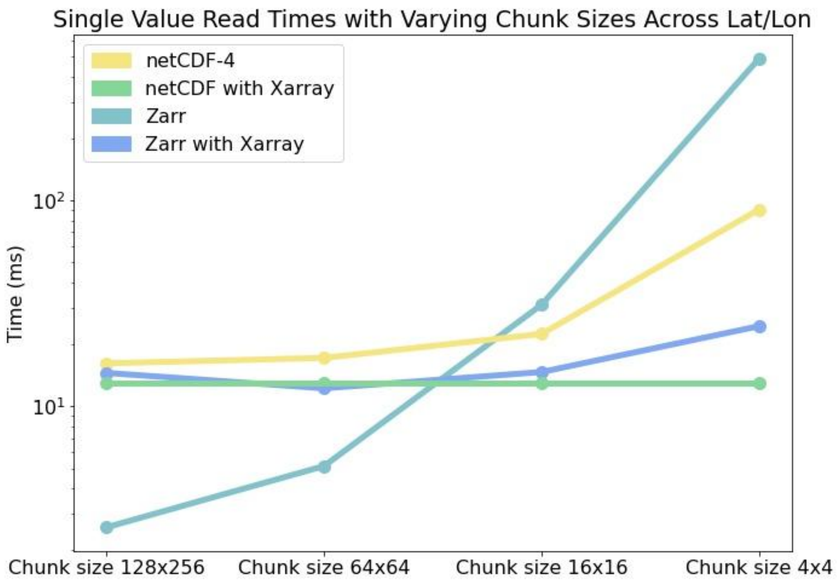


A netCDF-4 file stores all data in one .nc file, consequently more operations are needed to find the appropriate data, but only one open is required.

Zarr directory stores save chunked data as many subdirectories and files. The more chunks, the more individual files in one Zarr directory store. With small chunk size, this resulted in more time spent opening than reading.

In some cases, the Xarray Python package can read faster with both netCDF-4 and Zarr directory stores until loading is specified.

Note: read speeds will vary on an object store. These comparisons are for a posix file system only.



Future Work

- Compare reads for netCDF-Java HDF5 and Zarr implementations
- Test on datasets with varying dimensions and size

Scan to view code

Acknowledgements

Thank you to Unidata for the mentorship and community during this internship. This work was funded by the National Science Foundation (Grant AGS-1901712).