Cloud Analysis Runtime Comparison Between Native and Cloud-Optimized Data Formats: Project Plan

Intern: Matthew Thompson
Start Date: June 14th, 2021

Mentors:
Jorge Vazquez, jorge.vazquez@jpl.nasa.gov
Catalina Oaida, catalina.oaida@jpl.nasa.gov
Michelle Gierach, michelle.gierach@jpl.nasa.gov

Background

Motivation

The age of "Big Data" and "Cloud Computing" is upon us, and with these world-changing, new technologies comes new challenges for the user community. At PO.DAAC, the Surface Water and Ocean Topography (SWOT) mission (2022 expected launch date) will be one of the first NASA missions to deliver 20 TB of data per day! In order to handle this much information, PO.DAAC is transitioning its data storage to Amazon Web Services' (AWS) cloud servers. This means that users may have to transition from doing analyses on their laptops to instead on the cloud where they are charged for computation time. When working with extremely large datasets, these fees can add up quickly due to extensive analysis time. A potential solution is to provide datasets in a cloud-optimized format such as Zarr rather than the native netCDF4 with the hope that it will reduce cloud analysis time, and therefore user cost. The goal of this project is to use a proxy "big data" dataset, in this case the MUR 1-km Sea Surface Temperature dataset, to compare runtimes between native (netCDF4) and cloud-optimized (Zarr) formats across several cloud analysis scenarios.

Dataset

MUR 1-km Sea Surface Temperature (SST) dataset, global from June 2002 – Present

Objectives

Goals

To compare MUR 1-km SST dataset in…

- netCDF4
- Zarr, using netCDF4-to-Zarr converting services
- Zarr, data native to this format

…for several cloud analysis scenarios at different scales, including:

- Global SST time series from June 2002 – January 2020

- Global SST spatial plot averaged from June 2002 – January 2020
- Regional SST Anomaly time series – NW Pacific (Blob Region) from August 2019 – January 2020
- Regional SST Anomaly spatial plot – NW Pacific (Blob Region) averaged from August 2019 – January 2020
- Application of climatology to derive anomalies for Nino 3.4 box from January 2015 – March 2016 (2015–2016 El Nino)

Metrics

- Cloud analysis time
- Resource requirements (tools, cloud access libraries, etc. needed to employ each scenario)

Deliverables

- Feasibility assessment of using the cloud for analyzing MUR in scientific applications (i.e., advantages, disadvantages, pain points experienced, metrics captured for compute resources used, etc.)
- Runtime comparison of netCDF4 and Zarr capabilities to inform user requirements and PO.DAAC decisions

Approach

Overview

The deliverables for this project will be completed by testing analysis run times for different dataset formats and sizes in order to create benchmarks of data format effectiveness within the cloud. There are two principle steps to be accomplished within the 10-week timeline, the first of which likely being the most difficult. Learning how to analyze data on AWS and implementing the code that enables the planned tests will be significant challenges due to the steep learning curve of cloud computing. There have been 3 weeks allocated to this preliminary step allowing for ample time for research and the development of understanding. The remaining step involves running all the tests for each format and is estimated to take 6 weeks, or 2 weeks per format. Completion of this project will not depend on results from other projects.

Resources

- AWS PO.DAAC netCDF4 dataset (inherited)
- Earthdata Harmony netCDF4-to-Zarr conversion services (inherited)
- AWS PO.DAAC Zarr dataset (inherited)
- JPL funds to pay AWS fees for computation time while testing analyses (inherited)

Project Schedule

Week of 06/14/21:

- Get access to MUR 1-km SST dataset stored on AWS
- Get access to MUR 1-km SST in POCLOUD from Earthdata Cloud in AWS
- Set up access to AWS computing
- Begin learning how to use AWS for data analysis

Week of 06/21/21:

- Finish learning how to use AWS for data analysis
- Begin implementation of test code

Week of 06/28/21:

- Continue implementation of test code

Week of 07/05/21:

- Run netCDF4 tests:
    - Regional SST Anomaly time series – NW Pacific (Blob Region) from August 2019 – January 2020
    - Regional SST Anomaly spatial plot – NW Pacific (Blob Region) averaged from August 2019 – January 2020
    - Application of climatology to derive anomalies for Nino 3.4 box from January 2015 – March 2016 (2015–2016 El Nino)

Week of 07/12/21:

- Run netCDF4 tests:
    - Global SST time series from June 2002 – January 2020
    - Global SST spatial plot averaged from June 2002 – January 2020

Week of 07/19/21:

- Run tests with Zarr using netCDF4-to-Zarr converting services:
    - Regional SST Anomaly time series – NW Pacific (Blob Region) from August 2019 – January 2020
    - Regional SST Anomaly spatial plot – NW Pacific (Blob Region) averaged from August 2019 – January 2020
    - Application of climatology to derive anomalies for Nino 3.4 box from January 2015 – March 2016 (2015–2016 El Nino)

Week of 07/26/21:

- Run tests with Zarr using netCDF4-to-Zarr converting services:
    - Global SST time series from June 2002 – January 2020
    - Global SST spatial plot averaged from June 2002 – January 2020

Week of 08/02/21:

- Run tests on datasets native to Zarr:
    - Regional SST Anomaly time series – NW Pacific (Blob Region) from August 2019 – January 2020

    - Regional SST Anomaly spatial plot – NW Pacific (Blob Region) averaged from August 2019 – January 2020
    - Application of climatology to derive anomalies for Nino 3.4 box from January 2015 – March 2016 (2015–2016 El Nino)

Week of 08/09/21:

- Run tests on datasets native to Zarr:
    - Global SST time series from June 2002 – January 2020
    - Global SST spatial plot averaged from June 2002 – January 2020

Week of 08/16/21:

- Synthesize results from all tests into final takeaway
- Complete project report

References

MUR 1-km Dataset

- PO.DAAC netCDF dataset on premise → provided only for dataset reference materials and documentation, https://podaac.jpl.nasa.gov/MEaSUREs-MUR?tab=background&sections=about%2Bdata
- AWS PO.DAAC netCDF4 dataset, https://search.earthdata.nasa.gov/search?q=POCLOUD%20MUR
- AWS PO.DAAC Zarr dataset, https://registry.opendata.aws/mur/

PO.DAAC & NASA Earthdata Cloud-Based Services

- Harmony → offers netCDF-Zarr conversion service, https://harmony.earthdata.nasa.gov/
- Examples of using Harmony netCDF-Zarr conversion service, https://github.com/podaac/tutorials/blob/master/notebooks/SWOT-EA-2021/Estuary_explore_inCloud_zarr.ipynb
- PO.DAAC Cloud Migration Information, https://podaac.jpl.nasa.gov/cloud-datasets/about

Research Papers

- Vazquez-Cuervo, J.; Torres, H.S.; Menemenlis, D.; Chin, T.; Armstrong, E.M. Relationship between SST gradients and upwelling off Peru and Chile: model/satellite data analysis. *Int. J. Remote Sens*. **2017**, *38:23*, 6599-6622. DOI: 10.1080/01431161.2017.1362130
- Vazquez-Cuervo, J.; Gomez-Valdes, J.; Bouali, M. Comparison of Satellite-Derived Sea Surface Temperature and Sea Surface Salinity Gradients Using the Saildrone California/Baja and North Atlantic Gulf Stream Deployments. *Remote Sens.* **2020**, *12*, 1839. DOI: 10.3390/rs12111839
- Gentemann, C.L.; Fewings, M.R.; García-Reyes, M. Satellite sea surface temperatures along the West Coast of the United States during the 2014-2016 northeast Pacific marine heat wave. *Geophys. Res. Lett*. **2017**, *46:1*, 312-319. DOI: 10.1002/2016GL071039