NOAA
SCIENCE
ADVISORY
BOARD

DAARWG REPORT ON NESDIS CLOUD ARCHIVE PROJECT

2022 Nov 30

Jeff de La Beaujardière, DAARWG chair

General Comments

- DAARWG thanks Dr. Monica Youngman (NESDIS/NCEI) and her team for briefings and Q&A sessions.
- DAARWG supports NCEI's plan to use commercial cloud resources rather than on-prem infrastructure for primary storage and access of archival data.
- DAARWG believes this could yield lower operating costs and provide better proximity to scalable computing resources.

1. Clarify Motivation

- DAARWG recommends that NOAA **clarify which benefits it is seeking to maximize or optimize** in the NCCF cloud archive project.
  - Is it ability to compute in place, decommissioning of on-prem hardware, efficient I/O, broader data access for external customers, proximity to non-NOAA data, staff costs, or other attributes?
  - The goal should be stated along with quantitative metrics to assess whether they have eventually been met.

2. Refine Migration Plan

- DAARWG suggests that NOAA **consider more detailed planning regarding what data are migrated in what order**, including contingencies for unexpected delays.
  - 50 → 150 PB data volume evolution FY2022-2026 is significant
- 2(a) Clarify the method(s) to be used for the actual bulk data transfer and data integrity verification.
- 2(b) Provide more detail on the incremental solutions in terms of functionality. Copy as-is, with possible later optimizations? Or optimize data prior to initial migration?

3. Ensure Exit Plan

- DAARWG suggests NOAA include some **provision for an exit strategy** to be built into the contract with the initial cloud vendor.
- Concerns could include data egress costs, data transfer/copy/verification methods, time and effort to perform a bulk transfer to get everything out, and whether to retain on-premises disaster recovery copy of data on tape just in case.

4. Explain Data Catalog Options

- DAARWG suggests that **Spatio-Temporal Asset Catalog (STAC) be considered** as an alternative to NASA Consolidated Metadata Repository (CMR)

  - STAC may have broader community of support and use

  - Related question: What will happen to existing NOAA Data Catalog and NOAA OneStop projects?

5. Consider Optimized Data Formats

- In order to enhance data when transferred to the cloud, DAARWG recommends that NOAA **consider data optimizations** such as:
  - organizing datasets to provide a more holistic, multi-dimensional data "cube" view of data rather than individual files;
  - using cloud-optimized formats such as Zarr or COG; or
  - storing as-is with structural metadata such as ncZarr or kerchunk.

6. Define Terminology

- Recommendation 6: DAARWG recommends that NOAA **clarify what vAIP and KG are and indicate their value** to either NOAA or users of implementing these.

- The NCAP presentations include at least two concepts not normally found in archive reference model:
  - Virtual Archive Information Package (vAIP) - AIP is an OAIS Reference Model term; how does a vAIP differ?
  - Knowledge Graph (KG) - KG can be a semantic web concept; is this what is meant? How will users use them?

7. Use or Contribute to Open Source Software

- Recommendation 7: DAARWG urges NOAA to **consider releasing code as open source**, and to **contribute back to the community** any enhancements NOAA may make to existing open source projects.

- NOAA will need to write and leverage a considerable amount of code to support the Cloud Archive Project.

DISCUSSION