DAARWG report on NESDIS Cloud Archive Project

1. Introduction

The Data Archiving and Access Requirements Working Group (DAARWG) of the NOAA Science Advisory Board (SAB) greatly appreciates the briefings received from Dr. Monica Youngman and her team concerning the NESDIS Cloud Archive Project (NCAP) within the NESDIS Common Cloud Framework (NCCF). DAARWG supports NCEI's plan to use commercial cloud resources rather than on-prem infrastructure for primary storage. DAARWG believes this could yield lower operating costs including both hardware acquisition and effort to ensure data preservation, and will provide better proximity to scalable cloud computing resources.

DAARWG has several comments and questions and welcomes NOAA's thoughts and responses.

2. Findings and Recommendations

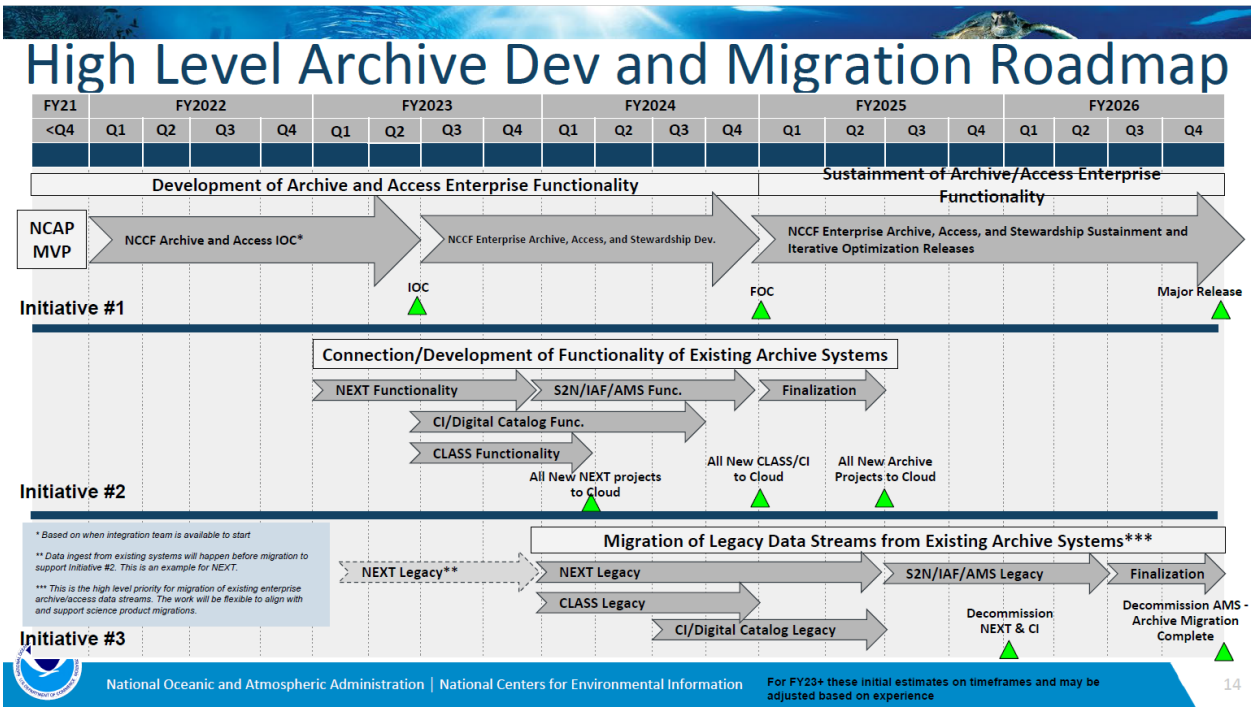
2.1. Clarify Motivation

There are various reasons why NOAA may wish to migrate its archives to the cloud. Is it the 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?

Recommendation 1: DAARWG recommends that NOAA clarify which benefits it is seeking to maximize or optimize in the NCCF project. The goal(s) should be stated along with quantitative metrics to assess whether they have eventually been met.

2.2. Refine Migration Plan

The National Centers for Environmental Information (NCEI) archive data volume is currently 40-50 PB, according to the briefings DAARWG received, with projected growth to ~150 PB during the time period FY 2022-2026 shown on the "High Level Archive Dev and Migration Roadmap."



That Roadmap does suggest an overall sequencing of datasets but with little detail and multiple data streams being migrated at the same time. This is a large data volume, and there will likely be complexities and optimizations discovered as the earlier datasets are transferred.

Recommendation 2: DAARWG suggests that NOAA consider some more detailed planning regarding what data are migrated in what order, including contingencies for unexpected delays.

2(a) DAARWG recommends that NOAA clarify the method(s) to be used for the actual bulk data transfer (basic S3 API, Globus, AWS Snowballs or Snowmobile) and how the data integrity will be verified after the transfer.

2(b) DAARWG recommends that NOAA provide more detail on the incremental solutions in terms of functionality. Will NOAA lift-and-shift data first, simply treating the cloud as storage for as-is data initially, with later optimizations to maximize the benefit of the cloud? Or will NOAA plan for and implement improved data and functionality prior or during the initial migration? The former approach may be faster but leave large collections of data not well organized; the latter might take more time but be better in the long run.

2.3. Ensure Exit Plan

DAARWG's understanding is that Amazon Web Services (AWS) will be the initial cloud vendor for this project, which is reasonable. Indeed, any of the major cloud vendors could doubtless support this project. What is the exit plan from the initial cloud provider should a change in future years be needed? Will there be an on-premises disaster recovery copy of data just in case, for example on tape?

Recommendation 3: DAARWG suggests NOAA include some provision for an exit strategy to be built into the contract with AWS. Concerns could include data egress costs (in particular, can they be waived in the event of a bulk migration out of the cloud), data transfer/copy/verification methods, and the time and effort to perform a bulk transfer to get everything out.

2.4. Explain Data Catalog Options

DAARWG understands that NCEI has decided to use the NASA Consolidated Metadata Repository (CMR) as the catalog of data in the cloud archive, and that NOAA would run a separate instance of NASA CMR. Potential other choices might have been the existing NOAA Data Catalog based on open-source CKAN software, or NOAA OneStop developed in-house by NCEI, or the Spatio-Temporal Asset Catalog (STAC). What is the rationale for choosing NASA CMR? Are there potential issues? What will happen to existing Catalog and OneStop projects?

Recommendation 4: DAARWG suggests that STAC be considered as an alternative to CMR, as it has a much broader community of support and use. Adopters include Google Earth Engine, Microsoft Planetary Computer, RadiantMLHub, and Sentinel Hub. While CMR can be translated to STAC, as exemplified by the NASA CMR-STAC proxy, the resulting STAC metadata is generalized. A direct STAC implementation could leverage community extensions and other specific STAC capabilities to better integrate with the community of tools made available to a STAC-compliant service. See more here: https://stacspec.org/en/.

2.5. Consider Optimized Data Formats

An important benefit of archiving data in the cloud is access to co-located scalable computing resources. However, data submitted to a traditional archive are often not organized, formatted, and documented in ways that best support machine readability, efficient I/O, and large-scale analysis. What are NOAA's plans regarding enhancement of data prior, during, or after transfer to the cloud?

Recommendation 5: In order to enhance data when transferred to the cloud, DAARWG recommends that NOAA consider optimizations of highly-utilized datasets, such as organizing data to provide a more holistic, multi-dimensional "datacube" view of data rather than individual files; using cloud-optimized formats such as Zarr or COG; storing as-is with structural metadata such as ncZarr or kerchunk; or other enhancements. As with the data migration step, any enhancement or format conversion should be followed by some kind of data verification. Documentation should also be provided with details of any enhancement or format conversion steps as well as how the verification step was done.

2.6. Define Terminology

The NCAP presentations include at least two concepts not normally found in archive reference model, notably Virtual Archive Information Package (vAIP) and Knowledge Graph (KG). AIP is an OAIS Reference Model term; how does a vAIP differ? KG can be a semantic web concept; is this what is meant, and in practice will the KGs be descriptive in nature (e.g., and RDF model) or operable functions in code?

Recommendation 6: DAARWG recommends that NOAA clarify what vAIP and KG are, with examples, and indicate the value to either NOAA or users of implementing these.

2.7. Use or Contribute to Open Source Software

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

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.

3. Conclusion

DAARWG applauds the NCAP effort and hopes NCEI will carefully consider these findings and recommendations.