



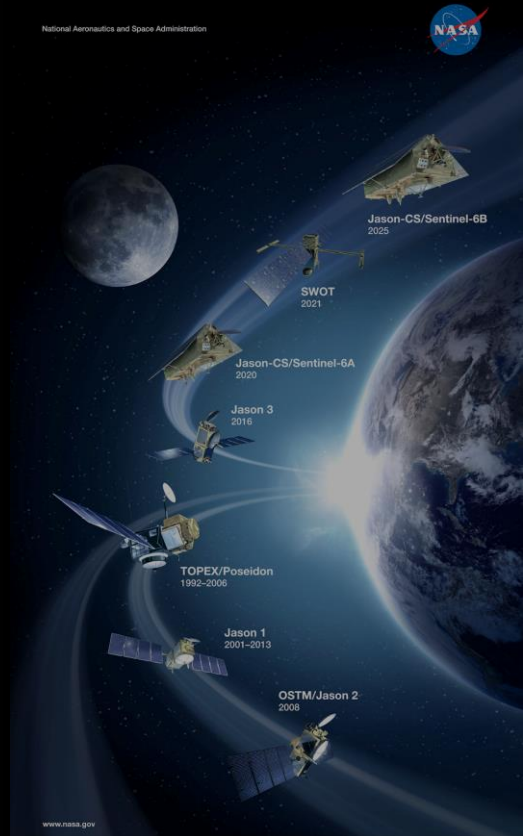
# Data Management in the Petabyte Era - PO.DAAC Journey to the Cloud

*Suresh Vannan, PO.DAAC team*

*Jet Propulsion Laboratory, California Institute of Technology*

*May 2nd, 2023*

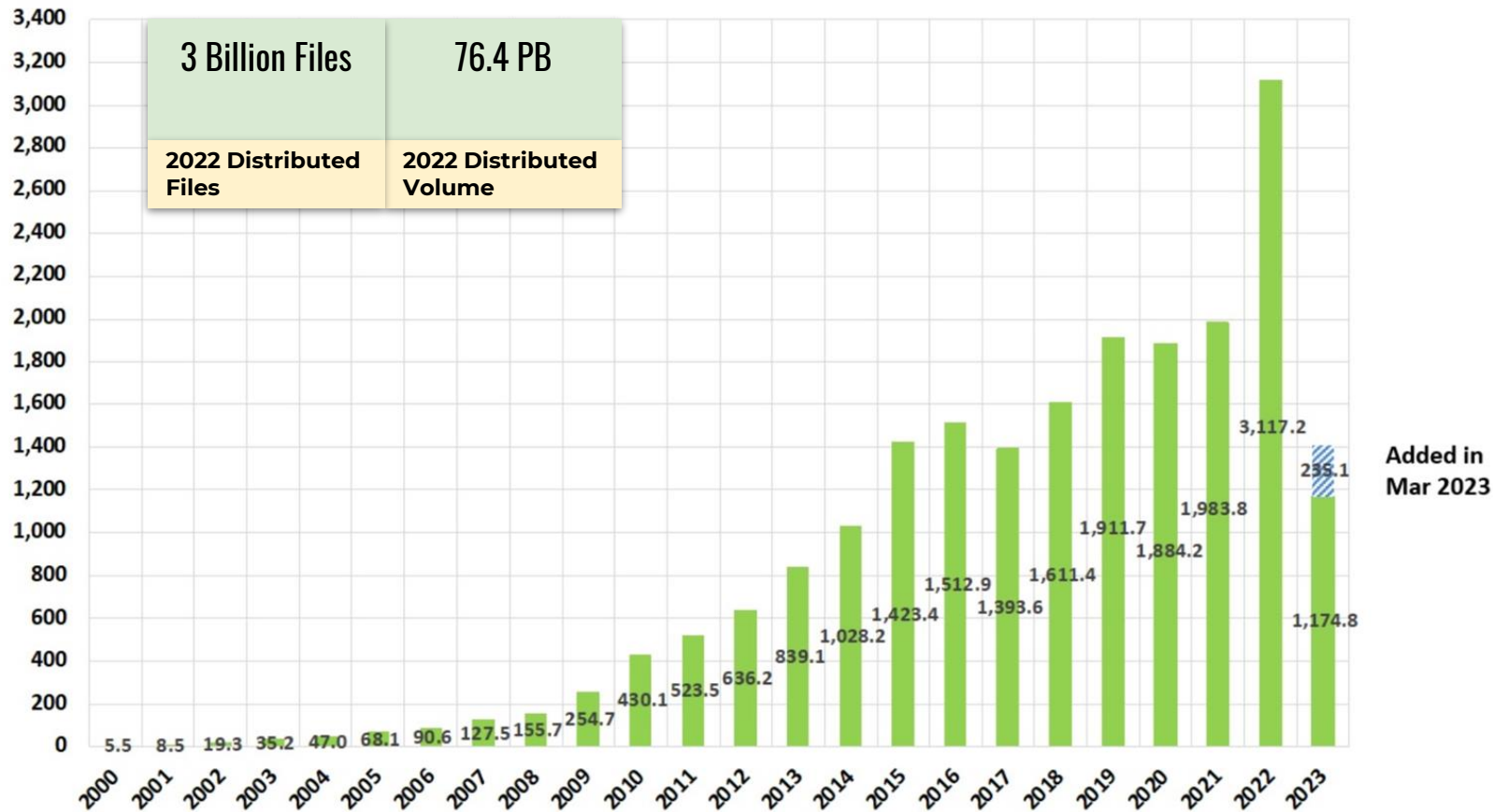
PV2023, European Organization for Nuclear Research  
(CERN) Geneva (Switzerland)



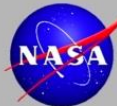


Apr 20 2023 16:48

Millions



# Distributed Active Archive Centers (DAACs)



## Alaska Satellite Facility DAAC

SAR Products,  
Sea Ice,  
Polar Processes,  
Geophysics

## Land Processes DAAC

Land Cover,  
Surface Reflectance,  
Radiance, Temperature  
Topography,  
Vegetation Indices

## Goddard Earth Sciences Data and Information Services Center

Global Precipitation,  
Solar Irradiance,  
Atmospheric Composition,  
and Dynamics, Global Modeling

## Socioeconomic Data and Applications Center

Human Interactions, Land Use,  
Environmental Sustainability,  
Geospatial Data

## National Snow and Ice Data Center DAAC

Frozen Ground,  
Glaciers, Ice Sheets,  
Sea Ice, Snow,  
Soil Moisture

## Physical Oceanography DAAC

Gravity, Sea Surface  
Temperature, Ocean  
Winds, Topography,  
Circulation & Currents

## Crustal Dynamics Data Information System

Space Geodesy,  
Solid Earth

## Oak Ridge National Laboratory DAAC

Biogeochemical Dynamics,  
Ecological Data,  
Environmental Processes

## Ocean Biology DAAC

Ocean Biology,  
Sea Surface  
Temperature

## Level 1 and Atmosphere Archive and Distribution System (LAADS) DAAC

MODIS Level-1 and  
Atmosphere Data Products

## Global Hydrometeorology Resource Center DAAC

Hazardous Weather,  
Lightning, Tropical Cyclones  
and Storm-induced Hazards

## Atmospheric Science Data Center

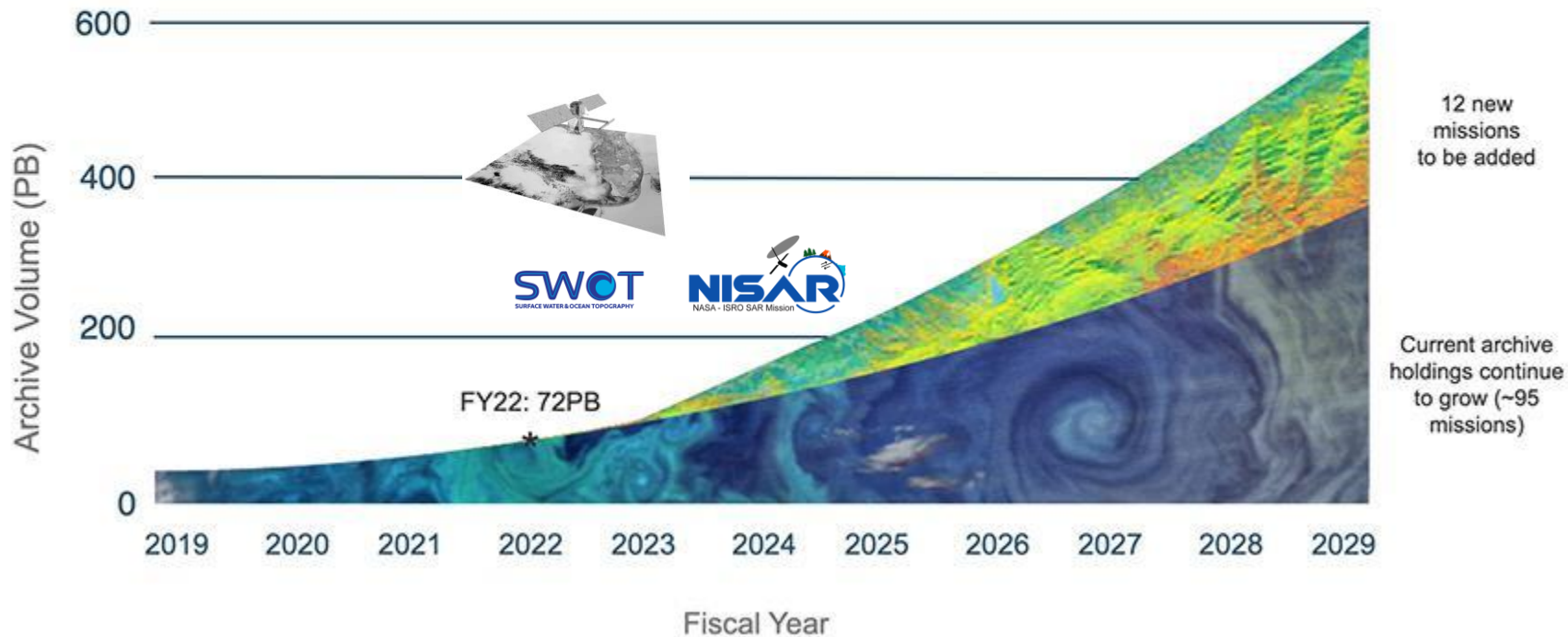
Radiation Budget,  
Clouds, Aerosols,  
Tropospheric Chemistry



<https://podaac.jpl.nasa.gov/>

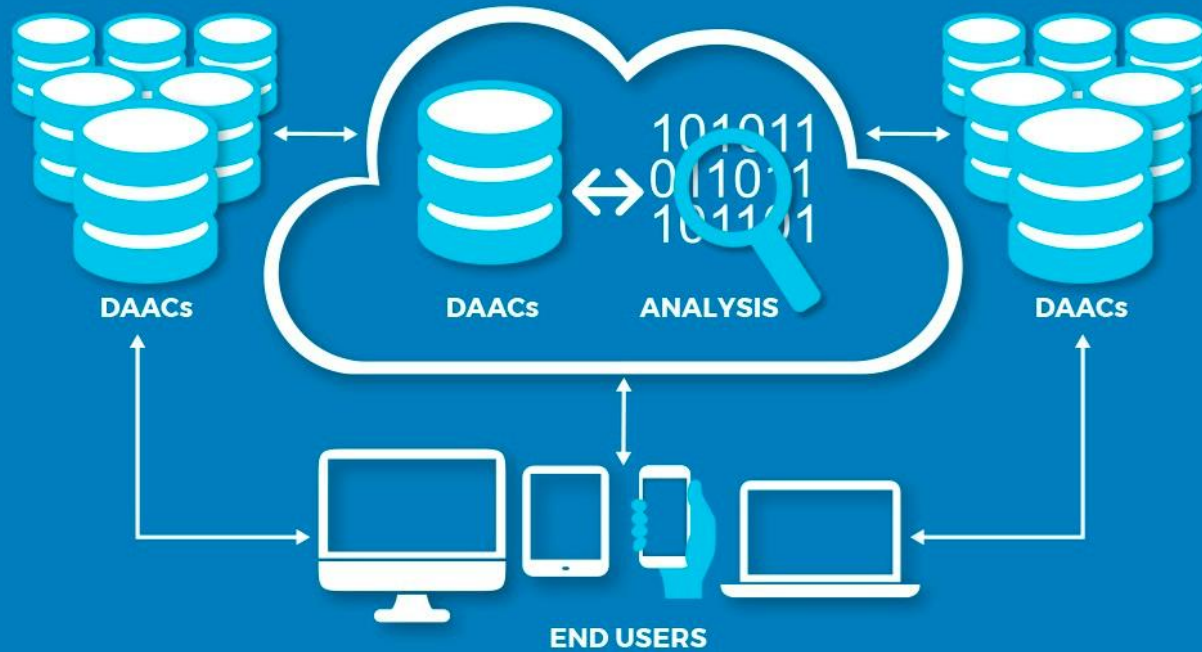
DAACs are custodians of EOS mission data and ensure that data will be easily accessible to users.

# Earth Science Data Archive Growth Projection



# A New Paradigm

The EOSDIS Cloud Evolution



# Benefits of the Cloud

- **Easy access to data:** Data users will be able to access data directly in the cloud, making the need to download volumes of data unnecessary.
- **Rapid deployment:** Users can bring their algorithms and processing software to the cloud and work directly with the data in the cloud
- **Scalability:** The size and use of the archive can expand easily and rapidly as needed.
- **Flexibility:** Mission needs can dictate options for selecting operating systems, programming languages, databases, and other criteria to enable the best use of mission data.
- **Reduced Duplication:** The use of a common infrastructure with cloud native services will reduce redundant tools and services.

# Challenges



## Cost

- Storage
- Egress
- Development
- Computational
- Labor



## Security

- Data protection
- Access control
- Cybersecurity



## Migration

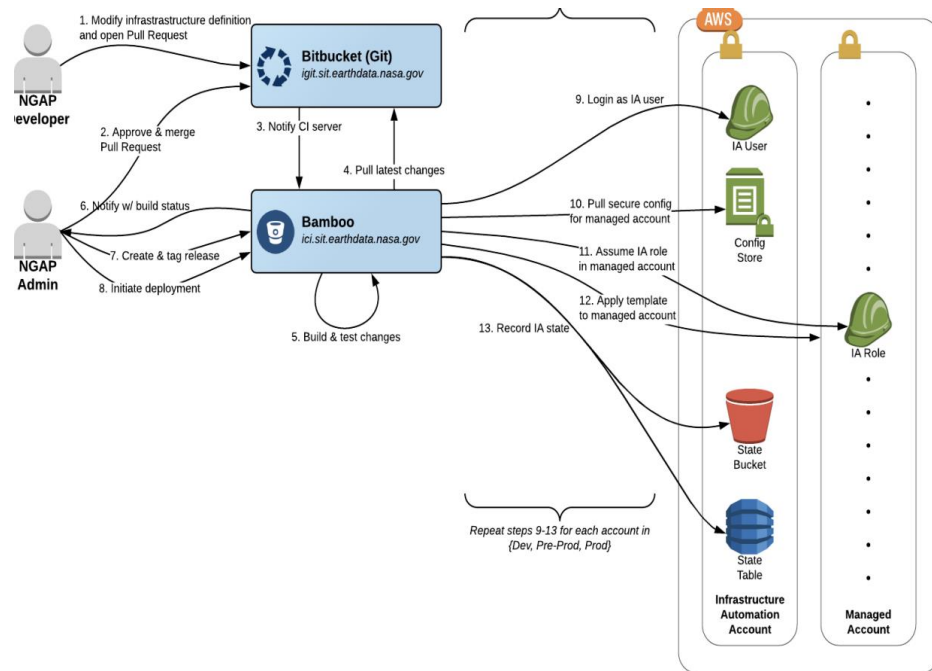
- Maintain existing system
- End-user
- Staffing
- Technical skill
- End-user migration

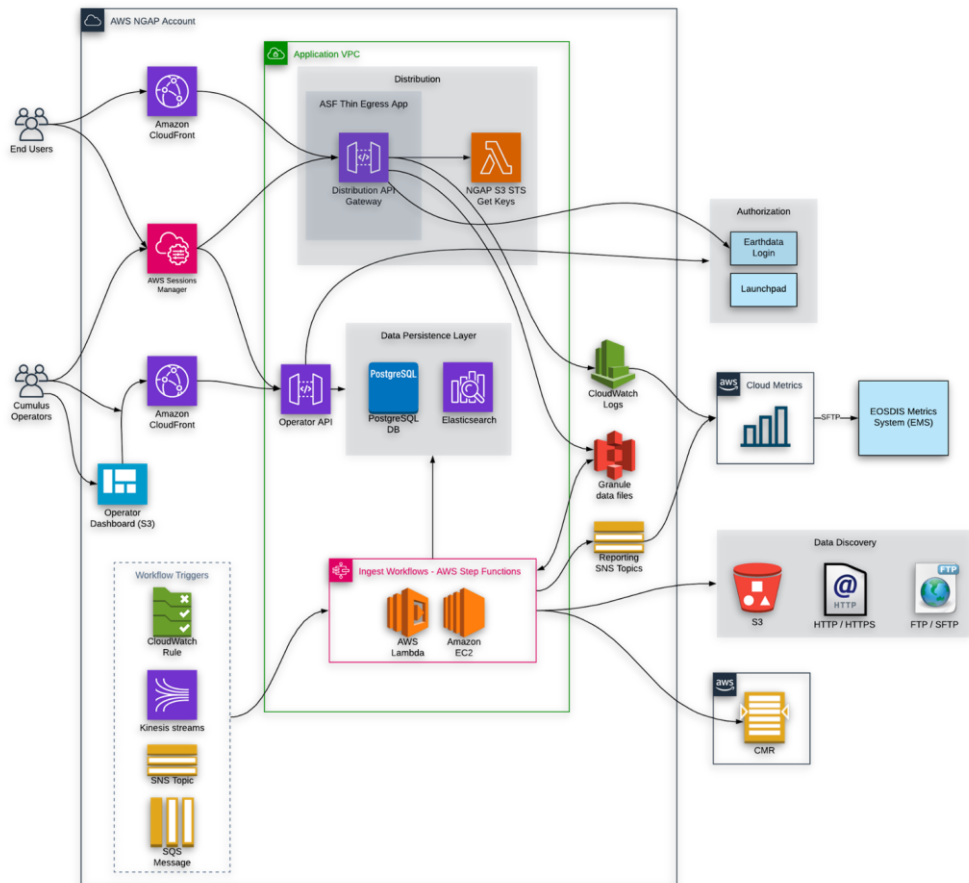


NGAP is a multi-account, Infrastructure-as-a-Service (IaaS) cloud platform operating on Amazon Web Services (AWS).

## Features:

1. NASA-Approved Amazon Web Services (AWS)
2. Code Deployment Services: DevOps Pipeline
3. Use of Infrastructure as Code: including reusable template





The screenshot shows the Cumulus Documentation v14.0.0 website. The navigation menu includes:

- Getting Started
- Introduction
- About Cumulus
- Overviews
- Deployment
- Configuration
- Development
- Workflow Tasks
- Features
- Troubleshooting
- Cumulus Development
- Integrator Guide
- Upgrade Notes
- External Contributions

The **Introduction** page content includes:

This Cumulus project seeks to address the existing need for a "native" cloud-based data ingest, archive, distribution, and management system that can be used for all future Earth Observing System Data and Information System (EOSDIS) data streams via the development and implementation of Cumulus. The term "native" implies that the system will leverage all components of a cloud infrastructure provided by the vendor for efficiency (in terms of both processing time and cost). Additionally, Cumulus will operate on future data streams involving satellite missions, aircraft missions, and field campaigns.

This documentation includes both guidelines, examples, and source code docs. It is accessible at <https://nasa.github.io/cumulus>.

**Navigation the Cumulus Docs**

- Get To Know Cumulus
  - Getting Started - here - If you are new to Cumulus we suggest that you begin with this section to help you understand and work in the environment.
  - General Cumulus Documentation - here - you're here
- Cumulus Reference Docs
  - Cumulus API Documentation - here
  - Cumulus Developer Documentation - here - READMEs throughout the main repository.
  - Data Cookbooks - here
- Auxiliary Guides
  - Integrator Guide - here
  - Operator Docs - here

<https://nasa.github.io/cumulus/docs/cumulus-docs-readme>





## Database



## Use Cases



## PO.DAAC Cloud Service Requirements

### • **SWOT Survey 2.0** (n=111)

- SWOT Science Team
- SWOT Early Adopters
- PO.DAAC User Working Group
- PO.DAA SOTO use cases
- SWOT Hydrology wishlist

### • **Application Journeys** (n=65)

- **Application data requirements and user capabilities**

- **User workflows** (use case traceability matrix)

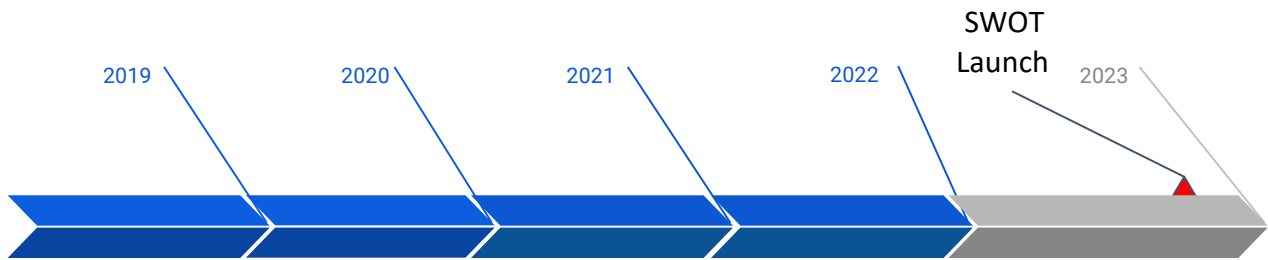
- Prioritized use cases based on % users impacted
- Use cases can be looked at by User Persona (e.g. oceans, hydrology, or coastal applications)
- Use cases complemented by user data preferences (e.g. data file format, projections, software & tools)

Functionality:  
Tools & Services  
on the Cloud

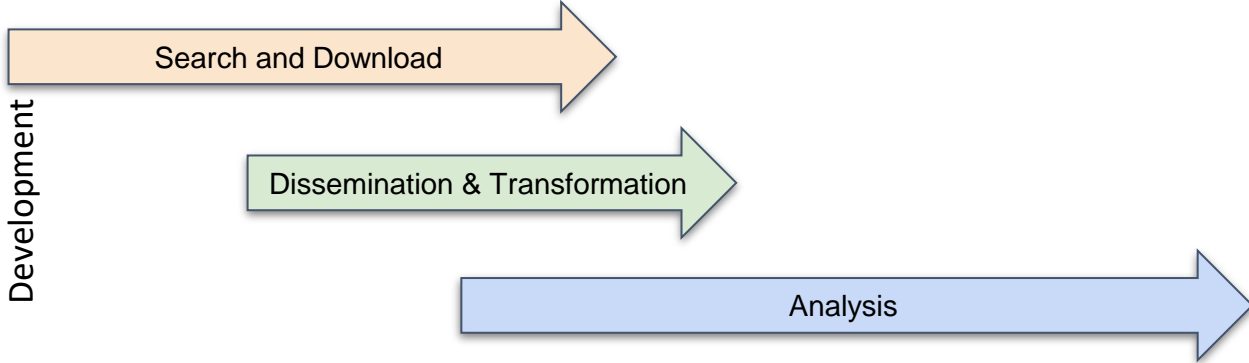
E. N. Stavros, C. M. Oaida, J. Hausman and M. M. Gierach, "A Quantitative Framework to Inform Cloud Data System Architecture and Services Requirements Based on User Needs and Expected Demand," in *IEEE Access*, vol. 8, pp. 138088-138101, 2020, doi: 10.1109/ACCESS.2020.3012054.



# Data Services Migration Timeline



End-User Capabilities Development



## DESIGN GOALS

**Users will get the same level of service**

**Data download will continue to be freely available to users**

**Leverage the power of co-located data** for processing large volumes of co-located datasets

**Enable new frontiers in science/applications**

**Search and Download**

- Feature based search
- Space/Time download

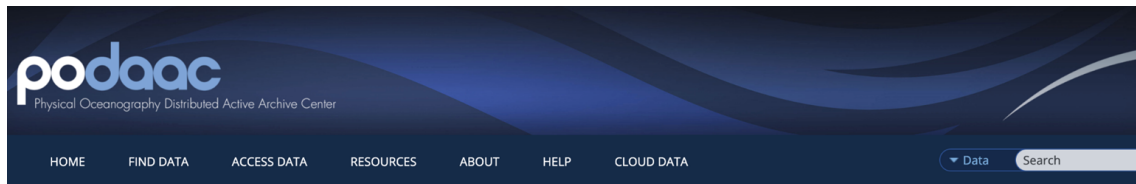
**Dissemination & Transformation**

- Subscriptions
- Subset
- Regrid , Reformat
- On-demand Raster Generation

**Analysis**

- Product space/time averaging
- Analysis-in-place (Cloud)
- Integration with other Datasets

# Data Migration



Home

## FINAL REMINDER - PO.DAAC Drive and Legacy Tools and Services RETIREMENT April 24, 2023 - Important Information for Users

Thursday, April 20, 2023

PO.DAAC has been preparing for the next generation of NASA Earth observing missions and the migration of the data archive, tools, and services to the Cloud. Part of this transition is determining which tools and services are best for our users in the new environment along with the challenges of delivering data faster and at higher volumes than ever before.

Last year, we informed our users that as a part of this transition, PO.DAAC Drive and certain legacy data access tools and services (specifically LAS, CWS, THREDDS) will be retiring.

**This is a FINAL REMINDER for all PO.DAAC users to complete the transition of their legacy data access scripts and methods for compatibility with cloud data access endpoints at their earliest convenience, as PO.DAAC Drive and the legacy data access tools and services (specifically LAS, CWS, THREDDS) will retire on April 24, 2023.**

To ensure a successful transition for all PO.DAAC users, the [PO.DAAC in the CLOUD Forum](#) can be used as the primary entry point to help address technical issues and concerns. Our User Services team ([podaac@podaac.jpl.nasa.gov](mailto:podaac@podaac.jpl.nasa.gov)) is also standing by to support.

For additional information on the PO.DAAC Data Migration to the Earthdata Cloud, please see the [PO.DAAC Cloud Data Page](#), or the list of resources below.

### Additional Cloud Resources

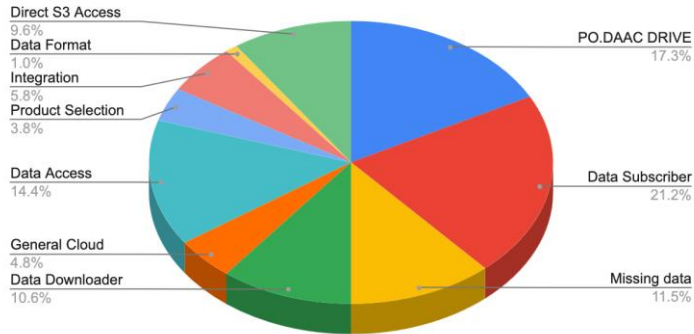
## User Migration:

## Training and Direct Outreach

- Workshops & Hackathons (Training) (12\*)
- Webinars (3\*)
- Present/participate at Science Team Meetings (20\*)
- Present/participate at Conferences (15\*)
- Tutorials/Notebooks (40\*)

\*estimate

# End User Migration



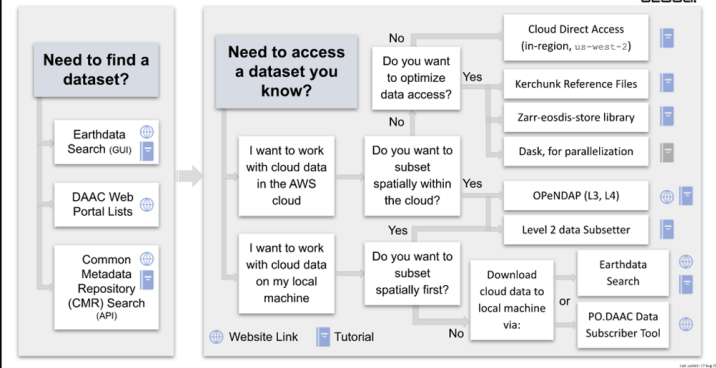
Q

- Welcome
- Cheatsheets & Guides
- How To's >
- Tutorials >
- In Development
- Workshops
- Webinars >
- Tech Guides >
- Contribute >
- Questions?

## Tools & Services Roadmap

Below is a practical guide for learning about and selecting helpful tools or services for a given use case, focusing on how to find and access NASA Earthdata Cloud-archived data from local compute environment (e.g. laptop) or from a cloud computing workspace, with accompanying example tutorials. Once you follow your desired pathway, click on the respective blue notebook icon to get to the example tutorial. Note: these pathways are not exhaustive, there are many ways to accomplish these common steps, but these are some of our recommendations.

### Working with NASA Earthdata Cloud data :: Tools & Services Roadmap



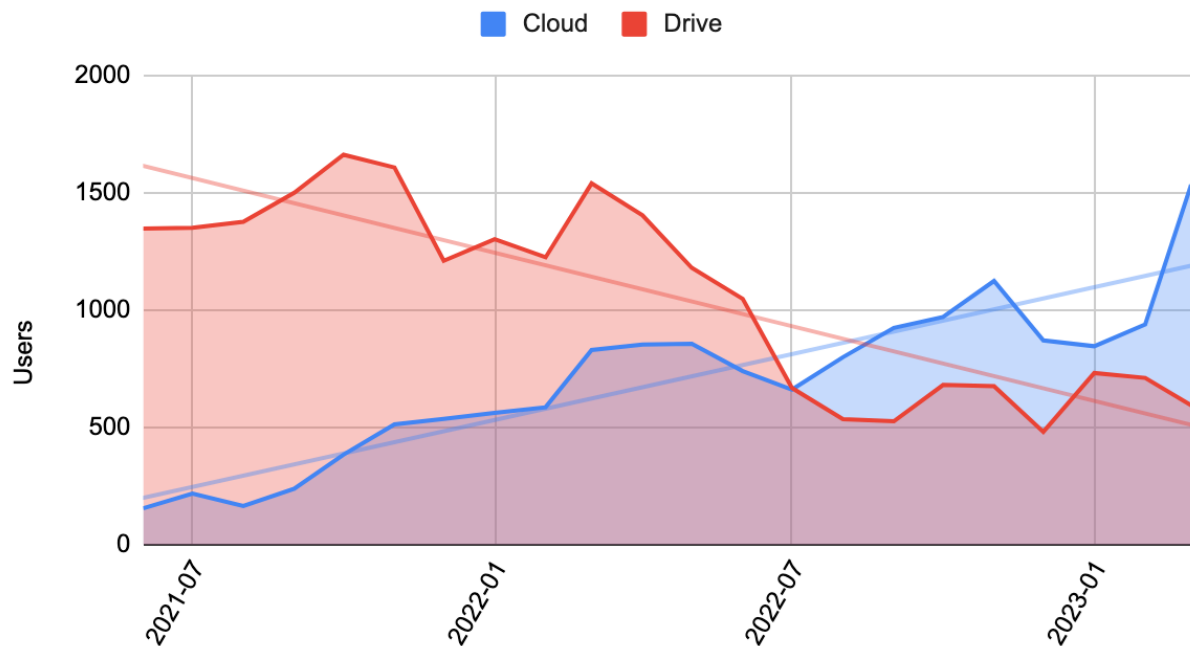
#### On this page

- Contents:
- What is the NASA Earthdata Cloud?
- Cloud Access Pathways
- Getting Started Roadmap
- Tools & Services Roadmap
- Cloud Terminology 101
- Workflow Cheatsheet
- Workflow Cheatsheet Terminology

- Edit this page
- Report an issue

# Measuring Success

## Number of Unique Users



- Since July of 2022, **PO.DAAC delivered data to more EarthData Cloud users than PO.DAAC Drive Users**
- Data Distribution and user adoption are trending in a positive direction

# Data backup

[Cumulus Documentation](#) [API Docs](#) [Distribution API Docs](#) [Developer Docs](#) [Data Cookbooks](#) [Operator Docs](#) Next ▾

- Getting Started ▾
  - Introduction
  - Getting Started
  - Glossary
  - Frequently Asked Questions
- About Cumulus ▾
  - Architecture
  - Interfaces
  - Cumulus Team
- Deployment ▾
  - How to Deploy Cumulus
  - Creating an S3 Bucket
  - Terraform Best Practices
  - Component-based Cumulus Deployment
- Databases ▾
  - PostgreSQL Database Deployment
  - RDS: Choosing and Configuring Your Database Type
- APIs ▾
  - Using the Thin Egress App (TEA) for Cumulus Distribution
  - Using the Cumulus Distribution

This is unreleased documentation for Cumulus Documentation **Next** version.

For up-to-date documentation, see the [latest version](#) (v15.0.0).

🏠 > Workflows > Workflow Tasks > LZARDS Backup

Version: Next

## LZARDS Backup

The LZARDS backup task takes an array of granules and initiates backup requests to the LZARDS API, which will be handled asynchronously by LZARDS.

### Deployment

The LZARDS backup task is not automatically deployed with Cumulus. To deploy the task through the Cumulus module, first you must specify a `lzards_launchpad_passphrase` in your terraform variables (e.g. `variables.tf`) like so:

```
variable "lzards_launchpad_passphrase" {  
  type = string  
  default = ""  
}
```

Then you can specify a value for your `lzards_launchpad_passphrase` in `terraform.tfvars` like so:

```
lzards_launchpad_passphrase = your-passphrase
```

Lastly, you need to make sure that the `lzards_launchpad_passphrase` is passed into the Cumulus module (in `main.tf`) like so:

- Deployment
- Task Inputs
- Input
- Configuration
- Task Outputs
- Output



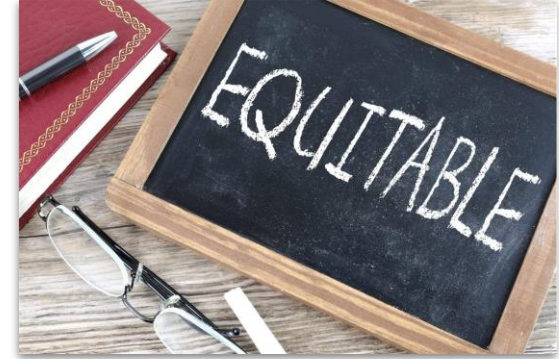
# We are not at the destination yet....



User Experience



Enable new Frontiers in Science



Open Data  
Open Source Science  
Equitable access to data