Jet Propulsion Laboratory
California Institute of Technology

Physical Oceanography Distributed Active Archive Center

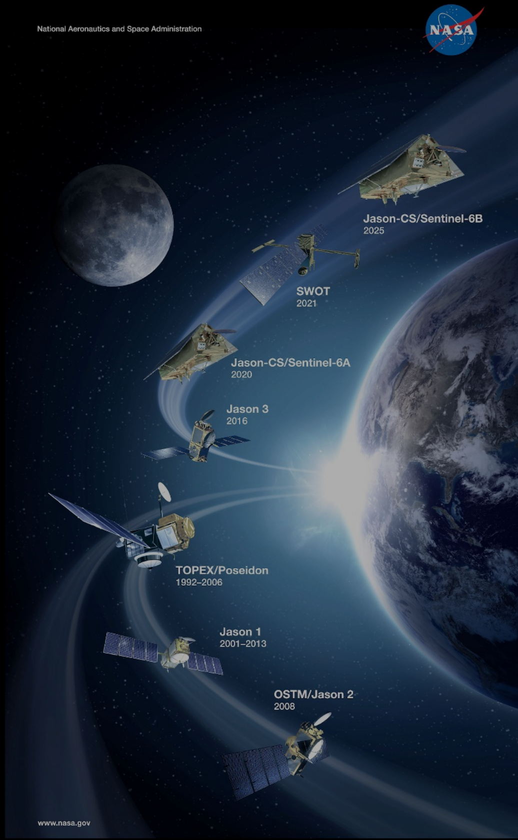
Earthdata in the cloud: PO.DAAC journey to the Cloud in support of SWOT

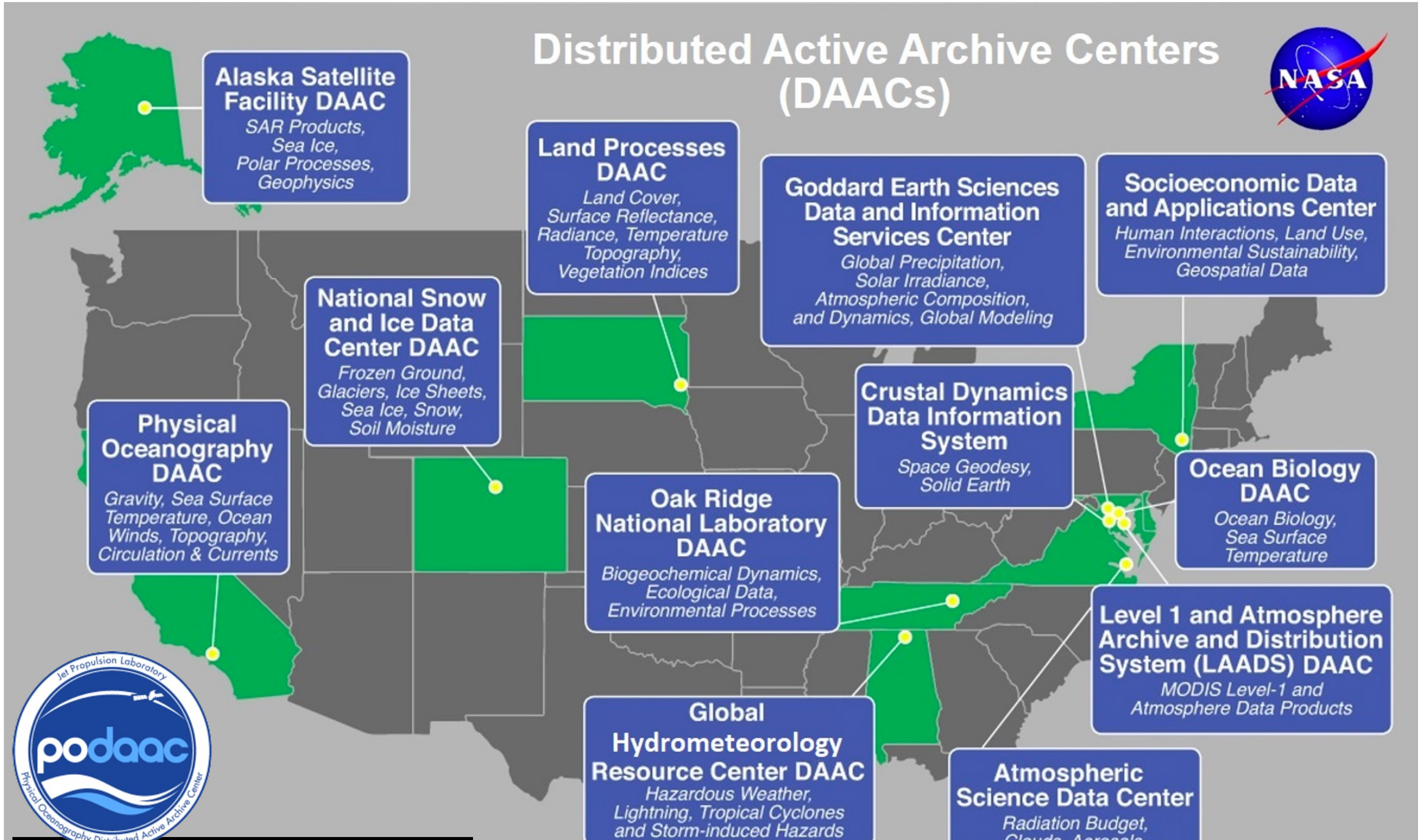
Suresh Vannan, PO.DAAC team

Jet Propulsion Laboratory, California Institute of Technology

February 27th, 2023

Ground System Architectures Workshop (GSAW)

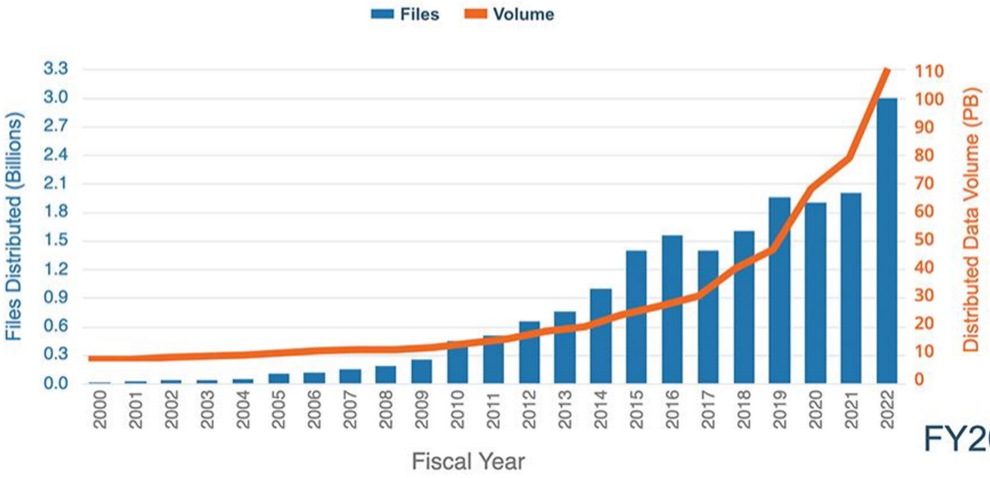




https://podaac.jpl.nasa.gov/

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

Total Data Volume and Data Files Distributed by Year



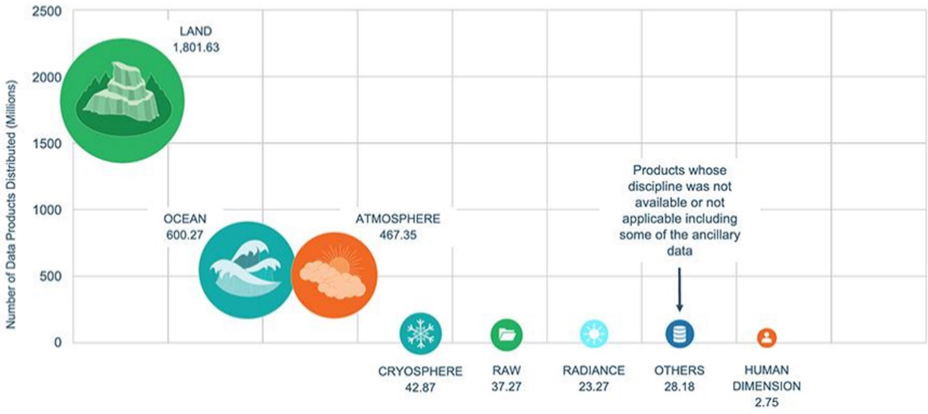
76.4 PB

Data Archive Volume (2022)

3 Billion Files

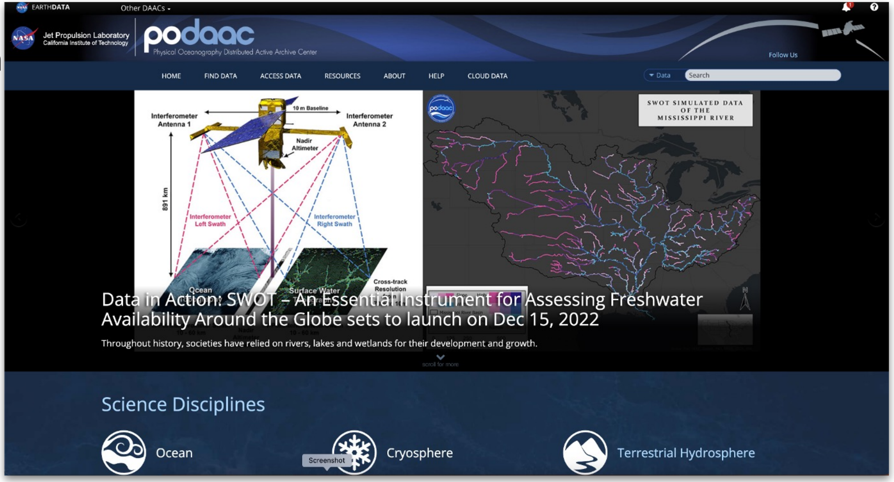
Distributed 2022

FY2022 Number of Data Products Distributed by Discipline



NASA PO.DAAC : Physical Oceanography Distributed Active Archive Center

The Physical Oceanography Distributed Active Archive Center (PO.DAAC) is a NASA Earth Observing System Data and Information System (EOSDIS) data center managed by the Earth Science Data and Information System (ESDIS) Project. The PO.DAAC is operated by Jet Propulsion Laboratory (JPL) in Pasadena, California



Mission:

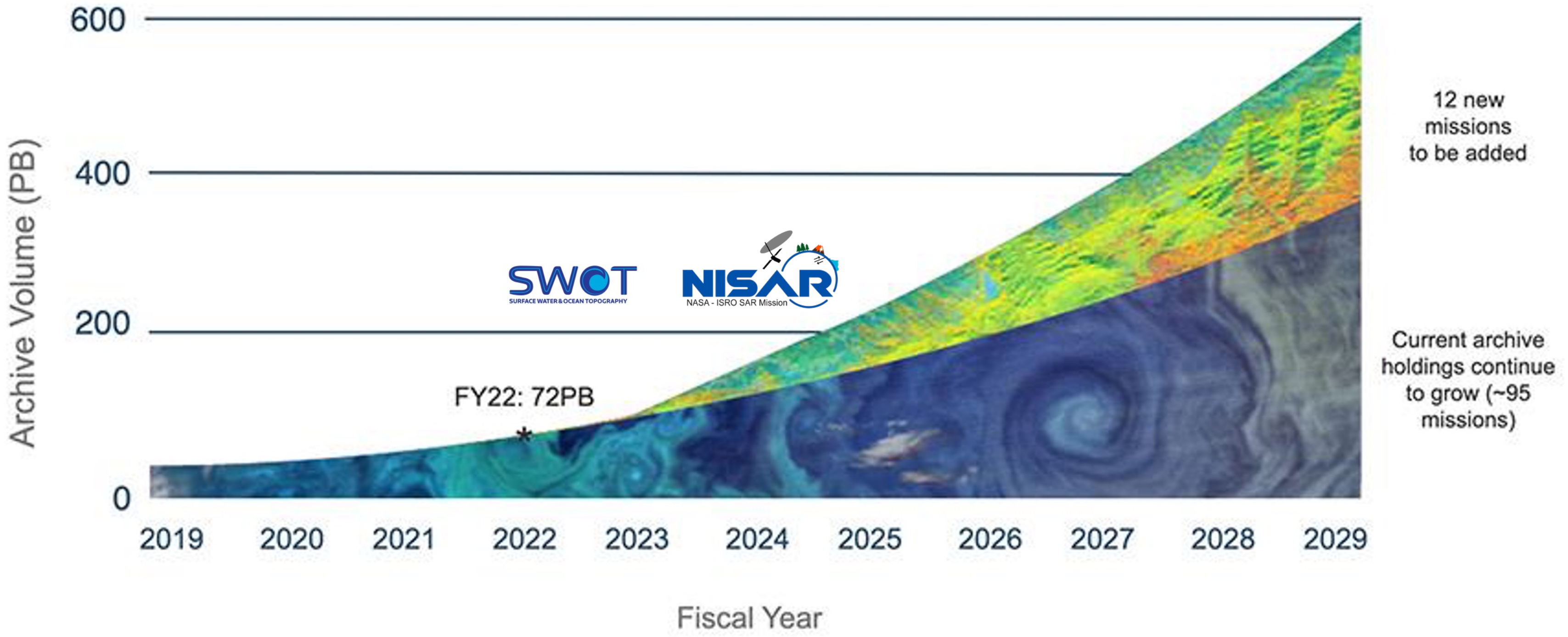
The mission of the PO.DAAC is to preserve NASA's ocean and climate data and make these universally accessible and meaningful.

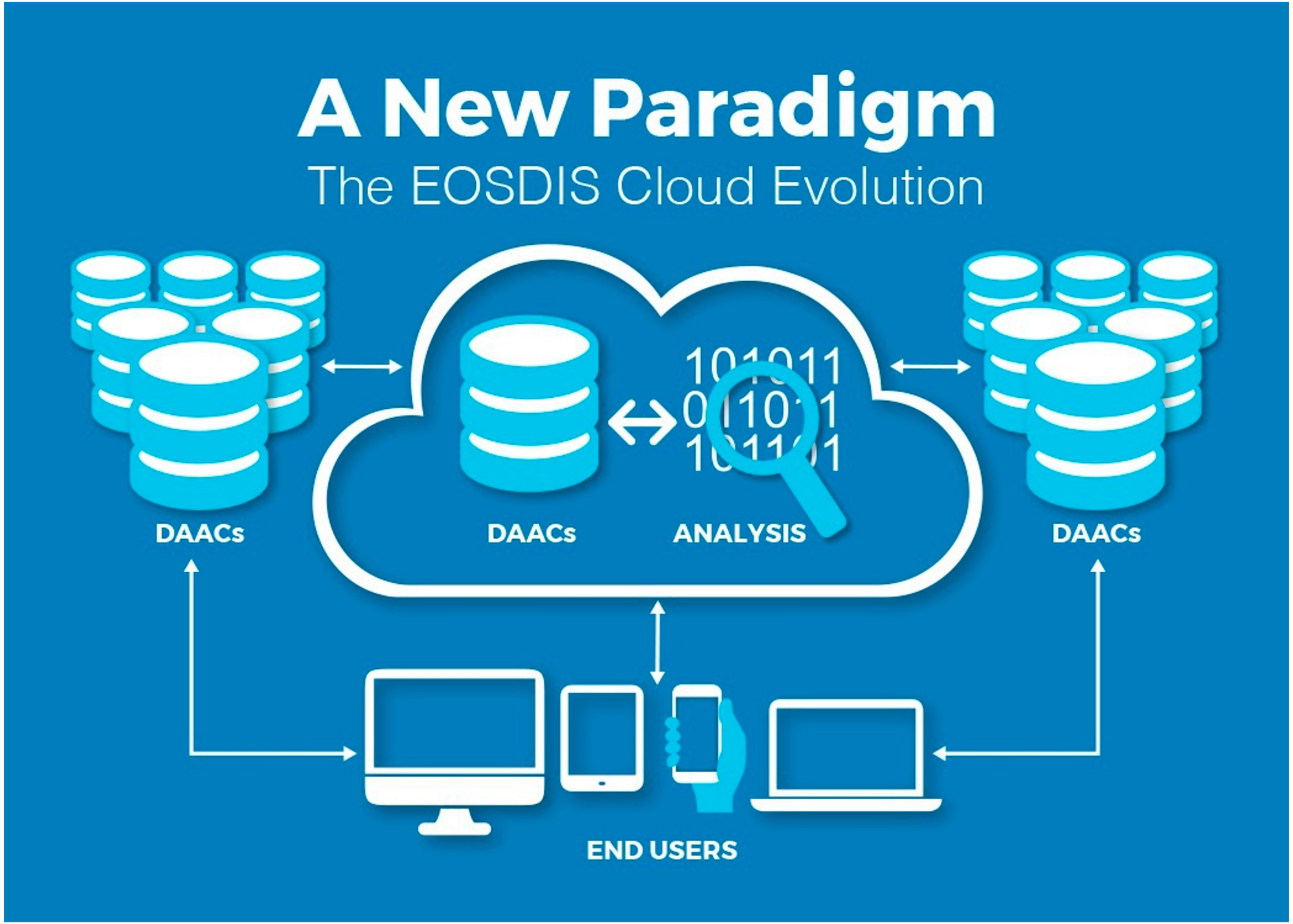
PO.DAAC in numbers:

- 40+ years of operations
- Free and open data
- 50,000+ unique users
- > 1 PB of data
- 650+ datasets

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

Earth Science Data Archive Growth Projection




A New Paradigm
The EOSDIS Cloud Evolution
DAACs
DAACs
ANALYSIS
DAACs
END USERS

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.

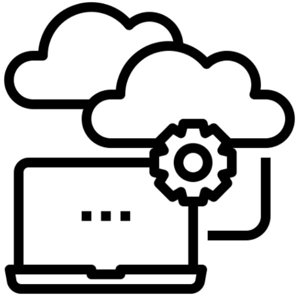
Challenges

Cost

- Storage
- Egress
- Development
- Computational
- Labor

Security

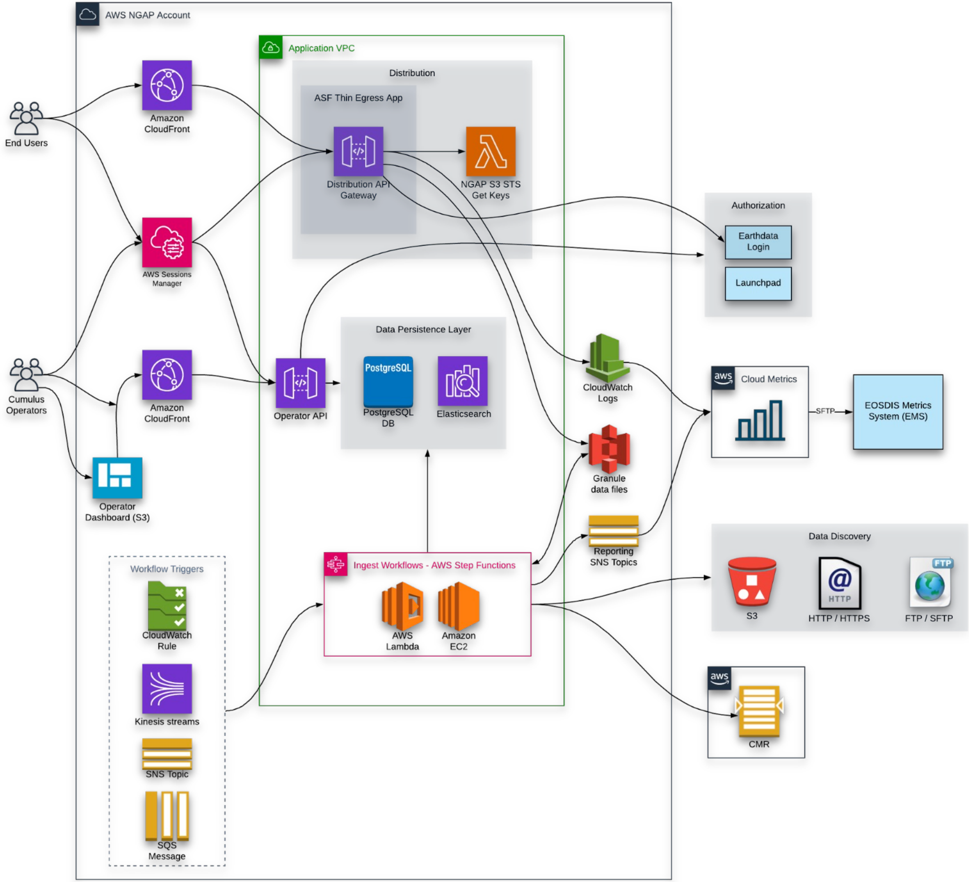
- Data protection
- Access control
- Cybersecurity

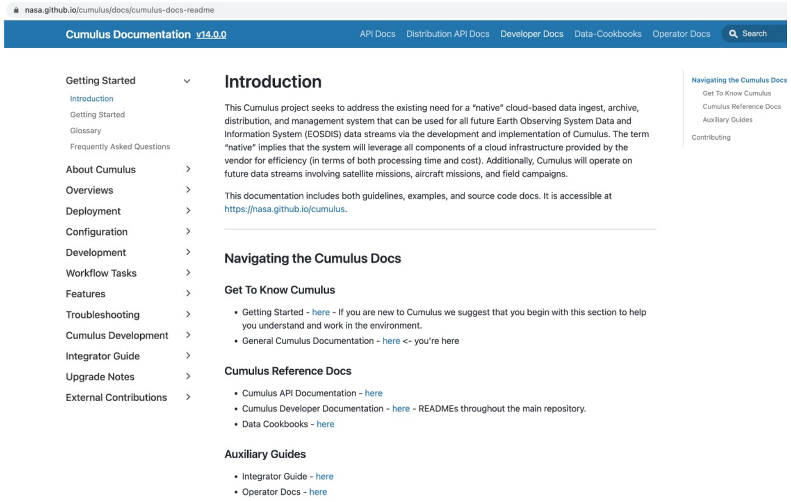


Migration

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

CUMULUS
EOSDIS NASA'S EARTH OBSERVING SYSTEM DATA AND INFORMATION SYSTEM




nasa.github.io/cumulus/docs/cumulus-docs-readme

Cumulus Documentation v14.0.0 | API Docs | Distribution API Docs | Developer Docs | Data-Cookbooks | Operator Docs | Search

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

Introduction

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.

Navigating 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

Navigating the Cumulus Docs
Get To Know Cumulus
Cumulus Reference Docs
Auxiliary Guides
Contributing

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

User Needs Assessment → Database → Use Cases → PO.DAAC Cloud Service Requirements

User Needs Assessment

- **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)

Database

- Application **data requirements** and **user capabilities**
- **User workflows** (use case traceability matrix)

Use Cases

- 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)

PO.DAAC Cloud Service Requirements

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 Migration Timeline

Jet Propulsion Laboratory
California Institute of Technology

podaac
Physical Oceanography Distributed Active Archive Center

Follow Us

HOME | FIND DATA | ACCESS DATA | RESOURCES | ABOUT | HELP | CLOUD DATA

Data | Search

Home

CLOUD DATA - MIGRATION

- CLOUD DATASETS
- ABOUT
- ACCESS DATA
- FAQ
- RESOURCES
- **MIGRATION**
  - Timeline
  - What to Expect
  - Tutorials
  - Migration FAQs

During 2021 PO.DAAC is in the process of migrating its data archive to the Earthdata Cloud, hosted in Amazon Web Services (AWS). During this transition, some data will continue to be available from the on premise archive, while some data will also be available from and within the Earthdata Cloud.

Timeline

- Phase 1 (Cloud Accessable Date: April 15, 2021; Cloud Only Access Date: January 31, 2022)
- Phase 2 (Cloud Accessable Date: October 1, 2021; Cloud Only Access Date: May 2, 2022)
- Phase 3 (Cloud Accessable Date: December 23, 2021; Cloud Only Access Date: November 14, 2022)
- Phase 4 (Cloud Accessable Date: April, 2022; Cloud Only Access Date: November 14, 2022)
- Phase 5 (Cloud Accessable Date: July, 2022; Cloud Only Access Date: March 31, 2023)

Data Migration

podaac
Physical Oceanography Distributed Active Archive Center

Follow Us

HOME | FIND DATA | ACCESS DATA | RESOURCES | ABOUT | HELP | CLOUD DATA

Data | Search

Home » Dataset Discovery

RSS SMAP Level 3 Sea Surface Salinity Standard Mapped Image 8-Day Running Mean V4.0 Validated Dataset

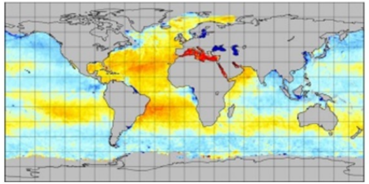
(SMAP_RSS_L3_SSS_SMI_8DAY-RUNNINGMEAN_V4)

6 Publications Cited this Dataset
Citation metrics available for years (2014-2021)

Information | Coverage | Data Access | Documentation | Citation | Version History

Version	4.0
Processing Level	3
Start/Stop Date	2015-Mar-27 to Present
Short Name	SMAP_RSS_L3_SSS_SMI_8DAY-RUNNINGMEAN_V4
Description	The version 4.0 SMAP-SSS level 3, 8-Day running mean gridded product is based on the fourth release of the validated standard mapped sea surface salinity (SSS) data from the NASA Soil Moisture Active Passive (SMAP) observatory, produced operationally by Remote Sensing Systems (RSS). Enhancements with this release include: use of an improved 0.125 degree land correction table with land emission based on SMAP TB; replacement of the previous NCEP sea-ice mask with one

SHARE THIS PAGE



CLOUD ENABLED

Status: ACTIVE

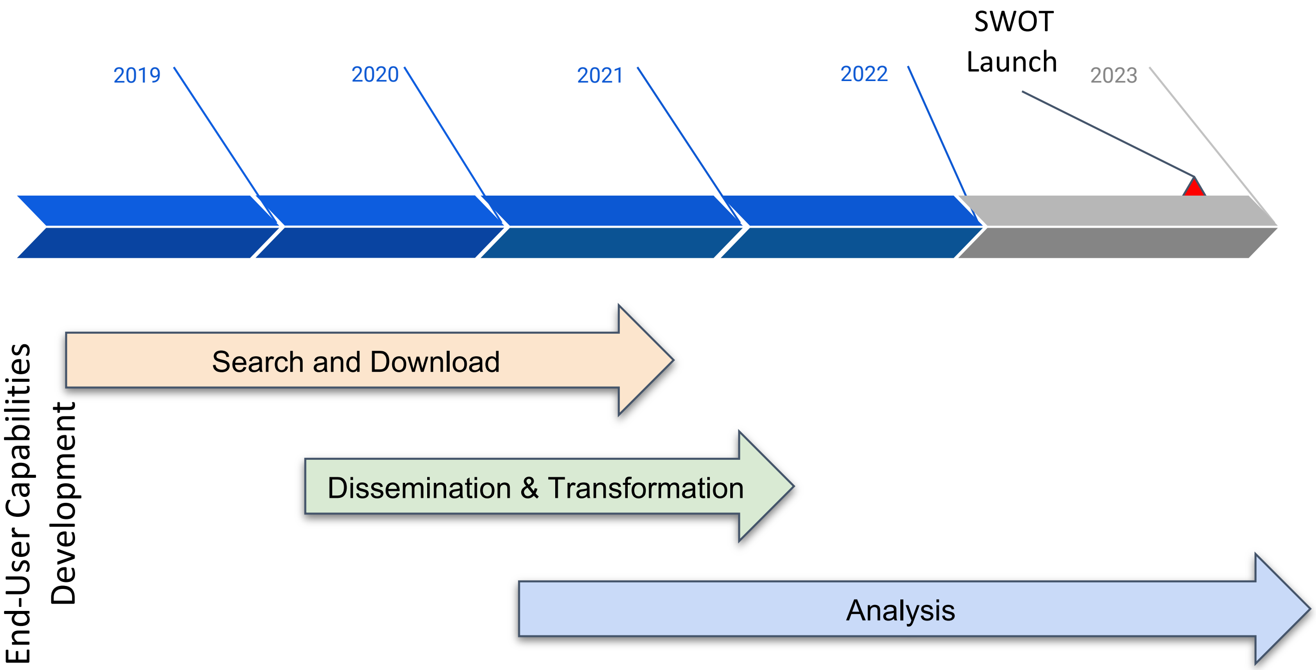
Short Name:
SMAP_RSS_L3_SSS_SMI_8DAY-RUNNINGMEAN_V4

Collection Concept ID:
C1940468263-POCLOUD

Spatial Coverage:
N: 90° **S:** -90°
E: 180° **W:** -180°

Access:

Data Services Migration Timeline



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

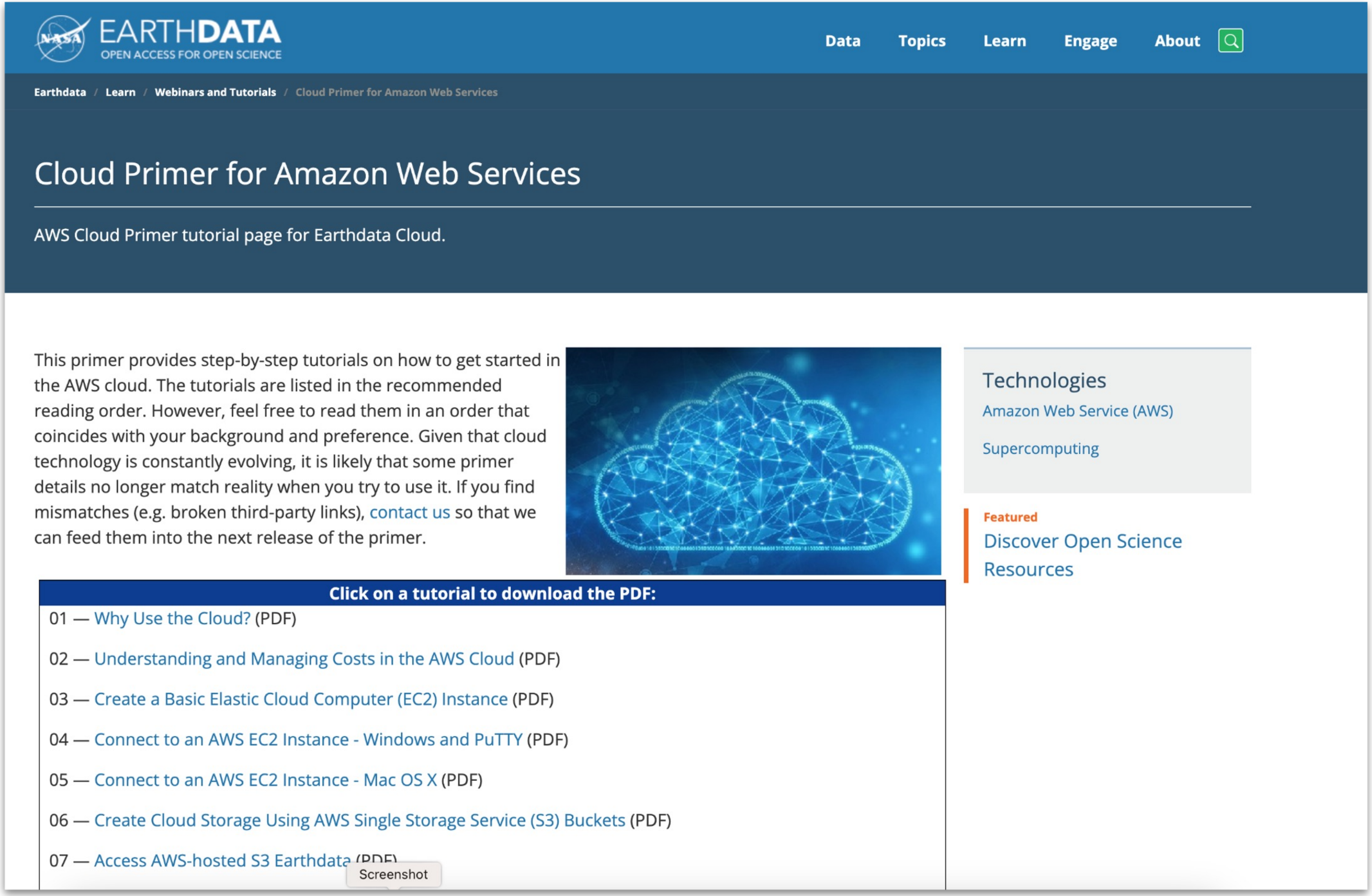
Dissemination & Transformation

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

Analysis

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

End User Migration


EARTHDATA
OPEN ACCESS FOR OPEN SCIENCE

Data Topics Learn Engage About

Earthdata / Learn / Webinars and Tutorials / Cloud Primer for Amazon Web Services

Cloud Primer for Amazon Web Services

AWS Cloud Primer tutorial page for Earthdata Cloud.

This primer provides step-by-step tutorials on how to get started in the AWS cloud. The tutorials are listed in the recommended reading order. However, feel free to read them in an order that coincides with your background and preference. Given that cloud technology is constantly evolving, it is likely that some primer details no longer match reality when you try to use it. If you find mismatches (e.g. broken third-party links), contact us so that we can feed them into the next release of the primer.

Click on a tutorial to download the PDF:

01 — Why Use the Cloud? (PDF)

02 — Understanding and Managing Costs in the AWS Cloud (PDF)

03 — Create a Basic Elastic Cloud Computer (EC2) Instance (PDF)

04 — Connect to an AWS EC2 Instance - Windows and PuTTY (PDF)

05 — Connect to an AWS EC2 Instance - Mac OS X (PDF)

06 — Create Cloud Storage Using AWS Single Storage Service (S3) Buckets (PDF)

07 — Access AWS-hosted S3 Earthdata (PDF)

Technologies
Amazon Web Service (AWS)
Supercomputing

Featured
Discover Open Science Resources

Screenshot

https://www.earthdata.nasa.gov/learn/webinars-and-tutorials/cloud-primer-amazon-web-services

Challenges

Cost

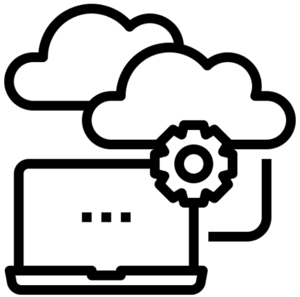
- ❑ Leveraged Space Act Agreement
- ❑ Negotiated contract with AWS for lowered cost
- ❑ Invested in building cost models
  - ❑ Past trends
  - ❑ Expected usage
- ❑ Built buffer in cost
- ❑ Support from organization

Challenges

Security

- Earthdata uses NGAP
  - NGAP - Next Generation Application Platform
  - NGAP is the NASA Compliant General Application Platform. It provides a cloud-based Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS) for EOSDIS applications.
- Implemented security training across teams
- Incorporated security as an integral part of the DevOPS process (synck tools)

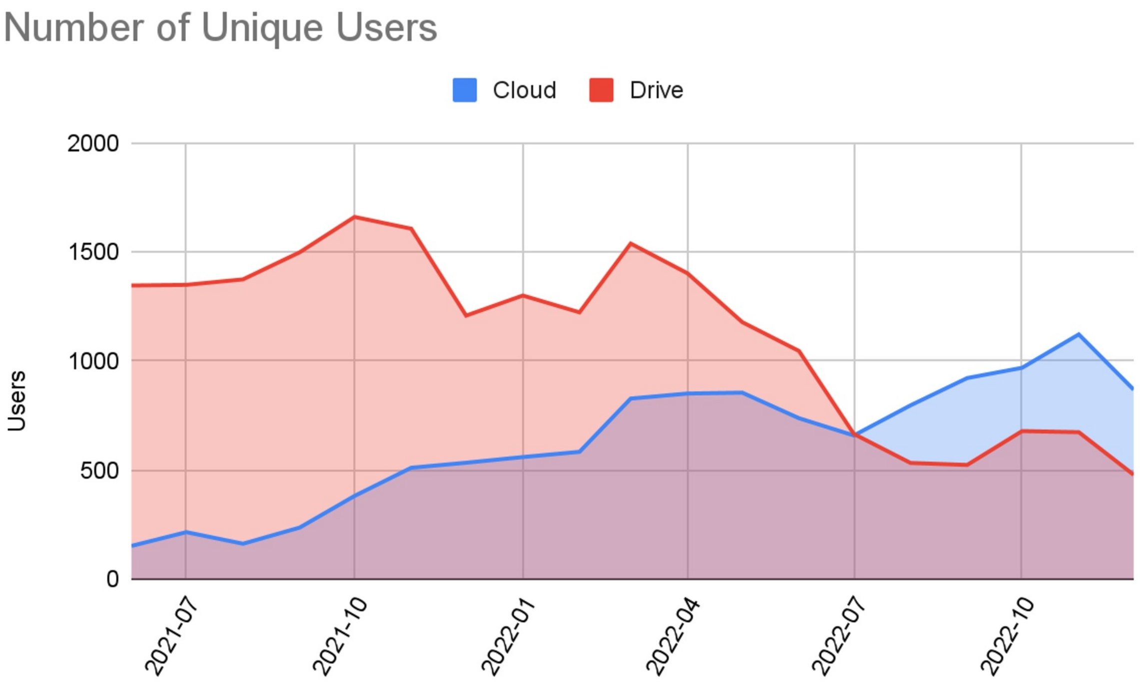
Challenges



Migration

- ❑ Utilized SAFe agile development model
- ❑ Encountered delays and unexpected challenges
  - ❑ Build margins
- ❑ Long tail issues take longer effort and resources
- ❑ Communication was very critical

Measuring Success

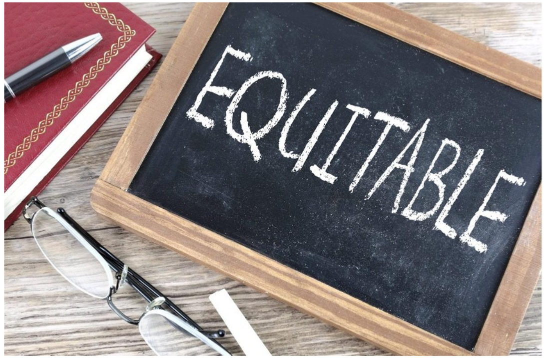


- Since July of 2022, **PO.DAAC delivered data to more EarthData Cloud users than PO.DAAC Drive Users**
- Data Distribution and user adoption are trending in a positive direction

We are not at the destination yet….

User Experience

Enable new Frontiers in Science



Open Data
Open Source Science
Equitable access to data