EOSDIS
NASA'S EARTH OBSERVING SYSTEM
DATA AND INFORMATION SYSTEM

# NASA Earthdata Cloud

WGISS-52 - October 20, 2021

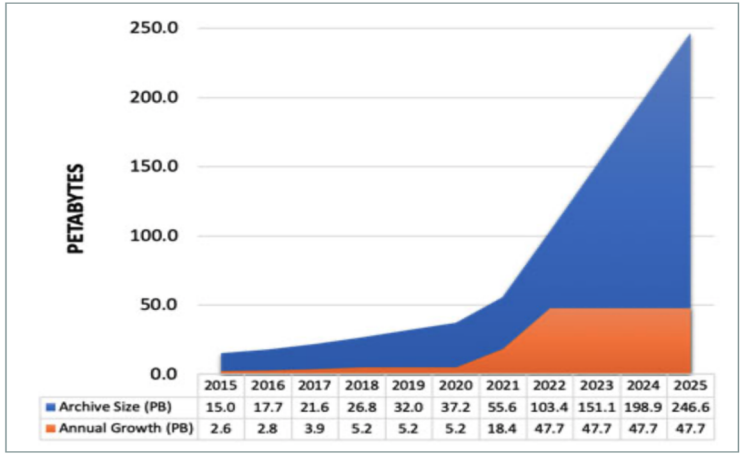
Dan Pilone
dan@element84.com

# Executive Summary

**Earthdata Cloud** (EDC) enables Earth science data and computations to move into the cloud, creating opportunities for innovation around new services, such as sequencing data to support machine learning and application of large scale analytics.

Earthdata Cloud will:

- Improve the efficiency of data systems operations
- Increase user autonomy
- Maximize flexibility
- Offer shared services and controls




| | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | 2025 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Archive Size (PB) | 15.0 | 17.7 | 21.6 | 26.8 | 32.0 | 37.2 | 55.6 | 103.4 | 151.1 | 198.9 | 246.6 |
| Annual Growth (PB) | 2.6 | 2.8 | 3.9 | 5.2 | 5.2 | 5.2 | 18.4 | 47.7 | 47.7 | 47.7 | 47.7 |

Between 2017 and 2025, the volume of data in the EOSDIS archive (blue area) is expected to grow dramatically, accompanied by an order of magnitude increase in the rate of data ingest (orange area). NASA EOSDIS graphic.

## Key End User Benefits

- **Power:** Any user can access big processing power "next to" Big Data.
- **Performance:** Data can be offered in a form enabling high-performance analysis.
- **Freedom from Data Transfers:** Users need not move Big Data.
- **Freedom from Data Management:** Users need not store and manage Big Data.
- **Data Co-location:** Users can easily work with multiple EOSDIS datasets together.
- **Choice:** Users can still download data if they prefer.

## Key Metrics

- ~800 products and 36M granules available in EDC
- > 70% of privileged users in EDC are developers
- ~2 TBs of egress monthly, >500M Lambda executions / month, ~2300 S3 buckets for data
- 77% of data stored in S3 Infrequent Access, 13% in Standard S3, 10% in Glacier

## Lessons Learned

1. Know why you're moving to the cloud & communicate
2. Embrace cloud risks
3. Cost controls need to be a tier-1 capability
4. Think about data transition like any other
5. Cloud development is hard

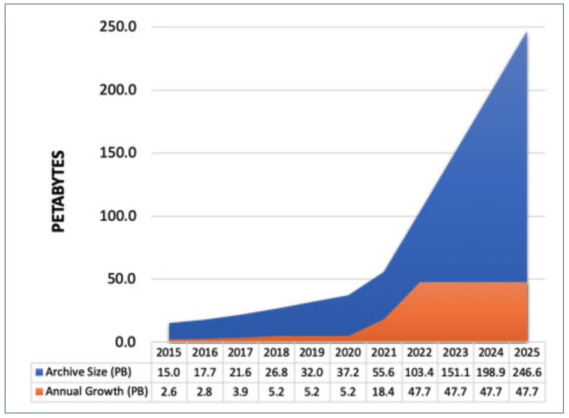
# Earthdata Cloud Overview

**Earthdata Cloud** enables Earth science data and computations to move into the cloud, creating opportunities for innovation around new services, such as sequencing data to support machine learning and application of large scale analytics.

Earthdata Cloud will:

- Improve the efficiency of data systems operations
- Increase user autonomy
- Maximize flexibility
- Offer shared services and controls

Researchers and commercial users of NASA Earth Science data will have increased opportunity to **access and process petabytes of EOS data quickly**, allowing new types of research and analysis. Data that was previously geographically dispersed will now be accessible via the cloud, saving time and resources.




| | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | 2025 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Archive Size (PB) | 15.0 | 17.7 | 21.6 | 26.8 | 32.0 | 37.2 | 55.6 | 103.4 | 151.1 | 198.9 | 246.6 |
| Annual Growth (PB) | 2.6 | 2.8 | 3.9 | 5.2 | 5.2 | 5.2 | 18.4 | 47.7 | 47.7 | 47.7 | 47.7 |

Between 2017 and 2025, the volume of data in the EOSDIS archive (blue area) is expected to grow dramatically, accompanied by an order of magnitude increase in the rate of data ingest (orange area). NASA EOSDIS graphic.

EOSDIS

Moving data and services to the cloud brings numerous benefits for both data users and EOSDIS, including:

- **Easy access:** Data users are able to access data directly in the cloud, removing the need to download volumes of data for use.
- **Rapid deployment:** With an established EOSDIS cloud platform, data users can bring their algorithms and processing software to the cloud and work directly with the data in the cloud, simplifying procurement and hardware support while expediting science discovery.
- **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 redundancy:** The use of a common infrastructure with cloud native services will reduce redundant tools and services, enable sharing, and enforce the use of community standards as well as uniform policies and processes.
- **Cost effectiveness:** EOSDIS and NASA pay only for the storage and services actually used. Along with scalability benefits, this allows the amount of storage or services to be continually adjusted to ensure that data and services are effectively provided at the lowest possible cost to NASA and EOSDIS.



More information: https://earthdata.nasa.gov/eosdis/cloud-evolution

# Lesson 1: Know why you're moving to the cloud

- What are the goals of the effort?
- What are the metrics (KPIs / Key Results) of progress?
- Where do the users want to be? How do you bridge that?
- Prioritize and communicate what you are and are not doing
- For example:
  - Provide scientific data stewardship for all data collections
  - Provide a unified & simplified environment for community to discover, obtain, and analyze data
  - Evolve & adapt to new sources of data and data systems technologies
  - Expand the community and engage with users
  - Partner with other organizations, agencies, etc. to share data and make it easier to integrate for science

# Elevator Pitch

By hosting EOS data in the cloud, EOSDIS is able to realize several end user benefits:

- **Power:** Any user can access big processing power "next to" Big Data.
- **Performance:** Data can be offered in a form enabling high-performance analysis.
- **Freedom from Data Transfers:** Users need not move Big Data.
- **Freedom from Data Management:** Users need not store and manage Big Data.
- **Data Co-location:** Users can easily work with multiple EOSDIS datasets together.
- **Choice:** Users can still download data if they prefer.

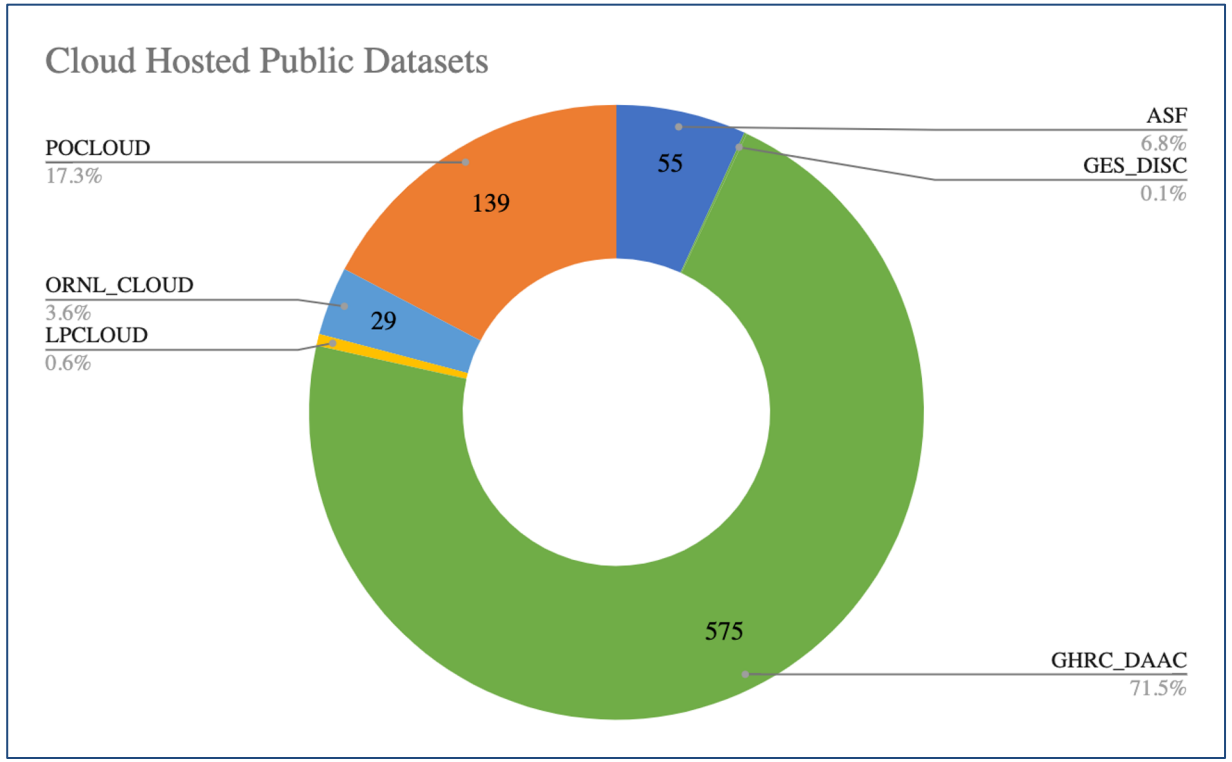
# Lesson 2: Embrace cloud risks

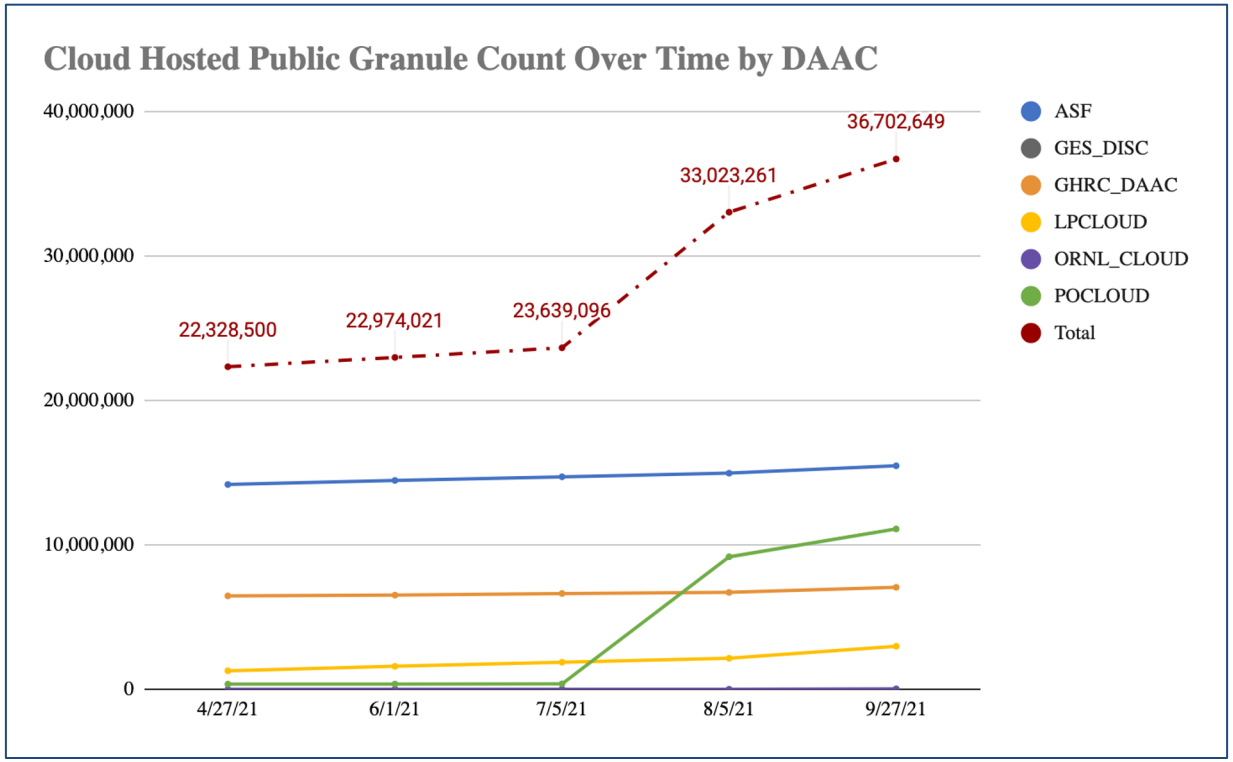
- Solicit, capture, triage, and talk about risks.  A LOT.
- Use risks to help drive progress and innovation
- Some cloud risks will look familiar – some should not.
- For example:
  - Vendor “Data Lock-in”
  - Vendor “Service and Software lock-in”
  - Open-source development and community security concerns
  - Limited experience with operational cloud-native systems
  - System X may not be able to scale with cloud archive needs
  - Authentication and authorization policy discrepencies
  - Data regionality and User regionality
  - Security and cost controls

# Lesson 3: Cost controls need to be a tier-1 capability

- Costs are variable in the cloud due to elastic nature.  This is GOOD.
- This is often very different than on-premises development.  This is BAD.
- Decide what indicators you can rely on and which actions you can take under various circumstances and automated them.
- Reliable cost capabilities and automated overrun mitigation techniques provide the confidence necessary to allow distributed innovation in the cloud.
- **Table stakes:** Leverage metrics and automated monitoring to watch for cost overruns and provide automated responses.
- **Pro-tier:** Leverage analytics / ML for real-time continuous monitoring, anomaly detection, and prediction.

# Cloud Migration Metrics




804 datasets and 36,702,649 granules are hosted in Earthdata Cloud and available for public discovery and distribution

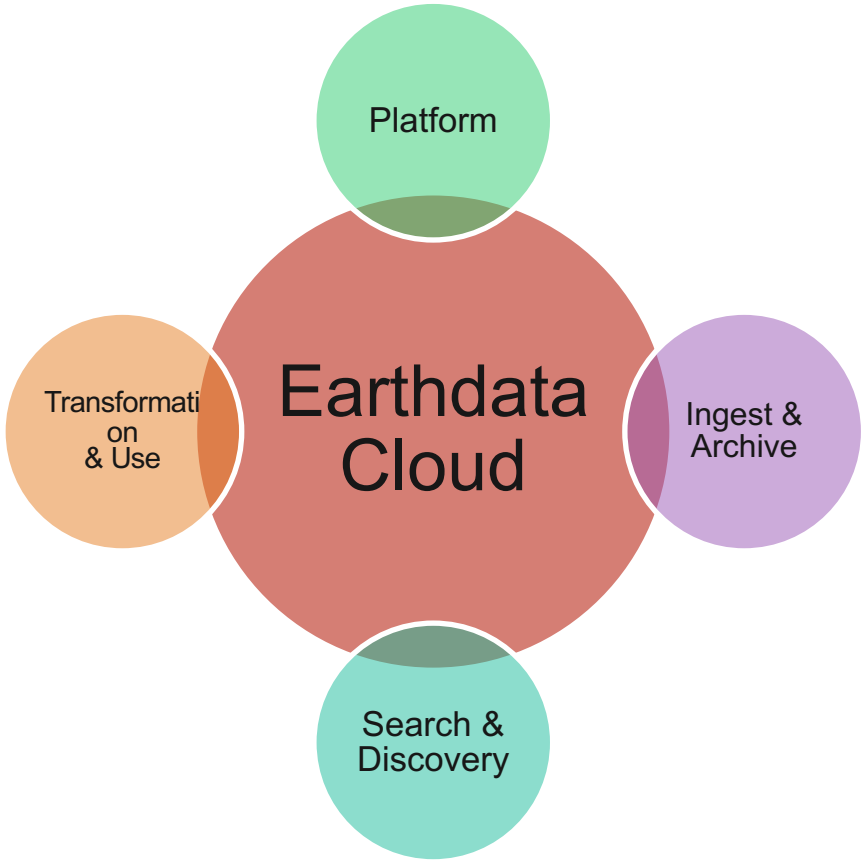
EOSDIS

Data gathered from CMR inventory on 9/27/21. Includes only publicly discoverable data in Earthdata Cloud.

# Lesson 4: Think about data transition like any other

- Traditional software / system transitions involve discussions around lift-and-shift vs. rebuild and cloud native. Think about data transitions the same way.
- Ask yourself:
  - What does it mean to have transformation services in the cloud vs. allowing users to compute next to the data?
  - What does “direct data access” really look like in a pseudo-controlled environment with real cost constraints?
  - How do you meet users where they are vs. bringing them to where you want them?
  - What patterns that are in use on-prem may be inefficient / cost prohibitive in the cloud? What are the implications of those?
  - What opportunities are there to be forward looking with how the data will be accessed in the cloud? Obvious: ARD, cloud-native file formats. Less obvious: data chunking, parallelism, data regionality, other agency and public data lakes, metadata support for byte offsets, etc.

# Earthdata Cloud Team Structure

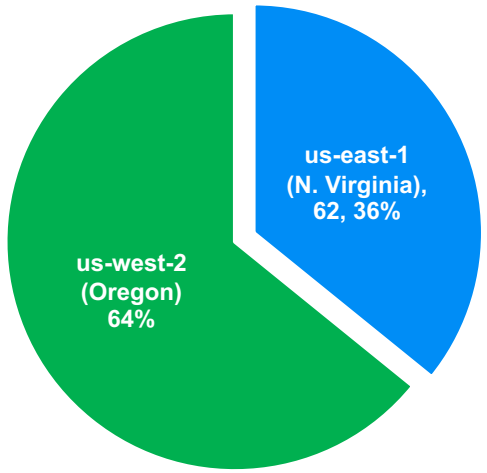



- **Platform** – fundamental cloud access and compliance capabilities
- **Ingest & Archive** – ingest, processing, metadata, archival pipeline support
- **Search & Discovery** – User and Machine interfaces for data discovery and access
- **Transformation & Use** – In-place access, common services, data interoperability support

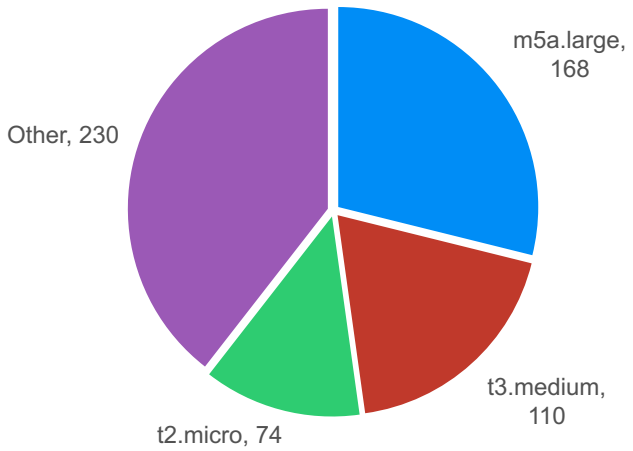
# Lesson 5: Cloud development is difficult

- While there are tremendous advantages to DevOps / DevSecOps / ML Ops, in addition to traditional software development complexity you now require developers to care about:
  - Cost Modeling, deployment automation, data storage systems and scaling, security, operations, networking..
- There is a real cost to recruiting, training, and retaining cloud engineers and may require cultural changes in the organization
- You *could* rework your entire application during a cloud transition, but it's critical to understand your goals and prioritize accordingly, e.g.:
  - Is your priority to retire legacy hardware? Lift-and-shift may be a great option
  - Is your priority to leverage cost savings and / or exploit cloud on-demand scaling? Lift-and-shift could be a failure from this perspective.
- Distributed and shared development is still hard and requires coordination across teams. Distributed and shared data is no different. The cloud does not make either problem go away.

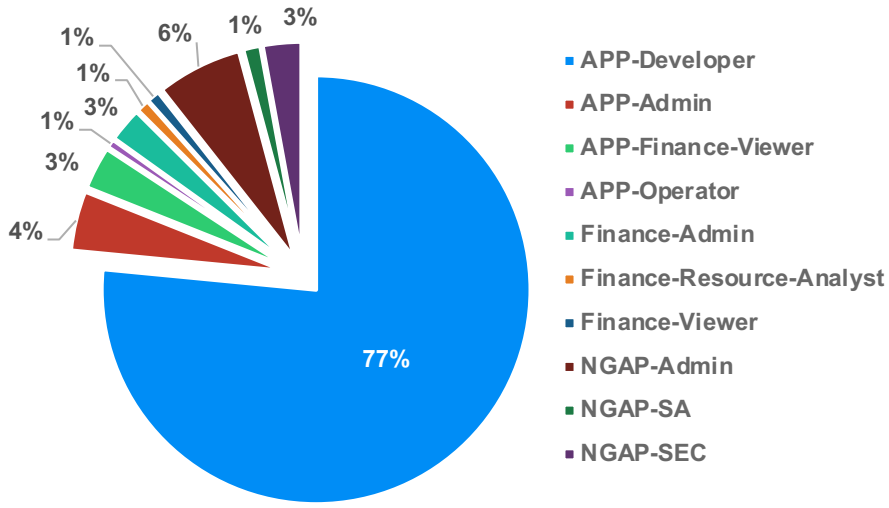
## AWS Regions by Accounts



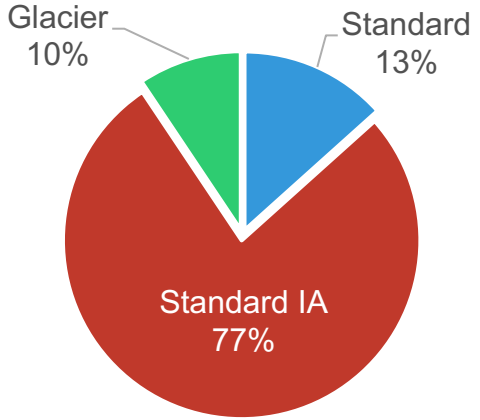

## EC2 Instance Type Distribution




## EDC Users by Role




## Object Storage Tiers




## Additional Stats

- **Monthly Egress:** ~2TBs
- **Unique S3 Buckets:** 2328
- **Cloudfront Endpoints:** 197
- **Lambda Executions:** 522M / month, 195 / second

EOSDIS

EOSDIS
NASA'S EARTH OBSERVING SYSTEM
DATA AND INFORMATION SYSTEM

# Looking ahead…

# Near-term activities and goals

- Continued migration of EOSDIS products and services
- Pilot-projects and movement towards Analysis Ready Data and more support for Analysis in Place
- Enhanced direct data access and simplification of user authentication / authorization
- Continued development and offering of “mix-and-match” services enabling easier data use for traditional users (e.g. subsetters, format conversions, etc.)
- Enhanced metrics analysis capabilities
- Continued education and outreach efforts
- Additional “self-services” offerings for end users in controlled security boundaries
- Additional interoperability enhancements including continued expansion of STAC use

EOSDIS

NASA'S EARTH OBSERVING SYSTEM
DATA AND INFORMATION SYSTEM

# Thank you!