BIG DATA, CLOUD, AND EARTH SCIENCE

AGU Fall Meeting
December 9, 2019
Who am I?

Katie Baynes, katie.baynes@nasa.gov

System Architect,
NASA Earth Science Data and Information Systems

Currently working on migrating NASA Earth Science Data
distribution to commercial cloud
NASA’s Earth Observing System Data and Information System

Our Work in Context

https://earthdata.nasa.gov
Applications
Capture and Clean
*Subset, reformat, reproject
Transform*
Distribute
Archive
Process
Research
Commercial
Applications
Education
satellite, airplane, compass icons by Nook Fulloption, database, transformation, decision, process, data mining, customer community by Becris via thenounproject.com (CC 3.0)
Data are Produced and Managed by Science Discipline Experts

- **ASF DAAC**: SAR Products, Sea Ice, Polar Processes
- **NSIDC DAAC**: Cryosphere, Polar Processes
- **PO.DAAC**: Ocean Circulation, Air-Sea Interactions
- **NCAR, U. of Co.**: MOPITT
- **JPL**: MLS, TES, SNPP Sounder
- **LPDAAC**: Land Processes and Features
- **U. of Wisc.**: SNPP Atmosphere
- **CDDIS**: Crustal Dynamics, Solid Earth
- **SEDAC**: Human Interactions in Global Change
- **GSFC**: SNPP, MODIS, OMI, OBPG
- **ORNL**: Biogeochemical Dynamics
- **ASDC**: Radiation Budget, Clouds, Aerosols, Tropo Composition
- **LAADS/MODAPS**: Atmosphere
- **OB.DAAC**: Ocean Biology and Biogeochemistry
Big Variety at EOSDIS

NASA Earthdata Datasets in 2019

- 12 NASA centers of domain expertise
- 8,900 distinct data collections online
- 420 million cataloged files
EOSDIS Date Holdings Evolution

Data Look-ahead

Now
~ 23 TBs/day generated

Soon
~126 TBs/day generated
NASA’s Earth Observing System Data and Information System

Our Goals and Motivations

https://earthdata.nasa.gov
→ Provide scientific **data stewardship** for all data collections and insure data integrity

→ Provide a **unified and simplified environment** for a diverse and distributed community of Earth Science and Applications users

→ **Evolve, grow and adapt** to new sources of data and new data systems technologies

→ Expand the user community and engage with users to **enhance and improve user access** to data and other resources.

→ **Partner** with other organizations, US agencies, and Nations to share data and make it easier to integrate for science
NASA’s
Earth Observing System Data and Information System

Our Vision

https://earthdata.nasa.gov
Benefits
- Optimized for archive, search and distribution
- Expert user support
- Easily add new data products and producers
- Predictable

Challenges
- Uneven levels of service and performance
- Significant time to coordinate interfaces
- Limited on-demand product generation and end-user processing capabilities
- Duplication of storage
- Duplication of services and software
We are on the cusp of opportunity. Can we do better? We are targeting:

- Better support for interdisciplinary Earth science researchers
- Reduced burden of data management/preparation for end-users
- More insightful, interactive data for research and commercial development
- More seamless interoperability with other institutional, international, and commercial providers
- Reducing overall monetary footprint and increasing efficiency across the system
Towards a Streamlined Cloud-Based Architecture

**Benefits**
- Collocated, pay-as-you-go processing for anyone
- All data available to DAACs and users
- Expert user support
- Streamlined product addition

**Challenges**
- Development coordination
- Cost Management
- Shifting Labor Needs
- Security/Export Compliance
- Vendor Lock In

**Cloud-Native Ingest/Archive/Distribution System**

**DAAC’s (12) Thematic Stewardship**

**SIPS (17)**

**Varied Data Sources**
EOSDIS DAACs all operate and maintain their own archive and distribution systems. This will also be how we operate in the future. However, as we work towards a cloud-based system, Cumulus is providing DAACs with a customizable
https://github.com/nasa/cumulus
Each DAAC maintains their own instances of Cumulus and other services.

Account owners have autonomy within their own AWS account. (more on this later)
Example Step Function Cloud Workflows

https://aws.amazon.com/step-functions/
Unifying Services in the Cloud: Harmony

Historically, EOSDIS DAACs have all provided their own tooling with diverse interaction patterns and APIs. Harmony is our ongoing effort to revisit these siloed capabilities in a more harmonized manner.
The Harmony Elevator Pitch

Service Execution Framework(s)
Enterprise Integration: Login, Metrics, Egress, Metadata Catalog
Common Interface: Common API, Earthdata Search UI
Data staging, Transformation code execution

End-Users

DAAC- or Core-Team-Supplied

DAAC-Supplied

DAAC-Unique Transformations
e.g., SWOT water feature averaging

Common Transformations
e.g., Subsetting L3 NetCDF

DAAC- or Core-Team-Supplied
NASA’s
Earth Observing System Data and Information System

Our Current Status

https://earthdata.nasa.gov
Current EOSDIS Systems Operating in the AWS Cloud

Common Metadata Repository

https://cmr.earthdata.nasa.gov

Earthdata Search

https://search.earthdata.nasa.gov

API-driven, standards-compliant, sub-second search of:

- 8,900 collections
- 420 million files
Current EOSDIS and Partner Data in the AWS Cloud

Global Hydrology Research Center

https://ghrc.nsstc.nasa.gov/home/

Alaska Satellite Facility

ESA’s Sentinel 1 Archive Mirror

https://search.asf.alaska.edu/
Planned Cloud Dataset Timeline in 2020

Jan

GES DISC Onboarding ➔ MERRA 2

ORNL Onboarding ➔ Daymet

LP Onboarding ➔ ASTER GDEM

→ Harmonized Landsat Sentinel

PODAAC Sentinel 6 Preparation

Feb

GES DISC Onboarding ➔ GPM IMERG L3

ORNL Onboarding ➔ DELTA-X

LP Onboarding ➔ MODIS 6.1

→ HLS Operational Rollout

Mar

Parallel Operations Prep

Jun

Parallel Operations Prep

Jul

Parallel Operations Prep

Aug

NSIDC Prototyping
NASA’s Earth Observing System Data and Information System

Our Challenges and Strategies

https://earthdata.nasa.gov
Challenge: Vendor Lock-In
“What if you have to move the data?”

Right now, AWS is the only NASA-approved commercial cloud vendor. As more options become available we will investigate them.
“<XYZ> AWS-specific product!”

Most of the tools we are using are not a unique problem that Amazon alone has solved. There are usually (many) free and open source, alternatives.

As we continue to evolve cloud functionality, we continue to examine trade-offs between out-of-the-box and vendor agnostic.
What about ECS, Lambda, SQS, etc

Again, these are not unique problems. Every major competitor in the cloud space has alternatives, or open source alternatives exist.

Serverless: Qinling, Google Cloud Functions
Queues: Zaqar, RabbitMQ
etc, etc
Knowledge Transfer Risk

“We are training everyone in AWS”

This is a real problem. Effectively leveraging the AWS console is its own skillset. People may become unwilling to be retrained if we have to migrate. But we have faced this problem before.
Challenge: End-User Adoption

If the data is in the cloud, we would like to encourage users to work with that data in place.
All DAACs are starting to look at this problem individually...

Data Recipes

How to Create and Unwrap an Interferogram with GMT5SAR Script in the Cloud — Windows

In this document you will find:

- Background
- Required Pre-Steps
- Prerequisites
- Steps
- Sample Images
- Appendix 1: Steps the Script Completes
- Appendix 2: Output files
- Appendix 3: Sample script run
We are Developing a “Cloud Primer” to aid user transition across all DAACs

POC: justin.rice@nasa.gov
Challenge: Cost Control

Photo by Fabian Blank on Unsplash
Budget and Cost Monitoring

Total Spending
Current Funding Sources

63%
8/1/2018 12/31/2020

$2,533.73 Spent

of $4,000.00 Planned

Monthly Spending
Individual Totals

https://www.cloudtamer.io/
Cost Conscious Development

76% S3
10% EC2
2.6% CloudTrail
1.7% VPC
1.7% ElasticSearch Service
1.3% Kinesis
1.2% DynamoDB
1.0% Config
0.9% CloudWatch
0.8% Lambda
2.2% All Other Services
EDC year-to-date cost percentages
Budget and Cost Monitoring

Total Spending

Current Funding Sources

63%

8/1/2018
12/31/2020

$2,533.73
Spent

of $4,000.00
Planned

Monthly Spending

Individual Totals

$137.71
$159.32
$197.61
$514.40
$157.33
$172.86
$275.96
$0.00

February 2019
March 2019
April 2019
May 2019
June 2019
July 2019
August 2019
September 2019

https://www.cloudtamer.io/
Challenge: Security

IT Security keeps us safe; keeping up-to-date and protected requires constant vigilance.
Security is every team member’s responsibility. We are working towards automating as much as possible to remain productive and safe as we migrate to the cloud.

We regularly test and implement new tools to protect our code bases, dependency trees, and operational systems and vulnerabilities.

- Code Security Working Group PMB Sub-Team
  - 2018-09-25 Meeting notes
  - 2018-10-03 Meeting notes
  - 2019-03-27 Meeting notes
  - 2019-05-29 Meeting notes
  - 2019-10-16 Meeting notes
  - Information: Current processes for deploying code
  - Proposal: Defense in depth for programmatic secrets
Limiting Exposure while Communicating with Users

### Lambda

<table>
<thead>
<tr>
<th>Full Access w/ Caveats</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Lambda functions must execute from within the NGAP provisioned private subnets of the Application VPC.</td>
</tr>
</tbody>
</table>

#### Lambda Networking Requirements

<table>
<thead>
<tr>
<th>VPC</th>
<th>Application VPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subnet</td>
<td>Private application-[xxx]</td>
</tr>
<tr>
<td>Security Group</td>
<td>Any</td>
</tr>
</tbody>
</table>

#### IAM Helper Policy

| NGAPShLambdaInVpcBasePolicy |

Contains all necessary permissions for a generic Lambda to execute from within the Application VPC. This policy should be attached, along with any other required policies, to your Lambda IAM execution role.

---

### IAM Identity and Access Management

#### Limited Permissions

Application owners may only manage IAM permissions for application components running in their AWS account(s), not for users. All IAM roles created by app owners are subject to NGAP-managed permissions boundaries.

#### IAM Requirements

<table>
<thead>
<tr>
<th>IAM Roles</th>
<th>NGAPShRoleBoundary / NGAPShNonProdRoleBoundary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MUST be assigned as Permissions Boundary to create a custom IAM Role.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IAM Policies</th>
<th>Full Access</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>IAM Users</th>
<th>Handled via CloudTamer. No access through AWS IAM Web Console.</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAM Access Keys</td>
<td>Handled via CloudTamer. No access through AWS IAM Web Console.</td>
</tr>
</tbody>
</table>
Other Opportunities for Big Data in the Cloud

Much of our work over the last 24-36 months has been about coping with an oncoming data onslaught. We still have much ground to tread to embrace our new found wealth.

https://earthdata.nasa.gov
Questions?
Backups
Project Board

sets the vision for

Project Leadership Team

provides the “what” for

Technical Leadership Team

guides the “how” for

Developers (Committers)

make changes to

Code Base

serves as chair for

serves as head of

Project Manager

sets the vision for

proposes changes to

Contributors

serves as head of

EOSDIS Funded
Tackling Egress Monitoring

1. DNS points to NGAP-controlled CloudFront distribution.

2a. If the egress cutoff limit for this month has already been reached, the request is rejected (403).

2b. Otherwise, the request is passed to the origin (a tenant app/distribution app, e.g. Cumulus), through the platform’s monitoring stack (Internet Services). If required, the application’s distribution application sends unauthenticated users and those with expired sessions to Earthdata Login for Auth. A signed S3 URL is returned.

3a. If the origin returns a redirect to S3, CloudFront then picks between the following download mechanisms:

   a. If the application user is in the same region as the S3 bucket, pass through the S3 redirect unchanged; the user downloads directly from S3.

   b. If the application user is not in the same region but throttling is turned off, the user is redirected to a CloudFront URL to download the data (unthrottled) from S3 w/ CloudFront cost savings in place.

   c. If the application user is not in the same region but throttling is turned on, the download from S3 will be proxied through the bandwidth limiter.

3b. Otherwise, the response will be returned to the application user unmodified.