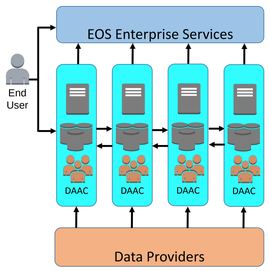
IN51B-0575
Fall AGU Meeting
December 10-14, 2018
Washington, D.C.

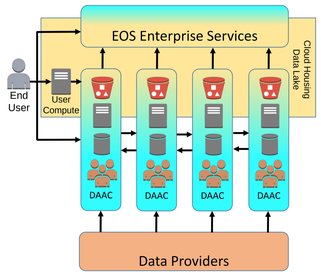
A Concept of Operations for Earth Science Data Archive and Distribution in the Cloud

John F Moses[1], Kathleen Baynes[1], Jeanne Behnke[1]
[1] NASA Goddard Space Flight Center, Greenbelt, MD, United States.

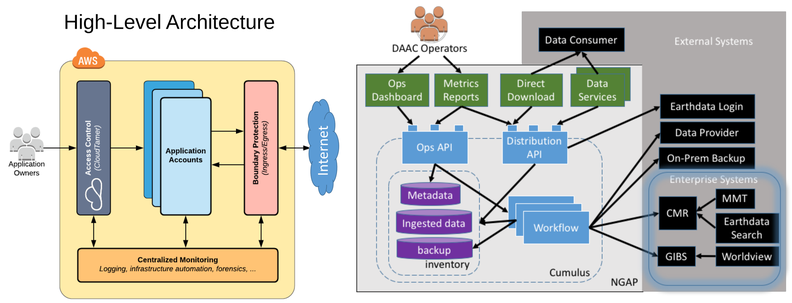
EOSDIS System of Systems Architecture circa 2018



EOSDIS System of Systems Architecture circa 2021



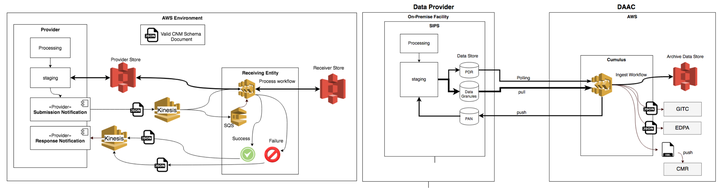
DAAC Data Collection in-the-cloud Target Architecture

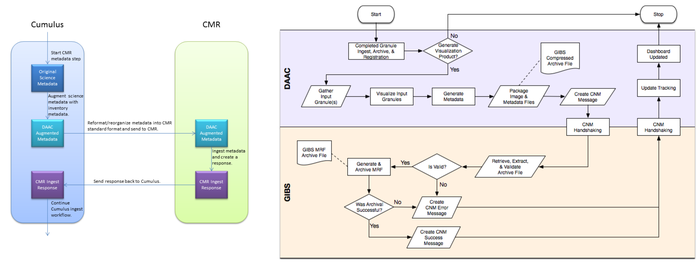


Major Functional Concepts and Scenarios

Ingest from Data Provider:

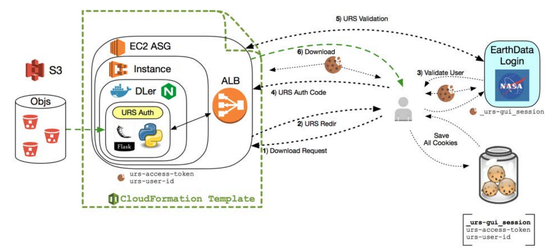
- Routine automated data transfers of standard products.
- Support multiple protocols both legacy and cloud-native.
- Include integrity checks, archive, browse imagery creation, metadata publication, backup.
- Continuous monitoring, troubleshooting workflows via Cumulus Dashboard.





Distribution:

- Download to user's on-premise facility.
  - The user is authenticated and authorize via Earthdata login.
  - The user requests and downloads data products.
  - Network throttling or circuit breakers may implemented to control egress costs.
  - Transfer rates are managed similar to DAAC on-premise bandwidth to the internet.
- Direct access to DAAC AWS archive via external AWS account within same AWS region.
  - The user is authenticated and authorize via Earthdata login
  - Temporary access is granted with no egress costs in the cloud-to-cloud transfer.
- Utilization is monitor & controlled by data collection or by registered user.



Concept of Operations Purpose and Objectives

- Describe how the DAACs will operate utilizing commercial compute services circa 2021-2026: Audience is developers, DAACs, EOSDIS, research community
- Identify the basic roles of ESDIS and DAACs personnel in operating a system-in-the-cloud
- Describe routine DAAC system operations; major automated flows and check & recovery points.
- Use scenarios to help distinguish what we expect the system will do and what operators will do.
- Drive out gaps in current requirements and development plans

Driving Assumptions

- Architect so systems of EOSDIS can be developed and evolve in parallel
  - Ongoing sustaining engineering and operations with coordinated parallel development efforts
  - Maintaining standard DAAC system and ESDIS service interfaces as much as possible
- Utilize NASA general application platform and services to manage cloud access
  - Amazon currently designated as NASA's authorized commercial vendor
- Transitions to cloud (i.e., 'onboarding') phased by science mission data collection

Science Data Systems Operations Tenets

- Fully automated ingest, archive & distribution with 24x7 system availability; 8hr weekday support staffing
- Efficient addition/sunset of missions, campaigns and products through automation
- More data services cross product domains
- Integrated EOSDIS-wide responses for help with products, services & tools

Roles and Responsibilities

- Personnel that will operate or control system resources and data
- Activities envisioned to require authorization to access
- Access may be restricted to specific system environments or data to limit exposure, reduce risk, and protect system or data integrity

Activity	DAAC Operator	DAAC Data Owner	DAAC Developer	DAAC App Owner	DAAC Mgmt	NGAP Admin	ESDIS Mgmt
Allocates AWS resources to the Organizational Unit (DAAC) and monitors utilization	-	-	-	C	C	R	A
Allocates, monitors AWS resources within DAAC to Projects: PROD, UAT, SIT and Sandbox (development) environments	C	C	-	R	A	I	I
Monitors data operations; Troubleshoots problems, restarts workflows, applications, services	R	C	C	C	A	-	-
Creates and tests new product workflows and data services. Verifies readiness for operations & data release. Maintains or corrects workflow config.	C	R	-	A	-	-	-
Develops and tests configurations and software to add new data providers, new products, new data services	-	C	R	A	-	-	-
Deploys new data providers, workflows, applications and service releases to operations	-	C	-	R	A	I	I
Integrates and tests new NGAP releases and new Cumulus framework releases	C	-	R	A	-	-	I
Deploys new releases of NGAP and Cumulus framework for DAAC operations	I	-	-	R	A	I	I

R = Responsible (shareable), A = Accountable, C = Consulted, I = Informed

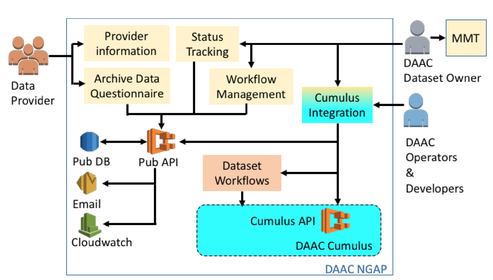
EOSDIS Component Systems and External Interfaces

- **Data Provider**: may be internal to EOSDIS such as SIPS or DAAC, or external such as a NASA Principle Investigator; may be from on-premise or in-cloud facility; DAAC on-premise facility may serve as proxy science data source.
- **Data Consumer**: may be internal to EOSDIS such as SIPS or DAAC, or external such as public user system; may be on-premise or in-cloud. The interface is agreed to in advance and can represent a one-time or an ongoing active data transfer.
- **Common Metadata Repository (CMR):** catalogs all DAAC science data and services; metadata records are registered, modified, and accessed via tools such as the MMT or via standard APIs.
- **Metadata Management Tool** (MMT): allows metadata authors (data owners) to create, update, publish, view, delete and manage their metadata records in CMR.
- **Earthdata Search Client**: Web application allowing users to search, discover, visualize, and access NASA Earth science data products using CMR and GIBS.
- **Earthdata Login**: single sign-on user registration and user profile management system for public users to get Earth science data form any of the DAACs.
- **Global Imagery Browse Services**: provides standard image services from DAAC science data.
- **Worldview**: public Web tool for interactively browsing global imagery from GIBS.
- **Data Preservation Archive**: Secure copy of unique and irreplaceable science data and information at a separate location.
- **NASA General Application Platform (NGAP)**: provides a cloud-based platform for DAAC applications. Managing accounts, monitoring, logging, security, storage, ingress/egress & networks.
- **Cumulus**: a framework for data ingest and archive management configurable to perform acquisition, ingest, validation, preprocessing, metadata harvesting & creation, publication to CMR, data distribution and metrics reporting.

Major Functional Concepts and Scenarios

Data Management:

- Setup and retirement of ingest, archive and distribution workflows & services for data collections.
- Metadata curation, product guide development, Website landing page development.
- Setup test environments, build/test/checkout metadata and data service workflows.
  - Cumulus Dashboard used to add/delete new dataset collections or individual granules.
  - MMT or discipline-specific metadata tool used to add collection metadata to the CMR.
  - Worldview used to test/check product imagery generation and delivery to GIBS.
  - Test/check dataset and documentation backup and restore configuration.
- Control data access based on guidance from the Principle Investigator or science team.
- Monitoring, troubleshooting workflow interrupts and checking quality of data services
  - auditing product granule availability, checking for data gaps, removing bad/duplicate data
  - augmenting collection and granule metadata as necessary



Clipart: https://aws.amazon.com/architecture/icons/

Maintenance:

- DAAC Application Owners run their own instance of Cumulus on their own NGAP account.
  - Deployments to operations accounts occur without downtime.
- User Acceptance Test environment is in separate isolated NGAP account.
  - Used by DAAC operators and selected Data Providers and Consumers.
  - Acceptance test new releases on realistic workflows.
- Development teams will have NGAP sandbox accounts to work on new code.
- DAAC Application Owners ensure integration of the latest releases of NGAP and Cumulus.