



EOSDIS

NASA'S EARTH OBSERVING SYSTEM
DATA AND INFORMATION SYSTEM

FOSS4G NA 2018: How NASA is Building a Petabyte Scale Geospatial Archive in the Cloud

Dan Pilone – NASA EED2 / Element 84, Inc.

Patrick Quinn - NASA EED2 / Element 84, Inc.

Alireza Jazayeri – Development Seed

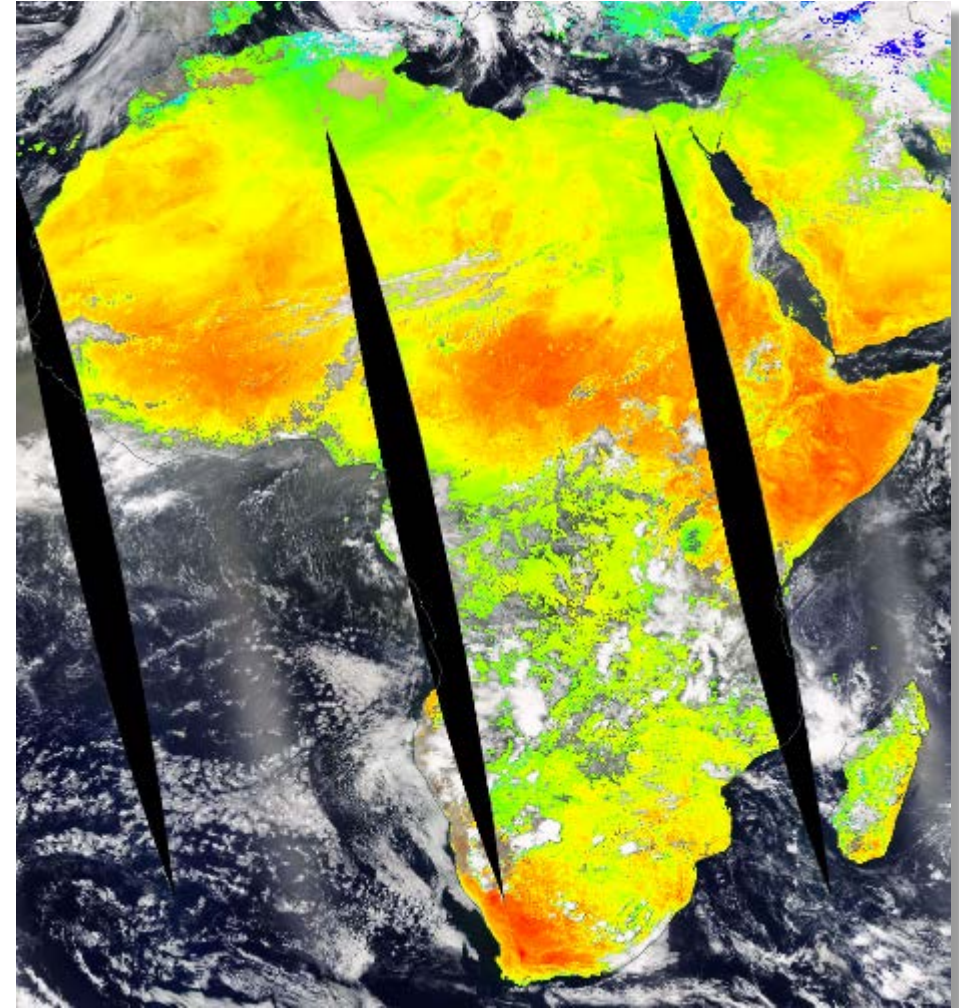
Katie Baynes - NASA / GSFC

Kevin Murphy – NASA / HQ

NASA's Earth Science Data Systems Program

- Actively manages NASA's Earth science data as a national asset (satellite, airborne, and field)
- Develops capabilities optimized to support rigorous science investigations
- Processes (and reprocesses) instrument data to create high quality long-term earth science data records.

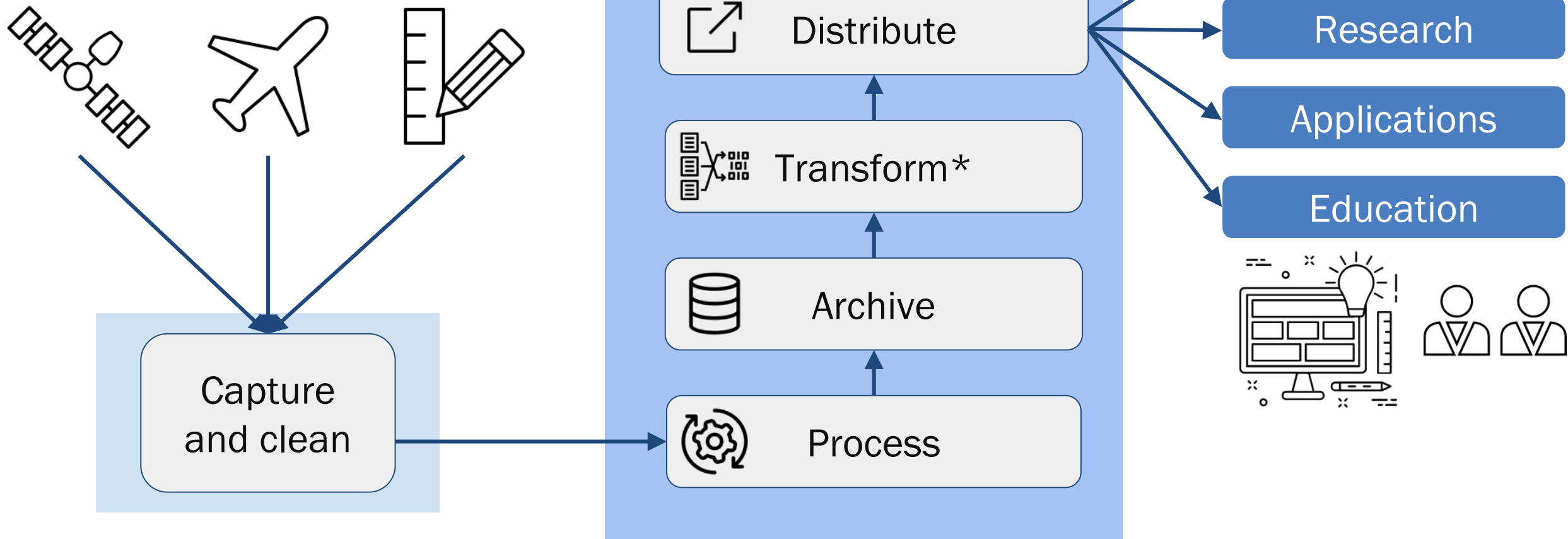
Single largest repository of Earth Science Data, integrating **multivariate/heterogeneous** data from diverse observational platforms.



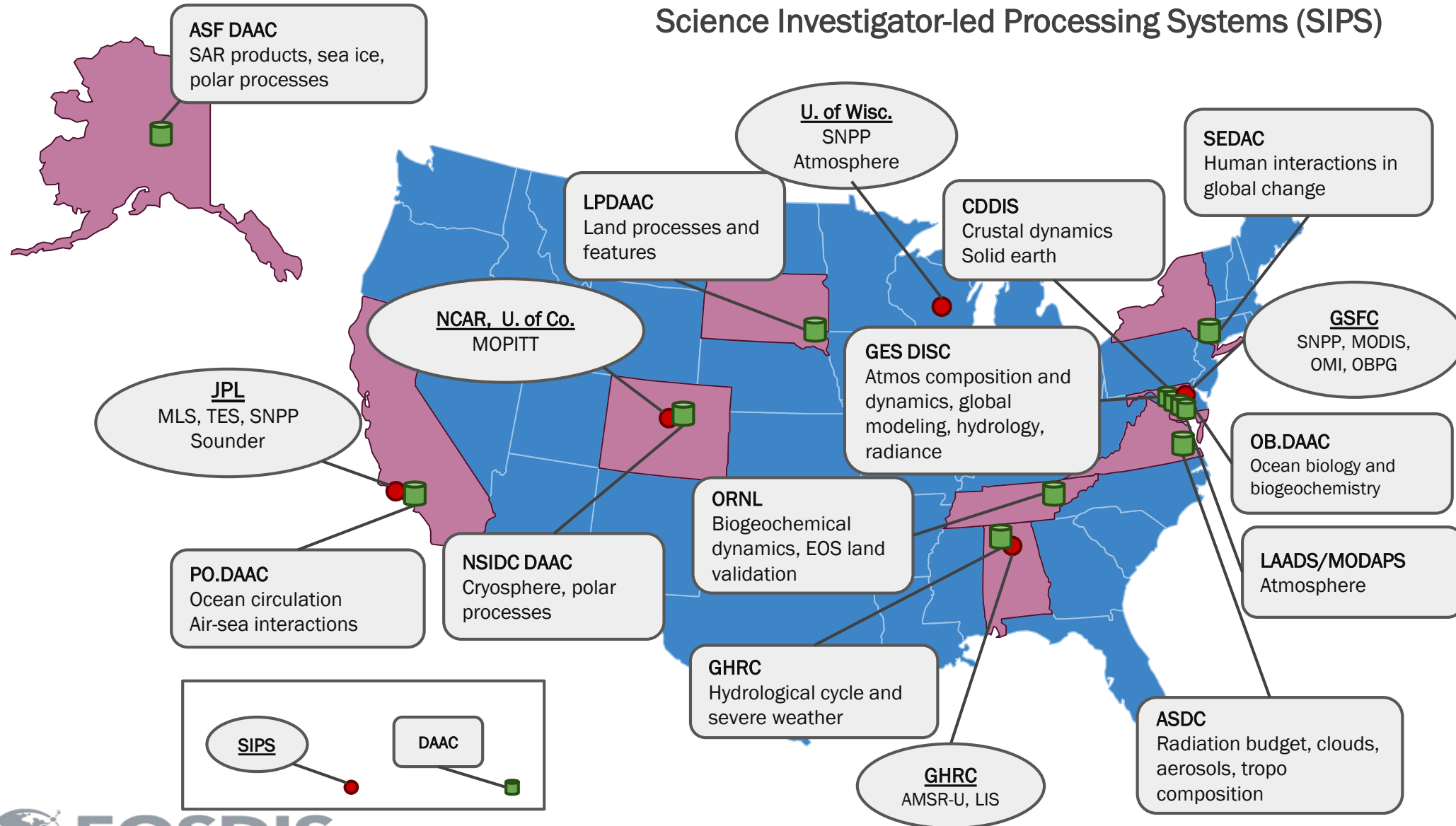
Earth science open data policy

- NASA's Earth Observation data is collected continuously. For over half a century these invaluable records of Earth processes have provided a critical resource for scientists and researchers.
- Since 1994 NASA Earth science data have been free and open to all users for any purpose as quickly as practical after instrument checkout and calibration.

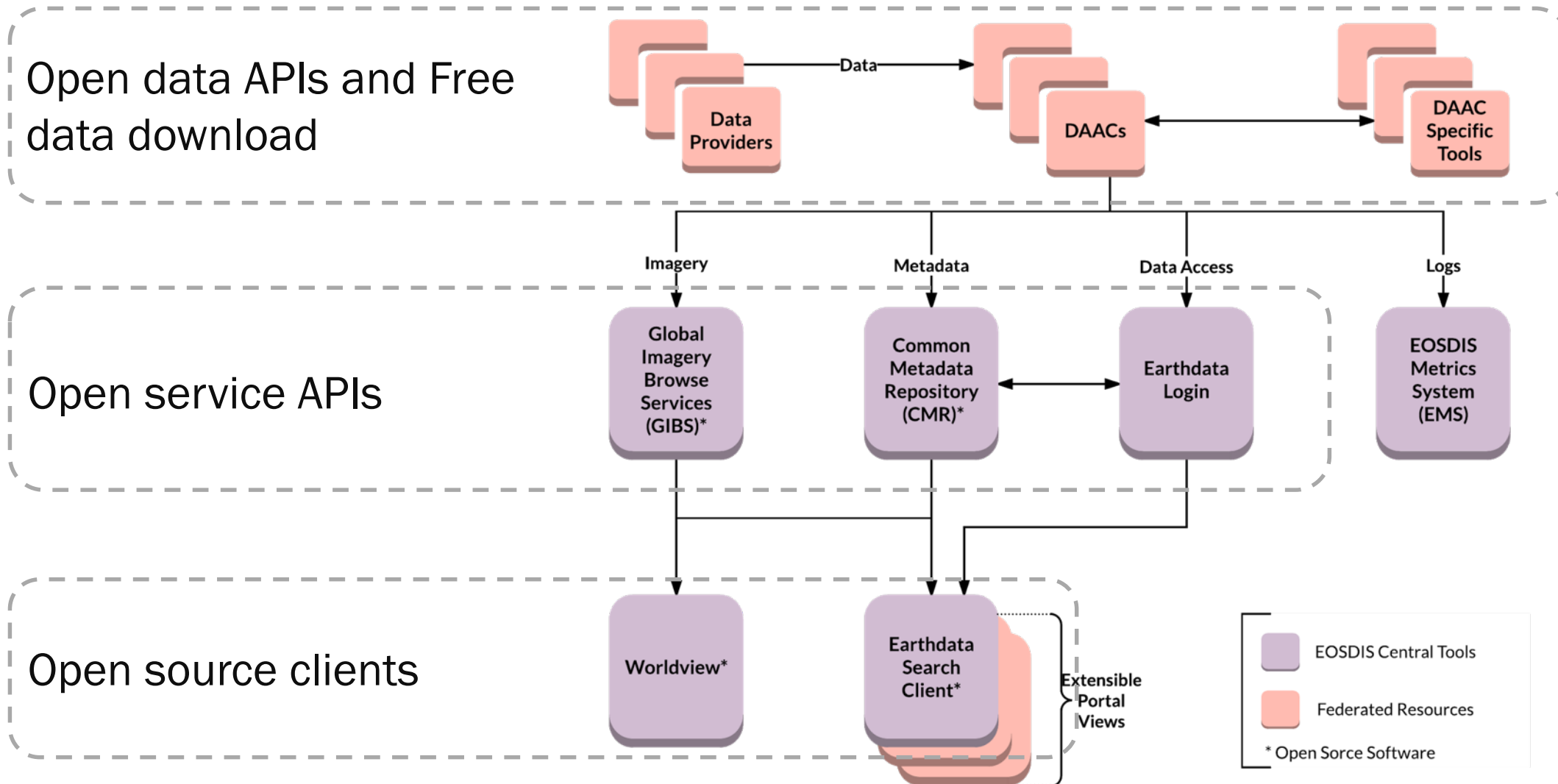
Earth Observing System Data and Information System (EOSDIS)

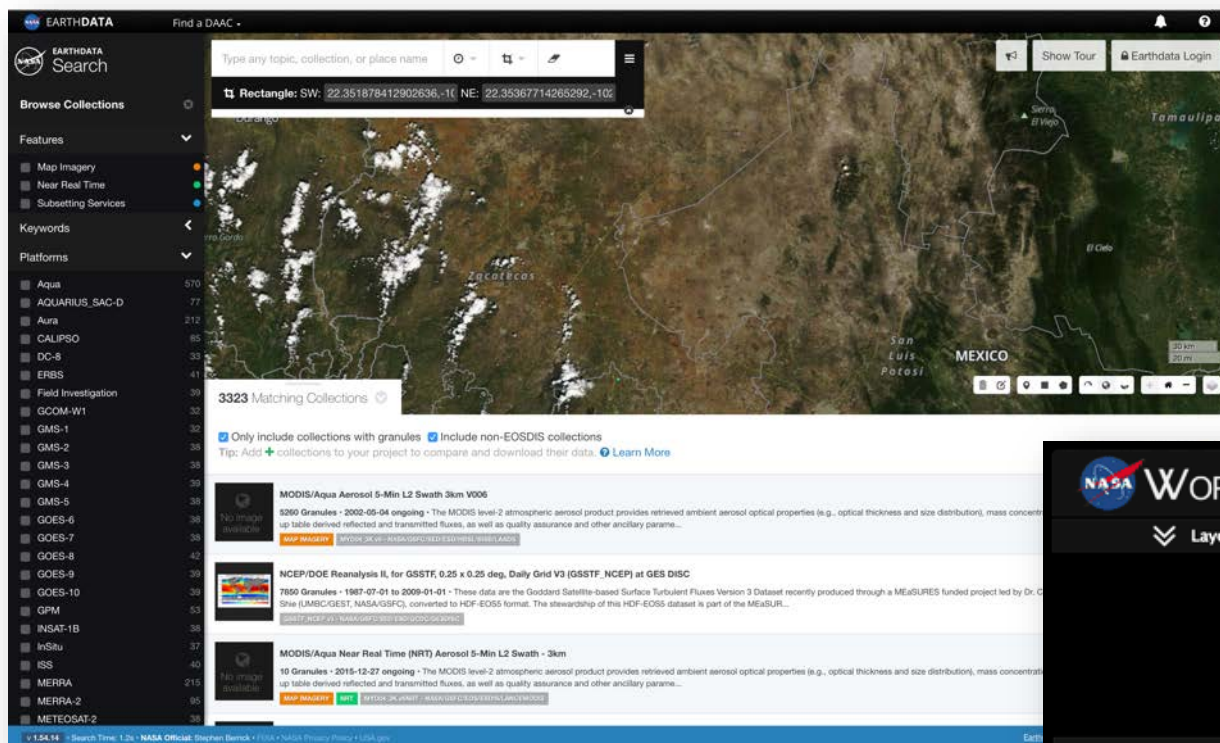


Distributed Active Archive Centers (DAACs), collocated with centers of science discipline expertise, archive and distribute standard data products produced by Science Investigator-led Processing Systems (SIPS)



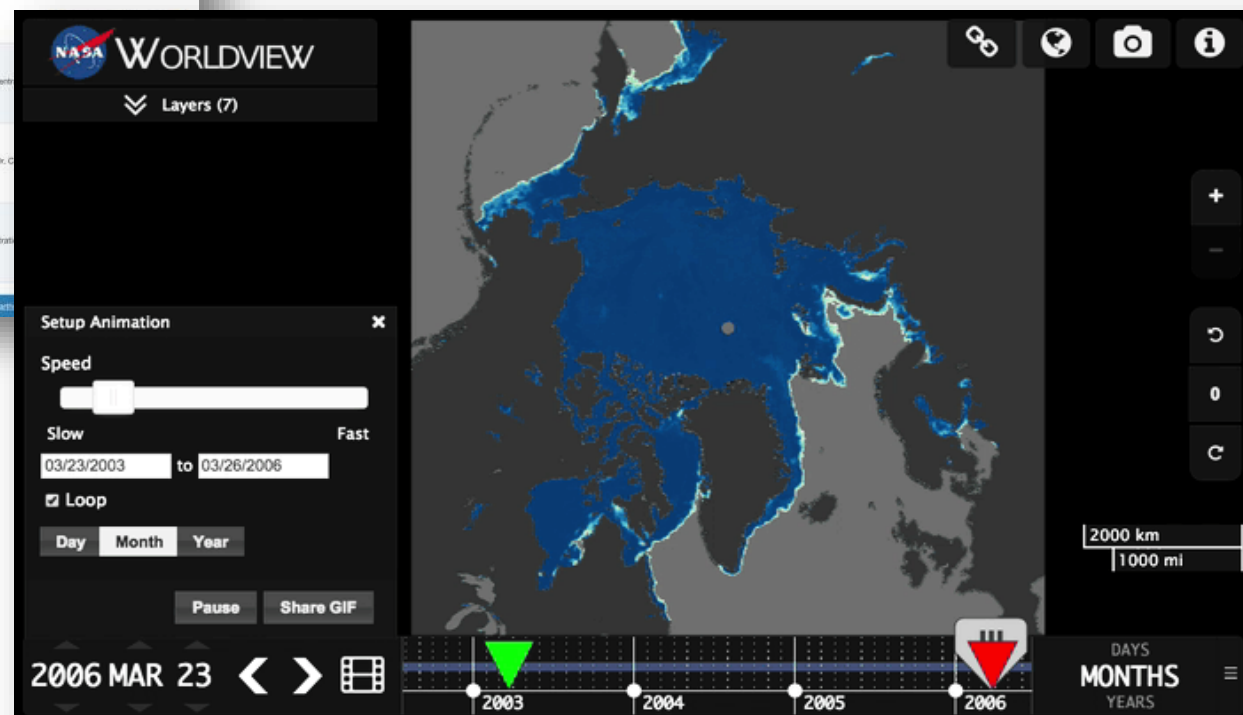
EOSDIS core services





Data-centric users

<https://search.earthdata.nasa.gov>



Imagery-centric users

<https://worldview.earthdata.nasa.gov>

OVERLAYS

- Place Labels
© OpenStreetMap (license), Natural Earth
- Coastlines / Borders / Roads
© OpenStreetMap (license), Natural Earth
- Coastlines
© OpenStreetMap (license)
- Corrected Reflectance (True Color)
Suomi NPP / VIIRS
- Corrected Reflectance (True Color)
Aqua / MODIS
- Corrected Reflectance (True Color)
Terra / MODIS

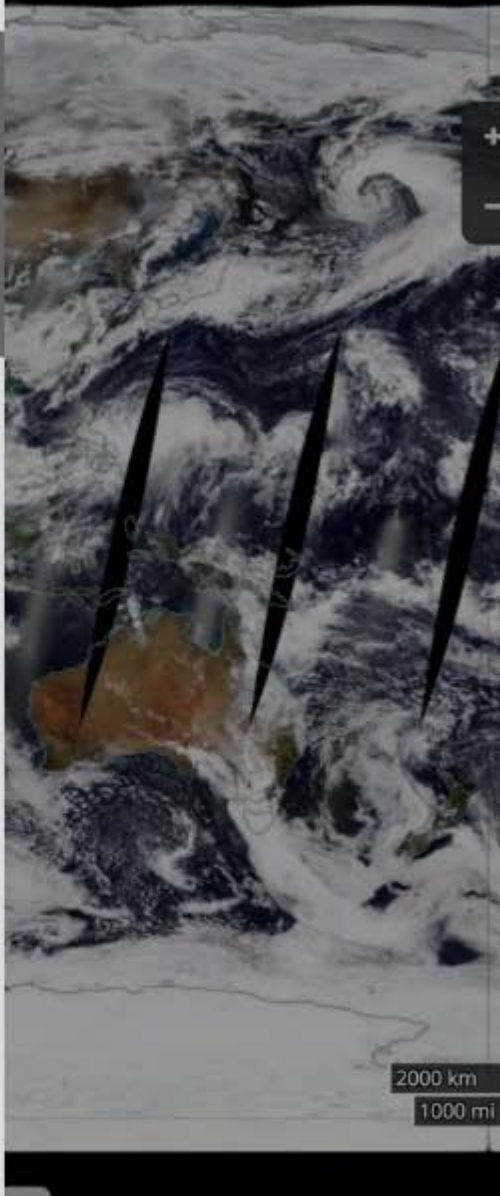
+ Add Layers

Search

Hazards And Disasters Science Disciplines

<p>All</p> <ul style="list-style-type: none"> Aerosol Optical Depth Aerosol Albedo Areas of No Data (mask) Blue Marble Brightness Temperature Carbon Monoxide ... 	<p>Air Quality</p> <ul style="list-style-type: none"> Aerosol Optical Depth Carbon Monoxide Corrected Reflectance Dust Score Fires and Thermal Anomalies Nitric Acid ... 	<p>Ash Plumes</p> <ul style="list-style-type: none"> Aerosol Optical Depth Corrected Reflectance Fires and Thermal Anomalies Land Surface Reflectance Sulfur Dioxide Volcano Hazard
<p>Drought</p> <ul style="list-style-type: none"> Corrected Reflectance Dams Drought Hazard Land Surface Reflectance Land Surface Temperature Precipitation Estimate ... 	<p>Dust Storms</p> <ul style="list-style-type: none"> Aerosol Optical Depth Dust Score Corrected Reflectance Land Surface Reflectance 	<p>Fires</p> <ul style="list-style-type: none"> Aerosol Optical Depth Fires and Thermal Anomalies Carbon Monoxide Corrected Reflectance Earth at Night Land Surface Reflectance ...
<p>Floods</p> <ul style="list-style-type: none"> Corrected Reflectance Cloud Fraction Cloud Multi Layer Flag Cloud Phase Cloud Pressure Cloud Effective Radius ... 	<p>Severe Storms</p> <ul style="list-style-type: none"> Corrected Reflectance Cloud Fraction Cloud Multi Layer Flag Cloud Phase Cloud Pressure Cloud Effective Radius ... 	<p>Shipping</p> <ul style="list-style-type: none"> Corrected Reflectance Brightness Temperature Land Surface Reflectance Sea Ice Sea Surface Temperature

Smoke Plumes, Vegetation, Other




OVERLAYS

Place Labels
© OpenStreetMap (license), Natural Earth

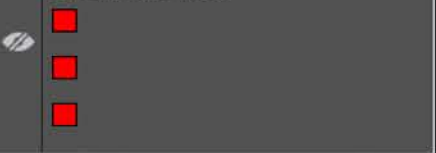
Coastlines / Borders / Roads
© OpenStreetMap (license), Natural Earth

Nighttime Imagery (Day/Night Band, Enhanced Near Constant Contrast)
Suomi NPP / VIIRS

Aerosol Optical Depth
Terra / MODIS



Fires and Thermal Anomalies (Day and Night)
Terra and Aqua / MODIS



Coastlines
© OpenStreetMap (license)

BASE LAYERS

Corrected Reflectance (True Color)
Suomi NPP / VIIRS

Corrected Reflectance (True Color)
Aqua / MODIS

Corrected Reflectance (True Color)
Terra / MODIS

+ Add Layers



50 km / 20 mi scale bar



OVERLAYS

- Place Labels
© OpenStreetMap (license), Natural Earth
- Coastlines / Borders / Roads
© OpenStreetMap (license), Natural Earth
- Nighttime Imagery (Day/Night Band, Enhanced Near Constant Contrast)
Suomi NPP / VIIRS
- Aerosol Optical Depth
Terra / MODIS
Color scale: < 0.000 to 5.000
- Fires and Thermal Anomalies (Day and Night)
Terra and Aqua / MODIS
Color scale: Red
- Coastlines
© OpenStreetMap (license)

BASE LAYERS

- Corrected Reflectance (True Color)
Suomi NPP / VIIRS
- Corrected Reflectance (True Color)
Aqua / MODIS
- Corrected Reflectance (True Color)
Terra / MODIS

+ Add Layers



2017 OCT 09

Navigation icons: back, forward, play, stop.

Timeline navigation: AUG 2017, SEP 2017, OCT 2017, NOV

DAYS MONTHS YEARS

OVERLAYS

Place Labels
 © OpenStreetMap (license), Natural Earth

Coastlines / Borders / Roads
 © OpenStreetMap (license), Natural Earth

Nighttime Imagery (Day/Night Band, Enhanced Near Constant Contrast)
 Suomi NPP / VIIRS

Aerosol Optical Depth
 Terra / MODIS

Fires and Thermal Anomalies (Day and Night)
 Terra and Aqua / MODIS

Coastlines
 © OpenStreetMap (license)

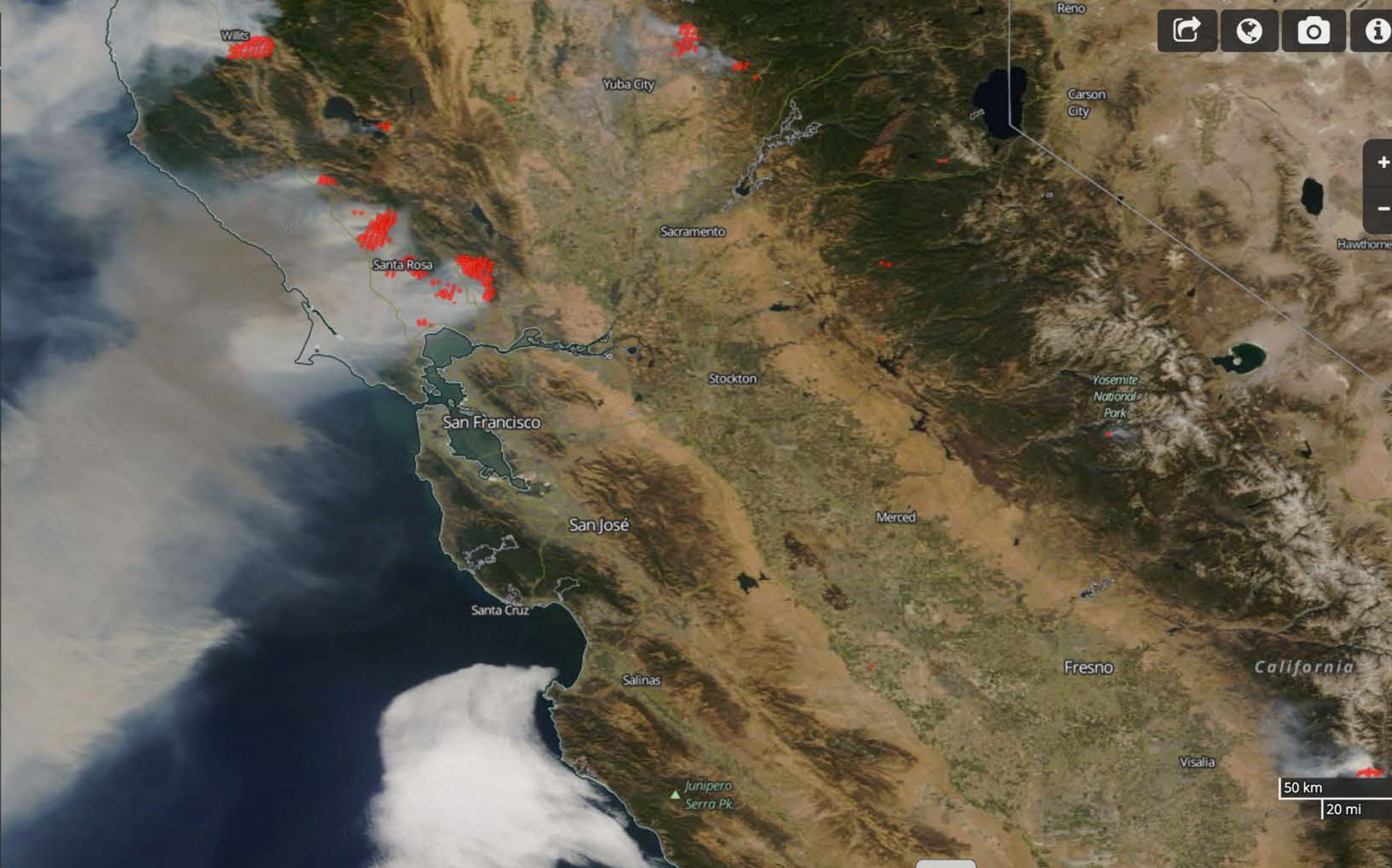
BASE LAYERS

Corrected Reflectance (True Color)
 Suomi NPP / VIIRS

Corrected Reflectance (True Color)
 Aqua / MODIS

Corrected Reflectance (True Color)

+ Add Layers



50 km
20 mi



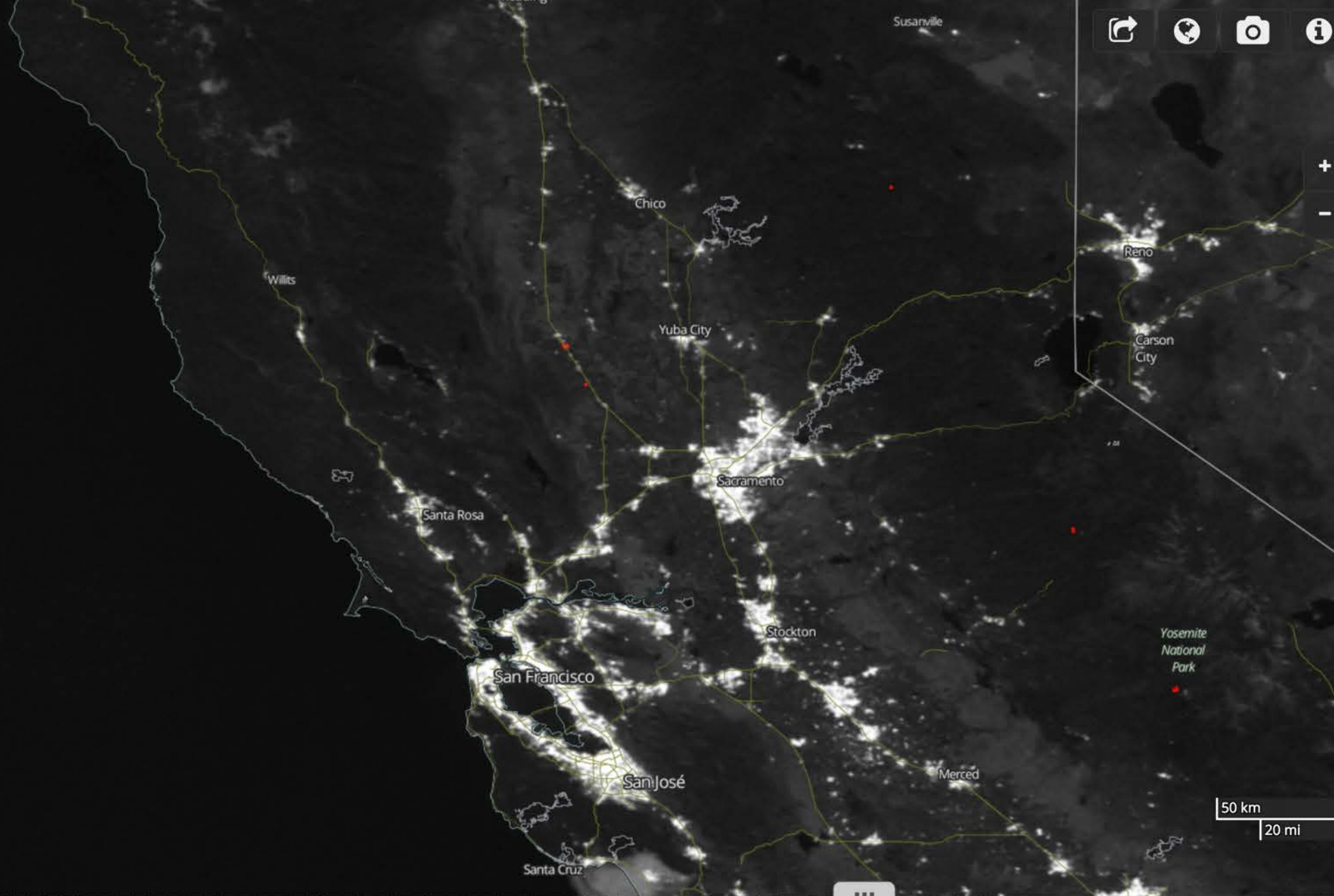
OVERLAYS

- Place Labels
© OpenStreetMap (license), Natural Earth
- Coastlines / Borders / Roads
© OpenStreetMap (license), Natural Earth
- Fires and Thermal Anomalies (Day and Night)
Terra and Aqua / MODIS
- Nighttime Imagery (Day/Night Band, Enhanced Near Constant Contrast)
Suomi NPP / VIIRS
- Aerosol Optical Depth
Terra / MODIS
- Coastlines
© OpenStreetMap (license)

BASE LAYERS

- Corrected Reflectance (True Color)
Suomi NPP / VIIRS
- Corrected Reflectance (True Color)
Aqua / MODIS
- Corrected Reflectance (True Color)
Terra / MODIS

+ Add Layers



2017 OCT 08



DAYS MONTHS YEARS

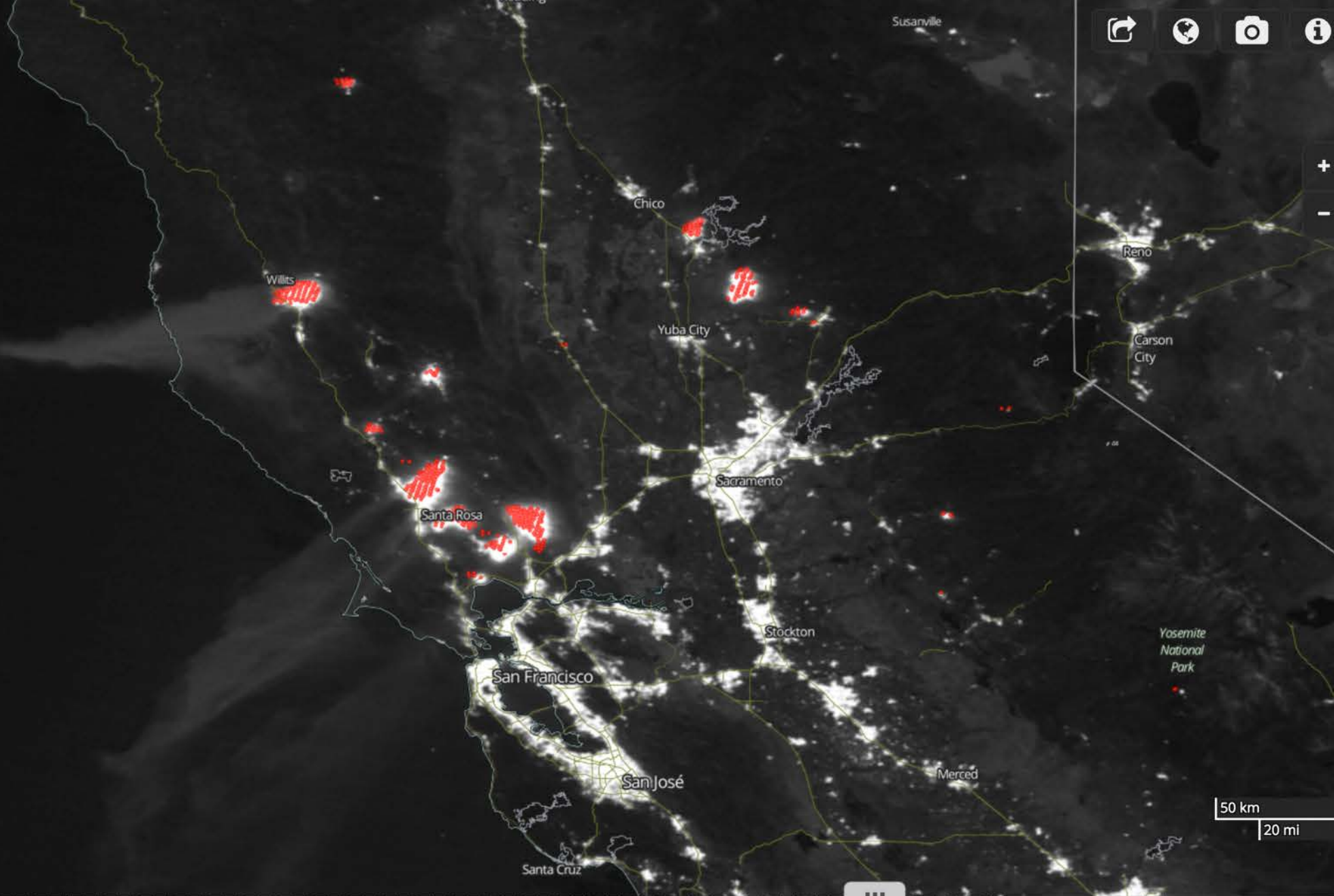
OVERLAYS

- Place Labels
© OpenStreetMap (license), Natural Earth
- Coastlines / Borders / Roads
© OpenStreetMap (license), Natural Earth
- Fires and Thermal Anomalies (Day and Night)
Terra and Aqua / MODIS
- Nighttime Imagery (Day/Night Band, Enhanced Near Constant Contrast)
Suomi NPP / VIIRS
- Aerosol Optical Depth
Terra / MODIS
- Coastlines
© OpenStreetMap (license)

BASE LAYERS

- Corrected Reflectance (True Color)
Suomi NPP / VIIRS
- Corrected Reflectance (True Color)
Aqua / MODIS
- Corrected Reflectance (True Color)
Terra / MODIS

+ Add Layers



Preparing for the future

5 Years from Today

New instruments and missions.

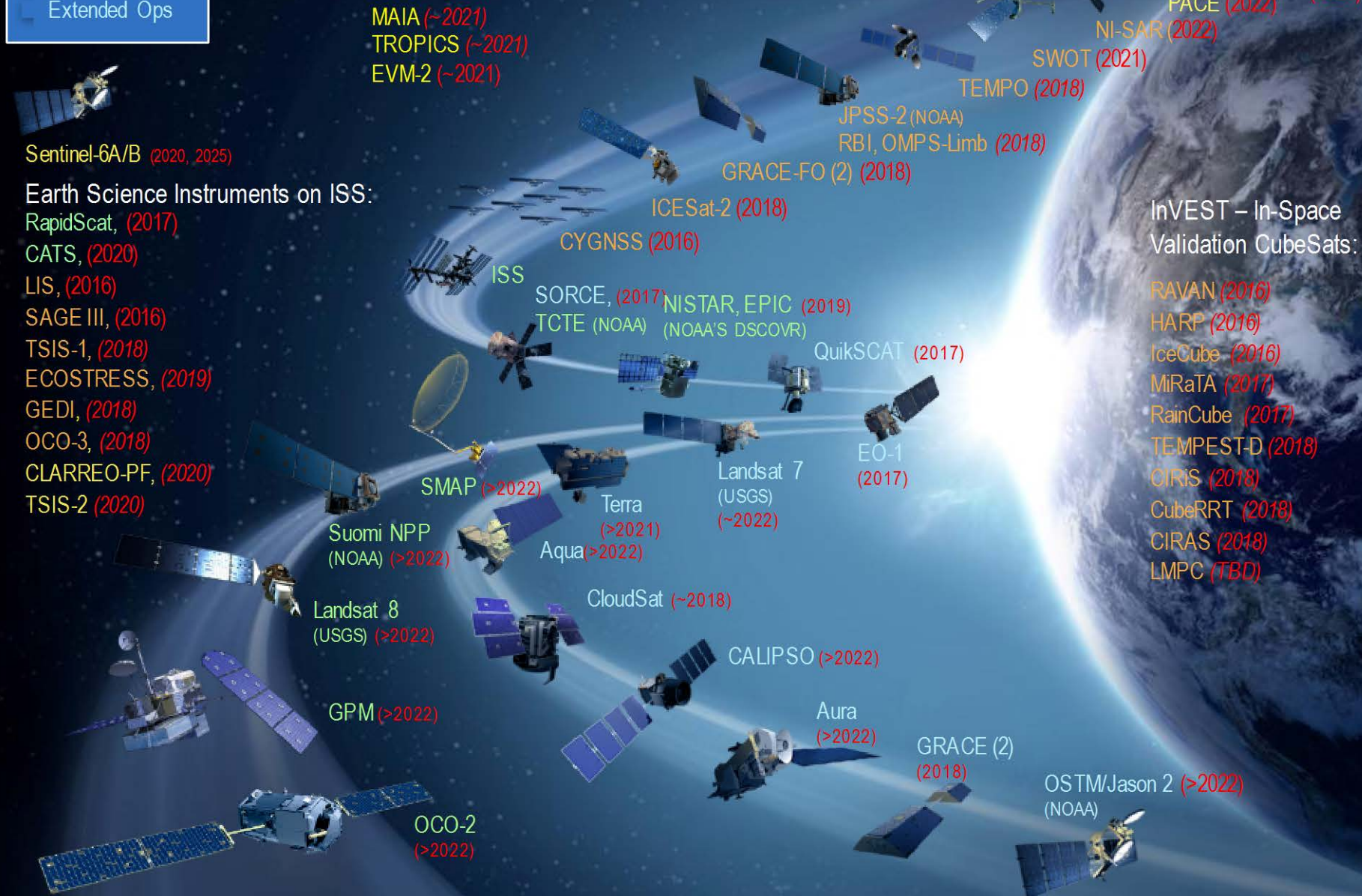
2017 NRC Decadal Survey - **Earth Science** and Applications from Space: National Imperatives for the Next Decade and Beyond

User expectations continue to evolve.



- Formulation
- Implementation
- Primary Ops
- Extended Ops

NASA Earth Science Missions: Present through 2023



EOSDIS is many interconnected systems...

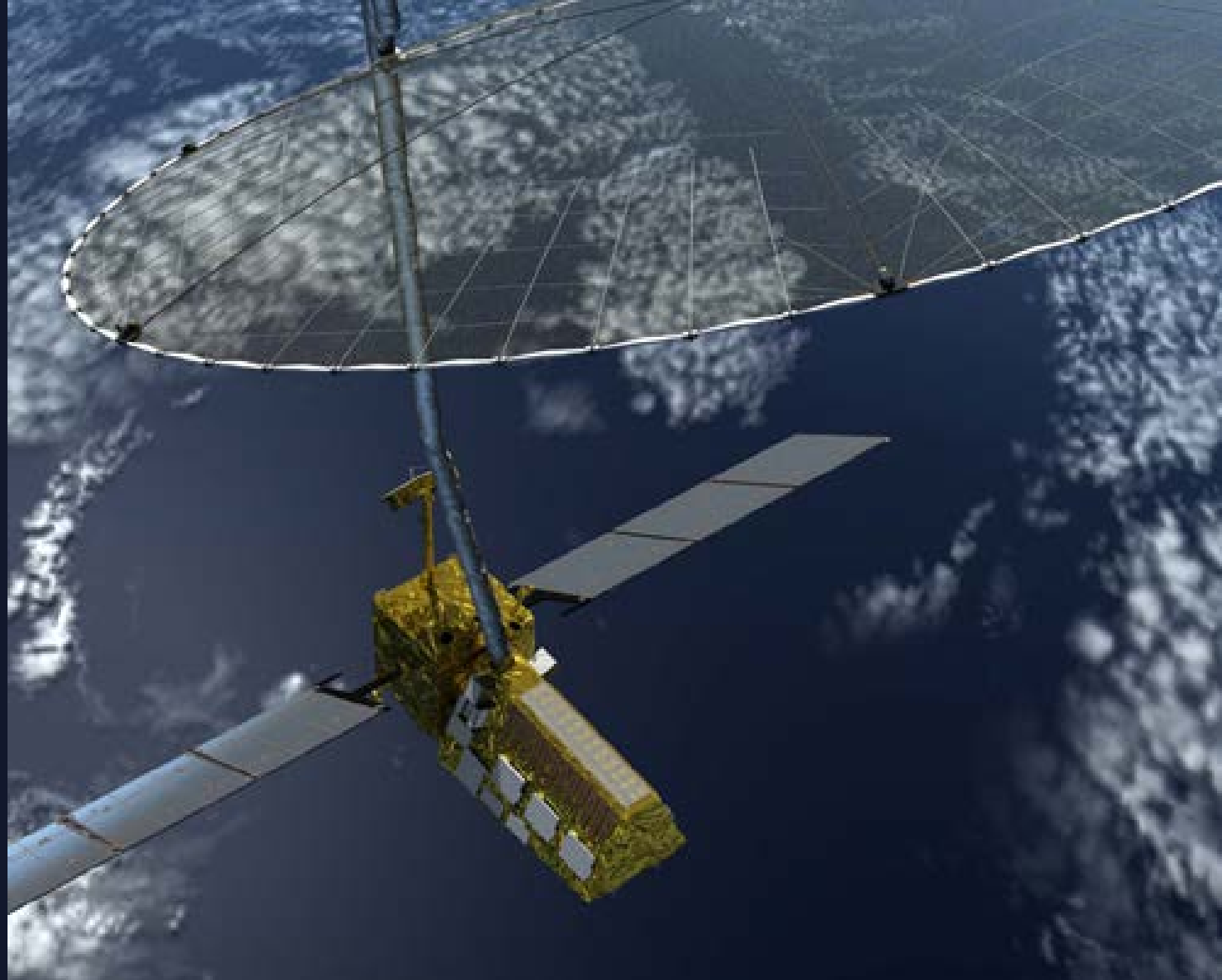
The image shows a screenshot of the NASA WorldView web application interface. On the left is a sidebar with various layers and controls. The main area is a satellite map of California with several red fire hotspots. Overlaid on the map are five circular icons with arrows pointing to specific features: a computer monitor icon, a document icon, a magnifying glass icon, a database icon, and a gear icon. To the right of the map, there are text annotations: 'Worldview, Earthdata Search web applications', 'Global Imagery Browse Services (GIBS)', 'Common Metadata Repository (CMR)', 'Data ingest, archive, and distribution', and 'Archives, announcements, monitoring, distribution services, etc.' at the bottom. The bottom of the screenshot shows a timeline for the year 2017, with the date '2017 OCT 09' and a 'DAYS MONTHS' selector.

80 TBs/day
generation

400 TBs/day
reprocessing

300 GB
Granules

150 PBs @ 50 Gbps
processing speed for months



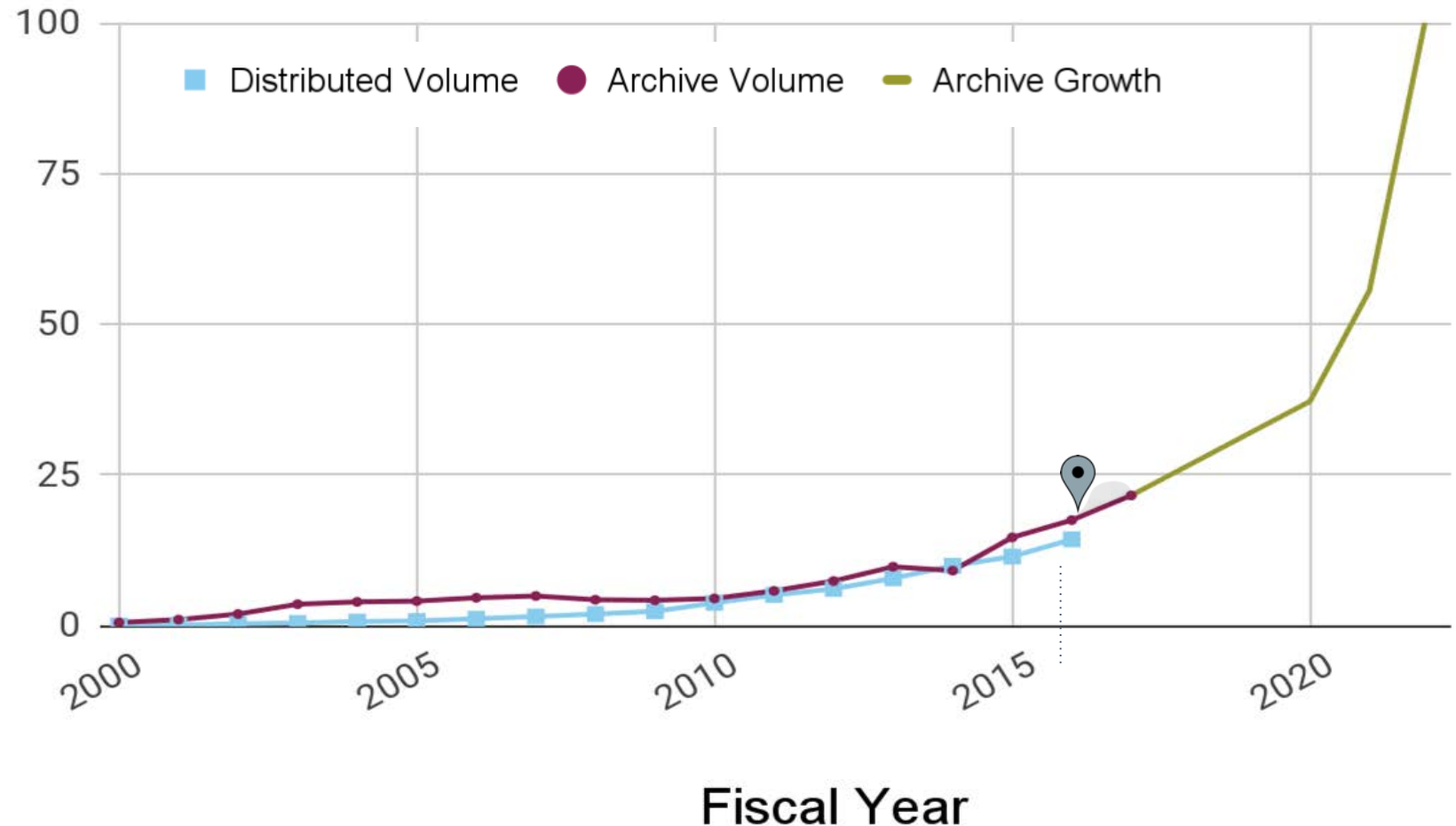
EOSDIS Data System Evolution

EOSDIS is the premier Earth science archive, but we are always looking for ways to improve

The current architecture will not be cost effective as the annual ingest rate increases from 4 to 50PB/year

It will become increasingly difficult and expensive to maintain and improve our current system as data volumes and research demands continue to increase exponentially

EOSDIS is developing **open source cloud native software** for reuse across the agency and throughout the government



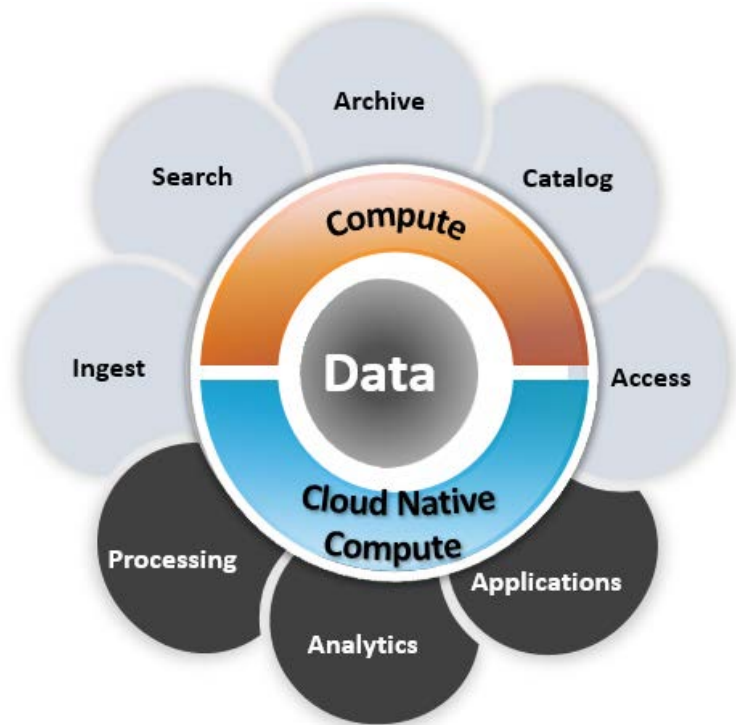
Cloud offers benefits like the ability to analyze data at scale, analyze multiple data sets together easily and avoid lengthy expensive moves of large data sets allowing scientists to work on data “in place”

We **have** to change the paradigm

EOSDIS works well, but can we do better?

- *Can we evolve NASA archives to better support interdisciplinary Earth science researchers?*
- *What system architecture(s) will allow our holdings to become interactive and easier to use for research and commercial users?*
- *Can we afford additional functionality?*
- *How will data from multiple agencies, international partners, and the private sector be combined to study the earth as a system?*
 - *GOES-R, CubeSats, Copernicus...*

Conceptual 'data close to compute'



The operational model of consolidating data—allowing users to compute on the data in place with a platform of common tools—is natural to cloud; it is a cost-effective way to leverage cloud and could be applicable to many businesses and missions

Bring customers to the data

Large volume data storage: Centralized mission observation and model datasets stored in auto graduated AWS object storage (Amazon S3, Amazon S3 IA, Amazon Glacier)

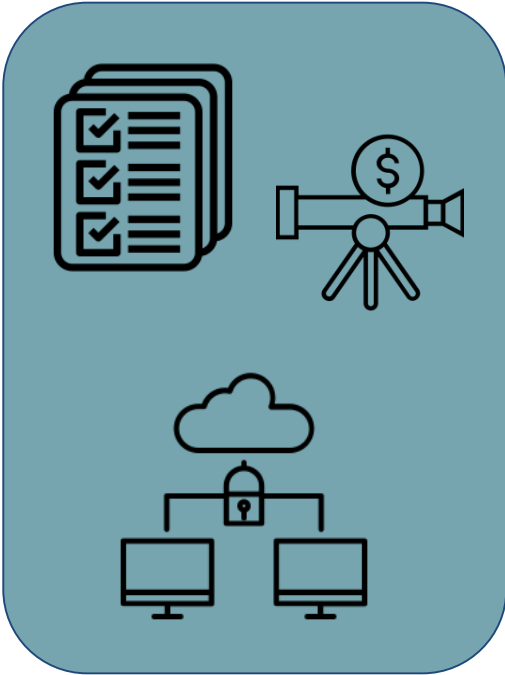
Scalable compute: Provision, access, and terminate dynamically based on need. Cost by use

Cloud Native Compute: Cloud vendor service software stacks and microservices easing deployment of user based applications

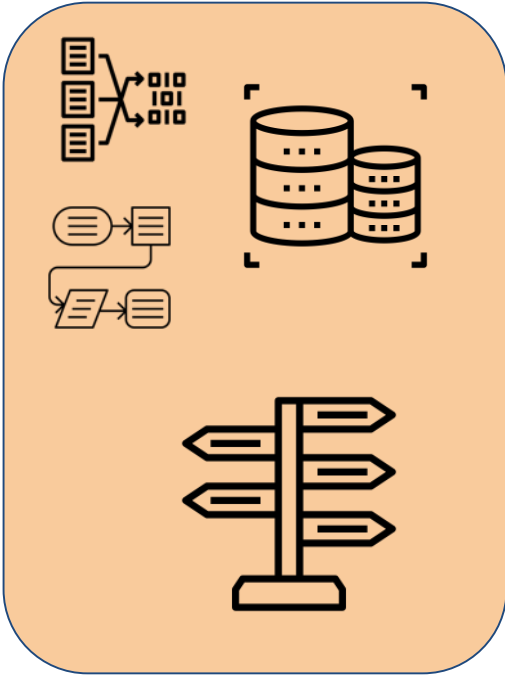
EOSDIS applications and services: Application and service layer using AWS compute, storage (Amazon S3, Amazon S3 IA, Amazon Glacier), and cloud native technologies

Non-EOSDIS/public applications and services: Science community brings algorithms to the data. Support for NASA and non-NASA

Past 24 Months: Focused on evaluation and planning for a cloud migration in 4 areas



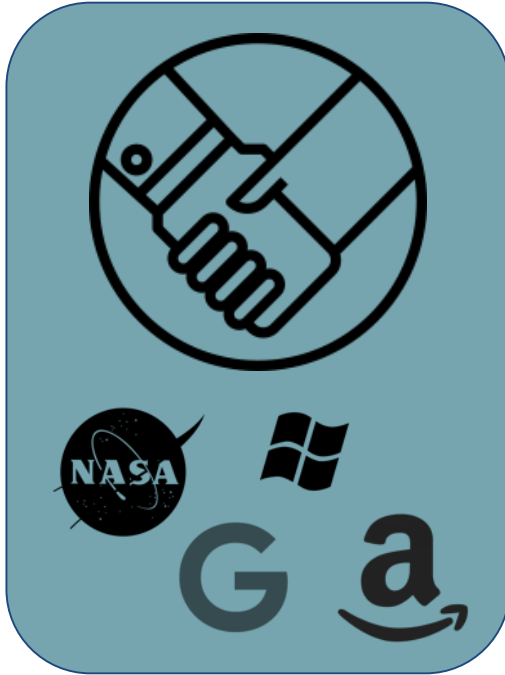
Compliance,
Security, Cost
Tracking



Core Archive
Functionality
and Processing



End-User
Application
Migration



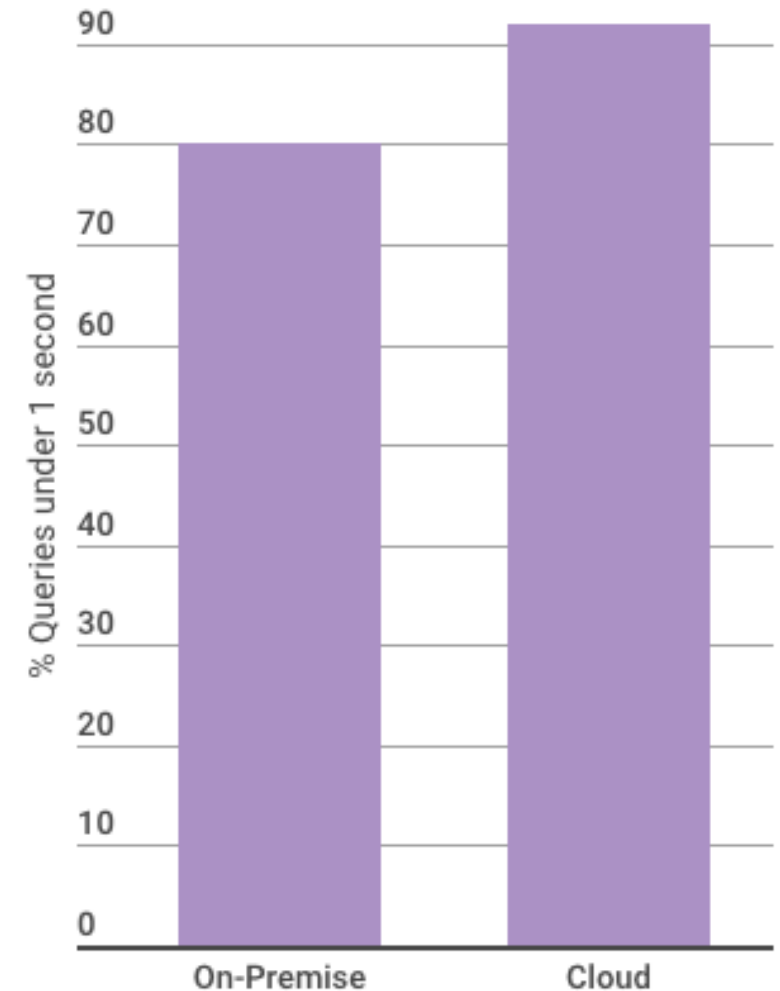
Pursuing Cloud
Partnerships

Starting AWS migration

Since September 2016, EOSDIS has migrated two of its core systems, Common Metadata Repository (CMR) and Earthdata Search, into the Amazon cloud to immense success



- One year migration effort
- Over 500K queries per day
- Open source
- Open access API



Layers Events Data

OVERLAYS

Place Labels © OpenStreetMap (license), Natural Earth

Coastlines / Borders / Roads © OpenStreetMap (license), Natural Earth

Nighttime Imagery (Day/Night Band, Enhanced Near Constant Contrast) Suomi NPP / VIIRS



Aerosol Optical Depth Terra / MODIS



Fires and Thermal Anomalies (Day and Night) Terra and Aqua / MODIS



Coastlines © OpenStreetMap (license)

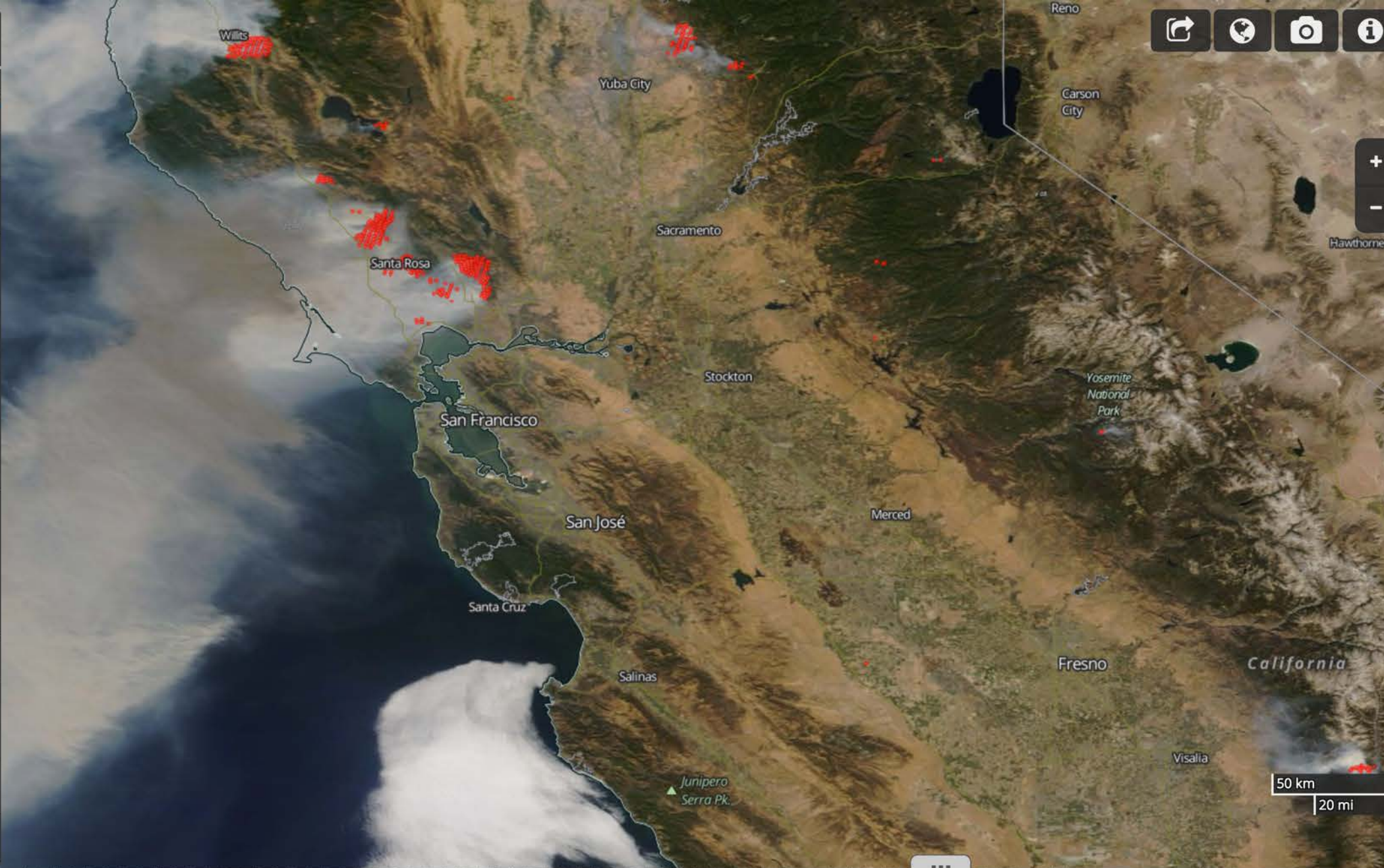
BASE LAYERS

Corrected Reflectance (True Color) Suomi NPP / VIIRS

Corrected Reflectance (True Color) Aqua / MODIS

Corrected Reflectance (True Color)

+ Add Layers

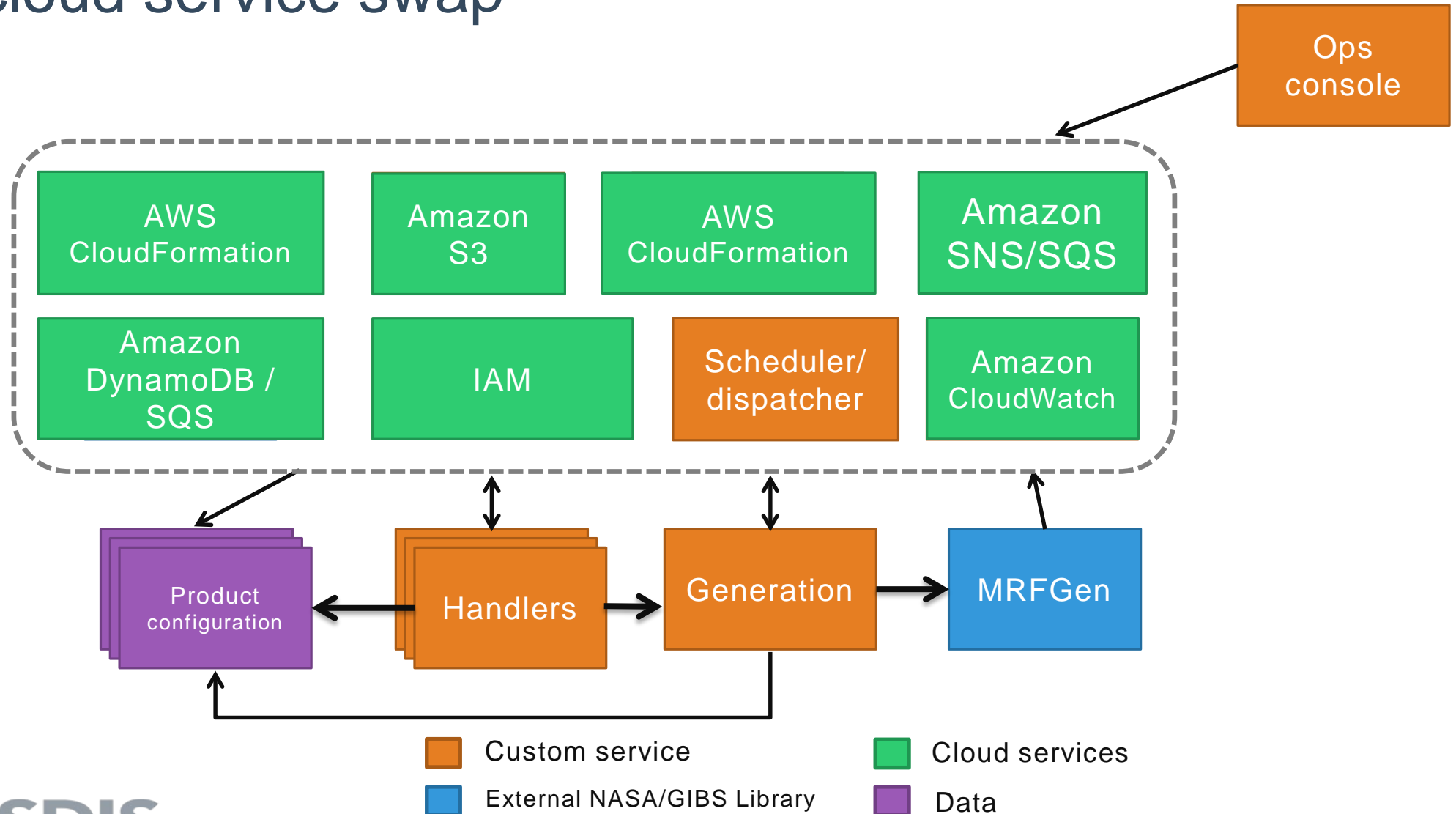


2017 OCT 09 Navigation arrows

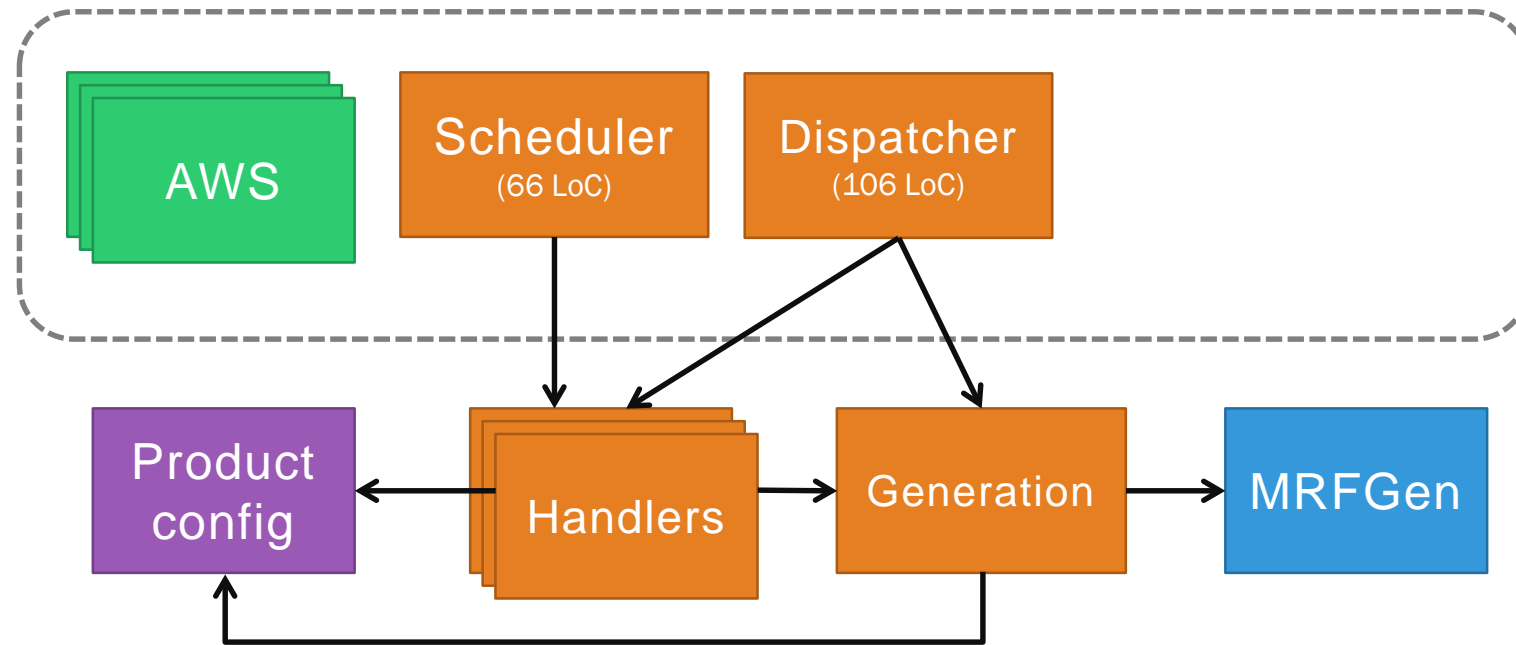


DAYS MONTHS YEARS

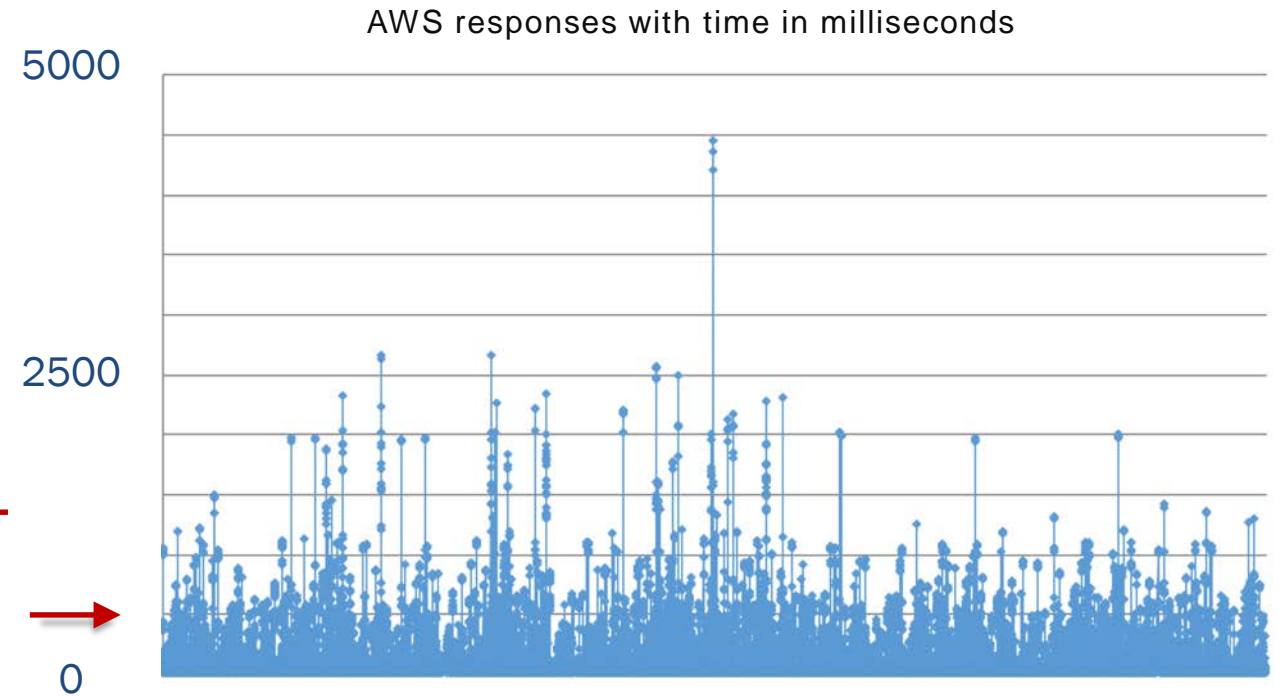
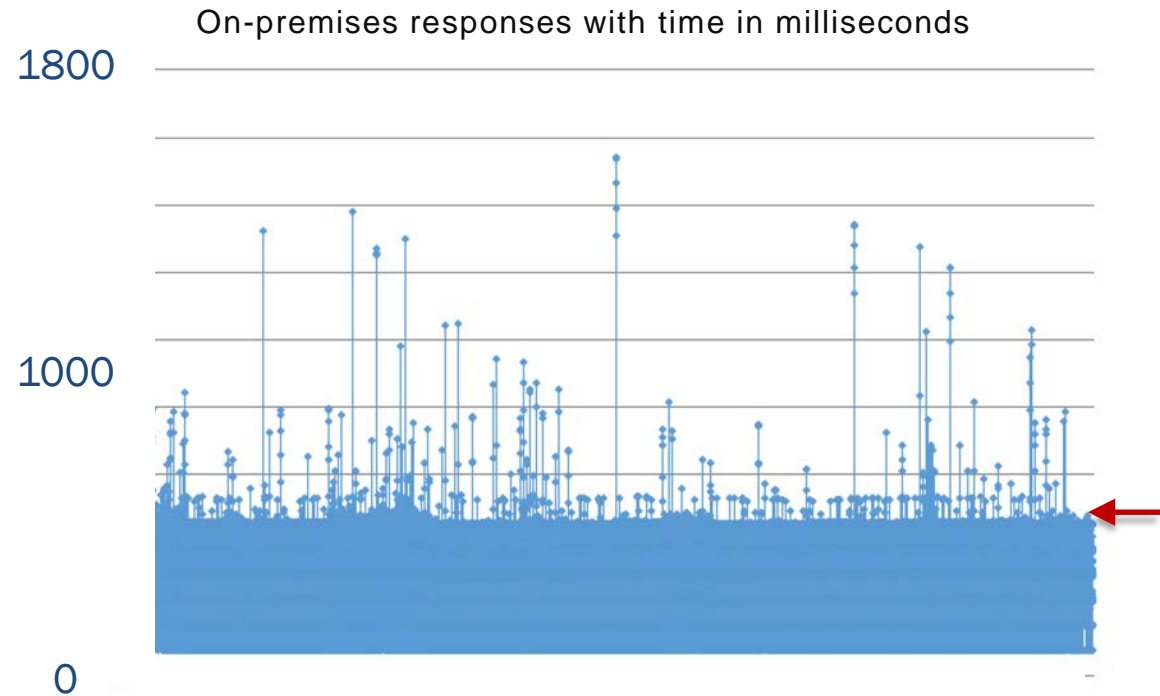
Global Imagery Browse Service (GIBS) in the cloud service swap



GIBS-in-the-cloud ingest & processing



Cloud performance affected architecture



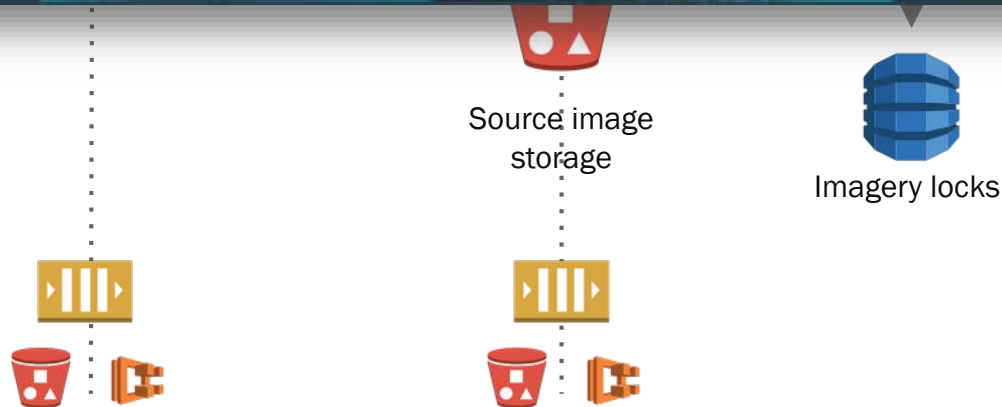
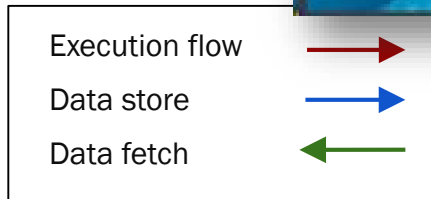
On-premises implementation showed consistent performance during load testing vs. more sporadic latencies in AWS

Ingest: Earth science Imagery Processing

Discover

Sync

Process

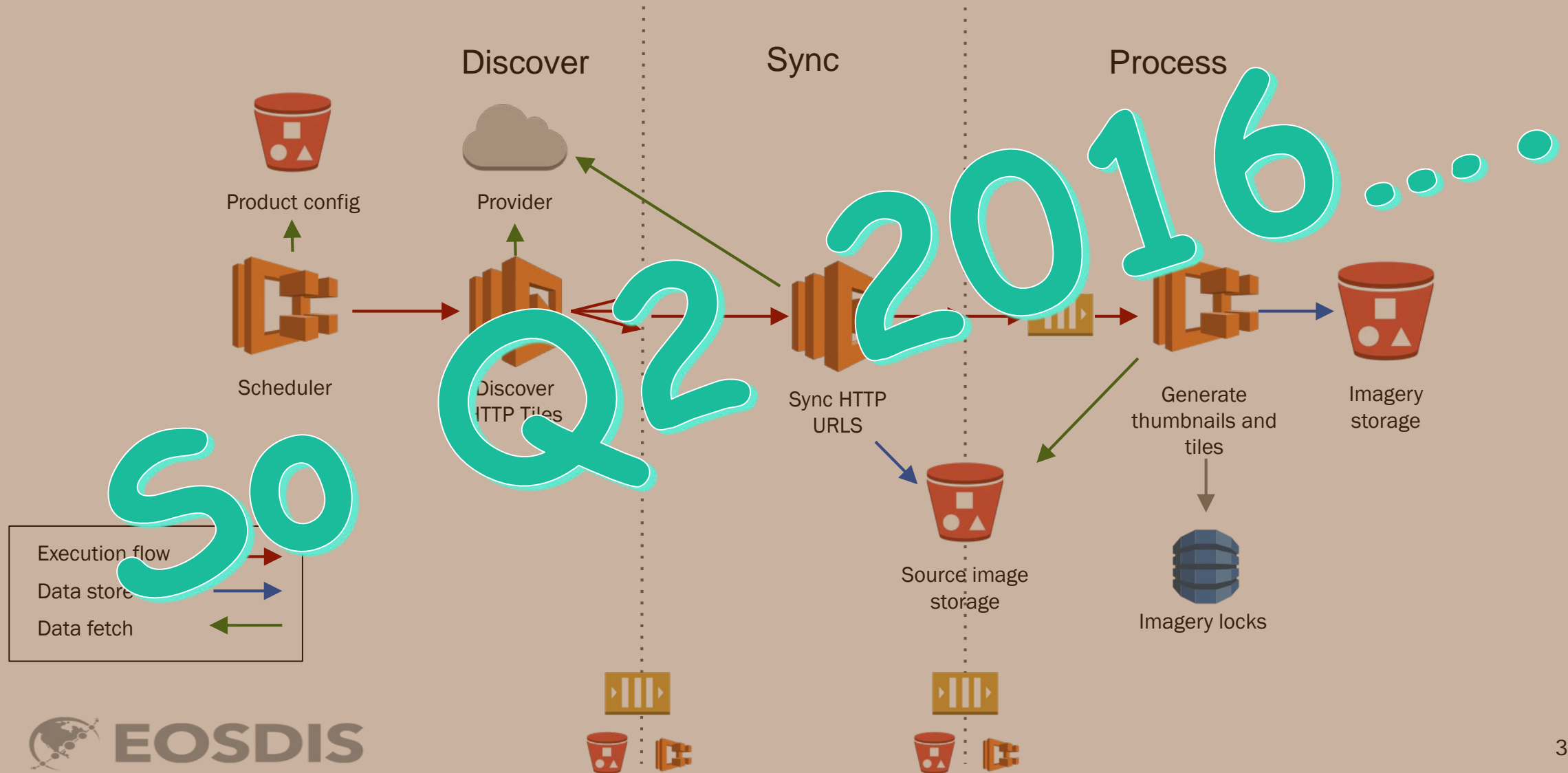


OCTOBER 17-20, 2017 AWS announcements!

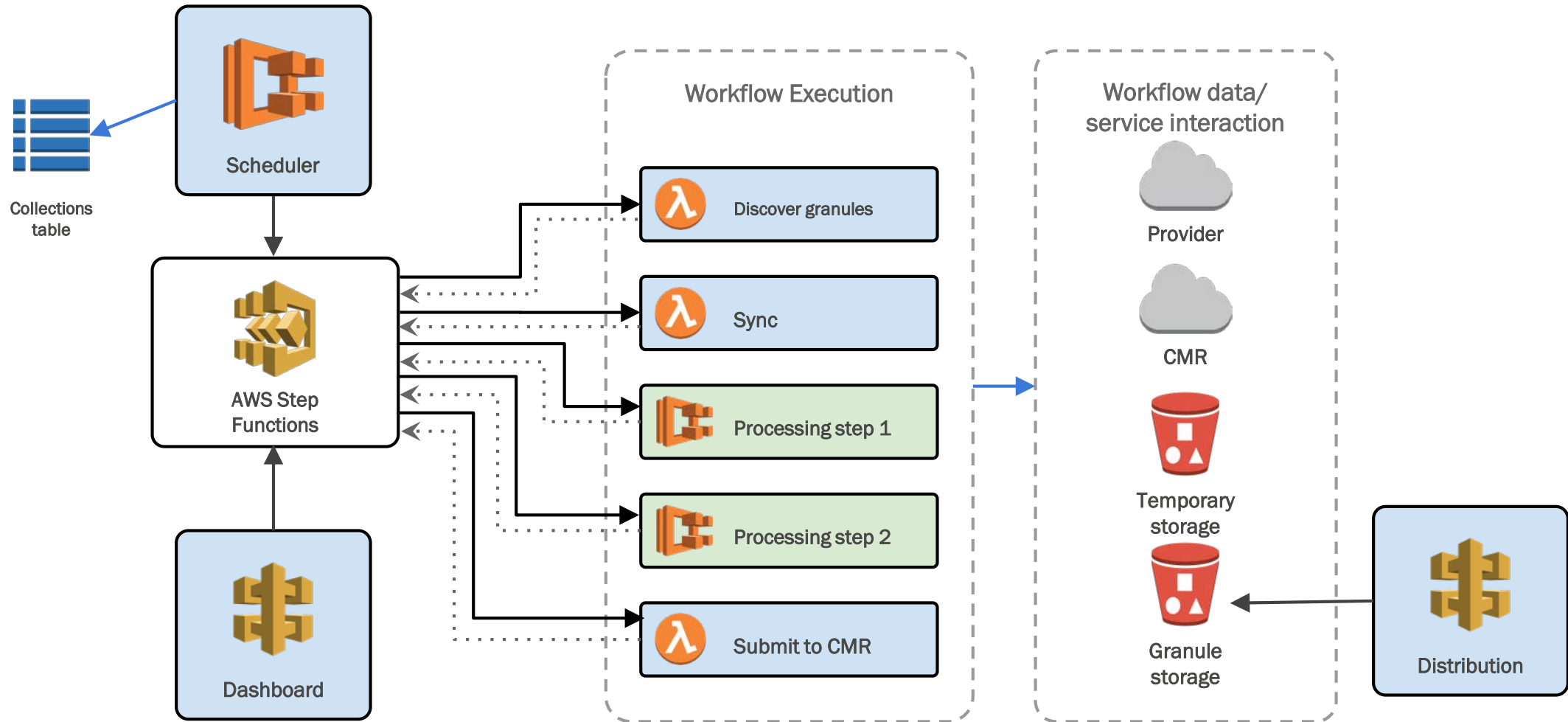
Most Recent Announcements from AWS

Date	Announcement
Oct 20	AWS Config Adds Support for AWS CodeBuild
Oct 20	Amazon QuickSight Adds Support for Combo Charts and Row-Level Security
Oct 19	AWS Direct Connect now live in Vancouver, Manchester and Perth
Oct 19	Manage Amazon Simple Queue Service costs using Cost Allocation Tags
Oct 19	Amazon Athena is now available in the EU (Frankfurt) region.
Oct 19	Amazon Redshift Spectrum is now available in Europe (Ireland) and Asia Pacific (Tokyo)
Oct 18	Amazon EC2 Spot Can Now Encrypt your EBS volumes at launch time
Oct 18	AWS Deep Learning AMI Now Supports PyTorch, Keras 2 and Latest Deep Learning Frameworks
Oct 17	Amazon Redshift announces Dense Compute (DC2) nodes with twice the performance as DC1 at the same price
Oct 18	AWS Marketplace: Announcing Availability of Multi-AMI Solutions.

Ingest: Earth science imagery processing...



Ingest & Archive with AWS Step Functions



Cumulus Major System Components

A lightweight framework consisting of:

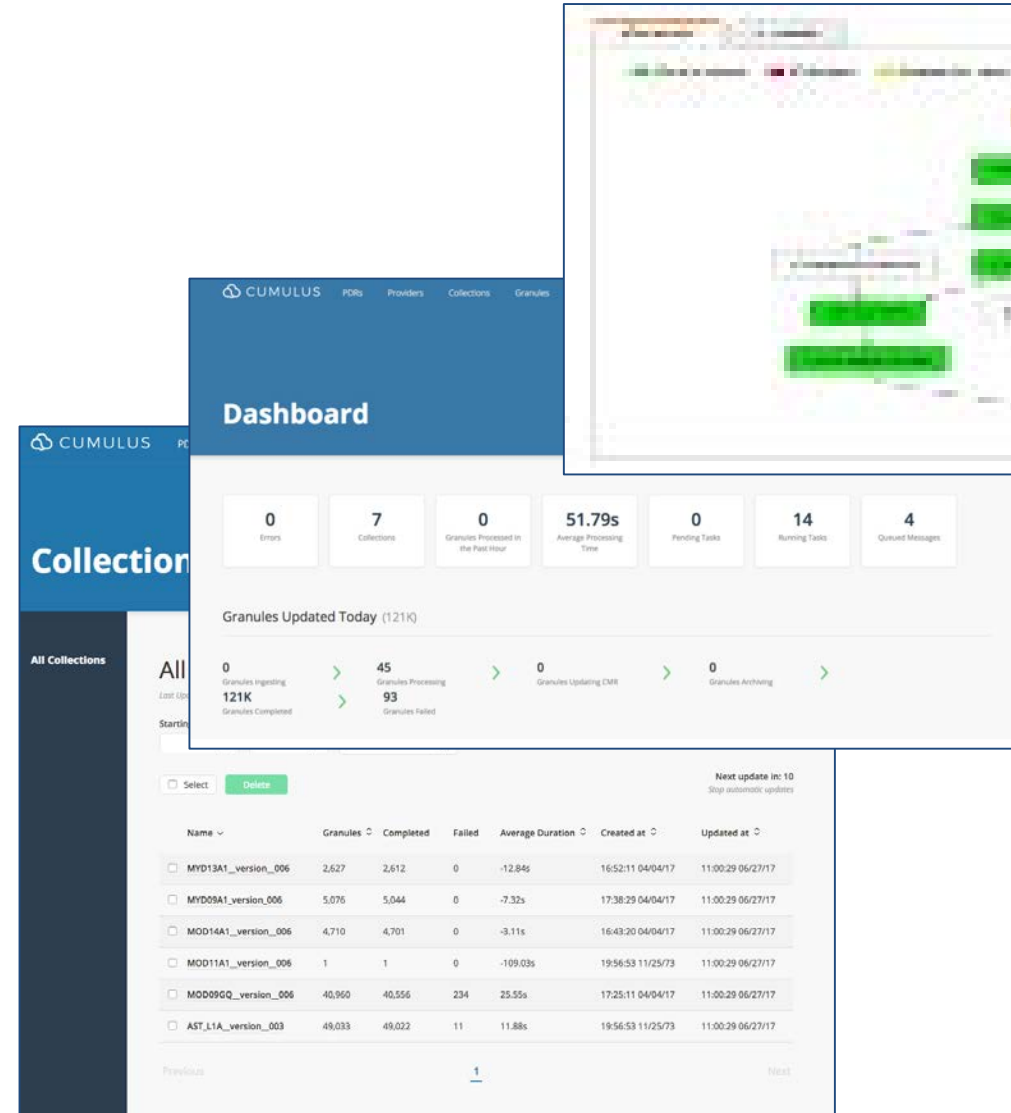
Tasks a discrete action in a workflow, invoked as a Lambda function or EC2 service, common protocol supports chaining

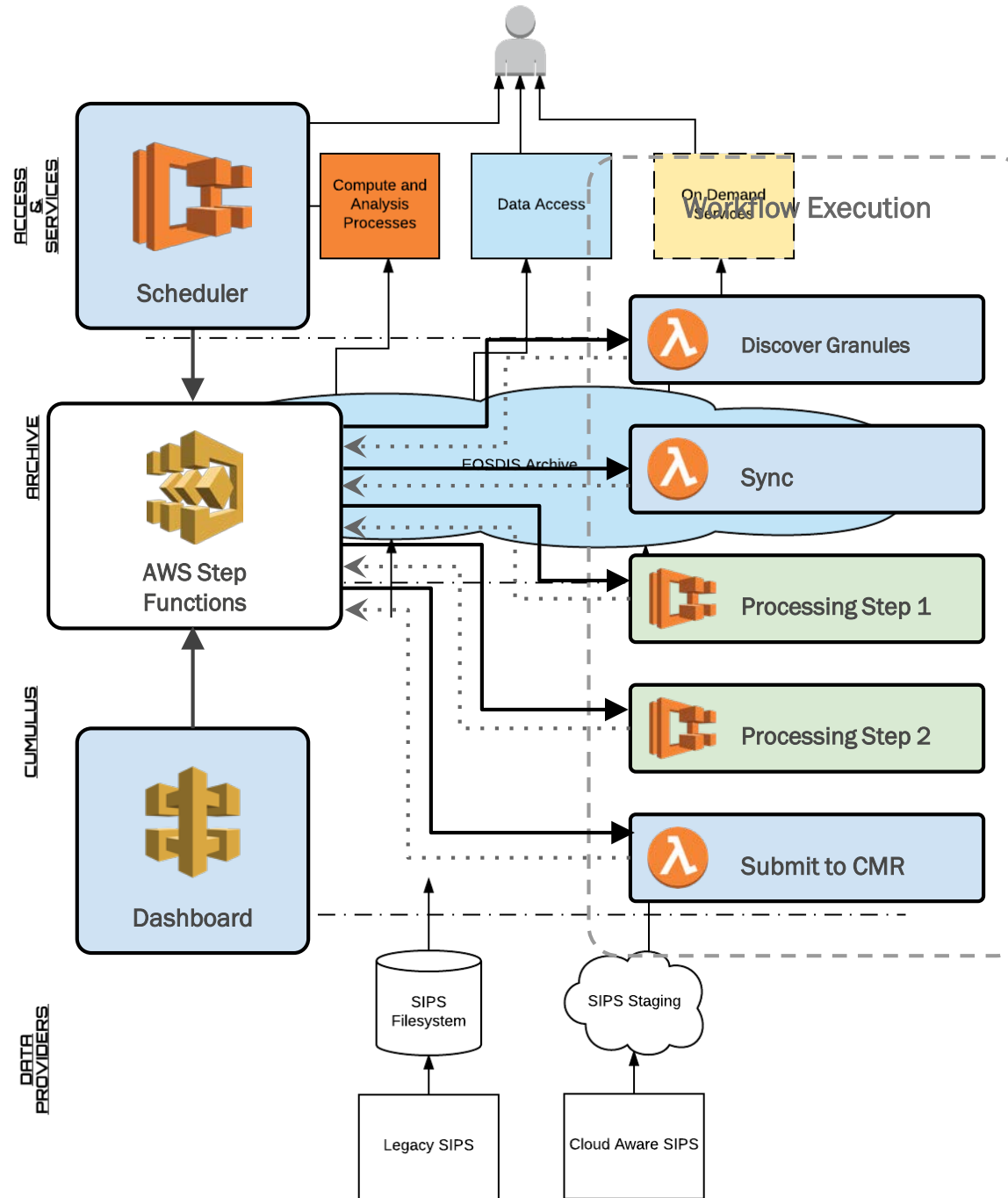
Orchestration engine (AWS Step Functions) that controls invocation of tasks in a workflow

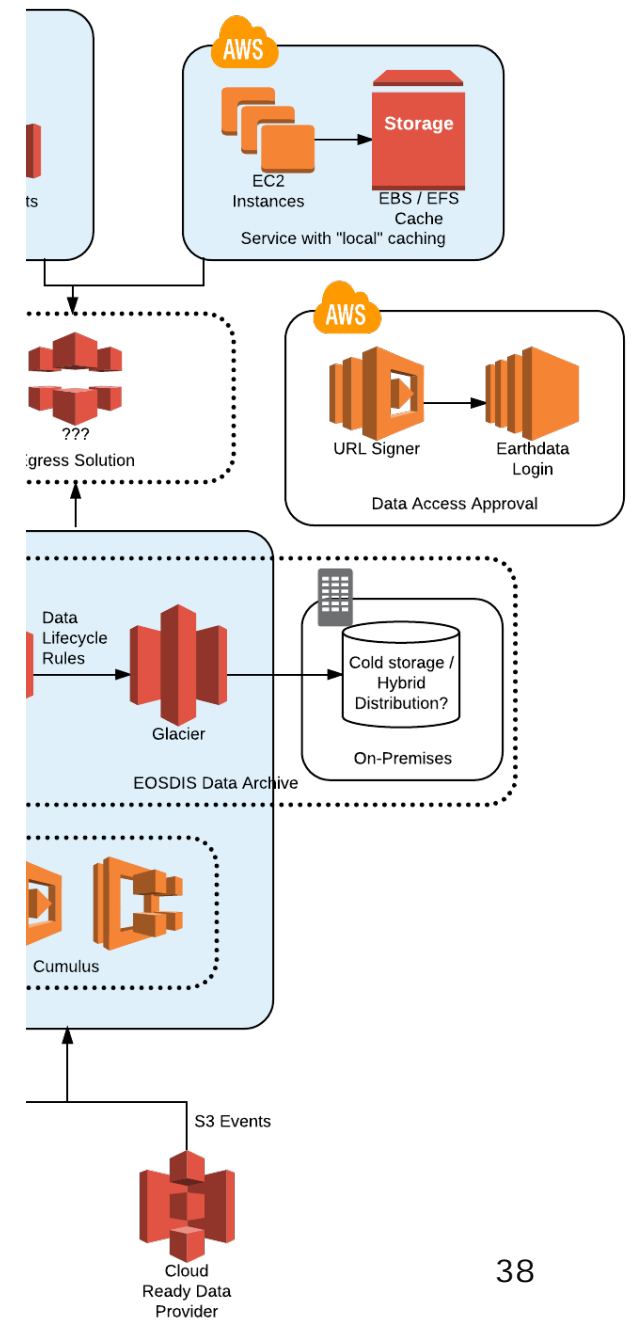
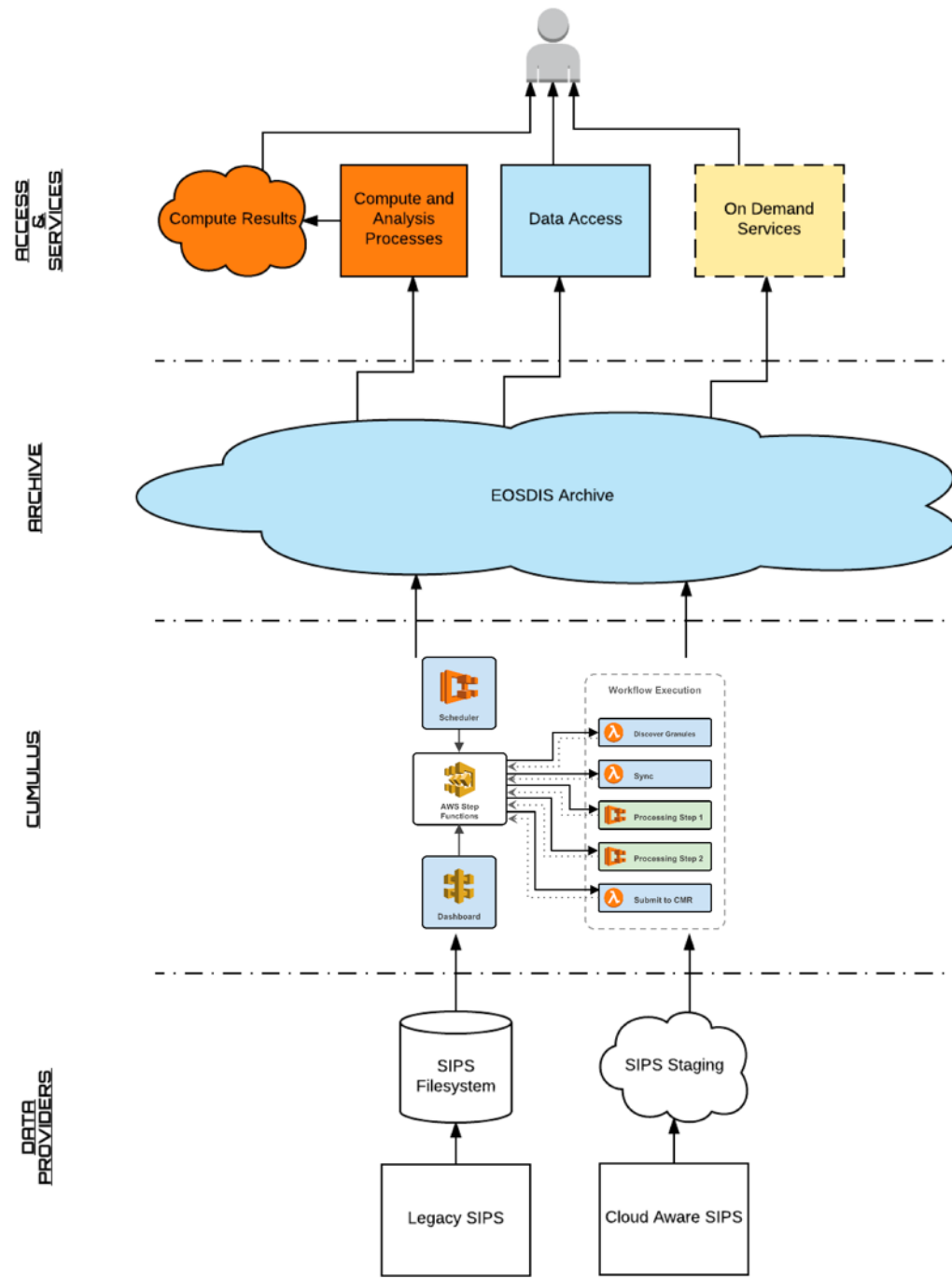
Database store status, logs, and other system state information

Workflows(s) file(s) that define the ingest, processing, publication, and archive operations (json)

Dashboard create and execute workflows, monitor system

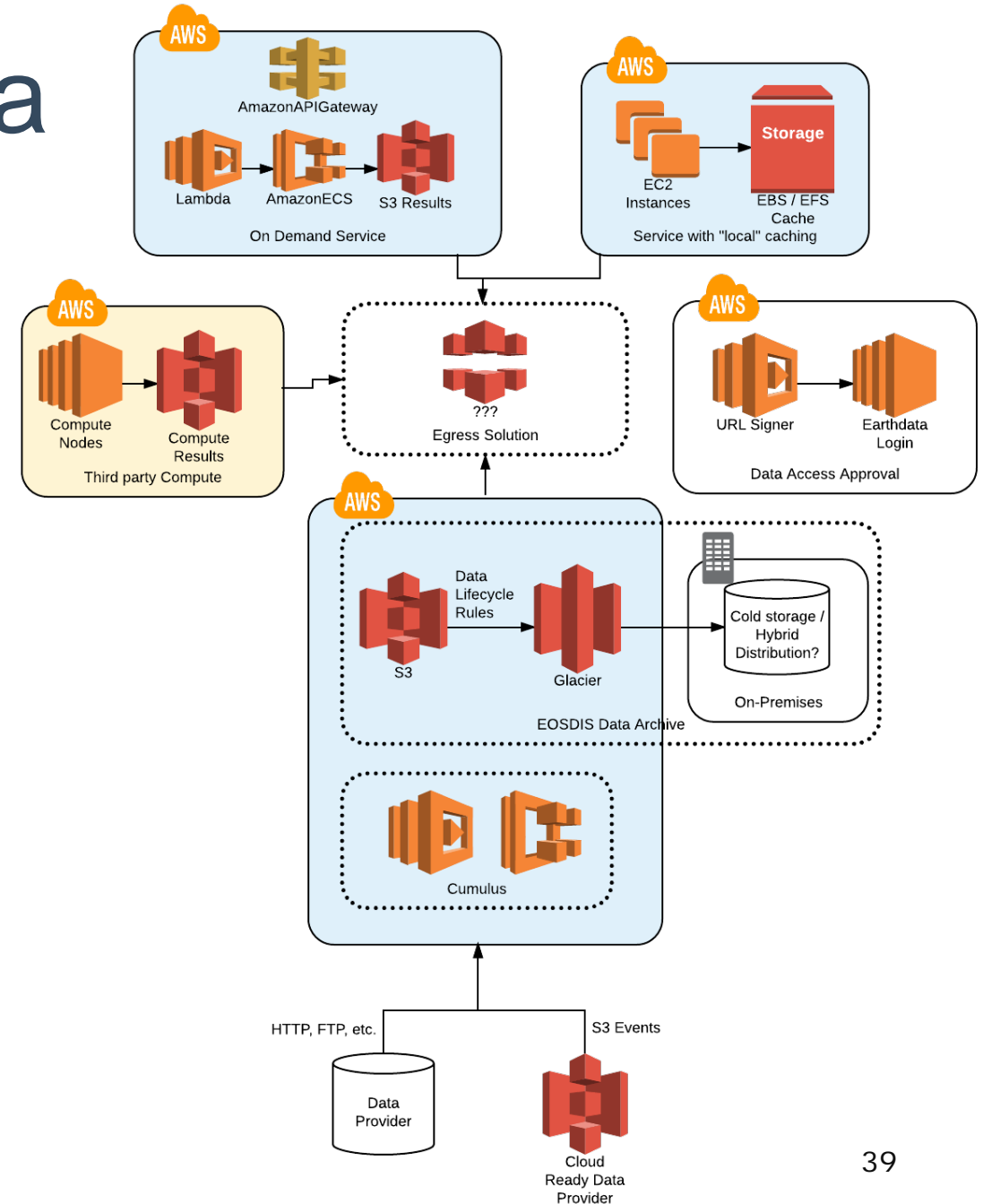




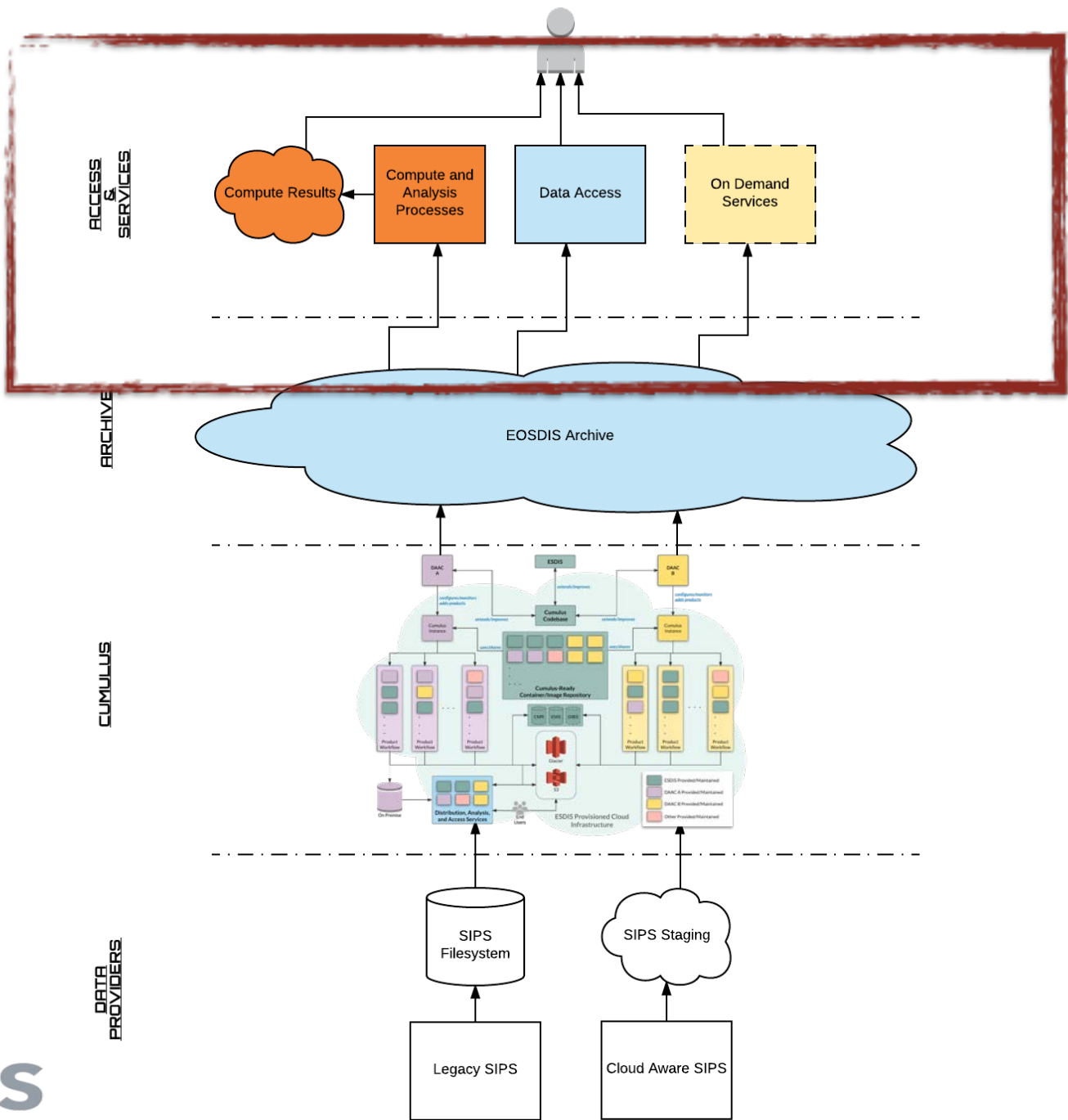


Cloud scale science data

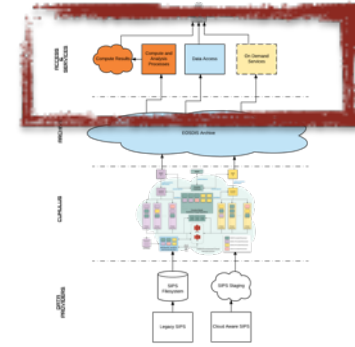
- **Data can be generated at scale** in AWS and placed in accessible buckets, avoiding massive data moves
- Ingest, archival, validation, processing, etc. can **scale dynamically** based on incoming data streams, reprocessing needs, etc.
- **Entire petabyte scale archive** is directly accessible, with no transfer time or costs, to science users in the same region for longtime series or multiproduct use
- **Data processing, transformation, and analysis services** can be spun up, NASA funded or completely independently, leveraging the data with scalable compute and cost and access-managed output targets.



HOW DO USERS *USE* THIS DATA?



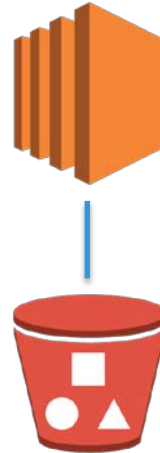
Data Access Use Cases



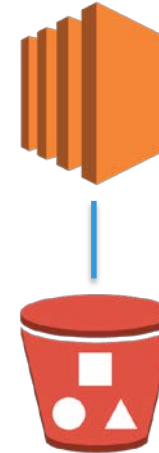
***Traditional file based
Data Access***



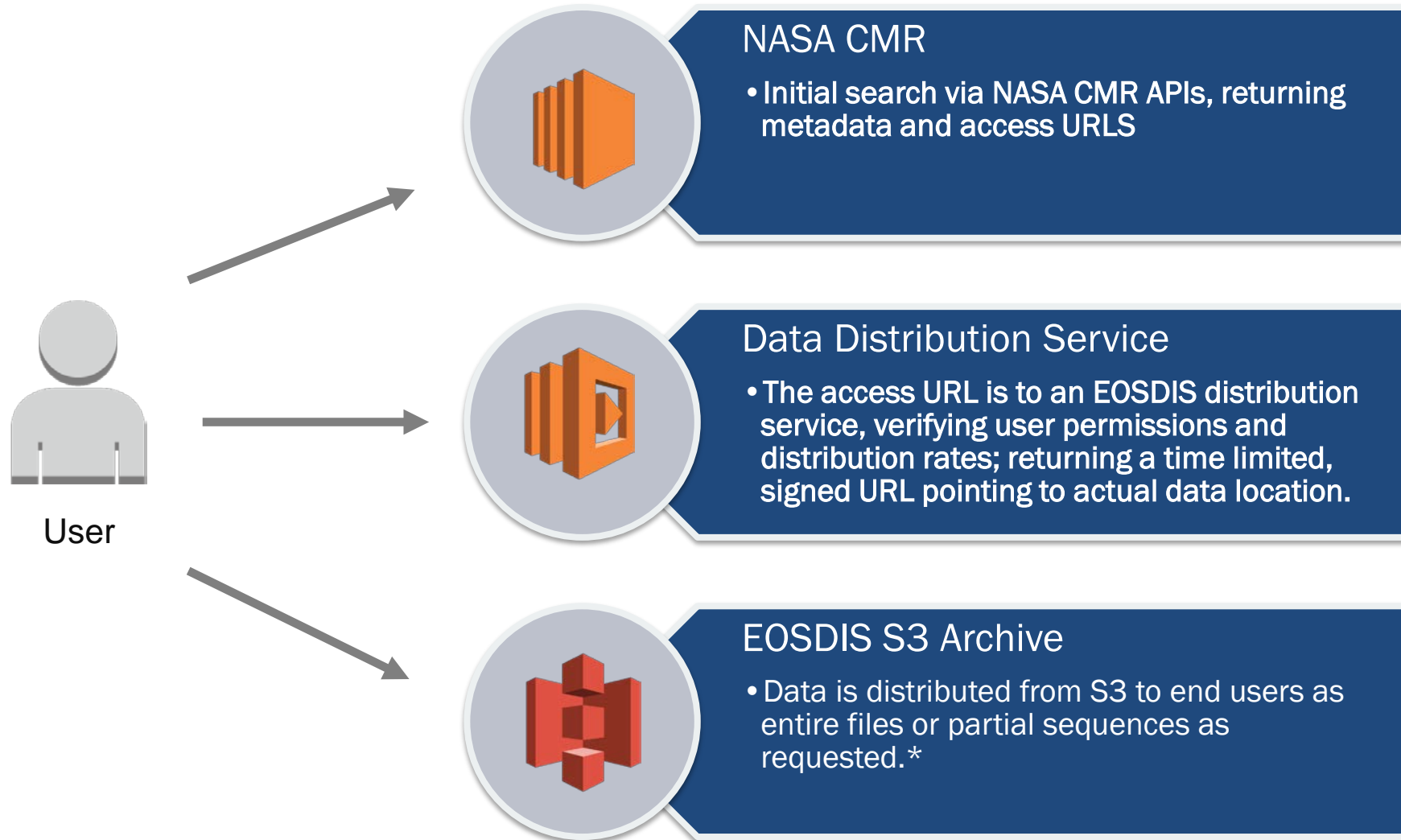
***Distribution from
Access Services***



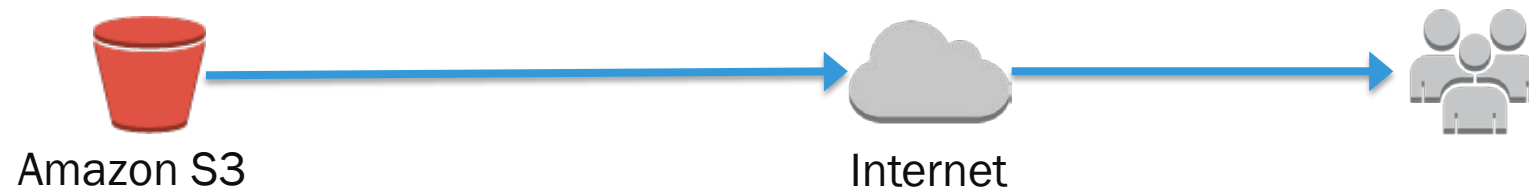
***Computing and Analysis
Near the Data***



Basic Data Access



Basic Amazon S3 egress



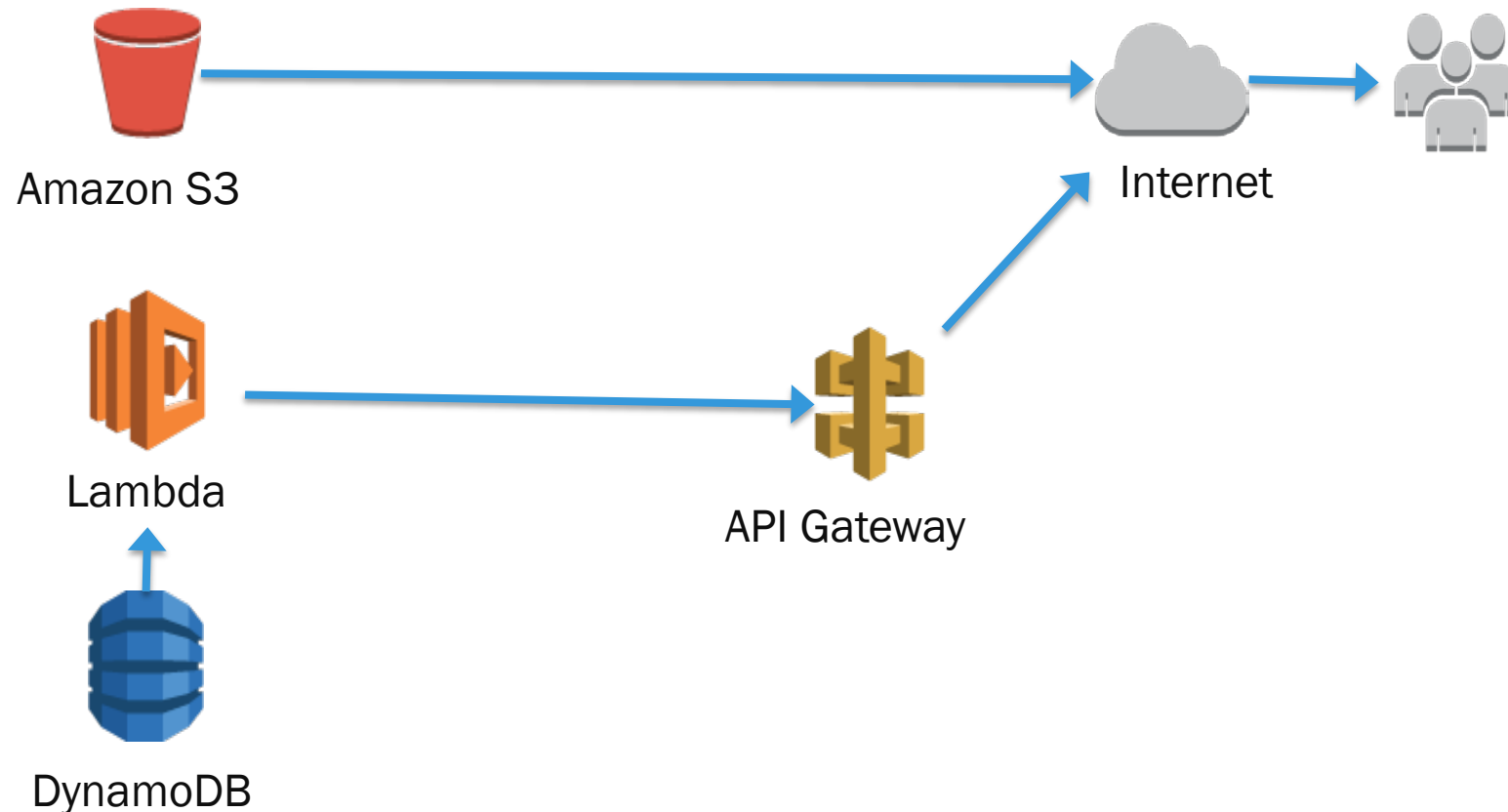
Amazon S3 with CloudFront



Amazon S3 through AWS Direct Connect to on-premises distribution pipe



Request limiting using Lambda and API Gateway



Egress costs range more than
13x across those models

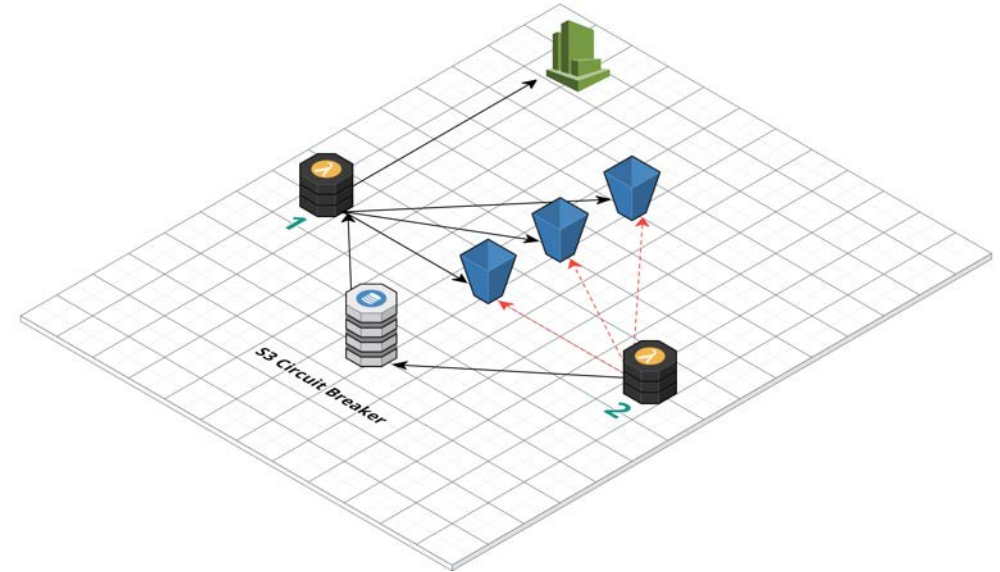
Egress costs are a **big** deal...
...but they weren't our only issue...

Hard cost controls are essential

- The Anti-Deficiency Act (ADA) disallows unbounded costs
- We needed a means of **absolutely** limiting egress costs

Circuit Breaker Conceptual design

- Lambda 1: Calculate Amazon S3 egress
 - Watch each bucket's "Bytes Downloaded" via CloudWatch
 - Post totals
- Lambda 2: Break the circuit (if needed)
 - If total from first billing period to now exceeds our threshold...
 - ...lock down Amazon S3 bucket policy

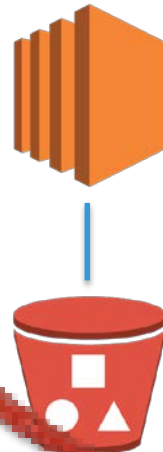


Data Access Use Cases

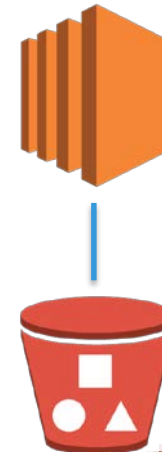
*Traditional file based
Data Access*



*Distribution from
Access Services*



*Computing and Analysis
Near the Data*



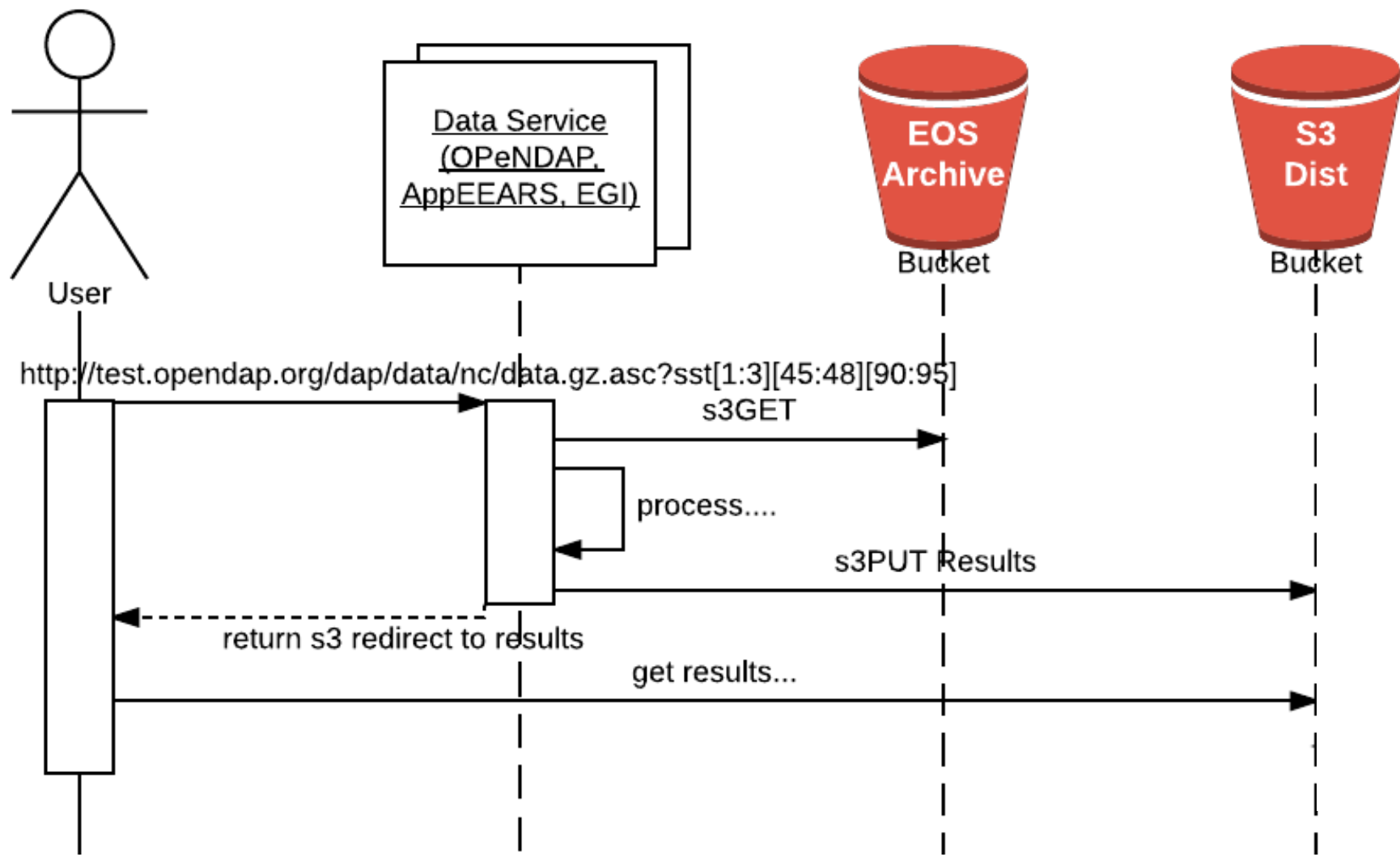
Distribution from Access Services

Example access services

- OPeNDAP access
- OGC WxS Implementations
- Format transformations (repackaging)
- Reprojection, mosiacing, etc

S3 is a Distribution Mechanism

- Mark Korver @ AWS



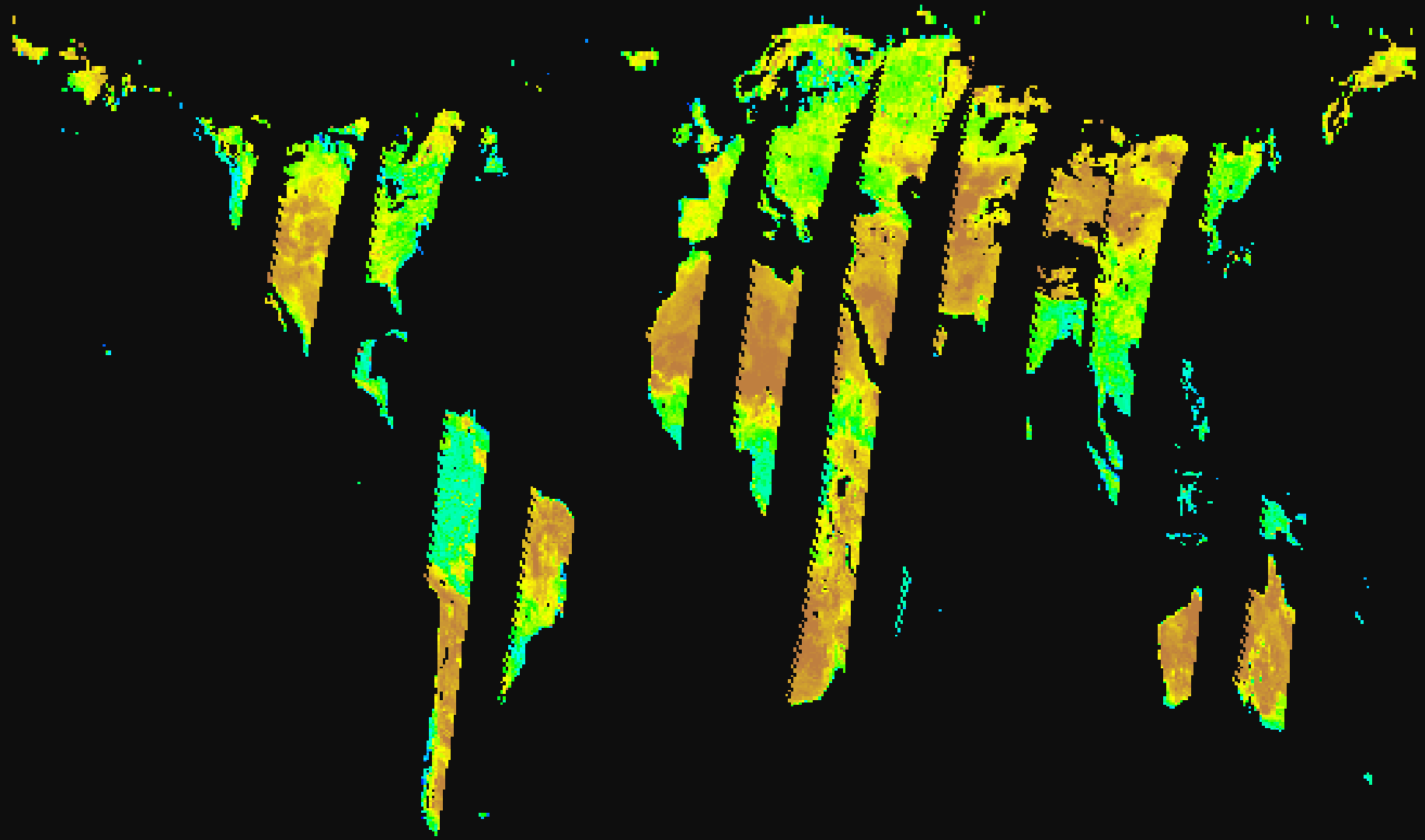
Favor Re-architecture over “just getting into the cloud”

- Natural Inflection Point
- Managed Services
- Opportunity for Innovation
- Better leverage cloud cost models

Psycho-Social

GO HANDS-ON QUICKLY





LOOKING FORWARD

What we're working on now...

- Efficient data services access and distribution
- Cost effective large archive storage
- Data disaster recovery and preservation approaches
- Third party cloud native data use at scale
- Expanding the paradigm of an established community

Here's where we want help

Ways to Compute Near the Data?

- EC2 Instances mounting data via **yas3fs**
- Jupyter Notebooks with **s3contents** or **boto3**
- Serverless implementations with **SNS/SQS** and containerized code (**ECS**)
- Managed solutions like **Athena** or **RedShift Spectrum**

Call for help

- What can we do to make it easier for you use to use the data in the cloud?
- What are barriers to you using the cloud for processing at scale?
- What kind of sample code, documentation, reference implementations, etc. would help you?
- Would you use / want to use the data as is on S3 or via some other access API?
- Right now discovery and getting a URL to the data goes through the CMR. Are there other ways you'd like to be able to find and access the data? Flat file catalogs?

Thank you!

Majority of code discussed today is Open Source:

<https://github.com/nasa>

Dan Pilone // dan@element84.com

This material is based upon work supported by the National Aeronautics and Space Administration under Contract Number **NNG15HZ39C**.

Raytheon

Acronym List

- AWS – Amazon Web Services
- CNES - Centre national d'études spatiales
- DAAC – Distributed Active Archive Center
- EC2 – Elastic Compute Cloud
- EED2 – EOSDIS Evolution and Development 2
- FOSS4G – Free and Open Source for Geospatial
- GOES-R - Geostationary Operational Environmental Satellite
- IAM – Identity and Access Management
- JSON – JavaScript Object Notation
- OGC – Open Geospatial Consortium
- OPeNDAP - Open-source Project for a Network Data Access Protocol
- MRF – Metadata Raster Format
- NRC – National Resource Council
- S3 – Simple Storage Service
- S3 IA – Simple Storage Service Infrequent Access
- SNS – Simple Notification Service
- SQS – Simple Queuing Service
- URS – User Registration Service