FROM ARDS TO AODS: FUTURE OF ANALYTICS FOR EARTH OBSERVATIONS

Dr. Rahul Ramachandran*, Kaylin Bugbee, Manil Maskey, Chris Lynnes

*Manager | Inter-Agency Implementation and Advanced Concepts Team (IMPACT)
Senior Research Scientist | Earth Science Branch (ST11)
Marshall Space Flight Center/NASA
Huntsville, Alabama 35812, USA

Acknowledgements: Hook Hua, George Change (JPL)
Aimee B (Dev Seed) + MAAP Team

IGARSS 2019, Yokohama, Japan

Outline

- Concept Definitions
  - Existing ARD Definitions
  - Missing Component – Big Data Perspective
  - AODS Definition
- Notional Analytics Architecture for the Future
  - Background on NASA’s Data and Information Systems
  - Drivers
  - Analytics Modes
  - Prototype: MAAP Project
  - Summary

Concept Definitions

Analysis Ready Data

“Consistently processed to the highest scientific standards and level of processing required for direct use in monitoring and assessing landscape change” - U.S Landsat ARD

“Time-series stacks of overhead imagery that are prepared so that a user can analyze the data without having to pre-process the imagery themselves” - Planet

“Analysis Ready Data for Land (CARD4L) are satellite data that have been processed to a minimum set of requirements and *organized into a form* that allows immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets” - CEOS

- Lower barrier to data use
- Focus on (pre)processing needs of users (subsetting, atmospheric corrections, co-registration, and sensor calibration)

- Not new - maps directly to NASA’s level 3 processing definition of providing “variables mapped on uniform space-time grid scales, usually with some completeness and consistency”

Missing Component: Big Data Perspective

- **Data engineering**, or restructuring, is important from the big data analytics perspective because restructuring easily enables mapping data into analytics tools

- Characteristics of big data analytics tools :
    - Exploit parallelism
    - Minimize data movement between compute nodes
    - Utilize commodity computing to reduce cost
    - Scale to handle growth in data volumes.

Analytics Optimized Data Stores (AODS)

- Addresses both the preprocessing requirements as well as the data engineering requirement.
  - Focus on restructuring and distributing the data in order to exploit parallelism and minimize data movement during computing.

- Characteristics:
  - Serve specific analytic goals
  - Preprocessed with the right steps to meet analytic goals
  - Engineered (data) to enable fast access patterns tuned to the expected analyses
  - Can be subsets of a total archive holding, preprocessed using consistent and well-known algorithms that cover a spectrum of steps

Notional Analytics Architecture

Earth Science – NASA's Strategic Goal

This ability to *observe our planet comprehensively* matters to each of us, on a daily level. Earth information—for use in Internet maps, daily weather forecasts, land use planning, transportation efficiency, and agricultural productivity, to name a few—is central to our lives, providing substantial contributions to our economies, our national security, and our personal safety. It helps ensure we are a thriving society. - NRC, 2018

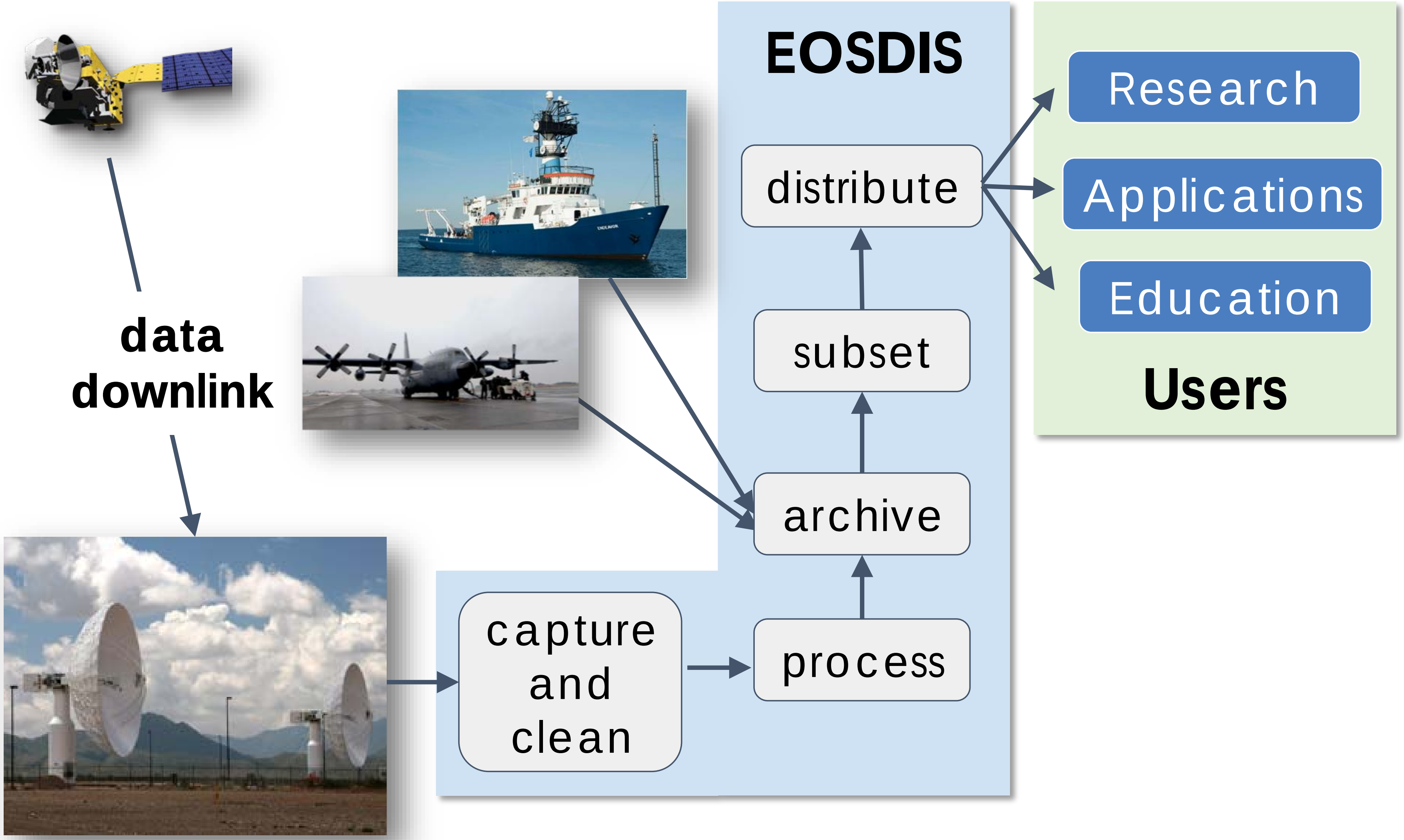
NASA's Strategic Goal 1.1:
"Understand The Sun, Earth, Solar System, And Universe."

Earth Science Data System Program

The Earth Science Data System Program is an essential *component of the Earth Science Division* and is responsible for:

- Actively *managing NASA's Earth science data* (Satellite, Airborne, and Field).
- *Developing unique* data system *capabilities* optimized to support *rigorous science investigations* and interdisciplinary research.
- *Processing* (and reprocessing) instrument data to create high quality long-term Earth science data records.
- Upholding NASA's policy of *full and open sharing of all data, tools, and ancillary information* for all users.
- Engaging members of the Earth science community in the *evolution of data systems*.

Earth Observing System Data and Information System (EOSDIS)



data downlink

EOSDIS

capture and clean

process

archive

subset

distribute

Research

Applications

Education

Users

Earth Observing System Data and Information System (EOSDIS)

EOSDIS is managed by the Earth Science Data and Information System (ESDIS) Project at GSFC and includes the following major core components:

Science Investigator-led Processing Systems (SIPS)

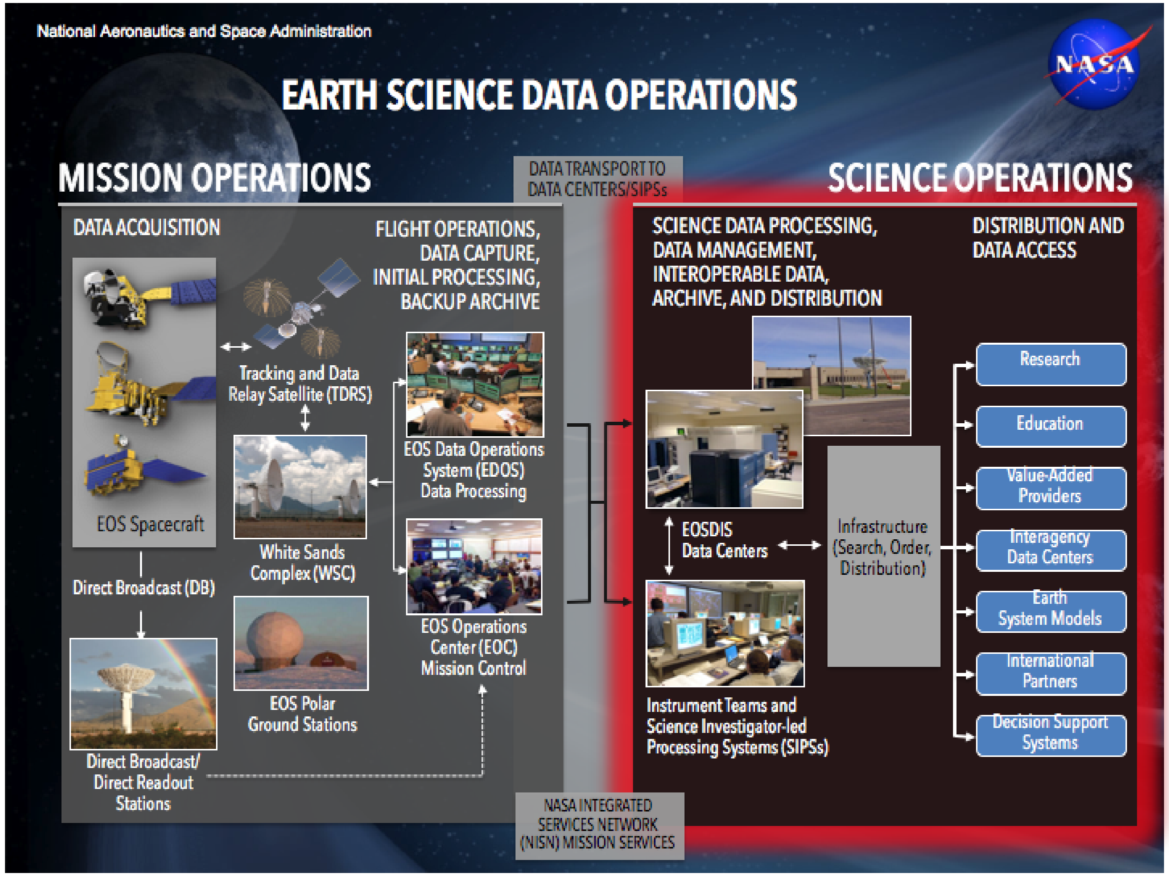
- Perform forward processing of standard data products and reprocess data to incorporate algorithm improvements

Distributed Active Archive Centers (DAACs)

- Co-located with centers of science discipline expertise; archive and distribute standard data products produced by the SIPS and others

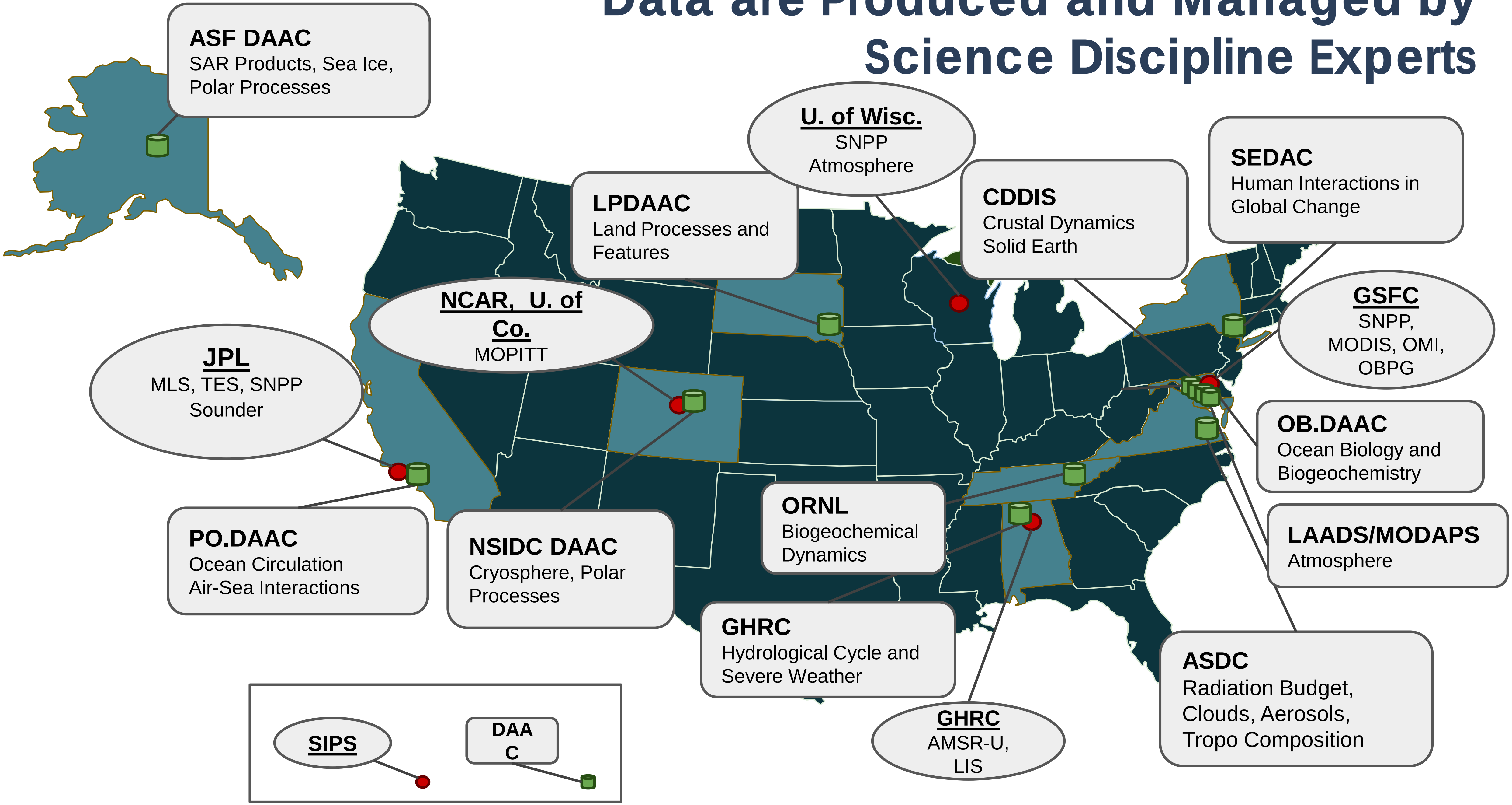
Earthdata and Core Services

- Allows users to search, discover, visualize, refine, and access NASA Earth Observation data. Includes networking and security


National Aeronautics and Space Administration
NASA
EARTH SCIENCE DATA OPERATIONS
MISSION OPERATIONS
DATA TRANSPORT TO DATA CENTERS/SIPSs
SCIENCE OPERATIONS
DATA ACQUISITION
FLIGHT OPERATIONS, DATA CAPTURE, INITIAL PROCESSING, BACKUP ARCHIVE
EOS Spacecraft
Tracking and Data Relay Satellite (TDRS)
White Sands Complex (WSC)
EOS Data Operations System (EDOS) Data Processing
EOS Operations Center (EOC) Mission Control
EOS Polar Ground Stations
Direct Broadcast (DB)
Direct Broadcast/ Direct Readout Stations
NASA INTEGRATED SERVICES NETWORK (NISN) MISSION SERVICES
SCIENCE DATA PROCESSING, DATA MANAGEMENT, INTEROPERABLE DATA, ARCHIVE, AND DISTRIBUTION
DISTRIBUTION AND DATA ACCESS
EOSDIS Data Centers
Infrastructure (Search, Order, Distribution)
Instrument Teams and Science Investigator-led Processing Systems (SIPSs)
Research
Education
Value-Added Providers
Interagency Data Centers
Earth System Models
International Partners
Decision Support Systems

**Red highlight indicates EOSDIS boundary.*

Data are Produced and Managed by Science Discipline Experts



ASF DAAC
SAR Products, Sea Ice, Polar Processes

U. of Wisc.
SNPP Atmosphere

LPDAAC
Land Processes and Features

CDDIS
Crustal Dynamics Solid Earth

SEDAC
Human Interactions in Global Change

NCAR, U. of Co.
MOPITT

JPL
MLS, TES, SNPP Sounder

GSFC
SNPP, MODIS, OMI, OBPG

OB.DAAC
Ocean Biology and Biogeochemistry

PO.DAAC
Ocean Circulation
Air-Sea Interactions

NSIDC DAAC
Cryosphere, Polar Processes

ORNL
Biogeochemical Dynamics

LAADS/MODAPS
Atmosphere

GHRC
Hydrological Cycle and Severe Weather

GHRC
AMSR-U, LIS

ASDC
Radiation Budget, Clouds, Aerosols, Tropo Composition

SIPS ●

DAA C 🟩

Extensive Data Collection

Started in the 1990s, EOSDIS today has 11,000+ data types (collections)

- Cover & Usage
- Surface temperature
- Soil moisture
- Surface topography

Land

- Surface temperature
- Surface wind fields & heat flux
- Surface topography
- Ocean color

Ocean

- Winds & Precipitation
- Aerosols & Clouds
- Temperature & Humidity
- Solar radiation

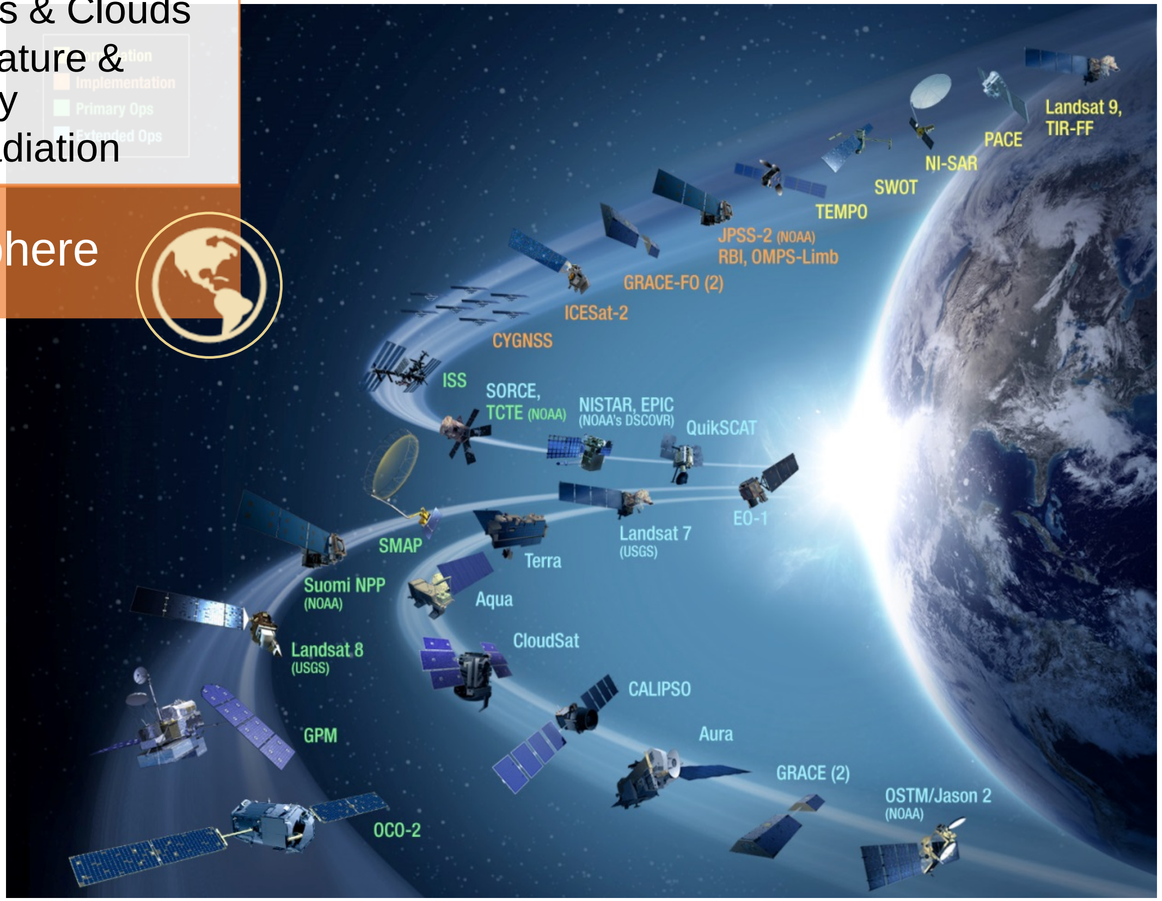
Atmosphere

- Population & Land Use
- Human & Environmental Health
- Ecosystems

Human Dimensions

- Sea/Land Ice & Snow Cover

Cryosphere



NASA’s Earth Science Data System in 2018

EOSDIS currently has over
27 Petabytes
of accessible Earth science data

Easy access and discovery of data to over
12,500 unique data products

… of which 95% of granule searches complete in less than
1 Second

EOSDIS delivered over
1.6 Billion data products
to over 4.1 Million
users from around the world

33,000 Data Collections
in the Common Metadata Repository (CMR)

EOSDIS also delivers near-real-time products in under 3 hours
from observation …

Over 330,000 users
have registered with EOSDIS to date

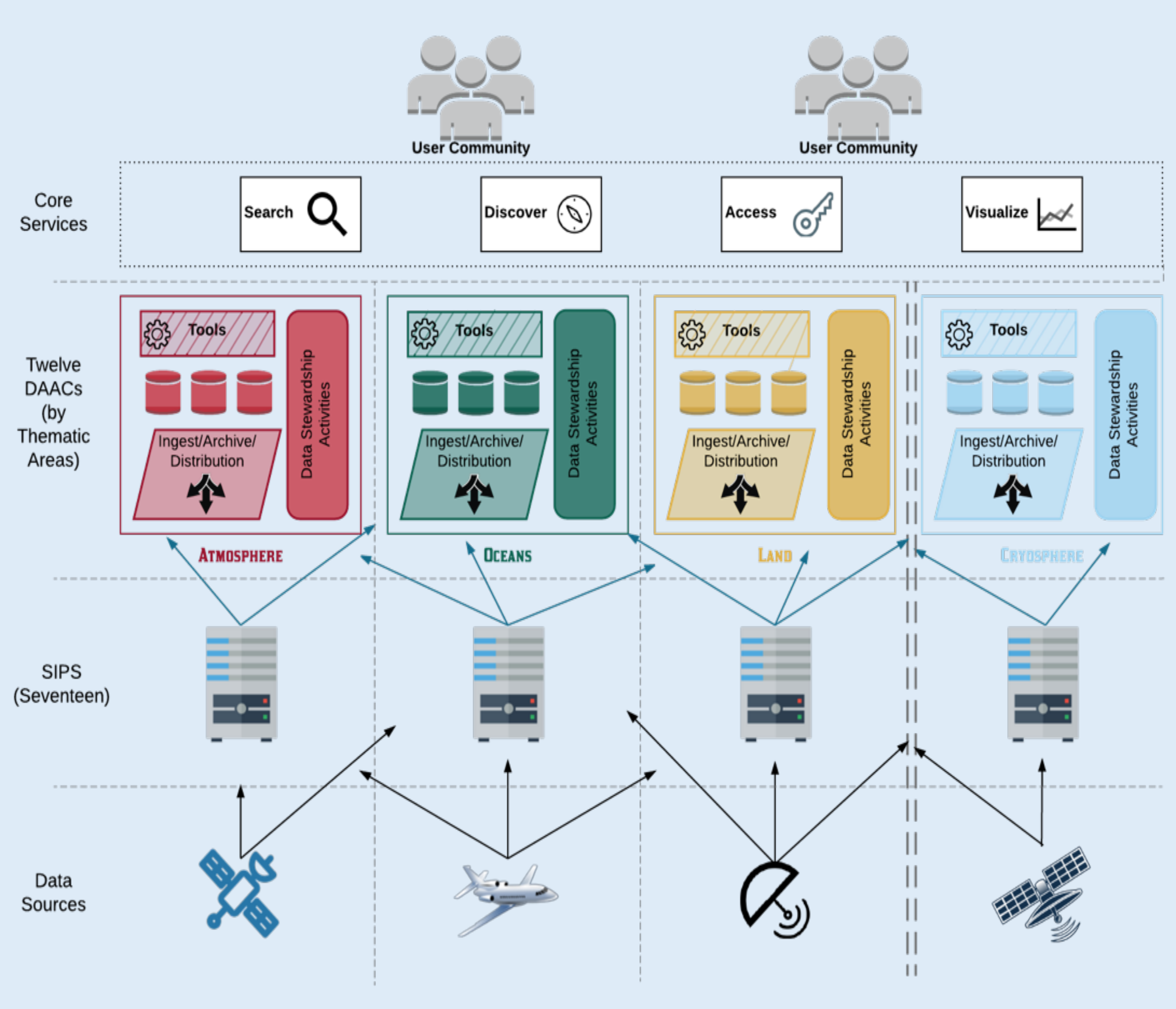
And Over 380 Million
data granules

ACSI American Customer Satisfaction Index®

American Customer Satisfaction Index (ACSI) survey scoring 79 from over 4,000 respondents

https://earthdata.nasa.gov/about/system-performance

Current EOSDIS Architecture



Strengths

- Discipline specific support and tools (DAAC data)
- Optimized for discipline specific archive, search and distribution
- Easily add new data products
- Supports millions of users

Weakness

- Uneven service and performance
- Significant interface coordination
- Limited on-demand product generation
- ***Fragmentation – duplication of services, software and storage***

NASA Earth Science

Missions: Present through 2023

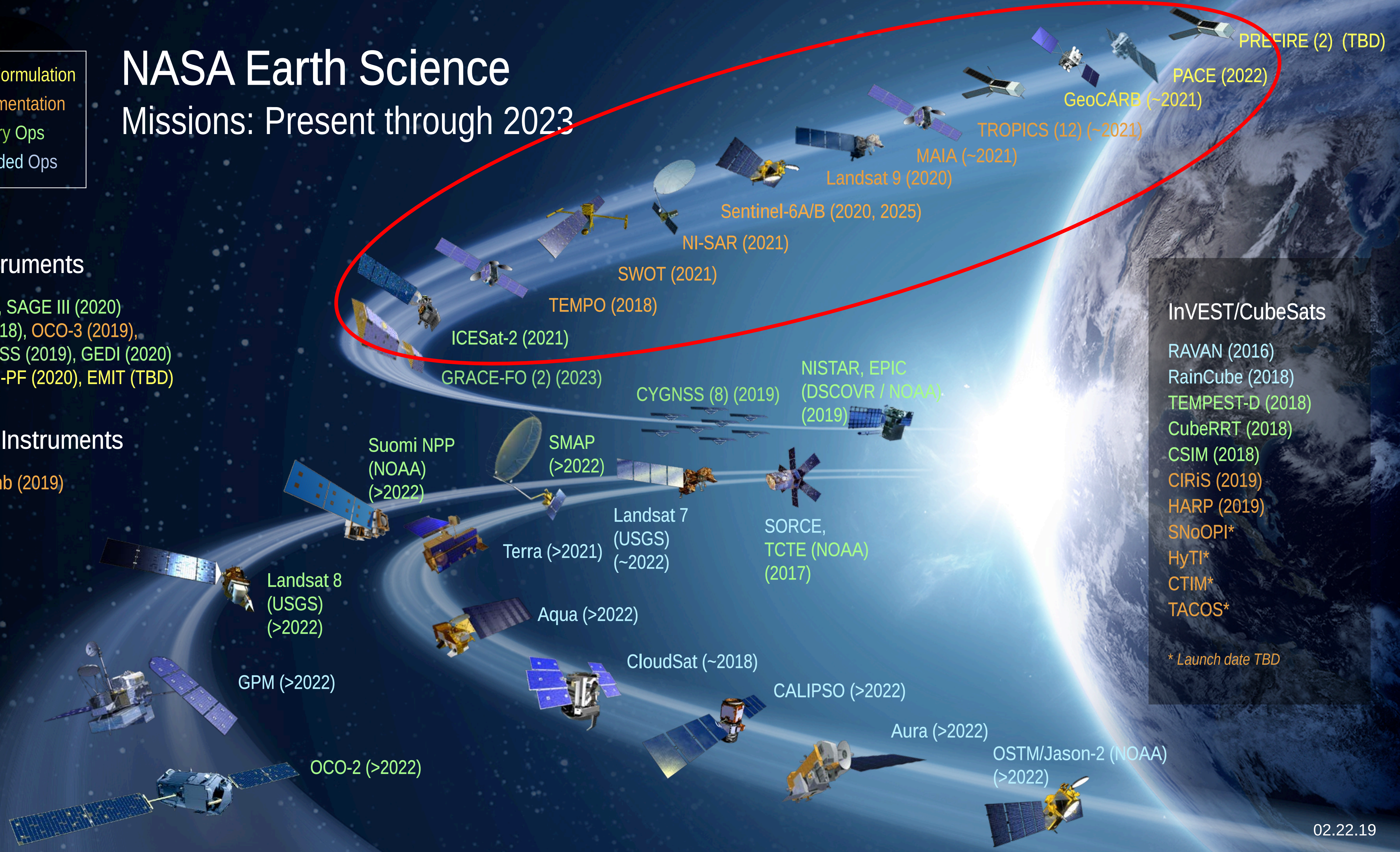
- 🟨 (Pre)Formulation
- 🟧 Implementation
- 🟩 Primary Ops
- 🟦 Extended Ops

ISS Instruments

LIS (2020), SAGE III (2020)
TSIS-1 (2018), OCO-3 (2019),
ECOSTRESS (2019), GEDI (2020)
CLARREO-PF (2020), EMIT (TBD)

JPSS-2 Instruments

OMPS-Limb (2019)



InVEST/CubeSats

RAVAN (2016)
RainCube (2018)
TEMPEST-D (2018)
CubeRRT (2018)
CSIM (2018)
CIRiS (2019)
HARP (2019)
SNoOPI*
HyTI*
CTIM*
TACOS*

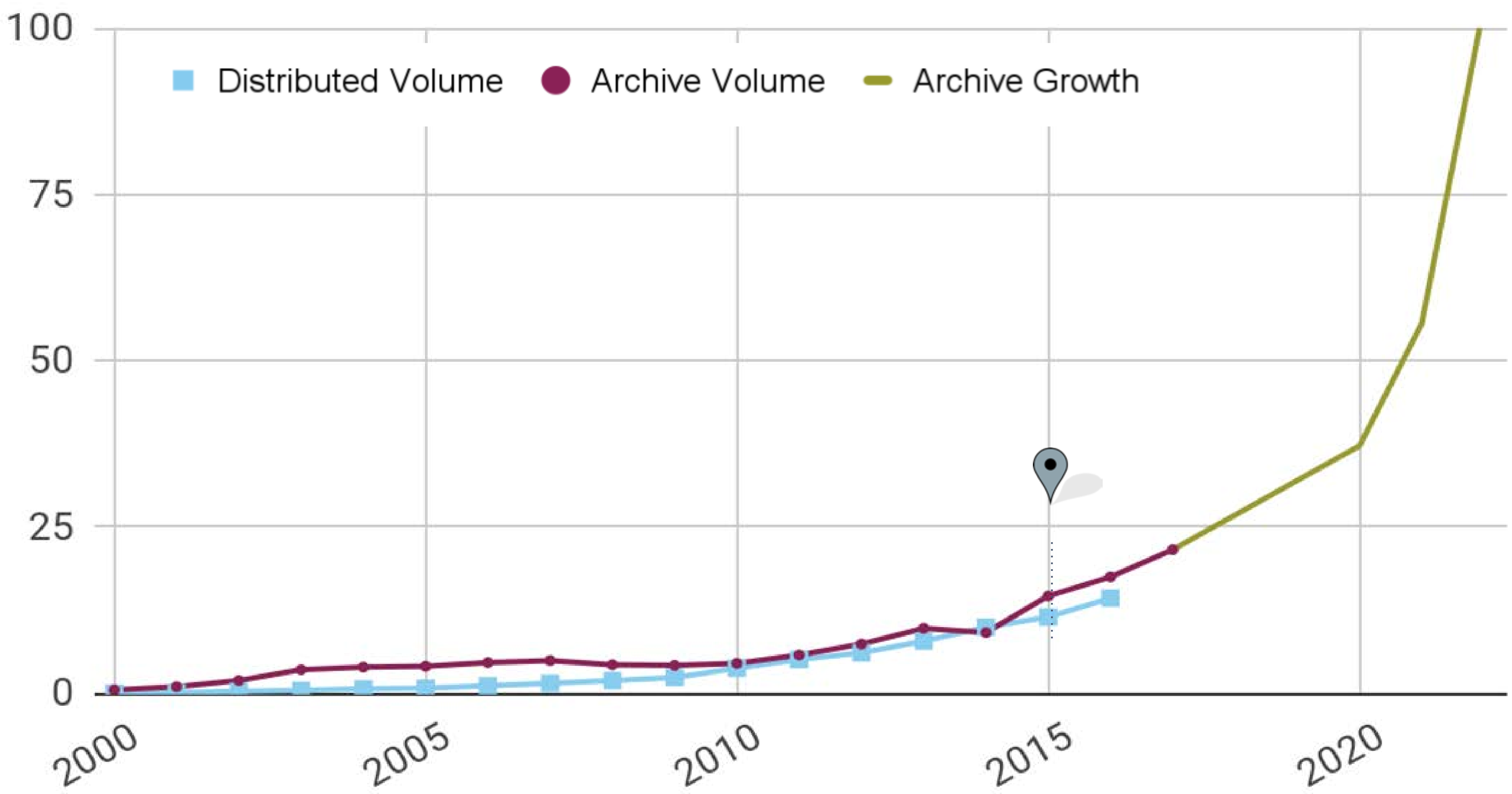
** Launch date TBD*

02.22.19

EOSDIS Data System Evolution

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

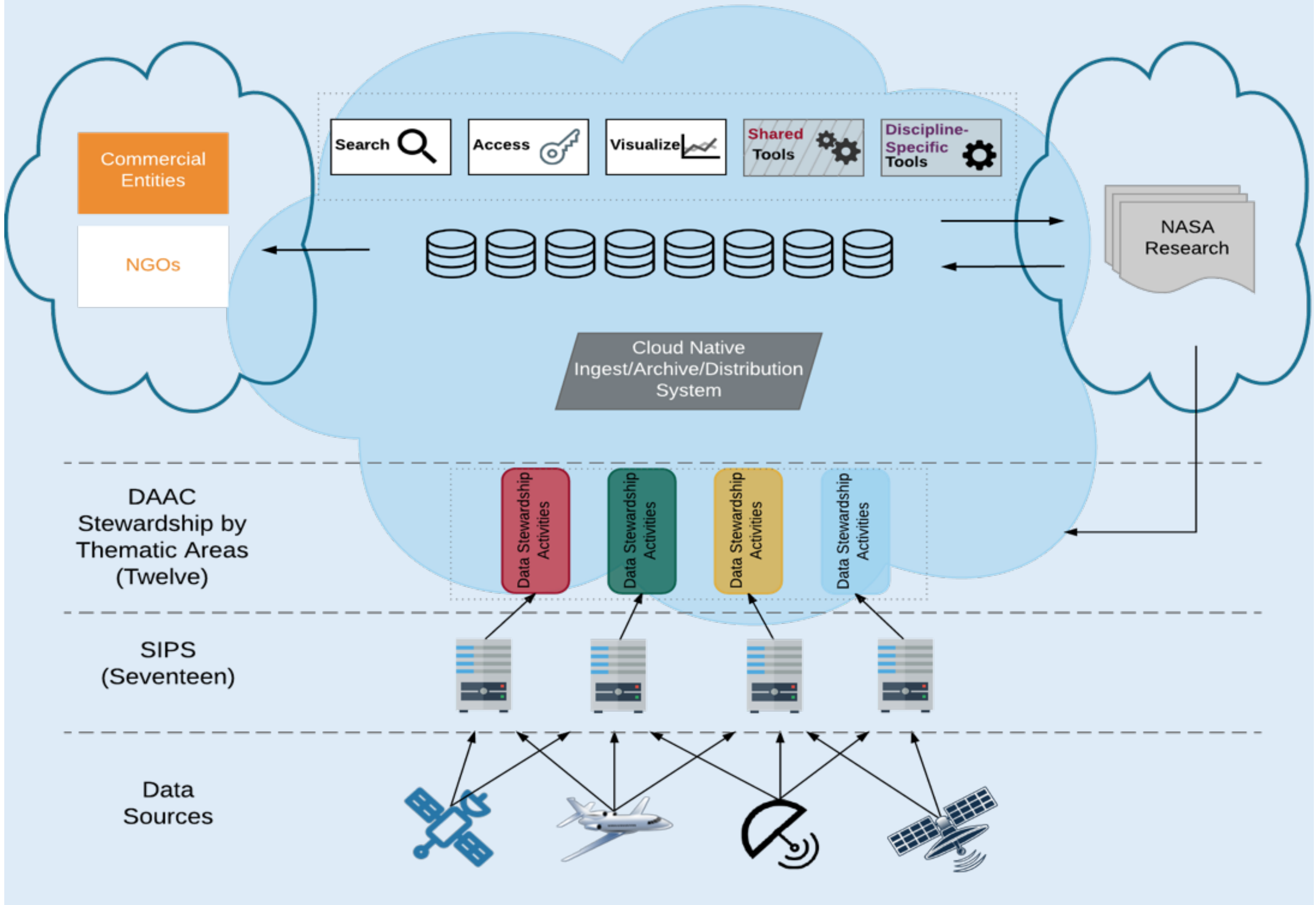
EOSDIS is developing open source cloud native software for reuse across the agency, throughout the government and for any other user.



***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”*

Conceptual Cloud Architecture - 2021

Open Science = Open Data + Open Source Software + Open Services



Commercial Entities
NGOs
Search
Access
Visualize
Shared Tools
Discipline-Specific Tools
NASA Research
Cloud Native Ingest/Archive/Distribution System
DAAC Stewardship by Thematic Areas (Twelve)
Data Stewardship Activities
Data Stewardship Activities
Data Stewardship Activities
Data Stewardship Activities
SIPS (Seventeen)
Data Sources

Advantages

- Scalability
- Processing next to data for anyone
- Optimized for multidisciplinary research

Challenges

- Develop coordination and documentation
- Cost management
- Business processes, security and skillsets
- Vendor lock-in

Moving towards the cloud: Cumulus

Lightweight, cloud-native framework for data ingest, archive, distribution and management

A lightweight framework consisting of:

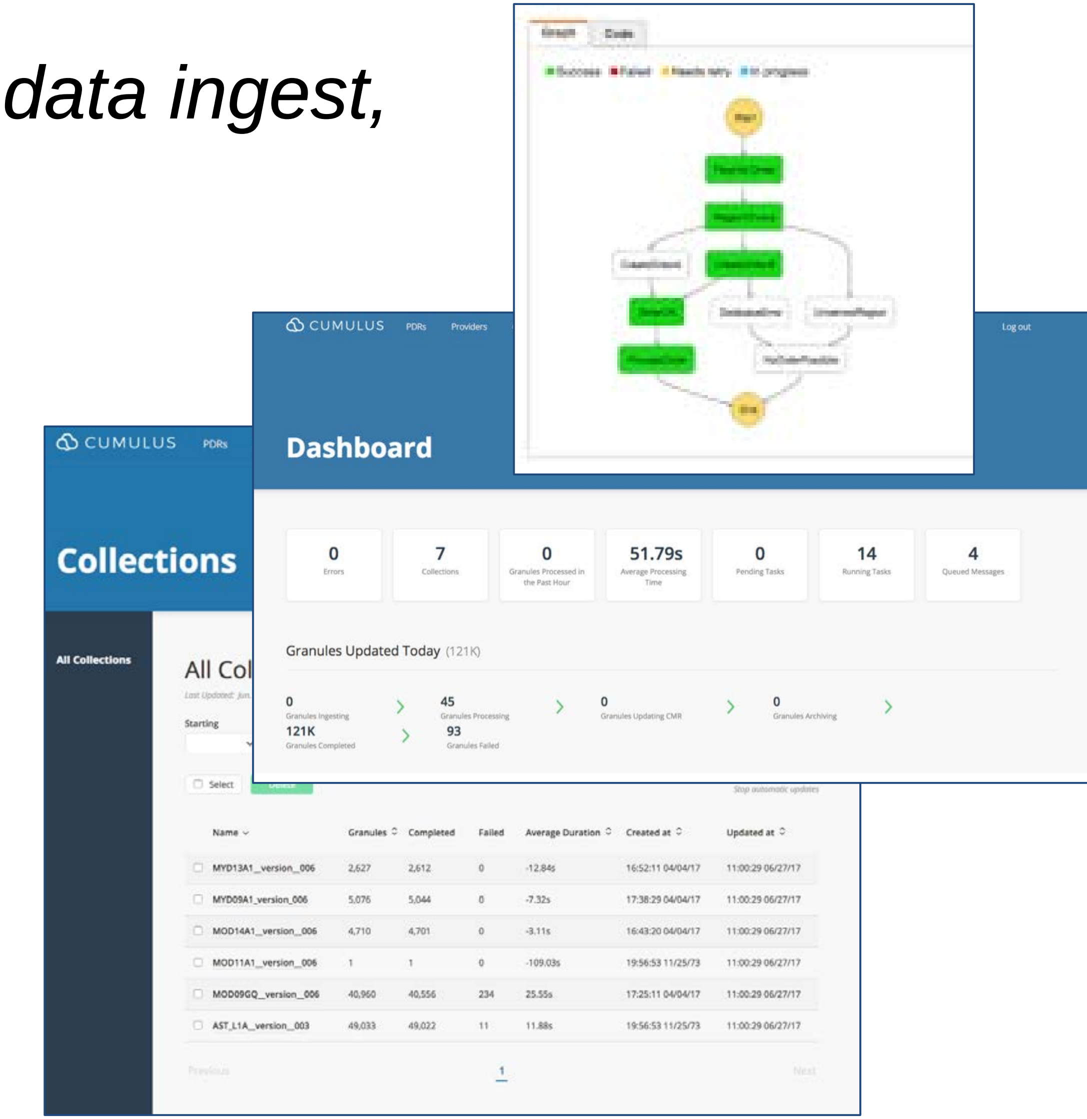
Tasks a discrete action in a workflow, invoked as a Lambda function or EC2 service, common protocol supports chaining

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



Dashboard

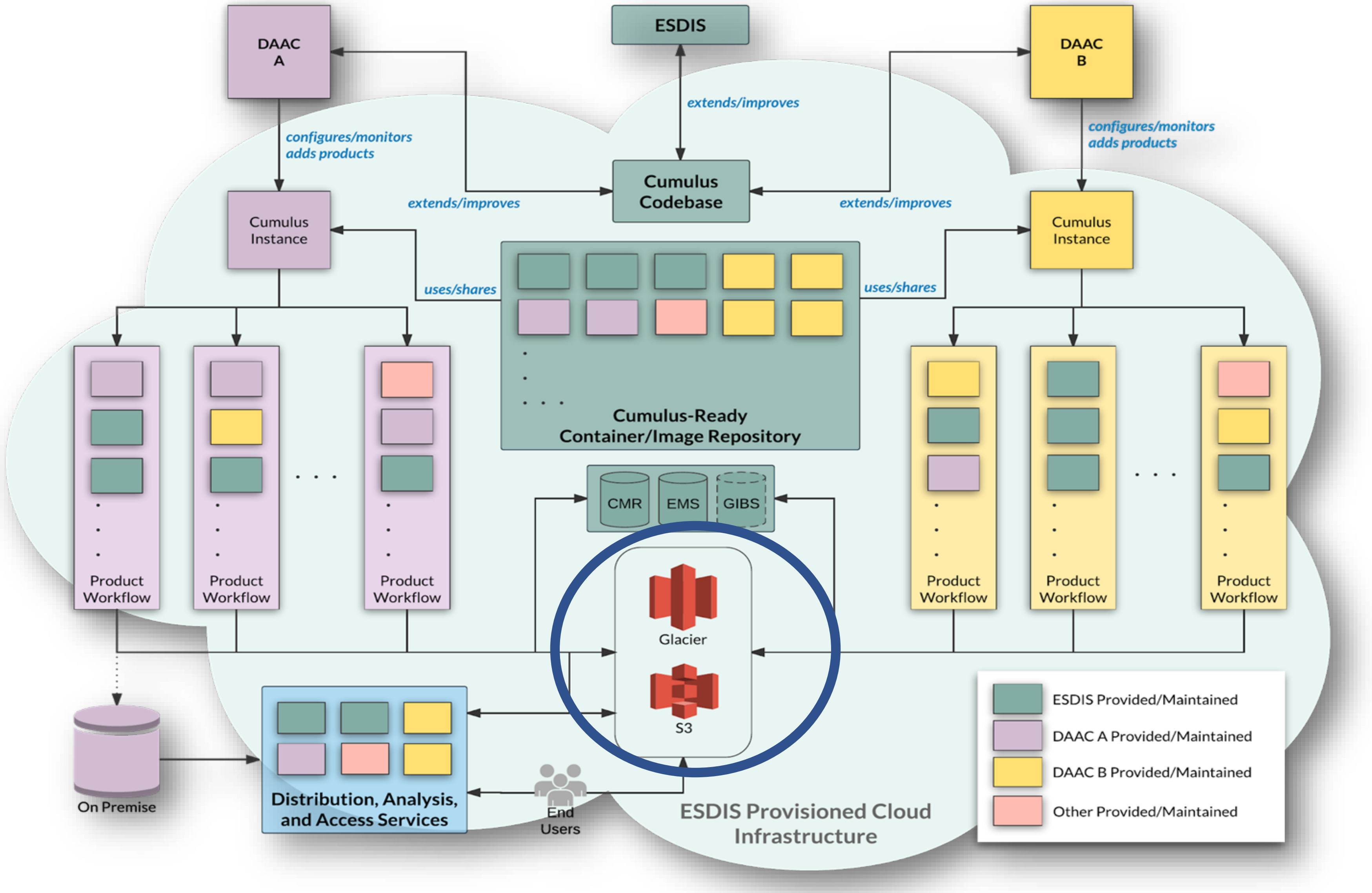
0	7	0	51.79s	0	14	4
Errors	Collections	Granules Processed in the Past Hour	Average Processing Time	Pending Tasks	Running Tasks	Queued Messages

Granules Updated Today (121K)

0 Granules Ingesting
121K Granules Completed
45 Granules Processing
93 Granules Failed
0 Granules Updating CMR
0 Granules Archiving

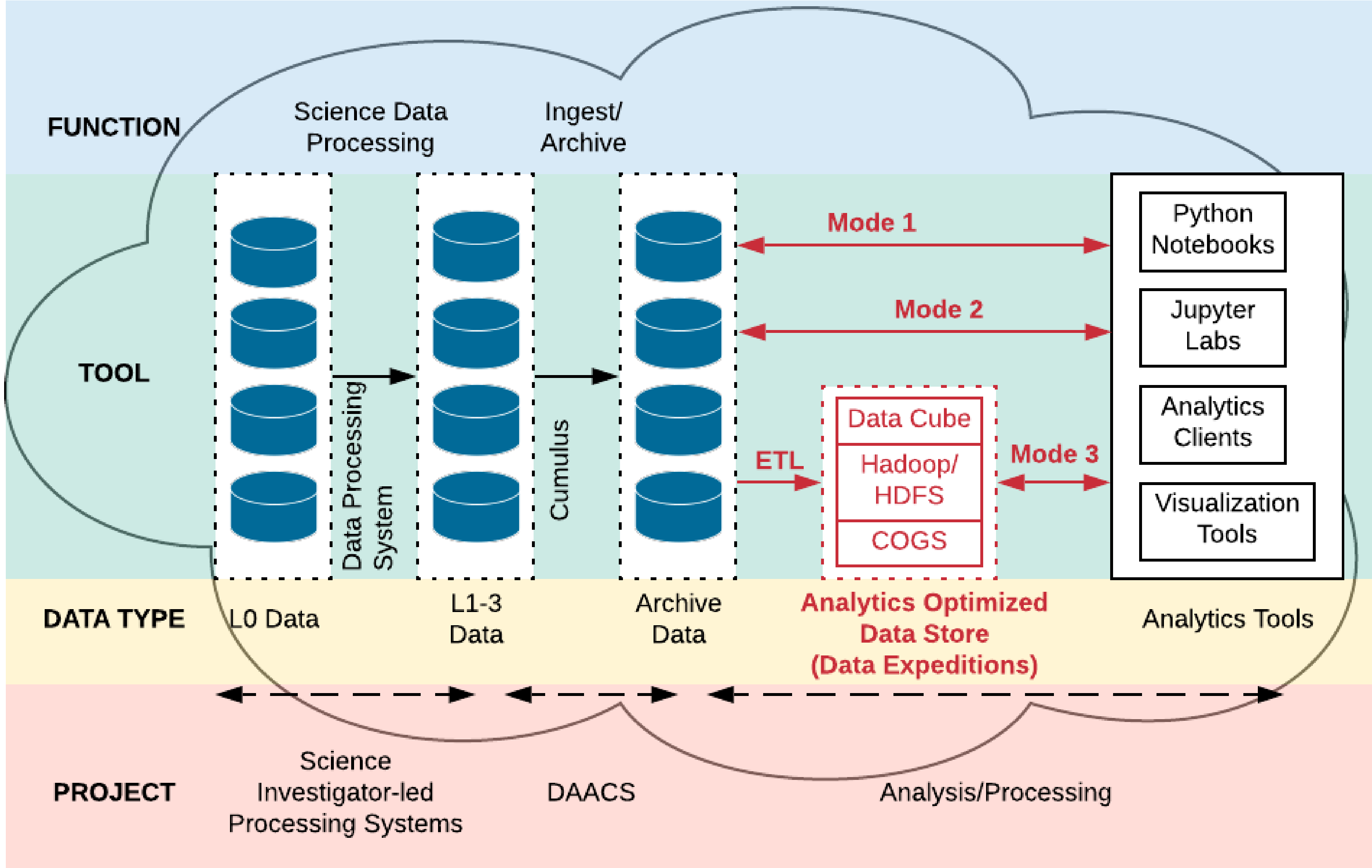
Name	Granules	Completed	Failed	Average Duration	Created at	Updated at
MYD13A1__version__006	2,627	2,612	0	-12.84s	16:52:11 04/04/17	11:00:29 06/27/17
MYD09A1_version_006	5,076	5,044	0	-7.32s	17:38:29 04/04/17	11:00:29 06/27/17
MOD14A1__version__006	4,710	4,701	0	-3.11s	16:43:20 04/04/17	11:00:29 06/27/17
MOD11A1__version__006	1	1	0	-109.03s	19:56:53 11/25/73	11:00:29 06/27/17
MOD09GQ__version__006	40,960	40,556	234	25.55s	17:25:11 04/04/17	11:00:29 06/27/17
AST_L1A__version__003	49,033	49,022	11	11.88s	19:56:53 11/25/73	11:00:29 06/27/17

Data is on the cloud: Now What?



How do we maximize the use of NASA data?

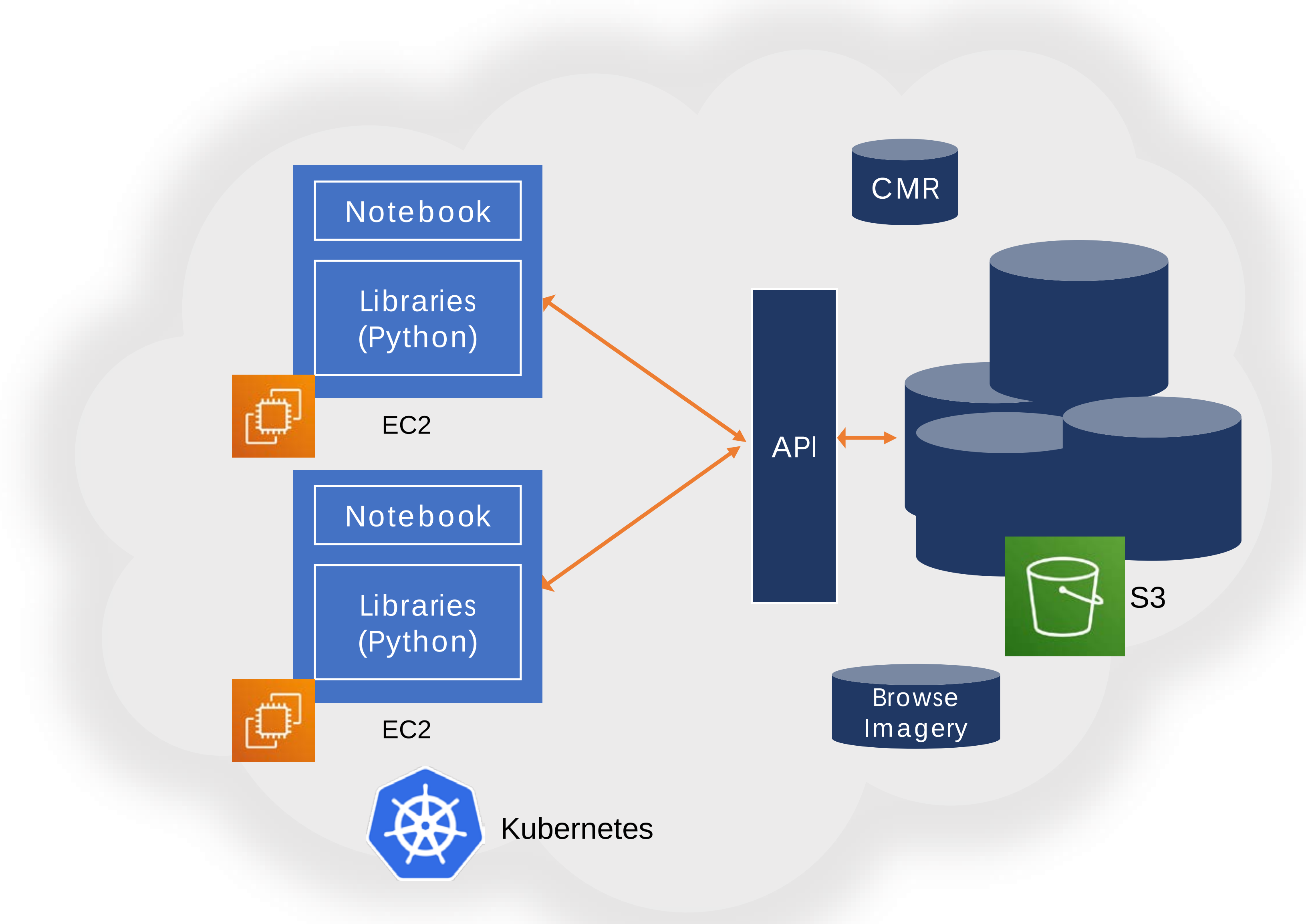
Conceptual Architecture Supporting Analytics



Three Modes of Data Interaction:

- **Small Data/Interactive analysis**
- **Batch Processing**
- **Ephemeral AODS enabling big data analytics**

Mode 1: Small Amounts of Data/Interactive Analysis



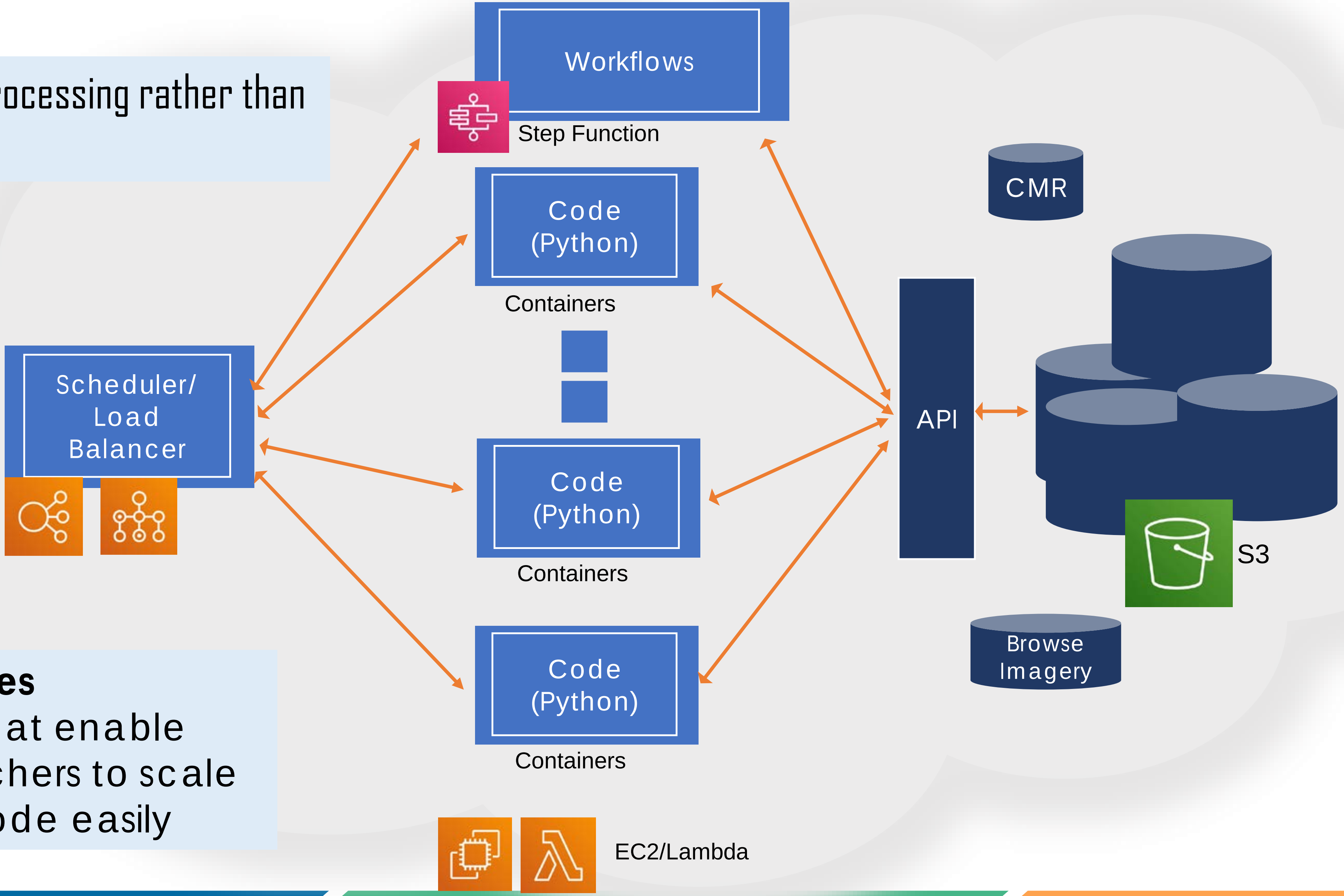
- Moving individual researchers to the cloud
- Shifting the paradigm from downloading files for data analysis

Challenges

- Who provides and pays for the computing environment (eg. Pangeo)
- Researchers need to learn to instantiate their instances and manage their cloud accounts

Mode 2: Scaling Out/Batch Processing

Focus is processing rather than analysis

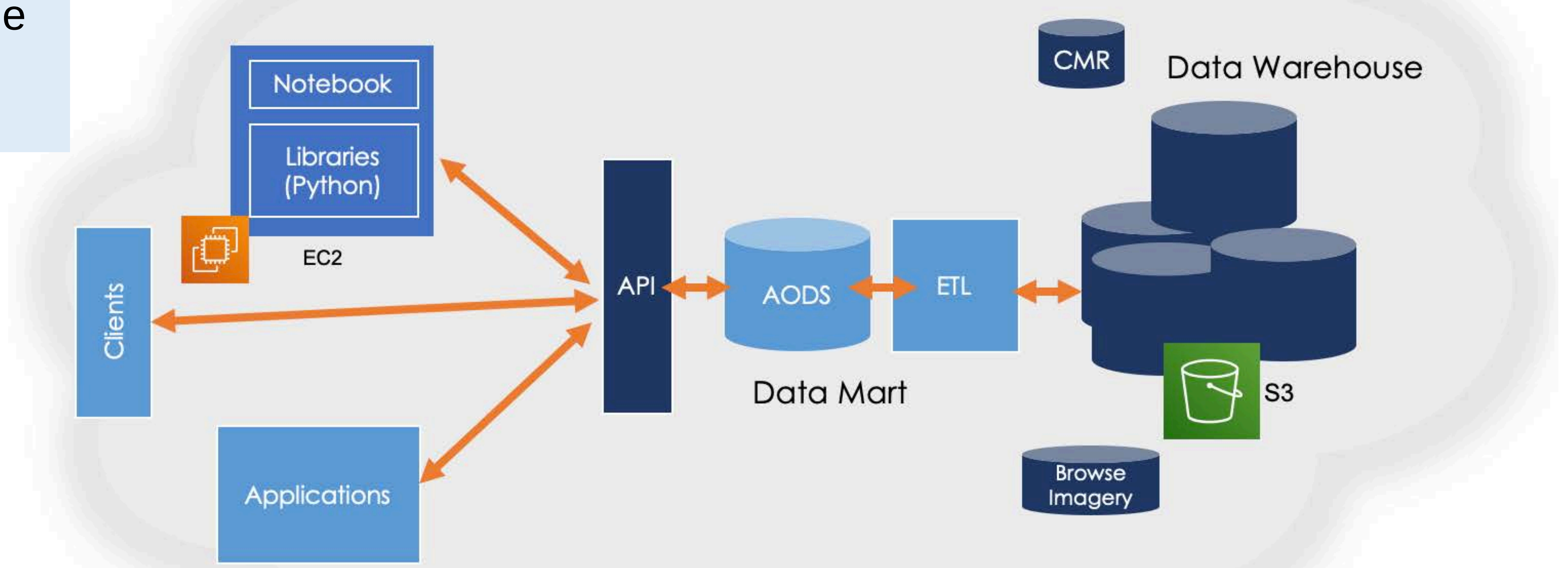


Challenges

- Tools that enable researchers to scale their code easily

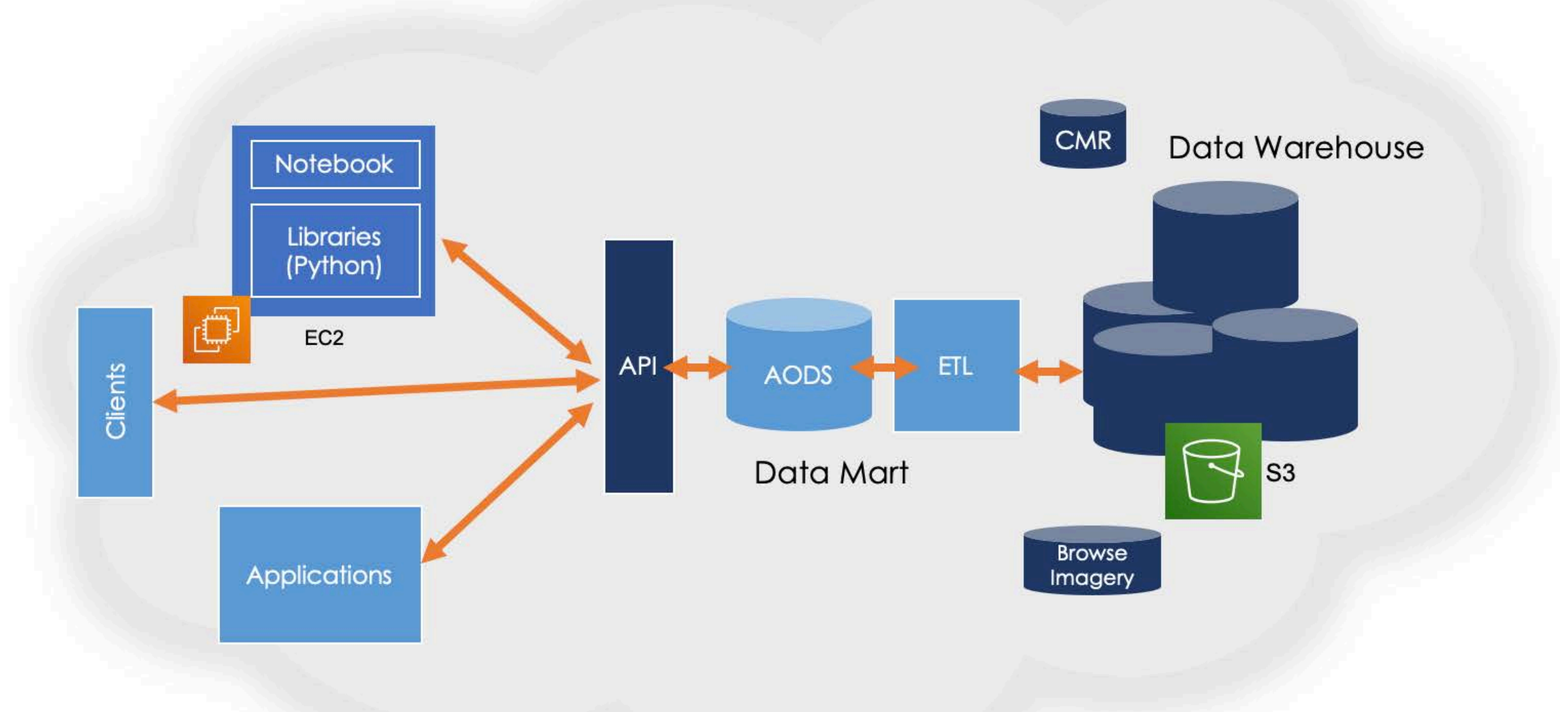
Mode 3: Big Data/Interactive Analysis

Enable researchers to iteratively query large amounts of data to perform an analysis.



- Akin to data marts but for the Earth science domain
- AODS can be constructed and an analytics software stack spun up to support the researchers' analysis (can be ephemeral)
- Data expeditions - “time-bound” analytics projects on answering specific science questions

Mode 3: Big Data/Interactive Analysis



Additional Extract Transform Load (ETL) costs:

- Data curation to meet data expedition needs
- Preprocessed using consistent and well-known algorithms that cover a spectrum of steps
- Data restructuring to enable data parallel computation analysis

Multi-Mission Algorithm and Analysis Platform (MAAP)

- The MAAP is a virtual environment dedicated to the unique needs of sharing and processing data from relevant field, airborne and satellite measurements related to ESA and NASA missions
  - Jointly managed by ESA and NASA and accessible to designated ESA and NASA scientists.
  - Initially populated with pre-launch and complimentary data from other projects.
- Science focus is to improve the understanding of global terrestrial carbon dynamics & to support algorithm development
- Addresses a need expressed by the science community to more easily share and process data collected by NASA and ESA activities

GEDI
ECOSYSTEM LIDAR

biomass

NISAR
NASA - ISRO SAR Mission

Multi-Mission Algorithm and Analysis Platform (MAAP)

Mode1 Stack:

- Eclipse Che
- Docker/Kubernetes

Mode 2 Stack:

- HySDS

Mode 3 Stack:

- Dask/Zarr
- AWS Athena

Acknowledgements: Hook Hua, George Change (JPL)
Aimee B (Dev Seed)

Summary

- Current ARD definitions are incomplete as they do not encompass the data engineering component needed to exploit parallelism and minimize data movement during compute.
- AODS extends ARD to encompass big data analytics to enable data-driven science.
- Adoption of cloud platforms by the science community provides an opportunity to provide different modes of analytics
- Ephemeral AODSs can be created to support data expeditions focused on addressing specific science goals by analyzing large data volumes.
- Need exists to ***develop a common set of ETL tools that enforces uniformity and transparency*** on preprocessing/data structuring is paramount for scientific analysis

Thank you!

Contact info: rahul.ramachandran@nasa.gov

Slides Attribution:
Kevin Murphy ESDS Program PE, NASA HQ
Andy Mitchell ESDIS Project Manager, NASA GSFC
IMPACT Team, NASA MSFC
MAAP Team, NASA JPL/NASA MSFC [UAH, Development Seed]